Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 3, July 2024, pp. 613∼624
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i3.3584

Classification of Cash Direct Recipients Using the Nave Bayes with Smoothing

Muhammad Faris Al-Adni , Eko Prasetyo , Rahmawati Febrifyaning Tias
Universitas Bhayangkara, Surabaya, Indonesia

Article Info

Article history:

Received November 23, 2023
Revised June 14, 2024
Accepted June 20, 2024

Keywords:

Classification
Direct Cash Assistance
Nave Bayes
Smoothing
Prediction

ABSTRACT

Direct Cash Assistance is a social program distributed to residents meeting specific requirements. The village government determines the recipients using a conventional system through village meetings. This approach is greatly influenced by the decision-holders' subjectivity with non-transparent thinking. **This research aims** to solve the problem of classifying Direct Cash Assistance recipients by applying probability-based classification. **The research method used** is smoothed Nave Bayes, which improves Nave Bayes by adding a constant to avoid zero classification. The datasets use variables such as age, type of work, and criteria for receiving assistance. The last variable includes five nominal data, which debilitates Nave Bayes by not obtaining a posterior probability as a prediction class result. We used Direct Cash Assistance data from the Sedati sub-district, Sidoarjo district, East Java. **The results of research** with original Nave Bayes and smoothed Nave Bayes classification show that smoothed Nave Bayes has good prediction performance with an accuracy of 95.9% with a data split of 60:40. Smoothed- Nave Bayes also solves the problem of 8 data without predictive classes. The prediction results show that Smoothed Nave Bayes performs better than standard Nave Bayes. **This research contributes** to refining Nave Bayes to complement probability-based classification by adding refinement constants to avoid zero classification.

Corresponding Author:

Eko Prasetyo, +628819314737
Departmenet of Informatics, Faculty of Engineering,
Universitas Bhayangkara, Surabaya, Indonesia,
Email: eko@ubhara.ac.id.

How to Cite:
Muhammad Faris Al-Adni, "Classification of Cash Direct Recipients Using the Nave Bayes with Smoothing", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 3, pp. xxx-xxx, July, 2024.

1. INTRODUCTION

Direct Cash Assistance (DCA) is also a program organized by the Indonesian government in the context of social security after the COVID-19 pandemic. Determining recipients of DCA is carried out by the village government using a conventional system, namely village deliberations involving all village residents and manual selection. This approach is relatively simple, but the influence of decision-maker subjectivity is quite high. A computer-based approach to determining the eligibility of DCA recipients using data mining, such as classification using Nave Bayes, is a solution to help determine DCA recipients objectively and fairly. The DCA problem has also been paid attention to by many researchers, such as [1] trying to digitize the non-cash DCA distribution system using bank accounts, mobile money, and prepaid cards. This solution is claimed to be effective in solving the problem of how, when, and where to distribute it. The same research by [2] also examines the digitization of the cash assistance and voucher system to solve the problems of transparency, scalability, speed, and inclusion for beneficiaries. Therefore, determining DCA using computerized data-based methods, such as data mining, is essential research, as [3] proves the importance of digitizing DCA distribution. Conversely, [4] also proves that DCA provides socio-economic benefits that should not be ignored.

Data mining collects large amounts of data and then extracts the data into valuable information. Research by [5] proves that the results of data mining processing can be used to make decisions in the future. Classification is one of the data mining jobs that discovers models or functions that describe and distinguish data classes. Classification allows us to use a model to predict the class of an object from unknown class labels. The classification model should differentiate each class from the others and present an organized diagram of the dataset. Research by [6] also, classification makes it easier to find models, letters, and concepts from the dataset for each class to help us make decisions. Naive Bayes (NB) is a statistical classifier that predicts a class's data membership probability. The NB algorithm is based on conditional probability [5]. There are several studies related to DCA that have been carried out previously, such as by [7] classifying loan eligibility on 1,200 data with an accuracy of 99, 8%, [5] classifying social assistance recipients for the Family Hope program with an accuracy of 84%, [8] classifying social assistance recipients based on income, family members, residence and vehicle, [9] classifying Hope Family Assistance Program recipients using decision trees and NB algorithms. Other research with NB achieved high accuracy, such as research by [10] conducted classification of determining recipients of basic food assistance with seven variables on 135 training data and 40 testing data, resulting in an accuracy of 86%. Research by [11] conducted predictions of agricultural production reached an accuracy of 99.64%, research by [6] conducted classification of student graduation to support decisions regarding graduation, research by [12] conducted classification of recipients of the Family Hope program, research by [13] conducted classification of students' academic potential based on character, academic activities, socio-economic status, and distance of residence 75% accuracy, and research by [14] conducted classification of eligibility for recipients of health insurance contribution assistance with 96.83% accuracy, research by [15] conducted classification of social assistance recipients based on business and employment with an accuracy of 95%, [16] conducted classification of family assistance based on the criteria of disability, children's school, toddlers, salary, and housing status with an accuracy of up to 96%, [17] classification of new students of junior high school with 87% precision, [18] expert system for early detection of COVID-19 with 96% accuracy and classification of building fire safety with 93% accuracy [19]. In similar cases, other research also proved that NB provided some unsatisfied performance, such as the research [20] classifying food aid recipients with 58% accuracy and classification of social aid recipients with 62% accuracy. In other one, NB was also combined with other methods to solve problems, such as [21] combined with the Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalance class, [22] classified public opinion with Term Frequency-Inverse Document Frequency (TF-IDF) and Count Bag of Word (BOW) variables, and [23] classified public opinion about COVID-19 with TF-IDF. These results depend on the problem, variables, and data used. Research carried out previously generally used variables with common values where numerical variables and categories were regular so that NB weakening did not occur.

From the previous explanation, the data shows that previous research used criteria that focused on houses only, such as ownership status, size, and floor type [12]. There are gaps that previous research has not addressed; we see these criteria are irrelevant when citizens have jobs with salaries guaranteed to be very high. Therefore, other critical criteria that should be considered include type of work and age. Therefore, we use several criteria, including type of work, age, and Criteria for Receiving Assistance (CRA). These criteria include entrance into Integrated Social Welfare Data (ISWD), not yet receiving a Social Safety Net (SSN), loss of livelihood, chronic disease, and poor family. This research aims to implement the smoothed NB to solve the DCA recipient's classification system using five variables with nominal data. The data have gone through pre-processing NB and may not get a posterior probability as a prediction class result. The system cannot make predictions as it should. Therefore, this research contributes to experimenting by adding smoothing to NB. Refinement is carried out by adding a constant of 0.001 to anticipate a class probability with a value of 0 (undefined prediction). The experiments in this study added smoothing to resolve predictions that could not be determined. The 8 data cannot be previously predicted due to NB weakening with a value of 0 in each class and cannot be predicted. Adding a smoothing constant of 0.001 (smoothing) can determine the prediction results for these 8 data.

The difference between this research and the previous one is that we involve variables for determining DCA recipients that previous researchers have not used, namely type of work, age, ISWD, not yet receiving an SSN, Loss of Livelihood, Chronic Disease, and Poor Families. The use of variables that are relevant to government regulations is an important issue that must be considered. To guarantee reliable results, we used 2600 population data from 16 villages. The data in this research are recipients and non-recipients of direct cash assistance from 16 villages in the Sedati District, Sidoarjo Regency, including Segoro Tambak, Banjar Kemuning, Gisik Cemandi, Tambak Cemandi, Cemandi, Kalanganyar, Buncitan, Pepe, Pulungan, Kwangsan, Betro, Sedati Gede, Sedati Agung, Semampir, Pabean, Pranti, with 1600 and 1000 data on recipients and non-recipients of direct cash assistance respectively. DCA acceptance classification performance before and after the addition of refinement reached an accuracy of 95.1% and 95.9%, respectively.

This research aims to develop a system for classifying DCA recipients using variables that comply with government regulations. Therefore, the classification results are not influenced by the subjectivity of the policyholder. We use variables that contribute to determining DCA recipients: type of work, age, and CRA. Experiments were carried out to improve NB. Adding a smoothing constant of 0.001 to NB can improve zero classification where the classifier cannot differentiate the class of predicted results.

2. RESEARCH METHOD

This research experiment is to determine the feasibility of receiving DCA using modified NB. This research was carried out by collecting data, pre-processing, separating, building experimental models with NB, and classifying test data. The final stage is evaluating the performance of the method. Our experiment uses data from 16 villages in Sedati District, Sidoarjo Regency, in 2022.

2.1. Research Flowchart

In this research, we applied three flowcharts: the research step, the method step, and the application flowchart, as presented in Figure 1, Figure 2, and Figure 1, respectively. Research steps described the steps used to complete the research. The method step explained the method we used and enhanced, while the application flowchart explained how the application was developed as an interface when running the system.

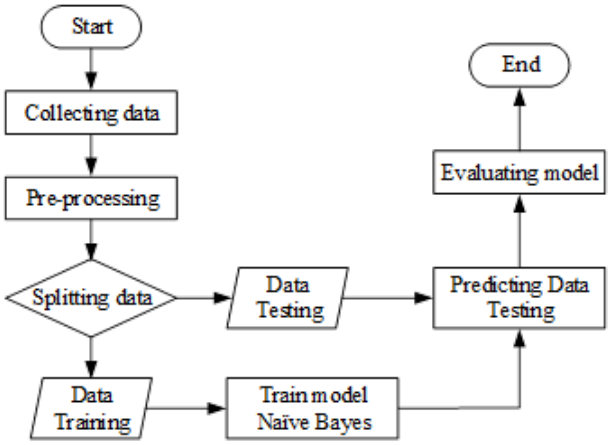

Figure 1. The Research Flowchart

In the flowchart of research steps, as presented in Figure 1, the first step is collecting data, where data is collected through visits to each village in the Sedati sub-district to obtain it directly. Next is preprocessing to binarize the 5 DCA criteria variables: included in Integrated Social Welfare Data (ISWD) has not yet received Social Safety Net (SSN), ISWD has not yet been recorded, loss of livelihood, chronic disease for low-income families. This variable takes the value 1 if it exists and 0 if it does not, then calculate age from date of birth. Next is the data split process, where the data division uses a 60% and 40% ratio. 60% of the data is training data, and 40% is test data. Next is the Nave Bayes training process. This process is a system calculation using the NB method. Next is prediction; this prediction process determines whether the input data is included in the appropriate or inappropriate class category. The final process is evaluating system performance assessment data with test data.

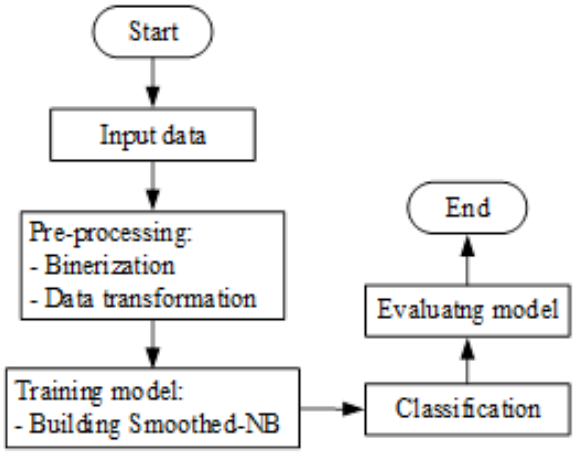

Figure 2. Flowchart Nave Bayes

As presented in Figure 2, the method flowchart has input data, binarizing data according to 7 variables: age, type of work, and DCA criteria. Then, the data is transformed according to system input, training the model using the NB method with smoothing. After that, the input data will be classified into eligibility or ineligibility classes. The last is the evaluation of the system with test data.

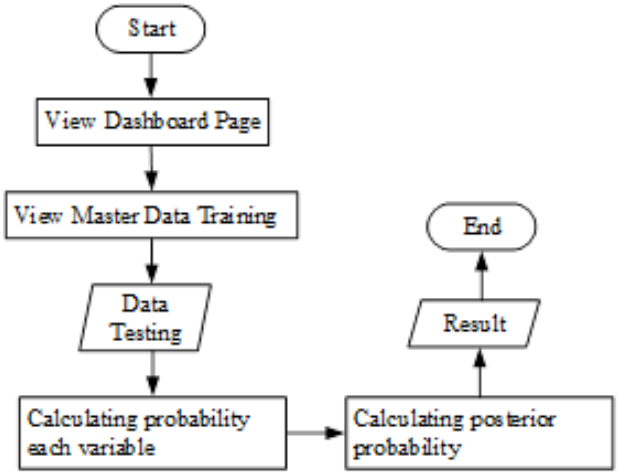

Figure 3. Application Flowchart

As presented in Figure 3, the application flowchart starts with a dashboard display with the main menu, which shows the amount of training data, the amount of test data, and the number of users on the system. Next is the display of the training data. This page manages training data by creating, reading, updating, and deleting. Next is the input of test data for Nave Bayes calculations. As a result, the system presents the probability of each test data class. After that, the class conclusion appears to the users.

2.2. Dataset

This research uses data from 16 villages in the Sedati sub-district, Sidoarjo district, in 2022. Data was obtained from a collection of population data in each village. The dataset consists of seven variables: type of work, age, and Criteria for Receiving Assistance (CRA). KPB consists of entrance into Integrated Social Welfare Data (ISWD), not yet receiving a Social Safety Net (JPS), Loss of Livelihood, Chronic Disease, and Poor Family. The sample of data is shown in Table 1.

Table 1. Data on Pre-Prosperous Families Beneficiaries of 2022 DCA Village Fund Banjarkemuning Village, Sedati District, Sidoarjo Regency

No	District	Village	Year	Nik	Recipient name	Date of birth	Address	Rt / Rw	Job	DCA Criteria
1	Sedati	Banjar kemu-ning	2022	351517*****	Abd. Kholiq	10/10/1978	Jl Tombro	001/001	Buruh Tambak	2
2	Sedati	Banjar kemu-ning	2022	351517*****	Afif Rosyidi	05/04/1976	Jl. Tombro	001/001	Buruh Bangu-nan	3
3	Sedati	Banjar kemu-ning	2022	351517*****	Aminatin	03/02/1945	Jl Tombro	001/001	Belum/Tidak Bekerja	1
4	Sedati	Banjar kemu-ning	2022	351517*****	Anita	27/12/1984	Jl. Wader	002/001	Pembantu Rumah Tangga	2
5	Sedati	Banjar kemu-ning	2022	351517*****	Dewi Rosyidah	30/06/1973	Jl Tombro	001/001	Buruh Pabrik	3
6	Sedati	Banjar kemu-ning	2022	351517*****	Fatin Komamah	19/04/1964	Jl Tombro	001/001	Pedagang Makanan	1
7	Sedati	Banjar kemu-ning	2022	351517*****	H Jamilah	09/10/1948	Jl Tombro	001/001	Belum/Tidak Bekerja	4
8	Sedati	Banjar kemu-ning	2022	351517*****	Khusnul Khotimah	04/02/1972	Banjar Kemuning	001/001	Belum/Tidak Bekerja	2
...						...				
88	Sedati	Banjar kemu-ning	2022	351517*****	Sumo Rejo	18/06/1955	Jl Tawes	008/004	Buruh Tambak	4
89	Sedati	Banjar kemu-ning	2022	351517*****	Zuhriyah	10/05/1971	Jl. Tawes	008/004	Pembantu Rumah Tangga	2

2.3. Pre-processing

After obtaining the actual data, pre-processing is carried out using binarization to obtain the age sourced from the date of birth. The DCA Criteria variables initially with values 1-5 were converted into binary variables with values 1 and 0, as in Table 2. These two variables are used in classification calculations using the NB method. In Table 2, the DCA criteria variable contains acceptance criteria, which are divided into 5 points as follows: I for included in ISWD who have not yet received the SSN, II for Not Having ISWD Recorded, III for losing livelihood, IV for Chronic Diseases, V for Poor Families.

Table 2. Dataset After Pre-Processing

Name	Address	Type Of Work	Age	DCA Criteria					Class
				I	II	III	IV	V	
Abd. Kholiq	Banjar Kemuning	Buruh Tambak	44	1	0	0	0	0	LAYAK
Afif Rosyidi	Banjar Kemuning	Buruh Bangunan	47	1	0	0	0	0	LAYAK
Aminatin	Banjar Kemuning	Belum/Tidak Bekerja	78	0	1	0	0	0	LAYAK
Anita	Banjar Kemuning	Pembantu Rumah Tangga	38	0	0	0	1	0	LAYAK
Dewi Rosyidah	Banjar Kemuning	Buruh Pabrik	50	0	0	1	0	0	LAYAK
Fatin Komamah	Banjar Kemuning	Pedagang Makanan	59	1	0	0	0	0	LAYAK
H Jamilah	Banjar Kemuning	Belum/Tidak Bekerja	74	0	0	1	0	0	LAYAK
Khusnul Khotimah	Banjar Kemuning	Belum/Tidak Bekerja	51	0	0	0	1	0	LAYAK
...									
Sumo Rejo	Banjar Kemuning	Buruh Tambak	63	1	0	1	0	0	LAYAK
Zuhriyah	Banjar Kemuning	Pembantu Rumah Tangga	52	0	1	0	0	0	LAYAK

2.4. Nave Bayes

NB is a statistical-based classification method that predicts data based on initial probabilities in the training data. Each variable involved is considered not to correlate (independence). Consequently, there is no visible dominance of one variable. The output obtained by this method is the final probability value, which combines the initial probability of the class with the probability of the influence of each variable on the class. The results achieved are probability values in the range 0 to 1. The Equation (1).

$$P(Xi = xj \mid C = ci) = \frac{1}{\sqrt{2\pi\sigma ij}} exp \frac{(xi - \mu ij)^2}{2\sigma^2 ij} \tag{1}$$

Where :
P : Probability
Xi : Variable i-th
xj : Attribute value i-th
C : Class
Ci : Sub class required
μ : Average of all attribute
σ : Standard deviation

$$P(ci) = \left(\frac{si}{s}\right) \tag{2}$$

Where :
Si : Number of training data from category Ci
s : Number of training data training

2.5. Smoothed Nave Bayes

NB on seven variables in this study weakens it by giving zero values to all classes. The results of this prediction cannot be determined by whether the prediction results are eligible or ineligible. To solve this problem, the author conducted experiments by adding smoothing to solve prediction problems whose values could not be determined. Refinement adds a constant (0.001) to each variable's prior probability to anticipate the value 0. Following are the smoothed NB. Formulas (1) and (2) are the original formulas for NB classification. At the same time, Equations (3) and (4) are modified formulas with the addition of a constant of 0.001, which, in this research, the researchers call smoothed Nave Bayes to avoid unidentified predictions. We use a small value, namely 0.001, to avoid 0 in the posterior probability by not significantly influencing the final probability.

$$P(Xi = xj \mid C = ci) = \left(\frac{1}{\sqrt{2\pi\sigma ij}} exp \frac{(xi - \mu ij)^2}{2\sigma^2 ij}\right) + 0.001 \tag{3}$$

$$P(ci)\left(\frac{si}{s}\right) + 0.001 \tag{4}$$

3. RESULT AND ANALYSIS

This research experimented with improving NB to complete the classification of DCA from the village government using seven criteria, including the type of work, age, and DCA criteria. An important finding achieved in the research is that NB requires smoothing to handle zero-class prediction. We tested using data on 1560 family heads from 16 villages in the Sedati sub-district, Sidoarjo district. Testing is carried out using the hold-out method, as explained in the following subsection.

3.1. Result

We experimented with hold-out to determine what portion of the data split provides high classification performance. Holt-out allows us to determine the percentage of data that becomes training and test data. We use the percentage that offers the best performance as the percentage data in this article. We use the following percentages: 50:50, 60:40, 70:30, and 80:20. These percentages are training and test data; for example, 60:40 means 60% of the data is training data, and 40% is test data, as presented in Table 3.

Table 3. Splitting data experiment

Splitting data	Accuracy
50:50 :00	94.85
60:40:00	95.90
70:30:00	94.87
80:20:00	95.06

The experimental results presented in Table 3 show that the highest accuracy was achieved by the 60:40 splitting option with an accuracy of 95.9%. So, the results we present next are the experimental results achieved by splitting the data. The other results are also reasonably good, with a slight difference. However, we should also consider the balance of training and testing data. Too much training data results in overfitting, while insufficient training data results in underfitting. Our best option is to use 60:40, which is the best accuracy.

In this research, actual data goes through a pre-processing stage for classification calculations in the system. In the testing session, the researcher split the data using the hold-out method using a ratio of 60:40, with 60% as training data and 40% as test data. The following displays the training data and test data presented in table format. Table 4 is the training data used in implementing the DCA acceptance classification application using the Nave Bayes method with a total of 1560 training data, divided into 960 eligible recipient class data and 600 ineligible recipient class data of DCA.

Table 4. Training Data

No.	Name	Address	Type of works	Age	DCA Criteria					Class
					I	II	III	IV	V	
1	Ahmad Suharsono	Segoro Tambak	Belum/Tidak Bekerja	51	1	0	0	0	0	Layak
2	Sumiati	Segoro Tambak	Belum/Tidak Bekerja	53	1	0	0	0	0	Layak
3	Fatimah	Segoro Tambak	Belum/Tidak Bekerja	61	0	1	0	0	0	Layak
4	Moch Yasak	Segoro Tambak	Belum/Tidak Bekerja	54	0	0	0	1	0	Layak
5	Bunaji	Segoro Tambak	Swasta	49	0	1	0	0	0	Layak
6	Mujiono	Segoro Tambak	Belum/Tidak Bekerja	26	1	0	0	0	0	Layak
7	Miah	Segoro Tambak	Belum/Tidak Bekerja	54	0	0	1	0	0	Layak
8	Afifah	Segoro Tambak	Buruh Pabrik	50	0	0	1	0	0	Layak
...										
1559	M Yusuf	Banjar Kemuning	Swasta	39	0	0	0	0	0	Tidak Layak
1560	Suwadi	Banjar Kemuning	Swasta	50	0	0	0	0	0	Tidak Layak

Table 5. Test Data

No	Name	Address	Type of works	Age	DCA Criteria					Class
					I	II	III	IV	V	
1	Haris Firmansyah	Pulungan	Buruh Pabrik	34	0	1	0	0	0	Layak
2	Sugi Hartono	Pulungan	Belum/Tidak Bekerja	49	0	0	0	1	0	Layak
3	Sunarsih	Pulungan	Belum/Tidak Bekerja	61	0	0	0	0	1	Layak
4	Agnes Suharti	Pulungan	Pedagang Makanan	66	0	0	1	0	0	Layak
5	Ida Ambarwati	Pulungan	Pedagang Barang	58	0	0	1	0	0	Layak
6	F. Charlyn Pranata H.S	Pulungan	Belum/Tidak Bekerja	52	0	0	0	1	0	Layak
7	Kartini	Pulungan	Buruh Tani	59	0	1	0	0	0	Layak
8	Efendi	Pulungan	Belum/Tidak Bekerja	62	0	1	0	0	0	Layak
9	Susiani	Pulungan	Buruh Pabrik	47	0	1	0	0	0	Layak
10	Dian Isnaniah	Pulungan	Pembantu Rumah Tangga	48	0	0	1	0	0	Layak
...										
1039	Aminin	Banjar Kemuning	Perangkat Desa	50	0	0	0	0	0	Tidak Layak
1040	H. Hasanudin	Banjar Kemuning	Pemilik Tambak	68	0	0	0	0	0	Tidak Layak

Table 5 is the data used to test the DCA acceptance classification system. The total training data is 1040, with 640 data for the class eligible to receive DCA and 400 data for the class not eligible to receive DCA. After preprocessing the data, it was tested by calculating the NB classification using the application the researchers created with the display in Figure 4.

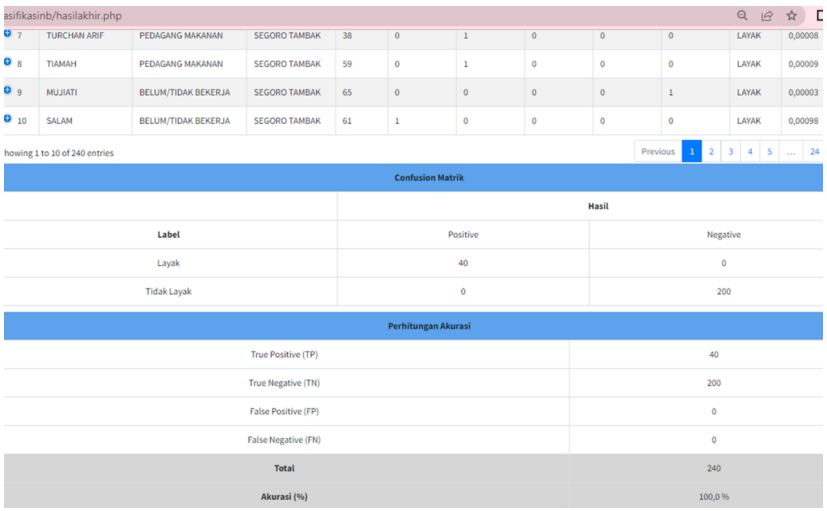

Figure 4. DCA Classification Application Dashboard Display

The data presented in Table 6 is data from the NB classification using all data in the Sedati sub-district, where the prediction results cannot be known because the final probability in each class is zero. These results are predictions of data using NB without smoothing. The final probability of each class is zero. As a result, the prediction class cannot be determined from the data whether it is included in the eligible and ineligible recipient class. Based on the results presented in Table 7, it can be seen that the number of data predicted correctly was 989, and the data mispredicted was 51. By knowing the amount of data that was predicted correctly, the level of accuracy can be determined. Thus, the accuracy is obtained as follows.

Table 6. Prediction Results with Original NB

No	Name	Address	Job	Age	DCA Criteria					Classification Nave Bayes				
					1	2	3	4	5	Eligible Prob.	Ineligible Prob.	Class Result	Actual Class	Suitability
86	Abdul Malik	Betro	Guru Agama	61	TIDAK	IYA	TIDAK	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
587	Nasir Junaidi	Pranti	Peadagang Barang	60	TIDAK	IYA	IYA	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
592	Widiarto Nugroho	Pranti	Peadagang Barang	46	TIDAK	TIDAK	IYA	IYA	TIDAK	0	0	LAYAK	LAYAK	SESUAI
613	Yuyun Yuliani	Pranti	Peadagang Barang	54	TIDAK	TIDAK	IYA	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
619	Mariani	Pranti	Peadagang Barang	45	TIDAK	IYA	IYA	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
629	Mudriah	Pranti	Peadagang Barang	72	TIDAK	IYA	IYA	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
632	Nur Aini	Pranti	Peadagang Barang	46	TIDAK	IYA	IYA	TIDAK	TIDAK	0	0	LAYAK	LAYAK	SESUAI
1032	Agus Mustofa	Banjar Kemu-ning	Karyawan Bumn	43	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	0	0	TIDAK LAYAK	TIDAK LAYAK	SESUAI

Table 7. Confusion Matrix Of Prediction Results With Original NB

Number of Data (1040)		Prediction	
		Eligible	Ineligible
Actual class	Eligible	633	7
	Ineligible	44	356

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% = \frac{633 + 356}{633 + 356 + 44 + 7} \times 100\% = \frac{989}{1040} \times 100\% = \textbf{95,1}\%$$

Table 8. Prediction Results Using NB With Smoothing

No	Name	Address	Job	Age	DCA Criteria					Classification Nave Bayes				
					1	2	3	4	5	Eligible Prob.	Ineligible Prob.	Class Result	Actual Class	Suitability
86	Abdul Malik	Betro	Guru Agama	61	TIDAK	IYA	TIDAK	TIDAK	TIDAK	0,0000065	0,0000002	LAYAK	LAYAK	SESUAI
587	Nasir Junaidi	Pranti	Peadagang Barang	60	TIDAK	IYA	IYA	TIDAK	TIDAK	0,0000018	0	LAYAK	LAYAK	SESUAI
592	Widiarto Nugroho	Pranti	Peadagang Barang	46	TIDAK	TIDAK	IYA	IYA	TIDAK	0,0000003	0	LAYAK	LAYAK	SESUAI
613	Yuyun Yuliani	Pranti	Peadagang Barang	54	TIDAK	TIDAK	IYA	TIDAK	TIDAK	0,0000057	0	LAYAK	LAYAK	SESUAI
619	Mariani	Pranti	Peadagang Barang	45	TIDAK	IYA	IYA	TIDAK	TIDAK	0,0000015	0	LAYAK	LAYAK	SESUAI
629	Mudriah	Pranti	Peadagang Barang	72	TIDAK	IYA	IYA	TIDAK	TIDAK	0,0000007	0	LAYAK	LAYAK	SESUAI
632	Nur Aini	Pranti	Peadagang Barang	46	TIDAK	IYA	IYA	TIDAK	TIDAK	0,0000016	0	LAYAK	LAYAK	SESUAI
1032	Agus Mustofa	Banjar Kemu-ning	Karyawan Bumn	43	TIDAK	TIDAK	TIDAK	TIDAK	TIDAK	0,0000136	0,0000457	TIDAK LAYAK	TIDAK LAYAK	SESUAI

Table 8 shows the test results after adding the NB method with smoothing, where each final probability has a value. Therefore, the interesting finding in this research is that we can find out the classification results for each piece of data. If the final value of the eligible probability is greater than the final probability value of ineligible, then the classification result is included in the eligible class, and vice versa. Based on Table 9, it can be seen that the number of data from the correct status is 997 data, and the data that was mispredicted is 43. By looking at the number of data predicted correctly, we can determine the accuracy level in the NB classification with the following refinement.

Table 9. Confusion Matrix Prediction Results with Smoothed NB

Number of Data (1040)		Prediction	
		Eligible	Ineligible
Actual class	Eligible	640	0
	Ineligible	43	357

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% = \frac{640 + 357}{640 + 357 + 43 + 0} \times 100\% = \frac{997}{1040} \times 100\% = \textbf{95,9}\%$$

The results were achieved from data that included variables for all types of DCA acceptance criteria, starting from the criteria for entering ISWD, not having received SSN, not having ISWD recorded, loss of livelihood, chronic illness, low-income family, and also all types of work and age, the accuracy of 95.9%. This accuracy shows high system performance based on the training data. The results above show that smoothed NB can classify all data with full nominal variables. A full nominal value causes zero prediction where the classifier cannot confirm the class assigned to the data. Our findings are a refinement of NB that can overcome this problem. The performance of smoothed NB is similar to the method used in previous research, with the advantage of completing zero classification.

3.2. Analysis

This research implements a refined NB to complete the classification system for DCA recipients using several variables based on government regulations. For this purpose, we use five variables, a mixture of numerical and nominal data. Preprocessing is done on nominal data by binarizing it into several binary digits. This step aims to ensure that the smoothed NB can use these variables. The NB standard's weakness is zero prediction due to zero probability values. Whatever the probability of another variable being non-zero, the prediction result becomes zero when one zero variable exists. This problem occurs in the Cash Transfer Assistance dataset, where there is data where the value of certain variables is zero. Classification with NB provides high classification performance, up to 95.1%. However, these results are still disturbed by the zero classification. We achieve zero classification by adding a smoothing constant with a small value. Although this smoothing value is small, it can prevent zero classification. The constant 0.001 is used to smooth the probability of each variable to avoid zero traps. This addition also increases the probability value but linearly for all variables. Therefore, adding a smoothing does not deviate from the direction of the classification.

NB smoothing performance has been proven and explained in the previous subsection, where there was an increase in accuracy of 0.8% from 95.1% to 95.9%. The improvement seems small but resolved the zero classification problem on 43 out of 1040 data. This performance is still better than similar studies, with standard NB achieving accuracy of 62% [20]. This finding is very reasonable because previous research has not used criteria for aid recipients. Thus, NB smoothing satisfactorily classifies data with a mixture of numerical and categorical variables.

4. CONCLUSION

Based on research on the classification of direct aid recipients using a DCA dataset with 2600 data from all villages in the Sedati sub-district with 7 variables, the NB method with smoothing classifies Eligible and Ineligible with satisfactory results. The dataset is divided into 2 groups of data, namely training data and test data, with a ratio of 60%:40% (1560 and 1040 data). System testing results with NB with smoothing achieved an accuracy of 95.9%. This experimental and analysis result concludes that Smoothed Nave Bayes can satisfactorily classify data using numeric and categorical variables.

Suggestions for further research are ongoing development maintenance and application monitoring by the responsible party. Developing a more straightforward and responsive application interface must also make users more comfortable using the system.

5. ACKNOWLEDGEMENTS

The author thanks the Sedati sub-district government, Sidoarjo, for data support in this research.

6. DECLARATIONS

AUTHOR CONTIBUTION

Muhammad Faris Al-Adni: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Original Draft
Eko Prasetyo: Formal analysis, Validation, Resources, Writing - Review & Editing, Visualization, Supervision
Rahmawati Febrifyaning Tias: Validation, Writing - Review & Editing, Supervision.

FUNDING STATEMENT

This research and scientific articles use independent funding from researchers and Faculty.

COMPETING INTEREST

This research uses original data from several villages in the Sedati sub-district government, Sidoarjo, East Java. There was no funding during data collection and curation. Funding for data processing uses independent funds from researchers and the Faculty of Engineering, Universitas Bhayangkara Surabaya. All authors contributed to the research and writing of the article following the authority and approval for submitting the manuscript.

REFERENCES

[1] A. A. Abdelgawad, A. Khan, and H. Baharmand, "Exploring gaps in using digital delivery mechanisms for cash-based assistance in refugee crises," *International Journal of Disaster Risk Reduction*, vol. 96, no. 1, pp. 1–27, Oct. 2023, https://doi.org/10.1016/j.ijdrr.2023.103907.

[2] A. Maghsoudi, R. Harpring, W. D. Piotrowicz, and D. Kedziora, "Digital technologies for cash and voucher assistance in disasters: A cross-case analysis of benefits and risks," *International Journal of Disaster Risk Reduction*, vol. 96, no. 1, pp. 1–16, Oct. 2023, https://doi.org/10.1016/j.ijdrr.2023.103827.

[3] E. Juntunen, C. Kalla, A. Widera, and B. Hellingrath, "Digitalization potentials and limitations of cash-based assistance," *International Journal of Disaster Risk Reduction*, vol. 97, no. 1, pp. 1–12, Oct. 2023, https://doi.org/10.1016/j.ijdrr.2023.104005.

[4] A. Agrawal, N. Kaur, C. Shakya, and A. Norton, "Social assistance programs and climate resilience: reducing vulnerability through cash transfers," *Current Opinion in Environmental Sustainability*, vol. 44, no. 1, pp. 113–123, Jun. 2020, https://doi.org/10.1016/j.cosust.2020.09.013.

[5] N. Alfiah, "Klasifikasi Penerima Bantuan Sosial Program Keluarga Harapan Menggunakan Metode Naive Bayes," *Respati*, vol. 16, no. 1, pp. 32–40, Mar. 2021, https://doi.org/10.35842/jtir.v16i1.386.

[6] A. Andika, S. Syarli, and C. R. Sari, "Data Mining Klasifikasi Kelulusan Mahasiswa Menggunakan Metode Nave Bayes," *Journal Peqguruang: Conference Series*, vol. 4, no. 1, pp. 423–428, May 2022, https://doi.org/10.35329/jp.v4i1.2358.

[7] N. K. A. Suarpurningsih, N. W. Utami, and N. M. Estiyanti, "Klasifikasi Penentuan Kelayakan Pemberian Kredit Menggunakan Metode Naive Bayes Classifier (Kasus: Koperasi Simpan Pinjam Artha Segara)," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 6, no. 1, pp. 391–404, Mar. 2022, https://doi.org/10.30645/j-sakti.v6i1.454.

[8] A. Junaidi, Y. Yunita, S. Agustyani, C. I. Agustyaningrum, and Y. T. Arifin, "Klasifikasi Penerima Bantuan Sosial Menggunakan Algoritma C 4.5," *Jurnal Teknik Komputer*, vol. 9, no. 1, pp. 77–82, Jan. 2023, https://doi.org/10.31294/jtk.v9i1.14378.

[9] E. Fitriani, "Perbandingan Algoritma C4.5 dan Nave Bayes untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," *Sistemasi: Jurnal Sistem Informasi*, vol. 9, no. 1, pp. 103–115, Jan. 2020, https://doi.org/10.32520/stmsi.v9i1.596.

[10] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Nave Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal Riset Komputer)*, vol. 8, no. 6, pp. 219–225, Dec. 2021, https://doi.org/10.30865/jurikom.v8i6.3655.

[11] D. N. Aini, B. Oktavianti, M. J. Husain, D. A. Sabillah, S. T. Rizaldi, and M. Mustakim, "Seleksi Fitur untuk Prediksi Hasil Produksi Agrikultur pada Algoritma K-Nearest Neighbor (KNN)," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, pp. 140–145, Sep. 2022, https://doi.org/10.30865/json.v4i1.4813.

[12] A. A. A. Arifin, W. Handoko, and Z. Efendi, "Implementasi Metode Naive Bayes Untuk Klasifikasi Penerima Program Keluarga Harapan," *J-Com (Journal of Computer)*, vol. 2, no. 1, pp. 21–26, Mar. 2022, https://doi.org/10.33330/j-com.v2i1.1577.

[13] N. K. W. Patrianingsih and I. K. A. Sugianta, "Penerapan Nave Bayes pada Potensi Akademik Siswa SD Negeri 5 Singakerta," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 8, no. 2, pp. 154–163, May 2023, https://doi.org/10.14421/jiska.2023.8.2.154-163.

[14] L. Budhy Adzy, A. Asriyanik, and A. Pambudi, "Algoritma Nave Bayes untuk Klasifikasi Kelayakan Penerima Bantuan Iuran Jaminan Kesehatan Pemerintah Daerah Kabupaten Sukabumi," *Jurnal Mnemonic*, vol. 6, no. 1, pp. 1–10, May 2023, https://doi.org/10.36040/mnemonic.v6i1.5714.

[15] D. Azlil Huriah and N. Dienwati Nuris, "Klasifikasi Penerima Bantuan Sosial UMKM Menggunakan Algoritma Nave Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 360–365, Mar. 2023, https://doi.org/10.36040/jati.v7i1.6300.

[16] A. W. S. Sriwibowo, A. Alwi, and S. Sugianti, "Application of the Nave Bayes Algorithm in Predicting Acceptance of Family Hope Program Assistance (PKH):," *JICTE (Journal of Information and Computer Technology Education)*, vol. 4, no. 1, pp. 1–8, Sep. 2020, https://doi.org/10.21070/jicte.v4i1.916.

[17] F. Santoso, Sunardi, and H. Z. Lukman, "Implementasi Data Mining dengan Metode Naive Bayes Untuk Memprediksi Penerimaan Siswa Baru di MTS NU Islamiyah Asembagus," *G-Tech: Jurnal Teknologi Terapan*, vol. 7, no. 4, pp. 1355–1366, Oct. 2023, https://doi.org/10.33379/gtech.v7i4.3086.

[18] N. Nurdin, E. Susanti, H. A.-K. Aidilof, and D. Priyanto, "Comparison of Naive Bayes and Dempster Shafer Methods in Expert System for Early Diagnosis of COVID-19," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 1, pp. 215–228, Nov. 2022, https://doi.org/10.30812/matrik.v22i1.2280.

[19] N. G. A. Dasriani, S. Hadi, and M. Syahrir, "Intelligent System for Internet of Things-Based Building Fire Safety with Naive Bayes Algorithm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 1, pp. 229–242, Nov. 2023, https://doi.org/10.30812/matrik.v23i1.3581.

[20] H. N. F. Fikrillah, S. Hudawiguna, and C. Juliane, "Klasifikasi Penerima Bansos Menggunakan Algoritma Naive Bayes," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 10, no. 1, pp. 683–695, Mar. 2023, https://doi.org/10.35957/jatisi.v10i1.3624.

[21] D. Kurniadi, F. Nuraeni, and M. Firmansyah, "Klasifikasi Masyarakat Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Nave Bayes dan SMOTE," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 309–320, Apr. 2023, https://doi.org/10.25126/jtiik.20231026453.

[22] Dedy Sugiarto, Ema Utami, and Ainul Yaqin, "Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng," *JURNAL TEKNIK INDUSTRI*, vol. 12, no. 3, pp. 272–277, Dec. 2022, https://doi.org/10.25105/jti.v12i3.15669.

[23] S. Supangat, M. Z. B. Saringat, and M. Y. F. Rochman, "Predicting Handling Covid-19 Opinion using Naive Bayes and TF-IDF for Polarity Detection," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 173–184, Mar. 2023, https://doi.org/10.30812/matrik.v22i2.2227.