ABSTRACT

The toddler age represents a period of significant growth and development for a child. The 2022 Indonesian Nutrition Status Survey (SSGI), conducted by the Central Java Health Office, reports a prevalence of stunting in Temanggung at 28.9%, equating to 29 stunted toddlers out of 100. This figure is higher than the average stunting prevalence in Central Java, which stands at 20.8%. The survey also indicates that the prevalence of stunting in toddlers aged 0-1 years is higher than the overall prevalence. Consequently, it is estimated that the number of toddlers aged 0-1 years affected by stunting in Temanggung Regency over the past three years (2021-2023) ranges from 15,000 to 20,000 toddlers. The objective of this study is to ascertain the nutritional status of toddlers by utilizing the K-Nearest Neighbors (KNN) algorithm and to evaluate its accuracy in making such determinations. The results of the test on 3,999 data points indicate that the KNN algorithm can accurately classify the stunting status of children based on age and height with the highest accuracy and the smallest error rate. This is observed at a k-value of 5, with an accuracy rate of 0.999.

Keywords: Detection, K-Nearest Neighbor (KNN), Artificial Intelligence, Stunting, Temanggung.

1. INTRODUCTION

Toddler age is a crucial period in the process of child growth and development, where all children's cognitive abilities develop according to their age stages (Sufa et al., 2023). If toddlers experience nutritional problems, they will automatically face problems in growth and development, often get sick, and if left unchecked can cause death (Kaesmitan & Johannis, 2017). The age range of toddlers who are most vulnerable to stunting in Indonesia is between 0-5 years, with the most critical period occurring at the age of 0-2 years, which is often referred to as the first 1000 days of life (Priyono, 2020). During this period, children experience rapid growth and development, so adequate nutritional intake is very important. Malnutrition in this period can cause irreparable damage to physical growth and brain development (Khulafa’ur Rosidah & Harsiwi, 2019). Based on data from the 2018 Basic Health Research (Riskesdas), the prevalence of stunting in toddlers in Indonesia reached around 30.8%, which indicates that almost one in three toddlers is stunted (Ramadhan, 2019).
According to data from the Central Java Health Office in 2022 on the results of the Indonesian Nutrition Status Survey (SSGI), the prevalence of stunting in Temanggung reached 28.9%, meaning that out of 100 toddlers, 29 toddlers were stunted (BPS, 2022). This figure is higher than the average stunting prevalence in Central Java of 20.8%. Based on the data from the Indonesian Nutrition Status Survey (SSGI) the prevalence of stunting in toddlers 0-1 years is higher than the overall prevalence of stunting, it is estimated that the number of toddlers 0-1 years affected by stunting in Temanggung Regency in the last 3 years (2021-2023) ranges from 15,000 to 20,000 toddlers (Tim Percepatan Penurunan Stunting Jateng, 2023).

The Temanggung District Government continues to strive to reduce stunting rates, including among 0–1-year-olds: Increasing the coverage of exclusive breastfeeding for the first 6 months of life, providing nutritious and balanced complementary foods (MPASI), Providing complete immunization to toddlers, increasing access to health services for toddlers, Providing education to parents about the importance of nutrition and child health. With continuous efforts, it is hoped that the prevalence of stunting in children under 0-1 years in Temanggung District can be reduced and all children under five can grow optimally (Lukito & Setyaningsih, 2023).

Based on the above problems, researchers will detect the nutritional status of stunted toddlers in Temanggung Regency through the K-Nearest Neighbor (KNN) algorithm approach. Currently, the determination of malnutrition status is still done by physical observation and recording in the KIA book. Some of the parameters used to determine the nutritional status of stunted toddlers include toddler age and height (Murti et al., 2020). The purpose of this study is to determine the accuracy produced by the K-Nearest Neighbor (KNN) algorithm in classifying stunted toddlers based on these characteristics. The KNN algorithm was chosen because of its simple computation to process a lot of data. In addition, KNN also produces high accuracy when the K value is chosen correctly because KNN calculates the shortest distance from the test sample to the training sample without taking into account the distribution of each class.

Many previous researchers have discussed stunting. Stunting can occur in infants who are tall and have poor parenting, so babies receive less energy and nutrient intake (Saeful Bachri & Herdian Bhakti, 2021). In addition, pregnant women who suffer from chronic energy deficiency (CED) and anemia during pregnancy can also cause toddlers to experience stunting (Ruaida & Soumokil, 2018). In the computer field, determining the nutritional status of toddlers has been done using an expert system approach with certainty factor to identify severe malnutrition such as kwashiorkor, marasmus, and marasmic-kwashiorkor (Pantaleon et al., 2016). Based on the results of trials on 120 test data, the accuracy produced by the certainty factor has a percentage above 70% (Anggraeni & Syafrullah, 2023). Naïve Bayes has also been used to determine malnutrition status and the percentage probability of its assessment through 3 types of malnutrition status and 24 types of symptoms (Wati & Sudrajat, 2022). The grouping of nutritional values of toddlers has also been calculated with k-means to group toddlers into malnutrition, undernutrition, good nutrition, and obesity (Irfiani & Rani, 2018). In addition, Fuzzy KNN has also been used to classify the nutritional status of toddlers through measurements of health status, parental education, parental knowledge, genetics, and parental income. The accuracy produced by this algorithm is 84.37 (Nugraha et al., 2017).
2. THEORY

Stunting describes chronic malnutrition during growth and development from childhood (Loka & Marsal, 2023). Globally, about one in four children under the age of five is stunted (Hastuty, 2020; Husna et al., 2023; Hatijar, 2023). Early malnutrition increases infant and child mortality and makes them susceptible to disease, while their posture may not be optimal as adults (Lecerf & Andreelli, 2022; Munthe, 2022). Cognitive abilities may also be affected, resulting in long-term economic losses for Indonesia. Stunting is more common in children aged 12-59 months than in those aged 0-24 months (Wulandari & Arianti, 2023). Short-term effects include impaired cognitive, motor, and language development, and increased risk of disability, infectious disease, and even death. Long-term effects include the risk of degenerative diseases such as hypertension, diabetes mellitus, coronary heart disease, and stroke, which can affect work productivity in adulthood (Siswati, 2018).

Research on stunting in Indonesia continues to show a significant problem, with prevalence remaining high: 27.7% in 2019, 24.4% in 2020, 24.4% in 2021, and 21.6% in 2022 (Rahman et al., 2023). Stunting, which indicates a lack of linear growth in children, occurs mainly in children under five years of age due to long-term chronic malnutrition. Contributing factors include inadequate nutrient intake during pregnancy and childhood, suboptimal parenting, and limited access to adequate health and sanitation services (Maryati et al., 2023).

In Temanggung Regency, Central Java, the situation is not much different. Research over the past five years shows that the prevalence of stunting in this area remains a major health concern, with the following data for 2019: 28.9%, 2020: 28.9%, 2021: 28.2%, 2022: 25.9%, 2023: 20.5% (Nurmandhani et al., 2023). Factors such as family economic status, maternal education, and child-feeding practices strongly influence the incidence of stunting in the region (Nurmandhani et al., 2023). Efforts to reduce the prevalence of stunting in Temanggung District include more intensive intervention programs to improve under-five nutrition, education on healthy parenting, and improved access to quality health services (Hapsari, 2023).

2.1 K-Nearest Neighbor (KKN)

K-Nearest Neighbor (KNN) is a machine learning algorithm that loads all case data and classifies them into new case data based on proximity (Sumarlin, 2015; Zhang, 2021). This algorithm is included in the supervised learning type because the algorithm works by connecting the available data patterns with new data sets to find a new pattern (Mukhlis et al., 2024). To calculate the distance, the KNN algorithm generally uses the Euclidean distance and the Manhattan distance. K-Nearest Neighbor (KNN) is a machine learning algorithm used for classification and regression (Peterson, 2009; Trivusi, 2023; IBM, 2024). This algorithm works by finding the $k$ nearest data points (neighbors) to the new data point that needs to be classified or whose value needs to be predicted, based on a specific distance metric such as Euclidean, Manhattan, or others. KNN is a simple and easy-to-implement algorithm, but it can become inefficient on very large datasets because it requires distance calculations from every data point in the training set each time a new data point needs to be predicted (Kramer & Kramer, 2013; Peterson, 2009; Steinbach & Tan, 2009). KNN is also sensitive to feature scales and requires
normalization or standardization of data for more accurate results (Adminlp2m, 2023). The steps to apply K-Nearest Neighbour are (Lestari, 2014):

1) Determine the parameters of the value of $k$. The value of $k$ can be chosen freely by the researcher. For an even number of classes, the value of $k$ must be odd to avoid the occurrence of the same number of distances or a tie in the classification process. On the other hand, for an odd number of classes, it is better to use an even value of $k$ (Iriantoro et al., 2018).

2) Calculate the distance between the test data to be classified and all the training data.
3) Give an order to the distance formed (sort up / distance is sorted from the smallest value to the largest value)
4) Determine the closest distance to the order of $k$ values.
5) Determine the matching class (determine the class classification of the test data based on the training data)
6) Find the number of classes with the closest neighborhood, then assign this class as the data class to be evaluated.

### 2.2 Euclidean Distance

Euclidean distance is the most popular or commonly used formula for measuring distance. The smaller the Euclidean distance, the more similar the two attributes are. This formula, commonly referred to as linear distance, uses the equation (Salsabila et al., 2024a) (1):

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n}(x_{il} - x_{jl})^2}$$

Description:
- $d(x_i, x_j)$: Manhattan Distance between $i$ object and $j$ object
- $x_{il}$: Value of the $i$ object on the $l$ variable
- $x_{jl}$: Value of the $j$ object on the $l$ variable
- $n$: The number of observed variables

### 2.3 Manhattan Distance

The Manhattan distance is a distance measurement that adjusts the distance between two locations in a city (Black, 2019). To illustrate, if one were to move three blocks to the right and four blocks down, the total distance would be seven blocks. Manhattan distance is employed as a method to measure the distance obtained from the absolute difference between two data objects (Sutirta & Noviandi, 2024). Manhattan distance has the ability to calculate distances more accurately in cases where the points to be measured have the same or almost the same coordinates on their axes (Mu & Tong, 2020; Elen & Avuçu, 2021; Salsabila et al., 2024b). This distance measure uses the equation (2):

$$d(x_i, x_j) = \sum_{l=1}^{n} |x_{il} - x_{jl}|$$

Description:
- $d(x_i, x_j)$: Manhattan Distance between $i$ object and $j$ object
A confusion matrix is a test designed to predict the likelihood of an object being declared true or false (Townsend, 1971; Nerbonne & Heeringa, 1997; Strauss & Von Maltitz, 2017). The order of the test is presented in the form of a confusion matrix, with the predicted class presented at the top and the observed class presented on the left (Tharwat, 2021). Each cell contains a numerical value indicating the number of actual cases in the observed class that are to be predicted (Romadloni et al., 2022).

### Table 1. Confusion Matrix

<table>
<thead>
<tr>
<th>Class</th>
<th>Prediction</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>Yes</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>No</td>
<td>FP</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

Description:
- **TP**: *True Positive* (The number of correctly labeled positive tuples)
- **TN**: *True Negative* (The number of correctly labeled negative tuples)
- **FP**: *False Positive* (The number of negative tuples that have been mislabelled and subsequently identified as positive)
- **FN**: *False Negative* (The number of positive tuples that have been mislabelled and subsequently identified as negative.)

The calculation formula for the confusion matrix is as follows:

- **Specificity** is used in measuring the success rate of a method in correctly classifying negative classes, with equation (3), namely:
  \[
  \text{Specificity} = \frac{TN}{FP+TN}
  \]  
  \( (3) \)

- **Recall or Sensitivity** is used to measure how good a method is at correctly classifying positive classes, using equation (4) as follows:
  \[
  \text{Specificity} = \frac{TP}{FP+TN}
  \]  
  \( (4) \)

- **Accuracy** or Recognition rate is useful for measuring the performance of a method, using equation (5), namely:
  \[
  \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+TN}
  \]  
  \( (5) \)

### 3. METHOD

The research procedure, which was conducted in order to identify vulnerability in infants aged between 0 and 1 years old, who were affected by stunting, is presented in Figure 1.
At the outset of the research process, investigators will seek to identify any issues pertaining to stunting and the objective of the findings resulting from the analysis of the conducted research. The objective of this research is to ascertain the efficacy of the K-Nearest Neighbor (KNN) algorithm in the detection of stunted toddlers. The potential benefits of this research include the ability to assist health workers in Temanggung district in detecting stunted toddlers. Once the objectives and benefits have been established, the subsequent step is to analyze the parameters that will be employed to ascertain the nutritional status of the toddlers. This will entail determining whether the toddlers are malnourished, malnourished, or healthy. The parameters employed in this study are anthropometric measurements of toddlers, specifically the age and height of the infant.

The subsequent stage is the design of classification patterns using the K-nearest neighbor (KNN) approach. This is achieved by determining the nutritional status of toddlers based on TB/U, which can be categorized as very short, short, normal, or high. Before implementation, the system is tested using a confusion matrix to ascertain the resulting accuracy and error values. This serves to further ensure that the pattern formed is in accordance with the KNN classification model.

4. RESULTS AND DISCUSSION

Dataset utilised in this research is comprised of data obtained from the Temanggung District Health Office, Central Java Province, pertaining to the nutritional status of toddlers. The dataset is in the form of a file with a CSV extension, comprising a total of 3,998 data points and 1,094 instances of stunting in 1,094 toddlers. The data sample employed in this study comprises toddlers aged between 0 and 12 months. The data set is presented in Table 2.

Table 2. Data on the number of stunting patients in Temanggung Regency

<table>
<thead>
<tr>
<th>Age (months)</th>
<th>Gender</th>
<th>Height (cm)</th>
<th>Nutrition Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>male</td>
<td>44.591973</td>
<td>stunted</td>
</tr>
<tr>
<td>0</td>
<td>male</td>
<td>56.705203</td>
<td>tall</td>
</tr>
<tr>
<td>0</td>
<td>male</td>
<td>46.863358</td>
<td>normal</td>
</tr>
<tr>
<td>0</td>
<td>male</td>
<td>47.508026</td>
<td>normal</td>
</tr>
<tr>
<td>0</td>
<td>male</td>
<td>42.743494</td>
<td>severely stunted</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>female</td>
<td>51.126175</td>
<td>normal</td>
</tr>
<tr>
<td>1</td>
<td>female</td>
<td>64.374831</td>
<td>tall</td>
</tr>
</tbody>
</table>
Data preprocessing is an important step in knowledge discovery when the data obtained is incomplete (missing value), has noise (contains an error), or is inconsistent/different data format. The toddler nutritional status examination data used in this study does not have missing values so that it can proceed to the next stage, this is shown in Table 3.

Table 3. Checking missing value

<table>
<thead>
<tr>
<th>Age (months)</th>
<th>Gender</th>
<th>Height (cm)</th>
<th>Nutrition Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>44.591973</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>56.705203</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>46.863358</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>47.508026</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>42.743494</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>51.126175</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>64.374831</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>68.200071</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>60.757330</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>58.819450</td>
<td>2</td>
</tr>
</tbody>
</table>

There are a total of 3999 rows

Prior to the application of the K-Nearest Neighbour algorithm, it is necessary to divide the dataset into two distinct subsets: training data and testing data. The training data comprises independent variables, or variable X, which are attributes, and dependent variables, or variable Y, which are labels or classes. These are then taken to be the input. Concurrently, the testing data is employed to classify the labels/classes by calculating the distance to the independent variable/variable X (attribute). Prior to the division of the training and testing data, the data was randomized using Google Colab. Subsequently, the randomized dataset was divided into two distinct subsets, namely the training data and the testing data. The proportion of the training data to the testing data was set at 80:20. The results of the division of the testing data and training data can be observed in Figure 2.
Once all the requisite data has been gathered, the next step is to perform the K-Nearest Neighbour (KNN) calculation process. The objective of this study is to identify the optimal value of k, which will result in the highest accuracy and lowest error rate in detecting the nutritional status of toddlers. The optimal K value is presented in Figure 3.

The figure above illustrates that the optimal value of K, with the highest accuracy, is found at K = 5, with the smallest error occurring at an error value of 0.001. This allows for the precise representation of the accuracy value of 0.999 in the confusion matrix, as illustrated in Figure 4.
Figure 4 presents the results of the KNN testing, which can be elucidated through the use of a confusion matrix. For a more detailed analysis, the average values of precision, recall, and accuracy can be observed in Table 5.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND SUGGESTIONS
The results of the test on 3999 data points indicate that the KNN algorithm is capable of accurately determining the stunting status of children based on the height attribute (TB) with the highest accuracy achieved at the k value of 5. The test results indicate that the highest accuracy rate and the smallest error rate value are produced at the k = 5 value, which is 99.9% with an error rate of 0.001%.

REFERENCES
Genetika, 2(8), 2926–2933.


Psychophysics, 9, 40–50. https://doi.org/https://doi.org/10.3758/BF03213026