ANALYSIS OF STUNTING PREVALENCE WITH UTILIZATION OF STUNTING REDUCTION

PERFORMANCE INDEX COMPONENTS IN INDONESIA USING K-MEANS AND SUPPORT VECTOR MACHINE

ABSTRACT

Stunting is a health issue that has received major attention from the government, as mandated in the 2020-2024 National Medium-Term Development Plan (RPJMN) in Indonesia. Based on the Indonesian Nutrition Status Survey (SSGI) in 2022, the national stunting prevalence rate reached 21.6%, while the target by the end of 2024 is 14%. Therefore, this study aims to analyze the factors that contribute to the stunting prevalence rate in the region. Data collected from the Indonesian Central Bureau of Statistics publication will be analyzed using the K-means and Support Vector Machine methods. There are 19 factors that are thought to influence the stunting prevalence rate in Indonesia. The results of the research using the K-means method gave rise to two clusters. The first cluster consists of 425 districts/cities characterized by an average stunting prevalence of 13.55%, which is lower than the second cluster consisting of 95 districts/cities. Meanwhile, the classification results with the Support Vector Machine method showed the highest accuracy rate of 88%, using the rbf kernel.

INTRODUCTION

1..1 BACKGROUND

Despite efforts to address stunting, SSGI 2022 data reveals that the prevalence is still high, reaching 21.6%.

two methods are used:

- K-Means Clustering Support = grouping districts based on certain characteristics,
- Vector Machine (SVM) = used to classify and further analyze the factors that influence the prevalence of stunting in each region.

This research effort aims to support the achievement of national targets to significantly reduce the prevalence of stunting. In line with the government's target to reduce stunting prevalence by 3.8 percent per year, this research is expected to provide a strong evidence base for progressive and sustainable policy planning until 2024.

INTRODUCTION

1.2 Literature Study

- K-Means Algroitma

$$D(i,j) = \sqrt{(X_{1i} - Y_{1j})^2 + (X_{2i} - Y_{2j})^2 + \cdots + (X_{ki} - Y_{kj})^2}$$

non-supervised learning used to group unlabeled data into different clusters

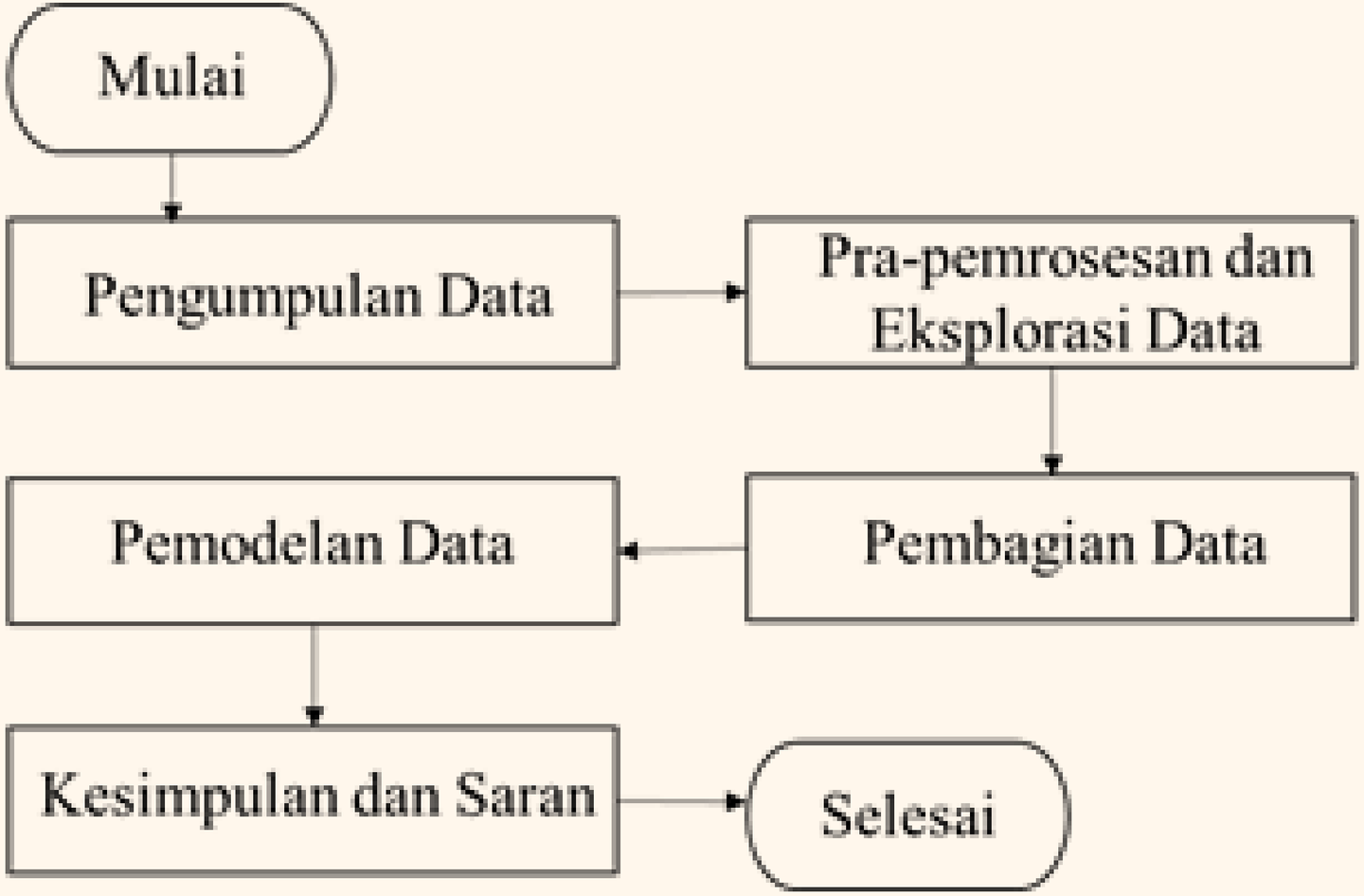
- Support Vector Machine (SVM) Algorithm

Supervised learning algorithm used to perform two-class classification of kernel SVMs of several types:

1. Linear Kernel
2. Polynomial Kernel
3. Radial Basis Function
4. Sigmoid Kernel

METHODOLOGY

2.1 Data Collection



data used = BPS publication

2022 Dataset consists of

METHODOLOGY

2.1 Data Collection

```
#Input Data
src =pd.read_excel("C:/Users/asus/Downloads/Stunt dataset.xlsx")
src.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Kabupaten/Kota Prov Indonesia   520 non-null    object
 1   Prevalensi Stunting (TB/U) %    520 non-null    float64
 2   K4                              520 non-null    float64
 3   Persalinan FASYANKES            520 non-null    float64
 4   KF Lengkap                      520 non-null    float64
 5   Vit A Ibu                       520 non-null    float64
 6   bumil TTD                       520 non-null    float64
 7   BBLR                            520 non-null    float64
 8   IMD                             520 non-null    float64
 9   ASI                             520 non-null    float64
 10  CPKB                            520 non-null    float64
 11  IDL                             520 non-null    float64
 12  A 611                           520 non-null    float64
 13  A 1259                          520 non-null    float64
 14  A 659                           520 non-null    float64
 15  mCPR                            520 non-null    float64
 16  Air Minum Layak                 520 non-null    float64
 17  Sanitasi Layak                  520 non-null    float64
 18  IKP                             520 non-null    float64
 19  BPNT 40%                        520 non-null    float64
 20  KKS 40%                         520 non-null    float64
 21  APK PAUD                        520 non-null    float64
 22  UMK                             520 non-null    float64
```

consists of 23 columns

22 float shapes

1 shape string

METHODOLOGY

2.2 Data Pre-processing and Exploration

```
#Cek missing value
src.isna().sum()
```

```
Kabupaten/Kota Prov Indonesia    0
Prevalensi Stunting (TB/U) %     0
K4                               0
Persalinan FASYANKES             0
KF Lengkap                       0
Vit A Ibu                        0
bumil TTD                        0
BBLR                             0
IMD                              0
ASI                              0
CPKB                             0
IDL                              0
A 611                            0
A 1259                           0
A 659                            0
mCPR                             0
Air Minum Layak                  0
Sanitasi Layak                   0
IKP                              0
BPNT 40%                         0
KKS 40%                          0
APK PAUD                         0
UMK                              0
dtype: int64
```

Missing Value Checking

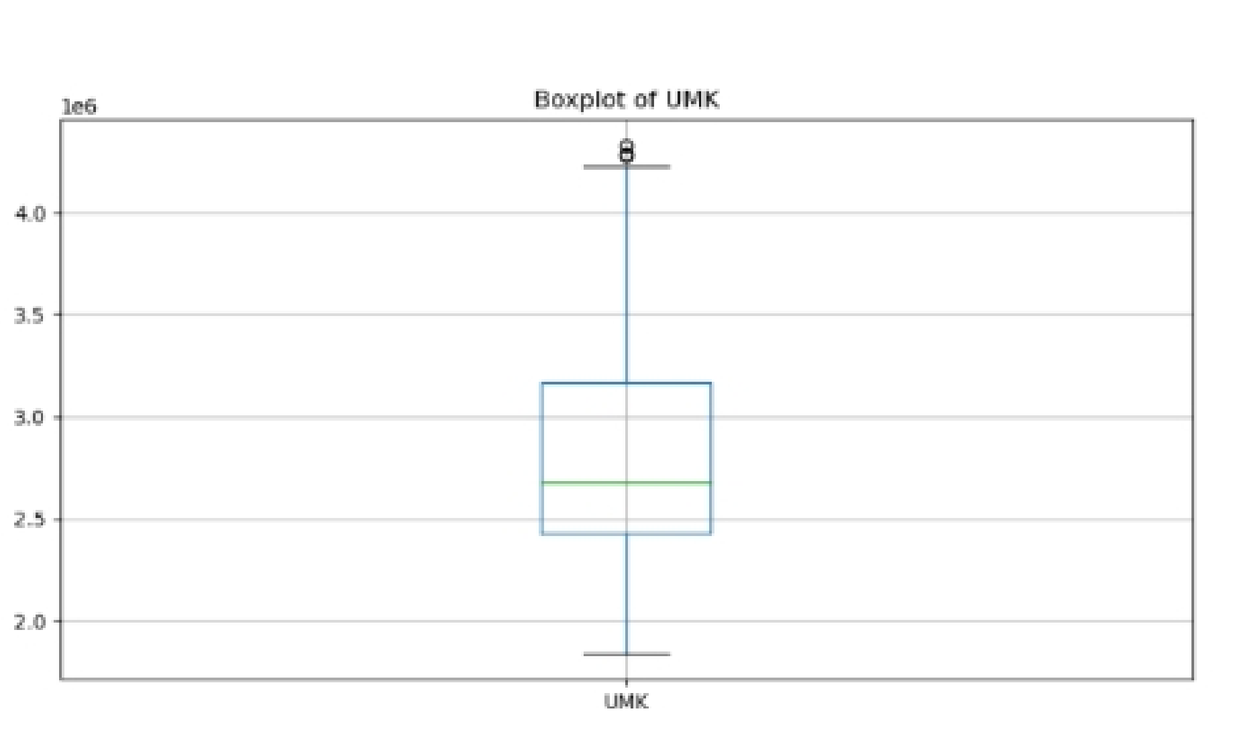
```
#cek outlier
Q1 = srcnokab.quantile(q=.25)
Q3 = srcnokab.quantile(q=.75)
IQR = Q3-Q1

data_iqr = srcnokab[-((srcnokab < (Q1-1.5*IQR)) | (srcnokab >(Q3+1.5*IQR))).a
data_iqr.shape

print("Dimensi dataset awal", srcnokab.shape)
print("Dimensi dataset setelah pengecekan outlier", data_iqr.shape)
```

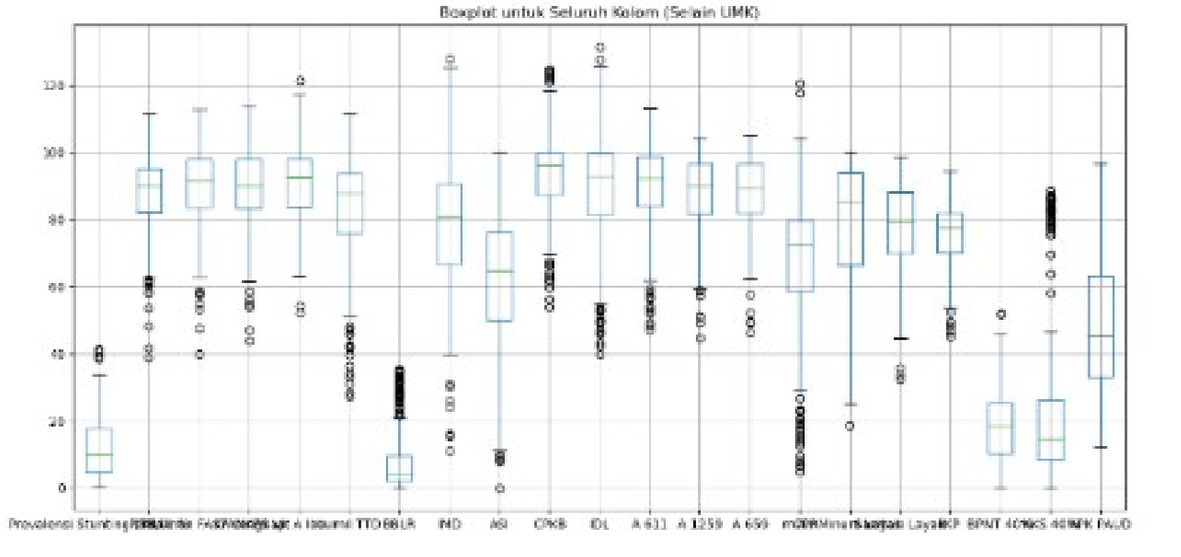
```
Dimensi dataset awal (520, 22)
Dimensi dataset setelah pengecekan outlier (274, 22)
```

Outlier Checking Using Interquartile

METHODOLOGY

2.2 Data Pre-processing and Exploration

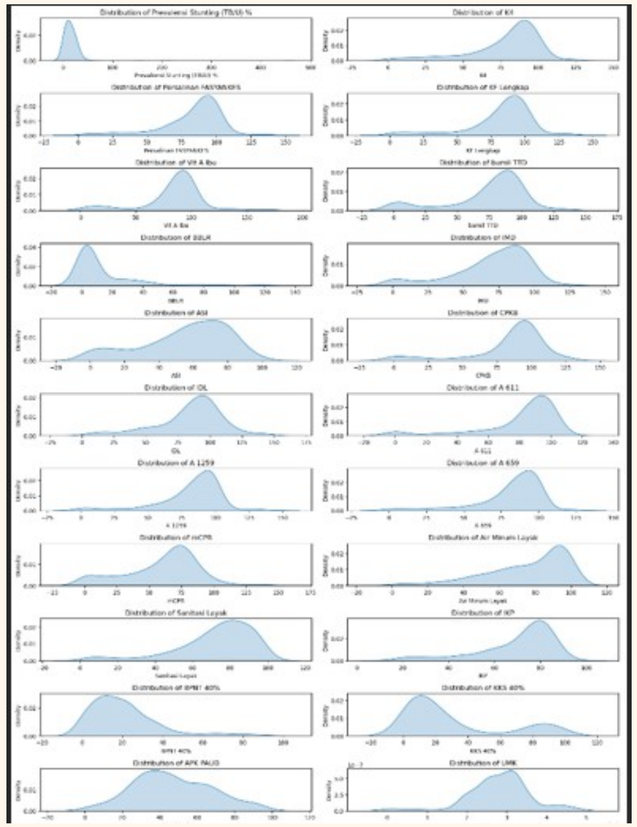




Boxplot of Numerical Variables

METHODOLOGY

2.2 Data Pre-processing and Exploration



```
Skewness of Prevalensi Stunting (TB/U) %: 13.017
Skewness of K4: -1.401
Skewness of Persalinan FASYANKES: -1.307
Skewness of KF Lengkap: -1.381
Skewness of Vit A Ibu: -0.849
Skewness of bumil TTD: -1.131
Skewness of BBLR: 2.633
Skewness of IMD: -1.027
Skewness of ASI: -0.681
Skewness of CPKB: -1.633
Skewness of IDL: -0.878
Skewness of A 611: -1.836
Skewness of A 1259: -1.513
Skewness of A 659: -1.733
Skewness of mCPR: -0.428
Skewness of Air Minum Layak: -1.1
Skewness of Sanitasi Layak: -1.383
Skewness of IKP: -1.382
Skewness of BPNT 40%: 1.64
Skewness of KKS 40%: 1.09
Skewness of APK PAUD: 0.242
Skewness of UMK: -0.638
```

. Kernel Density Estimation (KDE) Plot

METHODOLOGY

2.2 Data Pre-processing and Exploration

```
from sklearn import preprocessing
srcz = preprocessing.scale(srcnokab)
srcz
```

```
array([[ 0.43378231,  0.32265872,  0.2393182 , ..., -0.65467682,
         0.70923903, -1.06641861],
       [ 0.09986639,  0.70091003,  0.64587553, ..., -0.11416977,
        -0.05339914, -0.60080731],
       [-0.05678553, -0.04709256, -0.09803787, ...,  0.03831383,
         1.0150661 , -0.99347532],
       ...,
       [-0.23404954, -2.77560203, -1.27445906, ...,  2.16618955,
        -1.53477918, -0.02977162],
       [-0.11037697,  1.89941419, -0.49594504, ..., -0.75200336,
         0.0088275 ,  0.58191388],
       [ 0.30186491, -1.33059702,  1.45034002, ...,  1.7813729 ,
        -0.71473813, -0.31955558]])
```

. Data Snapshot After Standardization

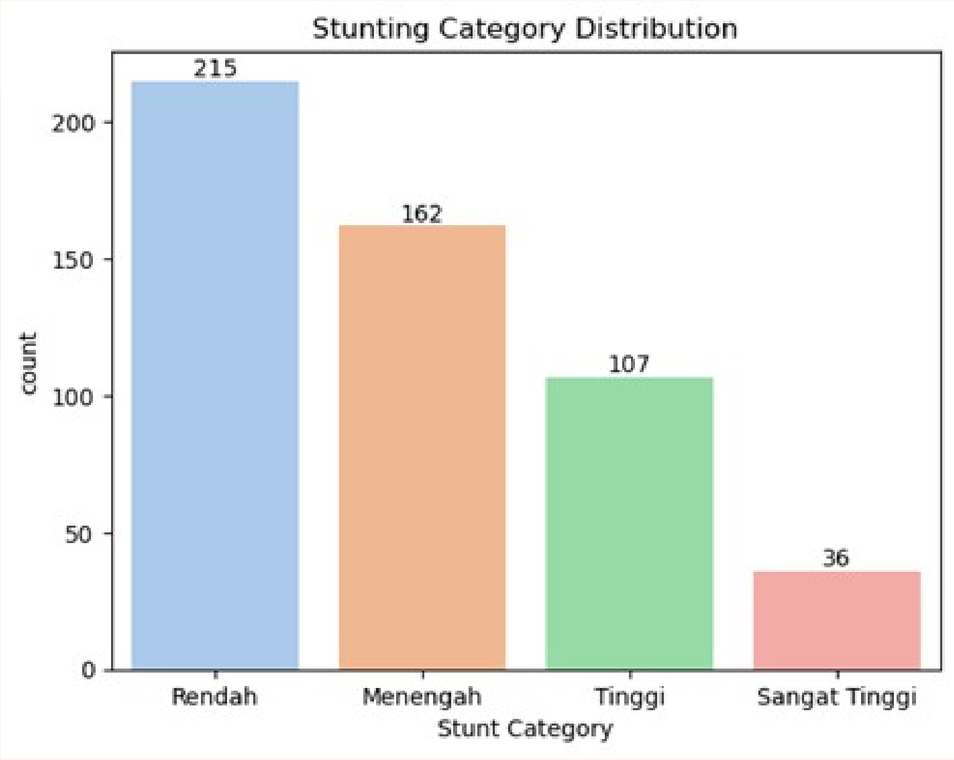
This process aims to make the data distribution more symmetrical and reduce the impact of extreme values.

METHODOLOGY

2.2 Data Pre-processing and Exploration

```
#Binning menurut WHO
categories = ['Rendah','Menengah', 'Tinggi','Sangat Tinggi']
src['Stunt Category'] = pd.cut(src['Prevalensi Stunting (TB/U) %'], bins=[-float('inf'), 10, 20, 30, float('inf')], labels=categories)


#Encoding
category_mapping = {'Rendah': 1, 'Menengah': 2, 'Tinggi': 3, 'Sangat Tinggi': 4}
src['Stunt CatNum'] = src['Stunt Category'].map(category_mapping)
print(src[['Kabupaten/Kota Prov Jawa Timur','Prevalensi Stunting (TB/U) %', 'Stunt Category', 'Stunt CatNum']])
```

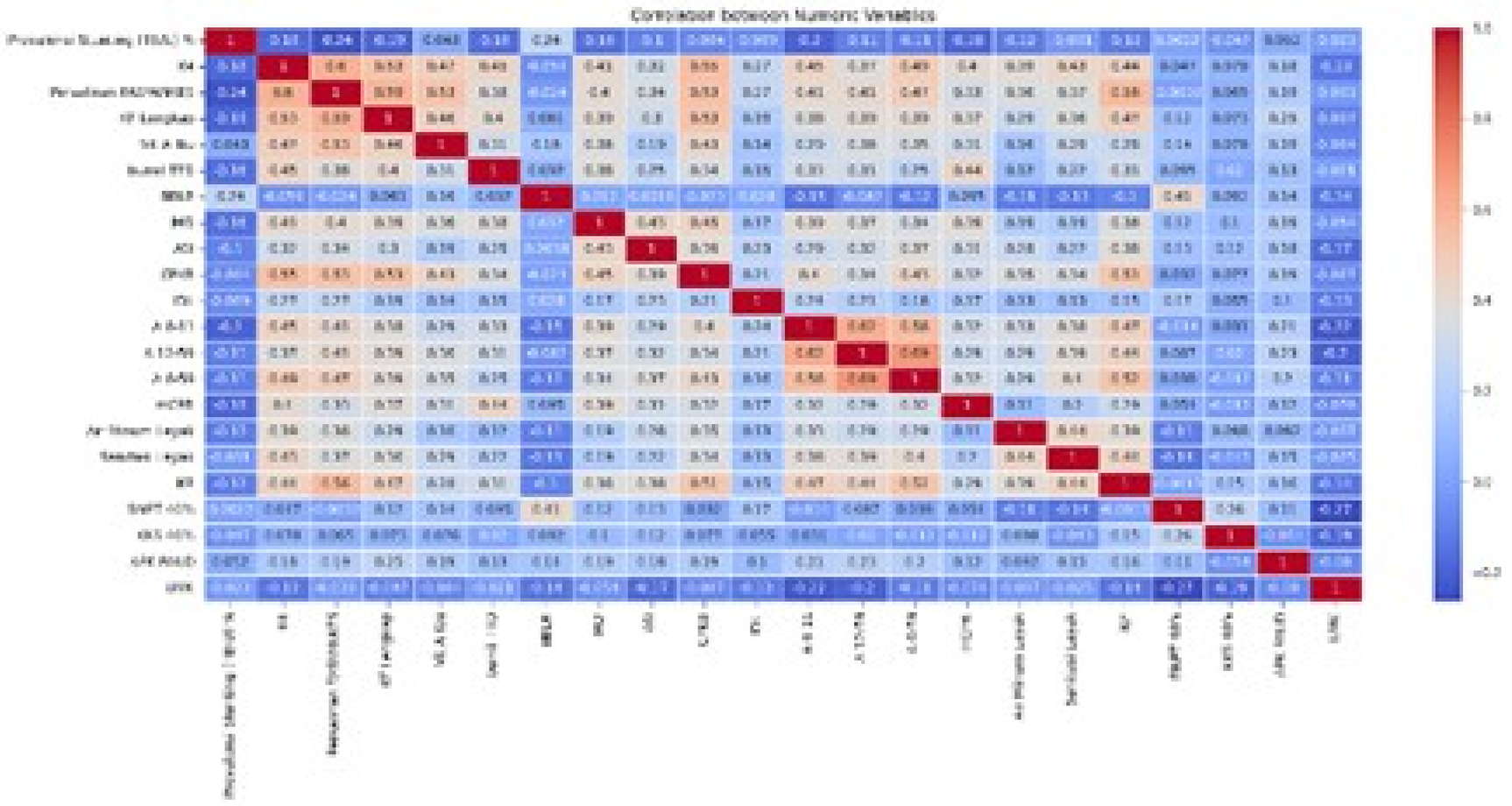
Binning and Encoding



. Distribution of Districts/Cities by Stunting Prevalence Category

METHODOLOGY

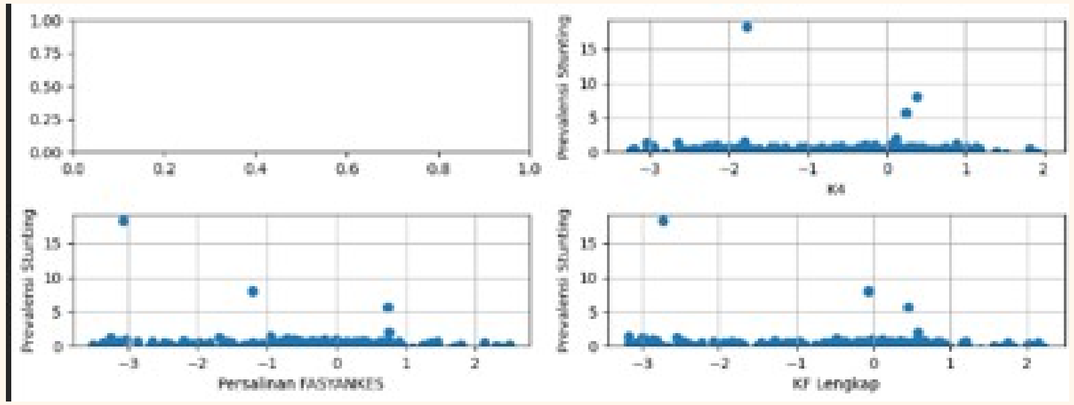
2.2 Data Pre-processing and Exploration

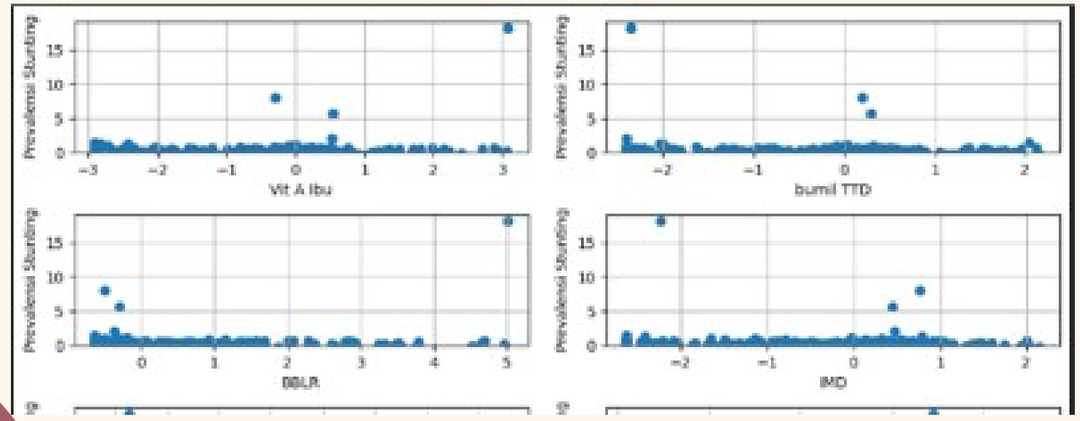


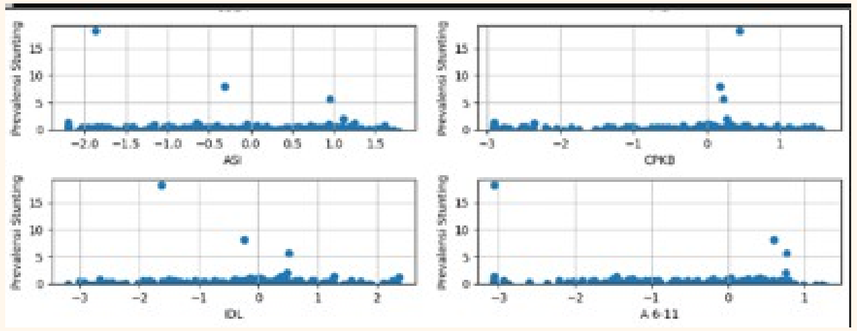
Correlation of Numerical Variables to Response Variables

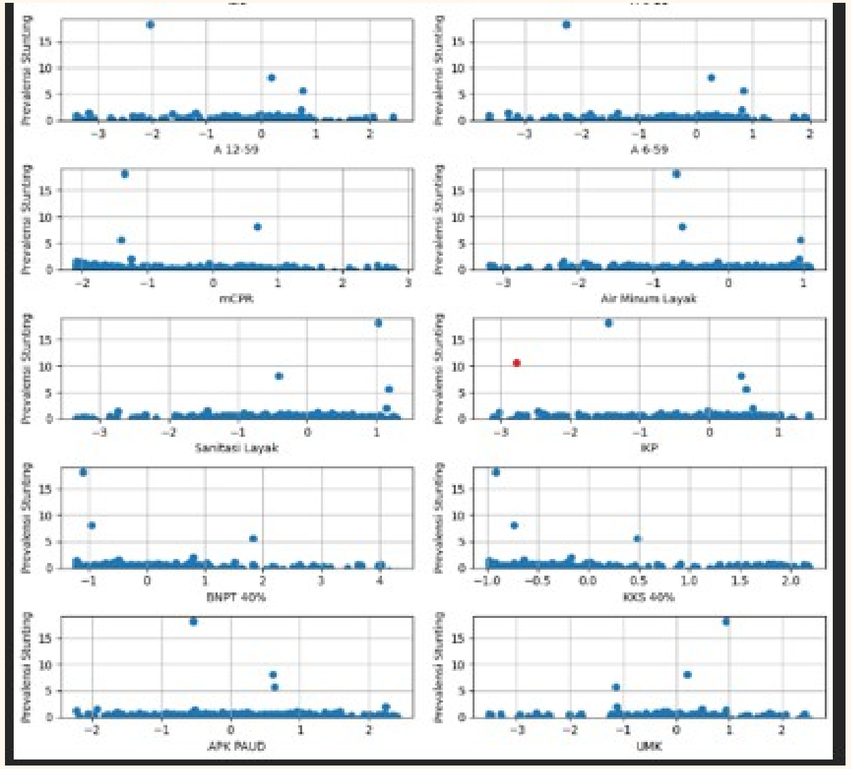
METHODOLOGY

2.2 Data Pre-processing and Exploration









Scatterplot of Numerical Variables against Stunting Prevalence

METHODOLOGY

2.3 Data Sharing

At this stage, the process of dividing data into train data and test data is carried out. The goal is to measure the goodness of the model formed.

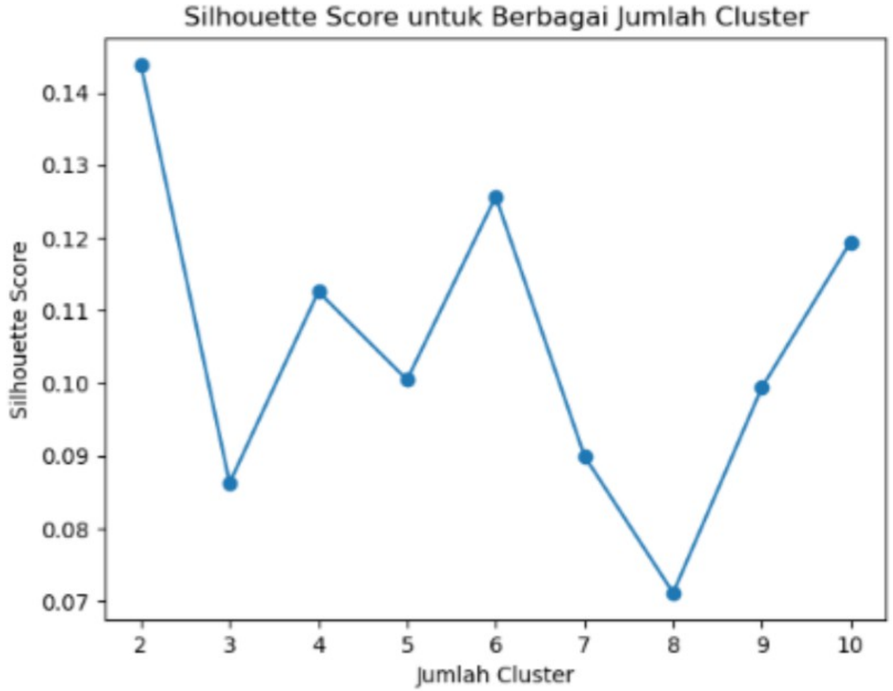
The 80:20 ratio is used in this research, which means 80% of the data will be used as train data and the remaining 20% as test data which is expected to train the machine learning model to improve accuracy.

METHODOLOGY

2.4 Data Modeling (K-Means Clustering)

Based on the Figure, the highest Silhouette value is obtained when using two clusters, with a Silhouette score of 0.144.

Thus, districts/cities in Indonesia will be grouped into two groups based on the similar characteristics of the observed variables.



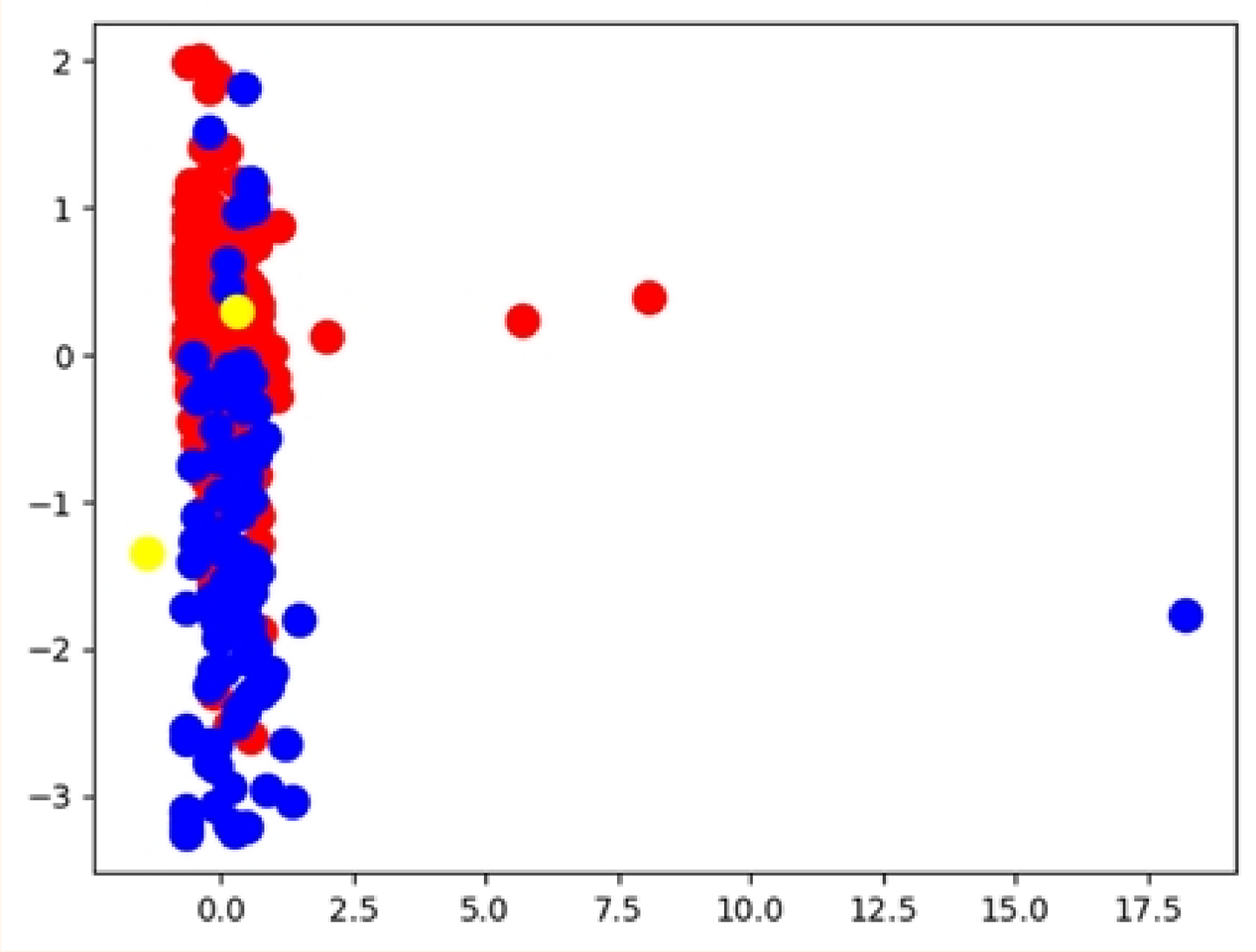

```
Silhouette Score for 2 clusters: 0.144
Silhouette Score for 3 clusters: 0.086
Silhouette Score for 4 clusters: 0.113
Silhouette Score for 5 clusters: 0.100
Silhouette Score for 6 clusters: 0.126
Silhouette Score for 7 clusters: 0.090
Silhouette Score for 8 clusters: 0.071
Silhouette Score for 9 clusters: 0.099
Silhouette Score for 10 clusters: 0.119
```

METHODOLOGY

2.4 Data Modeling (K-Means Clustering)

Cluster 0 has a lower Stunting Prevalence (TB/U) %, around 13.55%, which can be identified as a group with better public health conditions.

Cluster 1 which has a prevalence rate of 25.20%, stunting

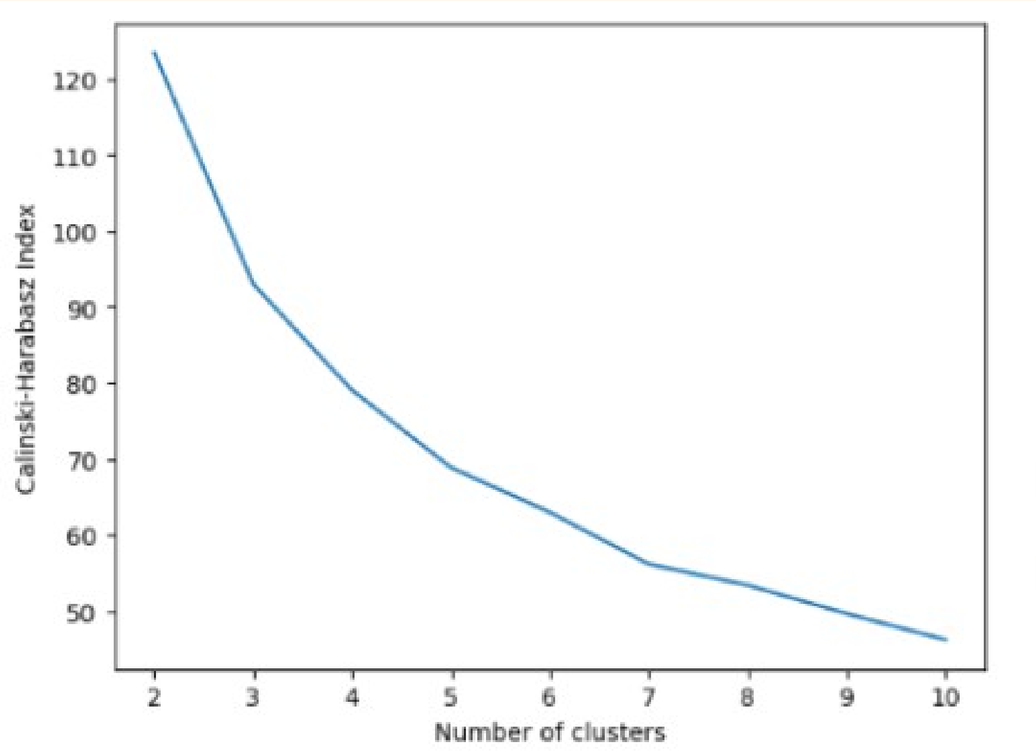


Grouping Districts/Cities into Two Clusters

METHODOLOGY

2.4 Data Modeling (K-Means Clustering)

Clustering performance according to the Calinski-Harabasz index also shows that the optimum number of clusters for clustering districts/cities in Indonesia based on predictor variables of stunting prevalence is two.



Calinski-Harabasz Index

METHODOLOGY

2.4 Data Modeling (Support Vector Machine (SVM))

	precision	recall	f1-score	support
0	1.00	0.80	0.89	20
1	1.00	0.83	0.91	6
2	1.00	1.00	1.00	2
3	0.70	1.00	0.82	7
4	0.90	0.90	0.90	31
6	0.90	1.00	0.95	9
7	1.00	0.71	0.83	7
8	1.00	0.71	0.83	7
9	0.75	1.00	0.86	15
accuracy			0.88	104
macro avg	0.92	0.89	0.89	104
weighted avg	0.91	0.88	0.89	104

rbf kernel

	precision	recall	f1-score	support
0	1.00	0.80	0.89	20
1	0.83	0.83	0.83	6
2	0.67	1.00	0.80	2
3	0.88	1.00	0.93	7
4	0.88	0.94	0.91	31
6	0.89	0.89	0.89	9
7	1.00	0.86	0.92	7
8	1.00	0.71	0.83	7
9	0.78	0.93	0.85	15
accuracy			0.88	104
macro avg	0.88	0.88	0.87	104
weighted avg	0.90	0.88	0.88	104

linear kernel

	precision	recall	f1-score	support
0	1.00	0.80	0.89	20
1	1.00	0.33	0.50	6
2	0.06	0.50	0.10	2
3	0.80	0.57	0.67	7
4	0.90	0.90	0.90	31
6	0.90	1.00	0.95	9
7	1.00	0.71	0.83	7
8	0.00	0.00	0.00	7
9	0.76	0.87	0.81	15
accuracy			0.75	104
macro avg	0.71	0.63	0.63	104
weighted avg	0.83	0.75	0.77	104

sigmoid kernel

	precision	recall	f1-score	support
0	1.00	0.50	0.67	20
1	1.00	0.67	0.80	6
2	1.00	1.00	1.00	2
3	1.00	0.86	0.92	7
4	0.49	1.00	0.66	31
6	1.00	0.44	0.62	9
7	1.00	0.43	0.60	7
8	1.00	0.86	0.92	7
9	0.50	0.20	0.29	15
accuracy			0.66	104
macro avg	0.89	0.66	0.72	104
weighted avg	0.78	0.66	0.65	104

poly kernel

METHODOLOGY

2.4 Data Modeling (Support Vector Machine (SVM))

The SVM model with rbf and linear kernels gives the highest accuracy value, which is 88%. The class with the highest recall value was class 2 (100%), followed by class 4 (90%) and class 6 (100%). The class with the lowest recall value was class 0 (80%), while class 3 and class 9 had a recall of 100%.

With the highest f1-score values belonging to class 2 (100%) and class 6 (95%), this model can be considered as an effective model in classifying the data into different classes.

CONCLUSIONS AND SUGGESTIONS

The study concluded that 520 districts/cities in Indonesia can be grouped into two clusters, with the first cluster consisting of 425 districts/cities, and the second cluster consisting of 95 districts/cities. The first cluster is characterized by a lower average prevalence rate than members of the second cluster.

prevalence rate of first cluster < second cluster

SVM with the highest accuracy value: Linear kernel

CONCLUSIONS AND SUGGESTIONS

Suggestions for the Indonesian government to consistently run programs that have succeeded in reducing the prevalence of stunting in Indonesia, especially giving focus to districts / cities that are members of cluster 1, namely in the form of 425 districts / cities.

THANK YOU