PREDICTION OF ANEMIA USING THE PARTICLE SWARM OPTIMIZATION (PSO) AND NAÏVE BAYES ALGORITHM

Septiana Tri Utami[1], Sriyanto[2*], Sri Lestari[3], Handoyo Widi Nugroho[4], Zarnelly[5]

[1,2,3,4]Institut Informatika dan Bisnis Darmajaya . Jl.Zainal Abidin PA No.93 Bandar lampung 35142,
[5]Prodi Sistem Informasi UIN SUSKA Riau, Jl H.R. Soebrantas Km 15, Pekanbaru, Riau
[1]Septiana.2221210030@mail.darmajaya.ac.id, [2]sriyanto@darmajaya.ac.id,, [3]srilestari@darmajaya.ac.id,
[4]handoyo.wn@darmajaya.ac.id, [5]zarnelly@uin-suska.ac.id

Abstract. Health is one of the factors that can influence human activities, but many people rarely or even do not pay attention to their own body health, so that diseases easily come without realizing it, so it is often too late to diagnose the disease they are suffering from. Of the many diseases, one disease that is often diagnosed late is anemia. Implementing the Naïve Bayes and PSO algorithms to predict anemia and evaluating the prediction results using the accuracy parameters of the Naïve Bayes and PSO algorithms. In this research, the Naïve Bayes and PSO algorithms will be applied by optimizing attributes derived from the dataset to predict anemia. PSO can be used to improve model performance or find the best combination of features. PSO can help adjust parameters or select the most informative features to increase accuracy. Naive Bayes model predictions. Once the model is trained, Naive Bayes can be used to predict whether a patient is anemic based on certain features. Naive Bayes calculates the probability for each class based on the given test data. The class with the highest probability will be considered as predicted. Based on the results of testing the Naïve Bayes and PSO algorithm models which were carried out through confusion matrix evaluation, it was proven that the tests carried out by the Naïve Bayes algorithm were 93.88% and the tests carried out with Naïve Bayes and PSO had a high accuracy value, namely 94.02%. The purpose of selecting information acquisition features is to select features or attributes that have a significant influence on anemia. Using PSO can increase a higher level of accuracy. The success of this model makes a positive contribution to efforts to prevent and treat anemia.

Keywords: Anemia, Naive Bayes, Particle Swarm Optimization (PSO), Prediction

INTRODUCTION

"Health is one of the factors that can influence human activities, however, many people rarely or even do not pay attention to their own body health, so that diseases can easily come without realizing it, so it is often too late to diagnose the disease they are suffering from. Of the many diseases, one that is often too late to diagnose is anemia, anemia is a disease that is considered normal by society" [1].

Anemia is a nutritional disorder that is still often found in Indonesia. The Household Health Survey found that around 57% of Indonesian adolescent girls aged 10-14 years and 39.5% of girls aged 15 years and over still suffer from anemia. The survey figures show that women are still susceptible to anemia, so extra vigilance is needed. And the results of the 2010 Population Census show that 63.4 million of Indonesia's population are teenagers, consisting of 31.2 million young women (49.30%). According to WHO, "the prevalence of anemia in the world ranges from 40-88% with the incidence of anemia in adolescent girls, especially in developing countries, reaching 53.7%" [2]. The main risk factors for iron deficiency anemia are low iron intake, poor iron absorption, and periods of life when the need for iron is high such as during growth, pregnancy, and breastfeeding. Anemia can generally occur in pregnant women, teenagers, the elderly and even babies who have anemia. Adolescent girls are a group at higher risk of experiencing anemia compared to adolescent boys, where the need for iron absorption peaks at the age of 14-15 years in adolescent girls, while in adolescent boys one or two years later. The impact of anemia on adolescent girls and poor nutritional status makes a negative contribution when pregnant as a teenager or as an adult, which can cause the birth of babies with low birth weight, morbidity and even death for mothers and babies. Apart from that, anemia also has a negative impact on the physical and cognitive development of adolescents [3]. Anemia is more common in pregnancy, this is because during pregnancy the need for food substances increases and changes occur in the blood and bone marrow, resulting in iron deficiency, or

what is called iron nutritional anemia. Anemia in pregnancy has an adverse effect on pregnancy, childbirth and the postpartum period, so anemia must be treated immediately so that it does not have an impact on the fetus. Anemia can also affect pregnancy and childbirth, one of which is premature birth. Food is one of the determining factors in the occurrence of anemia [4]. The condition of iron anemia can affect psychomotor development because it can inhibit the synthesis of fatty acids and cholesterol by oligodendrocytes for the production and breakdown of substances that act as transmitters that transmit stimulation from one neuron cell to another. The condition of iron anemia can also be indirectly this affects the level of socio-economic welfare of society, because it can result in sub-optimal cognitive development, which will cause a decrease in work productivity [5].

According to Riskesdas in 2013, "the prevalence of anemia reached 21.7% including 26.4% for ages 5-14 years and 18.4% for ages 15-24 years". Meanwhile, according to Riskesdas in 2018, "the prevalence of anemia in pregnant women reached 48.9%". One of the causes of the high prevalence of anemia is the public's lack of knowledge about detecting anemia. This is proven by research conducted on 30 respondents which resulted in 90% of them not knowing that the respondent had anemia [6]. Due to the high number of cases of anemia sufferers, to be able to overcome this problem, much research has been carried out in the field of computer science, including the Expert System for Diagnosing Anemia Using the Web-Based Bayes Theorem Method. The research results show that the accuracy value is 50.52% [7]. The second study, Classification of Anemia using Digital Images of Nails and Palms using the Naive Bayes Method, the accuracy produced by this study was 87.5% with different light intensities, and the accuracy level was 92.3% with a light intensity of 5362 Lux [8]. The Naive Bayes algorithm has a fairly high accuracy value when compared to the Bayes Theorem method.

Based on case analysis and journal review results in this research the author will use data mining techniques. The algorithms used in this research are Naïve Bayes and PSO by optimizing attributes to predict anemia. The tool used is a rapid miner, so it can predict anemia. This algorithm is good for detecting anemia caused by a lack of red blood cells.

METHODS

Methodology is a process carried out to solve a problem raised in a research so that accurate results can be found and conclusions can be drawn [9]. This research method explains in detail all the sequences in carrying out the research. The dataset used is the anemia dataset with a total of 1421 data records, consisting of 5 attributes, with the models used are Naive Bayes and PSO.

Anemia is a condition in which the number of red blood cells or the concentration of hemoglobin in them is lower than normal. Hemoglobin is needed to carry oxygen and if you have too few or abnormal red blood cells, or not enough hemoglobin, there will be a decrease in the blood's capacity to carry oxygen to the body's tissues. This causes symptoms such as fatigue, weakness, dizziness and shortness of breath, among others. The optimal hemoglobin concentration required to meet physiological needs varies based on age, gender, altitude of residence, smoking habits and pregnancy status. Anemia can be caused by several factors: nutritional deficiencies through inadequate diet or inadequate nutrient absorption, infections (e.g. malaria, parasitic infections, tuberculosis, HIV), inflammation, chronic diseases, gynecological and obstetric conditions, and red blood cell disorders, lowered. The most common cause of nutritional anemia is iron deficiency, although deficiencies of folate, vitamins B12 and A are also important causes [10].

Particle Swarm Optimization (PSO) is a very simple optimization technique for implementing and modifying several parameters. In Particle Swarm Optimization (PSO), there are several techniques for optimization, including increasing the attribute weights of all the attributes or variables used, selecting attributes (attribute selection), and feature selection [11]. Particle swarm optimization is an algorithm that is inspired by the social behavior of animals such as birds, bees and fish. An animal in the PSO algorithm will be considered a particle. These particles will be influenced by the intelligence of the individual animal itself and the intelligence of other particles in a group. If one particle finds the right and shortest path to a food source, then what happens is that the other particles will follow the particle that has found the right and shortest path [12]. The Naïve Bayes algorithm uses mathematical branching techniques by looking for the largest possible opportunity in classification based on the frequency of each classification of training data which is often called probabilistic theory. The calculation formula for Naïve Bayes is as follows [13].

$$P(X|Y) = \frac{P(Y|X)xP(X)}{P(Y)} \qquad (1)$$

The configuration matrix is a table consisting of the number of rows of test data that are predicted correctly and incorrectly by the classification model used. The Confusion Matrix table is needed to select the best performance of a classification model [14]. Confusion matrix is a 2x2 matrix that represents the results of binary classification on a dataset. There are several general formulas that can be used to calculate classification performance. The results of the accuracy, precision and recall values can be displayed in percentages [15].

Steps in the research process to be carried out. The main goal is to analyze and look for data patterns that will be used as a dataset. This will help make research easier and carry it out systematically according to the desired goals. To achieve this, the following steps were designed as a flow in the research stages to be carried out: The research steps taken in this study are shown in Figure 1.
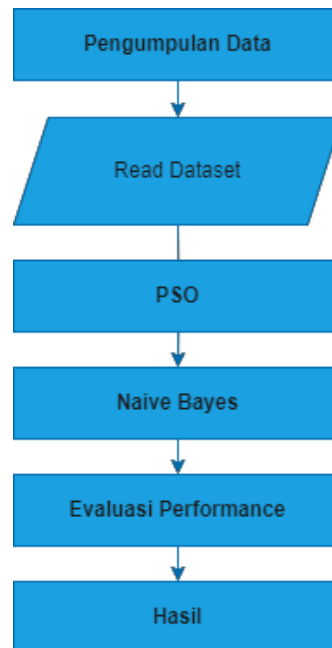

Figure 1. Research Stages

The following is an explanation of Figure 1 which is the flow of this research

1. **Dataset Collection:**
   Collect a dataset containing information about anemia, including relevant attributes. This dataset should include information about whether the patient is anemic or not.
2. **Read Dataset:**
   A step or operation that involves importing or reading a dataset from a file or external data source into the Rapidminer software.
3. **Particle Swarm Optimization (PSO):**
   PSO can be performed to improve model performance or find the best combination of features. PSO can help adjust parameters or select the most informative features to improve the prediction accuracy of Naive Bayes models.
4. **Naive Bayes Method:**
   Once the model is trained, Naive Bayes can be used to predict whether a patient has anemia based on certain features. Naive Bayes calculates the probability for each class based on the given test data. The class with the highest probability will be considered as predicted.
5. **Model Evaluation:**
   Researchers can use test sets to test model performance. Next, calculate evaluation metrics such as accuracy, precision, Recall, to evaluate the extent to which the model can accurately predict anemia
6. **Results**
   Results are results obtained from data analysis carried out in order to answer research questions. Research results reflect findings or information found by researchers based on data that has been collected and analyzed.

RESULT AND DISCUSSION

In this research, the Naive Bayes and PSO methods were tested, the tools used to make calculations easier were using Rapid Miner Studio software. The dataset used comes from kaggel.com Anemia dataset. The number of data records is 1421 data consisting of 5 attributes and 1 label. This data set is used to predict whether a patient is likely to have anemia. Anemia data can be seen in Figure 2 :

| Gender | Hemoglobin | MCH | MCHC | MCV | Result |
integer	real	real	real	real	integer
1	14.900	22.700	29.100	83.700	0
0	15.900	25.400	28.300	72.000	0
0	9.000	21.500	29.600	71.200	1
0	14.900	16.000	31.400	87.500	0
1	14.700	22.000	28.200	99.500	0
0	11.600	22.300	30.900	74.500	1
1	12.700	19.500	28.900	82.900	1

Figure 2. Anemia dataset

Application of the Naïve Bayes Algorithm

The application of data in Rapid Miner for Anemia Prediction using the Naïve Bayes algorithm is shown in the image below:
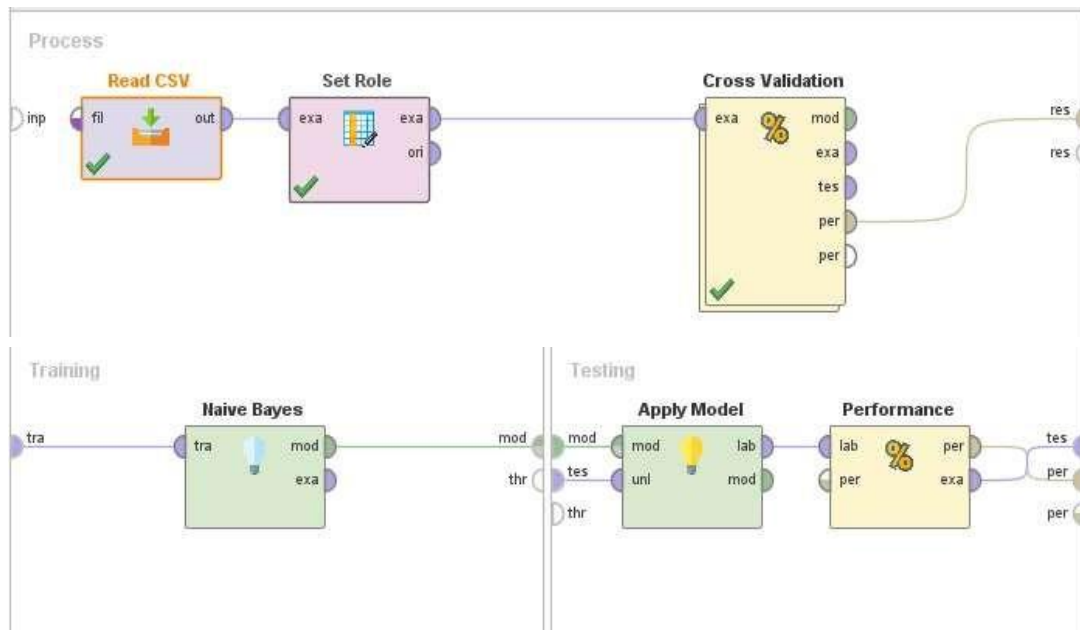

Figure 3 Application of Anemia Prediction Data Using the
Naïve Bayes Algorithm in RapidMiner

in the Rapid Miner application predictions by conducting experiments using cross validation which can directly divide the data into training data and testing data because the data used is supervised and the algorithm used is naïve Bayes. We can see the experimental results in table 1 below:

Tabel1. Confusion Matrix Anemia Prediction Results Using the
Naïve Bayes Algorithm in RapidMiner

	true 0	true 1	class precision
pred. 0	747	33	95.77%
pred. 1	54	587	91.58%
class recall	93.26%	94.68%	

Table 1 is a confusion matrix which shows the experimental results. In the confusion matrix we can see the results of accuracy, class precision and class recall. The resulting accuracy was 93.88% class recall 94.68% and class precision 91.70%.

Application of Particle Swarm Optimization (PSO) and Naïve Bayes Algorithm
The feature selection optimization method that can be carried out uses the PSO method with the Naïve Bayes algorithm. The application of methods and algorithms in Rapid Miner is shown in the image below:
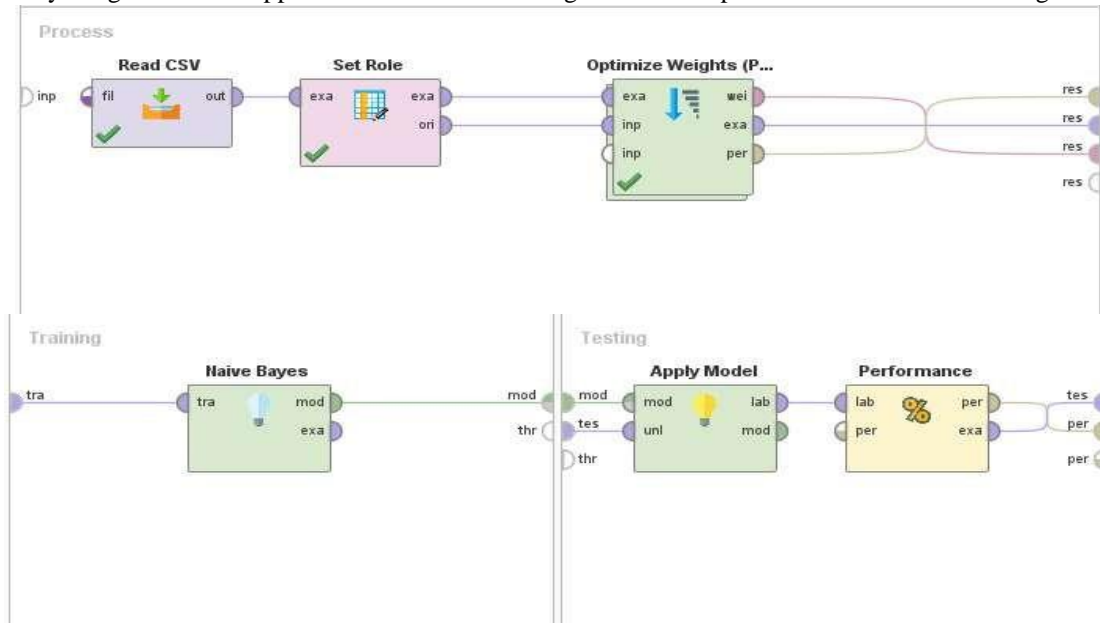

Figure 4 Application of Anemia Prediction Data Using the PSO Method and
Naïve Bayes Algorithm in Rapid Miner

In Figure 4 the data that has been prepared is applied to the Rapid Miner application in predicting anemia by conducting experiments using cross validation and using Particle Swarm Optimization (PSO) feature

selection optimization techniques and the algorithm used remains the same, namely Naïve Bayes. We can see the experimental results in table 2 below:

Tabel 2. Confusion Matrix Prediction Results for Anemia Using the Naïve Bayes and PSO Algorithm Methods

	true 0	true 1	class precision
pred. 0	746	30	96.13%
pred. 1	55	590	91.47%
class recall	93.13%	95.16%	

Evaluation with the Confusion Matrix Model of the Naïve Bayes Algorithm and PSO
Confusion matrix is a performance evaluation method used to measure the extent to which a machine learning model can classify data correctly. Confusion matrices are generally used in classification tasks where we have known labels and want to measure the model's ability to predict those labels. A confusion matrix is a two-dimensional table that shows the amount of data that is classified correctly and incorrectly by the model.

Table 3. Confusion Matrix Algorithm Naïve Bayes and PSO Algorithms

	true No	true Yes	class precision
pred. No	746	30	96.13%
pred. Yes	55	590	91.47%
class recall	93.13%	95.16%	

Based on Table 1 above, Prediction for Class "No":

Class Prediction "No":
True Negative (TN): 746 False Positive (FP): 30

So, the total number of predictions of the "No" class is 776 (746 + 30). This means that the model predicted the class "No" 776 times, and of these predictions, 746 were correct (True Negative) and 30 were incorrect (False Positive).
Class Prediction "Yes":
Class Prediction "Yes":
False Negative (FN): 55 True Positive (TP): 590
So, the total number of "Yes" class predictions is 645 (55 + 590). This means that the model predicted the "Yes" class 645 times, and of these predictions, 55 were wrong (False Negative) and 590 were correct (True Positive).

Recall (Sensitivity) for class "No" from the given Confusion Matrix Table:

$$\text{Recall (No)} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN) + False Negative (FN)}}$$

$$\text{Recall (No)} = \frac{746}{746 + 55}$$

6

$$\text{Recall (No)} = \frac{746}{801}$$

$$\text{Recall (No)} \approx 0.9313$$

Recall (No) ≈ 0.9313
So, Recall (Sensitivity) for class "No" is 93.13%.
Recall (Sensitivity) for class "Yes" from the given Confusion Matrix Table:

$$\text{Recall (Yes)} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$\text{Recall (Yes)} = \frac{590}{590 + 30}$$

$$\text{Recall (Yes)} = \frac{590}{620}$$

Recall (Yes) ≈ 0.9516
So, Recall (Sensitivity) for the "Yes" class is 95.16%.

CONCLUSION

In developing an anemia prediction model using the Naïve Bayes Algorithm and PSO, this model can be used as a predictive tool to predict who is at risk of experiencing anemia. Based on the results of testing the Naïve Bayes and PSO algorithm models which were carried out through confusion matrix evaluation, it was proven that the tests carried out by the Naïve Bayes and PSO algorithms had a high accuracy value, namely 94.02%. In the context of this research, the confusion matrix is used as an evaluation tool to measure model performance. Metrics such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) provide a deeper understanding of the model's ability to make predictions. An accuracy of 94.02% shows that the model can properly classify individuals who are at risk of developing anemia and those who are not.

The success of this model makes a positive contribution to efforts to prevent and manage anemia. This predictive model can provide early information to stakeholders, such as medical personnel or parties involved in public health, so that preventive and intervention steps can be taken quickly.

REFERENCES

[1] A. Informatics, U. Salamah, R. Ramadhan, and A. Info, "Analisis Perbandingan Metode Dempster Shafer dan Certainty Factor dalam Mendeteksi Penyakit Anemia," vol. 06, no. 02, pp. 138–146, 2023.

[2] J. Ilmiah, U. Batanghari, Y. Indrawatiningsih, E. P. Sari, and H. Listiono, "Jiubj 1,3,4,5," vol. 21, no. 1, pp. 331–337, 2021, doi: 10.33087/jiubj.v21i1.1116.

[3] V. Silalahi, E. Aritonang, and T. Ashar, "Potensi Pendidikan Gizi Dalam Meningkatkan Asupan Gizi Pada Remaja Putri Yang Anemia Di Kota Medan," *J. Kesehat. Masy.*, vol. 11, no. 2, p. 295, 2016, doi: 10.15294/kemas.v11i2.4113.

[4] N. H. Nurrahman, D. S. Anugrah, A. P. Adelita, and A. N. Sutisna, "Faktor dan Dampak Anemia pada Anak-Anak , Remaja , dan Ibu Hamil," 2020.

[5] I. Oktaviani, D. Rahmawati, and Y. N. R. Kana, "Prevalensi dan Faktor Risiko Anemia pada Anak di Negara Maju," *J. Kesehat. Masy. Indones.*, vol. 16, no. 4, p. 218, 2021, doi: 10.26714/jkmi.16.4.2021.218-226.

[6] R. A. Ningrum, A. Priyanto, U. Athiyah, F. Informatika, I. Teknologi, and T. Purwokerto, "Implementasi Algoritma Fuzzy Tsukamoto Untuk Diagnosis Penyakit Anemia (Studi Data :

Rekam Medis Pasien Ibu RSIA Bunda Arif Purwokerto) Jurnal Ilmiah Rekam Medis dan Informatika Kesehatan Jurnal Ilmiah Rekam Medis dan Informatika Kesehatan," vol. 11, no. 2, pp. 65–72, 2021.

[7] R. Rm and D. Binjai, "MENGGUNAKAN METODE TEOREMA BAYES BERBASIS WEB," vol. 03, no. 01, 2021.

[8] K. Anemia and D. Kuku, "Classification of Anemia with Digital Images of Nails and Palms using the Naive Bayes Method," vol. 18, no. 1, pp. 118–130, 2021, doi: 10.31515/telematika.v18i1.4587.

[9] Rawan and Triwidiastuti, "Pengantar Metode Penelitian modul 1," *J. Pendidik.*, vol. 2, no. 2, pp. 1–43, 2019.

[10] I. Z. Hidayati, "Penilaian Uji Validitas Instrumen Skrining Anemia Pada Siswa Madrasah Aliyah Islamic Center Baiturahman Banyuwangi," *Maj. Kesehat. Masy. Aceh*, vol. 2, no. 3, 2019, doi: 10.32672/makma.v2i3.1293.

[11] A. Mustopa, Hermanto, Anna, E. B. Pratama, A. Hendini, and D. Risdiansyah, "Analysis of user reviews for the pedulilindungi application on google play using the support vector machine and naive bayes algorithm based on particle swarm optimization," *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, vol. 2, 2020, doi: 10.1109/ICIC50835.2020.9288655.

[12] S. H. F. Hakim, I. Cholissodin, and A. W. Widodo, "Seleksi Fitur Dengan Particle Swarm Optimization Untuk Pengenalan Pola Wajah Menggunakan Naive Bayes (Studi Kasus Pada Mahasiswa Universitas Brawijaya Fakultas Ilmu Komputer Gedung A)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 10, pp. 1045–1057, 2017.

[13] R. Dahlia *et al.*, "Penerapan Data Mining Terhadap Data Covid - 19," *J. Inform.*, vol. 21, no. 1, pp. 44–52, 2021.

[14] M. R. Romadhon and F. Kurniawan, "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," *3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021*, pp. 41–44, 2021, doi: 10.1109/EIConCIT50028.2021.9431845.

[15] L. A. Andika, P. A. N. Azizah, and R. Respatiwulan, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," *Indones. J. Appl. Stat.*, vol. 2, no. 1, p. 34, 2019, doi: 10.13057/ijas.v2i1.29998.