PERBANDINGAN METODE KLASIFIKASI *MULTICLASS* UNTUK PEMETAAN ZONA RISIKO COVID-19 DI PULAU JAWA

Jesica Nauli Br. Siringo Ringo[1], Wahyu Joko Mursalin[2], Nisrina Citra Nurfadilah[3], Dwiky Rachmat Ramadhan[4] dan Wa Ode Zuhayeni Madjida[5]

[1,2,3,4,5]Politeknik Statistik STIS, Jl. Otto Iskandardinata No. 64C Jakarta Timur
[1]Email: 211709764@stis.ac.id
[2]Email: 211710048@stis.ac.id
[3]Email: 211709898@stis.ac.id
[4]Email: 211709648@stis.ac.id
[5]Email: zuhayeni@stis.ac.id

ABSTRAK

Penambahan kasus COVID-19 yang besar di Indonesia, khususnya Pulau Jawa, membutuhkan berbagai upaya untuk mengendalikannya. Salah satu upaya efektif yang dapat dilakukan adalah tindakan preventif dengan memberi informasi mengenai kondisi suatu wilayah. Sebagai peringatan kepada masyarakat dan sebagai upaya pengambilan kebijakan daerah, Indonesia mengeluarkan zona risiko sampai pada tingkat kabupaten/kota melalui Satgas Penanganan COVID-19. Pembentukan level zona risiko tersebut menggunakan teknik konvensional yaitu pembobotan skor menggunakan informasi dari tiga jenis indikator. Dengan mempertimbangkan bahwa zona risiko merupakan hal yang penting dalam penentuan kebijakan terkait COVID-19, penelitian ini bertujuan untuk membangun model klasifikasi zona risiko kabupaten/kota di Pulau Jawa menggunakan beberapa teknik klasifikasi *data mining* dan menentukan model klasifikasi terbaik berdasarkan hasil evaluasi. Teknik klasifikasi yang digunakan sebagai perbandingan dalam penelitian ini adalah *naive Bayes, decision tree, k-nearest-neighbor,* dan *neural network*. Sebelum dilakukan pemodelan, data disesuaikan terlebih dahulu pada tahap *preprocessing* di mana pada tahap tersebut teridentifikasi terdapat permasalahan *missing value* dan *imbalanced data*. Permasalahan tersebut diatasi dengan imputasi data dan teknik *oversampling*. Hasil penelitian menunjukkan bahwa model *k-nearest-neighbor* merupakan model terbaik dibandingkan tiga model lainnya. Hasil tersebut didasarkan pada ukuran evaluasi keempat model di mana model k-NN memiliki nilai *acccuracy*, nilai rata-rata makro untuk sensitivitas, spesifisitas, dan ukuran F1 paling tinggi dibandingkan model lainnya.
Kata kunci: zona risiko, klasifikasi, *data mining*, ukuran evaluasi

ABSTRACT

Various attempts are needed to control the increment of COVID-19 cases in Indonesia, especially Java Island. One of the effective attempt to do this is through the preventive act by providing news about a region. Indonesia, through Satgas Penanganan COVID-19, has built a risk zone of district/city as a warning system for the public and the substance of policy making for government in region level. The risk zone is built by three kinds of indicator using a conventional technique named score weighting. By considering the importance of the risk zone for policy making in the government, this study aims to build a risk zone classification model for districts / cities in Java using several *data mining* classification techniques and determine the best classification model based on evaluation results. This study uses several classification technique on the purpose of comparation. These techniques are naive Bayes, decision tree, k-nearest-neighbor, and *neural network*. Before entering the modeling stage, data is being adjustedat the preprocessing stage where missing value and imbalanced data problems are identifies. These problems is being overcome by doing data imputation and oversampling techniques. The result of this study indicates that k-nearest-neighbor is the best model compared to other three models. This result is based on the evaluation measures of the four models where the k-NN model has the highest accuracy value, the macro average value for sensitivity, specivicity, and F1-Measure compared to other models.
Keywords: risk zone; classification; *data mining*; evaluation measure

1. PENDAHULUAN

Penyakit virus corona (COVID-19) disebabkan oleh virus bernama *severe acutre respiratory syndrom coronavirus 2* (SARS-CoV-2). Kasus terjangkit pertama dilaporkan terjadi di kota Wuhan, provinsi Hubei, Cina pada Desember 2019, dan selanjutnya dideklarasikan sebagai pandemik global oleh *World Health Organization* (WHO) pada Maret 2020. Menurut [1], terdapat 58,2 juta kasus COVID-19 yang terkonfirmasi dan di antaranya terdapat 1,38 juta kematian per 23 November 2020. Pada waktu tersebut, Indonesia menduduki peringkat ke-21 sebagai negara dengan total kumulatif kasus terbanyak yaitu sebesar 497 ribu kasus. Kasus tersebut sudah menyebar di seluruh provinsi di Indonesia.

Pulau Jawa merupakan wilayah dengan data historis kasus terkonfirmasi positif yang tinggi di mana per 23 November 2020 terdapat sekitar 300 ribu kasus positif yang membuat Pulau Jawa menjadi wilayah dengan kasus positif tertinggi dibandingkan pulau lainnya [2]. Empat provinsi dengan kasus terkonfirmasi positif juga berada di Pulau Jawa, yaitu DKI Jakarta, Jawa Timur, Jawa Barat, dan Jawa Tengah. Pada minggu keempat bulan November, tidak terdapat satupun zona hijau di Pulau Jawa. Dari 5,45% daerah yang berisiko tinggi atau berzona merah, 3,69% kabupaten/kota berada di Pulau Jawa [2]. Kondisi tersebut membuat kasus COVID-19 di Pulau Jawa perlu dikaji lebih lanjut sebagai upaya untuk menekan penambahan kasus.

Dalam upaya membantu perencanaan sumber daya kesehatan sehingga dapat mencegah penyebaran COVID-19 semakin luas, dapat dilakukan berbagai pemodelan menggunakan informasi data yang tersedia. Di Indonesia, data harian mengenai kondisi COVID-19 dikeluarkan oleh berbagai tingkatan wilayah administrasi dimulai dari tingkat desa, kecamatan, kabupaten/kota, provinsi, dan nasional. Data harian tersebut berupa kasus positif, kasus *probable*, kasus *suspect*, dan kontak erat pasien. Ketersediaan data historis dapat menjadi bahan evaluasi pemerintah dalam mengkaji kebijakan terkait penanganan COVID-19.

Sebagai peringatan kepada masyarakat dan sebagai upaya pengambilan kebijakan daerah, Indonesia mengeluarkan zona risiko sampai pada tingkat kabupaten/kota. Berdasarkan penelitian [3], penyebab penyebaran COVID-19 yang tidak terkendali adalah kurangnya pengetahuan dan kesadaran masyarakat untuk melakukan pencegahan. Hal ini mengarah pada permasalahan selanjutnya di mana sulit untuk melakukan pengawasan, deteksi dini, dan penelusuran kontak pasien. Adanya zona risiko merupakan upaya untuk memberi gambaran kepada masyarakat mengenai situasi di wilayah tersebut.

Satuan Tugas (Satgas) Penanganan COVID-19 melakukan pembentukan zona risiko berdasarkan indikator epidemiologi, indikator surveilans kesehatan masyarakat, dan indikator pelayanan kesehatan. Ketiga indikator tersebut disusun menjadi sebuah indeks dan diterapkan untuk membentuk 5 level risiko kenaikan kasus. Pembentukan level zona risiko tersebut menggunakan teknik konvensional yaitu pembobotan skor. Mengingat zona risiko merupakan hal yang penting dalam penentuan kebijakan terkait COVID-19, penelitian ini berupaya untuk membentuk level zona risiko dengan menerapkan teknik *data mining*, di mana diharapkan model yang diperoleh dapat memberi sudut pandang lain bagi pembuat kebijakan. Penelitian ini bertujuan untuk melakukan prediksi zona risiko kabupaten/kota di Pulau Jawa melalui beberapa teknik klasifikasi *data mining* dan menentukan model klasifikasi terbaik berdasarkan hasil evaluasi. Beberapa teknik klasifikasi yang akan dibandingkan adalah *naive Bayes*, *decision tree*, *k-nearest-neighbor*, dan *neural network*.

Sebelum penelitian ini, terdapat beberapa peneliti yang sudah mempublikasikan kajian empirisnya mengenai kejadian penyebaran COVID-19. Penelitian oleh [4] bagian pertama bertujuan untuk mengkombinasikan metode *clustering* dan klasifikasi. Penelitian ini menggunakan kasus jumlah persebaran COVID-19 di Indonesia (34 provinsi). Variabel yang digunakan pada penelitian adalah jumlah kasus sembuh, jumlah kasus positif, dan jumlah kasus meninggal dunia. Metode *clustering* dan klasifikasi yang digunakan adalah *K-Medoids* dan C4.5 dengan label pemetaan berupa klaster tinggi (C1= zona merah), klaster waspada (C2= zona kuning), klaster rendah (C3= zona hijau). Hasil dari pemetaan diteruskan dengan pembentukan model klasifikasi menggunakan metode C4.5. Hasil pemetaan diperoleh 9 provinsi berada di klaster tinggi, 3 provinsi berada di klaster waspada, dan 22 provinsi berada di klaster rendah. Nilai yang diperoleh dari *decision tree* untuk klaster tinggi adalah jika jumlah kasus positif lebih kecil dari 9524 dan lebih besar dari 4329 ($4329 > x1 < 9524$).

Selanjutnya, penelitian yang dilakukan oleh [5] bagian kedua, bertujuan untuk membentuk model prediksi pada arsitektur *neural network* terbaik dengan mengkombinasikan metode *K-Medoids* dan *backpropagation (neural network)* pada kasus pandemi COVID-19 di Indonesia. Penelitian ini menggunakan lokus, variabel, dan jumlah klaster yang sama dengan bagian pertama di mana diperoleh hasil pemetaan klaster dilanjutkan ke metode *backpropagation* untuk memprediksi hasil akurasi dari klaster yang ada. Dengan menggunakan model arsitektur terbaik 3-2-1 diperoleh nilai akurasi 94,17% dengan *learning rate*= 0.696.

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

2. MATERI DAN METODE

COVID-19

Menurut informasi dari [6], virus Corona termasuk dalam kelompok virus yang menyebabkan penyakit baik pada manusia ataupun hewan. Saat menjangkit manusia, virus ini mengakibatkan penyakit infeksi saluran pernapasan, berupa flu biasa hingga penyakit yang sangat serius seperti *Severe Acute Respiratory Syndrome (SARS) dan Middle East Respiratory Syndrome (MERS).* Gejala COVID-19 mirip dengan SARS yang muncul pada tahun 2003, di mana angka kematian COVID-19 lebih rendah. Meskipun begitu, jumlah kasus COVID -19 jauh lebih banyak dan penyebarannya lebih luas dibandingkan dengan SARS. Terdapat beberapa istilah kasus dalam COVID-19, antara lain:

1. **Kasus *Suspect*** memiliki kriteria kasus infeksi saluran pernafasan akut di mana dalam 14 hari sebelum sakit, orang yang bersangkutan berasal/tinggal di daerah yang sudah terjadi *local transmission* atau pernah kontak dengan kasus terkonfirmasi positif / kasus *probable*. Pasien kasus *suspect* harus dirawat di rumah sakit, meskipun tidak ditemukan penyebabnya secara spesifik dan meyakinkan bahwa ini bukan penyakit COVID-19.
2. **Kasus *Probable*** yang merupakan kasus klinis yang diyakini sebagai COVID-19. Pasien berada dalam kondisi infeksi saluran pernafasan akut (ISPA) berat dan gangguan pernafasan yang sangat terlihat, namun belum dilakukan pemeriksaan laboratorium melalui RT-PCR untuk melakukan konfirmasi kasus positif atau tidak.
3. **Kontak Erat** atau *close contact* merupakan kasus di mana seseorang melakukan kontak dengan kasus konfirmasi positif atau dengan kasus *probable*.
4. **Kasus Konfirmasi** merupakan kasus di mana seseorang sudah terkonfirmasi positif setelah melalui pemeriksaan laboratorium RT-PCR. Terdapat dua kriteria dalam kasus konfirmasi yakni kasus konfirmasi dengan gejala dan kasus konfirmasi tanpa gejala.

Kondisi penyebaran virus Covid-19 di Indonesia dapat ditinjau berdasarkan peta risiko yang dapat diakses melalui *website* Satgas Penanganan COVID-19. Dalam peta risiko tersebut terdapat kelas zona resiko untuk mengklasifikasikan tiap wilayahnya, yaitu zona hijau (tidak terdampak dan tidak ada kasus), zona kuning (risiko rendah), zona oranye (risiko sedang), dan zona merah (risiko tinggi).

Klasifikasi dalam *Data mining*

Data mining merupakan suatu proses mencari pengetahuan atau pola dari sekumpulan data yang biasanya berukuran besar. Terdapat dua metode pembelajaran dalam *data mining* yang sering digunakan yaitu metode *supervised learning* dan *unsupervised learning*. *Supervised learning* adalah sebuah *task* dari fungsi yang memetakan *input-output* berdasarkan contoh pasangan *input-output* yang sudah ada untuk membangun sebuah model. Contoh algoritma dalam *supervised learning* adalah algoritma estimasi, peramalan, dan klasifikasi. *Unsupervised learning* merupakan fungsi yang memetakan *input-output* di mana label tidak ditentukan sebelumnya. Salah satu algoritma yang menggunakan *unsupervised learning* adalah teknik *clustering*.

Klasifikasi adalah sebuah proses menemukan model (atau fungsi) yang mendeskripsikan dan membedakan kelas data dengan tujuan agar dapat menggunakan model tersebut untuk prediksi kelas data yang label kelasnya tidak diketahui [7]. Terdapat beberapa algoritma dalam klasifikasi *seperti decision tree, Naive Bayes, k- nearest neighbor, neural network, adaboost, support vector machine* (SVM), dan lain-lain. Berikut penjelasan beberapa metode klasifikasi:

1. ***Neural network* (NN)**

   Neural network atau jaringan saraf adalah sekumpulan unit input/output yang terhubung di mana setiap koneksi memiliki bobot yang terkait dengannya. Selama fase pembelajaran, jaringan melakukan proses pembelajaran dengan menyesuaikan bobot sehingga dapat memprediksi label kelas dari input *tupel* dengan benar. *Backpropagation* adalah algoritma yang digunakan dalam pembelajaran NN [7]. Kelebihan dari metode klasifikasi NN adalah dapat memetakan berdasarkan *input-output* dan juga sangat fleksibel terhadap data *noisy.*
2. **Decision Tree**

   Decision tree merupakan metode yang sangat sering digunakan pada teknik klasifikasi. Menurut [7], *decision tree* berbentuk struktur pohon seperti diagram alir, di mana setiap *node* internal (*node* tanpa daun) menunjukkan pengujian pada atribut, setiap cabang mewakili hasil pengujian, dan setiap *node* daun merupakan label kelas. *Node* yang paling atas disebut dengan akar *node*. Kelebihan dari *decision tree* adalah hasil klasifikasi yang akurat, proses komputasi yang efisien, mampu menghindari hilangnya informasi pada atribut kontinu dan dapat menginterpretasikan prosesnya secara lebih mudah dibandingkan metode klasifikasi lainnya.

3. **K-Nearest Neighbor (k-NN)**
   k-NN adalah suatu metode untuk mengklasifikasikan suatu objek berdasarkan data latihan sebanyak k yang jaraknya paling dekat dengan objek tersebut. Ketepatan algoritma k-NN dan tingkat akurasi model sangat dipengaruhi oleh penentuan jumlah k. k-NN merupakan contoh dari teknik pembelajaran yang malas di mana proses pembelajaran dilakukan setelah adanya data baru.
4. **Naive Bayes**
   Naive Bayes merupakan metode yang paling populer dan sederhana yang digunakan untuk mengklasifikasikan data dalam jumlah yang besar dan dapat digunakan untuk memprediksi keanggotaan suatu kelas. [8] menjelaskan bahwa *naive Bayes* digunakan karena memiliki tingkat ketelitian dan kecepatan tinggi saat diaplikasikan untuk jumlah data yang besar. Selain itu, metode ini memiliki keuntungan yaitu hanya membutuhkan jumlah data latih yang kecil dalam proses pengklasifikasiannya. Metode ini memiliki asumsi independensi dari masing-masing kondisi.

Evaluasi Model

Untuk membandingkan kinerja setiap model, dibentuk matriks konfusi dari masing-masing model. Dari *confusion matrix*, dihitung masing-masing ukuran evaluasi model berupa *accuracy, precision, recall, dan F1-Measure.* Dalam membentuk ukuran evaluasi model tersebut terdapat empat komponen, yaitu TP merupakan jumlah kasus positif yang diklasifikasikan dengan benar, TN merupakan jumlah kasus negatif yang diklasifikasikan dengan benar, FP merupakan jumlah kasus positif yang diklasifikasikan dengan salah, dan FN merupakan jumlah kasus negatif yang diklasifikasikan dengan salah. Rumus dari ukuran evaluasi model tersebut ditunjukkan dalam persamaan 1, 2, 3 dan 4.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Pada klasifikasi non binar, matriks konfusi yang terbentuk lebih kompleks. Tabel 1 merupakan contoh matriks konfusi pada tiga kategori kelas.

Tabel 1. *Confusion matrix* 3x3

		Kelas Prediksi		
		Kelas A	Kelas B	Kelas C
Kelas Aktual	Kelas A	AA	AB	AC
	Kelas B	BA	BB	BC
	Kelas C	CA	CB	CC

Kondisi *Missing Data* dan Upaya Penanganannya

Little dan Rubin (2002) dalam [9] menyatakan bahwa terdapat tiga kondisi *missing data* atau data hilang, yaitu:

a. *Missing Not at Random* (MNAR) yaitu kejadian data hilang pada suatu variabel berkaitan dengan variabel itu sendiri, sehingga data hilang tidak dapat diprediksi dari variabel lain pada *dataset.*
b. *Missing Completely at Random* (MCAR) yaitu kejadian data hilang pada suatu variabel tidak berkaitan dengan seluruh variabel, baik variabel dari data yang hilang maupun dengan variabel lainnya. Dalam arti, hilangnya data terjadi secara acak.
c. *Missing at Random* (MAR) yaitu kejadian data hilang pada suatu variabel berkaitan dengan data lainnya, namun tidak berkaitan dengan variabel itu sendiri.

Imputasi merupakan prosedur pengisian data hilang menggunakan informasi pada *dataset* tersebut. Imputasi dapat dilakukan dengan mengisi data secara manual atau otomatis. Imputasi manual dilakukan berdasarkan justifikasi peneliti di mana hal ini sulit dilakukan, khususnya pada data berjumlah besar. Terdapat beberapa metode pengisian secara otomatis, yaitu:

a. Metode rata-rata
   Metode ini merupakan bagian dari imputasi tunggal atau melakukan imputasi berdasarkan nilai tunggal dari suatu penduga. Nilai tunggal dapat dihitung berdasarkan keseluruhan data, maupun berdasarkan kelas yang sama dari data yang hilang tersebut. Pada data numerik, penduga yang digunakan adalah nilai rata-rata, sedangkan pada data kategorik penduga yang digunakan adalah nilai modus. Menurut [9], metode rata-rata memiliki keunggulan dalam menghasilkan nilai harapan yang relatif stabil, namun kelemahannya adalah keragaman yang diperoleh tidak sesuai dengan data yang sebenarnya sehingga cenderung *underestimate.*

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

b. Metode *Expectation Maximisation* (EM)
Metode EM merupakan metode yang digunakan untuk memperkirakan parameter populasi yang tidak diketahui. Algoritma EM dilakukan dengan membentuk model *maximum likelihood* di mana dilakukan pengisian nilai yang hilang terlebih dahulu, kemudian mencari penduga *maximum likelihood*. Proses dilakukan secara iterative untuk menghasilkan suatu statistik yang cukup untuk menduga parameter.

c. Metode *Multiple Imputation*
Metode ini melakukan imputasi berdasarkan beberapa kemungkinan di mana nilai dari data yang hilang diperoleh berdasarkan sampel acak dari nilai-nilai yang hilang. Terdapat sebanyak m himpunan data yang lengkap dan setiap set dianalisis dengan metode data lengkap. Hasil yang diperoleh dari m himpunan data digabungkan untuk menghasilkan inferensi yang valid secara statistik dengan mempertimbangkan ketidakpastian dalam prosesnya.

Salah satu metode dalam imputasi ganda adalah *Multivariate Imputation by Chained Equations* (MICE). Penelitian oleh [10] menyatakan bahwa metode MICE merupakan prosedur imputasi dengan asumsi bahwa kondisi data hilang adalah MAR. MICE merupakan metode yang sangat fleksibel dan dapat digunakan pada berbagai kondisi data. Tahapan dalam melakukan prosedur MICE dapat dilakukan sebagai berikut:

1. Melakukan imputasi tunggal seperti imputasi rata-rata pada setiap nilai yang hilang
2. Ditetapkan suatu variabel hasil imputasi rata-rata kembali menjadi data yang hilang. Variabel tersebut dinyatakan sebagai "var"
3. Nilai "var" pada tahap kedua diregresikan dengan variabel lainnya di dalam model imputasi, di mana model dapat berisi seluruh atau sebagai variabel dalam himpunan data. Dalam arti, "var" digunakan sebagai variabel dependen dan variabel lainnya merupakan variabel independen. Regresi yang dilakukan memiliki proses dan asumsi sama dengan model umumnya, tergantung pada model yang digunakan seperti linear, logistic dan Poisson.
4. Data yang hilang pada "var" diganti dengan hasil prediksi model
5. Tahap 2-4 dilakukan secara iteratif untuk setiap variabel yang memiliki nilai hilang. Setiap variabel melalui satu iterasi atau satu siklus di mana pada siklus terakhir seluruh nilai yang hilang sudah diganti dengan hasil prediksi model.
6. Tahap 2-4 dilakukan secara berulang di mana setiap siklusnya menggunakan hasil imputasi dari siklus sebelumnya.

Imbalance Class

Suatu data dikatakan *imbalance* atau tidak seimbang ketika rasio observasi pada suatu kelas data lebih banyak daripada kelas yang lain. Kelas dengan rasio yang besar disebut dengan kelas mayor, sedangkan kelas dengan rasio yang kecil disebut kelas minor. Dalam rangka menangani permasalahan kelas yang tidak seimbang, terdapat berbagai metode yang dapat dilakukan. Salah satu teknik yang sering digunakan adalah teknik *resampling* yang terdiri atas *undersampling* dan *oversampling*. Teknik *oversampling* melakukan penyeimbangan kelas dengan melakukan duplikasi kelas minor secara acak. Di samping itu, teknik *undersampling* melakukan penyeimbangan kelas dengan mengurangi data pada kelas mayor agar memiliki jumlah observasi yang sama dengan kelas minor. Tidak seperti *undersampling*, teknik *oversampling* tidak membuat informasi dari kelas mayor hilang.

Kerangka Pikir

Alur kerangka pikir yang mendasari penelitian ini dimulai dari upaya pembentukan zona risiko melalui teknik *data mining* di Pulau Jawa.

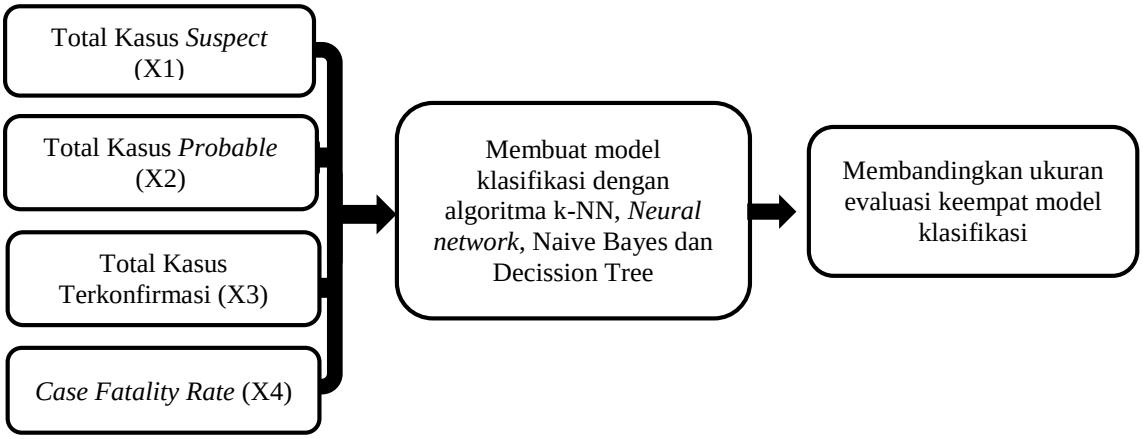
Pembentukan level zona risiko menggunakan variabel total kasus *suspect*, total kasus *probable*, total kasus terkonfirmasi, dan *case fatality rate*. Pelabelan unit observasi, dalam hal ini kabupaten/kota di Pulau Jawa, diperoleh melalui data zona risiko oleh Satgas Penanganan COVID-19. Dalam penelitian ini dilakukan pemodelan klasifikasi menggunakan k-NN, *neural network, naive Bayes* dan *decision tree*. Selanjutnya, pada model-model yang telah terbentuk, dilakukan perbandingan ukuran evaluasi dari ketiga model tersebut untuk mengetahui kombinasi model terbaik. Gambar 1 adalah alur kerangka pikir yang disajikan dalam bentuk diagram.

Metode Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Gugus Tugas Pusat, Kementerian Kesehatan, dan hasil rekapitulasi dari masing-masing pemerintah daerah yang dicuplik dan diolah di situs web www.andrafarm.com. Variabel yang digunakan adalah total konfirmasi, total

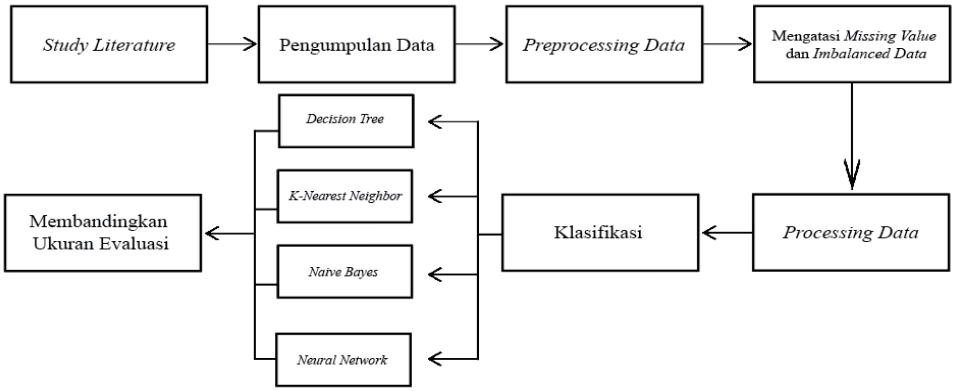
ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

suspect, total *probable*, dan *case fatality rate* (CFR). Dalam melakukan pengolahan data, digunakan perangkat lunak R Studio.



Gambar 1. Kerangka pikir

Berdasarkan pendahuluan yang telah dijelaskan, penelitian ini dilakukan melalui beberapa tahap. Tahap pertama adalah dilakukan studi literatur mengenai upaya pemanfaatan data COVID-19 dalam membangun suatu pengetahuan. Selanjutnya, proses pengumpulan data yang berasal dari beberapa sumber. Data yang diambil adalah data persebaran COVID-19 di Pulau Jawa pada tanggal 20 November 2020 yang terdiri atas 119 observasi. Input yang digunakan pada klasifikasi merupakan zona risiko oleh Satgas Penanganan COVID-19 yang terdiri atas zona hijau, zona kuning, zona oranye, dan zona merah. Namun, zona risiko seluruh kabupaten/kota di Pulau Jawa per 20 November 2020 hanya terdiri atas zona kuning, zona oranye, dan zona merah. Dalam hal ini berarti tidak terdapat kabupaten/kota yang berzona hijau. Sehingga, terdapat 3 kelas yang digunakan pada input klasifikasi. Data yang diperoleh kemudian memasuki tahap *preprocessing* di mana teridentifikasi bahwa terdapat permasalahan *missing value* dan ketidakseimbangan data. Data yang telah diolah pada tahap *preprocessing* dilanjutkan dengan tahap *processing* yang menggunakan beberapa teknik klasifikasi, yaitu *naive Bayes*, *decision tree*, *k-nearest neighbor*, *neural network*. Dilakukan perbandingan hasil ukuran evaluasi dari keempat metode klasifikasi tersebut untuk memperoleh model terbaik. Alur penelitian disajikan pada diagram alir pada gambar 2.



Gambar 2. Diagram alir kerangka kerja penelitian

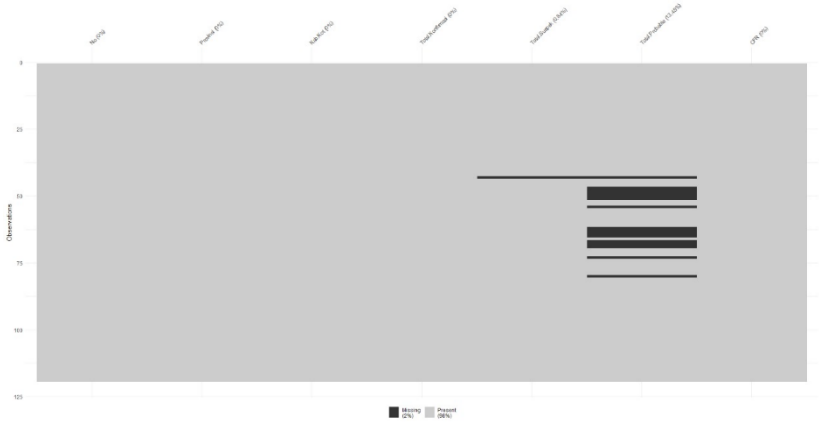
3. HASIL DAN PEMBAHASAN

Prosedur Imputasi Data

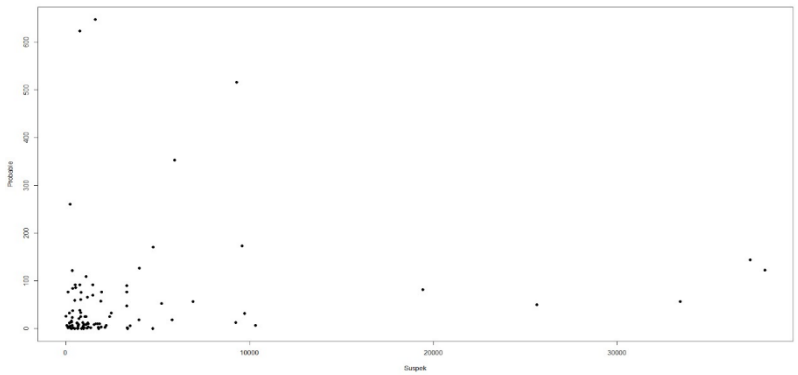
1. Untuk mengetahui jenis kondisi data hilang, dapat dilihat pola dari data hilang yang terbentuk berdasarkan data referensi yang diurutkan, dalam hal ini adalah variabel *suspect*. Terdapat sebanyak 2,1% data yang hilang di dalam himpunan data di mana satu nilai berada pada variabel *suspect* dan nilai lainnya berada pada variabel *probable*. Berdasarkan gambar 3 diketahui bahwa mayoritas data yang hilang hanya terdapat pada variabel *probable*, maka tidak terdapat kesamaan pola dari data yang hilang. Sehingga, kondisi data tersebut bukan merupakan data MNAR.

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

2. Untuk melihat hubungan variabel *probable* dengan variabel referensinya yaitu variabel *suspect*, maka dibentuk *scatter plot*. Berdasarkan gambar 4, dapat diketahui bahwa berdasarkan nilai-nilai observasi yang berkumpul kedua variabel berhubungan positif, meskipun terdapat nilai yang outlier. Sehingga, dapat disimpulkan bahwa kondisi data hilang merupakan MAR. Hal ini sesuai dengan kondisi kasus COVID-19 di mana sesuai dengan definisinya, kasus *probable* dan kasus *suspect* merupakan kasus pasien yang memiliki infeksi saluran pernapasan akut.
3. Berdasarkan kondisi data yang hilang yaitu MAR, maka metode imputasi yang dapat digunakan adalah MICE. Dalam melakukan metode MICE, parameter yang digunakan adalah m=2 di mana nilai m merupakan iterasi maksimum yang dapat dilakukan. Menurut [11], nilai m yang digunakan merupakan nilai persentase dari data yang hilang yaitu sebesar 2,1%. Sehingga, setelah melakukan prosedur imputasi MICE, data yang digunakan pada penelitian ini sudah lengkap.



Gambar 3 Pola Persebaran Data Hilang



Gambar 4. *Scatter plot* variabel *probable* dan variabel *suspect*

Penanganan *Imbalance Data*

Berdasarkan hasil deskripsi pada data, diketahui bahwa proporsi kabupaten/kota dengan zona risiko rendah sebesar 15,97%, zona risiko sedang sebesar 67,23%, dan zona risiko tinggi sebesar 16,80%. Kondisi tersebut menunjukkan bahwa data *imbalanced*. Penelitian oleh [12] menyatakan bahwa model yang dibuat dengan *imbalance class* menghasilkan prediksi kelas minor yang rendah. Informasi yang ada pada kelas mayor mendominasi kelas minor, sehingga informasi pada kelas minor cenderung diabaikan dalam sistem klasifikasi. Karena sulit untuk memprediksi model klasifikasi dengan baik, maka diperlukan penanganan pada data *imbalanced*. Penelitian ini menggunakan teknik *random oversampling* di mana dilakukan penyeimbangan kelas dengan melakukan duplikasi pada kelas minor secara acak agar memiliki jumlah anggota sama dengan kelas mayoritas.

Hasil Pemodelan Naïve Bayes

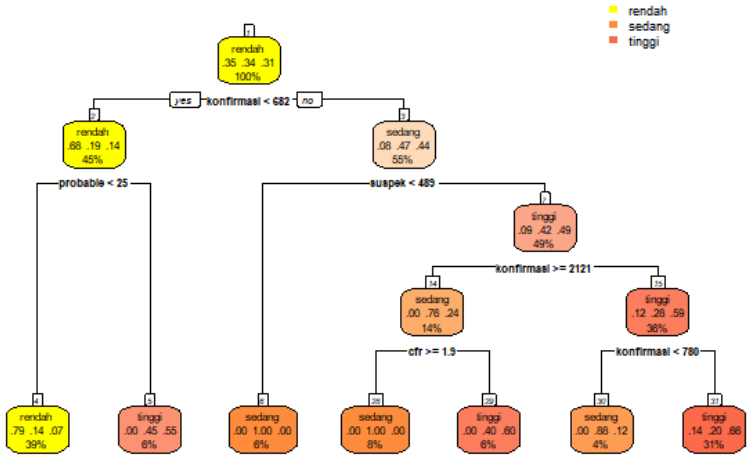
Naive Bayes merupakan suatu metode pengklasifikasian dengan metode probabilitas dan statistik di mana dilakukan *prediksi* peluang di masa depan berdasarkan kejadian di masa lalu dengan asumsi adanya *independensi* antar kondisi atau kejadian. Berdasarkan hasil pemgolahan data, diperoleh hasil probabilitas

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

prior kabupaten/kota diklasifikasikan zona rendah sebesar 0,3222, diklasifikasikan zona sedang sebesar 0,3556, dan diklasifikasikan zona tinggi sebesar 0,3222.

Hasil Pemodelan *Decision Tree*

Dalam pembentukan model *decision tree* digunakan fungsi *rpart* dalam *software* R untuk menghasilkan output *decision tree* secara otomatis. Gambar 5 adalah output *decision tree* yang dihasilkan.



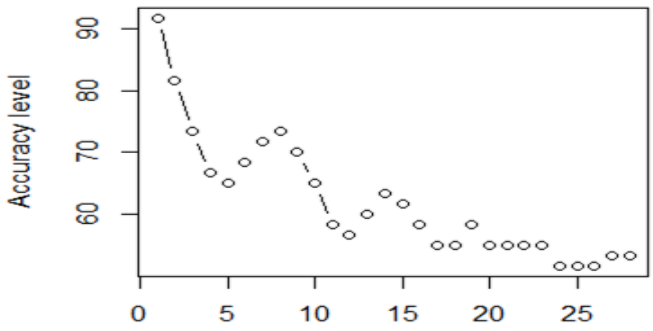
Gambar 5. *Output decision tree*

Gambar 5 menunjukkan bahwa jika kabupaten/kota di Pulau Jawa memiliki total kasus terkonfirmasi < 682 kasus, maka kabupaten/kota tersebut terklasifikasi sebagai zona risiko rendah. Kemudian, dilakukan peninjauan berdasarkan kasus *probable* di mana jika kasus < 25, maka terklasifikasi sebagai zona risiko rendah dan zona risiko tinggi untuk kasus > 25.

Jika suatu kabupaten/kota memiliki total kasus terkonfirmasi > 682 kasus, maka terklasifikasi sebagai zona risiko sedang yang kemudian terbagi berdasarkan total kasus *suspect*. Jika total kasus *suspect* < 489 maka terklasifikasi sebagai zona risiko sedang, dan tidak dapat dipartisi kembali. Jika kasus *suspect* > 489 maka terklasifikasi sebagai zona risiko tinggi. Pada tahap ini, kasus dapat dipartisi kembali berdasarkan total kasus terkonfirmasi di mana jika jumlahnya ≥ 2121 kasus dan memiliki CFR ≥1,9, maka termasuk zona risiko sedang. Namun, jika total kasus terkonfirmasi kurang dari sama dengan 2121 kasus, maka wilayah tersebut terklasifikasikan zona risiko tinggi dan dilakukan partisi kembali berdasarkan total terkonfirmasinya, di mana jika total konfirmasi < 780 kasus maka termasuk zona risiko sedang dan jika >780 termasuk zona risiko tinggi.

Hasil Pemodelan k-NN

Perhitungan k-NN dilakukan dengan jarak Euclidean untuk memperoleh jarak terdekat dengan data baru. Model dibentuk menggunakan nilai k optimum yang diinisiasi di awal. K optimum diperoleh dari hasil iterasi model k-NN di mana pada gambar 6 diketahui bahwa akurasi maksimum yang ditunjukkan dengan siku pada grafik berada pada k=2. Sehingga, parameter k=2 digunakan pada model k-NN.

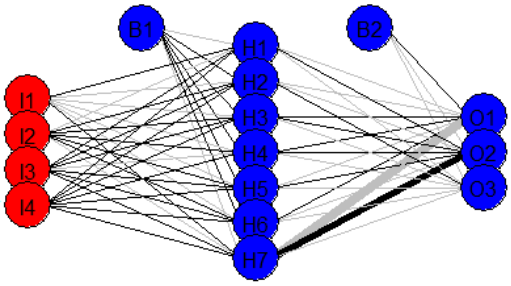


Gambar 6. *Plot* akurasi terhadap nilai k

Hasil Pemodelan *Neural network*

Model *neural network* yang digunakan dalam penelitian ini menggunakan 7 *hidden layer* dengan bobot masing-masing telah ditentukan secara otomatis oleh fungsi. Fungsi aktivasi yang digunakan adalah "logistic", dengan fungsi *error* yang digunakan adalah "ce". Gambar 7 adalah hasil dari pembentukan model menggunakan *neural network*.

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)



Gambar 7 *Output Model Neural network*

Evaluasi Model

Berdasarkan pemodelan klasifikasi dengan keempat metode tersebut, dilakukan evaluasi terhadap model yang terbentuk. Evaluasi model dilakukan menggunakan akurasi, presisi, sensitivitas, spesifisitas dan ukuran F1. Tabel 2 menunjukan hasil evaluasi dari setiap kelas yang diperoleh.

Tabel 2. Ukuran evaluasi model

Metode	Zona Risiko	Akurasi	Sensitivitas	Spesifisitas	Presisi	Ukuran F1
naive Bayes	Rendah	0,5333	0,7727	0,5789	0,7727	0,6182
	Sedang		0,3125	0,9546	0,3125	0,4348
	Tinggi		0,4545	0,7368	0,4546	0,4762
	Rata-rata Makro		**0,5132**	**0,7567**	**0,5133**	**0,5097**
decision tree	Rendah	0,7333	0,7059	0,9535	0,8571	0,7742
	Sedang		0,4444	0,9762	0,8889	0,5926
	Tinggi		0,9600	0,6286	0,6486	0,7742
	Rata-rata Makro		**0,7034**	**0,8528**	**0,7982**	**0,7149**
K-NN	Rendah	0,7500	0,8077	0,9412	0,9130	0,8571
	Sedang		0,5882	0,8372	0,5882	0,5882
	Tinggi		0,8235	0,8605	0,7000	0,7568
	Rata-rata Makro		**0,7398**	**0,8796**	**0,7337**	**0,7340**
neural network	Rendah	0.6833	0.7778	0.8571	0.7	0.7368
	Sedang		0.3158	0.9756	0.8571	0.4615
	Tinggi		0.9130	0.6757	0.6363	0.7500
	Rata-rata Makro		**0,6689**	**0,8361**	**0,7311**	**0,6494**

Berdasarkan ukuran evaluasi model yang diperoleh, dapat diketahui bahwa secara umum model k-NN menunjukkan performa yang lebih baik dibandingkan model lainnya di mana model *decision tree* menunjukkan hasil evaluasi model yang hampir sama baiknya dengan model k-NN. Berdasarkan nilai akurasi, dapat diketahui pula bahwa kedua model tersebut unggul dibandingkan model lainnya.

Dalam kasus data tidak seimbang, ukuran evaluasi yang lebih tepat untuk diinterpretasi adalah sensitivitas dan spesifisitas, Pada nilai sensitivitas diketahui bahwa kelas minoritas (zona rendah dan zona tinggi) bernilai lebih tinggi dibandingkan kelas mayoritas (zona sedang). Hal tersebut sejalan dengan nilai ukuran F1 dan terjadi pada seluruh model sehingga menunjukkan bahwa keempat model sudah cukup baik dalam mengklasifikasikan kelas minoritas. Nilai spesifisitas menunjukkan kondisi sebaliknya kecuali pada model k-NN yang memiliki nilai spesifisitas lebih kecil pada kelas minoritas dibandingkan kelas mayoritas. Pada nilai presisi, ditunjukkan bahwa model *decision tree* dan *neural network* bernilai lebih besar untuk kelas mayoritas dibandingkan minoritas. Berdasarkan hasil evaluasi model menurut masing-masing kelas, dapat disimpulkan bahwa model k-NN dan *decision tree* menunjukkan kinerja yang lebih baik dibandingkan model lainnya.

Rata-rata makro digunakan untuk menunjukkan hasil evaluasi model dengan satu nilai yang dapat dibandingkan antar model. Menurut [13], rata-rata makro diperoleh dari rata-rata aritmatika dari setiap kelas. Secara umum, model k-NN memiliki rata-rata makro yang lebih besar dari setiap nilai evaluasi model, kecuali nilai presisi. Hasil evaluasi yang baik pada sensitivitas dan spesifisitas tersebut menunjukkan bahwa model k-NN merupakan model terbaik dibandingkan ketiga model lainnya untuk melakukan klasifikasi zona risiko COVID-19.

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)

4. KESIMPULAN DAN SARAN

Berdasarkan hasil klasifikasi menggunakan 4 model, diperoleh kesimpulan bahwa model k-NN merupakan model terbaik untuk melakukan klasifikasi zona risiko COVID-19. Pertimbangan model terbaik didasarkan pada hasil evaluasi pada masing-masing kelas dan secara keseluruhan melalui rata-rata mikro. Adanya perbedaan hasil klasifikasi antara zona risiko oleh Satgas Penanganan COVID-19 menggunakan metode penimbang dengan penelitian menggunakan *data mining* dapat menjadi pengetahuan baru dalam menghasilkan zona risiko. Penggunaan variabel pada penelitian ini relatif lebih mudah didapatkan dan berfrekuensi lebih tinggi dibandingkan indikator yang digunakan oleh Satgas Penanganan COVID-19. Efisiensi dalam penentuan indikator dan penggunaan teknik pembentukan zona risiko dapat menjadi pertimbangan dalam menentukan metode alternatif untuk membentuk level zona risiko COVID-19.

DAFTAR PUSTAKA

[1] World Health Organisation, 'WHO Coronavirus Disease (COVID-19) Dashboard'. https://covid19.who.int/table (accessed Nov. 20, 2020).
[2] Satuan Tugas Penanganan COVID 19, 'Peta Sebaran | Covid19.go.id'. https://covid19.go.id/peta-sebaran (accessed Nov. 20, 2020).
[3] W. Wiguna and D. Riana, 'Diagnosis of Coronavirus disease 2019 (Covid-19) surveillance using C4. 5 algorithm', *Jurnal PILAR Nusa Mandiri*, vol. 16, no. 1, pp. 71–80, 2020.
[4] A. P. Windarto, U. Indriani, M. R. Raharjo, and L. S. Dewi, 'Bagian 1: Kombinasi Metode Klastering dan Klasifikasi (Kasus Pandemi Covid-19 di Indonesia)', *Jurnal Media Informatika Budidarma*, vol. 4, no. 3, pp. 855–862, 2020.
[5] A. P. Windarto, J. Naam, Y. Yuhandri, A. Wanto, and M. Mesran, 'Bagian 2: Model Arsitektur Neural Network Dengan Kombinasi K-Medoids dan Backpropagation pada kasus Pandemi Covid-19 di Indonesia', *Jurnal Media Informatika Budidarma*, vol. 4, no. 4, pp. 1175–1180, 2020.
[6] Kemenkes Indonesia, 'Kemenkes Siap Sosialisasikan Perubahan Istilah ODP, PDP dan OTG ke Seluruh Dinas Kesehatan - Sehat Negeriku'. https://sehatnegeriku.kemkes.go.id/baca/umum/20200714/3334463/kemenkes-siap-sosialisasikan-perubahan-istilah-odp-pdp-dan-otg-seluruh-dinas-kesehatan/ (accessed Nov. 20, 2020).
[7] M. Kamber and J. Pei, *Data Mining*. Morgan kaufmann, 2006.
[8] C. M. Rahman, M. Kabir, A. Hossain, and K. Dahal, 'Enhanced classification accuracy on naive bayes data mining models', 2011.
[9] T. Hendrawati, 'Kajian Metode Imputasi dalam Menangani Missing Data', 2015.
[10] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, 'Multiple imputation by chained equations: what is it and how does it work?', *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011.
[11] T. E. Bodner, 'What improves with increased missing data imputations?', *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 15, no. 4, pp. 651–675, 2008.
[12] C. Jian, J. Gao, and Y. Ao, 'A new sampling method for classifying imbalanced data based on support vector machine ensemble', *Neurocomputing*, vol. 193, pp. 115–122, 2016.
[13] B. Jeong *et al.*, 'Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data', *Diagnostics*, vol. 10, no. 6, p. 415, 2020.

ISSN: 2337-7631 (printed)
ISSN: 2654-4091 (Online)