

ANALISIS KLASIFIKASI TINGKAT PENYEBARAN COVID-19 MENGUNAKAN ALGORITMA NAÏVE BAYES BERBASIS RAPIDMINER

Anitaria Putri Rahayu, Minarwati

Sistem Informasi, STMIK El Rahma Yogyakarta
Jalan Karangajen No. 76, Yogyakarta, Indonesia
anitariarahayu@gmail.com

ABSTRAK

Pandemi Pandemi COVID-19 menghasilkan banyak data statistik yang dapat dimanfaatkan untuk proses analisis menggunakan teknik data mining. Penelitian ini bertujuan untuk melakukan klasifikasi tingkat penyebaran COVID-19 menggunakan algoritma Naïve Bayes berbasis RapidMiner. Dataset yang digunakan berupa data statistik COVID-19 beberapa negara dengan atribut seperti TotalCases, TotalDeaths, TotalRecovered, dan ActiveCases. Proses penelitian dilakukan melalui tahap preprocessing data, pembentukan label klasifikasi, pembagian data training dan testing, proses klasifikasi menggunakan algoritma Naïve Bayes, serta evaluasi model menggunakan confusion matrix dan accuracy. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memperoleh nilai accuracy sebesar 89.86%. Berdasarkan hasil tersebut dapat diketahui bahwa algoritma Naïve Bayes mampu melakukan klasifikasi tingkat penyebaran COVID-19 dengan cukup baik menggunakan aplikasi RapidMiner.

Kata kunci : Covid-19, data mining, klasifikasi, naïve bayes, rapidminer

1. PENDAHULUAN

COVID-19 merupakan penyakit menular yang menyebar dengan cepat di berbagai negara dan menyebabkan peningkatan jumlah kasus dalam waktu singkat. Banyaknya data statistik COVID-19 yang tersedia dapat dimanfaatkan sebagai sumber informasi melalui proses pengolahan data menggunakan teknik data mining. Salah satu teknik data mining yang dapat digunakan adalah klasifikasi.[1].

Meningkatnya jumlah kasus COVID-19 menghasilkan data statistik yang terus bertambah dari waktu ke waktu. Data tersebut meliputi jumlah total kasus, jumlah kematian, jumlah pasien sembuh, kasus aktif, serta kondisi pasien yang tergolong serius atau kritis. Banyaknya data COVID-19 yang tersedia dapat dimanfaatkan sebagai sumber informasi untuk membantu proses analisis tingkat penyebaran COVID-19 di berbagai wilayah atau negara. Namun, apabila data hanya disajikan dalam bentuk angka tanpa dilakukan pengolahan lebih lanjut, maka informasi yang diperoleh masih sulit dipahami dan kurang memberikan gambaran yang jelas mengenai kondisi penyebaran COVID-19[2].

Berdasarkan kondisi tersebut, diperlukan suatu metode yang mampu membantu proses pengolahan data agar data statistik COVID-19 dapat menghasilkan informasi yang lebih terstruktur dan mudah dianalisis. Salah satu metode yang dapat digunakan adalah data mining. Data mining merupakan proses pengolahan data untuk menemukan pola, hubungan, maupun informasi tertentu dari kumpulan data berukuran besar. Teknik data mining banyak digunakan dalam berbagai bidang, termasuk bidang kesehatan, karena mampu membantu proses analisis data secara lebih efektif.[3]. Salah satu teknik dalam data mining yang cukup sering digunakan adalah klasifikasi. Klasifikasi digunakan untuk mengelompokkan data ke dalam kategori

tertentu berdasarkan atribut yang dimiliki oleh data tersebut. Pada penelitian ini, klasifikasi digunakan untuk menentukan tingkat penyebaran COVID-19 ke dalam dua kategori, yaitu Tinggi dan Rendah berdasarkan jumlah TotalCases pada dataset. Dengan adanya proses klasifikasi, data COVID-19 tidak hanya menampilkan angka kasus, tetapi juga memberikan informasi mengenai kategori tingkat penyebaran sehingga lebih mudah dipahami.

Algoritma yang digunakan dalam penelitian ini adalah Naïve Bayes. Algoritma Naïve Bayes merupakan salah satu algoritma klasifikasi berbasis probabilitas yang memiliki proses perhitungan relatif sederhana dan mudah diterapkan. Meskipun menggunakan asumsi bahwa setiap atribut bersifat independen, algoritma Naïve Bayes tetap mampu memberikan hasil klasifikasi yang cukup baik pada berbagai penelitian data mining [4].

Penelitian ini menggunakan aplikasi RapidMiner sebagai alat bantu dalam proses klasifikasi data. RapidMiner dipilih karena memiliki tampilan visual berbasis operator yang mempermudah pengguna dalam melakukan proses pengolahan data mulai dari import dataset, preprocessing data, pembentukan model, hingga evaluasi hasil klasifikasi tanpa harus menggunakan banyak kode program.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk melakukan klasifikasi tingkat penyebaran COVID-19 menggunakan algoritma Naïve Bayes berbasis RapidMiner serta mengetahui performa model klasifikasi berdasarkan nilai accuracy dan confusion matrix yang dihasilkan.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian mengenai penerapan data mining dan algoritma klasifikasi pada data COVID-19 telah

banyak dilakukan sebelumnya. Salah satu penelitian dilakukan oleh [1] [1] membahas penerapan algoritma klasifikasi Naïve Bayes, C4.5, dan K-Nearest Neighbor terhadap data penyebaran COVID-19 di Korea Selatan menggunakan RapidMiner. Penelitian tersebut dilakukan melalui beberapa tahapan mulai dari preprocessing data, transformasi data, hingga evaluasi model klasifikasi. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memperoleh tingkat akurasi terbaik sebesar 80,79% dibandingkan algoritma lainnya sehingga dinilai cukup baik dalam melakukan klasifikasi data COVID-19. Penelitian tersebut juga menunjukkan bahwa penerapan data mining mampu membantu proses analisis penyebaran COVID-19 secara lebih efektif dan terarah.

Penelitian lain dilakukan oleh Hetty Rohayani dkk. mengenai prediksi penyebaran virus COVID-19 menggunakan metode Naïve Bayes berbasis RapidMiner. Penelitian tersebut menggunakan beberapa atribut kesehatan pasien seperti usia, jenis kelamin, kondisi kesehatan, dan vaksinasi untuk membantu proses klasifikasi status terkonfirmasi COVID-19. Berdasarkan hasil penelitian, algoritma Naïve Bayes mampu menghasilkan tingkat akurasi sebesar 97% dalam melakukan proses prediksi data kesehatan pasien COVID-19. Hasil tersebut menunjukkan bahwa metode probabilistik seperti Naïve Bayes cukup efektif digunakan pada bidang kesehatan, khususnya dalam membantu proses analisis data COVID-19 [2].

Selain itu, penelitian yang dilakukan oleh Dewi Yanti Liliana dkk. membahas penerapan data mining untuk memprediksi status pasien COVID-19 menggunakan pengklasifikasi Naïve Bayes. Dataset yang digunakan diperoleh dari Kaggle dan diolah menggunakan RapidMiner untuk membantu proses klasifikasi status kegawatan pasien berdasarkan usia dan jenis kelamin. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mampu menghasilkan tingkat akurasi sebesar 96,67% dalam mengklasifikasikan status pasien COVID-19 sehingga metode tersebut dianggap cukup baik digunakan dalam analisis data kesehatan berbasis machine learning [3].

Penelitian mengenai klasifikasi kasus COVID-19 juga dilakukan oleh Ahmad Firdaus dkk. menggunakan model Gaussian Naive Bayes Classifier. Penelitian tersebut bertujuan untuk mengetahui tingkat kasus COVID-19 di Kabupaten Pamekasan berdasarkan data pasien yang telah direkam selama periode tertentu. Hasil penelitian menunjukkan bahwa model Gaussian Naive Bayes menghasilkan nilai akurasi sebesar 96,88% dengan nilai precision, recall, dan f1-score yang cukup tinggi sehingga model dinilai mampu melakukan klasifikasi kasus COVID-19 dengan baik [4].

Penelitian lain yang berkaitan dengan klasifikasi kesehatan dilakukan oleh Ahmad Roni Husaini mengenai klasifikasi berbasis machine learning untuk

prediksi kondisi kesehatan berdasarkan gejala umum COVID-19 menggunakan RapidMiner. Penelitian tersebut menggunakan dataset gejala seperti demam, batuk, sakit tenggorokan, kelelahan, dan sesak napas untuk membantu proses klasifikasi kondisi kesehatan pasien. Hasil penelitian menunjukkan bahwa model Naïve Bayes memperoleh tingkat akurasi sebesar 92,5% sehingga metode tersebut dinilai efektif digunakan dalam membantu proses skrining kesehatan berbasis gejala COVID-19 [5].

Selain pada kasus COVID-19, algoritma Naïve Bayes juga banyak digunakan dalam penelitian kesehatan lainnya. Penelitian oleh Alfa Saleh dkk. membandingkan beberapa algoritma machine learning seperti KNN, SVM, Naïve Bayes, dan Decision Tree untuk memprediksi penyakit diabetes. Hasil penelitian menunjukkan bahwa algoritma machine learning mampu membantu proses prediksi penyakit secara lebih cepat dan akurat berdasarkan atribut kesehatan pasien [6].

Penelitian lain oleh Arum Dewi Ayu Safira Surya membahas penerapan algoritma Naïve Bayes untuk klasifikasi penyakit jantung menggunakan dataset kesehatan dari Kaggle. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mampu menghasilkan performa klasifikasi yang cukup baik dalam membantu proses deteksi penyakit jantung berdasarkan atribut kesehatan pasien [7].

Selain itu, penelitian oleh Eka Wulansari Fridayanthie membahas analisa data mining untuk prediksi penyakit hepatitis menggunakan metode Naïve Bayes dan Support Vector Machine. Penelitian tersebut menunjukkan bahwa algoritma klasifikasi dapat membantu proses diagnosis penyakit secara lebih efektif dan dapat digunakan sebagai sistem pendukung keputusan pada bidang kesehatan [8].

Berdasarkan beberapa penelitian terdahulu tersebut, dapat disimpulkan bahwa metode data mining dan algoritma Naïve Bayes memiliki kemampuan yang cukup baik dalam membantu proses klasifikasi data kesehatan, khususnya data COVID-19. Selain memiliki proses perhitungan yang sederhana, algoritma Naïve Bayes juga mampu menghasilkan tingkat akurasi yang cukup tinggi pada berbagai penelitian sebelumnya. Oleh karena itu, penelitian ini menerapkan algoritma Naïve Bayes menggunakan aplikasi RapidMiner untuk melakukan klasifikasi tingkat penyebaran COVID-19 berdasarkan data statistik yang tersedia agar dapat menghasilkan informasi yang lebih sederhana dan mudah dipahami.

2.2. Data Mining

Data mining merupakan suatu proses penggalian informasi atau pola tertentu dari kumpulan data dalam jumlah besar yang sebelumnya belum diketahui. Proses ini dilakukan agar data yang awalnya hanya berupa kumpulan angka dapat diolah menjadi informasi yang lebih berguna dan mudah dipahami. Data mining banyak digunakan dalam berbagai bidang seperti bisnis, pendidikan, pemerintahan,

kesehatan, hingga teknologi informasi karena mampu membantu proses analisis data secara lebih efektif dan efisien.

Didefinisikan Menurut Rizka Dahlia dkk. data mining digunakan untuk membantu proses analisis penyebaran COVID-19 melalui tahapan preprocessing, transformasi data, hingga evaluasi model klasifikasi menggunakan algoritma machine learning [1]. Selain itu, Ahmad Firdaus dkk. menjelaskan bahwa data mining sangat penting digunakan dalam proses pengolahan data COVID-19 karena banyaknya data pasien yang terus berkembang menyebabkan proses analisis manual menjadi sulit dilakukan [4].

Menurut Eka Wulansari Fridayanthie, data mining merupakan proses menemukan pola, hubungan, dan kecenderungan tertentu dari sekumpulan data menggunakan teknik statistik dan matematika sehingga mampu membantu proses prediksi maupun klasifikasi data kesehatan [8]. Sementara itu, Mohammad Nova Kurniawan dkk. juga menjelaskan bahwa data mining mampu membantu proses pengolahan data kesehatan menjadi informasi yang lebih sederhana dan mudah dipahami berdasarkan pola data yang tersedia [9].

Dalam penelitian ini, data mining digunakan untuk membantu proses klasifikasi tingkat penyebaran COVID-19 berdasarkan atribut pada dataset sehingga data statistik COVID-19 dapat diolah menjadi informasi yang lebih sederhana dan lebih mudah dipahami berdasarkan kategori tingkat penyebaran.

2.3. Algoritma Naïve Bayes

Naïve Bayes merupakan salah satu algoritma klasifikasi yang menggunakan konsep probabilitas berdasarkan Teorema Bayes. Algoritma ini bekerja dengan menghitung kemungkinan suatu data masuk ke dalam kelas tertentu berdasarkan atribut yang dimiliki oleh data tersebut. Naïve Bayes disebut “naïve” karena memiliki asumsi bahwa setiap atribut dianggap independen atau tidak saling memengaruhi satu sama lain [10].

Algoritma Naïve Bayes dikenal memiliki proses perhitungan yang sederhana, mudah diterapkan, serta mampu menghasilkan performa klasifikasi yang cukup baik pada berbagai kasus data mining. Menurut Hetty Rohayani dkk, algoritma Naïve Bayes mampu menghasilkan tingkat akurasi sebesar 97% dalam proses prediksi status pasien COVID-19 menggunakan RapidMiner [2]. Selain itu, Dewi Yanti Liliana dkk juga menunjukkan bahwa Naïve Bayes mampu menghasilkan tingkat akurasi sebesar 96,67% dalam mengklasifikasikan status pasien COVID-19 berdasarkan atribut usia dan jenis kelamin.

Penelitian lain oleh Ahmad Firdaus dkk menunjukkan bahwa model Gaussian Naïve Bayes mampu menghasilkan nilai akurasi sebesar 96,88% dalam melakukan klasifikasi kasus COVID-19 [4]. Selain itu, penelitian Arum Dewi Ayu Safira Surya juga menjelaskan bahwa algoritma Naïve Bayes

cukup efektif digunakan pada klasifikasi penyakit jantung karena mampu memberikan hasil klasifikasi yang cukup baik dengan proses perhitungan yang relatif sederhana.

Rumus umum Teorema Bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Keterangan :

$P(H|X)$ = probabilitas hipotesis H berdasarkan data X

$P(X|H)$ = probabilitas data X berdasarkan hipotesis H

$P(H)$ = probabilitas awal dari hipotesis H

$P(X)$ = probabilitas dari data X

Dalam penelitian ini, algoritma Naïve Bayes digunakan untuk membangun model klasifikasi tingkat penyebaran COVID-19 berdasarkan atribut yang terdapat pada dataset.

2.4. RapidMiner

RapidMiner merupakan aplikasi yang digunakan untuk melakukan proses data mining, machine learning, preprocessing data, klasifikasi, clustering, dan evaluasi model. RapidMiner menggunakan konsep operator visual sehingga pengguna dapat membangun proses pengolahan data dengan cara menghubungkan operator satu dengan operator lainnya tanpa harus menulis kode program secara manual.

Menurut Rizka Dahlia dkk, RapidMiner mampu membantu proses pengolahan data COVID-19 menggunakan berbagai algoritma klasifikasi seperti Naïve Bayes, C4.5, dan K-Nearest Neighbor secara lebih mudah dan terstruktur [1]. Selain itu, Hetty Rohayani dkk juga menjelaskan bahwa RapidMiner dapat digunakan untuk membantu proses prediksi penyebaran COVID-19 melalui tahapan preprocessing data, pembentukan model, hingga evaluasi hasil klasifikasi.

Penelitian Ahmad Roni Husaini juga menjelaskan bahwa RapidMiner merupakan perangkat lunak berbasis visual yang mendukung proses eksplorasi data, pembentukan model machine learning, hingga evaluasi performa model tanpa memerlukan kemampuan pemrograman yang kompleks [5]. Dengan adanya operator visual pada RapidMiner, proses pengolahan data menjadi lebih mudah dipahami terutama dalam penelitian berbasis data mining dan machine learning.

Dalam penelitian ini, RapidMiner digunakan untuk menjalankan seluruh proses klasifikasi tingkat penyebaran COVID-19 mulai dari import dataset, pembentukan label klasifikasi, pembagian data training dan testing, pembentukan model Naïve Bayes, hingga evaluasi hasil klasifikasi menggunakan confusion matrix dan accuracy.

2.5. Klasifikasi

Klasifikasi merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan atribut

yang dimiliki oleh data tersebut. Pada proses klasifikasi, data yang telah memiliki label digunakan sebagai data training untuk membangun model klasifikasi. Setelah model terbentuk, model tersebut dapat digunakan untuk memprediksi kelas atau kategori dari data testing [11].

Menurut Alfa Saleh dkk, teknik klasifikasi berbasis machine learning mampu membantu proses prediksi penyakit secara lebih cepat dan akurat berdasarkan atribut kesehatan pasien [6]. Selain itu, Mahmud Yunus dkk menjelaskan bahwa metode klasifikasi menggunakan Naïve Bayes mampu menghasilkan performa yang cukup baik dalam mengelompokkan data sentimen berdasarkan kategori positif dan negatif dengan tingkat akurasi yang tinggi [9]. Penelitian I Putu Wibina Karsa dkk juga menunjukkan bahwa Naïve Bayes memiliki performa klasifikasi yang cukup baik dibandingkan algoritma lainnya pada proses analisis sentimen.

2.6. COVID-19

COVID-19 merupakan penyakit menular yang disebabkan oleh virus SARS-CoV-2 dan mulai menyebar secara luas sejak akhir tahun 2019. Penyebaran virus yang sangat cepat menyebabkan peningkatan jumlah kasus di berbagai negara sehingga data statistik mengenai COVID-19 terus bertambah setiap harinya. Data tersebut meliputi jumlah total kasus, jumlah pasien sembuh, jumlah kematian, kasus aktif, hingga kondisi pasien kritis.

Menurut Rizka Dahlia dkk, COVID-19 merupakan virus yang menyerang sistem pernapasan manusia dan memiliki tingkat penyebaran yang sangat cepat di berbagai negara [1]. Selain itu, Hetty Rohayani dkk menjelaskan bahwa pandemi COVID-19 menjadi masalah kesehatan yang kompleks sehingga diperlukan dukungan teknologi untuk membantu proses klasifikasi dan analisis data pasien COVID-19 [2].

Dewi Yanti Liliana dkk juga menjelaskan bahwa pandemi COVID-19 memberikan dampak besar pada berbagai sektor kehidupan sehingga diperlukan proses analisis data untuk membantu proses mitigasi dan penanganan pasien secara lebih efektif [3]. Oleh karena itu, dalam penelitian ini data COVID-19 digunakan sebagai dataset utama untuk membantu proses klasifikasi tingkat penyebaran berdasarkan atribut statistik yang tersedia.

3. METODE PENELITIAN

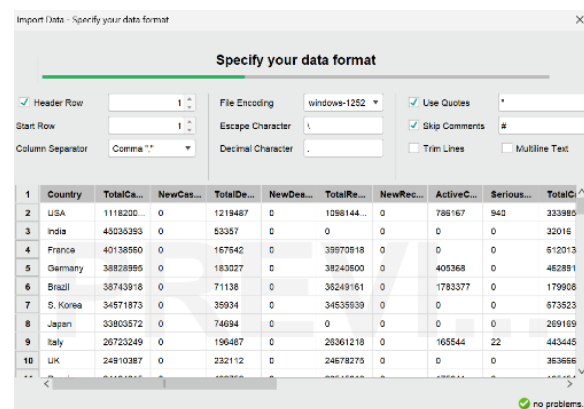
Penelitian ini dilakukan menggunakan aplikasi RapidMiner dengan dataset statistik COVID-19 beberapa negara. Dataset yang digunakan memuat atribut seperti TotalCases, TotalDeaths, TotalRecovered, ActiveCases, SeriousCritical, dan beberapa atribut pendukung lainnya. Penelitian ini bertujuan untuk melakukan klasifikasi tingkat penyebaran COVID-19 menggunakan algoritma Naïve Bayes berbasis RapidMiner sehingga data

statistik COVID-19 dapat diolah menjadi informasi yang lebih sederhana dan mudah dipahami.

Metode penelitian yang digunakan terdiri dari beberapa tahapan, yaitu import dataset, preprocessing data, pembentukan label klasifikasi, pembagian data training dan testing, proses klasifikasi menggunakan algoritma Naïve Bayes, penerapan model, dan evaluasi hasil klasifikasi. Seluruh tahapan dilakukan menggunakan operator yang tersedia pada aplikasi RapidMiner agar proses pengolahan data dapat dilakukan secara lebih terstruktur dan mudah dipahami.

3.1. Import Dataset

Tahap pertama dalam penelitian ini adalah melakukan import dataset ke dalam aplikasi RapidMiner. Dataset yang digunakan berbentuk file CSV yang berisi data statistik COVID-19 dari beberapa negara. Sebelum digunakan pada proses klasifikasi, dataset terlebih dahulu dilakukan pengecekan agar data dapat terbaca dengan baik dan tidak mengalami error saat proses pengolahan data dilakukan.



Gambar 1. Import Data

Pada tahap import dataset, dilakukan penyesuaian tipe data setiap atribut yang terdapat pada dataset. Atribut yang berisi teks seperti Country menggunakan tipe data polynominal, sedangkan atribut yang berisi angka seperti TotalCases, TotalDeaths, TotalRecovered, ActiveCases, dan SeriousCritical menggunakan tipe data integer atau numerical. Penyesuaian tipe data ini dilakukan agar proses klasifikasi dapat berjalan dengan baik pada aplikasi RapidMiner.

3.2. Preprocessing Data

Tahap preprocessing data dilakukan untuk memastikan data yang digunakan sudah sesuai dan siap digunakan pada proses klasifikasi. Pada tahap ini dilakukan pengecekan missing value, penyesuaian tipe data, serta pemilihan atribut yang akan digunakan dalam proses klasifikasi. Proses preprocessing sangat penting dilakukan karena kualitas data akan memengaruhi hasil klasifikasi yang diperoleh.

Selain itu, preprocessing data juga dilakukan agar dataset menjadi lebih terstruktur dan mudah diproses oleh algoritma Naïve Bayes. Data yang tidak sesuai atau memiliki kesalahan dapat menyebabkan proses klasifikasi menjadi kurang optimal sehingga tahap preprocessing menjadi salah satu tahapan penting dalam penelitian ini.

3.3. Generate Attributes

Operator Generate Attributes digunakan untuk membuat atribut baru pada dataset. Dalam penelitian ini, atribut baru yang dibuat adalah Tingkat_Penyebaran. Atribut tersebut digunakan sebagai label klasifikasi untuk menentukan apakah tingkat penyebaran COVID-19 termasuk kategori Tinggi atau Rendah.

Parameter yang digunakan pada operator Generate Attributes adalah sebagai berikut:

```
Name      : Tingkat_Penyebaran
Function   if(TotalCases                >=
Expression : 1000000,"Tinggi","Rendah")
```

Parameter Name digunakan untuk memberikan nama atribut baru, yaitu Tingkat_Penyebaran. Sedangkan Function Expression digunakan untuk menentukan aturan pembentukan kategori berdasarkan jumlah TotalCases pada dataset. Jika nilai TotalCases lebih dari atau sama dengan 1.000.000 maka data dikategorikan sebagai Tinggi, sedangkan jika nilai TotalCases kurang dari 1.000.000 maka data dikategorikan sebagai Rendah.

3.4. Set Role

Operator Set Role digunakan untuk menentukan atribut label pada proses klasifikasi. Pada penelitian ini, atribut Tingkat_Penyebaran ditentukan sebagai label atau target klasifikasi yang akan diprediksi oleh algoritma Naïve Bayes.

Parameter yang digunakan pada operator Set Role adalah sebagai berikut:

```
attribute name : Tingkat_Penyebaran
target role    : label
```

Tahap ini dilakukan agar RapidMiner dapat mengetahui atribut mana yang digunakan sebagai target prediksi pada proses klasifikasi. Tanpa adanya proses Set Role, algoritma Naïve Bayes tidak dapat menentukan atribut yang akan digunakan sebagai label klasifikasi.

3.5. Split Data

Operator Split Data digunakan untuk membagi dataset menjadi dua bagian, yaitu data training dan data testing. Data training digunakan untuk membangun model klasifikasi, sedangkan data testing digunakan untuk menguji performa model yang telah dibuat.

Parameter yang digunakan pada operator Split Data adalah sebagai berikut:

```
partions      : 0.7 dan 0.3
sampling type  : shuffled
use local random seed : true
local random seed : 1992
```

Nilai 0.7 menunjukkan bahwa 70% data digunakan sebagai data training, sedangkan nilai 0.3 menunjukkan bahwa 30% data digunakan sebagai data testing. Pembagian data dilakukan agar model dapat belajar dari sebagian data dan diuji menggunakan data yang berbeda sehingga hasil evaluasi menjadi lebih baik.

3.6. Naïve Bayes

Operator Naïve Bayes digunakan untuk membangun model klasifikasi berdasarkan data training yang telah dipersiapkan sebelumnya. Pada tahap ini algoritma Naïve Bayes akan mempelajari pola data berdasarkan atribut yang terdapat pada dataset untuk membantu proses klasifikasi tingkat penyebaran COVID-19.

Parameter yang digunakan pada operator Naïve Bayes adalah sebagai berikut:

```
laplace correction : true
calculate confidences : ttrue
```

Parameter laplace correction digunakan untuk mengurangi kemungkinan munculnya probabilitas nol pada proses klasifikasi. Sedangkan parameter calculate confidences digunakan untuk menampilkan tingkat keyakinan model terhadap hasil prediksi yang diperoleh.

3.7. Apply Model

Operator Apply Model digunakan untuk menerapkan model Naïve Bayes yang telah dibentuk ke dalam data testing. Tahap ini dilakukan untuk melihat kemampuan model dalam melakukan prediksi terhadap data baru yang sebelumnya tidak digunakan pada proses training.

Output dari operator Apply Model berupa hasil prediksi data testing yang meliputi prediction dan confidence dari setiap data yang diproses menggunakan model Naïve Bayes.

3.8. Performance Classification

Operator Performance Classification digunakan untuk melakukan evaluasi terhadap hasil klasifikasi yang diperoleh dari model Naïve Bayes. Evaluasi dilakukan untuk mengetahui tingkat performa model dalam melakukan klasifikasi tingkat penyebaran COVID-19.

Parameter yang digunakan pada operator Performance Classification adalah sebagai berikut:

```
main criterion      : accuracy
skip undefined labels : true
use example weights : true
```

Pada penelitian ini, accuracy digunakan sebagai parameter utama dalam proses evaluasi model. Selain accuracy, evaluasi juga dilakukan menggunakan confusion matrix untuk mengetahui jumlah data yang berhasil diklasifikasikan dengan benar maupun data yang masih mengalami kesalahan klasifikasi.

4. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan menggunakan aplikasi RapidMiner dengan menyusun alur proses klasifikasi menggunakan beberapa operator, yaitu Retrieve, Generate Attributes, Set Role, Split Data, Naïve Bayes, Apply Model, dan Performance Classification. Seluruh operator tersebut saling terhubung untuk membangun proses klasifikasi tingkat penyebaran COVID-19 berdasarkan dataset statistik COVID-19 beberapa negara.

Tahap awal penelitian dilakukan dengan mengimpor dataset COVID-19 ke dalam RapidMiner. Dataset yang digunakan memuat beberapa atribut seperti TotalCases, TotalDeaths, TotalRecovered, ActiveCases, dan SeriousCritical. Setelah dataset berhasil diimpor, dilakukan pengecekan tipe data untuk memastikan seluruh atribut dapat terbaca dengan benar pada aplikasi RapidMiner. Atribut berupa teks seperti Country menggunakan tipe data polynominal, sedangkan atribut numerik seperti TotalCases dan TotalDeaths menggunakan tipe data integer atau numerical.

Setelah proses import dataset selesai dilakukan, tahap berikutnya adalah preprocessing data. Pada tahap ini dilakukan pengecekan data yang kosong, penyesuaian tipe data, serta pemilihan atribut yang digunakan dalam proses klasifikasi. Tahap preprocessing dilakukan agar data yang digunakan menjadi lebih terstruktur dan siap digunakan pada proses pembentukan model klasifikasi menggunakan algoritma Naïve Bayes.

Format your columns.

Date format: Enter value... ☐ Replace errors with missing values

	Country	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecov...
	polynominal	Integer	Integer	Integer	Integer	Integer
1	USA	111620082	0	1219487	0	109614428
2	India	45035393	0	53357	0	0
3	France	40138560	0	167542	0	39970918
4	Germany	38828995	0	183027	0	38240600
5	Brazil	38743918	0	71138	0	36249161
6	S. Korea	34571873	0	35934	0	34535939
7	Japan	33803572	0	74694	0	0
8	Italy	28723249	0	196487	0	26361218
9	UK	24910387	0	232112	0	24678275
10	Russia	24124215	0	402756	0	23545818
11	Turkey	17232066	0	102174	0	0

no problems.

Gambar 2. Preprocessing dan penyesuaian data

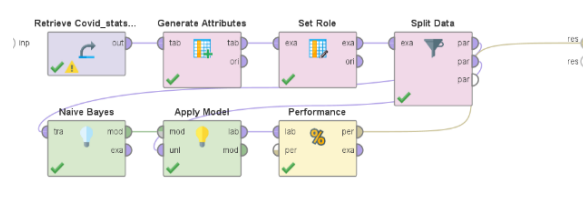
Tahap selanjutnya dilakukan pembentukan label klasifikasi menggunakan operator Generate Attributes. Label yang dibuat pada penelitian ini adalah Tingkat_Penyebaran dengan dua kategori, yaitu Tinggi dan Rendah. Kategori Tinggi diberikan pada data dengan nilai TotalCases lebih dari atau sama dengan 1.000.000, sedangkan kategori Rendah

diberikan pada data dengan nilai TotalCases kurang dari 1.000.000. Pembentukan label ini dilakukan agar dataset memiliki target klasifikasi yang dapat digunakan oleh algoritma Naïve Bayes.

Setelah label berhasil dibuat, operator Set Role digunakan untuk menentukan atribut Tingkat_Penyebaran sebagai label atau target klasifikasi. Tahap ini sangat penting karena RapidMiner harus mengetahui atribut mana yang akan digunakan sebagai target prediksi pada proses klasifikasi. Tanpa adanya proses Set Role, algoritma Naïve Bayes tidak dapat melakukan proses pembelajaran data secara optimal.

Tahap berikutnya dilakukan pembagian dataset menggunakan operator Split Data. Dataset dibagi menjadi dua bagian, yaitu 70% data training dan 30% data testing. Data training digunakan untuk membangun model klasifikasi, sedangkan data testing digunakan untuk menguji kemampuan model dalam melakukan prediksi terhadap data baru. Proses pembagian data dilakukan menggunakan metode shuffled agar data diacak terlebih dahulu sehingga pembagian data menjadi lebih merata.

Pada tahap pembentukan model, operator Naïve Bayes digunakan untuk mempelajari pola data berdasarkan atribut yang terdapat pada dataset. Algoritma Naïve Bayes bekerja menggunakan pendekatan probabilitas untuk menentukan kemungkinan suatu data masuk ke dalam kategori Tinggi atau Rendah. Selain memiliki proses perhitungan yang sederhana, algoritma ini juga mampu menghasilkan performa klasifikasi yang cukup baik pada berbagai penelitian sebelumnya.



Gambar 3. Tahapan Operator

Model yang telah dibangun kemudian diterapkan pada data testing menggunakan operator Apply Model. Tahap ini dilakukan untuk melihat kemampuan model dalam melakukan prediksi terhadap data yang sebelumnya tidak digunakan pada proses training. Hasil dari operator Apply Model berupa prediction dan confidence yang menunjukkan tingkat keyakinan model terhadap hasil prediksi yang diperoleh.

Proses evaluasi model dilakukan menggunakan operator Performance Classification. Evaluasi dilakukan menggunakan confusion matrix dan nilai accuracy untuk mengetahui tingkat performa model klasifikasi. Berdasarkan hasil pengujian menggunakan RapidMiner, algoritma Naïve Bayes memperoleh nilai accuracy sebesar 89.86%. Hasil tersebut menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model yang telah dibangun.

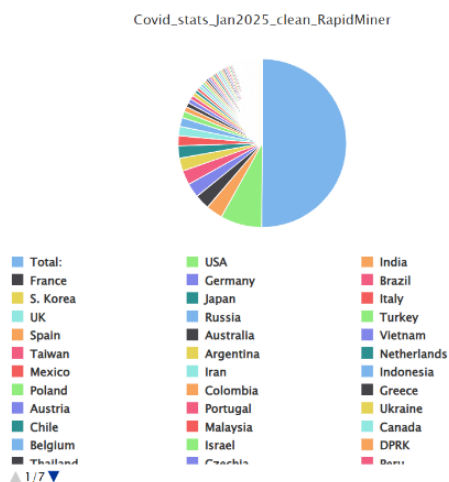
Berdasarkan hasil confusion matrix, terdapat sebagian besar data kategori Tinggi dan Rendah yang berhasil diprediksi dengan benar oleh model Naïve Bayes. Selain itu, hasil precision dan recall juga menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengenali masing-masing kategori tingkat penyebaran COVID-19. Nilai precision yang tinggi menunjukkan bahwa hasil prediksi model cukup akurat, sedangkan nilai recall menunjukkan bahwa model mampu mengenali sebagian besar data sesuai kategori sebenarnya.

accuracy: 89.86%

	true Tinggi	true Rendah	class precision
pred. Tinggi	19	1	95.00%
pred. Rendah	6	43	87.76%
class recall	76.00%	97.73%	

Gambar 4. Accuracy

Hasil penelitian ini menunjukkan bahwa algoritma Naïve Bayes mampu digunakan untuk melakukan klasifikasi tingkat penyebaran COVID-19 dengan cukup baik. Selain memiliki proses perhitungan yang sederhana, algoritma ini juga mampu menghasilkan tingkat akurasi yang cukup tinggi pada proses klasifikasi data kesehatan berbasis machine learning. Penggunaan RapidMiner juga membantu proses analisis data menjadi lebih mudah karena seluruh tahapan dapat dilakukan menggunakan operator visual tanpa harus melakukan pemrograman secara manual.



Gambar 5. Pie Chart Diagram

Berdasarkan hasil penelitian yang telah dilakukan, data statistik COVID-19 dapat diolah menjadi informasi yang lebih sederhana melalui proses klasifikasi. Dengan adanya kategori Tinggi dan Rendah, data penyebaran COVID-19 menjadi lebih mudah dipahami dibandingkan hanya berupa angka statistik. Oleh karena itu, penerapan algoritma Naïve Bayes berbasis RapidMiner dapat digunakan sebagai salah satu pendekatan dalam membantu proses analisis data COVID-19.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, algoritma Naïve Bayes dapat digunakan untuk melakukan klasifikasi tingkat penyebaran COVID-19 menggunakan aplikasi RapidMiner

Proses penelitian dilakukan melalui beberapa tahapan, yaitu import dataset, preprocessing data, pembentukan label klasifikasi menggunakan Generate Attributes, penentuan label menggunakan Set Role, pembagian data menggunakan Split Data, pembentukan model menggunakan Naïve Bayes, penerapan model menggunakan Apply Model, dan evaluasi menggunakan Performance Classification.

Berdasarkan hasil pengujian menggunakan RapidMiner, model klasifikasi memperoleh nilai accuracy sebesar 89.86%. Hasil tersebut menunjukkan bahwa algoritma Naïve Bayes mampu melakukan klasifikasi tingkat penyebaran COVID-19 dengan cukup baik berdasarkan dataset yang digunakan pada penelitian ini.

Selain itu, hasil confusion matrix, precision, dan recall juga menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar ke dalam kategori Tinggi maupun Rendah. Dengan demikian, algoritma Naïve Bayes dapat digunakan sebagai salah satu metode klasifikasi dalam membantu proses analisis data COVID-19 berbasis machine learning menggunakan RapidMiner.

Penelitian ini masih dapat dikembangkan kembali pada penelitian selanjutnya dengan menggunakan jumlah dataset yang lebih besar dan atribut yang lebih beragam agar hasil klasifikasi menjadi lebih optimal. Selain itu, penelitian selanjutnya juga dapat membandingkan algoritma Naïve Bayes dengan algoritma klasifikasi lainnya seperti Decision Tree, K-Nearest Neighbor, Support Vector Machine, maupun Random Forest untuk mengetahui algoritma yang memiliki performa terbaik dalam melakukan klasifikasi data COVID-19.

Selain menggunakan aplikasi RapidMiner, penelitian selanjutnya juga dapat dikembangkan menggunakan aplikasi atau bahasa pemrograman lain seperti Python, WEKA, maupun Orange Data Mining agar proses analisis data menjadi lebih luas dan lebih fleksibel. Dengan adanya pengembangan tersebut diharapkan hasil penelitian mengenai klasifikasi data COVID-19 dapat menjadi lebih baik dan lebih akurat.

DAFTAR PUSTAKA

- [1] Dahlia, R., Wuryani, N., Hadiani, S., Gata, W., & Selawati, A. (2021). Penerapan Data Mining Terhadap Data Covid-19 Menggunakan Algoritma Klasifikasi. *Jurnal Informatika*, 21(1), 44-52..
- [2] Rohayani, H., Nur, S. A., & Fauzi, M. (2022). Rico, "Prediksi Penyebaran Virus COVID-19 Dari Hasil PCR Menggunakan Metode Naïve Bayes,". *J. Comput. Syst. Informatics*, 4(1), 109-115.
- [3] Liliana, D. Y., Maulana, H., & Setiawan, A.

- (2021). Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes. *MULTINETICS*, 7(1), 48-53..
- [4] Firdaus, A., Walid, M., & Anwari, A. (2022). Klasifikasi Kasus Covid-19 Menggunakan Model Naive Bayes Classifier. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 6(2), 583-588..
- [5] Husaini, A. R. (2025). Klasifikasi Berbasis Machine Learning untuk Prediksi Kondisi Kesehatan Berdasarkan Gejala Umum COVID-19. *EJECTS: Journal Computer, Technology, and Informations System*, 5(1), 13-18.
- [6] Saleh, A., Sari, R. E., Ramadani, R., Fujiati, F., & Lestari, R. (2025). Analisis Komparatif Algoritma Klasifikasi Machine Learning untuk Memprediksi Diabetes. *Algoritma: Jurnal Ilmu Komputer dan Informatika*, 9(1), 26-35..
- [7] Surya, A. A., & Yamasari, Y. (2024). Penerapan Algoritma Naïve Bayes (NB) untuk Klasifikasi Penyakit Jantung. *Journal of Informatics and Computer Science (JINACS)*, 5(03), 447-455.
- [8] Fridayanthie, E. W. (2015). Analisa Data Mining Untuk Prediksi Penyakit Hepatitis Dengan Menggunakan Metode Naive Bayes dan Support Vector Machine. *Jurnal Khatulistiwa Informatika*, 3(1), 24-36..
- [9] Yunus, M., Husni, M., & Mufadhdhal, M. M. (2021). Klasifikasi Sentimen Terhadap Badan Penyelenggara Jaminan Sosial (BPJS) Pada Media Sosial Twitter Menggunakan Naive Bayes. *Smatika Jurnal*, 11(02), 81-91..
- [10] Surojudin, N. (2025). Perbandingan Metode Klasifikasi dalam Memprediksi Penyakit Ginjal Kronis. *Jurnal Informatika Ekonomi Bisnis*, 715-723..
- [11] Syafrianto, A. (2022). Perbandingan Algoritma Naïve Bayes dan Decision Tree Pada Sentimen Analisis. *The Indonesian Journal of Computer Science Research*, 1(2)..