Analisis Perbandingan Algoritma Gaussian Naive Bayes dan Categorical Naive Bayes dengan Laplace Smoothing dalam Deteksi COVID-19

Maulana Zhahran[1], Muhammad Hanif[2], Naufal Althafi[3]

[1,2,3]Informatika, Fakultas Teknik, Universitas Jenderal Soedirman, Indonesia
NIM : [1]H1D021070, [2]H1D021056, [3]H1D021087
Email: [1]maulana.zhahran@unsoed.ac.id, [2]muhammad.hanif056@mhs.unsoed.ac.id ,
[3]naufal.handoyo@mhs.unsoed.ac.id

(Artikel dikirimkan tanggal : dd mmm yyyy)

Abstrak

Pada Januari 2020, terkonfirmasi adanya transmisi manusia-ke-manusia dari SARS-CoV-2 dengan tingkat infeksi tinggi melalui saluran pernapasan atas, sementara jumlah total kasus COVID-19 di dunia terus meningkat dengan penyebaran cepat melalui kontak dekat, droplet, dan udara. Sebagai response, pemerintah dan WHO mengambil langkah pencegahan seperti mempersiapkan pengobatan COVID-19, meningkatkan kapasitas penanganan darurat di fasilitas kesehatan, serta mengatur skrining pasien. Oleh karena itu, deteksi dini COVID-19 menjadi penting dalam mempercepat tindakan dan pengobatan pasien, serta melindungi orang lain di sekitarnya. Tujuan dari penelitian ini adalah untuk membandingkan rata-rata tingkat akurasi algoritma berdasarkan proporsi data training yang berbeda dengan data training yang sama dalam mendeteksi COVID-19. Metode yan

Kata kunci: Covid 19, Klasifikasi, Laplace Smoothing, Python

Comparative Analysis of Gaussian Naive Bayes and Categorical Naive Bayes Algorithms with Laplace Smoothing in COVID-19 Detection

Abstract

Abstrak berbahasa Inggris diletakkan pada bagian ini. Gunakan font Times New Roman 10pt, italic. Setidaknya, tolong MINIMAL menggunakan aplikasi Grammarly untuk mengecek kesesuaian Bahasa Inggris.

Keywords:

1. PENDAHULUAN

Pada fase awal wabah COVID-19, keterkaitan antara pasien-pasien yang baru teridentifikasi dengan kunjungan mereka ke Pasar Grosir Makanan Laut menunjukkan kemungkinan adanya asal-usul penyakit ini yang bersifat zoonosis. Meskipun inang asli dan perantara dari SARS-CoV-2 belum secara pasti ditentukan, kedekatan filogenetik antara SARS-CoV-2 dan coronavirus yang berasal dari kelelawar mengindikasikan kemungkinan bahwa virus baru ini berhubungan dengan coronavirus yang ada pada kelelawar[1].

Pada Januari 2020, ada bukti klinis yang kuat yang mengkonfirmasi adanya transmisi manusia-ke-manusia dari SARS-CoV-2. Tingkat infeksi yang relatif tinggi, mode transmisi melalui saluran pernapasan atas (dan juga mungkin melalui kontak), periode inkubasi yang relatif panjang, dan periode pembuangan virus yang lama, bersama dengan pola perjalanan global saat ini, menjadi semua elemen kunci yang memungkinkan virus ini berevolusi menjadi pandemi dengan cepat[2].

Jumlah total kasus COVID-19 di dunia masih terus meningkat, yaitu sebanyak 504.571.336 kasus pada tanggal 24 Juni 2022. Berdasarkan bukti ilmiah bahwa penyebaran COVID-19 sangat cepat dan dapat ditularkan melalui kontak dekat atau droplet serta melalui udara, pemerintah dan Organisasi Kesehatan Dunia (WHO) telah mengambil beberapa langkah pencegahan untuk membantu mengurangi kasus COVID-19, seperti mempersiapkan pengobatan COVID-19 bagi pasien yang terinfeksi, meningkatkan kapasitas penanganan darurat di fasilitas kesehatan, dan mengatur skrining pasien[3].

Langkah-langkah preventif memiliki peran penting dalam menekan kasus COVID-19 jika terapi protokol (Protocol Therapy) diimplementasikan sejak tahap awal. Deteksi dini COVID-19 merupakan salah satu cara untuk membantu mempercepat tindakan bagi pasien, baik itu untuk memastikan kondisi kesehatannya atau untuk memerlukan pengujian lebih lanjut terkait COVID-19. Sistem deteksi dini COVID-19 dianggap sangat penting bagi pasien dan juga orang-orang di

sekitarnya agar dapat melawan pandemi COVID-19, karena jika pasien mendapatkan pengobatan yang tepat dan cepat, maka orang lain di sekitarnya juga akan terlindungi[4].

Naive Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada *teorema Bayes*. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Ciri utama dari *Naive Bayes Classifier* ini adalah asumsi yang sangat kuat (*naif*) akan independensi dari masing masing kondisi / kejadian[5].

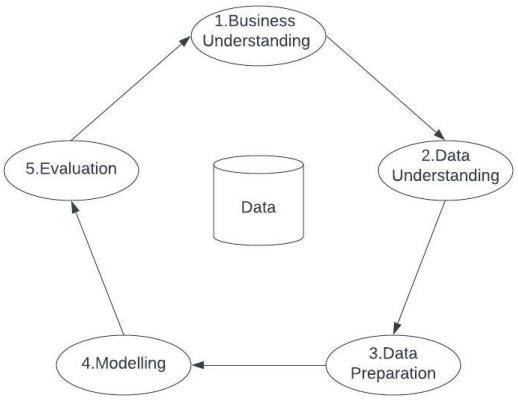
Gaussian Naive Bayes adalah salah satu metode klasifikasi yang termasuk dalam keluarga algoritma Naive Bayes. Metode ini digunakan untuk mengklasifikasikan data berdasarkan asumsi bahwa fitur-fitur yang ada dalam data tersebut mengikuti distribusi Gaussian (distribusi normal) independen satu sama lain. Dalam Gaussian Naive Bayes, diasumsikan bahwa setiap fitur dalam data memiliki distribusi Gaussian dengan rata-rata (mean) dan variansi (variance) yang berbeda untuk setiap kelasnya. Model ini menghitung probabilitas posterior kelas menggunakan teorema Bayes dan kemudian memprediksi kelas dengan probabilitas tertinggi[6].

Categorical Naive Bayes adalah implementasi dari algoritma Categorical Naive Bayes untuk data yang didistribusikan secara kategorikal. Algoritma ini mengasumsikan bahwa setiap fitur, yang dijelaskan oleh indeks i, memiliki distribusi kategorikal sendiri. Untuk setiap fitur i dalam set data pelatihan X, Categorical Naive bayes memperkirakan distribusi kategorikal untuk setiap fitur i dari X yang dikondisikan pada kelas y. Set indeks dari sampel-sampel didefinisikan sebagai {1, 2, ..., n}, dengan n sebagai jumlah sampel. Dalam algoritma Categorical Naive Bayes , estimasi distribusi kategorikal untuk setiap fitur i di komputasi menggunakan metode maximum likelihood estimation (MLE) dari data pelatihan. Selanjutnya, probabilitas posterior kelas dikomputasi menggunakan teorema Bayes dengan memperhitungkan distribusi kategorikal dari setiap fitur[7].

Tujuan dari penelitian ini adalah untuk membandingkan rata-rata tingkat akurasi algoritma Gaussian Naive Bayes, dan Categorical Naive Bayes dengan Laplace Smoothing berdasarkan proporsi data training dengan data testing yang sama dalam mendeteksi COVID-19 pada orang-orang dengan gejala tertentu.

2. METODE

Data mining adalah sebuah proses, sehingga dalam melakukan prosesnya harus sesuai dengan prosedur CRISP-DM (*Cross- Industry Standard Process for Data Mining*). CRISP-DM adalah standarisasi data mining yang disusun oleh tiga penggagas data mining market yaitu *Daimler Chrysler*, SPSS, NCR. CRISP-DM tidak menentukan standar atau karakteristik tertentu karena setiap data yang akan dianalisis akan diproses kembali pada fase-fase di dalamnya. Dalam penelitian ini peneliti menggunakan data mining dengan prosedur CRISP-DM, namun dalam penelitian ini hanya menggunakan lima tahapan dari enam tahapan yang ada[8]. Berikut tahapan yang digunakan seperti pada Gambar 1 pada penelitian ini:



Gambar 1. Tahapan prosedur CRISP-DM[9]

2.1. Business Understanding

Pada tahap ini berfokus pada pemahaman mengenai tujuan dari proyek dan kebutuhan secara perspektif bisnis, kemudian mengubah pengetahuan ini menjadi definisi masalah penambangan data dan rencana awal yang dirancang untuk mencapai tujuan. Tujuan utama dari penelitian ini adalah untuk membandingkan tingkat akurasi separasi data algoritma Gaussian Naive Bayes dan Categorical Naive Bayes dengan Laplace Smoothing berdasarkan proporsi separasi data training dan data testingnya. Agar didapatkan algoritma yang paling optimal dalam memprediksi COVID-19.

2.2. Data Understanding

Pada tahap ini dilakukan pengumpulan, pengidentifikasi, dan pemahaman data yang akan digunakan pada penelitian ini. *Dataset yang* digunakan dalam penelitian ini yaitu *Symptoms and COVID Presence (May 2020 data)*. Berisikan jenis gejala penyakit yang ada pada orang-orang yang diduga terkena COVID-19, ketika masa pandemi COVID-19. *Dataset* ini diperoleh secara online dari Kaggle[10].

2.3. Data Preparation

Selanjutnya akan dilakukan data preparation untuk menghasilkan data optimal pada saat modeling. Di dalam tahap data preparation juga

terdapat preprocessing, dimana merupakan proses membuang duplikasi data, memeriksa data yang tidak konsisten dan memperbaiki kesalahan pada penulisan kata[11]. Adapun beberapa tahapan yang ada pada data preparation, berikut ini contohnya:

a. Pembersihan Data untuk menghapus baris dengan nilai yang hilang atau mengisi dengan nilai yang sesuai
b. Transformasi Data untuk mengubah data kategorik menjadi data yang dapat dimengerti oleh algoritma.
c. Reduksi Dimensi untuk memilih fitur yang optimal yang akan digunakan untuk modeling.

2.4. Modeling

Pada tahap *modeling* dilakukan proses klasifikasi dengan model yang diusulkan dalam penelitian ini adalah *Gaussian Naive Bayes* dan *Categorical Naive Bayes* dengan *Google Collaboratory* untuk pengelompokan jenis gejala penyakit yang ada pada manusia, ketika masa pandemi COVID-19. Untuk Naive Bayes sendiri merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (aturan bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naive Bayes model yang digunakan adalah model fitur independen[12].

2.4.1. Gaussian Naive Bayes

Gaussian Naive Bayes adalah salah satu metode klasifikasi dalam data mining yang didasarkan pada asumsi bahwa fitur-fitur yang digunakan dalam klasifikasi mengikuti distribusi normal (*Gaussian*). Metode ini sering digunakan untuk mengklasifikasikan data dengan fitur-fitur yang kontinu. *Gaussian Naive Bayes* berbeda dengan *Naive Bayes* biasa, yakni *Gaussian Naive Bayes* memiliki distribusi *Gaussian*[13]. Rumus dari Gaussian Naive Bayes adalah sebagai berikut :

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (1)$$

Pada Persamaan 1, $P(X_i|y)$ adalah probabilitas variabel X_i pada kelas y, π adalah konstanta matematika dengan nilai sekitar 3.14159, σ_y adalah simpangan baku dari variabel X_i pada kelas y, μ_y adalah rata variabel X_i pada kelas y, dan exp adalah fungsi eksponensial yang menghitung e^x , dimana e adalah bilangan euler dengan nilia sekitar 2.71828.

Untuk menggabungkan var smoothing dalam rumus *Gaussian Naive Bayes*, kita dapat menambahkan variabel smoothing (biasanya disebut sebagai α) ke varian dalam persamaan 1 tersebut. Dengan demikian, persamaan 1 yang dimodifikasi akan menjadi:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi(\sigma_y^2+\alpha)}} \exp\left(-\frac{(x_i-\mu_y)^2}{2(\sigma_y^2+\alpha)}\right) \quad (2)$$

Penambahan α pada varian di persamaan 2 memastikan bahwa varian selalu lebih besar dari nol, sehingga probabilitas tidak menjadi tidak terdefinisi. Hal ini membantu menjaga stabilitas dan keandalan model dalam melakukan estimasi probabilitas dengan mempertimbangkan var smoothing. Nilai yang umum digunakan adalah α = 1, yang dikenal sebagai Laplace smoothing atau add-one smoothing.

2.4.2. Categorical Naive Bayes

Categorical Naive Bayes diterapkan untuk data yang didistribusikan secara kategoris. Dapat dimisalkan bahwa diasumsikan setiap fitur, yang dijelaskan oleh indeks *i*, memiliki distribusi kategorinya sendiri. Rumus dari *Categorical Naive Bayes* adalah sebagai berikut :

$$P(x_i = t \mid y = c\,;\alpha) = \frac{N_{tic}+\alpha}{N_c+\alpha n_i} \quad (3)$$

Pada persamaan 3 tersebut menyatakan bahwa probabilitas kondisional $P(x_i = t \mid y = c)$ dihitung dengan menambahkan α pada jumlah kemunculan nilai kategori t dalam kelas c, dan kemudian membaginya dengan jumlah total sampel dalam kelas c ditambah dengan α dikalikan dengan jumlah nilai kategori yang mungkin untuk fitur x_i.

Laplace Smoothing merupakan metode yang banyak digunakan, sekaligus smoothing yang disebut sebagai default smoothing dan smoothing tertua yang pernah diimplementasikan pada Naive Bayes[14]. Metode *Laplace Smoothing* digunakan untuk menghindari probabilitas yang nol ketika ada nilai kategori yang tidak muncul dalam kelas tertentu dalam data pelatihan. Dengan adanya penyesuaian α, probabilitas kondisional $P(x_i = t \mid y = c)$ akan selalu lebih besar dari nol, sehingga tidak ada probabilitas yang hilang. Untuk Metode Laplace Smoothing nilai α = 1.

2.5. Evaluation

Pada tahap *evaluation* dilakukan proses klasifikasi dengan beberapa algoritma yaitu *Naive Bayes, Gaussian Naive Bayes,* dan *Categorical Naive Bayes* dengan *Laplace Smoothing* untuk melihat hasil accuracy dengan menggunakan python

pada Google Colaboratory. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap pemodelan sesuai dengan sasaran yang ingin dicapai dalam tahap *business understanding.*

3. HASIL DAN PEMBAHASAN

Pada hasil dan pembahasan ini akan menggunakan tahapan-tahapan CRISP-DM yang sudah dilakukan :

3.1. Business Understanding

Pada tahap ini, fokusnya adalah pemahaman mengenai tujuan dari proyek dan kebutuhan dari perspektif bisnis dalam konteks prediksi COVID-19. Kemudian, pengetahuan ini akan diubah menjadi definisi masalah penambangan data dan dirancang rencana awal untuk mencapai tujuan tersebut. Tujuan utama dari penelitian ini adalah untuk memprediksi kasus COVID-19 dengan menggunakan algoritma Gaussian Naive Bayes dan Categorical Naive Bayes dengan Laplace Smoothing. Dalam konteks ini, penelitian ini bertujuan untuk mengidentifikasi algoritma yang paling optimal dalam memprediksi kasus COVID-19 berdasarkan proporsi separasi data training dan data testing. Proses evaluasi yang dilakukan pada model yakni interpretasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan pada tahap sebelumnya[15].

Pada tahap ini, akan dilakukan kegiatan untuk mempersiapkan strategi awal guna mencapai tujuan tersebut. Hal ini meliputi pengumpulan data yang relevan mengenai kasus COVID-19, pemilihan parameter yang tepat untuk kedua algoritma, serta merancang metode evaluasi yang sesuai untuk mengukur performa algoritma dalam memprediksi kasus COVID-19.

3.2. Data Understanding

Pada tahap ini dilakukan pengumpulan, pengidentifikasi, dan pemahaman data yang akan digunakan pada penelitian ini. *Dataset yang* digunakan dalam penelitian ini yaitu *Symptoms and COVID Presence (May 2020 data).* Berisikan jenis gejala penyakit yang ada pada orang-orang yang diduga terkena COVID-19, ketika masa pandemi COVID-19. Berikut ini merupakan data understanding pada penelitian ini :

	Breathing Problem	Fever	Dry Cough	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fatigue	Gastrointestinal	Abroad travel	Contact with COVID Patient	Attended Large Gathering	Visited Public Exposed Places	Family working in Public Exposed Places	Wearing Masks	Sanitization from Market	COVID-19
0	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	...	Yes	Yes	No	Yes	No	Yes	Yes	No	No	Yes
1	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	...	Yes	No	No	No	Yes	Yes	No	No	No	Yes
2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	...	Yes	Yes	Yes	No	No	No	No	No	No	Yes
3	Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	...	No	No	Yes	No	Yes	Yes	No	No	No	Yes
4	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	...	No	Yes	No	Yes	No	Yes	No	No	No	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5429	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	...	Yes	Yes	No	No	No	No	No	No	No	Yes
5430	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	...	Yes	No	No	No	No	No	No	No	No	Yes
5431	Yes	Yes	Yes	No	No	No	No	No	Yes	No	...	No	No	No	No	No	No	No	No	No	No
5432	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	...	No	No	No	No	No	No	No	No	No	No
5433	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	...	Yes	No	No	No	No	No	No	No	No	No

5434 rows × 21 columns

Gambar 2. Dataset gejala COVID-19

Pada gambar 2 data yang digunakan diantaranya yaitu *Breathing Problem, Fever, Dry Cough, Sore throat, Running Nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hyper Tension, Fatigue, Gastrointestinal, Abroad travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family working in Public Exposed Places, Wearing Masks, Sanitization from Market, COVID-19.* Dengan jumlah baris 5434 baris dan jumlah kolom 21 kolom. Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesis untuk informasi yang tersembunyi.

Untuk memudahkan pemahaman mengenai kolom pada tiap dataset yang digunakan, terdapat rincian variable pada tiap kolom di dataset dengan value atau nilai yang juga tertera pada tabel 1 dibawah ini.

Tabel 1. Indikator Pada Dataset *Symptoms and COVID Presence (May 2020 data)*

No	*Variabel*	Nama Variabel	*Value*
1	X_1	Breathing Problem	Yes/No
2	X_2	Fever	Yes/No
3	X_3	Dry Cough	Yes/No
4	X_4	Sore throat	Yes/No
5	X_5	Running Nose	Yes/No
6	X_6	Asthma	Yes/No
7	X_7	Chronic Lung Disease	Yes/No
8	X_8	Headache	Yes/No
9	X_9	Heart Disease	Yes/No
10	X_{10}	Diabetes	Yes/No
11	X_{11}	Hyper Tension	Yes/No
12	X_{12}	Fatigue	Yes/No
13	X_{13}	Gastrointestinal	Yes/No
14	X_{14}	Abroad travel	Yes/No

15	X_{15}	Contact with COVID Patient	Yes/No
16	X_{16}	Attended Large Gathering	Yes/No
17	X_{17}	Visited Public Exposed Places	Yes/No
18	X_{18}	Family working in Public Exposed Places	Yes/No
19	X_{19}	Wearing Mask	No
20	X_{20}	Sanitazion from Market	No
21	Y	COVID-19	Yes/No

3.3. Data Preparation

Pada tahap persiapan data *(data preparation)* ini dilakukan proses pembersihan dan transformasi data yang awalnya masih berupa data campuran berisi teks dan angka, dikonversi menjadi data numerik boolean. Hal ini dimaksudkan agar data tersebut lebih mudah dibaca dan dimengerti oleh algoritma. Berikut ini merupakan data preparation pada penelitian ini :

	Breathing Problem	Fever	Dry Cough	...	Family working in Public Exposed Places	Wearing Masks	Sanitization from Market
0	1	1	1	...	1	1	1
1	1	1	1	...	0	1	1
2	1	1	1	...	0	1	1
3	1	1	1	...	0	1	1
4	1	1	1	...	0	1	1
...	...	...	...	...	...	...	...
5429	1	1	0	...	0	1	1
5430	1	1	1	...	0	1	1
5431	1	1	1	...	0	1	1
5432	1	1	1	...	0	1	1
5433	1	1	1	...	0	1	1

5434 rows × 20 columns

Gambar 3. Dataset setelah dikonversi menjadi numerik boolean

Selanjutnya akan dilakukan seleksi fitur menggunakan SelectKBest dengan metode chi-square SelectKBest menggunakan metode statistik chi-square untuk mengukur hubungan antara setiap fitur dengan target variabel. Statistik chi-square digunakan untuk menguji independensi antara dua variabel kategorikal. Dalam konteks feature selection, SelectKBest dengan chi2 menghitung skor chi-square untuk setiap fitur dan mengambil K fitur dengan skor tertinggi. Skor chi-square mencerminkan sejauh mana fitur mempengaruhi target variabel. Untuk itu dilakukan analisis skor chi-square pada setiap fitur.

	Nama Fitur	Chi2 Scores
13	Abroad travel	587.725530
15	Attended Large Gathering	445.070586
3	Sore throat	374.479605
0	Breathing Problem	357.224357
14	Contact with COVID Patient	345.368317
2	Dry Cough	242.944097
1	Fever	144.581349
17	Family working in Public Exposed Places	81.414100
16	Visited Public Exposed Places	37.487883
10	Hyper Tension	29.155431
5	Asthma	23.614977
6	Chronic Lung Disease	9.268135
11	Fatigue	5.102075
9	Diabetes	4.697584
8	Heart Disease	2.133476
7	Headache	2.084083
4	Running Nose	0.079433
12	Gastrointestinal	0.032681
19	Sanitization from Market	0.000000
18	Wearing Masks	0.000000

Gambar 4. Skor chi-square

Seperti pada gambar 4, skor chi-square tertinggi dengan ambang skor 80, terdapat pada fitur *Abroad travel, Attended Large Gathering, Sore throat, Breathing Problem, Contact with COVID Patient, Dry Cough, Fever, Family working in Public Exposed Places*. Oleh karena itu kedelapan fitur ini akan dipilih untuk proses modeling algoritma Gaussian Naive Bayes dan Categorical Naive Bayes. Berikut ini adalah informasi dataset yang telah ditransformasi datanya dan dilakukan fitur seleksi pada tabel 2.

Tabel 2. Dataset *Symptoms and COVID Presence (May 2020 data) variabel X* setelah dilakukan transformasi data dan pemilihan fitur

No	*Variabel*	Nama Variabel	*Value*
1	X_{14}	Abroad travel	1 atau 0
2	X_{16}	Attended Large Gathering	1 atau 0
3	X_4	Sore throat	1 atau 0
4	X_1	Breathing Problem	1 atau 0
5	X_{15}	Contact with COVID Patient	1 atau 0
6	X_3	Dry Cough	1 atau 0
7	X_2	Fever	1 atau 0
8	X_{18}	Family working in Public Exposed Places	1 atau 0

3.4. Modeling

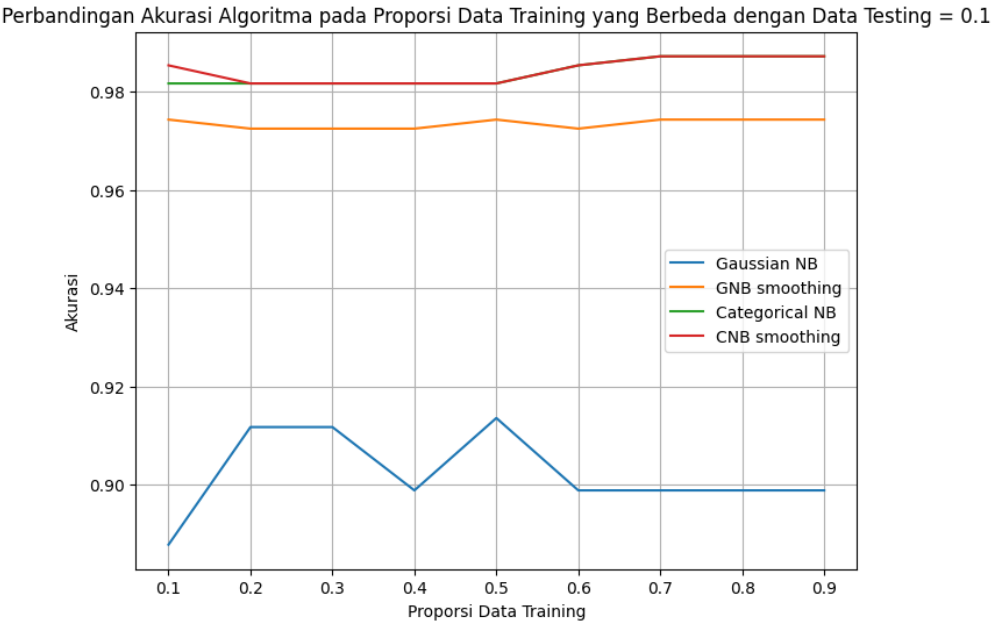
Tahap Pemodelan *(modelling)* dilakukan dengan menyiapkan dataset yang akan digunakan, kemudian melakukan uji akurasi algoritma Gaussian Naive Bayes dan Categorical Naive Bayes. Untuk model Gaussian Naive Bayes dibagi menjadi dua yaitu model tanpa laplace smoothing dan model dengan laplace smoothing. Begitu juga

dengan model Categorical Naive Bayes dibagi menjadi dua yaitu model tanpa laplace smoothing dan model dengan laplace smoothing. Untuk penerapan modelnya menggunakan python dapat dilihat pada gambar 5 berikut ini.

```
# Daftar algoritma yang akan diuji
no_laplace_smoothing = 1e-4 # tanpa laplace smoothing nilai alpha < 1
laplace_smoothing = 1 # dengan laplace smoothin nilai alpha = 1
algorithms = {
    'Gaussian NB': GaussianNB(var_smoothing=no_laplace_smoothing),
    'GNB smoothing': GaussianNB(var_smoothing=laplace_smoothing),
    'Categorical NB' : CategoricalNB(alpha=no_laplace_smoothing, force_alpha=True),
    'CNB smoothing': CategoricalNB(alpha=laplace_smoothing, force_alpha=True),
}

# Daftar proporsi pemisahan data yang akan diuji
test_size = 0.1 # teste size tetap
train_sizes = [i*0.1 for i in range(1 , 11 - int(test_size*10))] # train size berubah


# Inisialisasi dictionary untuk menyimpan akurasi masing-masing algoritma
accuracy_scores = {algorithm: [] for algorithm in algorithms}

# Lakukan pemisahan data dan evaluasi untuk setiap proporsi
for train_size in train_sizes:
    X_train, X_test, y_train, y_test = train_test_split(X_selected, y,
                                                        train_size=train_size, test_size=test_size, random_state=42)
    for algorithm_name, algorithm in algorithms.items():
        algorithm.fit(X_train, y_train)
        y_pred = algorithm.predict(X_test)
        accuracy = accuracy_score(y_test, y_pred)
        accuracy_scores[algorithm_name].append(accuracy)
```

Gambar 5. Penerapan Model Gaussian Naive Bayes dan Categorical Naive Bayes

Pengujian dilakukan dengan melatih model dengan proporsi data training yang berbeda tapi dengan data testing yang sama yaitu 10% dari keseluruhan dataset. Untuk proporsi data trainingnya sendiri adalah mulai dari 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, dan 90%. Berikut ini hasil analisis untuk tiap-tiap modelnya pada gambar 6 dan hasil akurasinya pada masing-masing data training pada tabel 3.



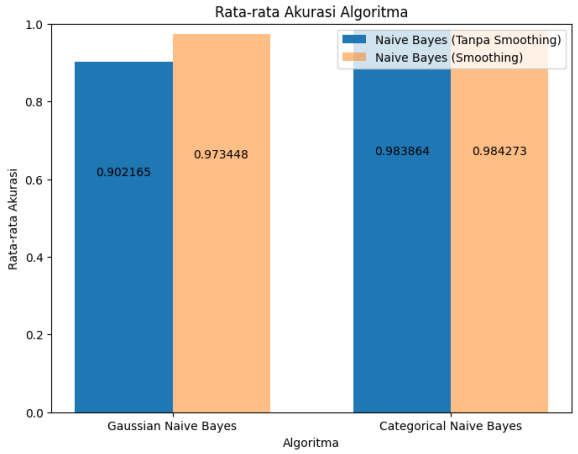
Gambar 6. Analisis Akurasi Antar Algoritma/Model

Tabel 3. Perbndinga akurasi antar Algoritma/Model dengan data training yang berbeda dan data testing yang sama

Data Training	Gaussian Naive Bayes	Gaussian Naive Bayes dengan Laplace Smoothing	Categorical Naive Bayes	Categorical Naive Bayes dengan Laplace Smoothing
90%	0.887868	0.974265	0.981618	0.985294
80%	0.911765	0.972426	0.981618	0.981618
70%	0.911765	0.972426	0.981618	0.981618
60%	0.898897	0.972426	0.981618	0.981618
50%	0.913603	0.974265	0.981618	0.981618
40%	0.898897	0.972426	0.985294	0.985294
30%	0.898897	0.974265	0.987132	0.987132
20%	0.898897	0.974265	0.987132	0.987132
10%	0.898897	0.974265	0.987132	0.987132

3.5. Evaluation

Klasifikasi yang dilakukan terhadap data jenis gejala penyakit yang ada pada orang-orang yang diduga terkena COVID-19, ketika masa pandemi COVID-19 menggunakan beberapa model algoritma diantaranya, algoritma *Gaussian Naive Bayes* dan *Categorical Naive Bayes* tanpa Laplace Smoothing dan dengan *Laplace Smoothing*. Dalam hal ini proses evaluasi akan menghasilkan nilai *accuracy* dan nilai rata-rata *accuracy* algoritma pada tabel 3 sebagai berikut :



Gambar 7. Perbandingan Rata-rata Akurasi Algoritma Gaussian Naive Bayes dan Categorical Naive Bayes

Berdasarkan gambar 7, dapat diketahui rata-rata akurasi Gaussian Naive Bayes tanpa Laplace Smoothing adalah 0.902165, sedangkan rata-rata akurasi Gaussian Naive Bayes dengan Laplace Smoothing adalah 0.973448. Selisih rata-rata akurasinya adalah 0.071323. Selanjutnya untuk Categorical Naive Bayes tanpa Laplace Smoothing rata-rata akurasinya adalah 0.983864, sedangkan untuk Categorical Naive Bayes dengan Laplace Smoothing rata-rata akurasinya adalah 0.984273. Selisih rata-rata akurasinya adalah 0.000409. Hal ini membuktikan bahwa Smoothing Laplace Smoothing dapat meningkatkan tingkat akurasi yang cukup besar pada Algoritma Gaussian Naive Bayes yaitu peningkatan rata-rata akurasi sebesar 7.9%. Sedangkan pada algoritma Categorical Naive Bayes, metode Laplace Smoothing dapat meningkatkan rata-rata akurasi sebesar 0.04%.

4. KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa penggunaan metode Laplace Smoothing berperan penting dalam meningkatkan akurasi Algoritma Naive Bayes seperi Gaussian Naive Bayes dan Categorical Naive Bayes. Untuk rata-rata akurasinya sendiri, Categorical Naive Bayes memperoleh nilai tertinggi dengan rata-rata akurasi 0.9840685. Sedangkan Gaussian Naive Bayes memperoleh rata-rata akurasi sebesar 0.947549.

Dengan demikian permasalahan mengenai pengelompokan data jenis penyakit yang terjadi ketika pandemi COVID-19 ini dianggap tercapai karena mendapatkan nilai akurasi tinggi.

DAFTAR PUSTAKA

[1] P. Zhou *dkk.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, hlm. 270–273, 2020, doi: https://doi.org/10.1038/s41586-020-2012-7.

[2] C. Wang, Z. Wang, G. Wang, J. Y. N. Lau, K. Zhang, dan W. Li, "COVID-19 in early 2021: current status and looking forward," *Signal Transduct. Target. Ther.*, vol. 6, no. 1, Des 2021, doi: 10.1038/s41392-021-00527-1.

[3] A. S. Fauci, H. C. Lane, dan R. R. Redfield, "Covid-19—Navigating the Uncharted," *N. Engl. J. Med.*, vol. 382, no. 13, hlm. 1268–1269, 2021.

[4] J. Riyono, A. Latifa, R. Putri, dan C. E. Pujiastuti, "Early Detection of COVID-19 Disease Based on Behavioral Parameters and Symptoms Using Algorithm-C5.0," *Indones. J. Artif. Intell. Data Min. IJAIDM*, vol. 6, no. 1, hlm. 47–53, 2023, doi: 10.24014/ijaidm.v2i2.22074.

[5] A. Felicia Watratan, A. B. Puspita, D. Moeis, S. Informasi, dan S. Profesional Makassar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," 2020. [Daring]. Tersedia pada: http://journal.isas.or.id/index.php/JACOST

[6] A. Ng, "CS229 Lecture Notes: Gaussian Naive Bayes." [Daring]. Tersedia pada: http://cs229.stanford.edu/notes2020spring/cs229-notes2.pdf

[7] F. Pedregosa *dkk.*, "Categorical Naive Bayes (CategoricalNB)." 2011. [Daring]. Tersedia pada: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.CategoricalNB.html

[8] A. Pambudi, "PENERAPAN CRISP-DM MENGGUNAKAN MLR K-FOLD PADA DATA SAHAM PT.TELKOM INDONESIA (PERSERO) TBK (TLKM)(STUDI KASUS: BURSA EFEK INDONESIA TAHUN 2015-2022)," *J. Data Min. Dan Sist. Inf.*, vol. 4, no. 1, hlm. 1–14, 2023.

[9] Nurahman, "Evaluasi Performa Algoritma C4.5 Dan C4.5 Berbasis PSO Untuk Memprediksi Penyakit Diabetes," *J. E-Komtek Elektro-Komput.-Tek.*, vol. 4, no. 1, hlm. 30–47, Jun 2020, doi: 10.37339/e-komtek.v4i1.230.

[10] H. Harikrishnan, "Symptoms and COVID Presence (May 2020 data)," *Kaggle*, 2020. https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence (diakses 25 Mei 2023).

[11] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, hlm. 213–219, Feb 2017, doi: 10.22202/ei.2016.v2i2.1465.

[12] E. Prasetyo, "Data mining konsep dan aplikasi menggunakan matlab," *Yogyak.*

Andi, vol. 1, 2012.

[13] C. C. Aggarwal, *Data mining: the textbook*, vol. 1. Springer, 2015.

[14] Z. H. Kilimci dan M. C. Ganiz, "Evaluation of classification models for language processing," dalam *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, IEEE, 2015, hlm. 1–8.

[15] G. Fiastantyo, "Perbandingan Kinerja Metode Klasifikasi Data Mining Menggunakan Naive Bayes dan Algoritma C4. 5 untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa," *Semantic J.*, 2014.