Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 5, Oktober 2022, hlm. 1397-1405

DOI: https://doi.org/10.20884/1.jutif.2022.3.5.561
p-ISSN: 2723-3863
e-ISSN: 2723-3871

APPLICATION OF DATA MINING FOR PREDICTION OF LONG COVID ON COVID-19 SURVIVAL WITH FEATURE SELECTION AND NAÏVE BAYES METHOD

Siti Rokhmah[*1], Nendy Akbar Rozaq Rais[2]

[1,2]Informatika, Fakultas Teknologi, Institut Teknologi Bisnis AAS Indonesia
Email: [1]sitirokhmah.itbaas@gmail.com , [2]abe.terate@gmail.ac.id

(Naskah masuk: 18 Agustus 2022, Revisi: 1 September 2022, diterbitkan: 24 Oktober 2022)

Abstract

Since it was declared a global pandemic in March 2020, Corona Virus Disease (Covid-19) has become the world's attention, a lot of research has focused on all things related to Covid-19. Covid-19 is an infectious disease caused by the acute respiratory syndrome Corona virus 2 (SARS-CoV-2). Several studies have shown that the symptoms of COVID-19 persist for a long period of time even though they have been declared cured of Covid-19, this is known as Long Covid. Complaints that are often experienced by patients who progress to Long Covid are fatigue, headaches, coughs, runny noses, sleep disturbances and even shortness of breath. Several risk factors for the occurrence of Long Covid include age, gender, patient congenital disease, condition during acute infection, ethnicity and the patient's Body Mass Index (BMI). To anticipate the occurrence of long covid, it is necessary to have a risk prediction system for the occurrence of long covid in covid-19 patients, this aims to anticipate and prepare for early handling and prevention efforts for covid-19 patients who are at risk of experiencing long covid. Prediction of the risk of long covid can be done by classifying long covid risk factors by utilizing data mining. The purpose of this study is to classify symptom data and patient history, so that data patterns can be obtained that can be used as predictions to estimate the risk of the occurrence of Long covid-19 in Covid-19 survivors. This study uses the Naïve Bayes classification method by classifying data based on the Long covid risk factor and the feature selection information gain method which is used as a technique in attribute selection to optimize the nave Bayes algorithm. The results of this study have a real contribution to the development of science and technology. The concept of the resulting prediction data pattern can be used as a reference in developing early detection of the risk of the occurrence of Long Covid.

Keywords: *, Data Clasification, Data mining, Featture Selection , Long covid, , naïve bayes.*

PENERAPAN DATA MINING UNTUK PREDIKSI *LONG COVID* PADA PENYINTAS COVID-19 DENGAN METODE *FEATURE SELECTION* DAN NAÏVE BAYES

Abstrak

Sejak ditetapkannya menjadi pandemi global pada maret 2020, *Corona Virus Disease* (Covid-19) mejadi perhatian dunia, banyak penelitian difokuskan pada segala hal terkait covid-19. Covid-19 adalah penyakit menular yang disebabkan oleh sindrom pernapasan akut *Corona virus 2* (SARS-CoV-2). Beberapa penelitian menunjukkan bahwa gejala covid-19 tetap ada dalam jangka waktu yang lama meskipun sudah dinyatakan sembuh dari Covid-19, hal tersebut dikenal sebagai *Long covid*. Keluhan yang sering dialami oleh pasien yang berlanjut ke *Long covid* adalah kelelahan, nyeri kepala, batuk, pilek, gangguan tidur bahkan sesak nafas. Beberapa faktor resiko terjadinya *Long covid* diantaranya adalah usia, jenis kelamin, Penyakit bawaan pasien, kondisi saat infeksi akut, etnis dan *Body Mass Index* (BMI) pasien. Untuk mengantisipasi terjadinya *Long covid*, perlu adanya sistem prediksi resiko terjadinya *Long covid* pada pasien covid-19, hal ini bertujuan untuk mengantisipasi dan mempersiapkan upaya penanganan dan pencegahan sejak dini terhadap pasien covid-19 yang beresiko mengalami *long covid*. prediksi resiko terjadinya long covid dapat dilakukan dengan pengklasifikasian faktor resiko long covid dengan memafaatkan Data mining. Tujuan dari penelitian ini adalah melakukan klasifikasi data gejala dan riwayat pasien, sehingga dapat diperoleh pola data yang dapat dijadikan prediksi untuk memperkirakan resiko terjadinya *Long covid*-19 pada pasien penyintas covid-19. Penelitian ini menggunakan metode klasifikasi Naïve Bayes dengan mengklasifikasikan data berdasarkan faktor resiko *Long covid* dan metode *feature selection information gain* yang digunakan sebagai teknik dalam pemilihan atribut untuk mengoptimalkan algoritma naïve bayes. Hasil penelitian ini memiliki kontribusi yang nyata bagi pengembangan ilmu pengetahuan dan teknologi. Konsep Pola data prediksi yang dihasilkan dapat digunakan sebagai acuan dalam mengembangkan deteksi dini resiko terjadinya *Long covid*.

Kata kunci: *Long covid*, d*ata mining, Feature selection, klasifikasi data, naïve bayes*

1. PENDAHULUAN

World Health Organization (WHO) resmi menetapkan Covid-19 sebagai pandemi global pada Maret 2020. Sejak saat itu banyak penelitian di fokuskkan pada segala sesuatu terkait Covid-19. Covid-19 adalah penyakit menular yang disebabkan oleh sindrom pernapasan akut *SARS-CoV-2* [1]. Gejala Covid-19 diantaranya demam, nyeri otot, menurunnya kemampuan membau dan *pneounomia* [2]. Virus Sar-Cov-2 dapat menyebabkan berbagai kerusakan organ, penelitian menunjukkan berbagai gejala tetap ada setelah seseorang dinyatakan sembuh dari covid-19 yang kemudian di kenal dengan istilah *Long covid* atau *post-syndrom* Covid-19.[3]

Long covid di definisikan sebagai sindroma pasca covid-19 dengan gejala sakit yang berkepanjangan meskipun sudah 12 minggu dari pertama terpapar dan telah dinyatakan sembuh[4]. Gejala yang sering dirasakan adalah kelelahan, pilek, batuk bahkan ada yang masih mengalami sesak nafas [5]. Beberapa data menunjukkan bahwa siapapun yang pernah terinfeksi Covid-19 beresiko mengalami *Long covid*, terutama orang dengan kekebalan rendah terhadap infeksi [6]. Pasien covid-19 yang mendapat perawatan intensif di ICU menunjukkan lebih berpotensi mengalami *Long covid* [7].

Penelitian *Long covid* sudah banyak dikembangkan, dari potensi resiko sampai gejala, namun belum ada penelitian yang melakukan klasifikasi data sebagai dasar prediksi resiko terjadinya *long covid,* sehingga penelitian ini perlu dilakukan sebagai salah satu upaya dalam membantu penanganan pandemi Covid-19. *Long covid* dapat diprediksi dari beberapa faktor resiko, diantaranya dari gejala awal saat terpapar covid-19, kondisi pasien dan penyakit penyerta pasien [8]. Data-data tersebut dapat diklasifikasikan dengan menggunakan teknik data mining untuk memperoleh gambaran prediksi resiko *Long covid* pada penyintas Covid-19. Data mining digunakan untuk menemukan pola dari suatu database dalam ukuran yang besar. Data mining merupakan proses menemukan korelasi, pola dan sebuah tren dengan cara menyaring data dan di olah mengunakan teknik matematis dan statistik.[9]. Prediksi terhadap resiko terajadinya *Long covid* penting untuk dilakukan agar dapat meminimalkan resiko bagi penyintas covid-19 yang mengalami *Long covid*.

Pada penelitian ini dilakukan pengklasifikasian data dari penyintas covid-19 dengan menggunakan algoritma naïve bayes, algoritma naïve bayes telah banyak digunakan dalam klasifikasi data dikarenakan naïve bayes memiliki tingkat akurasi yang lebih baik dibandingkan metode klasifikasi lainnya. Naïve bayes merupakan teknik pengklasifikasian dengan melakukan prediksi peluang di masa depan berdasarkan pengalaman dimasa lalu [10].

Tujuan khusus dari penelitian ini adalah mengembangkan konsep data mining untuk klasifikasi gejala Covid-19 yang berpotensi memiliki resiko *Long covid* dan Mendapatkan konsep pola data gejala coivd-19 untuk memperoleh prediksi resiko *Long covid*. Penelitian ini perlu dilakukan karena memiliki beberapa keutamaan. Peneletian terkait data sudah banyak dilakukan baik didalam maupun luar negeri, namun belum di jumpai penelitian terkait pemanfaatan data mininig untuk memprediksi resiko *Long covid*. beberapa keutamaan dari penelitian ini diantaranya 1)Penelitan ini memberikan informasi gambaran prediksi kemungkinan resiko terjadinya *Long covid* berdasarkan klasifikasi data; 2)Berdasarkan konsep prediksi pola data pada penelitian ini dapat dikembangkan untuk deteksi dini resiko *Long covid*; 3)Penelitian ini dapat memberikan kontribusi pada perkembangan dan inovasi di bidang teknologi informasi, hal tersebut sejalan dengan bidang fokus penelitian prodi Informatika.

2. METODE PENELITIAN

Ada beberapa tahapan dalam penelitian ini, yaitu

1. Studi pustaka

   Dalam penelitian ini mempelajari referensi berupa jurnal, buku maupun artikel lain yang terkait dengan penelitian. Adapun jurnal yang dijadikan referensi adalah jurnal yang membahas tentang Data mining, naïve bayes, teknik pediksi, Covid-19 dan *Long covid*. Selain itu juga buku dan atikel tentang Data mining dan metode klasifikasi Naïve Bayes dijadikan referensi dalam penulisan penelitian ini.
2. Pengumpulan data

   Data yang digunakan dalam penelitian ini adalah dataset gejala dan data pasien covid-19 yang di unduh dari kaggle.com dengan memasukkan kata kunci: “dataset covid-19 case”, “dataset *Long covid*” dan “datasate covid-19 symptomps”. Data yang diolah adalah data asal negara, usia, jenis kelamin, gejala awal, riwayat penyakit dan kontak penularan. Setelah data terkumpul dilakukan pembersihan data terhadap data yang inkonsisten atau dikenal dengan istilah pemberihan *noise data*.

3. Pra pemrosesan data

   Data dari tahap pengumpulan data diolah dengan prinsip KDD. Beberapa tahapan dalam penelitian ini adalah

   a. Seleksi data

Tidak semua data yang diperoleh akan digunakan, data yang diambil adalah data yang sesuai dengan faktor resiko *long covid*,. Untuk melihat gambaran data awal yang diperoleh dapat melihat pada tabel 1.

Tabel 1. Dataset *long covid*

Data ke	Usia	Jenis kelamin	Penyakit bawaan	Kondisi 2 minggu yang lalu				Kondisi saat ini (gejala *long covid*)			BMI
				Kondisi 1	Kondisi 2	…..	Kondisi 15	gejala 1		gejala10	
1	30	male	hipertensi	1	1	…	1	1	…	1	30
2	65	male	Tidak ada	0	0	…	0	…		….	18
…	….	….		…..	….	….		….		…..	…
1022	40	female		0	….	1		0	..	0	25

Keterangan dataset:

1. Kondisi 2 minggu yang lalu merupakan kondisi saat infeksi, yang terdiri dari 15 kondisi yaitu Sesak nafas, batuk terus menerus, demam, sakit kepala, nafas pendek, hilang penciuman, batuk kering, dahak, kehilangan nafsu makan, sembelit, diare, ageusia, sakit dada, mual muntah, lama waktu perawatan.
2. Kondisi saat ini (gejala *long covid*) ada 10 gejala yaitu pusing, batuk, sesak nafas, nafas pendek, nyeri dada, diare, lemas, kehilangan penciuman, kebingungan, gangguna kecemasan.

b. Data transformasi

Data yang sudah diseleksi selanjutnya di proses pada tahap transformasi, pada tahap ini dilakukan pengolahan menjadi beberapa kriteria sesuai dengan faktor resiko *long covid*. Data hasil proses transformasi dapat dilihat pada tabel 2.

Tabel 2. Data hasil transformasi

data ke	jenis kelamin	Usia	kondisi saat infeksi	Komorbid	BMI	status
1	male	49	berat	Tidak ada	30	YA
2	male	43	sedang	Tidak ada	20	Tidak
3	male	66	berat	Tidak ada	30	YA
4	male	45	ringan	Tidak ada	18	Tidak
5	female	34	berat	Ada	30	YA
6	female	37	berat	Ada	30	YA
7	male	48	ringan	Tidak ada	18	Tidak
8	male	62	berat	Tidak ada	30	YA
9	male	68	sedang	Ada	20	YA
…..	……	….	….	….	…	…
1022	female	29	berat	Tidak ada	18	Tidak

Berikut kriteria dalam pengklasifikasina data

- Kriteria jenis kelamin

Pada kriteria jenis kelamin di bedakan menjadi dua kategori yaitu *female* untuk pasien dengan jenis kelamin perempuan dan *male* untuk jenis kelamin laki-laki.

- Kriteria usia

Atribut pada data usia merupakan data kontinyu, dimana nilai P(Xi|C) di estimasi dengan fungsi densitas gauss. Penyelesaian perhitungan dilakukan dengan menghitung nilai rata-rata dan menghitung stadar deviasi dengan rumus persamaan (1)

$$f(x) = \frac{1}{2\pi\sigma} e^{\frac{-(x-\mu)2}{2\sigma 2}} \quad (1)$$

Keterangan :

- σ = standar Deviasi dari atribut
- μ = mean atau rata-rata dari atribut

- Kriteria kondisi saat infeksi.

Kondisi saat infeksi diperoleh dari data kondisi pasien 2 minggu yang lalu, atau saat terjangkit covid-19. Untuk seleksi data ke dalam kriteria ini dilakukan beberapa aturan, aturan dalam kriteria ini dapat dilihat pada tabel 3.

Tabel 3. Kriteria kondisi saat infeksi

kriteria	Kategori
Memenuhi kriteria gejala berat	Berat
Memenuhi kriteria gejala sedang	Sedang
Memenuhi kriteria gejala ringan	Ringan

Penentuan kriteria didasarkan pada buku pedoman tata laksana yang dikeluarkan oleh perhimpunan dokter indonesia [11]. Dalam pedoman tersebut terdapat 3 kriteria yang dijadikan dasar dalam penentuan kondisi saat infeksi diantaranya adalah gejala berat dengan kriteria mengalami *pneumonia* (demam, batuk, sesak nafas, nafas cepat), nafas berat dan gejala umum covid seperti anosmia,

diare dan kepala pusing. Kriteria sedang memiliki gejala demam, batuk, sesak nafas, nafas cepat, *anosmia* atau hilang kemampuan penciuman, diare namun tidak ditemui gejala *pneumonia* berat. Sedangkan derajat ringan adalah pasien dnegan gejala umum dan ringan seperti demam, batuk, diare, sakit kepala dan *anosmia.*

- Kriteria *Body Mass Index* (BMI)

Sama halnya dengan kriteria usia, kriteria BMI merupakan atribut kontinyu, dimana perhitungan dilakukan dengan menggunakan fungsi densitas gauss.

- Kriteria *komorbid*

Komorbid diperoleh dari kondisi sebelum infeksi covid-19, pada kondisi ini di peroleh kriteria yang terdapat pada tabel 4.

Tabel 4. Kriteria *Komorbid*

kriteria	Kategori
Ada riwayat penyakit sebelum covid	Berat
Tidak ada riwayat penyakit sebelum covid	sedang

- Kriteria status *long covid*

Didalam menentukan seleksi data pada kriteria status didasarkan pada gejala yang timbul pada pasien *long covid* diantaranya adalah kelelahan, batuk terus menerus, sesak nafas, nafas pendek, nyeri otot dan suara serak. Untuk data yang terdapat gejala tersebut maka statusnya iya dan jika tidak ada gejala tersebut maka status tidak. Untuk melihat penentuan kriteria status *long covid* -19 dapat melihat pada tabel 5.

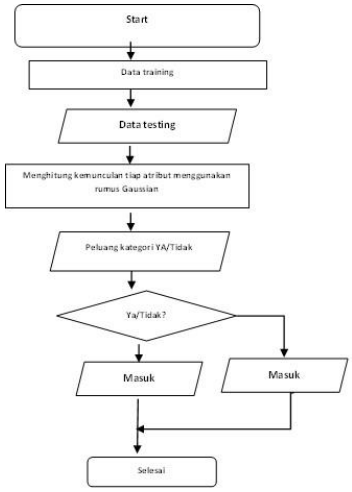
Tabel 5. Kriteria status *long covid*

kriteria	Kategori
Memiliki gejala *long covid*	Iya
Tidak memiliki gejala *long covid*	Tidak

4. Pemodelan dan Klasifikasi data

data yang diperoleh pada tahap pengumpulan data dan transformasi data dilakukan pengolahan data dengan metode klasifikasi naïve bayes. Hasil klasifikasi tersebut dapat dijadikan gambaran prediksi resiko terjadinya *Long covid* pada pasien penyintas Covid-19.

a. Algoritma naïve bayes.



Gambar 1. Flowchart naïve bayes

Algoritma naïve bayes merupakan metode pengklasifikasian yang didasarkan pada teorama bayes, teori ini menggunakan pendekatan statistik yang fundamental dalam pengenalan pola [12]. Naïve bayes dapat digunakan untuk memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Teorama bayes dikombinasikan dengan naïve dimana di asumsikan dengan kondisi antar atribut bisa saling bebas [13]. *Flowchart* naïve bayes digambarkan pada gambar 1.

b. Data *training* dan data *testing*.

Data yang telah di olah pada tahap transformasi kemudian di olah menjadi data *training* dan data *testing*.

Data *training* adalah data yang digunakan untuk melatih algotitma, sedangkan data *testing* adalah data yang digunakan untuk menguji algoritma. Persentase data *training* adalah 75 persen dari keseluruhan data dan data *testing* 25 peresen dari presentase data.

5. Pemilihan atribut (*Feature selection*)

Adanya atribut yang tidak relevan merupakan permasalahan dalam klasifikasi data. Terutama pada prediksi resiko terjadinya *long covid* pada penyintas covid-19. Untuk menghilangkan atribut-atribut yang tidak relevan diperlukan teknik seleksi atribut [14]. Feature Selection digunakan untuk mengurangi dimensi model dan menjadikan klasifikasi data menjadi lebih efektif dengan adanya pengurangan data yang dianalisa[15]. Pada penelitian sebelumnya terdapat dua teknik dalam metode *Feature selection* yaitu *wrapped* dan *filter*. Penelitian tersebut membandingkan dua teknik tersebut dan dihasilkan bahawa teknik filter lebih baik dibandingkan teknik *wrapped*. Sehingga dalam penelitian ini dalam melakukan seleksi atribut menggunakan teknik filter yaitu *information gain*[16].

Information gain merupakan metode dalam *feature selection* yang menggunakan teknik *scoring* untuk pembobotan sebuah fitur dengan menggunakan nilai entropi yang memiliki nilai maksimal. Untuk memghitung *information gain* dapat dihitung dengan persamaan (2), (3) dan (4)

$$Info(D) = \sum_{i=1}^{c} -pi\ log_2\ pi \quad (2)$$

Keterangan
C = jumlah nilai pada atribut target
Pi = jumlah sampai untuk kelas i

$$infoA(D) = \sum_{i=1}^{c} -pi\ log_2\ pi \quad (3)$$

Keterangan
A = atribut
|D| = jumlah dari seluruh sample data
|Dj|= jumah sample data untuk nilai j
V = suatu nilai yang mungkin untuk atribut A

Rumus *information gain* untuk mengukur efektivitas suatu atribut dapat menggunakan rumus persamaan (4)

$$Gain(A) = Info(D) - infoA(D) \quad (4)$$

6. Tahap pengujian

Tahap pengujian dilakukan dengan teknik *confussion matrix* untuk mengukur performa klasifikasi dengan membandingkan nilai aktual dan nilai prediksi. Pada teknik *confussion matrix* dilakukan perhitungan nilai akurasi, precission dan recall dengan melihat matriks berikut:

- *True Positif* (TP) = jumlah data yang bernilai positif dan diprediksi benar sebagai positif.
- *False positif* (FP) = jumlah data yang bernilai negatif, tapi diprediksi sebagai positif.
- *True Negatif*: jumlah data yang bernilai negatif dan diprediksi sebagai negatif.
- *False negatif*: jumlah data yang bernilai positif tapi diprediksi sebagai negatif

Rumus yang digunakan dalam *confussion matrix* adalah sebagai berikut

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$Precission = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{\frac{TP}{FN}} \quad (7)$$

bebera langkah dalam pengujian adalah

1. Melakukan pengujian dengan menghitung nilai akurasi dan presisi dari 6 atribut faktor resiko *long covid*.
2. Melakukan pengujian dengan melakukan seleksi atribut, yaitu pengujian dengan 3 atribut faktor resiko *long covid* 19

Pengujian dilakukan dengan 2 model yaitu dengan perhitungan secara manual dan dengan menggunakan aplikasi rapid manner.

3. HASIL DAN PEMBAHASAN

1. Menentukan data *training* dan data *testing*.

Jumlah data yang digunakan dalam penelitian ini adalah 1022, data tersebut dilakukan seleksi untuk menentukan data training yang digunakan untuk melatih algoritma sehingga diperoleh model klasifikasi. Data training berjumlah 773 dan data testing untuk menguji hasil pemodelan sebayak 250 data.

2. Menghitung *probabilitas* c*lass* atau P(C) untuk setiap kelas

$$P(C1) = \frac{\text{classStatus=YA}}{Jumlah\ Data\ Training} \quad (8)$$

P(Status|Ya) = 299
773
= 0,387
P(Status|Tidak) = 474
773
= 0,613

3. Menghitung *probabilitas* masing-masing kriteria.

Perhitungan menggunakan teorama bayes dengan rumus persamaan (9)

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)} \quad (9)$$

a. Kriteria Jenis kelamin

- Kriteria(jeniskelamin=female|status=Ya) =
  = 198 = 0,662
  299
- Kriteria(jeniskelamin=female|status=Tidak) =
  = 189 = 0,4
  474
- Kriteria(jeniskelamin=male|status=Ya) =
  = 101 = 0,338
  299
- Kriteria(jeniskelamin=male|status=Tidak) =
  = 284 = 0,6
  474

b. Kriteria usia

Kriteria usia merupakan atribut kontinyu, perhitungan dilakukan dengan distribusi *gauss*. Sehingga dilakukan perhitungan nilai rata-rata dan nilai standar deviasi.

- Nilai rata-rata = 44,49
- Standar deviasi = 15,1

c. Kriteria Kondisi saat infeksi

- Kriteria(kondisi=ringan|status=Ya) =
  = 7 = 0,023
  299
- Kriteria(kondisi=ringan|status=Tidak) =
  = 190 = 0,401
  474
- Kriteria(kondisi=sedang|status=Ya) =
  = 19 = 0,064
  299
- Kriteria(kondisi=sedang|status=Tidak) =

= 135 = 0,285
474

- Kriteria(kondisi=berat|status=Ya) =
  = 273 = 0,913
  299
- Kriteria(kondisi=berat|status=Tidak) =
  = 148 = 0,312
  474

d. Kriteria *komorbid*

- Kriteria(komorbid=Ada|status=Ya) =
  = 175 = 0,585
  229
- Kriteria(Komorbid=Ada|status=TIdak) =
  = 21 = 0,045
  474
- Kriteria(Komorbid=Tidak ada|status=Ya) =
  = 124 = 0,415
  299
- Kriteria(Komorbid=TidakAda|status=Tidak) =
  = 452 = 0,955
  474

e. Kriteria BMI

Sebagaimana kriteria usia, kriteria BMI merupakan atribut kontinyu, dengan menghitung nilai rata-rata dan standar deviasi

- Nilai rata-rata = 23,26
- Standar deviasi = 6,74

4. Seleksi atribut

Untuk mengoptimalkan metode naïve bayes, maka perlu dilakukan seleksi atribut, untuk menentukan atrubut mana yang paling efektif digunakan dalam klasifikasi data. Pemilihan atribut dengan menggunakan nilai entropi tertinggi. Persamaan yang digunakan adalah persamaan (2), diperoleh nilai entropi:

a. Entropi jenis kelamin

Info(male) = $\frac{101}{386}\log_2\frac{101}{386} - \frac{285}{386}\log_2\frac{285}{386}$
= 0,829
Info(female) = $\frac{198}{387}\log_2\frac{198}{387}-\frac{189}{387}\log_2\frac{189}{387}$
= 0,997

b. Entropi usia

Info(<30) = $\frac{32}{159}\log_2\frac{32}{159} - \frac{127}{159}\log_2\frac{127}{159}$
= 0,724
Info(30-49) = $\frac{117}{334}\log_2\frac{117}{334}-\frac{217}{334}\log_2\frac{189}{334}$
= 0,934
Info(>=50) = $\frac{150}{280}\log_2\frac{198}{280}-\frac{130}{280}\log_2\frac{189}{280}$
= 0,996

c. Entropi kondisi saat infeksi

Info(ringan) = $\frac{7}{198}\log_2\frac{32}{198} - \frac{191}{198}\log_2\frac{127}{198}$
= 0,221
Info(sedang) = $\frac{19}{154}\log_2\frac{19}{154}-\frac{135}{154}\log_2\frac{135}{154}$
= 0,539
Info(berat) = $\frac{273}{421}\log_2\frac{273}{421}-\frac{148}{421}\log_2\frac{148}{421}$
= 0,935

d. Entropi Komorbid

Inf (Ada) = $\frac{175}{196}\log_2\frac{175}{196} - \frac{21}{196}\log_2\frac{21}{196}$
= 0,491
Info(Tidakada)= $\frac{124}{577}\log_2\frac{124}{577}-\frac{453}{577}\log_2\frac{531}{577}$
= 0,751

e. Entropi BMI

Info(<18) = $\frac{8}{20}\log_2\frac{8}{20} - \frac{12}{20}\log_2\frac{12}{20}$
= 0,971
Info(18-25) = $\frac{48}{491}\log_2\frac{48}{491}-\frac{443}{334}\log_2\frac{443}{334}$
= 0,462
Info(>=25) = $\frac{243}{262}\log_2\frac{243}{262}-\frac{19}{262}\log_2\frac{19}{262}$
= 0,371

Selanjutnya dilakukan perhitungan entropi untuk seluruh data dengan rumus persamaan (3)
InfoA(total) =- $\frac{229}{773}\log_2\frac{229}{773}-=\frac{474}{773}\log_2\frac{474}{773}$
= 0,963

Setelah memperoleh entropi pada masing-masing atribut, berikutnya dilakukan perhitungan nilai gain, dengan rumus persamaan (4)

a. Gain atribut jenis kelamin

Gain(jeniskelamin) = Info(D) – infoA(D)
= 0,963-($\frac{386}{773}$*x*0,892*)+* $\frac{387}{773}$*x*0,997
= 0,0481

b. Gain atribut usia

Gain(usia) = Info(D) – infoA(D)
= 0,963-($\frac{159}{773}$*x*0,724*)+* $\frac{334}{773}$*x*0,934*)+* $\frac{280}{773}$*x*0,996
= 0,050

c. Gain atribut kondisi saat infeksi

Gain(usia) = Info(D) – infoA(D)
= 0,963-($\frac{198}{773}$*x*0,221*)+* $\frac{154}{773}$*x*0,539*)+* $\frac{421}{773}$*x*0,935
= 0,289

d. Gain atribut *komorbid*

Gain(usia) = Info(D) – infoA(D)
= 0,963-($\frac{196}{773}$*x*0,491*)+* $\frac{577}{773}$*x*0,751*)*
= 0,278

e. Gain atribut BMI

Gain(usia) = Info(D) – infoA(D)
= 0,963-($\frac{20}{773}$*x*0,971*)+* $\frac{491}{773}$*x*0,462*)+* $\frac{262}{773}$*x*0,376
=0,517

Dari perhitungan *information gain* tersebut, diperoleh nilai gain tertinggi, urutan nilai gain dapat dilihat pada tabel 5.

Tabel 5. Nilai atribut hasil perhitungan *information* ***gain***

No urut	Atribut	Nilai gain
1	BMI	0,517
2	Kondisi saat infeksi	0,289
3	Komorbid	0,278
4	Usia	0,050
5	Jenis kelamin	0,048

5. Pengujian data testing

a. Pengujian model naïve bayes

Tahap pengujian dilakukan dengan teknik *confussion matrix* untuk mengukur performa klasifikasi dengan membandingkan nilai aktual dan nilai prediksi. Pada teknik confussion matriks dilakukan perhitungan nilai *accuracy*, *precission* dan *recall*. Pada pengujian tahap ini dilakukan pengujian dengan menggunakan 6 atribut. Data *training* berjumlah 250 data, dapat dilihat pada tabel 6.

Tabel 6. Data training *long covid*

Data ke	jenis kelamin	age	kondisi saat infeksi	Komorbid	BMI	status	prediksi	keterangan
1	male	41	Berat	Tidak Ada	34	YA	YA	benar
2	female	43	Ringan	Tidak Ada	30	Tidak	YA	salah
3	female	76	Ringan	Tidak Ada	30	YA	YA	benar
4	male	29	Berat	Tidak Ada	20	YA	Tidak	salah
5	male	81	Ringan	Ada	27	YA	YA	benar
6	male	56	Berat	Ada	25	YA	YA	benar
7	male	26	Ringan	Tidak Ada	23	Tidak	Tidak	benar
8	male	27	Ringan	Tidak Ada	22	Tidak	Tidak	benar
9	male	21	Ringan	Tidak Ada	24	Tidak	Tidak	benar
10	female	22	Ringan	Tidak Ada	22	Tidak	Tidak	benar
…	….	….	….	….	…	….	…	….
250	female	37	Berat	Tidak Ada	29	YA	YA	benar

Dari data testing data dan prediksi dengan algoritma naïve bayes diperoleh nilai matriks dapat dilihat pada tabel 7

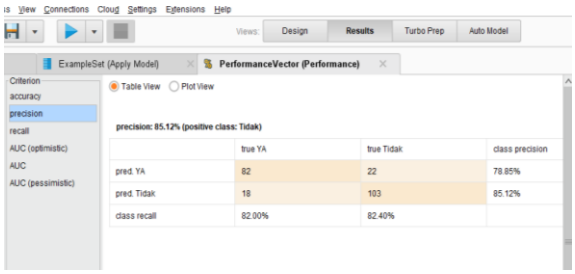
Tabel 7 *confessio matriks* pengujian 6 atribut

	True YA	True Tidak
Pred. YA	82	22
Pred. Tidak	18	103

Keterangan: TP=YA

- Menghitung nilai *accuracy.*
  Accuracy = (TP+TN)/(TP+FP+TN+FN)*100%
  = (82+103)/(82+103+22+18)
  = **82, 22 %**
- Menghitung nilai *precission.*
  Precission = TP/(TP+FP)
  = 82/(82+18)
  = 82,00 %
- Menghitung *recall.*
  Recall = TP/(TP+FN)
  = 82/(82+22)
  = **78, 85 %**

Selain pengujian dilakukan dengan cara manual, pengujian juga dilakkukan dengan menggunakan aplikasi Rapid manner dengan hasil *confussion matrix* yang sama. Hasil pengujian dengan rapid manner dapat dilihat pada gambar 2.



Gambar 2. Hasil pengujian dengan rapid manner

b. Pengujian naïve bayes + *Information gain*

Berdasarkan perhitungan seleksi atribut dengan *information gain* didapatkan nilai gain tertinggi dari masing-masing atribut. Sehingga pada pengujian tahap 2 dilakukan pengujian dengan 3 atribut, yaitu BMI, kondisi saat infeksi dan komorbid. Tabel nilai entropi tertinggi dapat dilihat pada tabel 8

Tabel 8. Nilai entropi

No urut	Atribut	Nilai gain
1	**BMI**	**0,517**
2	**Kondisi saat infeksi**	**0,289**
3	**Komorbid**	**0,278**
4	Usia	0,049
5	Jenis kelamin	0,048

Dari hasil prediksi data *training* diperoleh *confussion matrix* seperti pada tabel 9

Tabel 9. *confession Matrix* pengujian 3 atribut

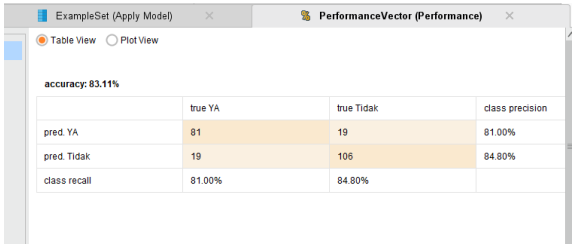
	True YA	True Tidak
Pred. YA	81	19
Pred. Tidak	19	106

Keterangan: TP=YA

- Menghitung nilai *accuracy*
  Accuracy = (TP+TN)/(TP+FP+TN+FN)*100%
  = (82+103)/(82+103+22+18)
  = **83, 11 %**

- Menghitung nilai *precission*
  Precission = TP/(TP+FP)
  = 81/(82+19)
  = 81,00 %
- Menghitung *recall*
  Recall = TP/(TP+FN)
  = 81/(82+19)
  = **81, 00 %**

Pengujian untuk 3 atribut juga dilakukan dengan aplikasi rapid manner dan diperoleh nilai yang sama, seperti terlihat pada gambar 3.



Gambar 3. Hasil pengujian 3 atribut dengan rapid manner

Untuk melihat perbandingan hasil uji coba algoritma naïve bayes dan algoritma naïve bayes dengan *feature selection* dengan teknik *information gain* dapat melihat pada tabel 10.

Tabel 10. Hasil pengujian dengan *information gain*

	Naïve bayes	Naïve bayes+IG
Accuracy	82,22 %	83,11 %
Precission	82,00 %	81,00 %
Recall	78,85 %	81,00 %

Naïve bayes merupakan metode dengan akurasi yang cukup tinggi. Beberapa penelitian membuktikan bahwa penggunaan metode naïve bayes memiliki akurasi yang cukup tinggi dibandingkan penggunaan metode lain, diantaranya adalah penelitian yang dilakukan oleh aristin chusnul khotimah, dkk yang membandingkan tingkat akurasi algoritma naïve bayes, k-nearest neighbour dan support vector machine, menunjukkan bahwa naïve bayes memiliki tingkat akurasi yang lebih baik dibandingkan yang lainnya [17].

4. DISKUSI

Pada penelitian ini dilakukan pengolahan data yang terdiri dari berbagai kriteria yang diolah menjadi kriteria yang didasarkan pada faktor gejala *long covid*. Penentuan faktor gejala *long covid* didasarkan atas beberapa jurnal referensi yang membahas kriteria faktor resiko *long covid*.

Dalam melakukan transformasi data, penentuan atribut dalam kriteria didasarkan pada beberapa kajian, diantaranya adalah jurnal tentang gejala covid-19, gejala *long covid* dan buku panduan tata laksana penanganan covid-19 yang dikeluarkan oleh persatuan dokter dan dokter spesialis indonesia.

Hasil dari pengolahan data kemudian diolah dengan algoritma naive bayes untuk menghasilkan model klasifikasi data yang dijadikan dasar dalam prediksi resiko terjadinya *long covid* pada penyintas covid-19.

Pada tahap pengujian dilakukan dengan 2 tahapan yaitu: pengujian model naive bayes dan pengujian model naive bayes dengan *information gain*. Hasil perhitungan *information gain* diperoleh nilai entropi tertinggi, dan diambil 3 atribut yang memiliki nilai entropi tertinggi. hasil pengujian penggunaan 3 atribut diperoleh nilai *accuracy* dan *recall* yang lebih tinggi dibandingkan dengan mengguanakan naive bayes tanpa pemilihan atribut.

5. KESIMPULAN

Dari penelitian ini diperoleh kesimpulan bahwa, ada beberapa faktor yang mempengaruhi resiko terjadinya *long covid*, diantaranya adalah faktor jenis kelamin, usia, kondisi saat infeksi, komorbid dan *body mass index* (BMI). Data *training* yang digunakan pada penelitian ini sebanyak 773 dan data *testing* sebanyak 250. Dengan menggunakan algoritma naive bayes dihasilkan model klasifikasi data yang dapat digunakan prediksi resiko terjadinya *long covid*. Pengujian dilakukan dengan menggunakan teknik *confussion matrix* dan diperoleh nilai accuracy 82, 22 % dan 83,11 % setelah dilakukan pemilihan atribut dengan *information gain*. Penelitian ini diharapkan dapat dikembangkan menjadi sebuah aplikasi prediksi *long covid*, sehingga dapat memberikan sumbangsih pada masyarakat, dunia kesehatan dan pengembangan teknologi.

UCAPAN TERIMA KASIH

Ucapan terima kasih kami tujukan kepada Kementerian Ristekdikti atas pemberian dana penelitian melalui hibah penelitian dosen pemula (PDP) periode 2021/2022.

DAFTAR PUSTAKA

[1] N. R. Yunus and A. Rezki, “Kebijakan Pemberlakuan Lock Down Sebagai Antisipasi Penyebaran Corona Virus Covid-19,” *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 3, 2020, doi: 10.15408/sjsbs.v7i3.15083.

[2] M. Siahaan, “Dampak Pandemi Covid-19 Terhadap Dunia Pendidikan,” vol. 1, no. 1, pp. 1–3, 2020.

[3] A. Crispo *et al.*, “Strategies to evaluate outcomes in long-COVID-19 and post-COVID survivors,” *Infect. Agent. Cancer*, vol. 16, no. 1, pp. 1–20, 2021, doi: 10.1186/s13027-021-00401-3.

[4] I. Visan, “Long COVID,” *Nat. Immunol.*, vol.

22, no. 8, pp. 934–935, 2021, doi: 10.1038/s41590-021-00992-4.

[5] P. Zimmermann, L. F. Pittet, and N. Curtis, "How Common is Long COVID in Children and Adolescents?," *Pediatr. Infect. Dis. J.*, vol. 40, no. 12, pp. e482–e487, 2021, doi: 10.1097/INF.0000000000003328.

[6] L. Mertz, "Researchers Seek Answers for Millions with Long COVID-19," *IEEE Pulse*, vol. 12, no. 2, pp. 17–21, 2021, doi: 10.1109/MPULS.2021.3066718.

[7] H. Crook, S. Raza, J. Nowell, M. Young, and P. Edison, "Long covid - Mechanisms, risk factors, and management," *BMJ*, vol. 374, pp. 1–18, 2021, doi: 10.1136/bmj.n1648.

[8] C. H. Sudre *et al.*, "Attributes and predictors of long COVID," *Nat. Med.*, vol. 27, no. 4, pp. 626–631, 2021, doi: 10.1038/s41591-021-01292-y.

[9] A. H. Ardiansyah, W. Nugroho, N. H. Alfiyah, R. A. Handoko, and M. A. Bakhtiar, "Penerapan Data Mining Menggunakan Metode Clustering untuk Menentukan Status Provinsi di Indonesia 2020," *Semin. Nas. Inov. Teknol.*, vol. 4, no. 3, pp. 329–333, 2020.

[10] S. Rizal, P. Studi, T. Informatika, and U. Yudharta, "Penerapan Algoritma Naïve Bayes Untuk Prediksi Penerimaan Siswa Baru Di Smk Al-Amien Wonorejo," *Explor. IT J. Keilmuan dan Apl. Tek. Inform.*, vol. 10, no. 1, pp. 14–21, 2018, doi: 10.35891/explorit.v10i1.1671.

[11] PDPI, PERKI, PAPDI, PERDATIN, and IDAI, *Pedoman tatalaksana COVID-19 Edisi 3 Desember 2020*. 2020. [Online]. Available: https://www.papdi.or.id/download/983-pedoman-tatalaksana-covid-19-edisi-3-desember-2020

[12] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *J. EECCIS*, vol. 7, no. 1, p. pp.59-64, 2013.

[13] Bustami, "Penerapan Algoritma Naive Bayes Untuk Nasabah Asuransi," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[14] S. A. Putri and D. Larasati, "Penerapan Feature Selection Pada Bayesian Network Untuk," *PILAR Nusa Mandiri*, vol. 13, no. 2, pp. 275–280, 2017.

[15] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "COMPARISON OF CLASSIFICATION ALGORITHM AND FEATURE SELECTION IN PERBANDINGAN ALGORITMA KLASIFIKASI DAN FEATURE SELECTION," *JUTIF*, vol. 3, no. 3, pp. 739–744, 2022.

[16] M. R. Hasibuan and Marji, "Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)," vol. 3, no. 11, pp. 10435–10443, 2019.

[17] A. C. Khotimah *et al.*, "COMPARISON NAÏVE BAYES CLASSIFIER , K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE IN THE CLASSIFICATION OF INDIVIDUAL ON PERBANDINGAN ALGORITMA NAÏVE BAYES CLASSIFIER , K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI," vol. 3, no. 3, pp. 673–680, 2022