Jurnal Ilmu Komputer dan Informatika (JIKI)
P-ISSN: 2807-6664
E-ISSN: 2807-6591
Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

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

Dila Saputra[1], Abdul Aziz Fahmi 'Alauddin[2], Mochamad Azizan[3]

[1,2,3]Informatics, Universitas Jenderal Soedirman, Indonesia

Email: [1]dila.saputra@mhs.unsoed.ac.id, [2]abdul.alauddin@mhs.unsoed.ac.id, [3]mochamad.azizan@mhs.unsoed.ac.id

Received: Feb 12, 2025; Revised: Aug 14, 2025; Accepted: Aug 15, 2025; Published: Aug 29, 2025

Abstract

In January 2020, it was confirmed that COVID-19 can be transmitted from human to human through the upper respiratory tract with a high infection rate. The number of COVID-19 cases worldwide continued to increase rapidly through close contact, droplets, and airborne transmission. In response, governments and the WHO implemented preventive measures, including COVID-19 treatment preparation, increased emergency healthcare capacity, and patient screening. Early detection of COVID-19 became crucial in taking action, providing treatment, and protecting others. In the Naïve Bayes algorithm, a potential issue arises with the possibility of zero probabilities for some features or attributes in the COVID-19 prediction training data. Therefore, Laplace Smoothing is used to address this problem. This study aims to compare the average accuracy rates of Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms using different proportions of training data but the same testing data for COVID-19 detection. The methods used in this research are Gaussian Naïve Bayes and Categorical Naïve Bayes with Laplace Smoothing implemented using the Python library called scikit-learn. The research results show that the Gaussian Naïve Bayes algorithm without Laplace Smoothing has an average accuracy of 0.902165, while with Laplace Smoothing, it has an average accuracy of 0.973448. For the Categorical Naïve Bayes algorithm, without Laplace Smoothing, it has an average accuracy of 0.983864, while with Laplace Smoothing, it has an average accuracy of 0.984273. In conclusion, Laplace Smoothing plays a significant role in improving the average accuracy of Naïve Bayes algorithms. Categorical Naïve Bayes achieves the highest average accuracy of 0.9840685 (with and without Laplace Smoothing), while Gaussian Naïve Bayes achieves 0.947549 (with and without Laplace Smoothing). Categorical Naïve Bayes has a higher average accuracy compared to Gaussian Naïve Bayes.

Keywords: *COVID-19, Laplace Smoothing, Naïve Bayes, Python*

1. INTRODUCTION

In the early phase of the COVID-19 outbreak, the association of newly identified patients with their visits to the Seafood Wholesale Market suggested a possible zoonotic origin of the disease. Although the original and intermediate hosts of SARS-CoV-2 have not been definitively determined, the phylogenetic closeness between SARS-CoV-2 and coronaviruses of bat origin indicates the possibility that this new virus is related to coronaviruses in bats [1].

As of January 2020, there is strong clinical evidence confirming human-to-human transmission of SARS-CoV-2. The relatively high rate of infection, the mode of transmission through the upper respiratory tract (and also possibly through contact), the relatively long incubation period, and the long shedding period of the virus, together with the current global travel pattern, have all been key elements that have allowed this virus to evolve quickly became a pandemic [2].

The total number of COVID-19 cases in the world is still increasing, namely 504,571,336 cases on June 24 2022. Based on scientific evidence that the spread of COVID-19 is very fast and can be transmitted through close contact or droplets as well as through the air, the government and Health Organizations The world (WHO) has taken several preventive steps to help reduce COVID-19 cases, such as preparing COVID-19 treatment for infected patients, increasing emergency treatment capacity in health facilities, and organizing patient screening [3].

Preventive measures have an important role in suppressing COVID-19 cases if protocol therapy is implemented from an early stage. Early detection of COVID-19 is one way to help speed up action for patients, whether to confirm their health condition or to require further testing related to COVID-19. The COVID-19 early detection system is considered very important for patients and the people around them to be able to fight the COVID-19 pandemic, because if the patient gets appropriate and fast treatment, other people around them will also be protected [4].

Naïve Bayes Classifier is a classification method that is rooted in Bayes' theorem. The classification method uses probability and statistical methods proposed by the British scientist Thomas Bayes, namely predicting future opportunities based on previous experience, so it is known as Bayes' Theorem. The main characteristic of the Naïve Bayes Classifier is a very strong (naive) assumption of the independence of each condition/event [5].

Gaussian Naïve Bayes is a classification method that is included in the Naïve Bayes algorithm family. This method is used to classify data based on the assumption that the features in the data follow a Gaussian distribution (normal distribution) independent of each other. In Gaussian Naïve Bayes, it is assumed that each feature in the data has a Gaussian distribution with a different mean and variance for each class. This model calculates the posterior probability of the class using Bayes' theorem and then predicts the class with the highest probability [6].

Categorical Naïve Bayes is an implementation of the Categorical Naïve Bayes algorithm for categorically distributed data. This algorithm assumes that each feature, described by index i, has its own categorical distribution. For each feature i in the training data set X, Categorical Naïve Bayes estimates the categorical distribution for each feature i of X conditioned on class y. The index set of samples is defined as {1, 2, ..., n}, with n being the number of samples. In the Categorical Naïve Bayes algorithm, the estimated categorical distribution for each feature i is computed using the maximum likelihood estimation (MLE) method from training data. Next, the class posterior probability is computed using Bayes' theorem by taking into account the categorical distribution of each feature.

The aim of this research is to compare the average accuracy level of the Gaussian Naïve Bayes algorithm, and Categorical Naïve Bayes with Laplace Smoothing based on the proportion of training data with the same testing data in detecting COVID-19 in people with certain symptoms.

2. METHOD

Data mining is a process, so carrying out the process must comply with the CRISP-DM (Cross-Industry Standard Process for Data mining) procedure. CRISP-DM is a data mining standardization prepared by three initiators of the data mining market, namely Daimler Chrysler, SPSS, NCR. CRISP-DM does not determine certain standards or characteristics because each data to be analyzed will be reprocessed in the phases within it. In this study, researchers used data mining with the CRISP-DM procedure, but in this study only used five stages out of the six existing stages [7]. The following stages are used as in Figure 1 in this research:

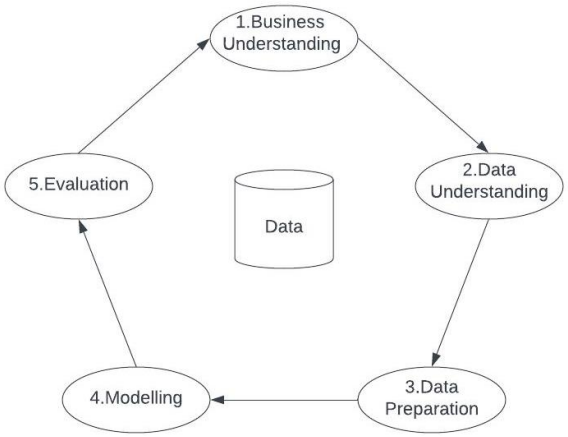


Figure 1. Stages of the CRISP-DM procedure [8]

2.1. Business Understanding

This stage focuses on understanding the project's goals and requirements from a business perspective, then turning this knowledge into a data mining problem definition and an initial plan designed to achieve the goals. The main aim of this research is to compare the level of data separation accuracy of the Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms with Laplace Smoothing based on the proportion of separation between the training data and the testing data. In order to obtain the most optimal algorithm in predicting COVID-19.

2.2. Data Understanding

At this stage, data is collected, identified and understood to be used in this research. The dataset used in this research is Symptoms and COVID Presence (May 2020 data). Contains the types of disease symptoms present in people suspected of having COVID-19, during the COVID-19 pandemic. This dataset was obtained online from Kaggle.

2.3. Data Preparation

Next, data preparation will be carried out to produce optimal data during modeling. In the data preparation stage there is also preprocessing, which is the process of removing duplicate data, checking inconsistent data and correcting errors in writing words [9]. There are several stages in data preparation, here are examples:

a. Data Cleaning to delete rows with missing values or fill with appropriate values
b. Data Transformation to change categorical data into data that can be understood by algorithms.
c. Dimensional Reduction to select optimal features to be used for modeling.

2.4. Modelling

At the modeling stage, a classification process is carried out with the models proposed in this research, namely Gaussian Naïve Bayes and Categorical Naïve Bayes with Google Collaboratory for grouping types of disease symptoms that exist in humans, during the COVID-19 pandemic. Naïve Bayes itself is a simple probabilistic-based prediction technique based on the application of Bayes' theorem (Bayes' rule) with strong (naive) assumptions of independence. In other words, in Naïve Bayes the model used is an independent feature model [10].

2.4.1. Gaussian Naïve Bayes

Gaussian Naïve Bayes is a classification method in data mining which is based on the assumption that the features used in classification follow a normal (Gaussian) distribution. This method is often used to classify data with continuous features. Gaussian Naïve Bayes is different from ordinary Naïve Bayes, namely that Gaussian Naïve Bayes has a Gaussian distribution [11]. The formula for Gaussian Naïve Bayes is as follows:

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

In equation 1, $P(X_i|y)$ is the probability of variable X_i given class y, π is a mathematical constant with a value of approximately 3.14159, σ_y is the standard deviation of variable X_i in class y, μ_y is the mean of variable X_i in class y, and exp is the exponential function that calculates e^x, where e is Euler's number with a value of approximately 2.71828.

To combine the smoothing variance in the Gaussian Naïve Bayes formula, we can add a smoothing variable (usually referred to as α) to the variance in equation 1. Thus, the modified equation 1 will be:

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

Adding α to the variance in equation 2 ensures that the variance is always greater than zero, so that the probability does not become undefined. This helps maintain the stability and reliability of the

model in carrying out probability estimates by considering smoothing variance. The commonly used value is α = 1, which is the value for the Laplace Smoothing method.

2.4.2. Categorical Naïve Bayes

Categorical Naïve Bayes is applied to categorically distributed data. It can be assumed that each feature, described by index i, has its own categorical distribution. The formula for Categorical Naïve Bayes is as follows:

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

In Equation 3, the conditional probability $P(x_i = t \mid y = c)$ is calculated by adding α to the count of occurrences of the categorical value t in class c, and then dividing this by the total number of samples in class c plus α multiplied by the number of possible categorical values for the feature x_i.

Laplace Smoothing is a method that is widely used, as well as smoothing which is called default smoothing and the oldest smoothing ever implemented in Naïve Bayes [12]. The Laplace Smoothing method is used to avoid zero probability when there are category values that do not appear in a particular class in the training data. With the α adjustment, the conditional probability $P(x_i = t \mid y = c)$ will always be greater than zero, so there is no missing probability. For the Laplace Smoothing Method, the value of α = 1.

2.5. Evaluation

At the evaluation stage, a classification process was carried out using several algorithms, namely Naïve Bayes, Gaussian Naïve Bayes, and Categorical Naïve Bayes with Laplace Smoothing to see accuracy results using Python on Google Colaboratory. Evaluation is carried out in depth with the aim that the results at the modeling stage are in line with the targets to be achieved in the business understanding stage.

3. RESULT

3.1. Business Understanding

In this stage, the focus is on understanding the objectives of the project and the requirements from a business perspective within the context of COVID-19 case prediction. This understanding is then translated into a clear data mining problem definition and an initial plan is designed to achieve the stated objectives. The primary goal of this research is to predict COVID-19 cases using the Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms with Laplace Smoothing. In this context, the study aims to identify the most optimal algorithm for predicting COVID-19 cases based on different proportions of training and testing data separation. The evaluation process involves interpreting the results of the data mining models, as demonstrated during the modeling phase in the previous stage [13].

At this stage, activities are carried out to prepare an initial strategy for achieving the research objectives. This includes collecting relevant data related to COVID-19 cases, selecting appropriate parameters for both algorithms, and designing suitable evaluation methods to measure their predictive performance.

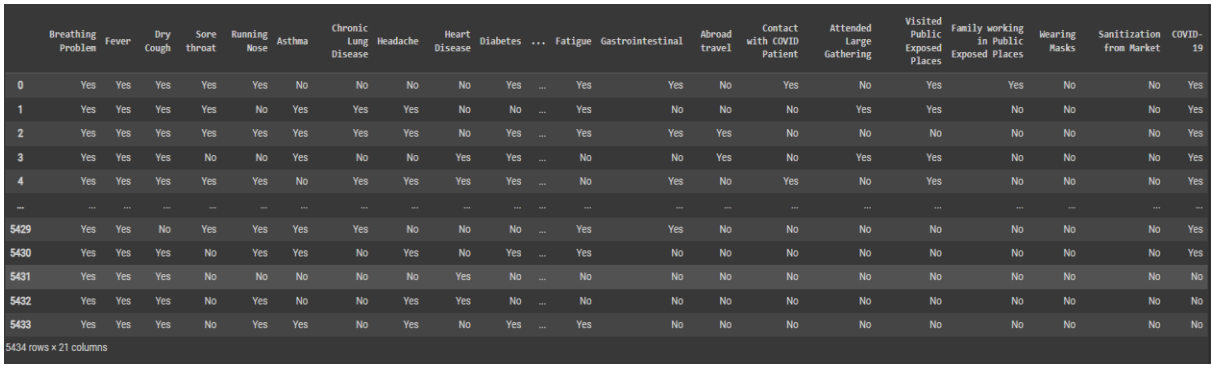
3.2. Data Understanding

At this stage, data collection, identification, and understanding are carried out for the research. The dataset used in this study is "Symptoms and COVID Presence (May 2020 data)." It contains types of symptoms experienced by individuals suspected of having COVID-19 during the pandemic period. Figure 2 represent an understanding of the data used in this research.

In Figure 2, the data used includes Breathing Problem, Fever, Dry Cough, Sore Throat, Running Nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hypertension, Fatigue, Gastrointestinal issues, Abroad Travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family Working in Public Exposed Places, Wearing Masks, Sanitization from Market, and COVID-19. The dataset consists of 5434 rows and 21 columns. This stage begins with data

Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

collection, followed by processes to gain a deep understanding of the data, identify data quality issues, or detect interesting aspects of the data that can be used to hypothesize hidden information.



	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

Figure 2. COVID-19 Symptoms Dataset

To facilitate understanding of the columns in each dataset used, there is a detailed breakdown of the variables in each column of the dataset along with their values or data listed in Table 1.

Table 1. Indicators in the Symptoms and COVID Presence Dataset (May 2020 data)

No	Variabel	Nama Variabel	Tipe Data	Value
1	X1	Breathing Problem	Binomial	Yes/No
2	X2	Fever	Binomial	Yes/No
3	X3	Dry Cough	Binomial	Yes/No
4	X4	Sore throat	Binomial	Yes/No
5	X5	Running Nose	Binomial	Yes/No
6	X6	Asthma	Binomial	Yes/No
7	X7	Chronic Lung Disease	Binomial	Yes/No
8	X8	Headache	Binomial	Yes/No
9	X9	Heart Disease	Binomial	Yes/No
10	X10	Diabetes	Binomial	Yes/No
11	X11	Hyper Tension	Binomial	Yes/No
12	X12	Fatigue	Binomial	Yes/No
13	X13	Gastrointestinal	Binomial	Yes/No
14	X14	Abroad travel	Binomial	Yes/No
15	X15	Contact with COVID Patient	Binomial	Yes/No
16	X16	Attended Large Gathering	Binomial	Yes/No
17	X17	Visited Public Exposed Places	Binomial	Yes/No
18	X18	Family working in Public Exposed Places	Binomial	Yes/No
19	X19	Wearing Mask	Binomial	Yes/No
20	X20	Sanitazion from Market	Binomial	Yes/No
21	Y	COVID-19	Binomial	Yes/No

3.3. Data Preparation

This is an example of the use of sub-chapters in a paper. Sub-chapters are allowed to be included in all chapters, except in the conclusion.

In the data preparation stage, a process of cleaning and transforming the data is carried out. Initially, the data consists of a mix of text and numbers, which is then converted into numerical boolean data. This is intended to make the data easier for algorithms to read and understand. The following is the data preparation for this research:

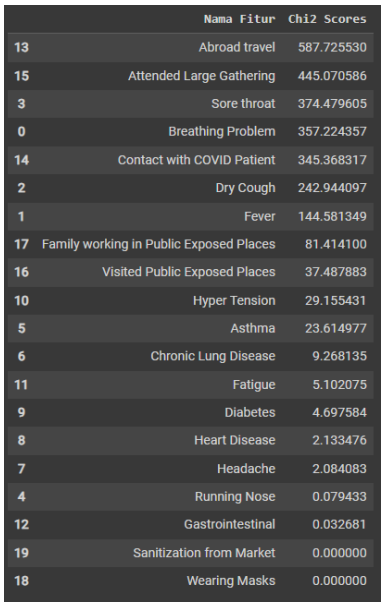
Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

	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

Figure 3. Dataset after conversion to numerical boolean

Next, feature selection will be performed using the Python library SelectKBest with the chi-square method. SelectKBest uses the chi-square statistical method to measure the relationship between each feature and the target variable. The chi-square statistic is used to test the independence between two categorical variables. In the context of feature selection, SelectKBest with chi-square calculates the chi-square score for each feature and selects the top K features with the highest scores. The chi-square score reflects the extent to which a feature influences the target variable. Therefore, a chi-square score analysis will be conducted on each feature.



	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

Figure 4. *Chi-square's Score*

As shown in Figure 4, the highest chi-square scores with a threshold score of 80 are found in the features Abroad travel, Attended Large Gathering, Sore throat, Breathing Problem, Contact with COVID Patient, Dry Cough, Fever, and Family working in Public Exposed Places. Therefore, these eight features will be selected for the modeling process using the Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms. Below is the information on the dataset that has been transformed and feature selection has been performed, as presented in Table 2.

Table 2. Symptoms and COVID Presence Dataset (May 2020 data) - Variable X after Data Transformation and Feature Selection

No	*Variabel*	Variable Name	Data Type	*Value*
1	X_{14}	*Abroad travel*	Binomial	1 or 0
2	X_{16}	*Attended Large Gathering*	Binomial	1 or 0
3	X_4	*Sore throat*	Binomial	1 or 0
4	X_1	*Breathing Problem*	Binomial	1 or 0
5	X_{15}	*Contact with COVID Patient*	Binomial	1 or 0

Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

6	X_3	*Dry Cough*	Binomial	1 or 0
7	X_2	*Fever*	Binomial	1 or 0
8	X_{18}	*Family working in Public Exposed Places*	Binomial	1 or 0

3.4. Modelling

The modeling stage involves preparing the dataset for testing the accuracy of Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms. For Gaussian Naïve Bayes, two models are developed: one without Laplace Smoothing and one with Laplace Smoothing. Similarly, for Categorical Naïve Bayes, two models are created: one without Laplace Smoothing and one with Laplace Smoothing. The implementation of these models utilizes the Python library scikit-learn [14], which includes implementations of Gaussian Naïve Bayes and Categorical Naïve Bayes algorithms. This setup can be seen in Figure 5 below.

```
# 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)
```

Figure 5. Application of Gaussian Naïve Bayes and Categorical Naïve Bayes Models

Testing is conducted by training the models with different proportions of training data while keeping the testing data consistent at 10% of the entire dataset. The proportions of training data range from 10% to 90% of the entire dataset. Below are the analysis results for each model as shown in Figure 6 and the accuracy results for each iteration of training data in Table 3.

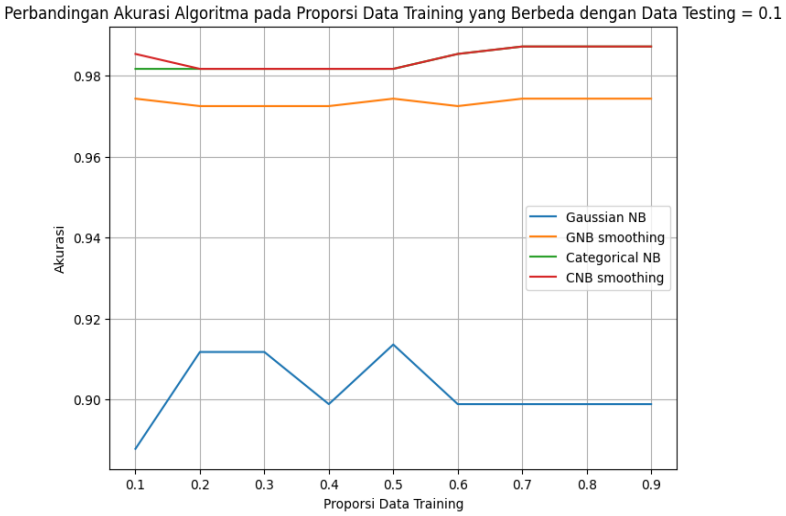


Figure 6. Accuracy Analysis Across Algorithms/Models

Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

Table 3. Comparison of Accuracy Among Algorithms/Models with Different Training Data Proportions and Consistent Testing Data

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

In Figure 6 and Table 3, Gaussian Naïve Bayes without Laplace Smoothing shows fluctuating accuracy across iterations of training data from 10% to 60%, achieving a highest accuracy of 0.913603 and a lowest of 0.887868. From 70% to 90% of training data, the accuracy remains consistent at 0.898897. Gaussian Naïve Bayes with Laplace Smoothing improves its accuracy compared to without Laplace Smoothing, with accuracy stabilizing across iterations of training data from 10% to 90%, ranging from a highest of 0.974265 to a lowest of 0.972426. Both Categorical Naïve Bayes without Laplace Smoothing and with Laplace Smoothing show identical accuracy values from iterations of training data from 20% to 90%, with a slight difference observed at 10% training data: 0.981618 accuracy for Categorical Naïve Bayes without Laplace Smoothing and 0.985294 accuracy for Categorical Naïve Bayes with Laplace Smoothing.

3.5. Evaluation

Classification conducted on the types of symptoms present in individuals suspected of having COVID-19 during the COVID-19 pandemic period utilized several algorithm models, including Gaussian Naïve Bayes and Categorical Naïve Bayes with and without Laplace Smoothing. The evaluation process yields accuracy values and average accuracy values for the algorithms in Table 3 as follows:

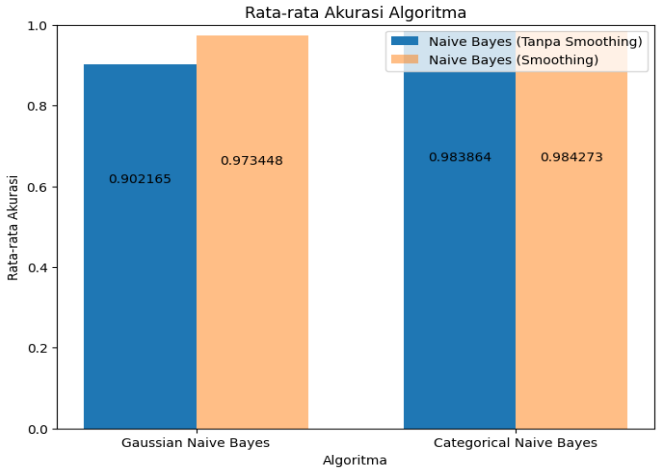


Figure 7. Comparison of Average Accuracy Between Gaussian Naïve Bayes and Categorical Naïve Bayes Algorithms

Based on Figure 7, the average accuracy of Gaussian Naïve Bayes without Laplace Smoothing is 0.902165, while the average accuracy of Gaussian Naïve Bayes with Laplace Smoothing is 0.973448. The difference in average accuracy is 0.071323. For Categorical Naïve Bayes, the average accuracy without Laplace Smoothing is 0.983864, and with Laplace Smoothing, it is 0.984273. The difference in

average accuracy is 0.000409.These results demonstrate that Laplace Smoothing significantly improves the accuracy of the Gaussian Naïve Bayes algorithm, with an average accuracy increase of 7.9%. On the other hand, Laplace Smoothing has a minimal impact on improving the average accuracy of the Categorical Naïve Bayes algorithm, with an increase of only 0.04%.

4. DISCUSSIONS

In analyzing the Symptoms and COVID Presence (May 2020 data) using Naïve Bayes classifiers, specifically Categorical Naïve Bayes (CNB) and Gaussian Naïve Bayes (GNB), several key observations and considerations emerge:

a. Nature of the Data:
   The dataset comprises symptoms categorized as either present or absent, represented in a binary or categorical format (1 or 0). This categorical nature inherently aligns well with the assumptions of Categorical Naïve Bayes, which models features as discrete probabilities based on their frequencies.
b. Distribution Assumptions:
   - Categorical Naïve Bayes: Assumes each feature is independent and calculates probabilities based on the frequency of category values. This model is suitable for the binary or categorical representation of COVID-19 symptoms in the dataset.
   - Gaussian Naïve Bayes: Assumes a normal distribution of numerical features. Given that COVID-19 symptoms are encoded as binary or categorical variables, Gaussian Naïve Bayes may not accurately capture the true distribution of the data.
c. Model Performance:
   - Categorical Naïve Bayes: In this experiment, Categorical Naïve Bayes demonstrates higher accuracy compared to Gaussian Naïve Bayes. This is evident from the higher average accuracy values observed in CNB without Laplace Smoothing, compared to GNB with Laplace Smoothing. CNB leverages simplicity and is well-suited for the categorical structure of the data.
   - Gaussian Naïve Bayes: Although Gaussian Naïve Bayes can improve accuracy with Laplace Smoothing, its performance remains lower than Categorical Naïve Bayes. This suggests that the normal distribution assumption of numerical features in GNB may not fully match the actual distribution of COVID-19 symptom data that is categorical in nature.
d. Effect of Laplace Smoothing:
   In Gaussian Naïve Bayes, Laplace Smoothing helps mitigate issues with zero or very low probabilities that can impact model accuracy. However, its impact is not as pronounced as seen in Categorical Naïve Bayes, indicating that the effectiveness of Laplace Smoothing depends on how well the data distribution aligns with the model assumptions.

5. CONCLUSION

Based on the results and discussions conducted, it can be concluded that the use of Laplace Smoothing plays a crucial role in improving the accuracy of Naïve Bayes algorithms such as Gaussian Naïve Bayes and Categorical Naïve Bayes. In terms of average accuracy, Categorical Naïve Bayes achieved the highest average accuracy of 0.9840685 (both without Laplace Smoothing and with Laplace Smoothing). On the other hand, Gaussian Naïve Bayes achieved an average accuracy of 0.947549 (both without Laplace Smoothing and with Laplace Smoothing). The increase in training data also proved to be a factor in enhancing the accuracy of Naïve Bayes algorithms with Laplace Smoothing. Therefore, the issue of COVID-19 detection based on symptom data during the COVID-19 pandemic is considered successfully addressed due to achieving high accuracy values.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

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

Vol. 5, No. 1, June 2025, Page. 69-78
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.286

[2] C. Wang, Z. Wang, G. Wang, J. Y.-N. Lau, K. Zhang, and W. Li, "COVID-19 in early 2021: current status and looking forward," *Signal Transduct Target Ther*, vol. 6, no. 1, p. 114, Mar. 2021, doi: 10.1038/s41392-021-00527-1.

[3] A. S. Fauci, H. C. Lane, and R. R. Redfield, "Covid-19 — Navigating the Uncharted," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1268–1269, Mar. 2020, doi: 10.1056/NEJMe2002387.

[4] C. Matrix and J. Riyono, "Early Detection of COVID-19 Disease Based on Behavioral Parameters and Symptoms Using Algorithm-C5. 0," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 6, no. 1, pp. 47–53, 2023.

[5] Alvina Felicia Watratan, Arwini Puspita. B, and Dikwan Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *Journal of Applied Computer Science and Technology*, vol. 1, no. 1, pp. 7–14, Jul. 2020, doi: 10.52158/jacost.v1i1.9.

[6] A. Ng, "CS229 Lecture notes," *CS229 Lecture notes*, vol. 1, no. 1, pp. 1–3, 2000.

[7] 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)," *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, pp. 1–14, 2023.

[8] Nurahman, "Evaluasi Performa Algoritma C4.5 dan C4.5 Berbasis PSO untuk Memprediksi Penyakit Diabetes," *Jurnal E-Komtek (Elektro-Komputer-Teknik)*, vol. 4, no. 1, pp. 30–47, Jun. 2020, doi: 10.37339/e-komtek.v4i1.230.

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

[10] E. Prasetyo, "Data mining konsep dan aplikasi menggunakan matlab," *Yogyakarta: Andi*, vol. 1, 2012.

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

[12] Z. H. Kilimci and M. C. Ganiz, "Evaluation of classification models for language processing," in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, IEEE, Sep. 2015, pp. 1–8. doi: 10.1109/INISTA.2015.7276787.

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

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.