Instal : Jurnal Komputer

E-ISSN: 2808-683X
Edisi : Volume 15 Nomor 02 | 2023

Available online at https://journalinstal.cattleyadf.org/index.php/Instal/index

Prediction of Stunted Toddlers Using K-Nearest Neighbor Algorithm in Kamarang Lebak Village

Anggi Fitria Amida[1], Sandy Eka Permana[2], Denni Pratama[3],Khaerul Anam[4],Ade Rizki Rinaldi[5]
[1,2,3,4,5]STMIK IKMI Cirebon

ARTICLE INFO

Article history:

Received : 02 November 2023
Revised : 24 November 2023
Accepted : 31 December 2023

Keywords:
Data mining,
K-Nearest Neighbor,
Stunting

ABSTRACT

Stunting refers to a condition where toddlers (under five years old) experience growth failure, resulting in height and weight below the average for their age. The focus of this research is on the situation in Kamarang Lebak Village, where the number of stunted toddlers is notably significant. However, there has yet to be a study accurately predicting the factors differentiating stunted toddlers from those growing normally, thus lacking clarity on how accurate such predictions are in identifying toddlers vulnerable to stunting. The data collection method employed in this study involves observational techniques, with researchers visiting the Kamarang health center in Greged Sub-District, Cirebon Regency, to gather necessary information and data. This research implements the K-Nearest Neighbor Algorithm method to predict stunted toddlers and is supported by the Knowledge Discovery in Database approach, involving steps such as data selection, collection, transformation, data mining processes, and evaluation. It is anticipated that this research will serve as a foundation for public health practitioners, especially community health workers and village midwives in the area, to plan more focused and efficient intervention programs addressing toddler stunting issues. The results of this study indicate that the K-nearest neighbor algorithm demonstrates good performance with an accuracy of 97.16%. Stunting precision reaches 95.60%, normal precision reaches 98.82%, stunting recall reaches 98.86%, and normal recall reaches 95.45%.

Corresponding Author:
Anggi Fitria Amida
STMIK IKMI Cirebon
Email: anggiamida19@gmail.com

INTRODUCTION

Over the past few decades, rapid advancements in the field of Information Technology have fundamentally altered the way we interact with the world around us. Information technology has influenced various aspects of life, including technology itself, business, education, and healthcare[1]. Data mining, which is the process of extracting valuable information from various data sources, has played a key role in this revolution[2]. Stunting, referring to the

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

condition of inhibited physical growth and development in toddlers, is a serious global health issue that affects millions of children worldwide. Addressing stunting involves highly complex factors, including nutrition, environment, social, and health aspects. Hence, data mining becomes relevant as it can identify patterns and key factors associated with stunting by integrating various data sources such as health records, nutritional data, environmental data, and social factors [3]. One of the causes of stunting is the level of education of parents. If the education level of both parents, father and mother, is higher, the risk of children experiencing stunting can decrease by about 3-5%. Although the level of education has the potential to influence the occurrence of stunting, its impact is not always significant [4]. In this case, data mining offers a potential solution to unravel this complexity and generate valuable insights in efforts to address stunting [5]. This research will employ data mining techniques, specifically the K-Nearest Neighbor method, to predict stunting in toddlers based on parameters such as gender, birth weight, birth height, basic immunization status, nutritional status, and parental education level. Additionally, a relevant Knowledge Discovery in Database (KDD) approach will be applied to process classification outcomes and extract useful insights from the analyzed data. This method is a proven data mining technique effective in classifying based on data characteristic similarities[6]. With this specific objective, this research aims to predict the occurrence of stunting in toddlers based on an analysis of stunting causative factors and optimize child health programs. The study intends to provide a deeper insight into the issue of stunting in toddlers, offer more effective guidelines for interventions, and contribute to the fields of science and informatics by integrating information technology in the comprehensive understanding and management of child health issues

METHODS

The research method in this study is structured as follows:

1. Problem Identification
   Conducting direct observations on the issues surrounding stunting cases in Kamarang Lebak Village.
2. Setting Literature Review
   The literature review in this research involves exploring and understanding the problems and solutions found through references from journals or articles published within the last five years. The aim is to develop effective methods in addressing the specific issues that are the focus of the study.
3. Data Selection
   The data collection technique used in this research involves conducting the data collection process directly or through primary means.
4. Descriptive Analysis.
   This consists of data selection, pre-processing, data transformation, data mining, and evaluation.
5. Results Analysis
   Evaluating the outcomes to assess the accuracy resulting from the implemented data mining process.

Data Mining

This stage involves the data analysis process used to predict potential situations or conditions based on available information and data [7]. In the context of this research, the data mining technique used is classification utilizing RapidMiner and employing the K-Nearest Neighbor Algorithm. At this step, the researcher conducts two processes: performing classification and testing or evaluating the results.

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

K-Nearest Neighbor.

The K-Nearest Neighbor (k-NN) method is an algorithm used to classify new objects based on the values of k nearest neighbors [8]. k-NN is utilized in classification approaches in a simple and efficient manner. The fundamental concept of k-NN involves finding the closest distance among evaluated data and a specified number of K neighbors.

The applied model equation for the K-Nearest Neighbor algorithm is presented below :

$$D = \sqrt{(x1 - y1)^2 + (x2 - y2)^2}$$

Description :

D represents training data

X represents testing data

Y represents distance

$$fknn(x') = \frac{1}{k}\sum_{i \in N_{k(x')}} y_i$$

Description :

X' is the estimation or prediction

K is the number of nearest neighbors taken

Nk(_x') is the identified nearest neighbors

Yi a is the output of those nearest neighbors [9].

Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) is where processed data is visualized to be more user-friendly and is expected to enable users to take actions based on the analysis. This facilitates the presentation of data mining analysis results, making them more easily understandable.

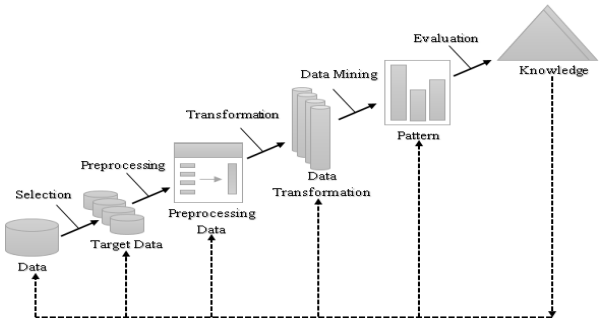

Figure 1. Stages of the KDD Method [10]

Evaluation

After applying the K-NN algorithm to classify data, the next step is to evaluate the final results. The evaluation of the K-NN algorithm is carried out using a confusion matrix to obtain accuracy, precision, and recall rates [11]. The higher the accuracy value, the better and more precise the model is in performing classification [12].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

$$Recall = \frac{TP}{TP + FN}$$

The data used in this research is sourced from the Toddler Data of Kamarang Lebak Village for the years 2018-2022. The selection of accurate and relevant data sources will be key to understanding the risk factors of stunting. The dataset consists of 1037 records and 13 attributes. The attributes in this dataset include No, Child's Name, Gender, Age (months), Birth Weight, Birth Height, Whether the Child has Completed Immunization, Parent's Name, Parent's Educational Level, Integrated Health Service Post (Posyandu), Village, Subdistrict, and District, as indicated in the following image.

DESKRIPSI ANAK	Nama Anak	Jenis Kelamin	Usia (bulan)	Berat Saat Lahir	Tinggi Saat Lahir	Apakah Anak Melakukan Imunisasi Dasar Lengkap (Ya = 1, Tidak = 0)	Riwayat penyakit yang pernah diderita (Hepatitis B) 1. Ya 2. Tidak	Riwayat penyakit yang pernah diderita (Difteri) 1. Ya 2. Tidak	Riwayat penyakit yang pernah diderita (Demam berdarah) 1. Ya 2. Tidak	Riwayat penyakit yang pernah diderita (Campak) 1. Ya 2. Tidak	Disabilitas (Tidak = 0, Tunanetra = 1, Tunarungu = 2, Tunawicara = 3, Cacat ganda (fisik dan mental) = 4	Status Gizi (N=Normal, S=Stunting)	DESKRIPSI ORANGTUA	Nama Orang Tua	Pendidikan Orang tua: 1. Tidak Tamat Sekolah 2. SD dan sederajat 3. SMP dan sederajat 4. SMA dan sederajat 5. Diploma 1-3 6. S1 dan sederajat 7. S2 dan sederajat 8. S3 dan sederajat 9. Pesantren, seminari, wihara dan sejenisnya 10. Lainnya	No. Wil:	Jaminan sosial kesehatan: 1. Peserta 2. Bukan peserta	DESKRIPSI ALAMAT	Posyandu	Desa	Kecamatan	Kabupaten
DESKRIPSI ANAK	DEVANO	L	42	2.2	49	0	2	2	2	2	0	S	DESKRIPSI ORANGTUA	ROSIN	2	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	DEYA	P	36	3.4	49	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	endang/aisah	2	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	GIBRAN	L	39	3.6	49	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	aan/lis	4	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	KHALEED	P	20	3.3	47	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	saepul/widia	2	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	KIRANA	P	59	3.6	48	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	Candika neng	2	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	LINDA	P	44	2.3	45	1	2	2	2	2	0	S	DESKRIPSI ORANGTUA	jaja/imas	4	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	M. KHOLID	L	52	3.1	46	0	2	2	2	2	0	N	DESKRIPSI ORANGTUA	masdi/minin	4	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	NIZAM	L	55	2.8	49	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	Jaja nur	2	0	2	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	NOVAL	L	45	2.3	39	1	2	2	2	2	0	S	DESKRIPSI ORANGTUA	RIKI	2	0	1	DESKRIPSI ALAMAT	Dadali	Kamarang Lebak	Greged	Cirebon
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
DESKRIPSI ANAK	MIMA SRI AYU	P	19	2.8	49	1	2	2	2	2	0	N	DESKRIPSI ORANGTUA	MUSA / SAROH	3	0	1	DESKRIPSI ALAMAT	Camar	Kamarang Lebak	Greged	Cirebon
DESKRIPSI ANAK	M. RIZKI FAUZAN	L	31	2.1	49	0	2	2	2	2	0	N	DESKRIPSI ORANGTUA	JAHIDIN / IIS	3	0	1	DESKRIPSI ALAMAT	Camar	Kamarang Lebak	Greged	Cirebon

Figure 2. Toddler Dataset of Kamarang Lebak Village 2018-2022

The data collection technique applied in this research is through observation. In this project, the researcher conducted direct visits to the Kamarang Health Center, Greged Subdistrict, Cirebon Regency, to gather information and data related to the toddler population in Kamarang Lebak Village. Subsequently, data was obtained from the Kamarang Health Center and utilized as the material for this research project.

RESULTS AND DISCUSSION

Data Selection

The first step is data selection. In this stage, the Read Excel operator is used to import the toddler dataset of Kamarang Lebak Village 2018-2022 from an Excel file. The next step involves using the Select Attributes operator, which plays a role in choosing or filtering attributes for analysis. Through the RapidMiner tools and using the Select Attributes operator, these attributes will be chosen by selecting a subset through parameters in the Select Attributes operator.

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

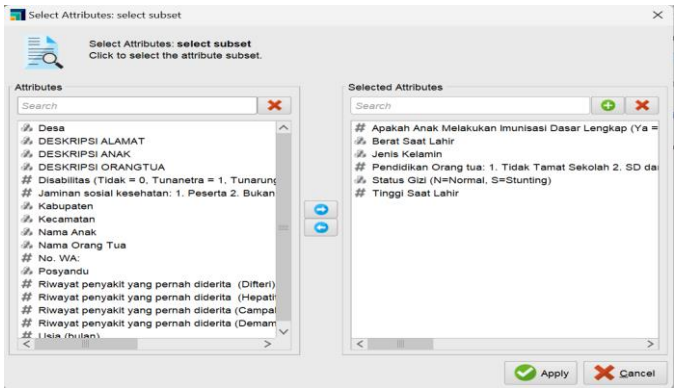

Figure 3. Select Attributes

At the beginning of the dataset, there are 13 attributes. After attribute selection, it is narrowed down to 6 attributes, namely gender, birth weight, birth height, nutritional status, whether the child has completed basic immunization, and parental education level.
The model process for the data selection stage in RapidMiner is shown in the image below.

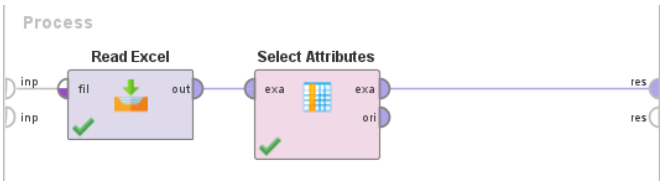

Figure 4. The model proses for the data selection

Pre-processing Data
The data pre-processing stage is a process of cleaning, organizing, arranging, and removing invalid or empty data to prepare the dataset for use in the data mining process. In this stage, two operators are used: replace missing values and set role.

1. Replace Missing Values Operator
   This operator aims to clean the data from null or missing values before proceeding to the data mining stage. Cleaning the data in this manner facilitates analysis and produces more accurate results.

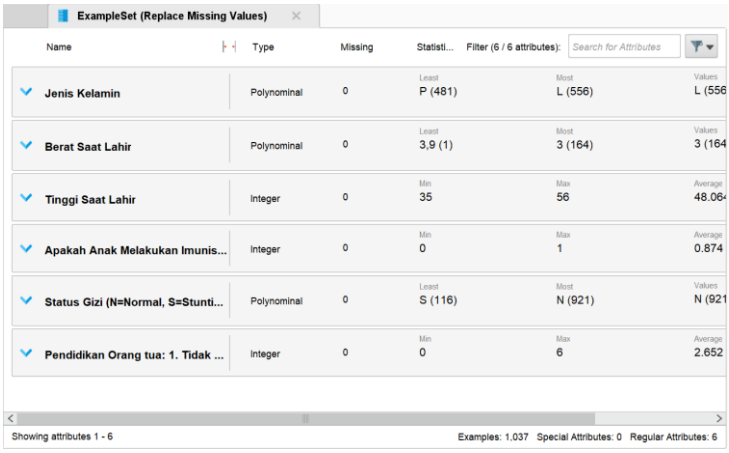

Figure 5. The result of the Replace Missing Values operator

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

After the Replace Missing Values operator completes the data cleaning, there are no more empty data, duplicate data, or inconsistent data in the dataset.

2. Set Role Operator
   The set role operator is used to transform regular attributes into special attributes, especially for attributes that play the role of labels.

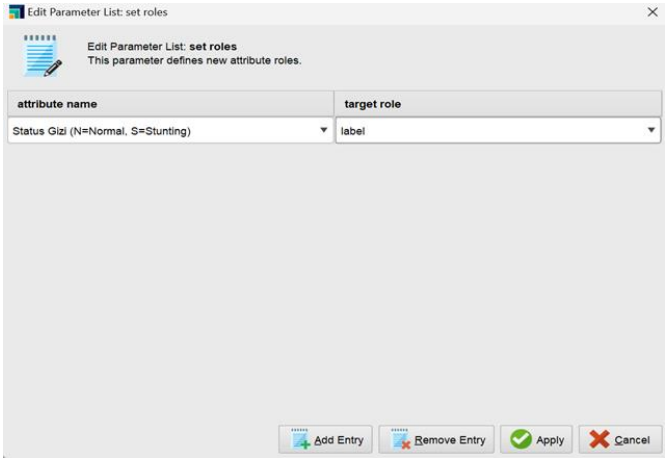

Figure 6. The parameter of set role operator

In this project, the determining attribute is the nutritional status of toddlers. Therefore, the nutritional status attribute needs to have its role transformed from a regular attribute to a special attribute, namely, the label. The model process for the data pre-processing stage in RapidMiner is shown in the image below.

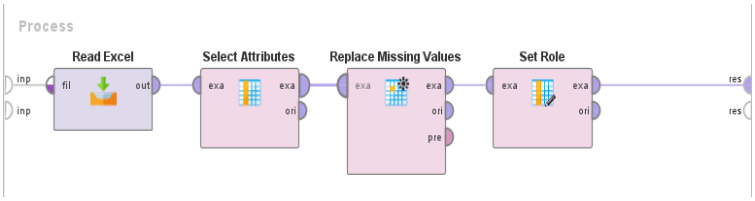

Figure 7. The model of the data pre-processing stage

Transformation Data

In the data transformation stage, the SMOTE Upsampling operator is employed. The SMOTE (Synthetic Minority Over-sampling Technique) Upsampling operator aims to address the imbalance between classes in the dataset.

Tabel 1. Comparison of data before and after upsampling.

SMOTE	Normal Toddlers	Stunted Toddlers
Before SMOTE	878	159
After SMOTE	878	878

The result of using the SMOTE Upsampling operator is an increase in the number of samples in the minority class, thus transforming imbalanced data into a more balanced state. After performing upsampling to balance the data, the next step is the use of the Split Data operator.

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

In the classification into two separate parts, namely data for training and data for testing. In the parameters of the Split Data operator, the sampling type is changed to automatic, and the data is partitioned with various ratios such as 70%:30%, 80%:20%, and 90%:10%. The model process for the transformation data stage in RapidMiner is shown in the image below.

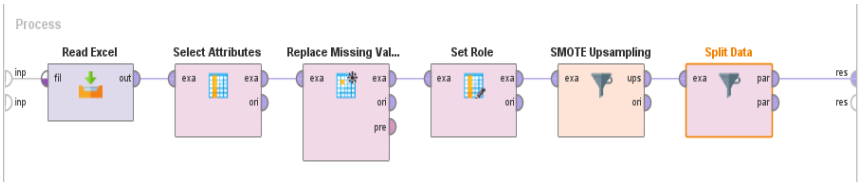

Figure 8. The model of transformation data stage

Implementation of the K-Nearest Neighbor Algorithm

The classification process in data mining using the k-nearest neighbor algorithm is designed using RapidMiner software. In this stage, several operators available in RapidMiner are used, namely:

1. k-NN Operator
   The algorithm applied in this research project is the K-Nearest Neighbor (KNN). After the previous dataset processing stage, the dataset will be processed using the K-Nearest Neighbor operator in RapidMiner.

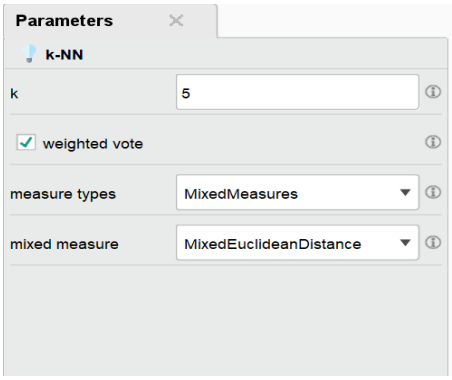

Figure 9. The parameter of k-NN operator

2. Apply Model Operator
   The Apply Model operator is a component that complements the K-Nearest Neighbor algorithm, designed to apply the previously learned model using training data to the testing data. This operator generates predictions for the labels in the testing data. In this data mining stage, the researcher conducted 7 experiments to find the number of nearest neighbors, including values for k = 3, 5, 9, 20, 25, 30, 35. The model process for the data mining stage in RapidMiner is shown in the image below.

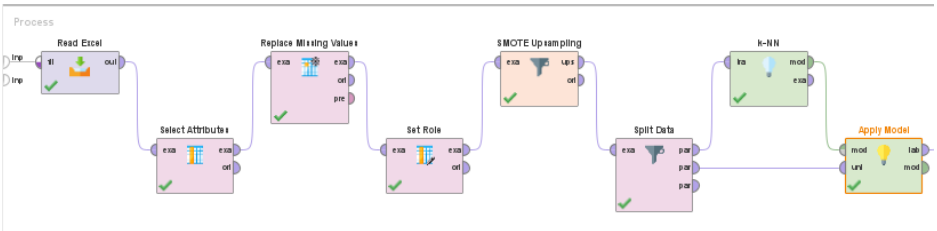

Figure 10. The model of data mining stage

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

Evaluation

In the evaluation process, the Performance (Classification) operator is used. The Performance (Classification) operator in RapidMiner is employed to obtain the accuracy level of the model on the classified data. The higher the accuracy value, approaching 100%, indicates that the model is performing better in predicting data.

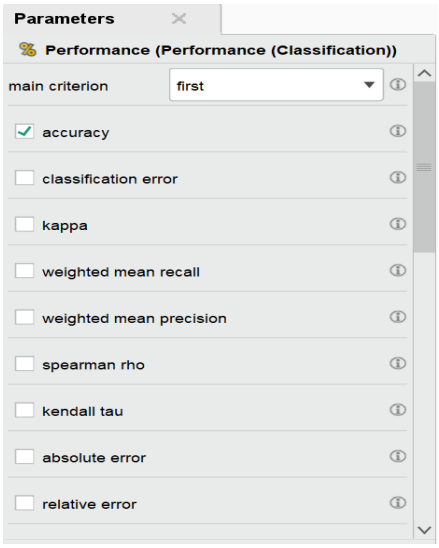

Figure 11. The parameter of performance (Classification) operator

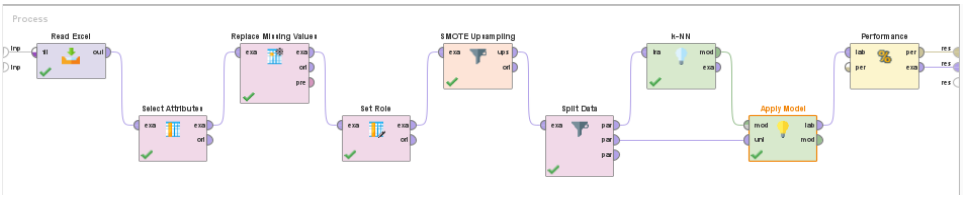

Figure 12. The Model of evaluation stage

Evaluation is conducted to determine the accuracy level in the classification process. After completing the data mining classification stage, the evaluation in this research involves calculating the values of accuracy, precision, and recall for various experiments with different values of K. Tables 4.2 to 4.4 display the accuracy, precision, and recall results for different values of K and three different data split scenarios.

This information is summarized in figures 4.43 to 4.45, which provide an overview of the evaluation results for three different data split ratios.

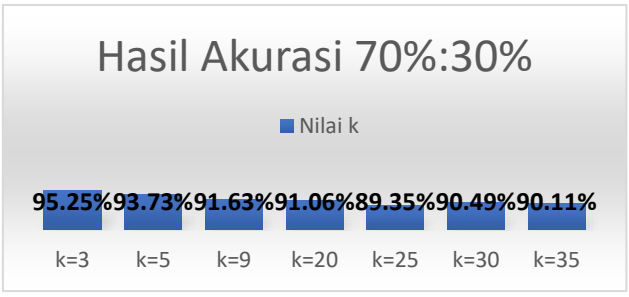

Figure 13. Accuracy results with a 70%:30% data split

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

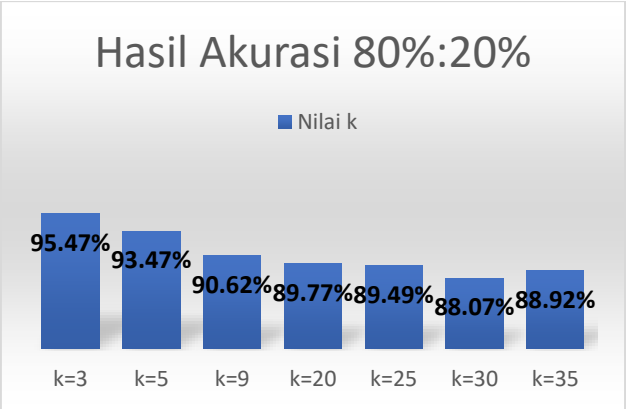

Figure 14. Accuracy results with a 80%:20% data split

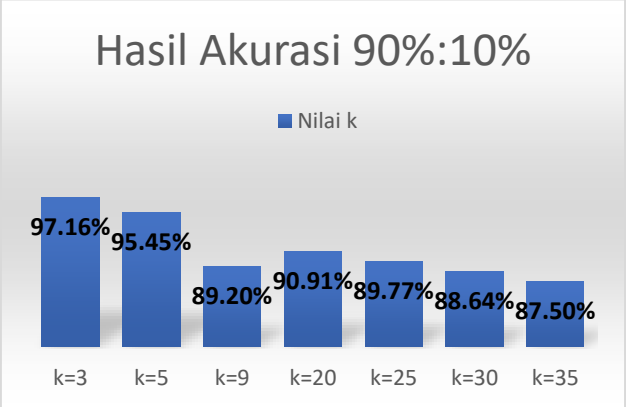

Figure 15. Accuracy results with a 90%:10% data split

Based on the chart above, the evaluation results for a 70%:30% data split with 1,228 training data and 526 testing data show that k=3 has the highest accuracy rate at 95.25%, with a precision of predicted S (stunting) at 93.75%, predicted N (normal) precision at 96.85%, recall true S (stunting) at 96.96%, and recall true N (normal) at 93.54%.
For an 80%:20% data split with 1,404 training data and 350 testing data, the evaluation results indicate that k=3 has the highest accuracy rate at 95.47%, with predicted S (stunting) precision at 93.48%, predicted N (normal) precision at 97.62%, recall true S (stunting) at 97.73%, and recall true N (normal) at 93.18%.
Meanwhile, the evaluation results for a 90%:10% data split with 1,578 training data and 176 testing data show that the highest k value is k=3 with an accuracy of 97.16%, predicted S (stunting) precision at 95.60%, predicted N (normal) precision at 98.82%, recall true S (stunting) at 98.85%, and recall true N (normal) at 95.45%.
The next step is to compare the evaluation results with the highest accuracy in each data split ratio. This step aims to determine the data split scenario with the highest accuracy values.

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

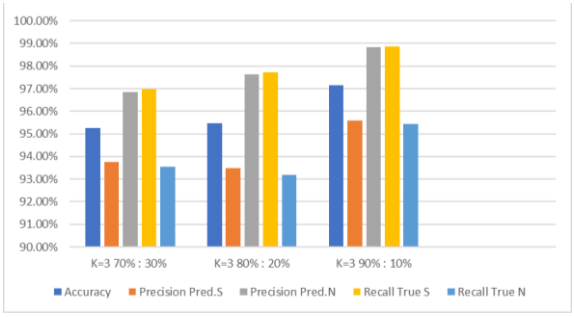

Figure 16. Comparison of accuracy results with the three data split ratios

Based on the comparison of classification tests conducted on the three data split scenarios, it can be concluded that the 90%:10% ratio with k=3 provides the highest accuracy level.

accuracy: 97.16%

	true S	true N	class precision
pred. S	87	4	95.60%
pred. N	1	84	98.82%
class recall	98.86%	95.45%	

Figure 17. Confusion matrix for k=3 with a 90%:10% data split

In figure 4.46, it is shown that the performance achieved by the K-NN algorithm is detailed as follows:

1. Accuracy reaches 97.16%
2. Precision results for predicting S (stunting) reach 95.60%, while Precision results for predicting N (normal) reach 98.82%.
3. Recall results for true S (stunting) reach 98.86%, while Recall results for true N (normal) reach 95.45%.

CONCLUSION

Based on the findings of this study, it can be concluded that using the k-nearest neighbor algorithm to predict stunted toddlers in Kamarang Lebak village yields good performance. In the evaluation results for the 90%:10% data split scenario, optimal performance was found with a 90%:10% data split ratio at k=3, achieving an accuracy of 97.16%. The precision for stunting is 95.60%, normal precision is 98.82%, stunting recall is 98.86%, and normal recall is 95.45%.

REFERENCES

[1] M. Jajuli, T. Hidayat, and Susilawati, "Clustering daerah rawan stunting di Jawa Barat menggunakan algoritma k-means Clustering stunting-prone areas in West Java using k-means algorithm," 2023, doi: 10.37373/infotech.v4i2.642.

[2] Ali, D. Ade Kurnia, M. A. Pratama, and F. Al Ma'ruf, "Klasifikasi Status Stunting Balita Di Desa Slangit Menggunakan Metode K-Nearest Neighbor," 2021, [Online]. Available: http://jurnal.kopertipindonesia.or.id/index.php/kopertip

[3] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," Hello World Jurnal Ilmu Komputer, vol. 2, no. 1, pp. 20–33, Mar. 2023, doi: 10.56211/helloworld.v2i1.230.

[4] Y. R. Rachman, A. P. N. Larassasti, A. S. Nanda, M. Rachsanzani, and R. Amalia, "HUBUNGAN PENDIDIKAN ORANG TUA TERHADAP RISIKO STUNTING PADA BALITA: A SYSTEMATIC REVIEW," vol. 2, no. 2, 2021.

Page : 345-355
Doi. 10.54209/jurnalkomputer.v15i02.156

[5] S. K. N. Julyantari, K. I. Budiarta, and K. D. M. N. Putri, "Implementasi K-Means Untuk Pengelompokan Status Gizi Balita (Studi Kasus Banjar Titih)," Jurnal Janitra Informatika dan Sistem Informasi, vol. 1, no. 2, pp. 92–101, 2021, doi: 10.25008/janitra.
[6] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 6, no. 1, p. 49, Jan. 2022, doi: 10.30865/mib.v6i1.3312.
[7] Faqih, T. Suprapti, and M. Hidayat, "IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOUR UNTUK PREDIKSI KETEPATAN KELULUSAN", [Online]. Available: https://ejournal.stmikgici.ac.id/
[8] H. Saleh, M. Faisal, and R. I. Musa, "KLASIFIKASI STATUS GIZI BALITA MENGGUNAKAN METODE K-NEAREST NEIGHBOR," vol. 4, no. 2, 2019.
[9] Colanus, R. Drajana, and A. Bode, "Prediksi Status Penderita Stunting Pada Balita Provinsi Gorontalo Menggunakan K-Nearest Neighbor Berbasis Seleksi Fitur Chi Square," Jurnal Nasional Komputasi dan Teknologi Informasi, vol. 5, no. 2, 2022.
[10] T. Hartati and W. Arie, "ANALISIS DATA LALU LINTAS JARINGAN DI KANTOR CANGEHGAR CYBER OPERATION CENTER MENGGUNAKAN ALGORITMA K-MEANS NETWORK TRAFFIC DATA ANALYSIS AT CANGEHGAR CYBER OPERATION CENTER OFFICE USING K-MEANS ALGORITHM," 2022.
[11] N. P. Fajarini, I. M. I. Subroto, and A. Riansyah, "Klasifikasi Kepakaran Reviewer Menggunakan Algoritma K-Nearest Neighbor (KNN)," 2022.
[12] H. M. Diki and N. Sari, "Penerapan Algoritma K-Nearest Neighbor Penerapan Algoritma K-Nearest Neighbor dalam Klasifikasi Judul Berita Hoax," 2022.