Classification System for the Level of Spread Covid-19 Cases in Bandung Regency Using the Naive Bayes Method

Elsa Dwiyanti
Department of Informatics
Universitas Jenderal Achmad Yani
West Java, Cimahi, Indonesia
elsadwiyanti18@if.unjani.ac.id

Wina Witanti, Gunawan Abdilah
Department of Informatics
Universitas Jenderal Achmad Yani
West Java, Cimahi, Indonesia
witanti@gmail.com, mgunawan12@gmail.com

Abstract

COVID-19 is a disease caused by a new type of corona virus, namely SARS-CoV-2. The most common symptoms of COVID-19 victims are fever, dry cough, and fatigue. This virus has spread to almost all countries, including West Bandung Regency. Transmission of this virus can be directly transmitted from one person to another, of course this is what causes the number of COVID-19 cases to continuously increase. So far, the use of COVID-19 vaccines is being developed, some of which are still in the experimental stage. However, a vaccine that is truly effective against the COVID-19 virus has not yet been found. Therefore, this study aims to classify the level of COVID-19 cases in Bandung Regency to predict the spread of the corona virus and suppress the number of COVID-19 cases so that it does not continue to increase. The method used in this research is to use one of the data mining algorithms, namely the Naive Bayes algorithm as a classification.

Keyword

COVID-19; Klasifikasi ; Text Mining ; Kabupaten Bandung ; Naïve Bayes

1. Introduction

Since 2019, the world has been shocked by the innovation of a new disease in the City of Wuhan, China which is claimed to use COVID-19. Coronavirus is an RNA virus that has a particle size between 120-160 nm. This virus usually attacks certain animals, including bats and camels. There are several types of coronavirus, including alphacoronavirus 229E, betacoronavirus HKU1, Severe Acute Respiratory Illness Coronavirus (SARS-CoV), and Middle East Respiratory Syndrome Coronavirus (MERS-CoV). Coronavirus belongs to the betacoronavirus genus (Susilo et al., 2020).

Previous studies have grouped the level of COVID-19 cases at various levels. West Bandung is one of the regencies in Indonesia, which consists of 16 sub-districts and 165 villages. Based on information from the West Bandung Regency COVID-19 Monitoring Data, as of February 2, 2022, the total number of cases was 24453 while 19327 positive cases were confirmed. Therefore, patients with a high or positive level of urgency are prioritized for treatment compared to patients with moderate or asymptomatic symptoms. (Liliana et al., 2021) Of course, medical personnel need to help classify patient status automatically to reduce fatigue of the medical staff on duty and minimize the risk of patient delays in treatment. Therefore, there is a need for a data-based automated technology solution that can help classify emergencies according to patient data. (Rahmayadi et al., 2021) Classifying using the nave Bayes algorithm is the main choice because its accuracy and simplicity are one of its advantages. Although simple, the results obtained are always comparable to other algorithms (Nurdiana & Algifari, 2015) . Research that has been carried out by Nunu Nurdiana, Abijar Algifari proves that the results obtained using the nave Bayes algorithm obtain an accuracy value of 2%. Naïve Bayes obtained 76% accuracy and 74% ID3. (Saputra et al., 2018)

This study aims to provide a solution to the classification of COVID-19 cases, especially in the Bandung Regency area. Using the Naïve-Bayes algorithm method for classification. (Ikbal et al., 2021) By

choosing the right algorithm, it is hoped that it will produce a good and useful classification model to be implemented when dealing with and determining areas that have the highest potential for the spread of COVID-19.

2. Research Methods

2.1 Data Mining

For data owners, data mining is data analysis to find correlations and draw conclusions that were previously unknown. Predictive information can be extracted from databases using data mining, a technology that can have a significant impact on businesses. The first step is to identify predictive information in the database. (Ratino et al., 2020) It is important to recalculate the new centroids as barycenter's of our newly formed ensemble before moving on to the next step. When the k-centroid group changes, it is retested to see if it improves. (Harianto Kristanto et al., 2016)

2.2 Classification

The purpose of classification is to predict the appropriate class for a certain point (. et al., 2019). On the other hand, classification can also be understood as a component of a learning system whose main purpose is to recognize patterns through representation and generalization of data. When there are multiple inputs from multiple methods in an environment with heterogeneous data, the challenge is how to determine which method produces the most accurate prediction of classification results. The combination determination can be used in a number of different classification methods to obtain output recommendations. (Zaki & Meira, 2013) In order to incorporate the Unified Model Language, we decided to use a method involving voting and meta-learning (UML).

2.3 Confusion Matrix

Confusion Matrix is used to calculate accuracy, confusion matrix is a data mining concept used in the process. The evaluation will produce accuracy, precision, and recall values based on the use of a confusion matrix. The percentage of correct data records classified as a result of tests performed on the classification results is referred to as accuracy in classification. (Sengkey et al., 2020) the proportion of cases that are expected to be positive but also turn out to be real positive based on actual data is called precision or confidence. The portion of actual positive cases that is accurately identified as positive in the prediction is referred to as recall or sensitivity. In addition, the confusion matrix is a table that can be used to record the results of the categorization operation. The Confusion Matrix for the classification of the two classes can be found in the following Table 1:

Table 1 Confusion Matrix

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False Negatives*
-	*False positives*	*True Negatives*

Confusion Matrix Formula:

- Accuracy is a calculation of the comparison between data that has been classified correctly from all data.
  Accuracy = (TP+TN)/(TP+TN+FP+fN) x100% ...(1)
- The precision value is a description of the number of positive data categories that are classified correctly and then divided by the total data that are classified as positive. Precision = TP/(TP+FP)x100% ...(2)
- Recall is a calculation that shows a few percent of positive category data that is classified correctly by the system.
  Recall = TP/(TP+FN) x100% ...(3)
- Errors are cases that are identified incorrectly in a number of data, therefore it can be seen how big the error rate is in the system used.
  Error = FP/TPx100% ... (4)

2.4 Bayes Method and Naïve Bayes Algorithm

Naive Classifier of Bayes the Naive Bayes method is a classification system that is easy to use and can be processed quickly. The basis of the Naive Bayes approach is conditional probability. Use Bayes' theorem. This formula estimates probability by calculating the frequency and combination of values in historical data. By assuming the probability of another event, the Bayes theorem estimates the probability of a single event occurring (Darwis et al., 2021) . The Naive Bayes Classifier (NBC) method classifies data with a statistical model. Using the generated or training data, the NBC technique calculates probability values for the test data. (Rosandy, 2016) Therefore, each attribute plays a role in the decision-making procedure. Each property is different from the others, and attribute weights are equally important. The basic Naive Bayes equation. Required: (Figure 1)

$$\mathrm{P(H|X)}\ \frac{P(x|H)P(H)}{\mathrm{P(x)}}\ldots\ (1)$$

Is known:
X : Data with classes that have not been Is known
H : Class Label
P(H) : Probability of hypothesis H based on condition X
P(X) : Probability X
P(H|X) : Probability of Hypothesis H based on condition X.
P(X|H) : Probability X, based on condition hypothesis H.

The description of the image below is as follows:

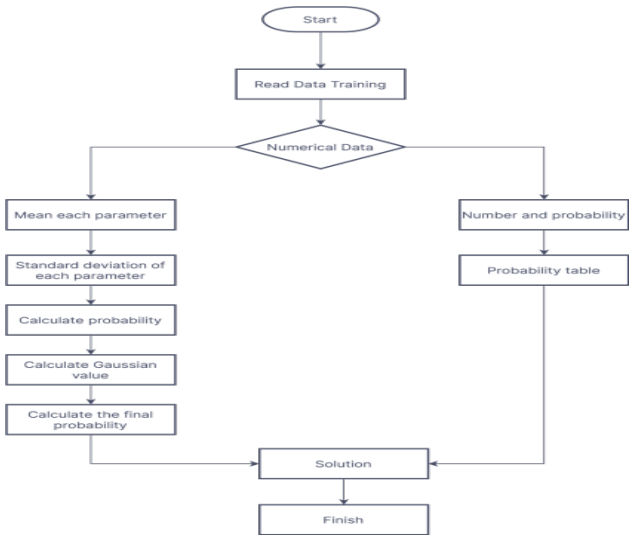

Figure 1 naive bayes algorithm steps

1. The first thing to do is read the Training Data
2. Then calculate the Amount and Probability, but if the data is numeric then
   a) Determine the mean and standard deviation for each numeric parameter using the appropriate formula.
   b) Calculating a probabilistic value involves calculating the number of corresponding data from the same category divided by the total number of data in that category
3. Get the mean, standard deviation, and probability values in the table.
4. Find a solution

- **Mean**
  The mean is used to show the average of all attributes. To calculate the average for each attribute, the following equation is used:

$$\mu = \frac{\sum_{i=1}^{n} x_1}{n}\ldots(2)$$

Is known
μ: average count (mean)

xi: value of sample to -i
n: number of samples

- **Standard Deviation**
  Standard deviation or standard deviation is a static value that serves to determine the distribution or distribution of data in the sample. To calculate the standard deviation, you can use the following formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x1-\mu)^2}{n-1}} \ldots(3)$$

Is known:
σ: standard deviation
Xi : value of x to -i
μ: average count
n : number of samples

- **Probability**
  Probability calculations are numbers used to determine the probability of an event occurring. To calculate the probability value, multiply the number of data according to the category or group you are looking for by the number of data in that category.

- **Gaussian Value**
  Gaussian value functions as a classification calculation with continuous or numeric data type. The function of density is to determine the probability of a given interval. Here is the equation formula:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{\left(x_i-\mu_{ij}\right)^2}{2\sigma_{ij}}} \ldots(4)$$

Is known:
P : Opportunity
Xi: Attribute to i
xi: Attribute value to i
Y : The class you are looking for
yi: Sub class Y you are looking for
μ : mean, states the average of all attributes
σ : Standard deviation, represents the variance of all attributes.

- **Final Probability Score**
  The way to calculate the final probability value is by multiplying the gaussian value that has been calculated for each attribute and the same class. The smallest probability value used is 0 which means that there is no chance if the event occurs, and the highest value used is 1 which indicates that the probability of an event must occur.

3. Results and Discussion

a) Calculation Process Determining the Percentage of Covid

Determination of the label is based on certain calculations, namely by involving the number or accumulation of residents who are confirmed positive for Covid-19 in the subdistrict and then adjusted to the population of the subdistrict, the following is the formula for determining the label:

$$\frac{Number\ of\ confirmed\ positive}{number\ of\ residents\ sub\ district} \text{x}100\% = \ldots..(5)$$

The process of calculating the percentage of covid is a provision as a reference for labeling, this calculation uses two numerical datasets, namely population data in West Bandung Regency and accumulated Covid-19 data in West Bandung Regency. (Table 2)

Table 2 Calculating Process Determining the Percentage of Covid

No	District Name	Calculation
1.	Batujajar	$\frac{1632}{105.773} x100\% = 1.5\%$
2.	Cihamplas	$\frac{1116}{129.203} x100\% = 0.8\%$
3.	Cikalong Wetan	$\frac{1100}{121.766} x100\% = 0.9\%$
4.	Cililin	$\frac{924}{93.415} x100\% = 0.9\%$
5.	Cipatat	$\frac{907}{137.984} x100\% = 0.9\%$
6.	Cipendeuy	$\frac{407}{85.373} x100\% = 0.4\%$
7.	Cipongkor	$\frac{323}{96.167} x100\% = 0.3\%$
8.	Cisarua	$\frac{1645}{77.373} x100\% = 2.1\%$
9.	Gunung Halu	$\frac{277}{75.337} x100\% = 0.3\%$
10.	Lembang	$\frac{5877}{189.789} x100\% = 3.09\%$
11.	Ngamprah	$\frac{2866}{172.522} x100\% = 1{,}6\%$
12.	Padalarang	$\frac{4370}{179.172} x100\% = 2.4\%$
13.	Parongpong	$\frac{4248}{108.148} x100\% = 3.9\%$
14.	Rongga	$\frac{225}{57.325} x100\% = 0.3\%$
15.	Saguling	$\frac{316}{16.797} x100\% = 1.8\%$
16.	Sindangkerta	$\frac{605}{70.726} x100\% = 0.8\%$

b) Stages of data grouping.

The data grouping stage is one of the stages carried out in the data processing process. The goal is that the data used looks clearer and organized. This makes the data easier to read and makes comparisons between data clearer. Table 3.2 describes how the form of data grouping is used in this study

Table 3 Stages of data grouping.

No	Domisili	Close Contact	Suspect	Porable	Covid Levels	Percentage affected by Covid
1	**BATUJAJAR**	168	51	5	Low	1.50%
2	**CIHAMPELAS**	146	91	2	Low	0.80%

3	**CIKALONGWETAN**	234	235	3	Low	0.90%
4	**CILILIN**	239	144	8	Low	0.90%
5	**CIPATAT**	78	29	4	Low	0.90%
6	**CIPEUNDEUY**	51	34	4	Low	0.40%
7	**CIPONGKOR**	44	36	0	Low	0.30%
8	**CISARUA**	490	113	2	Medium	2.10%
9	**GUNUNGHALU**	33	25	0	Low	0.30%
10	**LEMBANG**	855	286	9	High	3.09%
11	**NGAMPRAH**	216	175	8	Low	1.60%
12	**PADALARANG**	1308	223	5	Medium	2.40%
13	**PARONGPONG**	294	32	7	High	3.90%
14	**RONGGA**	22	4	1	Low	0.30%
15	**SAGULING**	112	3	0	Low	1.80%
16	**SINDANGKERTA**	130	40	2	Low	0.80%
TOTAL		**4420**	**1521**	**60**		21.99%

Information:
High= Total Percentage > 3.00%
Medium = Total Percentage < 3.00% - 2.00%
Low = Total Percentage > 2.00%

c) Calculating Probability

The dataset will be processed using the Naive Bayes method with several stages so it is necessary to determine the value of the data, namely from the 16 training data used, it is known that there are 2 data for high class, 2 data for middle class and 12 data for low class. Then the results of the calculation of the prior probability are as follows: (Table 4)

Table 4 Calculating Probability

Class label probability	
Label	**probability value**
Low	0.75
Meidum	0.125
High	0.125

d) Calculating Mean

The calculation of the mean is done after looking for probabilities using the training data in table 2, because the data used uses numeric data, it is necessary to find the average value of each attribute, here is the table of the calculation result (Table 5)

Table 5 Calculating Mean

LABEL	**Close Contact**	**Suspect**	**Porable**
Low	122.75	72.25	3.083333333
Meidum	899	168	3.5
High	574.5	159	8

e) Calculating Standard Deviation

Calculation of standard deviation or standard deviation is done after calculating the Mean of each attribute. By using equation 3, here is a table of the results of the calculation of the standard deviation (Table 6)

Table 6 Calculating Standard Deviation

LABEL	**Close Contact**	**Suspect**	**Porable**
Low	75.77392361	70.92851448	2.722080495
Meidum	409	55	1.5
High	280.5	127	1

f) Calculation of Gaussian Value and Final Probability

In calculating the Gaussian value, there are test data used, namely Cipayung District with close contact data of 180 people, for suspect data as many as 70 people and portable data as many as 8 people. Regarding the prediction results for the Covid level in Cipayung District, which is categorized as "High", with a total number of 9,79561E-05 . (Table 7)

Table 7 Calculating of Gaussian Value and Final Probability

	Close Contact	**Suspect**	**Portable**	**Covid Levels**
CIPAYUNG	180	70	8	?
Low	0.008862057	0.027699802	0.399043442	9.79561E-05
Meidum	0.004208159	0.011000775	0.003619507	1.67558E-07
High	0.034459169	0.035403823	0.047330886	5.7743E-05
The highest score				**9.79561E-05**

g) Naive Bayes Method Accuracy Test

Based on table 4.3, the calculation of the nave Bayes accuracy is carried out by the system using Jupiter Notebook tools, after going through data processing, the results of the accuracy value obtained from the predictions of the nave Bayes system are 0.938 if the percentage is 94%. (Figure 2)

```
              precision    recall  f1-score   support

      RENDAH       0.92      1.00      0.96        12
      SEDANG       1.00      1.00      1.00         2
      TINGGI       1.00      0.50      0.67         2

    accuracy                           0.94        16
   macro avg       0.97      0.83      0.88        16
weighted avg       0.94      0.94      0.93        16

AKURASI Naive Bayes: 0.938
```

Figure 2 Naive Bayes Accuracy

h) Classification Comparison Table

Prior to testing, the data in table 3.2 is converted into an xlsx file that functions as a dataset to be tested using Jupiter Notebook tools with the Python programming language. After that, the initial data of type string will be converted into data of numeric type. The data that will be converted into a numeric type is the "Covid Level" attribute, where the covid level is the reference point as a prediction result. It is known that the parameter labeled "Low" was changed to 0, the parameter labeled "Medium" became 1 and the parameter labeled "High" was changed to 2. After processing the data using Jupiter notebook tools, it turned out that there was one data with predictions that did not match the original dataset. , namely in Parongpong District where the system prediction results state that the level of Covid-19 cases in the area is "Medium", the following is a table from the results of classifying the level of Covid-19 cases in West Bandung Regency by District, (Table 8)

Table 8 Classification Comparison Table

No	Domisili	Kontak Erat	Suspek	Probable	Klasifikasi	
					Asli	NB
1	BATUJAJAR	168	51	5	0	0
2	CIHAMPELAS	146	91	2	0	0
3	CIKALONGWETAN	234	235	3	0	0
4	CILILIN	239	144	8	0	0
5	CIPATAT	78	29	4	0	0
6	CIPEUNDEUY	51	34	4	0	0
7	CIPONGKOR	44	36	0	0	0
8	CISARUA	490	113	2	1	1
9	GUNUNGHALU	33	25	0	0	0
10	LEMBANG	855	286	9	2	2
11	NGAMPRAH	216	175	8	0	0
12	PADALARANG	1308	223	5	1	1
13	PARONGPONG	294	**32**	**7**	**2**	**1**
14	RONGGA	22	4	1	0	0
15	SAGULING	112	3	0	0	0
16	SINDANGKERTA	130	40	2	0	0
TOTAL		4420	1521	60		

i)System Evaluation

Evaluation is used to measure the performance of a system, in this study evaluation is used to measure the accuracy of the text classification method. Evaluation techniques commonly used to measure the accuracy of text classification methods include precision, recall, and f1-score.

In the table there are 12 data that are predicted to have a low level of covid, 2 data that are predicted to have a moderate level of covid and 1 data that is predicted to have a high level of covid. At a high level of covid the model makes an inaccurate prediction because there should be 2 data with a high level of covid. (Table 9)

Table 9 System Evaluation

Current	Prediction		
	Low	**Medium**	**High**
Low	12	0	0
Medium	0	2	0
High	1	0	1

ii) Matrix Confusion Calculation

1) Calculating Accuracy

Rumus = TP / Amount of data
= (12+2+1)/16
= 0.9375 x 100%
= 93,75%

2) Calculating Prescision

= TP/(TP+FP)

All Precision = Precision A + B + C / Number of Classes
= (0.83 + 1 + 0.5) /3
= 0.7766 x 100%
= 77.66%

Table 10 Calculating Prescision

	Low	**Medium**	**High**
Low	12	2	1
Medium	0+0	0+0	1+0
High	12/(12+0)=0.83	2/(2+0)=1	1/(1+1)=0.5

3) Calculating Recall

= TP/(TP+FN)

All Recall = Precision A + B + C / Number of Classes
=(0.75+1+1)/3
= 0.9166 x 100%
= 91.66%

Table 11 Calculating Recall

	Low	**Medium**	**High**
Low	12	2	1
Medium	0+1	0+0	0+0
High	12/(12+4)=0.75	2/(2+0)=1	1/(1+0)=1

4) Calculating F1 Score

= 2 x (Recall x Precision)/ (Recall + Precision)
= 2 x (0.9166 x 0.7766)/ (0.9166 + 0.7766)
= 1.4326 / 1.6932
= 0.872

4. Conclusion

The Naive Bayes Algorithm method used to classify Covid-19 cases in West Bandung Regency resulted in a fairly good accuracy of 94% where out of 16 training data, 1 of them was classified as incorrect. It can be seen that the Covid-19 cases in West Bandung Regency tend to be high in only a few sub-districts. And it can also be concluded that the level of COVID-19 cases in West Bandung district is in a fairly high category for several sub-districts.

References

. M., Syarif, S., & . A. Penerapan Algoritma Naïve Bayes Pada Penilaian Kinerja Pemerintah Desa Dalam Pengelolaan Dana Desa. *Jurnal It*, *10*(1), 11–23. (2019). https://doi.org/10.37639/jti.v10i1.92

Darwis, D., Siskawati, N., & Abidin, Z. Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131.

(2021).https://doi.org/10.33365/jtk.v15i1.744

Harianto Kristanto, N., Christopher, A. LA, & Budi, H. S. Implemantasi K-Means Clustering untuk Pengelompokan Analisis Rasio Profitabilitas dalam Working Capital. *Juisi*, *02*(01). (2016).

Ikbal, M., Andryana, S., & Komala Sari, R. T. Visualisasi dan Analisa Data Penyebaran Covid-19 dengan Metode Klasifikasi Naïve Bayes. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *5*(4), 389. (2021). https://doi.org/10.35870/jtik.v5i4.233

Liliana, D. Y., Maulana, H., & Setiawan, A. (*Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes*. *7*(1), 48–53. 2021).

Nurdiana, N., & Algifari, A. *Naive Bayes Untuk Klasifikasi Penyakit*. 18–23. (2015).

Rahmayadi, A. P. U., Enri, U., & Purwantoro, P. Klasifikasi Kinerja Asisten Laboratorium Selama Pandemi Covid-19 Menggunakan Algoritma Naïve Bayes. *Journal of Applied …*, *5*(2), 122–127. (2021). https://103.209.1.42/index.php/JAIC/article/view/3261

Ratino, Hafidz, N., Anggraeni, S., & Gata, W. Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *Jurnal JUPITER*, *12*(2), 1–11. (2020).

Rosandy, T. PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA. *Jurnal Teknologi Informasi Magister Darmajaya*, *2*(01), 52–62. (2016).

Saputra, M. F. A., Widiyaningtyas, T., & Wibawa, A. P. Illiteracy classification using K means-naïve bayes algorithm. *International Journal on Informatics Visualization*, *2*(3), 153–158. (2018). https://doi.org/10.30630/joiv.2.3.129

Sengkey, D. F., Kambey, F. D., Lengkong, S. P., Joshua, S. R., & Kainde, H. V. F. Pemanfaatan Platform Pemrograman Daring dalam Pembelajaran Probabilitas dan Statistika di Masa Pandemi CoVID-19. *Jurnal Informatika*, *15*(4), 217–224. (2020).

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yunihastuti, E. Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, *7*(1), 45. (2020). https://doi.org/10.7454/jpdi.v7i1.415

Zaki, M. J., & Meira, M. J. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. (2013). https://books.google.com.tr/books?id=Gh9GAwAAQBAJ&lpg=PR9&dq=Data Mining and Analysis: Foundations and Algorithms&hl=tr&pg=PR9#v=onepage&q=Data Mining and Analysis: Foundations and Algorithms&f=false