Optimizing Data Preprocessing on Stunting Datasets: Identifying Relevant Attributes for Machine Learning Analysis

Devi Sartika[1], Febie Elfaladonna[2], Ayu Octarina[3]

[,1,2,3,4] *Department of Informatics Management, Politeknik Negeri Sriwijaya, Palembang, Indonesia* febie_elfaladonna_mi@polsri.ac.id

Abstract. In Indonesia, approximately 8.9 million children are affected by stunting, with a prevalence rate of 30.8%. This issue is more prevalent in children over 12 months of age due to increased nutritional needs. Health centers, particularly the UPTD Puskesmas Mariana in Banyuasin Regency, play a crucial role in stunting monitoring efforts. However, data collection is still conducted manually, limiting efficiency and responsiveness. Additionally, low community awareness and limited skills among *posyandu* cadres in performing physical measurements contribute to the persistently high stunting rates. This study aims to preprocess data from 122 under-five observations collected in August 2024, with a primary focus on identifying relevant variables. The dataset includes 32 variables, all of which are outlier-free, allowing for more precise analysis. The results of this analysis are expected to highlight key factors that contribute to stunting and to inform more effective health policies and interventions. From this research, 29 attributes were selected after data preprocessing, and these new attributes are expected to aid in future cluster selection using K-means Clustering. This data will serve as a foundation for developing a toddler stunting monitoring application that incorporates K-means Clustering.

Keywords: *Stunting Dataset, Preprocessing Data, Relevant Variables*

1. Introduction

Globally, about one in four children under five are stunted, caused by chronic malnutrition and frequent exposure to disease during childhood. This condition can permanently hinder a child's physical and cognitive development, and has long-lasting effects [1].

Based on data from the World Health Organization (WHO) in 2020, the prevalence of stunting in children under five in Indonesia is in the second highest position in the Southeast Asia region, reaching 31.8%. The first position is held by Timor Leste with a prevalence of 48.8%, followed by Laos which recorded 30.2% and Cambodia with 29.9% [2]. In contrast, the lowest prevalence of stunting was found in Singapore with

N. L. Husni et al. (eds.), *Proceedings of the 8th FIRST 2024 International Conference on Global Innovations (FIRST-ESCSI 2024)*, Advances in Engineering Research 261,
https://doi.org/10.2991/978-94-6463-678-9_45

only 2.8%. Data from the 2016 Nutrition Status Determination (PSG) also shows a significant prevalence of stunting.

In Indonesia, around 8.9 million children are stunted, meaning that one-third of all children under five are below normal height. About 30.8% of children under five in the country are stunted, with a higher prevalence in children older than 12 months compared to those younger than 12 months. This is due to increased nutrient requirements as children age, which are necessary for the body's energy metabolism.

In 2022, the stunting rate in South Sumatra was recorded at 24.8%, but has now decreased to 18.6%. Nevertheless, there are five out of 17 districts and cities in South Sumatra that still have a stunting prevalence above 20%. One of them is Mariana Village in Banyuasin Regency, which reached 20.9%.

Community Health Centers (Puskesmas) play an important role in monitoring and addressing stunting at the community level. UPTD Puskesmas Mariana in Banyuasin Regency is one of the Puskesmas that is active in this effort. However, data collection and monitoring of stunting prevalence among under-fives is still often done manually, which is time-consuming and inefficient.

In addition, low public awareness of the dangers of stunting contributes to the high stunting rate in this kelurahan. Other obstacles hindering stunting prevention include the lack of regular socialization from health authorities and the absence of follow-up diagnoses to pediatricians or nutritionists. Complicated procedures, such as the need to use special referrals, are also a barrier. Posyandu cadres often lack the skills to accurately measure height, arm circumference, head circumference and weight. As a result, analyzing stunting conditions in children under five takes a long time. The stunting problem in this kelurahan is not only experienced by the low-income community, but also by the middle to upper class. Therefore, a more structured and efficient effort is needed to reduce the stunting rate in this area.

By involving the role of technology, developing an application to monitor the prevalence of stunting in toddlers using the K-Means Clustering machine learning algorithm is an innovative step that can improve the monitoring and handling of stunting at the community level, especially in UPTD Puskesmas Mariana, Banyuasin Regency. This approach is expected to make monitoring more effective, decision-making more accurate, and resource management more efficient in dealing with the problem of stunting in children under five years old.

Clustering is a data grouping technique aimed at discovering structures within a dataset without the need for prior labels, with K-Means being one of the most widely used algorithms. K-Means has several advantages, including its simplicity, which makes it easy to understand, and its efficiency in processing, making it well-suited for large datasets. With its intuitive concept, this algorithm allows for clear visualization of results, while its flexibility in determining the number of clusters (k) and distance metrics makes it adaptable to various types of data. Compared to other algorithms such as DBSCAN and hierarchical clustering, K-Means is generally faster and more stable.

Before conducting an in-depth analysis of the prevalence of stunting in children under five, the first step is data preprocessing to identify relevant variables. Dataset preprocessing is a series of steps aimed at cleaning and preparing data so that it is ready for analysis. [3]. The contribution of this research is to perform data preprocessing on the toddler stunting analysis dataset.

This process includes removing incomplete or inconsistent data, handling missing values, and transforming the data to ensure that the format and scale of the variables match the needs of the analysis [4]. By preprocessing the data, researchers can ensure that the data used is accurate and relevant so that analysis of the prevalence of stunting can be carried out effectively and generate useful insights for handling this problem.

The data used in this study includes under-five information from various villages in Mariana Regency with a total of one hundred and twenty-two data taken in August 2024.

2. Method

In this study, a series of systematic steps were conducted, starting with the data collection stage and ending with data interpretation by exploratory data analysis. The specific stages of this research are illustrated in Figure 1.

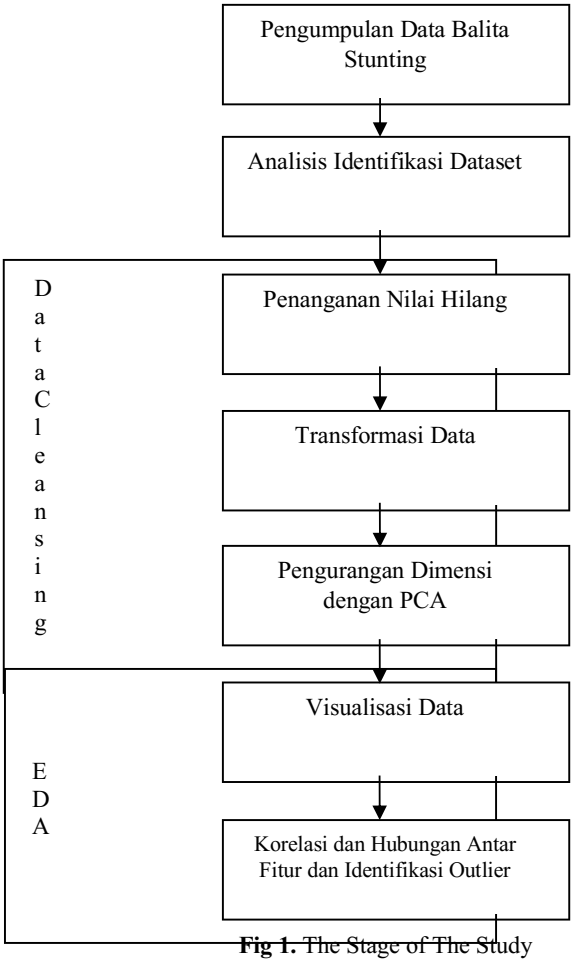


Fig 1. The Stage of The Study

The following is a technical explanation related to the stages contained in Figure 1.

2.1. Data Collection for Stunting Toddlers

In the data collection stage for stunting cases in Mariana Health Center, the first step is to determine the type of data needed for effective analysis. The data needed includes child health information such as height and weight, nutritional data on food consumption patterns, and socio-economic data such as family income and parental education level. In addition, it is important to collect information on access to health services, including the frequency of visits to health centers and the type of care received. Data can be obtained from health center medical records, community health surveys, and government data. This process should ensure that the data collected is high

quality, complete and relevant to build an effective machine learning model to address stunting.

In the dataset that has been obtained, there are several columns with a total of 122 rows. Demographic data includes information on age, gender, economic status, parental education, and more. Nutrition data includes breastfeeding, calorie intake, macronutrients (such as protein, fat, and carbohydrates), micronutrients (vitamins and minerals), and other information. Health Data includes disease history, immunization status, weight, height, and so on. Environmental data includes access to health facilities, sanitation conditions, and other relevant factors.

Figure 2 below displays the data that has been obtained from partners.

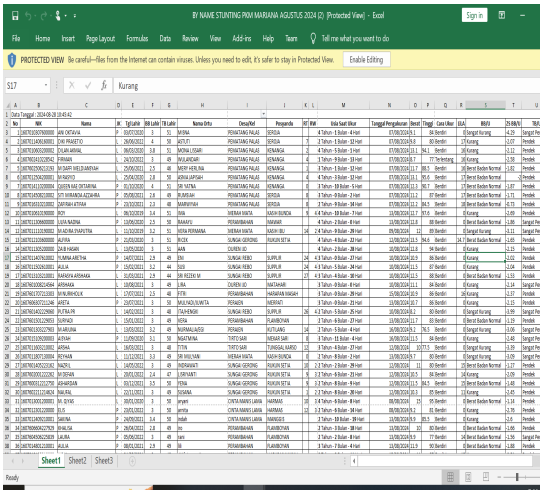


Fig 2. Data Stunting

2.2. Data Identification Analysis

Some preliminary analysis needs to be done to understand the condition of the dataset and identify potential problems. Typical analysis steps include examining the structure of the dataset, which includes looking at the dimensions of the dataset, the number of rows and columns, and the data type of each column to provide an overview of the data to be analyzed. Next, descriptive statistics are calculated, including the mean, median, mode, minimum, maximum and standard deviation for numerical variables, which helps in understanding the distribution of the data.

In the dataset provided by the partner, it can be seen that the total dataset is 122 rows with 38 attributes.

The attributes of the stunting dataset obtained are:

NIK, Name, Gender, Date of Birth, BW, TB, Village, Posyandu, RT, RW, Age at Measurement, Date of Measurement, Weight, Height, Measurement Method, LiLA, BB/U, ZS BB/U, TB/U, ZS TB/U, BB/TB, ZS BB/TB, Exclusive Breastfeeding, Appropriateness of MPASI Menu, Clean Water, Helminthiasis, Healthy Latrine, Immunization, Smoking, Birth, Participant's Disease, Micronutrient Adequacy, Family Head Occupation, Total Family Income (Month), Father's Education, Mother's Education, Father's Height, Mother's Height.

After conducting the analysis, it was found that there were some missing values and a number of attributes that were related to each other. The presence of these missing values can introduce bias in the analysis and affect the results obtained, so it is important to handle them properly. Missing values may be due to various factors, such as errors in data collection or lack of response from the respondents.

In addition, the interrelationships between attributes indicate significant relationships within the dataset, which can be utilized for more in-depth analysis. For example, attributes such as age are closely related to nutritional status or height, suggesting that understanding one attribute can provide insight into another. Recognizing these relationships is crucial for building more accurate and effective models, as well as formulating appropriate intervention strategies to address the stunting problem..

Therefore, the next step in the data cleaning process is to fill in or remove missing values, as well as use correlation analysis to identify interrelated attributes. With these steps, the dataset will become more comprehensive and informative, allowing for better analysis and more appropriate recommendations in addressing the issue of stunting in children under five years old.

2.3. Data Cleansing

Data cleansing, or data cleansing, is the process of detecting and correcting (or removing) corrupt or inaccurate datasets, tables, and databases. The term refers to the identification of incomplete, incorrect, imprecise, and irrelevant data. This “dirty” data will then be replaced, modified, or deleted as needed [5].

The data cleaning process is carried out using a library in python with Google Collab tools. There are several stages carried out in the data cleaning process, namely:

1. Missing Value Handling

Missing values are problems that arise when there are parts of the data that are incomplete or missing. The presence of missing values is common in datasets, and can be caused by a variety of factors, such as tool malfunctions, errors in calculations, missing data, and other technical issues [6].

In the stunting dataset, there are several columns that are manually deleted because the data in these columns is considered irrelevant. These columns are NIK and Nama. There are several reasons why NIK (Population Identification Number) and name are not included in the dataset, especially in the context of health or stunting analysis. First,

maintaining individual privacy is important, so personal information such as NIK and name are often omitted to prevent data misuse. Also, in certain analyses, the main focus is on variables that are directly related to the outcome.

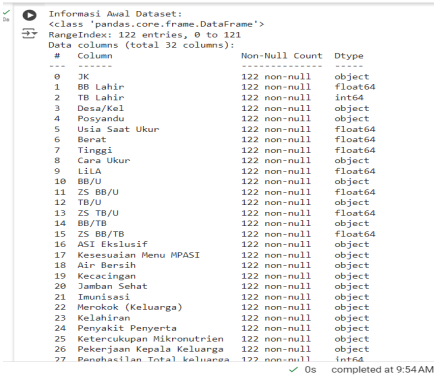
The following is a missing value analysis using the python language in Google Collab.

```python
import pandas as pd

# Membaca dataset
data = pd.read_csv('data_baru_ok.csv')

# Menampilkan informasi awal dataset
print("Informasi Awal Dataset:")
print(data.info())

# 1. Menampilkan persentase missing values per kolom
missing_values = data.isnull().mean() * 100
print("\nPersentase Nilai Hilang:\n", missing_values)

# 2. Mengisi missing values
# Mengisi nilai hilang dengan rata-rata untuk kolom numerik
for column in data.select_dtypes(include=['float64', 'int64']).columns:
    data[column].fillna(data[column].mean(), inplace=True)

# Mengisi missing values dengan median
# data['kolom_numerik'] = data['kolom_numerik'].fillna(data['kolom_numerik'].median())

# Mengisi missing values dengan modus (contoh untuk kolom kategori)
# data['kolom_kategori'] = data['kolom_kategori'].fillna(data['kolom_kategori'].mode()[0])

# 3. Menghapus baris dengan missing values (jika kritis)
# data.dropna(subset=['kolom_kritis'], inplace=True)

# 4. Menghapus kolom dengan persentase missing values di atas ambang batas tertentu
threshold = 0.3  # 30%
```

✓ 0s completed at 9:54 AM

Fig 3. Missing Value Handling

The results of the python code can be seen in Figure 4.



```
Informasi Awal Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122 entries, 0 to 121
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   JK                              122 non-null    object
 1   BB Lahir                        122 non-null    float64
 2   TB Lahir                        122 non-null    int64
 3   Desa/Kel                        122 non-null    object
 4   Posyandu                        122 non-null    object
 5   Usia Saat Ukur                  122 non-null    float64
 6   Berat                           122 non-null    float64
 7   Tinggi                          122 non-null    float64
 8   Cara Ukur                       122 non-null    object
 9   LiLA                            122 non-null    float64
 10  BB/U                            122 non-null    object
 11  ZS BB/U                         122 non-null    float64
 12  TB/U                            122 non-null    object
 13  ZS TB/U                         122 non-null    float64
 14  BB/TB                           122 non-null    object
 15  ZS BB/TB                        122 non-null    float64
 16  ASI Ekslusif                    122 non-null    object
 17  Kesesuaian Menu MPASI           122 non-null    object
 18  Air Bersih                      122 non-null    object
 19  Kecacingan                      122 non-null    object
 20  Jamban Sehat                    122 non-null    object
 21  Imunisasi                       122 non-null    object
 22  Merokok (Keluarga)              122 non-null    object
 23  Kelahiran                       122 non-null    object
 24  Penyakit Penyerta               122 non-null    object
 25  Ketercukupan Mikronutrien       122 non-null    object
 26  Pekerjaan Kepala Keluarga       122 non-null    object
 27  Penghasilan Total keluarga      122 non-null    int64
```

✓ 0s completed at 9:54 AM

Fig 4. Dataset Initial Information

```
Persentase Nilai Hilang:
 JK                                0.0
BB Lahir                           0.0
TB Lahir                           0.0
Desa/Kel                           0.0
Posyandu                           0.0
Usia Saat Ukur                     0.0
Berat                              0.0
Tinggi                             0.0
Cara Ukur                          0.0
LiLA                               0.0
BB/U                               0.0
ZS BB/U                            0.0
TB/U                               0.0
ZS TB/U                            0.0
BB/TB                              0.0
ZS BB/TB                           0.0
ASI Ekslusif                       0.0
Kesesuaian Menu MPASI              0.0
Air Bersih                         0.0
Kecacingan                         0.0
Jamban Sehat                       0.0
Imunisasi                          0.0
Merokok (Keluarga)                 0.0
Kelahiran                          0.0
Penyakit Penyerta                  0.0
Ketercukupan Mikronutrien          0.0
Pekerjaan Kepala Keluarga          0.0
Penghasilan Total keluarga         0.0
Pendidikan Ayah                    0.0
Pendidikan Ibu                     0.0
Tinggi Ayah (cm)                   0.0
Tinggi Ibu (cm)                    0.0
dtype: float64
```

Fig 5. Persentage Missing Value

```
Informasi Setelah Penanganan Missing Values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122 entries, 0 to 121
Data columns (total 32 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   JK                          122 non-null    object
 1   BB Lahir                    122 non-null    float64
 2   TB Lahir                    122 non-null    int64
 3   Desa/Kel                    122 non-null    object
 4   Posyandu                    122 non-null    object
 5   Usia Saat Ukur              122 non-null    float64
 6   Berat                       122 non-null    float64
 7   Tinggi                      122 non-null    float64
 8   Cara Ukur                   122 non-null    object
 9   LiLA                        122 non-null    float64
 10  BB/U                        122 non-null    object
 11  ZS BB/U                     122 non-null    float64
 12  TB/U                        122 non-null    object
 13  ZS TB/U                     122 non-null    float64
 14  BB/TB                       122 non-null    object
 15  ZS BB/TB                    122 non-null    float64
 16  ASI Ekslusif                122 non-null    object
 17  Kesesuaian Menu MPASI       122 non-null    object
 18  Air Bersih                  122 non-null    object
 19  Kecacingan                  122 non-null    object
```

Fig 6. Information after Missing Value Handling

After the missing values handling process, the dataset now consists of 122 entries with a total of 32 columns, all free of missing values. All columns have a total of 122 non-null values, indicating no more missing data. The data types in this dataset vary, including 8 columns with float64 types that include numeric data such as weight and height, 4 columns with int64 types for variables such as parents' age and height, and 20 columns with object types that contain categorical information such as gender and immunization status. Some important fields include demographic, health and access to resources information. With a memory usage of approximately 30.6 KB, this dataset demonstrates efficiency in storage. Overall, the cleaned dataset was ready for further analysis, with no missing values that could affect the results of the study. The dataset has also been saved under the name 'data_stunting_cleaned.csv' for use in the next steps of the analysis.

2. Data Transformation

Data that has several variables with varying ranges of values can cause variables with larger values or ranges to have a more significant impact on the cost function than variables with smaller values or ranges. This problem can be solved by rescaling the data, so that all variables have similar ranges [7]. The following image is a code snippet for the data transformation process.

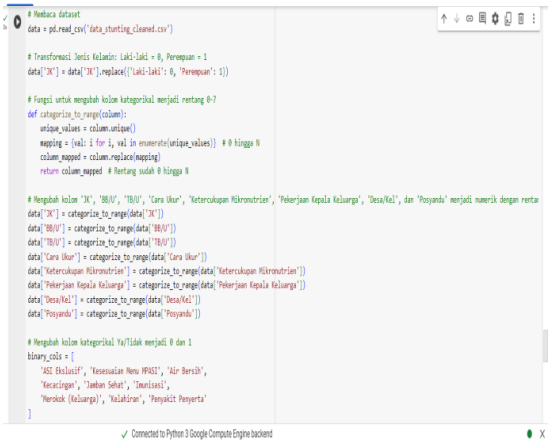
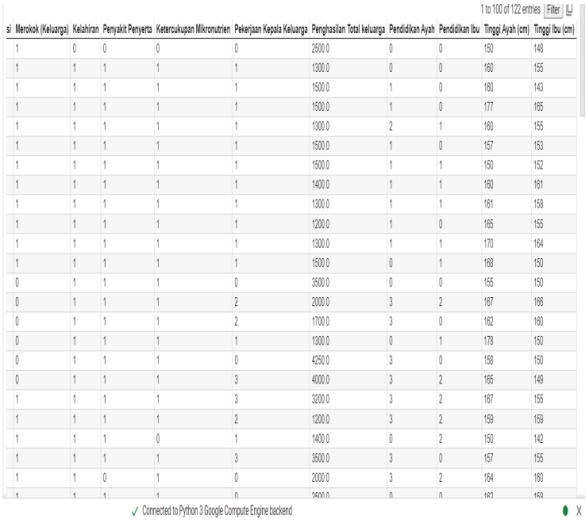


Fig 7. Data Transformation Code

This code aims to transform the stunting dataset. First, pandas and MinMaxScaler libraries from sklearn.preprocessing are imported to facilitate data processing and normalization. The dataset was read from a CSV file named 'data_stunting_cleaned.csv'. The categorize_to_range function was created to convert other categorical columns into numerical values in the range of 0 to N, corresponding to the number of unique categories. The columns include 'BW/U', 'TB/U', 'Measurement Method', 'Micronutrient Adequacy', 'Family Head Occupation', 'Village/Kel', and 'Posyandu'.

Furthermore, the columns containing the values 'Yes' and 'No' were converted to 1 and 0. Additional columns that were also categorized to a range of 0 to 7 include 'LBW', 'Clean Water', 'Immunization', 'Birth', 'Comorbidities', 'Father's Education', and 'Mother's Education'. In addition, the Total Family Income column was converted by dividing it by 1000 and rounding the result. Finally, the height of the father and mother was normalized so that the values ranged from 0 to 700. The transformed dataset was saved into a new CSV file named 'data_stunting_transformed_final.csv', and the program displayed a confirmation message on the screen.

The results of the data transformation can be seen in the following image.



1 to 100 of 122 entries | Filter

si	Merokok (Keluarga)	Kelahiran	Penyakit Penyerta	Ketercukupan Mikronutrien	Pekerjaan Kepala Keluarga	Penghasilan Total keluarga	Pendidikan Ayah	Pendidikan Ibu	Tinggi Ayah (cm)	Tinggi Ibu (cm)
1	0	0	0	0	2500.0	0	0	150	149	
1	1	1	1	1	1300.0	0	0	160	155	
1	1	1	1	1	1500.0	1	0	160	143	
1	1	1	1	1	1500.0	1	0	177	165	
1	1	1	1	1	1300.0	2	1	160	155	
1	1	1	1	1	1500.0	1	0	157	153	
1	1	1	1	1	1500.0	1	1	150	152	
1	1	1	1	1	1400.0	1	1	160	161	
1	1	1	1	1	1300.0	1	1	161	158	
1	1	1	1	1	1200.0	1	0	165	155	
1	1	1	1	1	1300.0	1	1	170	164	
1	1	1	1	1	1500.0	0	1	169	150	
0	1	1	1	0	3500.0	0	0	155	150	
0	1	1	1	2	2000.0	3	2	167	169	
0	1	1	1	2	1700.0	3	0	162	160	
0	1	1	1	1	1300.0	0	1	178	150	
0	1	1	1	0	4250.0	3	0	159	150	
0	1	1	1	3	4000.0	3	2	165	149	
1	1	1	1	3	5200.0	3	2	167	155	
1	1	1	1	2	1200.0	3	2	159	159	
1	1	1	0	1	1400.0	0	2	150	142	
1	1	1	1	3	3500.0	3	0	157	155	
1	1	0	1	0	2000.0	3	2	164	160	

✓ Connected to Python 3 Google Compute Engine backend

Fig 8. Data Transformation Results

3. Dimensionality Reduction with PCA

Dimensionality reduction is the process of reducing the number of features or dimensions in a data set without sacrificing the important information contained in it [8]. Dimensionality reduction can be done through two approaches, namely feature selection and feature extraction. In this research, the method used is feature extraction, where new features are generated from the initial data set. One of the feature extraction techniques that can be applied is Principal Component Analysis (PCA). PCA is a method that aims to simplify data through linear transformation so that a new coordinate system is formed with minimized variance. PCA enables data dimensionality reduction with very little risk of information loss [9].

Here is the PCA code to reduce the dimension of the dataset.

```python
# 1. Standarisasi Data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)

# 2. Penerapan PCA
pca = PCA()
pca.fit(scaled_data)

# 3. Varians yang Dijelaskan
explained_variance = pca.explained_variance_ratio_
print("Rasio Varians yang Dijelaskan:", explained_variance)

# 4. Membuat DataFrame untuk Komponen Utama
pca_components = pca.transform(scaled_data)
pca_df = pd.DataFrame(data=pca_components, columns=[f'PC{i+1}' for i in range(pca.n_components_)])

# 5. Visualisasi Varians yang Dijelaskan
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(explained_variance) + 1), explained_variance, marker='o', linestyle='--')
plt.title('Varians yang Dijelaskan oleh Setiap Komponen Utama')
plt.xlabel('Komponen Utama')
plt.ylabel('Rasio Varians yang Dijelaskan')
plt.xticks(range(1, len(explained_variance) + 1))
plt.grid()
plt.show()

# 6. Loading Scores
loading_scores = pca.components_.T * np.sqrt(pca.explained_variance_)
```

✓ 0s completed at 12:32 PM

Fig 9. PCA Code

The PCA results can be seen in Figure 10 below.

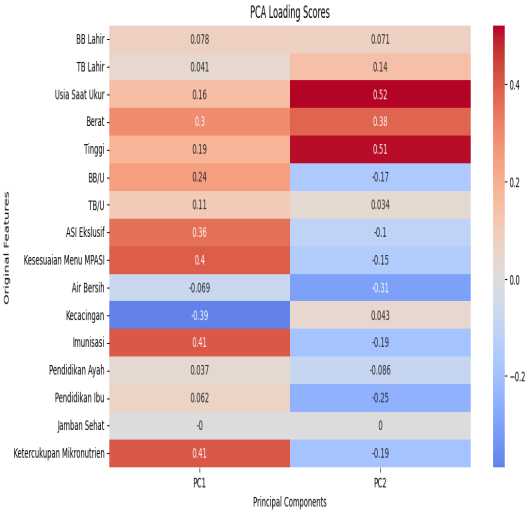


Fig 10. PCA Loading Scores

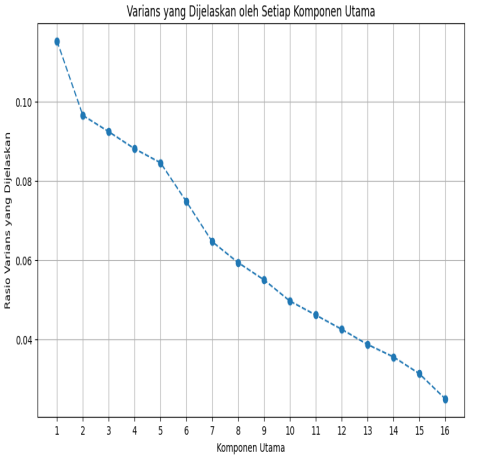


Fig 11. Variants based on principal components

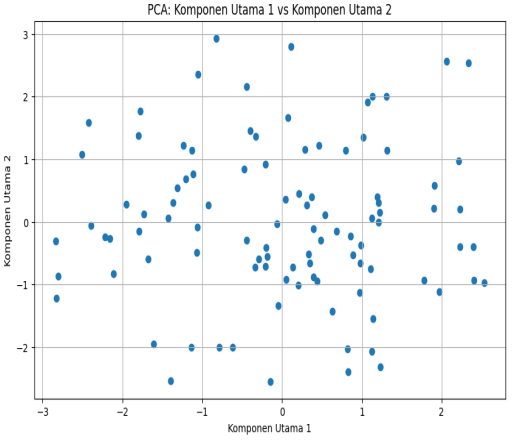


Fig 12. Component 1 vs Component 2

The following are the results of the loading scores:

Loading Scores:

	PC1	PC2	PC3	PC4	PC5
BB Lahir	0.218600	-0.348004	-0.050908	0.647921	0.170078
TB Lahir	-0.468668	0.138640	0.044028	-0.212296	0.258287
Usia Saat Ukur	0.076700	-0.344886	-0.446180	-0.470536	-0.229023
Berat	-0.392146	-0.042466	-0.438692	0.276567	0.235155
Tinggi	0.070563	0.482446	-0.330067	0.167075	-0.464912
BB/U	0.269307	0.280817	0.371371	-0.491351	0.033423
TB/U	-0.321961	0.369637	0.194895	0.244479	-0.090226
ASI Ekslusif	0.261772	0.347505	0.175141	0.030324	0.597797
Kesesuaian Menu MPASI	0.351369	-0.258770	0.039654	0.231383	-0.351758
Air Bersih	0.141290	0.677867	-0.257142	0.265128	-0.109083
Kecacingan	0.097867	-0.103837	0.607661	0.272477	-0.493834
Imunisasi	-0.496033	-0.019743	0.099394	0.109306	-0.087049
Pendidikan Ayah	0.548330	0.247424	0.174981	0.104587	0.196001
Pendidikan Ibu	0.064721	-0.026204	0.273404	-0.265424	-0.170018
Jamban Sehat	0.440003	-0.331506	-0.022391	0.115317	0.333019
Ketercukupan Mikronutrien	-0.537790	-0.142489	0.478674	0.162506	0.124431

	PC6	PC7	PC8	PC9	PC10
BB Lahir	0.107098	0.167801	0.115573	0.095771	-0.392166
TB Lahir	0.497621	-0.242174	0.146417	0.085618	0.086551
Usia Saat Ukur	-0.066840	0.035688	0.475653	0.113690	0.237715
Berat	0.382424	0.309338	-0.163432	0.154274	0.147794
Tinggi	-0.133099	-0.079712	0.204598	0.040579	-0.083395

Fig 13. Loading Scores

The loading scores from the PCA analysis illustrate the relative contribution of each feature to the resulting principal component, where each row in the table represents a feature from the dataset and the columns indicate the principal components (PC1, PC2, and so on). The values in the table reflect how much each feature contributes to the component. For example, BB Born has a positive loading score on PC1 (0.218600) and PC4 (0.647921), indicating a positive contribution to the principal component, but a negative loading score on PC2 (-0.348004) indicates a less favorable contribution. TB at Birth with a negative loading score on PC1 (-0.468668) shows a negative contribution to variation in this component, although it is slightly positive on PC2 (0.138640). Age at Measurement showed a negative contribution in PC2 (-0.344886) and PC3 (-0.446180), while Father's Education with the highest loading score in PC1 (0.548330) was an important feature that contributed significantly to variation in the data. Micronutrient Adequacy shows a negative loading score on PC1 (-0.537790), contributing negatively to variation in the first principal component, while Clean Water with a positive loading score on PC2 (0.677867) indicates that access to clean water contributes positively to variation in this component. From these results, it can be concluded that Paternal Education and BW at birth have a significant influence on the variation captured by the principal components, with Paternal Education being the key factor. Some features such as Micronutrient Adequacy and TB at Birth showed negative contributions, associated with unfavorable patterns. PCA provides insights into the interactions between features and their contribution to variation in the dataset, which can be utilized for decision-making and feature selection in follow-up analyses, as well as helping to understand the factors influencing observed outcomes and potential interventions needed to improve relevant conditions.

2.4. Analysis Data Explotarory

Exploratory Data Analysis (EDA) is a vital step in the data analysis process that focuses on inspecting datasets to outline their key features, typically through visual means. The main objective of EDA is to uncover insights regarding the data's structure, patterns, and interrelationships before engaging in any statistical modeling or machine learning applications.

1. Data Visualization

Data visualization is a way to represent information graphically, which aims to help people understand the context and meaning of the data. The purpose of data visualization is to convey information in a concise and clear way, so that it is easily understood by readers. Data visualization makes it very easy to convey information effectively. Analysts who deal with large amounts of raw data will usually make use of data visualization systems. [10].

Figure 14 below shows the code to perform data visualization using the seaborn library.

```
plt.ylabel('PC2 (Komponen Utama 2)')
plt.legend(title='Cluster')
plt.grid(True)

# Tampilkan plot
plt.show()

# Menjelaskan hasil visualisasi
cluster_counts = pca_df['Cluster'].value_counts()
print("Jumlah data per cluster:")
print(cluster_counts)

# Menghitung rata-rata nilai PC1 dan PC2 untuk masing-masing cluster
cluster_means = pca_df.groupby('Cluster')[['PC1', 'PC2']].mean()
print("\nRata-rata nilai PC1 dan PC2 per cluster:")
print(cluster_means)

# Definisikan fungsi untuk interpretasi cluster
def interpret_cluster(cluster):
    if cluster == 0:
        return "Cluster ini menunjukkan karakteristik tertentu yang terkait dengan nilai tinggi pada PC1 dan rendah pada PC2."
    elif cluster == 1:
        return "Cluster ini mencerminkan pola yang berbeda dengan nilai sangat negatif pada PC1."
    elif cluster == 2:
        return "Cluster ini menunjukkan nilai positif pada kedua komponen utama, menandakan karakteristik yang kuat."

# Menyediakan interpretasi hasil
for cluster in cluster_means.index:
    pc1_mean = cluster_means.loc[cluster, 'PC1']
```

Fig 14. Data Visualization Code

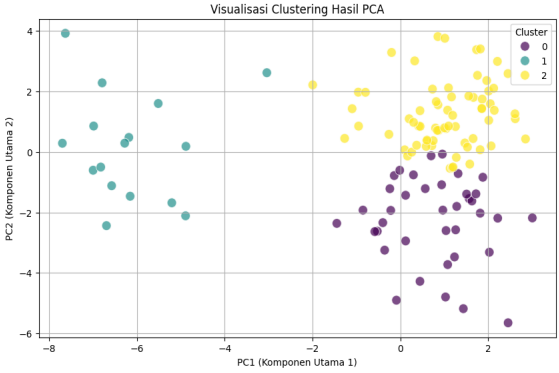


Fig 15. Data Visualization Results

The amount of data in each cluster shows a varied distribution, where cluster 2 has 68 data points, cluster 0 consists of 38 data points, and cluster 1 includes only 16 data points. Based on the mean values of the two principal components (PC1 and PC2), cluster 0 has an average PC1 of 0.78 and PC2 of -2.25, indicating characteristics with high scores on PC1 but low scores on PC2. Cluster 1 has a strongly negative average PC1 of -6.15, and a PC2 close to zero (0.16), reflecting a unique pattern with a large influence from PC1. Meanwhile, cluster 2 shows positive values on both components, with an average PC1 of 1.01 and PC2 of 1.22, indicating strong and balanced characteristics on both dimensions.

2. Correlation and Relationship between Features and Outlier Identification

Correlation is a method used to analyze the extent of the relationship between two or more variables. To understand how strong the relationship is, we can look at the value of the resulting correlation number. To illustrate the correlation and relationship between the features of the selected dataset, we used a correlation matrix.

The principle of the correlation matrix is to identify the strongest relationship between attributes, and then remove one of the attributes from the correlated pair [11].

The following figure shows the correlation matrix on the stunting dataset that has been preprocessed previously.

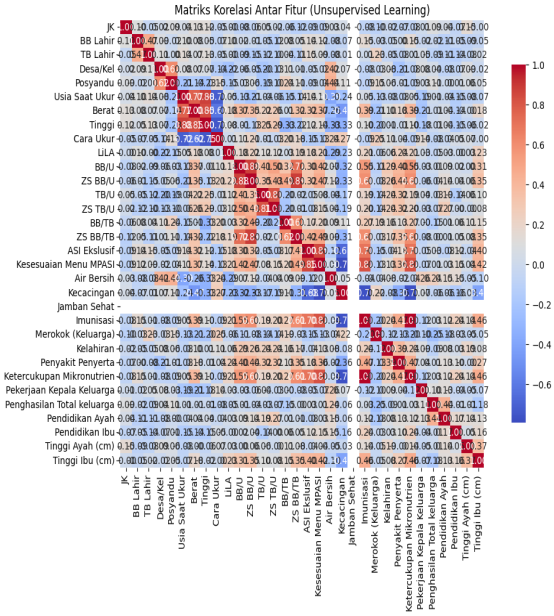
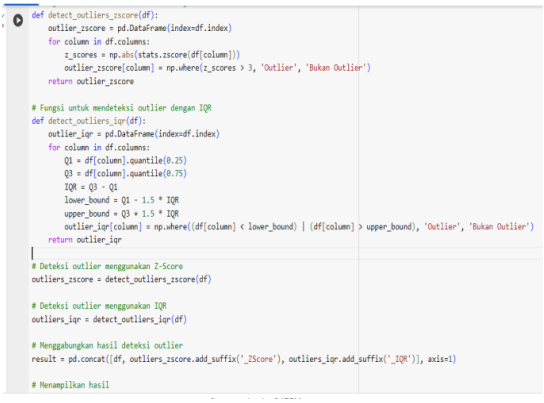


Fig 16. Correlation Matrix

After compiling the correlation matrix, the next step is to detect outliers. Outlier detection is crucial as these extreme values can have a major impact on data analysis. Outliers are values that are very different from the general pattern in the data and can arise due to several factors, such as errors in measurement, errors in recording, or natural variations present in the data.

The following figure represents the syntax for outlier detection on the dataset.



```
def detect_outliers_zscore(df):
    outlier_zscore = pd.DataFrame(index=df.index)
    for column in df.columns:
        z_scores = np.abs(stats.zscore(df[column]))
        outlier_zscore[column] = np.where(z_scores > 3, 'Outlier', 'Bukan Outlier')
    return outlier_zscore

# Fungsi untuk mendeteksi outlier dengan IQR
def detect_outliers_iqr(df):
    outlier_iqr = pd.DataFrame(index=df.index)
    for column in df.columns:
        Q1 = df[column].quantile(0.25)
        Q3 = df[column].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        outlier_iqr[column] = np.where((df[column] < lower_bound) | (df[column] > upper_bound), 'Outlier', 'Bukan Outlier')
    return outlier_iqr

# Deteksi outlier menggunakan Z-Score
outliers_zscore = detect_outliers_zscore(df)

# Deteksi outlier menggunakan IQR
outliers_iqr = detect_outliers_iqr(df)

# Menggabungkan hasil deteksi outlier
result = pd.concat([df, outliers_zscore.add_suffix('_ZScore'), outliers_iqr.add_suffix('_IQR')], axis=1)

# Menampilkan hasil
```

Fig 17. Outlier Detection Code

3. Result And Discussion

3.1. Correlation Matrix Results

In Cluster 0, there is a “Village/Kel” attribute with a PC1 value of 0.70 and PC2 of -0.13. The average PC1 and PC2 values in this cluster are 0.78 and -2.25 respectively, which indicates that in the first component (PC1), this cluster tends to have positive values, while in the second component (PC2), the values are more negatively skewed. Some data could not be identified from the original attributes because their indices were out of range.

Cluster 1 consists of attributes such as “JK” with a PC1 of -7.63 and PC2 of 3.92, and “Healthy Latrine” with a PC1 value of -6.98 and PC2 of 0.86. The average PC1 value in this cluster is -6.15, while the average PC2 is 0.16. This cluster displays characteristics with very negative PC1 values and PC2 that are close to zero or positive, indicating a strong linkage on the first dimension (PC1) with significant differences compared to other clusters. Some data also could not be identified because the index was outside the range of the PC1 and PC2 values.

For Cluster 2, some of the attributes included “Birth Weight” with a PC1 value of 0.86 and PC2 of 1.56, and “Birth TB” with a PC1 of 0.85 and PC2 of 3.81. The average PC1 value in this cluster is 1.01, while the average PC2 is 1.22. This cluster shows a balanced trend with positive values on both principal components, indicating that the

attributes included in it have a fairly even distribution on the PC1 and PC2 dimensions. Some data in this cluster also cannot be identified because the index is out of range.

Overall, each cluster shows a distinct pattern of characteristics based on the principal component values (PC1 and PC2), reflecting significant differences in the variables present in each cluster.

3.2. Outlier Detection Results

The data provided consists of 32 variables, where each variable, such as Variable_1 to Variable_10 and so on, contains numerical values representing various physical measurements, health conditions, or environmental factors relevant to the stunting analysis. There are 122 rows of data, indicating 122 observations or observed entities. Most of the columns contain numerical values, while the column ending in _IQR shows information about outliers, where all values in the column indicate "Not Outliers", indicating that the data is free of extreme values. Correlation analysis between these numerical variables is essential for identifying relationships between variables, helping to determine which variables have the most influence on the target variable, such as stunting prevalence.

Handling outliers is an important aspect of this analysis as outliers can affect the results of statistical analysis and prediction models. With all values declared as "Not Outliers", this dataset allows for more accurate analysis results. This data can potentially be used to identify factors that influence stunting, develop predictive models, and formulate health policies or interventions based on the analysis findings. Overall, this data can provide deeper insights into the factors contributing to stunting and support more effective mitigation efforts.

The following are the results of the analysis obtained on the selection of relevant attributes for further modeling with Kmeans Clustering.

```
Atribut dengan Korelasi Tinggi:
Index(['JK', 'BB Lahir', 'TB Lahir', 'Desa/Kel', 'Posyandu', 'Usia Saat Ukur',
       'Berat', 'Tinggi', 'Cara Ukur', 'LiLA', 'BB/U', 'ZS BB/U', 'TB/U',
       'ZS TB/U', 'BB/TB', 'ZS BB/TB', 'ASI Ekslusif', 'Kesesuaian Menu MPASI',
       'Air Bersih', 'Kecacingan', 'Imunisasi', 'Merokok (Keluarga)',
       'Kelahiran', 'Penyakit Penyerta', 'Ketercukupan Mikronutrien',
       'Pekerjaan Kepala Keluarga', 'Penghasilan Total keluarga',
       'Pendidikan Ayah', 'Pendidikan Ibu', 'Tinggi Ayah (cm)',
       'Tinggi Ibu (cm)'],
      dtype='object')
```

Fig 18. Attribute Selection Result

4. Conclusion And Future Work

In this analysis, a number of highly correlated attributes were identified, which could potentially have a significant effect on the prevalence of stunting in children under five. The following is a description of the twenty-eight attributes listed:

1. JK (Gender): Distinguishes between boys and girls, which can show differences in growth and development.
2. Birth Weight: Weight at birth is an important indicator of an infant's early health, which is associated with the risk of stunting later in life.
3. Birth TB (Height at Birth): Height at birth is also an important factor that can affect a child's future nutritional status.
4. Village/Kel (Village or Kelurahan): Indicates geographical location, which can affect access to health and nutrition services.
5. Posyandu: The presence of a nearby service post to monitor children's health and nutrition.Usia Saat Ukur: Usia saat pengukuran dilakukan memberikan konteks penting untuk penilaian pertumbuhan.
6. Weight: Current weight, as an indicator of current nutritional status.
7. Height: Current height, also an important indicator of nutritional status.
8. Measurement Method: The method used to measure weight and height, which may affect the accuracy of the data.
9. LiLA (Upper Arm Circumference): Measures body fat and provides additional information on nutritional status.
10. BB/U (Weight for Age): A ratio used to assess whether a child's weight is appropriate for their age.
11. ZS BB/U (Weight-for-Age Z-score): Describes how far a child's weight is from the average weight of children of the same age.
12. TB/U (Height for Age): A ratio that indicates whether a child's height is appropriate for their age.
13. TB/U ZS (Height-for-Age Z-score): Describes how far a child's height is from the average height of children of the same age.
14. BB/TB (Weight for Height): An indicator to assess a child's nutritional status based on the ratio between body weight and height.
15. ZS BB/TB (Weight-for-height Z-score): Describes how far a child's weight-to-height ratio is from the norm.
16. Exclusive breastfeeding: Indicates whether the child is receiving exclusive breastfeeding, which is essential for healthy growth.
17. Appropriateness of Complementary Feeding Menu: The suitability of complementary foods to the nutritional needs of the child.
18. Clean Water: The availability of clean water that plays a role in child health and prevents infections.

19. Worm infestation: The presence of worm infections can affect nutrient absorption and child health.
20. Immunisation: A child's immunisation status is important to prevent diseases.sss
21. Smoking (Family): Smoking habits in the family environment that may affect children's health.
22. Birth: Birth history, whether born prematurely or with complications that may affect health.
23. Comorbidities: The presence of diseases that may affect a child's nutritional status.
24. Micronutrient Adequacy: Ensuring that the child is getting enough vitamins and minerals necessary for growth.
25. Head of Household Occupation: Indicates economic stability which may affect access to nutrition.
26. Total Family Income: Total family income that may affect the ability to provide nutritious food.
27. Education of Father and Mother: The education level of parents, which is related to knowledge about nutrition and child health.
28. Father's Height (cm) and Mother's Height (cm): Indicates genetic potential that may affect a child's growth.

This data was obtained from a dataset entitled ‘data_stunting_transformed_final.csv’. This dataset includes information on child health and nutrition, as well as environmental and socioeconomic factors that may influence the prevalence of stunting. Using analytical techniques such as correlation matrix calculation and visualisation, attributes that have significant relationships can be identified and used for the next steps in clustering analysis to understand the patterns among the data and formulate appropriate interventions to address stunting in children under five.

By applying these preprocessing techniques, the dataset becomes better suited for machine learning algorithms like K-means clustering, which relies on clean and scaled data for optimal performance. As a result, the final model is more accurate, interpretable, and capable of providing insights that are relevant to real-world health policies and stunting interventions. This preprocessing also enables the model to generalize better to new data, supporting more effective monitoring and decision-making in stunting prevention efforts.

References

1. G. Classico, A. Thompson And Sentonastaso, "The Impact Of Stunting On Child Development: A Review," *Journal Of Pediatric Health Care,* Pp. 503-509, 2019.

2. World Health Organization, "Global Nutrition Report 2020: Action On Equity To End Malnutrition," No. Retrieved From Https://Globalnutritionreport.Org/Reports/2020-Global-Nutrition-Report/, 2020.

3. R. Jain And A. Yadaf, "A Survey On Data Preprocessing Techniques In Data Mining," *International Journal Of Computer Applications,* Vol. 175, No. 10, Pp. 21-25, 2021.

4. D. Kurniawan And S. Indratno, "Application Of K-Means Clustering In The Analysis Of Stunting Data In Indonesia," *International Journal Of Health Science And Research,* Vol. 10, No. 2, Pp. 36-42, 2020.

5. T. Ahmad And M. Aziz, "Data Preprocessing And Feature Selection For Machine Learning Intrusion Detection Systems," *Icic Express Letters,* Vol. 13, No. 2, Pp. 93-101, 2019.

6. S. Sunneng, B. Haeni, Jatmika And G. Fornieli, "Data Preprocessing Approach For Machine Learning-Based Sentiment Classification," *Jurnal Infotel,* Vol. 15, No. 4, Pp. 317-325, 2023.

7. G. Ainayya And Hari Setiaji, "Data Cleansing Pada Data Rumah Sakit," In *Proceeding Sintak 2019*, Medan, 2019.

8. R. Anshori, A. Mudi And Nurhaeni, "Penanganan Imputasi Missing Values Pada Data Time Series Dengan Menggunakan Metode Data Mining," *Jurnal Informasi Dan Teknologi,* Vol. 5, No. 2, Pp. 56-62, 2023.

9. A. Rahman, B. Warsito And A. Prahutama, "Pengaruh Transformasi Data Pada Metode Learning Vector Quantization Terhadap Akurasi Klasifikasi Diagnosis Penyakit Jantung," *Jurnal Gaussian,* Vol. 10, No. 1, Pp. 21-30, 2021.

10. C. Yu, F. Gao, S. Lin And J. Wang, "Quantum Data Compression By Principal Component Analysis," *Quantum Inf Process,* Vol. 18, No. 8, 2019.

11. D. Syarafina, A. Magdalena And I. Pakereng, "Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang," *Jipi (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika),* Vol. 8, No. 4, Pp. 1186-1195, 2023.

12. R. A. Ghivary, Mawar, N. Wulandari, N. Srikandi And A. Naziatul, "Peran Visualisasi Data Untuk Menunjang Analisa Data Kependudukan Di Indonesia," *Pentahelix: Jurnal Administrasi Publik,* Vol. 1, No. 1, Pp. 57-62, 2023.

13. E. N. R. Khaki, A. Hermawan And D. Avianto, "Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine," *Jiko (Jurnal Informatika Dan Komputer),* Vol. 7, No. 1, Pp. 158-166, 2023.