ILKOM Jurnal Ilmiah Vol. 15, No. 1, April 2023, pp. 11-20
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

Determining Eligible Villages for Mobile Services using K-NN Algorithm

Anton Yudhana [a,1]; **Imam Riadi** [a,2]; **M Rosyidi Djou** [a,3*]
[a] *Universitas Ahmad Dahlan, Jl. Kapas No.9, Semaki, Kec. Umbulharjo, Kota Yogyakarta, Daerah Istimewa Yogyakarta, 55166, Indonesia*
[1] *eyudhana@ee.uad.ac.id;* [2] *imam.riadi@is.uad.ac.id;* [3] *djou2107048009@webmail.uad.ac.id**
* *corresponding author*

Article history: Received December 06, 2022; Revised December 14, 2022; Accepted January 21, 2023; Available online April 07, 2023

Abstract

To maximize and get population document services closer to the community, the Disdukcapil district of Alor provides mobile services by visiting people in remote villages which difficult-to-reach service centres in the city. Due to a large number of villages and limited time and costs, not all villages can be served, so the kNN algorithm is needed to determine which villages are eligible to be served. The criteria used in this determination are village distance, difficulty level, and document ownership (Birth Certificate, KIA, family card, and KTPel). The classes that will be determined are "Very eligible", "Eligible", and "Not eligible". By applying Z-Score normalization with the value of K=5, the classification gets 94.12% accuracy, while non-normalized only gets 88.24% accuracy. Thus, applying normalization to training data can improve the kNN algorithm's accuracy in determining eligible villages for "ball pick-up" or mobile services.

Keywords: Data Mining; kNN; Determining Villages; *Dukcapil* Mobile Service; Public Service.

Introduction

The Indonesian government, primarily through the Director General of Population and Civil Registration, continues to improve its performance to increase the ownership scope of population documents. In 2014, the total ownership of Electronic Identity Card or known as KTPel (*Kartu Tanda Pendududuk Elektronik*) only reached 80.86% of the Indonesian population. At the end of 2021, it reached 99.15%. Likewise, the ownership of Birth Certificates from 31.25% in 2014 and it reached 96.57% by the end of 2021, Child Identity Card or KIA (*Kartu Identitas Anak*) 41.98%[1].

Many innovations have been made to support the acceleration of population document ownership, including Electronic signature[2], the Use of white HVS paper A4 80gr[3], and the Independent Dukcapil Platform or ADM (*Anjungan Dukcapil Mandiri*). The electronic signature is a QR Code that replaces signatures and stamps on population documents. It makes it easier for the head of the service to sign all population documents without having to be present at the office. Signing can be done anywhere, anytime, so that service can be faster. The use of white HVS paper size A4 80gr makes it easier for people to print their documents without having to come to the office. People also can download the documents sent to their email or other media and print them. While ADM is a kind of ATM that can print documents, especially KTPel and KIA, ADM is usually placed in public access. Mentioned innovations are carried out by the Director General of Population and Civil Registration to facilitate services, to bring services closer to the community, to provide happy services, and to maximize the achievement of population documents' ownership, Another innovation that is often carried out by the Population and Civil Registration Service or Disdukcapil (*Dinas Kependudukan dan Pencatatan Sipil)* in the district is with "a ball pick-up service" or direct service to villages using mobile services[4][5]. This mobile service aims to bring services closer to the community and to increase the scope of population documents' ownership

Likewise Disdukcapil Alor Regency, East Nusa Tenggara (NTT) in which Alor Regency is one of the archipelago districts in east Nusa Tenggara and whose topography consists of mountains, valleys and several islands, to reach people who live in villages far away from the city and have difficulty accessing service centers,

http://dx.doi.org/10.33096/ilkom.v15i1.1546.11-20

Disdukcapil Kab. Alor provides "ball pick-up services" or direct services to the people in the villages. However, not all villages can be served due to the huge number of villages and also limited time and budget, so it must be selected and sorted out villages that are eligible to be served by mobile services. So far, the selection of these villages is made manually, there are no particular criteria in the assessment, so it cannot be separated from the element of subjectivity. Therefore, to overcome this issue, this study will discuss the method of selecting eligible villages for "ball pick-up services" or mobile services using one of the data mining classification methods using the kNN algorithm.

Data mining is the process of discovering patterns, models and other types of knowledge from big data[6] first appeared in the 1990s. Data mining is analogous to looking for gold grains among sand and soil. From data that has not been patterned, specific characteristics are analyzed to generate something new. Data mining is part of the steps in Knowledge Discovery at the Database (KDD)[7]. It is the process of identifying the discovery of knowledge, extracting patterns and information considered necessary from a large piece of data, and resulting in new knowledge by applying artificial intelligence patterns.

Many applications have been specifically designed to facilitate data mining processing and modeling, such as RapidMiner, Xplenty, Datawatch, Oracle Data Mining, KNIME, Orange, SAS Data Mining, and WEKA, etc.

There are several types of methods in data mining, including the classification method. The classification method is a process of predicting unknown data by taking a collection of patterns from previous data[8] through two processes, the first is the training process, namely the process of building a model from data that has been known to be labeled/classed, and the second is the testing process using known data labeled to determine the accuracy of the modeling results.

Several algorithms are used for classification methods, including K-Nearest Neighbor (kNN). kNN is a supervised machine learning algorithm technique that is the most powerful, simple, and easy to implement[9]. The concept is that, ideally, adjacent objects will usually have the same characteristics. So, by knowing the characteristics of an object, the characteristics of nearby objects can be predictable[10]. Who learns the function of data training to create a class from new data[11]. Euclidean Distance is one of the most well-known and frequently used methods in arithmetic operations. Euclidean Distance is a straight-line interval between two entities within the Euclidean region, both unidirectional and opposite, both of which are of equal value because the Euclidean Distance value is of absolute value[12]

There have been many studies conducted by previous researchers using classification techniques with the kNN algorithm. In the world of health, by using several methods, including kNN to diagnose diseases by detecting white blood cells[13], it is also used to analyze EEG signals using Neurosky Mindwave to classify human feelings[14]. In the wood industry, kNN is used to predict the number and dimensions of sengon wood pieces[15]. Meanwhile, kNN is also used to detect fish freshness by comparing fish eye images to predict fish freshness[16]. Moreover, kNN method is also used in voice research as login authentication[17], image comparison to determine the authenticity of a logo[18], and the authenticity of rupiah currency[19]. From existing research, raising issues related to improving public services is rarely done. In this study, researchers tried to use the kNN algorithm to determine villages that deserve services for traveling population documents and civil registration.

Method

The method used in this research refers to the Cross-Industry Standard Process for Data Mining (CRISP-DM) method, a method that is quite widely recognized in data mining projects. CRISP-DM is a method in the form of a hierarchy and is goal-oriented[20][21], which divides the data mining stages into six stages[22][23]. **Figure 1** shows the flow of six stages in CRISP-DM :

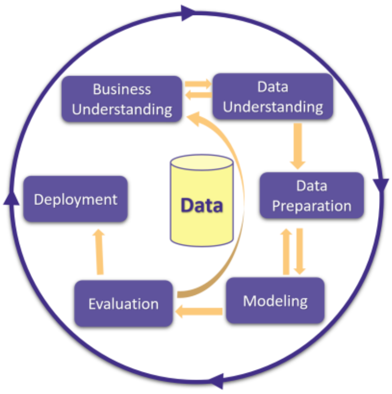

Figure 1. CRISP-DM Flow Diagram

A brief description of CRISP-DM is as follows[20]:

1. Business Understanding. Understanding the objectives and needs of a data mining project. At this stage, the optimal goals and results to be achieved are determined, the infrastructure to be used, and the plan of data mining project implementation.
2. Data Understanding. Collecting and understanding the necessary data. Choosing a data source, describing, and differentiating data quality. At this stage, data improvement can be made if errors are found, and the data is empty when exploring the data.
3. Data Preparation. At this stage, the data collected and verified in the previous stage is cleaned, compiled, and integrated (when there is more than one data source). The data is then formatted to be a ready dataset to be implemented in the model to be used. In supervised learning projects, the process of compiling label attributes and dividing datasets into data training and testing is carried out at this stage
4. Modeling. At this stage, selecting and applying a model that matches the data prepared in the previous stage. Analyzing the modeling results, if needed, it can return to the data preparation stage to adjust the data to the appropriate model.
5. Evaluation. At this stage, the quality, efficiency, and effectiveness of previously applied models are evaluated before being applied in the field. Evaluation is not only on the applied model but also on results adapted to the goals that have been determined at the initial stage.
6. Deployment. The deployment stage is where the information obtained from the data mining process which compiled and presented in a specific form to be useful for users. At this stage, it can be made in the form of a report or a software component.

A. Business Understanding

This study aims to determine eligible villages for mobile population document services. Considering the characteristics of the village's distance, access difficulty, and percentage of ownership population documents such as Birth Certificates, Kartu Identitas Anak (KIA), Electronic Kartu Tanda Penduduk (KTPel), and Kartu Keluarga. This determination is crucial in determining eligible villages, because Disdukcapil has just only used manual determination that cannot be separated from the element of subjectivity. This eligibility determination uses classification techniques in data mining; the algorithm is k-nearest neighbor (kNN), and RapidMiner is an application tool for modeling and measuring its performance.

This project uses a laptop with the specifications: AMD Ryzen 7 5800h Processor, 8Gb RAM, and 500GB SDD, with Microsoft Windows 11 System operation.

B. Data Understanding

The attributes used in this study consist of 10 attributes, as described in **Table 1**. The data type used in this table is based on the data type in the RapidMiner application

Table 1. The List of Attributes used in Determining Eligible Villages for Mobile Services

Attributes	Data Type	Description
No (*id*)	Int	No. Id Record
District_ Name (*Exclude*)	Polynomial	Name of District
Village_ Name (*Exclude*)	Polynomial	Name of Village
K_AL	Real	Percentage of birth certificate ownership (%)
K_KIA	Real	Percentage of KIA ownership (%)
K_KTP	Real	Percentage of KTP ownership (%)
K_KK	Real	Percentage of family card (Kartu Keluarga) ownership (%)
Ak_lok	Real	The difficulty level of access from the village to the city (scale)
Jrk	Real	Mileage from village to city (km)
Eligibility (*Label*)	Polynomial	Eligibility levels for mobile servicing, consist of three categories: "Not Eligible", "Eligible", "Very Eligible"

However, from these ten attributes, as shown in Table 1, two attributes are not included in the process because they have no contribution to the calculation. The two attributes are “District_Name” and "Village_Name”. So that the remaining eight attributes, consist of six regular attributes: "K_AL", "K_KIA", "K_KTP," "K_KK", "Ak_Lok", and "Jrk", the other two attributes are special, "No" as the "id" attribute and the "Eligibility” attribute as the “label” attribute. The "label" attribute consists of three categories, namely, "Not Eligible", "Eligible", and "Very Eligible".

The data used in this study were 175 records taken from Dukcapil service data from April 2022, with the proportion of the eligibility level as shown in **Figure 2**.: Not Eligible = 30%, Eligible = 48%, and Very Eligible = 22%.

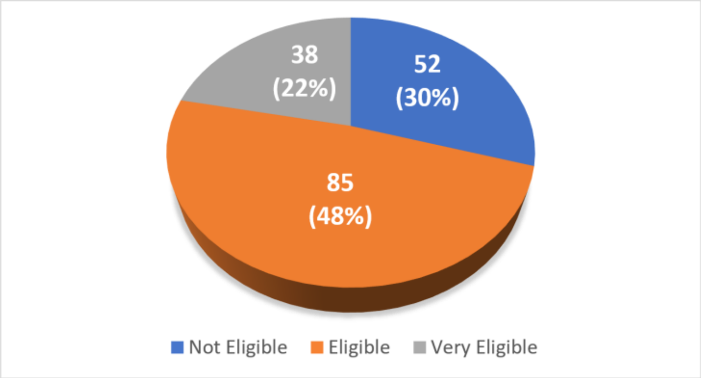

Figure 2. Distribution of Eligibility Variable

C. Data Preparation

Data Preparation is to prepare the data to be processed, standardize and normalize the data, and choose the attributes used in the calculation because not all existing attributes contribute to the calculation. Two of the ten attributes in the dataset have no contribution, so only eight will be used, consisting of two special attributes and six common attributes.

The data on these attributes was carried out through a normalization process to maximize the calculation. The function of this normalization is to avoid data inconsistencies and anomalies. The purpose is to produce more minor data than the original but not eliminate its characteristics[23]. The normalization method chosen in this study is Z-Score Normalization (ZSN)[24], by using mean values and standard deviations to recalculate data $x_{i,n}$ to $x'_{i,n}$ using the **Equation 1**:

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sigma_i} \tag{1}$$

$x'_{i,n}$: Normalized data
$x_{i,n}$: Data before normalization
μ_i : The mean value of the attribute to be searched
σ_i : The standard deviation value of the attribute

The mean of attributes (μ_i) is obtained by **Equation 2**

$$\mu_i = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{x1 + x2 + x3 + \cdots + x_n}{n} \tag{2}$$

μ_i : The mean value of the attribute to be searched
n : Number of records
x_j : Data of j-record

The standard deviation is obtained by **Equation 3**:

$$\sigma_i = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(x_j - \mu_i\right)^2} \tag{3}$$

$$\sigma_i = \sqrt{\frac{(x_1 - u_i)^2 + (x_2 - u_i)^2 + (x_3 - u_i)^2 + \cdots + (x_n - u_i)^2}{n}}$$

σ_i : Standard deviation value of the attribute
n : Total record of training data
x_j : The value of the attribute on the row to be searched for
μ_i : The mean value of the attribute to be searched

The eligibility dataset with 175 records is then divided into two parts using the stratified sampling method, with a composition of 90% (158 records) as training data and 10% (17 records) as testing data.

D. Modeling

The next step is to prepare dataset, training data, and testing data implemented into the model using the RapidMiner tool. The chosen model in this study is K-Nearest Neighbor (kNN), using two forms of modeling, the first modeling uses non-normalized data, and the second uses normalized data using the Z-Score method. The aim is to determine how much influence normalization has on algorithm performance. This modeling is done using the RapidMiner Studio application. The steps in the kNN algorithm are as follows[25]:

1. Determination of the value of K. K is the amount of data of the nearest neighbor, there is no standard union in determining the value of K,
2. Calculate the distance from the test data with each row of training data. The calculation uses the distance formula. The distance formula used in this study is Euclidean distance with Equation 4 [26][27]:

$$D(p,q) = \sqrt{\sum_{j=1}^{n}(p_j - q_j)^2} \quad (4)$$

$$D(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2}$$

Where

D(p,q) : Distance between new testing data and training data
p : New testing data
q : Training data
n : Number of rows from the training data

3. Sort the distance measurement results from smallest value to large
4. Take the most majority class set from the five training data with the shortest distance as a class of the new test data. **Figure 3** Is a visualization of modeling kNN algorithms using the RapidMiner application, from data retrieval to performance measurement.

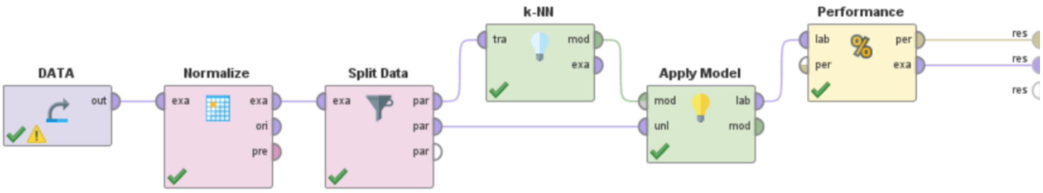

Figure 3. Visualization Modeling using RapidMiner Application

E. Evaluation

Evaluation measures the performance of the modeling that has been made in this study. The measurement uses using confusion matrix, and from this confusion, the matrix is analyzed for how much accuracy, average precision class, and average recall class. The larger these values, the better the modeling performed.

a. Accuracy is a value that describes how accurately the model classifies correctly or the success ratio of the model in predicting the actual data. In other words, accuracy is how close the model's predicted results to the actual data. This process is called the process of measuring accuracy and loss. The measurement formula is based on Mao (2020) that can be seen in **Equation 5** [28]

$$Accuracy\ (\%) = \frac{Number\ of\ true\ classification}{Total\ Data} x100\% \quad (5)$$

b. Precision is a description of the accuracy level between the requested data and the results given by the model, which is the ratio of true positive prediction results to all positive prediction results, precision values obtained by **Equation 6**[29]:

$$Precission\ (\%) = \frac{True\ Positive}{true\ positive + False\ Positive} x100\% \quad (6)$$

c. Recall or sensitivity is the value of a model's success in rediscovering information or comparing the number of positive true predictions with the number of positive correct data. Recall can be formulated by **Equation 7**[29]:

$$Recall\ (\%) = \frac{True\ Positive}{True\ Positive + False\ Negative} x100\% \quad (7)$$

Results and Discussion

A. Data Preparation

For modeling using normalization, a normalization process is held before the split process. The normalization technique is Z-Score or zeroes score normalization by converting values to a broad scale, where the average value is zero and the standard deviation is one.

The first step is determining each attribute's mean value and standard deviation.

Referring to **Equations 2 and 3**. For example, to obtain the mean value of K_AL (μ_{K_Al}) and standard deviation of K_AL (σ_{K_Al}) can be searched by:

- The mean value of K_AL (μ_{K_Al}):

$$\mu_{K_{AL}} = \frac{93.38 + 85.60 + 89.23 + \cdots + 90.48}{175}$$

$$\mu_{K_{AL}} = \frac{15{,}221.31}{175} = 86.98$$

- Standard deviation of K_AL (σ_{K_Al})

$$\sigma_{K\_AL} = \sqrt{\frac{(93.38-86.98)^2+(85.60-86.98)^2+(89.23-86.98)^2+\cdots+(90.48-86.98)^2}{175}}$$

$$\sigma_{K\_AL} = \sqrt{\frac{9{,}303.54}{175}} = \sqrt{53.16} = 7.29$$

This process is applied to all existing attributes so the mean and standard deviation values are summarized in **Table 2**

Table 2. Mean Value and Standard Deviation of Each Attribute

	K_AL	K_KIA	K_KTP	K_KK	Ak_Lok	Jrk
mean value (μ)	86.98	10.98	79.77	57.35	3.19	43.19
standard deviation (σ)	7.29	11.04	5.85	10.01	1.34	28.96

After determining each attribute's mean value and standard deviation, the next step is to calculate the normalization value of each attribute from each record to the end using the Z-Score formula. As in **Table 3**, the formula used refers to **Equation 1** to produce a normalized dataset. An example to get the normalization result of the K_AL attribute is as follows:

$$x'_{K\_AL,1} = \frac{93.38-86.98}{7.29} = 0.88$$

$$x'_{K\_AL,2} = \frac{85.60-86.98}{7.29} = -0.19$$

This calculation is continued until all data on the K_AL attribute is normalized. Furthermore, in the same way, it is done for the other attributes.

Table 3. Data Before and After Normalization

NO	K_AL		K_KIA		K_KTP		K_KK		Ak_Lok		Jrk	
	Before	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*
1	93.38	0.88	18.48	0.68	86.26	1.11	74.94	1.76	6	2.11	2.1	-1.42
2	85.60	-0.19	25.05	1.28	69.67	-1.73	58.86	0.15	6	2.11	0	-1.50
3	89.23	0.31	16.55	0.51	83.32	0.61	70.12	1.28	6	2.11	2.1	-1.42
4	89.78	0.38	18.78	0.71	80.53	0.13	68.40	1.11	6	2.11	3.4	-1.38
5	90.46	0.48	13.22	0.20	79.14	-0.11	68.51	1.12	6	2.11	0.5	-1.48
…	…	…	…	…	…	…	…	…	…	…	…	…
174	80.00	-0.96	8.62	-0.21	77.28	-0.43	47.37	-1.00	1.5	-1.26	82	1.34
175	90.48	0.48	10.58	-0.04	76.16	-0.62	66.67	0.93	1.5	-1.27	47	0.13

For the first model (modeling without prior data normalization), The data is directly divided without a normalization process. While the second model, the data is divided after the normalization process. This split process produces two different data, training and testing data, with a ratio of 9:1, 90% (158 rows) for training data and 10% (17 rows) for testing data. Test data results from split data as shown in **Table 4**

Table 4. Data Testing was Generated by Splitting using the Stratified Sampling Method.

No	Id	K_AL	K_KIA	K_KTP	K_KK	Ak_Lok	Jrk	Eligibility
1	5	0.48	0.20	-0.11	1.11	2.10	-1.47	Not eligible
2	20	0.39	-0.91	-0.20	0.02	0.98	-1.00	Not eligible
3	28	0.93	-0.75	1.11	1.73	-0.14	-0.11	Eligible

No	Id	K_AL	K_KIA	K_KTP	K_KK	Ak_Lok	Jrk	Eligibility
4	29	1.12	2.25	-1.24	0.18	-0.14	-1.14	Eligible
5	55	-0.93	-0.68	-0.10	-0.68	-0.14	0.06	Eligible
6	76	1.02	0.56	-0.14	0.60	-0.14	0.82	Eligible
7	78	-2.06	-0.72	-2.34	-2.00	-1.26	0.10	Very eligible
8	89	-0.08	-0.57	0.71	0.15	0.98	-1.14	Not eligible
9	93	0.28	-0.19	-0.43	-1.09	-0.14	-0.59	Eligible
10	94	0.50	-0.41	1.08	0.94	0.98	-1.18	Not eligible
11	116	0.40	-0.68	0.88	1.37	0.98	-1.11	Not eligible

.

B. *KNN Algorithm*

KNN is one of the machine learning methods to predict something, predict and give a class to something that does not yet have a class by looking at the closeness of its characteristics[30]. Based on the stages of data mining for the K-Nearest Neighbor algorithm[25], the steps are:

1. Set the value k=5.
2. Calculate the distance between the testing data attribute and all training data attributes. In this case, the Measurement Equation for Euclidean Distance is used as **Equation 4**. For example, looking for the distance between the third testing data (id=28) attributes with the first (id=1) of the training data attributes

$$d = \sqrt{\left(\begin{array}{c}(0.93-0.87)^2 + \left((-0.75)-0.68\right)^2 + (1.106-1.105)^2 \\ +(1.73-1.75)^2 + \left((-0.14)-2.09\right)^2 + \left((-0.109)-(-1.41)\right)^2\end{array}\right)}$$

$$d = \sqrt{8.747}$$

$$d = 2.957$$

Next, repeat this calculation for all attributes in all training data, resulting in comparison data from the distance calculation between the new test data and all training data, as shown in **Table 5**

3. Sort the data by the smallest Euclidean Distance. Table 5 describes the results of measuring the distance between the 3rd test data (id=28) and all training data that the distance value has been sorted as ascending.

Table 5. Data Distance Between Attributes in the 3rd Testing Data (Id=28) With All Attributes of Training Data that have been sorted

No	Id	Eligibility	K_AL	K_KIA	K_KTP	K_KK	Ak_Lok	Jrk	Euclidean Distance
1	25	Eligible	0.05	0.02	0.00	0.06	0.00	0.01	0.38
2	26	Eligible	0.22	0.06	0.02	0.07	0.00	0.03	0.63
3	48	Not eligible	0.02	0.17	0.01	0.47	0.00	0.20	0.93
4	27	Not eligible	0.14	0.11	0.06	0.02	0.00	0.68	1.00
5	24	Not eligible	0.00	0.12	0.01	0.08	0.00	1.14	1.16
6	17	Not eligible	0.00	0.09	0.01	0.03	1.25	0.17	1.25
7	23	Not eligible	0.05	0.05	0.23	1.15	0.00	0.12	1.26
…	…	…	…	…	…	…	…	…	…
158	145	Very eligible	29.62	0.01	16.72	20.94	1.25	0.17	8.29

4. Determine the class of testing data by taking the majority class from a class collection in the number of K records. And referring to **Table 5**, the results of measuring the distance between the third test data (id=28) and all training data. It can be concluded that the class of the data being tested is "Not Eligible" because of the five nearest neighboring classes. There are two classes of "Eligible" and three classes of "Not eligible"

C. *Evaluation of Modeling Results*

Evaluation using confusion matrix[31] concluded that modeling the kNN algorithm using unnormalized data, with values k=5, the system was only able to correctly classify 15 data from 17 data, with an accuracy of 88.24%, an average precision of each class of 89.17% and an average class recall value of 89.17% as seen in **Table 6**

Table 6. Confusion Matrix Evaluation of Modeling Without Normalization

	True Not eligible	True Eligible	True Very eligible	Class precision (%)
Pred. Not eligible	4	1	0	80.00
Pred. Eligible	1	7	0	87.50
Pred. Very Eligible	0	0	4	100

	True Not eligible	True Eligible	True Very eligible	Class precision (%)
Class recall (%)	80.00	87.50	100	

Meanwhile, modeling with data that has been normalized using the Z-Score technique, the system was able to classify 16 data from 17 test data with an accuracy of 94.12%, an average precision of each class of 94.44%, and a class recall value of 95.83% as seen in the **Table 7**. Comparison of modeling performance using normalization and without normalization, as shown in **Figure 4**

Table 7. Confusion Matrix Evaluation of Modeling With Normalization

	True Not eligible	True Eligible	True Very eligible	Class precision (%)
Pred. Not eligible	5	1	0	83.33
Pred. Eligible	0	7	0	100
Pred. Very Eligible	0	0	4	100
Class recall (%)	100	87.50	100	

Model comparison using normalized data with non-normalized data, as shown in **Figure 4**. Data normalization increased the accuracy value by 5.88% of the data without normalization. Similarly, the average precision and average recall values increased by 5.27% and 6.66%.

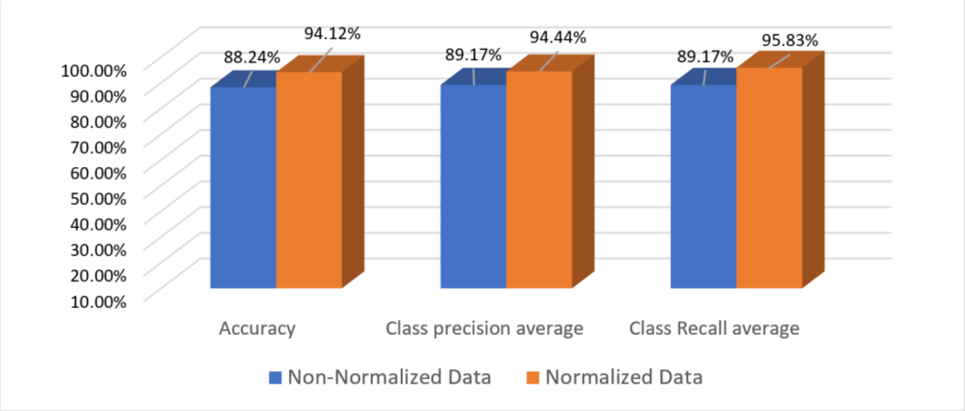

Figure 4. Graphic Comparison of kNN Modeling Performance of Mobile Service Eligibility Data

Conclusion

The data is taken from the service data of the Population and Civil Registration Service in Alor NTT Regency. The number of data used in this case was 175 data with seven assessment attributes. The 175 data were divided into two parts 158 data were used as training data, and 17 were used as testing data, with the value of K set to 5. Based on the results of the discussion on determining the eligibility of villages to get mobile services using the kNN method with RapidMiner software, this study concludes that modeling using normalized data using Z-Score produces higher accuracy than data without normalization. Normalized data get an accuracy of 94.12%. In comparison, data without normalization only reaches 88.24%, the average class precision on normalized data is 94.44% while without normalization is 89.17%, and the average class recall on normalized data is 95.83% while without normalization is only 89.17%, thus getting maximum results, the dataset must be normalized.

References

[1] Dirjen Dukcapil, “Tutup tahun 2021, Kinerja Dukcapil torehkan catatan positif,” 2021. https://dukcapil.kemendagri.go.id/berita/baca/985/tutup-tahun-2021-kinerja-dukcapil-torehkan-catatan-positif (accessed Oct. 18, 2022).

[2] Permendagri, “Peraturan menteri dalam negeri Republik Indonesia Nomor 7 Tahun 2019 tentang Pelayanan Administrasi Kependudukan secara *Daring*.” Jakarta, 2019.

[3] Permendagri, “Peraturan menteri dalam negeri No. 109 Tahun 2019 tentang formulir dan buku yang digunakan dalam Administrasi Kependudukan.” Jakarta, 2019.

[4] Dirjen Dukcapil, “Jemput Bola Dukcapil Terbitkan 2.578 Dokumen Kependudukan bagi Masyarakat Baduy,” 2022. https://dukcapil.kemendagri.go.id/berita/baca/995/jemput-bola-dukcapil-terbitkan-2578-

dokumen-kependudukan-bagi-masyarakat-baduy (accessed Nov. 20, 2022).

[5] Dirjen Dukcapil, “Terus dan Terus, Jemput Bola Dukcapil Jangkau Pulau Kera di Kupang,” 2022. https://dukcapil.kemendagri.go.id/berita/baca/1213/terus-dan-terus-jemput-bola-dukcapil-jangkau-pulau-kera-di-kupang (accessed Nov. 20, 2022).

[6] H. T. Jiawei Han, Jian Pei, *Data Mining: Concepts and Techniques*, 4th ed. Cambridge: Katey Birtcher, 2022.

[7] S. Suprianto, “Implementasi Algoritma Naive Bayes untuk menentukan lokasi strategis dalam membuka usaha menengah ke bawah di kota Medan (Studi Kasus: Disperindag kota Medan),” *J. Sist. Komput. dan Inform.*, vol. 1, no. 2, pp. 125–130, 2020, doi: 10.30865/json.v1i2.1939.

[8] R. Yendra, L. Marifni, and I. Suryani, “Klasifikasi Data Mining untuk seleksi penerimaan calon pegawai negeri sipil tahun 2017 menggunakan metode Naïve Bayes,” *J. Sains Mat. dan Stat.*, vol. 6, no. 1, pp. 65–78, 2020, doi: 10.24014/jsms.v6i1.9254.

[9] S. Kiran, J. Guru, R. Kumar, N. Kumar, D. Katariya, and M. Sharma, “Credit card fraud detection using Naïve Bayes model based and KNN classifier,” *Int. J. Adv. Res.*, vol. 4, no. 3, pp. 44–47, 2018, [Online]. Available: www.IJARIIT.com

[10] H. S. Khamis, “Application of K-Nearest Nieghbour Classification in medical data mining in the context of kenya,” *Int. J. Inf. Commun. Technol. Res.*, vol. 4, no. December, pp. 990–1000, 2014.

[11] C. L. Liu, C. H. Lee, and P. M. Lin, “A fall detection system using k-nearest neighbor classifier,” *Expert Syst. Appl.*, vol. 37, no. 10, pp. 7174–7181, 2010, doi: 10.1016/j.eswa.2010.04.014.

[12] S. P. Patel and S. H. Upadhyay, “Euclidean distance based feature ranking and subset selection for bearing fault diagnosis,” *Expert Syst. Appl.*, vol. 154, pp. 1–16, 2020, doi: 10.1016/j.eswa.2020.113400.

[13] A. S. Musliman, A. Fadlil, and A. Yudhana, “Identification of White Blood Cells using Machine Learning Classification Based on Feature Extraction,” *JOIN (Jurnal Online Inform.*, vol. 6, no. 1, pp. 63–72, 2021, doi: 10.15575/join.v6i1.704.

[14] A. Yudhana, A. Muslim, D. E. Wati, I. Puspitasari, A. Azhari, and M. M. Mardhia, “Human emotion recognition based on EEG signal using fast fourier transform and K-Nearest neighbor,” *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 6, pp. 1082–1088, 2020, doi: 10.25046/aj0506131.

[15] A. Yudhana, S. Sunardi, and A. J. S. Hartanta, “Algoritma K-NN dengan Euclidean Distance untuk prediksi hasil penggergajian kayu sengon,” *Transmisi*, vol. 22, no. 4, pp. 123–129, 2020, doi: 10.14710/transmisi.22.4.123-129.

[16] A. Yudhana, R. Umar, and S. Saputra, “Fish Freshness Identification using Machine Learning: Performance Comparison of k-NN and Naïve Bayes Classifier,” *J. Comput. Sci. Eng.*, vol. 16, no. 3, pp. 153–164, 2022, doi: 10.5626/JCSE.2022.16.3.153.

[17] R. Umar, I. Riadi, A. Hanif, and S. Helmiyah, “Identification of speaker recognition for audio forensic using k-nearest neighbor,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 3846–3850, 2019.

[18] R. Umar, I. Riadi, and D. A. Faroek, “A komparasi image matching menggunakan metode K-Nearest Neightbor (KNN) dan Support Vector Machine (SVM),” *J. Appl. Informatics Comput.*, vol. 4, no. 2, pp. 124–131, 2020, doi: 10.30871/jaic.v4i2.2226.

[19] M. Miladiah, R. Umar, and I. Riadi, “Implementasi local binary pattern untuk deteksi keaslian mata uang rupiah,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, pp. 197–201, 2019, doi: 10.26418/jp.v5i2.32721.

[20] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[21] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, and J. R. Coronado-Hernández, “Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test,” *Procedia Comput. Sci.*, vol. 198, no. 2020, pp. 512–517, 2021, doi: 10.1016/j.procs.2021.12.278.

[22] A. Morais, H. Peixoto, C. Coimbra, A. Abelha, and J. Machado, “Predicting the need of Neonatal Resuscitation using Data Mining,” *Procedia Comput. Sci.*, vol. 113, pp. 571–576, 2017, doi: 10.1016/j.procs.2017.08.287.

[23] M. F. Rifai, H. Jatnika, and B. Valentino, “Penerapan algoritma Naïve Bayes pada sistem prediksi tingkat kelulusan peserta Sertifikasi Microsoft Office Specialist (MOS),” vol. 12, no. 2, pp. 131–144, 2019, doi: https://doi.org/10.33322/petir.v12i2.471.

[24] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, p. 105524, 2020, doi: https://doi.org/10.1016/j.asoc.2019.105524.

[25] N. Dengen, “Comparison Performance of C4 . 5 , Naïve Bayes and K-Nearest Neighbor in,” *Int. Conf. Sci. Inf. Technol.*, pp. 112–117, 2019.

[26] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, pp. 1–7, 2016, doi: 10.21037/atm.2016.03.37.

[27] K. Polat, S. Şahan, and S. Güneş, “Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing,” *Expert Syst. Appl.*, vol. 32, no. 2, pp. 625–631, 2007, doi: 10.1016/j.eswa.2006.01.027.

[28] A. Fadlil, R. Umar, Sunardi, and A. S. Nugroho, “Comparison of Machine Learning Approach for Waste Bottle Classification,” *Emerg. Sci. J.*, vol. 6, no. 5, pp. 1075–1085, 2022, doi: 10.28991/ESJ-2022-06-05-011.

[29] J. Wang, X. Jing, Z. Yan, Y. Fu, W. Pedrycz, and L. T. Yang, “A Survey on Trust Evaluation Based on Machine Learning,” *ACM Comput. Surv.*, vol. 53, no. 5, pp. 107:1-107:37, 2020, doi: 10.1145/3408292.

[30] S. Saputra, A. Yudhana, and R. Umar, “Identifikasi Kesegaran Ikan Menggunakan Algoritma KNN Berbasis Citra Digital,” vol. 10, no. 1, pp. 1–9, 2022, doi: 10.32832/kreatif.v10i1.6845.

[31] N. L. Ratniasih, “Penerapan Algoritma K-Nearest Neighbour (K-Nn) Untuk Penentuan Mahasiswa Berpotensi Drop Out,” *J. Teknol. Inf. dan Komput.*, vol. 5, no. 3, pp. 314–318, 2019, doi: 10.36002/jutik.v5i3.804.