International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jan-24)

http://dx.doi.org/10.12785/ijcds/150111

Optimization of Population Document Services in Villages using Naive Bayes and k-NN Method

Imam Riadi[1], Anton Yudhana[2] and M. Rosyidi Djou[3]

[1]*Department of Information System, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia*
[2]*Departement of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia*
[3]*Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta, 55164 Indonesia*

Received 24 Jul. 2023, Revised 20 Oct. 2023, Accepted 17 Nov. 2023, Published 1 Jan. 2024

Abstract: The Indonesian government strives to improve its registration and the issuance of the population documents program. However, several obstacles are faced, such as the complicated topography of the area and the distance from the village. Therefore, ball pick-up services are urgently needed. The government of Alor Regency, East Nusa Tenggara Province, is one of the regions that has implemented this program. Nonetheless, due to constraints in both time and financial resources, not every village can served. Therefore, villages must be selected fairly so the program can run well. Machine learning, a classification technique using data mining concepts, is expected to address this problem. This research aims to identify the most effective method for classifying eligible villages. The experimental process includes preprocessing, model training using K-Nearest Neighbor (K-NN) and Naïve Bayes, and performance evaluation. The results show that both methods provide good results, albeit with slightly different levels of accuracy. Comparative analysis shows that K-NN has a higher accuracy rate of 97.14% for k=1 and k=2 on the Min-Max-normalized dataset but has the lowest accuracy of 77.1% at k=11 and k=13 on the raw dataset. In comparison, the NB method has an accuracy of 94.29% but is stable on raw and normalized datasets.

Keywords: Data mining, Machine learning, k-NN, Naïve bayes, Comparative analysis, village selection

1. INTRODUCTION

The registration data and population and civil registration documents are essential in maintaining the validity and accuracy of a country's population data. Civil registration documents verify and identify citizens' identities, forming the basis for implementing public services such as healthcare, education, aid distribution, etc. Citizens can access the services they need by possessing valid civil registration documents. Registration data and civil registration documents are also utilized to calculate population statistics, such as population size, age composition, educational levels, and more. These are vital for development planning, public policies, and decision-making across various sectors [1]. Most importantly, population and civil registration documents provide a strong foundation for safeguarding individual rights and interests and supporting overall national development and security.

The *Direktorat Jenderal Kependudukan dan Pencatatan Sipil* (Dirjen Dukcapil) is the leading institution of the Indonesian government, overseeing, managing, and ensuring the smooth functioning of population registration services in Indonesia. Numerous innovations have been implemented to facilitate these services, including digital signatures [2], replacing conventional signatures and stamps on civil registration documents. Additionally, 80g HVS paper [3] has been adopted as a substitute for security paper, making it easier for the public to print documents without visiting government offices. Moreover, the Self-Service Civil Registration Platform or *Anjungan Dukcapil Mandiri* (ADM) which is an ATM-like platform that can print population administration documents, enables individuals to print their documents independently [4], along with the implementation of digital population identity or *Identitas Kependudukan Digital* (IKD) which is a digital form of KTP-Elektronik and other documents that contains electronic information used to represent Population Documents and can be accessed through digital applications on smartphones, displaying Personal Data as an individual identity.

A. Problem Statement

Indonesia is an archipelagic country with a total area of 1,892,555.47 square kilometres, consisting of 16,772 islands and a population of 272,229,372 individuals. Its administrative divisions are structured into 34 provinces, 416 regencies, 98 cities, 7,266 sub-districts, 8,506 urban villages, and 74,961 rural villages [5]. 52.89% of the villages are on flat terrain, 45.77% on slopes, 1.03% in valleys, and 0.31% on cliffs [6]. The geographical conditions, with their

E-mail address: imam.riadi@is.uad.ac.id, eyudhana@ee.uad.ac.id, djou2107048009@webmail.uad.ac.id

influence on the difficulty in access and the long distances, have impacted service development, resulting in an uneven distribution of document ownership.

Similarly, the Alor Regency is the archipelagic regency in the East Nusa Tenggara Province, covering an area of 2,864.64 square kilometres. It consists of 17 sub-districts and 175 rural villages and urban villages. Geographically, the region comprises nine inhabited islands and 11 other islands. The landscape is distinguished by elevated mountainous regions, valleys and cliffs. Approximately 63.94% of the area in Alor Regency has slopes steeper than 40º. Such topographical conditions greatly influence the progress of population services and the ownership of civil registration documents. As a result, tDinas Kependudukan dan Pencatatan Sipil (Disdukcapil) or the Regional Office of Population and Civil Registration, particularly in Alor Regency, must exert extra effort to achieve the target of document ownership.

One of the policies implemented is a mobile service or "ball pick-up" service [7] [8]. The "ball pick-up" service is a strategy to bring population services closer to the community by relocating service teams to specific locations. With this service, the delivery of services can be more optimal.

However, a problem arises as not all villages can be served within a fiscal year due to limited time and resources, requiring selecting eligible villages for the service. To achieve the goal of population services, the selection of these villages must be based on principles of fairness and accountability. Currently, village selection is made manually, which hinders the optimal functioning of the services.

B. Proposed Solution

The solution proposed in this research is to utilize machine learning techniques with data mining classification methods to enhance the village selection process. This study aims to develop a more efficient and practical approach to identifying villages that require population services. This approach will assist in the village selection process by prioritizing villages based on criteria such as distance, accessibility, and service coverage. By doing so, limited resources and time can be optimized, and services can operate at their best.

C. Related Work

The application of classification techniques in machine learning (ML) has been investigated by past researchers, especially in the healthcare domain, and have proven to be significant in the detection and diagnosis of various medical conditions [9]. For instance, ML methods have been applied in the detection of infectious diseases [10], tumors [11] [12], chronic kidney disease, hepatitis C [13], and many other medical conditions. Using ML techniques such as K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Decision Tree, medical data can be analyzed efficiently and accurately to identify specific patterns or characteristics associated with particular diseases. ML enables early diagnosis, risk prediction, and better medical decision-making. The application of ML methods in healthcare has resulted in the development of advanced tools such as artificial immune recognition systems (AIRS) and automated detection tools [14], which assist doctors and medical personnel diagnose diseases with a higher degree of accuracy. That speeds up the diagnosis process and improves treatment and patient care.

The K-NN method combined with the Fast Fourier Transform (FFT) method is also used to classify human brain impulses to detect emotions experienced by a person, whether regular, focused, sad, or surprised. This research was conducted by Yudhana et al. [15]; brain impulses were recorded using a Neurosky Mindwave single-channel electroencephalogram (EEG). The K-NN method produces an average accuracy rate of 83.33%, with the highest accuracy rate reaching 93.33%.

In agriculture, ML techniques using the K-NN algorithm have also been applied to detect plant diseases, including those affecting peanut leaves. With this technique, the system can recognize characteristic patterns of diseased leaves and differentiate them from healthy ones. The K-NN algorithm can accurately detect plant diseases through analysis and classification. That enables farmers to identify diseases in peanut plants early, allowing them to take appropriate control measures to enhance production and prevent the further spread of the diseases [16]. And the Naïve Bayes algorithm also was used in mapping the nitrogen element in the soil to support the increased rice production in Lendah sub-district. Utilizing the Wemos D1 Mini microcontroller in conjunction with the TCS3200 sensor, soil sample data is transmitted to the ceerduad.com website. Subsequently, this soil sample data undergoes classification through the Naïve Bayes algorithm, resulting in an accuracy rate of 87.5% [17].

In the fisheries industry, machine learning techniques, particularly the K-Nearest Neighbors (K-NN) and Naïve Bayes algorithms, are employed to differentiate and choose between fresh and spoiled fish. By examining images of fish eyes, these two algorithms can effectively discern fresh fish from spoiled ones, achieving an accuracy rate of 94% for Naïve Bayes and 97% for K-Nearest Neighbors. [18]

The realm of cybersecurity also benefits from the use of ML methods. Distributed Denial of Service (DDoS) attacks significantly threaten network service providers. DDoS attacks involve flooding networks with high-intensity "illegitimate" data packets, causing congestion or even the complete shutdown of network traffic. The consequences can be highly detrimental. ML methods detect DDoS attacks using artificial neural network algorithms and Naïve Bayes. This approach can detect DDoS attacks with high accuracy (99.9%) [19].

ML methods, such as pattern recognition with K-NN,

are crucial against online banking and nancial industry fraud. With ML, intelligent machines can recognize suspicious transaction patterns and take early detection actions against fraudulent practices. This detection helps reduce losses and enhances overall security in banking operations and nancial transactions [20].

Previous research has shown that ML methods can optimize work in various elds. In terms of public services, particularly in registration programs and issuing population identi cation and civil registration documents, ML can optimize services by detecting or classifying villages that require more intervention and attention and are eligible for more optimal services, such as "ball pickup service" or mobile services. By analyzing historical data and considering the level of di culty accessibility and distance, ML can accelerate and improve the accuracy of determining villages that deserve to be served. Therefore, it is expected to enhance the eciency and eectiveness of public services in the required villages.

2. RESEARCH METHODOLOGY

A. Research Flowchart

Figure 1. Research Flowchart

A research ow is made to facilitate understanding and a clear ow in this research, as shown in Figure 1.

A brief explanation of the research owchart:

1) Data collected from the Population and Civil Registration O ce of Alor Regency.
2) Feature selection is performed. Irrelevant attributes are discarded to produce a relevant dataset.
3) Perform normalization using Z-Score and Min-Max.
4) The datasets, both raw datasets, Z-Score and Min-Max normalized datasets are divided into training subset and test subset.
5) Na ve Bayes and K-NN modeling (k=1 to k=15) using the three subsets of data that has been prepared in the previous stage.
6) Test the K-NN and Naive Baye performance models and compare them to nd the best performance.

B. Dataset Collection and Description

TABLE I. Data Attributes from Disdukcapil

No	Attribute	Description
1.	Seqn	Sequence number. Primary key
2	SubDistrict_name	Sub District name
3	Village_name	Village name
4	OBC	Ownership of a Birth Certi cate, (%)
5	OCID	Ownership of Kartu Identitas Anak (KIA)) or Child Identity Card, (%)
6	OID	Ownership of Kartu Tanda Penduduk Elektronik(KTPel)) or Electronic Identity card, (%)
7	OFC	Ownership of Kartu Keluarga (KK)) or Family Card. (%)
8	LAD	Level of access di culties. (Scale)
9	Dist	Distance from the village (km)
10	Eligibility	Eligibility for mobile service

The datasets used in this study are the percentage of ownership of population and civil registration documents such as birth certi cates, KIA, KTPel, and KK;these four documents are important documents that every resident must own. In addition, the level of dicuulty of access and distance from the village to the place of service in the district and the level of eligibility of villages to receive mobile services, as shown in TABLE I. The data source is processed data from the Disdukcapil of Alor Regency, East Nusa Tenggara Province.

175 data were collected from the Disdukcapil of Alor Regency, as shown in TABLE II, with the data distribution depicted in Figure 2. The villages classi ed as "Not Eligible" accounted for 30% (52 villages), the "Eligible" villages accounted for 48% (85 villages), and 22% (38 villages) were classi ed as "Not Eligible".

C. Data preprocessing

The dataset received from the Disdukcapil of Alor Regency was subsequently subjected to preprocessing.

1) Feature selection

Feature selection, or a variable subset or attribute selection, involves removing irrelevant attributes from the dataset [9][10]. This method reduces complexity in data mining processes [21]. Not all attributes contribute to the processing, and selecting the wrong attributes can lead to over tting

TABLE II. Eligibility Data

Seqn	SubDistrict_name	Villages_name	Eligibility	OBC	OCID	OID	OFC	LAD	Dist
1	Teluk Mutiara	Kalabahi Barat	Not Eligible	93.38	18.48	86.26	74.94	1.50	2.10
2	Teluk Mutiara	Kalabahi Kota	Not Eligible	85.60	25.05	69.67	58.86	1.50	-
3	Teluk Mutiara	Kalabahi Tengah	Not Eligible	89.23	16.55	83.32	70.12	1.50	2.10
...	...	...	...	...	...	...	...	...	...
25	Alor Barat Laut	Ternate	Eligible	95.50	4.05	86.19	72.22	4.50	37.00
26	Alor Barat Laut	Ternate Selatan	Eligible	90.37	5.32	85.50	72.17	4.50	35.00
...	...	...	...	...	...	...	...	...	...
173	Abad Selatan	Wakapsir Timur	Very Eligible	70.51	3.23	75.56	52.56	6.00	66.00
174	Abad Selatan	Kuifana	Very Eligible	80.00	8.62	77.28	47.37	6.00	82.00
175	Abad Selatan	Margeta	Eligible	90.48	10.58	76.16	66.67	6.00	47.00

Figure 2. Distribution of Eligibility Data

and poor predictions. Therefore, out of the ten attributes listed in TABLE I, two attributes (subdistrict_name and village_name) are not included in the modeling process as they do not contribute. Only six attributes (OBC, OCID, OID, OFC, LAD) are used as predictor attributes, one attribute (Seqn) is used as the id attribute or key attribute, and one attribute (Eligibility) is used as the label attribute or class attribute.

2) Normalization

Normalization is a technique to make the dataset into a standard scale. TABLE II shows that the six predictor attributes represent different values and have different scales or value ranges, and the statistic predictor attributes as shown in TABLE III. OBC, OCID, OID, and OFC are percentage values with varying minimum and maximum values. Minimum values range from 0 to 55.72, and maximum values range from 68.47 to 96.88. LAD represents data on a scale from 1.5 to 6. On the other hand, Dist represents the distance (in kilometres) from the village to the service location, with the closest distance being 0 km and the farthest distance being 130 km. These diverse value ranges can have an impact on the model's performance once it's constructed. Consequently, in order to mitigate data variance, normalization should be carried out.

TABLE III. Predictor Attribute Statistics

Atribut	Data presentation	Mean	Std.dev	Min	Max
OBC	%	86.98	7.29	54.01	96.88
OCID	%	10.98	11.07	-	68.47
OID	%	79.77	05.87	55.72	89.68
OFC	%	57.35	10.04	28.80	79.64
LAD	Scale	04.31	01.34	01.50	06.00
Dist	Kilometer	43.19	29.04	-	130.00

The normalization techniques chosen in this research are Z-Score normalization (ZSN) and Min-Max Normalization (MMN) techniques.

a. Z-Score Normalization (ZSN): This method standardizes the attribute's mean value to 0 and its standard deviation to 1, employing the formula (1) [22]:

$$x_{i;n} = \frac{x_{i;n} - \mu_i}{\sigma_i} \quad (1)$$

In formula (1), $x_{i;n}$ indicates the normalized value of the i-th attribute in the n-th record, which results from normalization. $x_{i;n}$ is the original or raw value of the i-th attribute in the n-th record before normalization, μ_i indicates the mean value of the i-th attribute, while σ_i indicates the standard deviation of the i-th attribute.

b. Min-Max Normalization (MMN): This method also scales the data to a specified lower and upper bound, usually ranging from 0 to 1 or -1 to 1, using the formula (2) [22]:

$$x_{i;n} = \left(\frac{x_{i;n} - \min(x_i)}{\max(x_i) - \min(x_i)} \right) (nMax - nMin) + nMin \quad (2)$$

In formula (2), $x_{i;n}$ indicates the normalized value of attribute i in record n, which is the result after normalization. $x_{i;n}$ is the original value of attribute i in record n before normalization. $\min(x_i)$ is the minimum value of attribute i in the dataset, indicating the minimum value of that attribute. $\max(x_i)$ is the maximum value of attribute i in the dataset, indicating the maximum value of that attribute. $nMax$ and $nMin$ are the designated upper and lower bounds within the specified normalization

range. These values determine the desired scale of the data within the normalization process. These values can vary depending on the speci c data needs and context.

In this study, three kinds of modeling data will be provided, One of these datasets will employ non-normalized data, while the others will use data that has undergone normalization using Z-Score Normalization (ZSN) and Min-Max Normalization (MMN). The objective is to assess the impact of normalization on the performance of the classi cation algorithm.

D. Classi er Algorithms

Classi cation in data mining is categorizing or grouping data into prede ned classes or labels based on speci c patterns or characteristics. The goal is to develop a model or algorithm to predict the class or label of unknown data based on the available information [23]. Classi cation involves a two-step process [24]. The initial stage involves constructing a model through the application of a classi cation algorithm to a training dataset. Subsequently, in the second phase, the model that has been generated is evaluated using a prede ned test dataset to gauge the performance and accuracy of the trained model.

So, classi cation is the procedure of assigning a class label to a dataset when the class label is not initially known. Following this, the model's performance is assessed by comparing its predictions to the actual labels present in the test data. This evaluation enables the measurement of the model's performace or e ectiveness in classifying new data. Several algorithms are commonly used in classi cation, such as Naïve Bayes, Random forest, decision tree, K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM). However, this research only uses two classi cation algorithms: Naïve Bayes and K-Nearest Neighbors (K-NN). The selection of Naïve Bayes is based on its simplicity [25][26], e ectiveness [27], speed [28], and high accuracy rate [29]. While K-NN was also chosen because of its simplicity [30], e ciency [31], and ability to work with limited information [29], KNN is also included in the top-10 data mining algorithms [21]. The speed of implementation in the real world [24][32] is also a consideration in choosing these two algorithms.

1) Naive Bayes (NB)

Naïve Bayes is a supervised learning method based on Bayes Theorem by calculating the probability and statistical [33] of some categories with some conditions [26] a simple, e ective, fast method with high accuracy, and insensitive to missing data [27]. NB works under a strong assumption of independence among each attribute [19]. Hence the model is called an "independent feature model" [34]. This assumption makes the method popular for handling various types of data, whether continuous or categorical [35]. Bayes' theorem determines the posterior The probability $P(C|x)$ is calculated from the prior class probability $P(C)$, the prior predictor probability $P(x)$, and the likelihood $P(x|C)$, as shown in Equation (3).

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} \quad (3)$$

Suppose the predictors are not discrete but have continuous values and are assumed to be Gaussian (normal) distribution samples. In that case, the formula used in determining the likelihood of $P(x|C)$ is as in equation (4).

$$P(x_i|y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{(x_i - \mu_k)^2}{2\sigma_k^2}} \quad (4)$$

The Gaussian NB, which assumes a Gaussian distribution, is used in this case. In implementing Gaussian NB, each feature's mean and standard deviation values for each class in the training dataset are rst calculated [36]. Then, this formula is used to calculate the posterior probability based on the previously calculated mean and standard deviation values.

2) K-Nearest neighbors (K-NN)

K-NN is a classi cation method that works by measuring the similarity between a new instance and existing instances (training data) based on weights assigned to each attribute [37]. Data is presented regarding distance measures, using formulas such as Euclidean, Manhattan, etc. The nearest data points are termed neighbors and are assembled into a neighbor group according to a speci ed 'k' value. The classi cation of the new object is determined by assessing the prevailing class within this neighbor group.

This K-NN method is known as a "lazy learner" [38] because it considers local data and postpones all computation until the classi cation process [39], K-NN is an algorithm that does not process data during training, only calculates the distance between input data and existing data, and then performs classi cation based on the majority of votes from the nearest neighbours. It means it tests relevant data without creating a generalized training data model. The generalization result is also postponed until an input request or test data arrives [40]; Induction or model building is also postponed until runtime or when test data is available [41]. K-NN has no pre-generated classi cation model, and whenever there is an input request, it recalculates the class based on the nearest neighbours, this makes the algorithm exible and adaptive to data changes, but also makes it slower in handling new requests, as it has to calculate the distance to the nearest neighbour each time.

Limitation of this method is the di culty in determining the optimal value of k [31]. Choosing the wrong value of k can lead to over tting or even under tting in the classi cation, so training with varying values of k is often required to achieve maximum accuracy. Several distance formulas are commonly used, including Euclidean, Manhattan, Minkowski, Chebyshev, and Hamming, depending

on the problem characteristics and data type being solved. In this study, the Euclidean Distance formula is used:

$$D(p;q) = \sqrt{\sum_{j=1}^{n} (p_j - q_j)^2} \quad (5)$$

Where D(p, q) represents the distance between the new data (point p) and the training data (point q), p refers to the new data, q refers to the training data, and n is the number of rows in the training data.

E. Evaluation

To compare the performance of the classification algorithms, the modeling will be repeated with a try-error system with several combinations of raw and normalized data and a selection of k values from 1 to 15 for K-NN. The evaluation method uses a confusion matrix for all models [13]. A confusion matrix, sometimes called as an error matrix, is a dedicated tabular format used to visually illustrate the performance of an algorithm. Typically, this matrix is widely utilized in the context of supervised learning. The way this matrix operates is by comparing the predicted results with the actual values, which leads to the creation of a matrix containing metrics such as accuracy, precision, recall, and F1-Score. The terms commonly used in the Confusion Matrix include:

True Positive (TP): The count of data points correctly predicted as positive.

True Negative (TN): The count of data points correctly predicted as negative.

False Positive (FP): The count of data points incorrectly predicted as positive.

False Negative (FN): The count of data points incorrectly predicted as negative.

The evaluation of the model will be based on standard data mining evaluation metrics. These metrics typically include [42]:

1) Accuracy
Accuracy is an evaluation model that describes how much the classification model predicts the correct overall. Accuracy measures the model's ability to classify instances accurately. It quantifies the percentage of correctly predicted instances relative to the overall number. The range of accuracy values is between 0 and 1, with a value of 1 indicating perfect accuracy where there are no errors in prediction. The equation obtains accuracy values:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (6)$$

2) Precision
Precision is an evaluation model that describes the extent to which a classification model provides correct predictions for positive classes. Precision provides information about the model's accuracy in identifying positive examples from the overall positive predictions. Precision has a range of values between 0 and 1, with a value of 1 indicating perfect precision where there are no false positive misclassifications. The equation obtains precision:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

3) Recall
Recall, also known as true positive rate or sensitivity, is an evaluation model that describes how many classifications the model can accurately identify the number of positive data. Its value ranges from 0 to 1, where 1 indicates that the model can correctly identify all positive data. Conversely, a value of 0 indicates that the model cannot identify any positive data. Recall is obtained using the equation:

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

4) F1-score
F1-Score is an evaluation model that shows the balance between precision and recall. F1-score combines the precision and recall values into a single number that describes the classification quality of the model. Its value ranges from 0 to 1, where 0 signifies poor performance, and 1 signifies excellent performance. The F1-Score value is obtained using the equation:

$$F1\ Score = \frac{2\ (Precision\ Recall)}{Precision + Recall} \quad (9)$$

F. Tools

The tools used in this research are Rapidminer studio Version 10.1 software for modeling and accuracy testing and Microsoft Excel 2021 for data processing and graphics. The hardware used is a laptop with specifications: Asus Vivobook, Processor Ryzen 7 5800h, SSD 500GB, Memory 8GB, with Windows 11 64bit operating system.

3. RESULTS AND DISCUSSIONS

A. Data Preprocessing

The data collected from Disdukcapil as TABLE I went through feature selection, where attributes that contributed to the classification were retained, while attributes that did not contribute were removed. Two attributes (SubDistrict_Name and Village_Name) were removed, leaving eight attributes; six predictor attributes (OBC, OID, OCID,

TABLE IV. Comparison of Raw Data With ZSN Data

Seqn	OBC (Raw)	OBC (ZSN)	OCID (Raw)	OCID (ZSN)	OID (Raw)	OID (ZSN)	OFD (Raw)	OFD (ZSN)	LAD (Raw)	LAD (ZSN)	Dist (Raw)	Dist (ZSN)
1	93.384	0.876	18.479	0.678	86.261	1.105	74.937	1.751	1.500	-2.095	2.100	-1.415
2	85.598	-0.189	25.051	1.271	69.668	-1.722	58.861	0.150	1.500	-2.095	0.000	-1.487
3	89.232	0.308	16.549	0.503	83.317	0.604	70.118	1.271	1.500	-2.095	2.100	-1.415
4	89.775	0.382	18.777	0.704	80.533	0.129	68.399	1.100	1.500	-2.095	3.400	-1.370
5	90.465	0.477	13.221	0.203	79.135	-0.109	68.507	1.111	1.500	-2.095	0.500	-1.470
6	90.351	0.461	24.386	1.211	76.688	-0.526	61.503	0.413	1.500	-2.095	1.200	-1.446
...	...	...	...	...	...	...	...	...	...	...	...	...
175	90.476	0.478	10.582	-0.036	76.164	-0.615	66.667	0.928	6.000	1.258	47.000	0.131

TABLE V. Comparison of Raw Data With MMN Data

Seqn	OBC (Raw)	OBC (MMN)	OCID (Raw)	OCID (MMN)	OID (Raw)	OID (MMN)	OFD (Raw)	OFD (MMN)	LAD (Raw)	LAD (MMN)	Dist (Raw)	Dist (MMN)
1	93.384	0.919	18.479	0.270	86.261	0.899	74.937	0.907	1.500	0.000	2.100	0.016
2	85.598	0.737	25.051	0.366	69.668	0.411	58.861	0.591	1.500	0.000	0.000	0.000
3	89.232	0.822	16.549	0.242	83.317	0.813	70.118	0.813	1.500	0.000	2.100	0.016
4	89.775	0.834	18.777	0.274	80.533	0.731	68.399	0.779	1.500	0.000	3.400	0.026
5	90.465	0.850	13.221	0.193	79.135	0.689	68.507	0.781	1.500	0.000	0.500	0.004
6	90.351	0.848	24.386	0.356	76.688	0.617	61.503	0.643	1.500	0.000	1.200	0.009
...	...	...	...	...	...	...	...	...	...	...	...	...
175	90.476	0.851	10.582	0.155	76.164	0.602	66.667	0.745	6.000	1.000	47.000	0.362

TABLE VI. Statistical Comparison of Raw and Normalized Data Datasets (ZSN and MMN)

ATTRB	DATASET	MIN	MAX	MEAN	STDEV
OBC	Raw	54.011	96.875	86.979	7.312
	ZSN	-4.509	1.353	0.000	1.000
	MMN	0.000	1.000	0.769	0.171
OCID	Raw	0.000	68.475	10.976	11.074
	ZSN	-0.991	5.192	0.000	1.000
	MMN	0.000	1.000	0.160	0.162
OID	Raw	55.723	89.683	79.774	5.869
	ZSN	-4.098	1.688	0.000	1.000
	MMN	0.000	1.000	0.708	0.173
OFD	Raw	28.796	79.643	57.355	10.039
	ZSN	-2.845	2.220	0.000	1.000
	MMN	0.000	1.000	0.562	0.197
LAD	Raw	1.500	6.000	4.311	1.342
	ZSN	-2.095	1.258	0.000	1.000
	MMN	0.000	1.000	0.625	0.298
Dist	Raw	0.000	130.000	43.186	29.044
	ZSN	-1.487	2.989	0.000	1.000
	MMN	0.000	1.000	0.332	0.223

OFC, LAD, and Dist), one ID attribute (Seqn), and one label or target attribute (Eligibility).

The data from the selected attributes have different scales, as shown in TABLE III. This difference in scale can have an impact on classification performance. Therefore, a normalization method is needed to transform the dataset with a uniform scale.

This research uses two normalization methods: Z-Score Normalization (ZSN) and Min-Max Normalization (MMN). ZSN refers to Equation (1), and the normalized dataset is presented in Table IV. MMN refers to Equation (2) and produces normalized data, as seen in TABLE V.

TABLE IV and TABLE V illustrate the comparison of data transformation from raw data to normalized data using ZSN and MMN. TABLE VI summarizes the statistics of the raw and normalized data. The raw data has a broad and diverse range of attributes, as seen from the irregular minimum, maximum, mean, and standard deviation values. On the other hand, in ZSN normalization, the scale becomes more regular, with smaller minimum and maximum values, mean = 0, and standard deviation = 1. Meanwhile, in MMN normalization, the minimum value = 0, the maximum value = 1, and the mean and standard deviation values range between 0 and 1. The patterns indicate that normalization has successfully made the dataset reach a more regular scale, and its effect on classification will be further discussed in the following analysis.

B. Data Training and Data Testing

The dataset is partitioned into two subsets to initiate the classification process: training and testing data. This partition is performed in a ratio of 80:20, where 80% of the dataset is allocated as training data, and the remaining 20% is designated as testing data. The partitioning follows the split data stratified sampling method, which ensures that the distribution of class labels in both subsets maintains the same proportions as the original dataset.

C. Implementation of Naïve Bayes

In this research, the Naïve Bayes algorithm is applied to three dataset types, namely raw dataset and the ZSN and MMN dataset . The algorithm used in this study applies a variation of the NB method known as Gaussian Naïve Bayes. This choice was made because the predictor attribute consists of continuous data, and the calculation refers to Equation (4). From the modeling results that have been carried out , the same accuracy value is obtained, both using the original dataset as well as ZSN and MMN, with an accuracy rate of 94.29% and Precesion, Recall, and F1-score value of 94.71%.

D. Implementation of K-NN

The K-NN algorithm is a classification method that measures the distance between new data and training data within a set of k nearest neighbors. To start modeling the K-NN algorithm, we need to select a distance metric and determine the value of k. This study uses the Euclidean distance formula as the distance metric, as shown in Equation (5). The influence of various 'k' values on the performance of the K-NN algorithm is assessed through multiple iterations. Experiments were conducted with three types of datasets: raw data, data normalized using ZSN, and data normalized using MMN. The k values explored in these experiments ranged from 1 to 15. From the modeling results, the accuracy level varies; the highest value of accuracy is 97.14% on the MMN dataset with the values of K = 1 and K = 2, while the lowest is 77.14% on the original dataset with the values of K = 11 and K= 13, and the highest F1-Score value is 97.40

E. Performance Comparison of NB and K-NN Methods

1) Accuracy

Accuracy is an evaluation metric that quantifies the model's capability to make correct predictions across the entire dataset.

Based on the analysis results, as shown in Figure 3, it can be concluded that:

Figure 3. Graph of accuracy values of NB and K-NN

RAW dataset: On the raw dataset, the Naive Bayes (NB) algorithm obtained an accuracy rate of 94.29%. The average accuracy for K-NN was 84.00%, with the highest accuracy achieved at k= (1 and 2), which was 94.29%. However, there was a significant decrease in the accuracy rate, reaching the lowest value of 77.14% when k= (11 and 13).

ZSN datase When the ZSN dataset was used, NB produced the same accuracy rate as the raw dataset, which was 94.29%. The average accuracy of K-NN on the ZSN dataset was 92.76%. The highest accuracy rate achieved by K-NN was also 94.29%, but this time, it occurred at a more varied number of k values, namely k=(1, 2, 3, 4, 10, 11, 12). The lowest accuracy rate was 91.43% at k=(5, 6, 7, 8, 9, 13, 14, 15).

MMN dataset: In using the MMN dataset, NB again showed the same accuracy rate of 94.29%. The average accuracy of K-NN on the MMN dataset was 93.14%. K-NN achieved a higher maximum accuracy, 97.14%, at k=(1 and 2), the minimum accuracy was recorded at k=(3, 7, 9, 10, 11, 12, 13, 15) with an accuracy of 91.43%. .

Highest performance: The highest accuracy performance is K-NN with k=(1 and 2) using dataset normalized using the Min-Max technique reaching 97.14%. It shows that using MMN normalization can provide higher accuracy.

2) Precision

Precision is an evaluation metric that measures the ability of the model to provide correct predictions for positive samples.

The results of the precision value analysis, as shown in Figure 4, can be summarized as follows:

Figure 4. Graph of precision values of NB and K-NN

RAW dataset: On the RAW dataset, the Naive Bayes (NB) algorithm obtained a precision rate of 94.71%. The average precision for K-NN was 84.41%, with the highest precision achieved at k=(1 and 2) of 94.71%. However, there is a significant decrease in the precision level by reaching the lowest value of 77.99% when the value of k= 11.

ZSN dataset When the ZSN dataset is used, NB produces the same precision rate as the RAW dataset, which is 94.71%. The average precision of K-NN on the ZSN dataset was 93.94%. The highest precision level achieved by K-NN is 96.49% with k=(3, 4, 12). The lowest precision level is 92.59% at k=(5, 6, 7, 8, 9, 13, 14, 15).

MMN dataset: In using the MMN dataset, NB again showed the same precision rate of 94.71%. The average precision of K-NN on the MMN dataset was 94.79%. K-NN achieved a higher maximum precision of 98.15% with k=(1 and 2). However, the minimum precision was recorded at k=(7, 9, 10, 11, 12, 13, 15) with a precision of 92.59%.

Highest performance: The highest precision performance was K-NN with k=(1 and 2) using a normalized dataset using the Min-Max technique with 98.15%.

3) Recall

Recall (sensitivity or true positive rate) is an evaluation metric measuring the model's ability to identify the most true positive samples.

Figure 5. Graph of recall values of NB and K-NN

The results of the recall value analysis, as shown in Figure 5, can be summarized as follows:

RAW dataset: On the raw dataset, the Naive Bayes (NB) algorithm obtained a recall rate of 94.71%. The average recall for K-NN was 84.99%, with the highest recall rate achieved at k=(1 and 2) of 94.71%. However, there is a signi cant decrease in the recall rate by reaching the lowest value of 75.78% when the value of k=13.

Dataset ZSN When the ZSN dataset is used, NB produces the same recall rate as the raw dataset, which is 94.71%. The average recall of K-NN on the ZSN dataset was 92.65%. The highest recall rate achieved by K-NN is 94.71%, with the value of k=(1, 2, 10, 11). The lowest recall rate is 91.37% at k=(5, 6, 7, 8, 9, 13, 14, 15).

MMN dataset: ˆ In using the MMN dataset, NB again showed the same recall rate of 94.71. The average recall of K-NN on the MMN dataset was 92.59%. K-NN achieved a higher maximum recall rate of 96.67% at k=1 and k=2. However, the minimum recall rate was recorded at k=3 with a recall rate of 89.17%.

Highest performance: The highest recall performance is K-NN with k=(1 and 2) using the normalized dataset using Min-Max technique at 96.67% This shows that using MMN normalization can provide higher Recall.

4) F1-Score

F1-score is an evaluation metric that combines Precision and Recall.

Figure 6. Graph of F1-score values of NB and K-NN

The results of the F1-Score value analysis, as shown in Figure 6, can be concluded that:

RAW dataset: On the raw dataset, the Naive Bayes (NB) algorithm obtained an F1-Score rate of 94.71The average F1-Score for K-NN was 84.69%, with the highest F1-Score level achieved at k=1 and k=2, which was 94.71%. However, there is a signi cant decrease in the F1-Score level by reaching the lowest value of 77.23% when the value of k=13.

ZSN dataset When the ZSN dataset was used, NB produced the same F1-Score rate as the raw dataset, which was 94.71%. The average F1-Score of K-NN on the ZSN dataset was 93.29%. The highest F1-Score level achieved by K-NN was 94.89%, at a number of k values, namely k=(3, 4, 15). The lowest F1-Score rate was 91.98% at k=(5, 6, 7, 8, 9, 13, 14, 15).

MMN dataset: In using the MMN dataset, NB again showed the same F1-Score level of 94.71%. The average F1-Score of K-NN on the MMN dataset was 93.14%. K-NN achieved a higher maximum F1-Score level of 97.40% at k=1 and k=2. The minimum F1-Score was recorded at k=(7, 9, 10, 11, 12, 13, 15) with an F1-Score of 91.98%.

Highest performance: The highest F1-Score performance was K-NN with k=(1 and 2) using normalized dataset using Min-Max technique at 97.40% This shows

that using MMN normalization can sometimes provide a higher F1-score.

The performance test results of NB and K-NN show that NB has a stable level of performance on all datasets, both normalized and non-normalized. The accuracy, precision, recall, and F1-score values achieved by NB are 94.29

On the other hand, K-NN performance varies depending on the k value and the dataset used. Dataset normalization has a signi cant impact on K-NN performance. On the raw dataset, K-NN achieved the same highest accuracy, precision, recall, and F1-score as NB at k= (1 and 2). However, on the MMN-normalized dataset, K-NN showed a signi cant performance improvement. The highest accuracy value achieved was 96.67%, while the highest precision was 98.15%, the highest recall was 96.67%, and the highest F1-score was 97.40%. These results show that dataset normalization can improve the performance of K-NN, especially MMN with k=(1 and 2), which produces the highest performance.

In the context of algorithm selection, the decision depends on the speci c situation and needs. If performance stability and no major changes are expected after dataset normalization are priorities, NB can be a suitable choice. However, if dataset normalization is required to improve performance, especially on MMN datasets with k= (1 and 2), K-NN may be a better alternative.

4. CONCLUSION AND FUTURE WORK

The optimization of the population data registration program, the provision of NIK, and the issuance of population registration and civil registration documents in villages through the implementation of mobile services or ball pick-up services, which face obstacles in determining which villages are eligible for services, can be resolved through the application of machine learning using classi cation techniques in data mining. From the results of classi cation experiments conducted in villages in Alor Regency, it can be concluded that K-NN has a reasonably high classi cation accuracy at k= 1 and k= 2 with datasets that have been normalized using MMN with an accuracy value of 97.14%. At the same time, NB is more stable in performance but has a slightly lower accuracy than K-NN, namely 94.29%.

The determination of these villages is expected to be extended to the sub-district and even the regency levels. By analyzing service history, distance, difficulty, and other attributes, it is possible to determine whether a sub-district or regency is eligible for funding and particular interventions to optimize its services.

References

[1] "Republic of indonesia law number 23 of 2006 on population administration (Undang-undang Republik Indonesia nomor 23 tahun 2006 tentang administrasi kependudukan)," Online, 2006, accessed on July 01, 2023. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/138575/permendagri-no-109-tahun-2019

[2] Ministry of Home A airs, "Minister of home aairs regulation number. 7 of 2019 on online administrative services for population administration (Peraturan Menteri Dalam Negeri Republik Indonesia nomor 7 tahun 2019 tentang pelayanan administrasi kependudukan secara dalam jaringan)," Online, 2019, accessed on July 01, 2023. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/127856/permendagri-no-7-tahun-2019

[3] ——, "Minister of home a airs regulation number 109 of 2019 regarding forms and books used in population administration (Peraturan Menteri Dalam Negeri Republik Indonesia nomor 109 tahun 2019 tentang formulir dan buku yang digunakan dalam administrasi kependudukan)," Online, 2019, accessed on July 01, 2023. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/138575/permendagri-no-109-tahun-2019

[4] Indonesiabaik.id. (2021) Self-service civil registry booth, making document printing for population registration easier (Anjungan Dukcapil Mandiri, Cetak Dokumen Kependudukan Jadi Makin Mudah). Accessed on July 01, 2023. [Online]. Available: https://indonesiabaik.id/infografis/anjungan-dukcapil-mandiri-cetak-dokumen-kependudukan-jadi-makin-mudah

[5] Ministry of Home A airs, "Minister of home aairs decision number 050-145 of 2022 regarding the granting and updating of codes, data on administrative regions, and islands for the year 2021 (Keputusan Menteri Dalam Negeri no 050-145 Tahun 2022 tentang Pemberian dan Pemutakhiran Kode, Data Wilayah Administrasi Pemerintahan, dan Pulau Tahun 2021)," Online, 2022, accessed on July 01, 2023. [Online]. Available: https://archive.org/details/kepmendagri-050-145-tahun-2022-lampiran%20Kepmendagri%20050-145%20Tahun%202022

[6] D. H. Jayani. (2022) Urban villages or rural villages according to topography of the region (2021) (Desa Kelurahan Menurut Topogra Wilayah (2021)). databoks.katadata.co.id. [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/03/24/lebih-dari-50-persen-desa-indonesia-di-wilayah-dataran

[7] Dirjen Dukcapil. (2022) On and on, Dukcapil's "picks up the ball" reaches Kera Island in Kupang (Terus dan Terus, "Jemput Bola" Dukcapil Jangkau Pulau Kera di Kupang). [Online]. Available: https://dukcapil.kemendagri.go.id/dhln/read/terus-dan-terus-jemput-bola-dukcapil-jangkau-pulau-kera-di-kupang

[8] RakyatNTT.com. (2022) Optimising services, TTS Dukcapil Ce 'picks up the ball' in villages (Optimalisasi Pelayanan, Dinas Dukcapil TTS 'Jemput Bola' di Desa-desa). [Online]. Available: https://rakyatntt.com/optimalisasi-pelayanan-dinas-dukcapil-tts-jemput-bola-di-desa-desa

[9] E. Houby and M. Enas, "A survey on applying machine learning techniques for management of diseases," Journal of Applied Biomedicine, vol. 16, no. 3, pp. 165–174, 2018. [Online]. Available: https://doi.org/10.1016/j.jab.2018.01.002

[10] S. Mishra, R. Kumar, S. K. Tiwari, and P. Ranjan, "Machine learning approaches in the diagnosis of infectious diseases: a review," Bull. Electr. Eng. Informatics, vol. 11, no. 6, pp. 3509–3520, 2022. [Online]. Available: https://doi.org/10.11591/eei.v11i6.4225

[11] M. M. Hossin, F. M. J. Mehedi Shamrat, M. R. Bhuiyan, R. Akter Hira, T. Khan, and S. Molla, "Breast cancer detection: an e ective comparison of different machine learning algorithms on the wisconsin dataset," Bull. Electr. Eng. Informatics, vol. 12, no. 4, pp. 2446–2456, 2023. [Online]. Available: https://doi.org/10.11591/eei.v12i4.4448

[12] W. N. L. Wan Hassan Ibeni, M. A. Mohd Salikon, Mohd Zaki, S. A. Daud, and M. N. Mohd Salleh, "Comparative analysis on bayesian classification for breast cancer problem," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1303–1311, December 2019. [Online]. Available: https://doi.org/10.11591/eei.v8i4.1628

[13] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis c virus infection," *Intelligent Medicine*, vol. 2, no. 4, pp. 193–198, 2022. [Online]. Available: https://doi.org/10.1016/j.imed.2021.12.003

[14] 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 Systems with Applications*, vol. 32, no. 2, pp. 625–631, 2007. [Online]. Available: https://doi.org/10.1016/j.eswa.2006.01.027

[15] 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," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 1082–1088, 2020. [Online]. Available: https://doi.org/10.25046/aj0506131

[16] M. Vaishnnave, K. S. Devi, P. Srinivasan, and G. A. P. Jothi, "Detection and classification of groundnut leaf diseases using knn classifier," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–5.

[17] A. Yudhana, D. Sulistyo, and I. Mufandi, "Gis-based and naïve bayes for nitrogen soil mapping in lendah, indonesia," *Sensing and Bio-Sensing Research*, vol. 33, p. 100435, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214180421000404

[18] A. Yudhana, R. Umar, and S. Saputra, "Fish freshness identification using machine learning: Performance comparison of k-nn and naïve bayes classifier," *Journal of Computing Science and Engineering*, vol. 16, no. 3, pp. 153–164, 2022. [Online]. Available: https://doi.org/10.5626/JCSE.2022.16.3.153

[19] A. Yudhana, I. Riadi, and F. Ridho, "Ddos classification using neural network and naïve bayes methods for network forensics," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2018.091125

[20] A. Kannagi, J. G. Mohammed, S. S. G. Murugan, and M. Varsha, "Intelligent mechanical systems and its applications on online fraud detection analysis using pattern recognition k-nearest neighbor algorithm for cloud security applications," *Materials Today: Proceedings*, vol. 81, pp. 745–749, 2023, international Virtual Conference on Sustainable Materials (IVCSM-2k20). [Online]. Available: https://doi.org/10.1016/j.matpr.2021.04.228

[21] S. Zhang, "Cost-sensitive knn classification," *Neurocomputing*, vol. 391, pp. 234–242, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.11.101

[22] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020. [Online]. Available: https://doi.org/10.1016/j.asoc.2019.105524

[23] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level k-nearest neighbor algorithm and support vector machine algorithm in classification water quality status," in *2016 6th International Conference on System Engineering and Technology (ICSET)*, 2016, pp. 137–141.

[24] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 7, pp. 58–63, 2017. [Online]. Available: http://computerscijournal.org/pdf/vol8no1/vol_8_no_01_13-19.pdf

[25] W. Hadi, Q. A. Al-Radaideh, and S. Alhawari, "Integrating associative rule-based classification with naïve bayes for text classification," *Applied Soft Computing*, vol. 69, pp. 344–356, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494618302564

[26] J. Wang, X. Jing, Z. Yan, Y. Fu, W. Pedrycz, and L. T. Yang, "A survey on trust evaluation based on machine learning," *ACM Computing Surveys*, vol. 53, no. 5, 2020. [Online]. Available: https://doi.org/10.1145/3408292

[27] M. Yucel, Z. Aslan, and M. Burunkaya, "Classification of the temperature-dependent gain of an erbium-doped fiber amplifier by using data mining methods," *Optik*, vol. 208, p. 164515, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402620303491

[28] M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, "Comparative analysis for topic classification in juz al-baqarah," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 406–411, 2018. [Online]. Available: http://doi.org/10.11591/ijeecs.v12.i1.pp406-411

[29] 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," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, no. 3, pp. 44–47, 2018. [Online]. Available: https://www.ijariit.com/manuscripts/v4i3/V4I3-1165.pdf

[30] A. P. Pawlovsky, "A knn method that uses a non-natural evolutionary algorithm for component selection," *Journal of Fundamental and Applied Sciences*, vol. 9, no. 4S, pp. 173–192, 2018.

[31] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. Abo-Elsoud, "A new covid-19 patients detection strategy (cpds) based on hybrid feature selection and enhanced knn classifier," *Knowledge-Based Systems*, vol. 205, p. 106270, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120304573

[32] Sıcakyüz and O. Yüregir, "Comparison of the accuracy of classification algorithms on three data-sets in data mining: Example of 20 classes," *International Journal of Engineering Science and Technology*, vol. 12, 09 2020.

[33] A. Yudhana, A. D. Cahyo, L. Y. Sabila, A. C. Subrata, and I. Mufandi, "Spatial distribution of soil nutrient content for sustainable rice agriculture using geographic information system and naïve bayes classifier," *International Journal of Smart Sensing and Intelligent Systems*, vol. 16, no. 1, pp. 1–14, 2023.

[34] N. Dengen, "Comparison performance of c4.5, naïve bayes and k-nearest neighbor in determination drug rehabilitation," in *Int. Conf. Sci. Inf. Technol.*, 2019, pp. 112–117.

[35] T. S. S. Rao and B. P. Battula, "A framework for hospital readmission based on deep learning approach and naive bayes