Jurnal Mantik
Volume 4 Number 1 May 2020, pp. 267-275 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Application of Data Mining Classification for Covid-19 Infected Status Using Algortima Naïve Method

Puji Hari Santoso[1], Fauziah[2], Nurhayati[3]

Universitas Nasional, Jl. Sawo Manila, Jakarta Selatan, Jakarta, Indonesia 12520

Email: [1]pujihari.sss@gmail.com, [2]fauziah@civitas.unas.ac.id, [3]nurhayati@civitas.unas.ac. id

ARTICLE INFO

Article history:
Received: 14/04/2020
Revised: 28/04/2020
Accepted: 01/05/2020

Keywords:
Data Mining,
Covid - 19,
Naïve Bayes Algorithm,
Classification,
Rapidminer,
Prediction

ABSTRACT

The number of virus infected status known as covid -19 is increasing in the southern Jakarta area, namely Pondok Labu, West Cilandak, Jagakarsa, Lenteng Agung, Pasar Minggu and Ragunan, it is necessary to classify the data to find out the negative status of covid-19 virus infection or positive infected with covid virus -19. The technique of classifying positive or negative covid-infected virus status with the naïve bayes classification method. To manage the data, software rapid miner 9. 6 is used, the infected status dataset covid -19 is obtained from the websitejeo. compass. CalculationPrediction shows the classification of the Naïve Bayes method obtained a positive prediction that shows a figure of 55.48% and a negative prediction result of 44.52%. From the results of the classification of data that has been obtained can be seen that the largest prediction found in the positive status infected with covid -19 virus reached 55.48%.

1. Introduction

Corona virus or COVID-19, the case began with pneumonia or mysterious pneumonia in December 2019. This case is allegedly related to the huana animal market in Wuhan that sells various types of animal meat, including those that are not commonly consumed, such as snakes, bats, and various types of mice.

Cases of infection with this mysterious pneumonia are indeed commonly found in the animal market. Corona virus or COVID-19 allegedly brought bats and other animals that humans eat until transmission occurs. Corona virus is actually no stranger to the world of animal health, but only a few types are able to infect humans to become pneumococcal disease.

Based on the outbreak, research was conducted using data mining techniques to classify data using the Naïve Bayes algorithm.

2. Literature Review

2.1. Covid Virus Infected Status Classification -19

In order to assist handling in classifying the status of the population infected or not in the case of the covid-19 virus. Aiming to more quickly classify those infected with the covid-19 virus can be separated and directly isolated in order to further medical treatment.

2.2. Covid Virus -19

Corona virus is a collection of viruses that can infect the respiratory system. In many cases, this virus only causes mild respiratory infections, such as flu. However, this virus can also cause severe respiratory infections, such as lung infections (pneumonia),*Middle-East Respiratory Syndrome*(MERS), and Severe Acute Respiratory Syndrome (SARS).

2.3. Data Mining

Data Mining is a process or activity to collect large data and then extract the data into information - information that can later be used.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2.4. Naïve Bayes

Naive Bayes is a statistical classification that can be used to predict the probability of a class. Naive Bayes is based on Bayes' theory that has the same classification capabilities as decision trees and neural networks. Naive Bayes is proven to have high accuracy and speed when applied to databases with large data [5]. Bayes' prediction is based on the Bayes theorem formula with the following general formula:

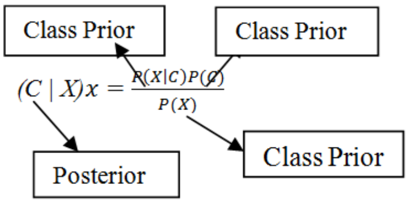


Fig. Naïve Bayes formula

Information :

X : Data with unknown classes
C : Data hypothesis is a specific class
P (c | x) : Hypothesis probability based on conditions (posteriori probability)
P (c) : Probability hypothesis (priorprobability)
P (x | c) : Probability based on conditions on the hypothesis
P (x) : Probability

In Figure 1, the formula explains that the chance of entering certain characteristic samples in class C (Posterior) is the chance of the emergence of class C (before the sample entry, often called prior), multiplied by the chance of the appearance of characteristic characteristics of the sample in class C (also called likelihood), divided with the opportunity for the emergence of global sample characteristics (also called evidence). Therefore, the formula in Figure 1 can also be written as follows:

$$Posterior = \frac{Prior\ x\ likelihood}{evidence} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [1]$$

Evidence values are always fixed for each class in one sample. The value of the posterior will then be compared with the value of the posterior grades of other classes to determine to which class a sample will be classified. Further elaboration of the Bayes formula is carried out by describing (c | x1, ..., xn) using the multiplication rules as follows:

$$\begin{array}{ll} P\ (C \mid x1, ..., xn = & P\ (C)\ P\ (x1, ..., xn \mid C) \\ = & P\ (C)\ P\ (X1 \mid C)\ P\ (X2, ..., Xn \mid C, X1) \\ = & (C)\ P\ (X1 \mid C)\ P\ (X2 \mid C, \\ & X1)\ P\ (X3, ...\ Xn \mid C, X1, X2\ (C)\ P\ (X1 \mid C)\ P\ (X2 \mid C, X1)\ P\ (X3 \mid \\ = & C, X1, X2)\ P\ (X4, ..., Xn \mid C, X1, X2, X3) \\ & P\ (C) \\ = & P\ (X1 \mid C)\ P\ (X2 \mid C, X1)\ P\ (X3 \mid C, \\ & X1, X2) ...\ P \\ & (Xn \mid C, X1, X2, X3, ..., Xn\text{-}1 .... [2] \end{array}$$

It can be seen that the results of the elaboration cause more and more complex factor conditions that affect the probability value, which is almost impossible to analyze one by one. As a result, the calculation becomes difficult to do. Here is used the assumption of independence, which is very high (naive), that each of the instructions is independent of each other. With these assumptions, the following similarities apply:

$$P\,(C/X1,\ldots\ldots,Xn) = P(C) \prod_{i=1}^{n} P(Xi|C) \ldots. [3]$$

$$P(C|X) = P(X_1|C)P(X_2|\ C)\ P\ (X_n|c)p(X_n|c)p(c)..\ [4]$$

The equation above is a model of the Naive Bayes theorem which will then be used in the classification process. For continuous data the Gauss Density formula is used:

$$P = (X_i = X_i | Y_i = = e \text{ - } \ldots [5Y_i) \frac{1}{\sqrt{2\pi\sigma ij}} \frac{(X_i - \mu_{ij})2}{2\sigma^2 ij}]$$

Information :

[1]: Opportunity
Xi : Attribute to i

Xi : Attribute value to i
Y : Class sought
Y j : Sub class Y sought
u : Mean, states the average of all attributes
o : Standard deviation, representing variants of all attributes
The mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} xi \ldots\ldots [6]$$

Standard Deviation:

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)\dot{2}\right]^{0,5} \ldots.\ [7]$$

Naive Bayes data mining algorithm is also widely used in previous studies, especially to classify data. Some examples of previous studies using the Naive Bayes algorithm include research titled opinion mining for product reviews with Naive Bayes classification. In this study, the accuracy rate of data processing using the Naive Bayes method was 77.7% (Jeremy, A., Christanti, V., & Mulyawan, 2018). In addition to these studies, another study using the Naive Bayes algorithm is entitled the Naive Bayes Classifier algorithm in web-based data mining applications. Naive Bayes algorithm researchers know that there are no studies that also prove the results with manual testing with rapid miners.

3. Research Methods

This study uses the concept of data mining techniques with the help of rapid miner studio software. Data used are data from the official website obtained from a website that provides public data about information about covid-19 in Jakarta, namely jeo. compass. Information data about covid-19 is sourced from the Jakarta Ministry of Health. The data will be used as training data and testing data. The attributes contained in the covid-infected status dataset are -19, namely age, sex, district, kelurahan, ODP status and PDP status. As for the class contained in the dataset, Infected Status with Positive and Negative values.

4. Results and Discussion

To determine the data that will be classified by the Naïve Bayes method, the first step is to read the data taken dated April 21, 2020. The training data displayed 11 data from 249 data used can be seen in table 1 below:

Table 1
Training data

No	Age	Gender	Sub-District	Village	ODP Status	PDP status	Status
1	21	Male	Cilandak	Pondok Labu	Finished Monitoring	Healthy and may go home	Positive
2	21	Male	Cilandak	Pondok Labu	Monitoring Process	In care	Positive
3	21	Male	Cilandak	West Cilandak	Finished Monitoring	In care	Positive
4	21	Female	Jagakarsa	Jagakarsa	Monitoring Process	Healthy and may go home	Positive
5	21	Male	Jagakarsa	Great Lenteng	Finished Monitoring	In care	Positive
6	22	Female	Jagakarsa	Great Lenteng	Finished Monitoring	Healthy and may go home	Positive
7	22	Female	Sunday market	Sunday market	Monitoring Process	In care	Positive
8	22	Female	Sunday market	Ragunan	Finished Monitoring	In care	Negative
9	22	Male	Sunday market	Ragunan	Finished Monitoring	In care	Negative
10	21	Female	Cilandak	Sunday market	Monitoring Process	In care	Positive
....	………	………..	………..	………….	………….	………..	………..
249.	24	Female	Jagakarsa	Jagakarsa	Finished Monitoring	In care	Negative

Source: jeo. compass

Information :

[1] Attribute 1 explains about "Age / Age"

[2] Attribute 2 explains about "Gender"

[3] Attribute 3 explains about "Sub-district"

[4] Attribute 4 explains about "Kelurahan"

[5] Attribute 5 explains about "ODP Status"

[6] Attribute 6 explains about "PDP status"

Early stages of the calculation process

Naive Bayes is to take training data from the data that has been obtained. The variable to be used in the classification status is covid -19 infection.

4.1 Calculating Priority Probabilities (P (Ci))

Calculate the number of classes of covid-infected virus status based on the classification formed (prior probability):

1. C0 (Class Status = "Negative) = Number of Negatives 126/249 = 0. 50
2. C1 (Class Status = "Positive) = Positive number 123/249 = 0. 49

4.2 Calculation of PosteriorX Probability (P (X | Ci))

The calculation of the posterior probability is carried out on 249 exercises by using X as a vector. The selection of covid-19 attribute classifications are X Age, X type sex, X sub-district, X village, X status STP, X status STP ie every X is calculated the probability of each Ci. To calculate every possibility. Possible results of attribute P (XX age | CI) can be seen in table 2.

Table 2

Age Probability

No	Age	Positive	Negative	P (positive)	P (negative)
1	21	34	32	0. 285714	0. 266446
2	22	14	26	0. 117647	0. 214876
3	23	30	27	0. 252100	0. 223140
4	24	26	23	0. 218487	0. 190082
5	25	15	13	0. 126050	0. 107438
Total		119	121		

To calculate every possible result of attribute P (XX Gender | Ci) can be seen from table 3

Table 3

Gender Probability

No	Gender	Positive	Negative	P (positive)	P (negative)
1	Male	62	57	0. 508196	0. 459677
2	Girl	60	67	0. 491803	0. 540322
Total		122	124		

To calculate every possible attribute P (XX District | Ci) can be seen from table 4.

Table 4

District Probability

No	sub-district	Positive	Negative	P (positive)	P (negative)
1	Cilandak	48	50	0. 421052	0. 403225
2	Jagakarsa	46	49	0. 403508	0. 395161
3	Sunday market	20	25	0. 175438	0. 201612
Total		114	124		

To calculate each possible attribute P (XX Village | Ci) can be seen from table 5

Table 5

Village Probability

No	Kelurahan	Positive	Negative	P (positive)	P (negative)
1	Pondok Labu	38	36	0. 319327	0. 297520
2	Cilandak Barat	15	14	0. 126050	0. 115702
3	Jagakarsa	23	22	0. 193277	0. 206611
4	Lenteng Agung	23	25	0. 193277	0. 206611
5	Pasar Minggu	13	14	0. 109243	0. 115702
6	Ragunan	7	10	0. 058823	0. 082644
TOTAL		119	121		

To calculate each possible attribute P (XX ODP | Ci) can be seen in table 6

Table 6
ODP Status Probability

No	ODP Status	Positive	Negative	P (positive)	P (negative)
1	Finished Monitoring	64	52	0. 512396	0. 429752
2	Monitoring Process	57	69	0. 471074	0. 570247
Total		121	121		

To calculate every possible result of attribute P (XX Status PDP | Ci) can be seen from table 7

Table 7
PDP Status Probability

No	PDP status	Positive	Negative	P (positive)	P (negative)
1	Healthy & may go home	70	44	0. 555555	0. 360655
2	In care	56	78	0. 444444	0. 639344
Total		126	122		

4.3 Manual Calculation

The following is a manual calculation using test data which can be seen in. Table 8 using the naïve bayes method.

Table 8
Test Data

Age	Gender	sub-district	Kelurahan	ODP Status	PDP status	Status
21	Male	Cilandak	Pondok Labu	Monitoring Process	Healthy & may go home	?

4.4 Calculation of Probability of Test Data

Based on the test data in table 8 classification can be carried out into the Negative status (C0) status of the infected virus covid -19 with the provisions of the value of each attribute, namely: 0. 266446, 0. 459677, 0. 403225, 0. 297520, 0. 570247, 0. 360655.

Then the value of each attribute is multiplied.

P (21 | C0) * P (Male | CO) * P (Cilandak | C0) * P (PondokLabu | C0) * P (Monitoring Process | C0) * P (Healthy & may go home | C0).

= 0. 266446 x 0. 459677 x 0. 403225 x 0. 297520 x 0. 570247 x 0. 360655

= 0. 00302190329

To calculate the classification into Positive status (C1), the infection status is covid -19 with the following values: 0. 285714, 0. 508196, 0. 421052, 0. 319327, 0. 471074, 0. 555555.

Then the value of each attribute is multiplied.

P (21 | C1) * P (Male | C1) * P (Cilandak | C1) * P (PondokLabu | C1) * P (Monitoring Process | C1) * P (Healthy & may go home | C1).

= 0. 285714 x 0. 508196 x 0. 421052 x 0. 319327x 0. 471074 x 0. 555555

= 0. 00510916978

4.5 Maximizing P (X | Ci) P (Ci)

The maximization calculation for the classification of Negative status (C0) classes is to multiply P (X | C0) with P (C0):

P (C0 | X) = P (X | C0) xP (C0)

= 0. 00302190329 x 0. 50

= 0. 00151095164

Then for the Positive status class (C1) is by switching P (X | C1) with P (C1) P (C1) X) = P (X | C1) xP (C1)

= 0. 00510916978 x 0. 49

= 0. 00250349319

From the calculation of Maximization can be produced

P (C0 | X) = 0. 00151095164.

P (C1 | X) = 0. 00250349319.

Based on these values conclusions can be drawn that. P (C0 | X) <P (C1 | X). then the test data contained in table 8 can be classified into Positive status in this covid-19 infected status.

4.6 Implementation with Rapid Miner

Following is the management of data sets and test data using naïve bayes

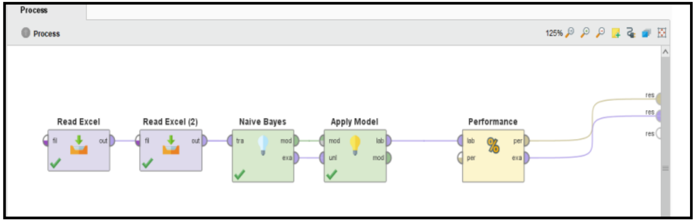
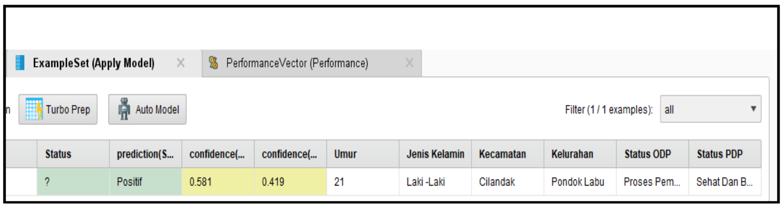


Fig 2. Process training data and test data

The main purpose of this study is to determine the prediction of the value of the naïve bayes algorithm which is used to classify the infected status of the covid -19 virus, in the read excel training there is a classification algorithm applied, namely naïve bayes. While in the testing column there is an Apply Model to run the naïve bayes model and performance to measure the predicted performance of the naïve bayes model.

4.7 Testing and Trial Results

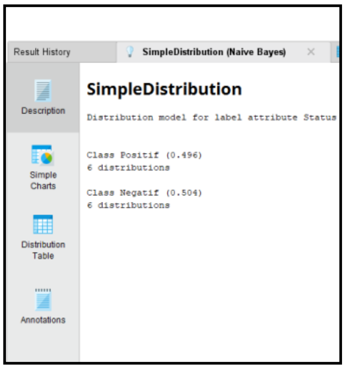


Status	prediction(S...	confidence(...	confidence(...	Umur	Jenis Kelamin	Kecamatan	Kelurahan	Status ODP	Status PDP
?	Positif	0.581	0.419	21	Laki -Laki	Cilandak	Pondok Labu	Proses Pem...	Sehat Dan B...

Fig 3. Predictive Classification Results

The results of the prediction of the naïve bayes model classification that the test data with the status classification model infected with the covid-19 virus in Figure 3 show that the prediction of the infection status of the virus is stated to be Positive.

With this classification the naïve bayes method uses rapid miners, the results are the same and in accordance with manual calculations.



SimpleDistribution

```
Distribution model for label attribute Status

Class Positif (0.496)
6 distributions

Class Negatif (0.504)
6 distributions
```

Fig 4. Simple Distribution

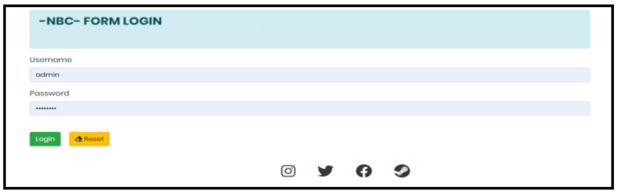
The distribution model for the status class attribute label is as follows:

Positive Class: 6 Distribution

Negative Class: 6 Distribution

The experiments in this study used rapid miner 9. 5. 00. 1. Algortima used data testing and Apply models to run the algorithm or the naïve bayes model to predict the classification of the naïve bayes model.

4.8 Implementation with a website

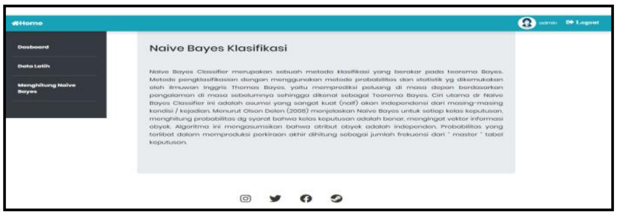


-NBC- FORM LOGIN

Username: admin

Password: ••••••••

Fig 5. Logic form index page

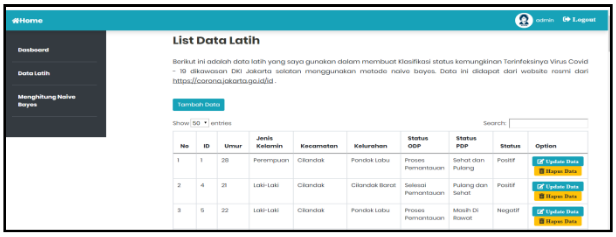


Naive Bayes Klasifikasi

Naive Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dr Naive Bayes Classifier ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi / kejadian. Menurut Olson Delen (2008) menjelaskan Naive Bayes untuk setiap kelas keputusan, menghitung probabilitas dg syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari " master " tabel keputusan.

Fig 6. The main menu page or dashboard

Inside a dashboard there is a function that is the dashboard, which contains the contents of the nbc or naïve bayes classification. And there are other functions, namely training data, and naïve Bayes calculations using web-based applications.



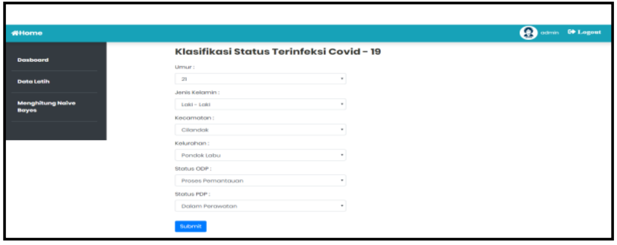
List Data Latih

Berikut ini adalah data latih yang saya gunakan dalam membuat Klasifikasi status kemungkinan Terinfeksinya Virus Covid – 19 dikawasan DKI Jakarta selatan menggunakan metode naive bayes. Data ini didapat dari website resmi dari https://corona.jakarta.go.id/id .

Show 50 entries Search:

No	ID	Umur	Jenis Kelamin	Kecamatan	Kelurahan	Status ODP	Status PDP	Status	Option
1	1	28	Perempuan	Cilandak	Pondok Labu	Proses Pemantauan	Sehat dan Pulang	Positif	Update Data Hapus Data
2	4	21	Laki-Laki	Cilandak	Cilandak Barat	Selesai Pemantauan	Pulang dan Sehat	Positif	Update Data Hapus Data
3	5	22	Laki-Laki	Cilandak	Pondok Labu	Proses Pemantauan	Masih Di Rawat	Negatif	Update Data Hapus Data

Fig 7. In Figure 7 this is the display of training data or training data



Klasifikasi Status Terinfeksi Covid – 19

Umur : 21

Jenis Kelamin : Laki - Laki

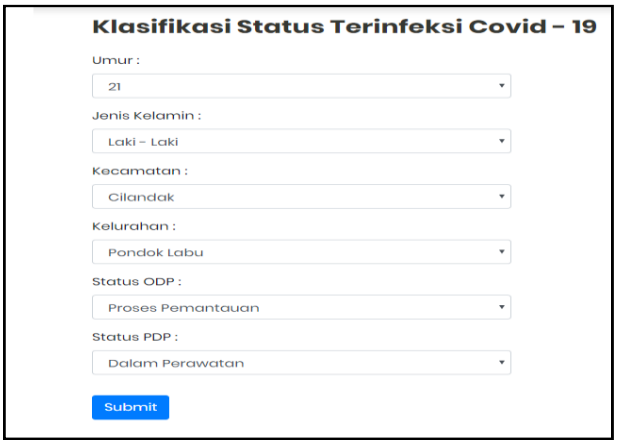
Kecamatan : Cilandak

Kelurahan : Pondok Labu

Status ODP : Proses Pemantauan

Status PDP : Dalam Perawatan

Fig 8. In Figure 8 this is the display form to calculate the naïve bayes classification



Klasifikasi Status Terinfeksi Covid – 19

Umur : 21

Jenis Kelamin : Laki – Laki

Kecamatan : Cilandak

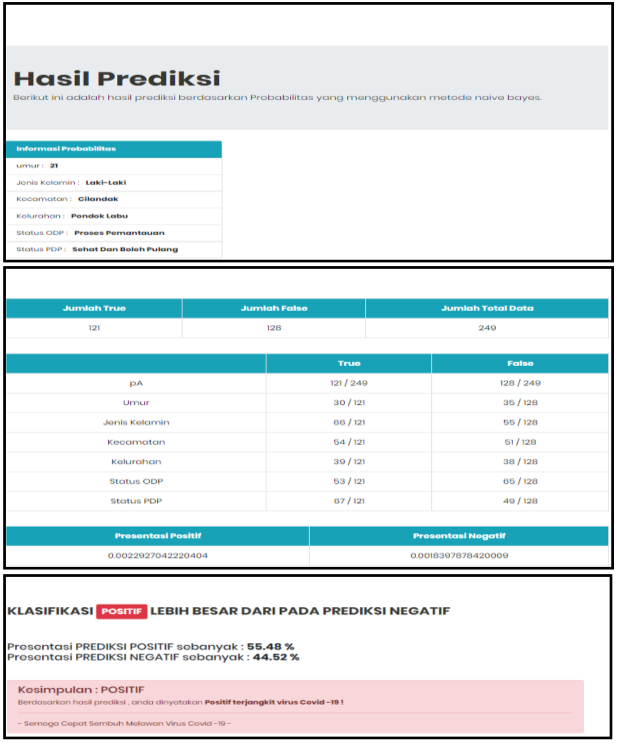
Kelurahan : Pondok Labu

Status ODP : Proses Pemantauan

Status PDP : Dalam Perawatan

Fig 9. In Figure 9 this is the display form to calculate the naïve bayes classification

In the calculation of the naïve bayes classification method using the web by entering the data according to the test data in table 8, namely with attribute 21, Men, Cilandak, Pondok pumpkin, monitoring process and in maintenance.



Hasil Prediksi

Berikut ini adalah hasil prediksi berdasarkan Probabilitas yang menggunakan metode naive bayes.

Informasi Probabilitas
umur : **21**
Jenis Kelamin : **Laki-Laki**
Kecamatan : **Cilandak**
Kelurahan : **Pondok Labu**
Status ODP : **Proses Pemantauan**
Status PDP : **Sehat Dan Boleh Pulang**

Jumlah True	Jumlah False	Jumlah Total Data
121	128	249

	True	False
pA	121 / 249	128 / 249
Umur	30 / 121	35 / 128
Jenis Kelamin	66 / 121	55 / 128
Kecamatan	54 / 121	51 / 128
Kelurahan	39 / 121	38 / 128
Status ODP	53 / 121	65 / 128
Status PDP	67 / 121	49 / 128

Presentasi Positif	Presentasi Negatif
0.0022927042220404	0.0018397878420009

KLASIFIKASI POSITIF LEBIH BESAR DARI PADA PREDIKSI NEGATIF

Presentasi PREDIKSI POSITIF sebanyak : **55.48 %**
Presentasi PREDIKSI NEGATIF sebanyak : **44.52 %**

Kesimpulan : POSITIF
Berdasarkan hasil prediksi , anda dinyatakan **Positif terjangkit virus Covid -19 !**

- Semoga Cepat Sembuh Melawan Virus Covid -19 -

Fig 10. Figure 10 shows the results of the calculation of the web-based naïve bayes method

From the interim research on corona as of 21 April 2020 it can be concluded that, in Figure 10 is the result of the calculation of test data in table 8 with the Naïve Bayes method using a web-based application. The prediction results of the covid-19 infected status classification obtained positive and negative data with each data ie positive 55. 48% of 249 data and negative 44. 52% of 249 data. In other words, if classified as numbers, the positive classification is 139 people and the negative classification is 110 people from a total of 249 data.

5. Conclusion

Based on the results of research conducted it can be said, that the processing of data to classify the infected status of covid -19 using the naïve bayes algorithm has good results and can be used as a reference for people who know the classification of infected status covid -19. In addition to using rapid miner studio software, researchers also try to manage data with manual calculations and use a web-based application to reference the prediction of classification classification coinfected infection status. From the results of data processing manual calculation, rapid miner and web-based applications, it can be concluded that the results of the prediction of the classification of data processing status are infected with covid-19 virus.

6. Reference

[1] Karisani, N. and Karisani, P., 2020. Mining Coronavirus (COVID-19) Posts in Social Media. arXiv preprint arXiv: 2004. 06778.
[2] Kusrini, Lutfi, Emha Taufiq. (2013). Data Mining Algorithm.
[3] Narin, A., Rich, C. and Pamuk, Z., 2020. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv: 20
[4] Apostolopoulos, ID and Mpesiana, TA, 2020. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engin

[5] Ardabili, Sina F., Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson. "COVID-19 Outbreak Prediction wit

[6] Murdiansyah, AO, & Siswanto, S. (2018). NaiveBaiyes Classsifier Algorithm in WEB Based Data Mining Applications. SCANIKA, 1 (1), 284-290.

[7] Rahmani, ME, Amine, A. and Hamou, RM, 2020. Supervised machine learning for plants identification based on images of their leaves. In Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 1314-1330). IGI Global

[8] Farid, Ahmed Abdullah, Gamal Ibrahim Selim, and Hatem Awad A. Khater. "A Novel Approach of CT Images Feature Analysis and Prediction to Screen for Corona Virus Disease (COVID-19

[9] Kononenko, Igor, Matjaz Bevk, Sasa Sadikov, and Luka Sajn. "DIFFERENT TYPES OF CORONAS AND MACHINE LEARNING."