Mutual information-MOORA based feature weighting on naive bayes classifier for stunting data

Citra Nurina Prabiantissa[1], Maftahatul Hakimah[1], Nanang Fakhrur Rozi[1], Ira Puspitasari[2], Laura Navika Yamani[3], Victoria Lucky Mahendra[1]
[1]Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Surabaya, Indonesia
[2]Department of Information System, University of Airlangga, Surabaya, Indonesia
[3]Department of Public Health, University of Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received Dec 9, 2023
Revised Jul 5, 2024
Accepted Jul 14, 2024

Keywords:

Feature weighting
MOORA method
Mutual information
Naive Bayes classifier
Stunting prediction

ABSTRACT

One effort to reduce stunting rates is to predict stunting status early in toddlers. This study applies Naive Bayes (NB) to build a stunting prediction model because it is simple and easy to use. This study proposes a filter-based feature weighting technique to overcome the NB assumption, which states that each feature has the same contribution to the target. The frequency of an event in a dataset influences the feature weighting using mutual information criteria. This is the gap in the filter-based ranking highlighted in this study. Therefore, this study proposes a feature-weighting method that combines mutual information with the MOORA (MI-MOORA) decision-making method. This technique makes it possible to include external factors as criteria for ranking important features. For stunting cases, the external consideration for ranking purposes is the assessment of nutrition experts based on their experience in dealing with stunted toddlers. The MI-MOORA technique makes the availability of clean water the most influential feature that contributes to the stunting status. In the ten best features, the MI-MOORA ranking results are dominated by family factors. Based on the performance evaluation results of NB and other classifiers, MI-MOORA can improve the performance of stunt prediction models.

Corresponding Author:

Maftahatul Hakimah
Department of Informatics Engineering, Faculty of Electrical Engineering and Information Technology
Institut Teknologi Adhi Tama Surabaya
Surabaya, Indonesia
Email: hakimah.mafta@itats.ac.id

1. INTRODUCTION

Stunting is a condition of malnutrition that starts in the womb. Stunting in children begins with weight loss due to a lack of calorie intake. When the lack of calorie intake is chronic, the growth rate decreases, leading to stunting [1]. The condition of stunted children is measured based on height-for-age Z-scores below −2 standard deviation (SD), and severely stunted conditions have height-for-age Z-scores below −3 SD. In 2022, the stunting prevalence rate in Indonesia is 21.6%, which decreased by 2.8% from 2021 [2]. Nevertheless, this number falls short of the World Health Organization (WHO) target of less than 20%.

Chronic malnutrition in children can have an impact on cognitive development in childhood, such as low levels of intelligence, mental retardation, and the risk of developing degenerative diseases when they grow [3]. Evidence that stunting has a direct impact on future adverse outcomes is incomplete. Still, increasing evidence shows that stunting is a risk factor for impaired health, education, and economic performance in the future [4]. Stunting and wasting are the dominant manifestations of malnutrition which have negative impacts and increase the risk of death [3].

The factors that cause stunting are complex, multifactorial, and interconnected [2]. According to the WHO, impaired growth and development in children are influenced by family and country. As stunting increases the likelihood of future harmful effects, preventive measures need to be implemented. The preventive measure discussed here is the early detection of stunting status at five through a prediction system based on risk factors. Machine learning is useful for forming prediction models to identify risk factors that together predict the stunting status of toddlers [5], [6].

Naive Bayes (NB) is a machine learning method widely used to build stunting prediction models. The NB classifier is a simple approach that exhibits low algorithm complexity and high algorithm performance [7], [8]. The inherent NB method assumes that each feature has the same level of influence on the target. Numerous NB techniques have been developed to address presumptions that are occasionally unrelated to actual cases. Previous research has carried out a feature weighting approach for NB improvement [9], [10].

The foundation of our research is the filter-based feature weighting for NB. The weighting scheme begins with a filter-based feature selection stage, and the metric score is then used as the feature weight value. The filter approach ranks the features based on specific measurement metrics. Filter-based feature weighting can use several measurement metrics, for example, mutual information (MI) [11], [12], F-criterion [13], and distance correlation coefficient [14]. The advantage of filter-based weighting, similar to feature selection, is that it does not depend on the classifier, so the computational cost is low [15]. However, this leads to a common weakness of filter approaches in that the performance of different classification methods varies when a selected subset of features is applied [15].

An important step in filter-based weighting is feature scoring, which evaluates the significance of a feature when developing a model. Improving NB through feature weighting requires selecting highly predictive features correlated with the class but not with other features. Regarding MI, a feature should have maximum relevance to the class and minimum redundancy with other features [11]. Based on this, we argue that feature scoring using MI is data-dependent. The state of the experimental data impacts the level of dependence between features and their relationship to the target. The MI score depends on the probability distribution of instances in the dataset. This paper presents a filter-based feature-weighting strategy using the multi-objective optimization by ratio analysis (MOORA) decision-making process. We propose feature scoring based on MI scores and incorporate external criteria, such as expert assessments based on experience or other relevant criteria.

The contributions of this study are twofold. First, it introduces a straightforward and adaptable filter-based feature scoring technique utilizing the MOORA approach, which allows the inclusion of assessment criteria to determine important features, called MI-MOORA. Second, the feature selection method employs the preference value as a measurement metric. The remainder of this paper is structured as follows: section 2 details the methodology, including an explanation of the experimental data, thorough preprocessing instructions, and the recommended technique stages. In section 3 presents and discusses the results obtained from several test cases conducted to demonstrate the effectiveness of the proposed strategy. Finally, conclusions and recommendations based on the findings are provided in the last section.

2. METHOD

Figure 1 shows the entire stages of this research. The aim is to obtain an NB model for predicting stunting risk factors with good performance. The MI-MOORA weighting method is proposed to achieve this goal. A feature selection scheme is applied before the NB classifier and other classifiers are employed for prediction. The best model is obtained based on performance evaluation.

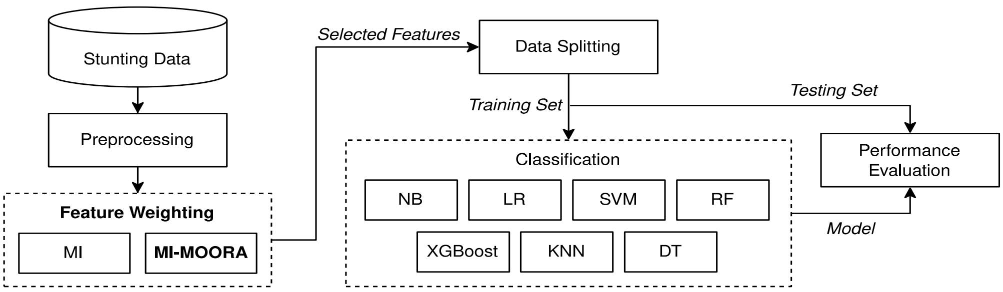

Figure 1. Research methodology

2.1. Data description

The research data was gathered from two distinct sources. The first source was secondary data collected in 2020 from community health centers (*Posyandu*) in East Surabaya, Indonesia. This dataset provided foundational information for the study. The second source involved a questionnaire administered at the *Posyandu* observation site in 2023. The questionnaire was structured based on variables derived from the first dataset. The respondents in this study were mothers with toddlers. Table 1 provides an overview of the observed variables, which total 27 factors. There were 121 data rows gathered and categorized into three classes: 71 rows were classified as normal, 32 as short (indicating stunting), and 18 as very short (indicating severely stunted). These categorizations were essential for analyzing and understanding the factors contributing to toddler growth and development within the studied population.

Table 1. Overview of observed variables related to stunting data

No.	Feature	Description	Type
1	Father's last education	Father's highest level of education	Discrete
2	Father's occupation	Activities that fathers do to earn a living	Discrete
3	Smoker dad	Father's smoking habit	Discrete
4	Mother's last education	Mother's highest level of education	Discrete
5	Mother's job	Activities that mothers do to earn a living	Discrete
6	Maternal age at pregnancy	Age at birth	Continuous
7	Monthly household income	Income received by the household, whether from the head of the household or other household members	Discrete
8	Type of delivery method	Method of childbirth used by the mother (normal/cesarean)	Discrete
9	Age	Toddler's age when measured	Continuous
10	Gender	Toddler's gender	Discrete
11	Birth weight	Toddler's weight at birth (kg)	Continuous
12	Birth height	Toddler's height at birth (cm)	Continuous
13	Birth number	Birth order in the family	Continuous
14	Early initiation of breastfeeding	The process of initiating breastfeeding within 1 hour after the baby is born	Discrete
15	Exclusive breastfeeding	Provide only breast milk to toddlers as their sole source of nutrition for the first 6 months after birth, without additional food or drink	Discrete
16	Length of exclusive breastfeeding	Duration of exclusive breastfeeding (month)	Continuous
17	UAC	Upper arm circumference measurement to assess the nutritional status of toddlers	Continuous
18	HC	Head circumference and crown size are used to assess a child's growth and development, reflecting the size and growth of their brain	Continuous
19	Immunization status	History of receiving basic immunization according to Ministry of Health programs (complete/no)	Discrete
20	Toddler caregiver	The caregiver responsible for the toddler	Discrete
21	Go to "*Posyandu*" regularly	Toddlers are routinely brought to Posyandu (a facility providing maternal and child health services) to monitor their growth and development	Discrete
22	Frequent illnesses	Congenital diseases in toddlers that frequently recur	Discrete
23	Availability of drinking water	The availability of clean drinking water in the surrounding environment	Discrete
24	The presence of any family member smoking in the house	The presence of smokers within the family	Discrete
25	Availability of clean water	The quality of clean water in the surrounding environment	Discrete
26	Daily fruit and vegetables consumption	The frequency of food and fruit consumption	Discrete
27	Stunting status	Stunted, Severely Stunted, or Normal	Discrete

2.2. Data preprocessing

The preprocessing stage is crucial for overall data analytics. It involved verifying missing values and encoding. Missing values often occurred due to respondents' unfamiliarity with the requested information in the instrument. Psychological factors toward their partners also influenced mothers of toddlers, leading them to omit answers to certain questions. These blank data were imputed using the K-nearest neighbors (KNN) imputer, which fills missing data by considering other features and selecting the closest value from the dataset.

Encoding involves treating discrete data features. This procedure was manually performed by grouping each category. Continuous data features were handled differently. In the NB classification procedure, probabilities were calculated using a Gaussian distribution.

2.3. MI-MOORA algorithm

This section is part of the proposed feature-weighting method. The two primary processes are MI-based feature selection and ranking using MOORA. The proposed algorithm will be presented at the end of this section.

2.3.1. Feature selection based on mutual information

MI-based feature selection is based on probability theory. MI measures the reciprocal relationship between two random variables. Using MI, the uncertainty of one variable can be decreased by presenting data from other variables [16]. Consider two random variables *X* and *Y*, with MI denoted as $I(X,Y)$, given by (1) [16]:

$$I(X,Y) = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

where *p* (*x*, *y*) is the joint probability distribution of *X* and *Y*. The symbols *p* (*x*) and *p* (*y*) represent the marginal distributions of *X* and *Y*, respectively.

Feature relevance is determined by the degree of MI between the feature and the target class. From (1), the MI score of the *k*-th feature, $I(X_k)$, is given by (2):

$$I(X_k) = I(X,C) \tag{2}$$

where C is the target class. The $I(X_k)$ score becomes the basis for ranking key features in determining the target class. The higher the $I(X_k)$ score, the higher the relevance of the feature to the class. The feature with the highest $I(X_k)$ score is then selected by applying the best *k*-feature selection threshold to determine the combination of features that best represents the model.

Feature redundancy is measured by the magnitude of the reciprocal relationship between features. The higher the I score between two features, the higher the degree of feature redundancy. A good generalization model in machine learning is indicated by a small level of redundancy between features [17]. If there are m features in a dataset, the average level of redundancy of a feature relative to other features, $R(X_k)$, is determined using (3):

$$R(X_k) = \frac{1}{m-1}\sum_{\substack{i=1\\k\neq i}}^{m} I(X_k, X_i) \tag{3}$$

2.3.2. Feature weighting by nutrition experts

Stunting cases have become a significant focus of the Indonesian government over the last decade. The government has appointed nutritionists to directly handle toddlers and children with stunted status in the field. The experience of these nutrition experts is valuable in identifying risk factors for stunting. Their evaluations provide crucial insights into the various factors contributing to stunting.

The assessments made by stunting nutrition specialists serve as an important input in determining the weighting of characteristics that are risk factors for stunting. To gather these insights, questionnaires were administered to nutrition experts who had extensive experience working with stunting cases. The feature weight criteria provided by these experts were computed based on the average score percentage of respondents, denoted as $PG(X_i)$. This value is then used to rank the influential features, ensuring the most critical factors are prioritized.

2.3.3. Calculation of preference values using MOORA

This study proposes a MOORA technique to re-rank features that have high relevance based on MI. MOORA is a multi-objective, conflicting optimization technique for solving complex decision-making problems using very simple computational steps [18]. The description of the MOORA ranking is as follows [18].

a) Create a decision matrix that shows the performance of various options based on the assessment criteria. In general, the decision matrix has the following form:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

where x_{ij} is the performance value of the *i*-th alternative on the *j*-th criterion. The symbol *n* represents the number of criteria and m indicates the number of options. The alternatives in this study are all dataset features. Meanwhile, the evaluation standards for identifying the best option (feature) are given as follows:

C1. The amount of MI of the *i*-th feature on the target ($I(X_i)$)

C2. The average amount of MI between i-th feature and other features ($R(X_i)$)
C3. The average level of influence of i-th feature according to nutrition experts ($PG(X_i)$)

The three criteria are categorized into two types according to the purpose of the ranking: benefit and non-benefit criteria. Important characteristics that contribute to determining stunting status are high mutual informant values and high weight percentages according to nutrition specialists. Conversely, these important features have low MI levels of MI with other features. Consequently, C1 and C3 are the benefit criteria, whereas C3 is a non-benefit criterion.

b) Determine the system ratio using (4) as follows:

$$x_{ij}^* = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}};\ j = 1, 2, \dots, n \tag{4}$$

where x_{ij}^* is the normalized value of the i-th alternative on the j-th criterion. In (4) is used for all values in the three criteria vectors so that the value processing is uniform in the interval [0,1].

c) Calculating feature preference values involves addressing an optimization issue with two goals: maximizing the benefit criteria and minimizing the non-benefit criteria. The overall optimization problem is given by (5):

$$y_i = \sum_{j=1}^{g} x_{ij}^* - \sum_{j=g+1}^{n} x_{ij}^* \tag{5}$$

where y_i is the preference value of the i-th alternative, g is the number of benefit criteria, and (n–g) is the number of non-benefit criteria. Specifically, the preference value for this study is given by (6):

$$y_i = I^*(X_i) + PG^*(X_i) - R^*(X_i);\ 0 \le I^*, PG^*, R^* \le 1 \tag{6}$$

where I^*, PG^*, and R^* represent the normalized values I, PG, and R. The y_i value is the basis for determining alternative rankings, and the best alternative is determined from the highest y_i value.

2.3.4. Determination of MI-MOORA feature weights

The feature weighting proposed in this study uses preference values y_i in (6). However, we cannot use the y_i value directly. There are two requirements for the y_i obtained, specifically:
y_i is non-negative if $I^*(X_i) + PG^*(X_i) \ge R^*(X_i)$
y_i is negative if $I^*(X_i) + PG^*(X_i) < R^*(X_i)$

The condition that y_i is non-negative is our objective. The greater the feature weight, the greater is the influence of the feature on the target. Nevertheless, when y_i is negative, it becomes an improper weight. Based on (6), we can get the most negative value of y_i which is -1 if $I^*(X_i) = PG^*(X_i) = 0$ and $R^*(X_i) = 1$. Therefore, to avoid this negative value, y_i is translated by one unit. Consequently, this y_i value becomes non-negative, so that the feature weights created using the preference value y_i can be written as follows:

$$w_i = 1 + y_i \tag{7}$$

we normalize (7) using (8).

$$w_i = \frac{1 + y_i}{\sum_{j=1}^{m} 1 + y_j} \tag{8}$$

The primary procedures of MI-MOORA weighting are summarized in Algorithm 1.

Algorithm 1. Feature weighting algorithm

```
Input:  n—number of features
             m—number of selected features
Output: wi—the weight of the i-th feature

1. Calculate the MI of each feature against the target class.
2. Arrange features according to the highest MI score.
3. for i  ← 1 to m do
     for j ← (i + 1) to m do
       Determine the MI between the i-th feature and the j-th feature.
     end for
      end for
4. Calculate the average MI of each feature with all features.
5. Provide the average percentage of nutrition expert ratings for each feature.
6. Create a decision matrix.
7. for xij ∈ X, i = 1, 2, …, m;  j = 1, 2, 3 do
```

```
     Normalize the x_ij using Equation 4
       end for
8. for i ← 1 to m do
     Determine the value of y_i using (6).
       end for
9. Calculate the weight of the i-th feature, w_i, using (8).
10. Sort features by greatest weight order.
```

2.4. Selected features

The subset of selected features is chosen based on their highest weight, which indicates their importance in the model. This selection process begins with determining an appropriate feature selection threshold. By setting this threshold, only the features with the highest weights are retained, ensuring that the most significant features are included in the analysis. This approach helps reduce dimensionality and improve the model's efficiency and performance.

This study employs three thresholds to select the best features: 30%, 50%, and 75%. These percentages represent the proportion of top-weighted features retained for further analysis. By experimenting with these varying thresholds, the study aims to identify the optimal subset of features that balances between including enough significant attributes and excluding less relevant ones. This method ensures the resulting model is robust and efficient, enhancing its predictive accuracy and interpretability.

2.5. Data splitting

The data-splitting stage is essential for preparing the dataset for model training and evaluation. This stage divides the dataset into two parts: the training set and the testing set. The training set, which is 80% of the data, is used to train the model, helping it learn patterns within the data. The remaining 20% (testing sets) were used to evaluate the model's performance, ensuring it generalizes well to new data. This 80:20 ratio balances sufficient data for training with enough data for accurate evaluation.

Stratified random sampling is used to achieve a representative dataset division. This method ensures that each class in the target variable is proportionally represented in the training and testing sets. By maintaining the original class distribution, stratified sampling enhances the reliability of model evaluation, preventing bias from uneven class representation. This careful data division supports the development of a robust model that can effectively handle diverse real-world instances.

2.6. Classification process

Weighted NB (WNB) is an expansion of the NB technique that assigns distinct weights to features [19]. This feature weighting method overcomes the assumption that each feature has a uniform distribution. The magnitude of the contribution of each feature represented by the weighting proposed in this research is examined to determine whether it can enhance the performance of traditional NB. An instance is classified in class C_i using a NB classifier based on (9).

$$c_j = \arg max\, P(C) \prod_{i=1}^{n} P(x_i|C) \quad (9)$$

The feature weights suggested in this study are employed in the WNB equation approach as follows [11].

$$c_j = \arg \max_{c \in C} P(C) \prod_{i=1}^{n} P(x_i|C)^{w_i} \quad (10)$$

Several classifiers were used to form a prediction model for determining stunting and to evaluate the effect of the proposed method on classifier performance. These classifiers include logistic regression (LR) [20], support vector machine (SVM) [21], XGBoost [22], random forest (RF) [23], KNN [24], and decision tree (DT) [25].

2.7. Method performance evaluation

The classification performance in this study is measured using (11) to (14) [26].

a) Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

b) Precision

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

c) Recall

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

d) F1-score

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision + Recall} \tag{14}$$

Where TP and TN indicate the number of accurate predictions for the *positive* and *negative* classes, respectively. FP and FN are the numbers of incorrect predictions in the *positive* and *negative* classes, respectively.

3. RESULTS AND DISCUSSION

Four schemes will be implemented to evaluate the performance of the NB method, which has been improved with the proposed weighting feature techniques. In the first testing scheme, the method performance was measured without any feature selection, meaning that all weighted features were deployed to build a stunting status model. The next three schemes set the number of feature selection thresholds at 30%, 50%, and 75% of the best features. The weighting technique proposed in Algorithm 1 for the WNB is referred to as the WNB-MI-MOORA. The effectiveness of WNB-MI-MOORA is compared to that of NB without feature weighting and WNB using mutual information scores as weights, called WNB-MI. In addition, six classifiers were also used to obtain a stunting prediction model and compare the effect of MI-MOORA on the model's performance. The name prefixing mechanism is also employed in other methods to differentiate each other.

3.1. Results

Table 2 presents the feature ranking order obtained based on the feature weights using the MI-MOORA framework and MI itself. Feature ranking can be seen in the ten best features. As shown in Table 2, there are variations in the ranking results for the 10 best features. The trend in the MI-MOORA ranking results showed that the stunting status of toddlers is based on family factors. Meanwhile, the results from MI demonstrate that the characteristics of the toddler dominate the important factors impacting stunting conditions.

Table 2. Feature ranking resulted by MI-MOORA and MI

Ranking	Feature name	
	MI-MOORA	MI
1	Availability of clean water	Age
2	Immunization status	UAC
3	Monthly household income	HC
4	Exclusive breastfeeding	Maternal age at pregnancy
5	Early initiation of breastfeeding	Birth weight
6	Go to *Posyandu* regularly	Length of breastfeeding
7	Toddler caregiver	Frequent illnesses
8	Length of exclusive breastfeeding	Birth height
9	Smoker dad	Father's last education
10	Maternal age at pregnancy	Father's occupation

Table 3 summarizes the performance of the NB models and the six classifier models. From these results, we can see that WNB-MI-MOORA is significantly better compared to NB. The combination of WNB-MI-MOORA improved standard NB's accuracy (Acc), precision (Pre), and F1-score (F1) by 88%, 88%, and 37.5%, respectively. However, the standard NB remains superior in the recall (Rec) term. Our result also shows that WNB-MI-MOORA has a similar performance to WNB-MI. Apart from NB, applying MI-MOORA increased the prediction accuracy for models without feature selection compared to standard classifier models, except SVM and KNN. The average improvement in the prediction model increased based on measurements of accuracy, recall, precision, and F1-score, each reaching 13.46%, 20.7%, 13.46%, and 10.75%, respectively. Unexpectedly, we found that MI-MOORA could improve the performance of tree-based classifiers, namely RF, XGBoost, and DT.

Table 3. Performance of the classification model

Models	Performance															
	% of features=100				% of features=75				% of features=50				% of features=30			
	Acc	Rec	Prec	F1	Acc	Rec	Prec	F1	Acc	Rec	Prec	F1	Acc	Rec	Prec	F1
NB (standard)	0.36	0.49	0.36	0.40	-	-	-	-	-	-	-	-	-	-	-	-
WNB-MI-MOORA	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55
WNB-MI	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55
LR	0.64	0.63	0.64	0.63	-	-	-	-	-	-	-	-	-	-	-	-
LR-MI-MOORA	0.72	0.63	0.72	0.66	0.72	0.63	0.72	0.66	**0.72**	0.63	**0.72**	0.66	0.72	0.63	0.72	0.66
LR-MI	0.72	0.65	0.72	0.67	0.72	0.63	0.72	0.66	**0.72**	0.63	**0.72**	0.66	0.76	**0.72**	0.76	**0.71**
SVM	0.68	0.46	0.68	0.55	-	-	-	-	-	-	-	-	-	-	-	-
SVM-MI-MOORA	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55
SVM-MI	0.68	0.46	0.68	0.55	0.68	0.46	0.68	0.55	0,68	0.46	0.68	0.55	0.72	0.71	0.72	0.64
RF	0.64	0.53	0.64	0.57	-	-	-	-	-	-	-	-	-	-	-	-
RF-MI-MOORA	**0.76**	**0.70**	**0.76**	**0.69**	**0.76**	**0.70**	**0.76**	**0.69**	**0.72**	0.60	**0.72**	0.63	**0.76**	0.70	**0.76**	0.69
RF-MI	0.68	0.62	0.68	0.64	0.64	0.56	0.64	0.60	0.64	0.56	0.64	0.60	0.64	0.56	0.64	0.60
XGBoost	0.56	0.56	0.56	0.56	-	-	-	-	-	-	-	-	-	-	-	-
XGBoost-MI-MOORA	0.64	0.56	0.64	0.60	0.64	0.56	0.64	0.60	0.64	0.56	0.64	0.60	0.64	0.56	0.64	0.60
XGBoost-MI	0.56	0.56	0.56	0.56	0.72	0.67	0.72	**0.69**	0.68	**0.64**	0.68	**0.66**	0.64	0.60	0.64	0.62
KNN	0.60	0.54	0.60	0.56	-	-	-	-	-	-	-	-	-	-	-	-
KNN-MI-MOORA	0.60	0.54	0.60	0.56	0.60	0.54	0.60	0.56	0.60	0.54	0.60	0.56	0.60	0.54	0.60	0.56
KNN-MI	0.60	0.46	0.60	0.52	0.60	0.46	0.60	0.52	0.60	0.46	0.60	0.52	0.64	0.53	0.64	0.57
DT	0.48	0.53	0.48	0.49	-	-	-	-	-	-	-	-	-	-	-	-
DT-MI-MOORA	0.52	0.58	0.52	0.54	0.56	0.55	0.56	0.56	0.56	0.55	0.56	0.56	0.52	0.50	0.52	0.50
DT-MI	0.52	0.56	0.52	0.54	0.52	0.53	0.52	0.52	0.44	0.44	0.44	0.44	0.40	0.48	0.40	0.44

3.2. Discussion

The proposed method, MI-MOORA, is used as a feature weighting algorithm to develop a stunting prediction model. This research utilizes external assessments to determine feature weights. We discuss nutrition expert judgment as a feature weighting criterion for NB model performance on stunting data. Feature weights indicate important features. Therefore, it can be used for feature selection. The MOORA approach allows for including external sources as feature weighting criteria alongside mutual information criteria.

Our research demonstrates that the MI-MOORA technique enhances the performance of the standard NB model and six other classifiers. However, based on the MI baseline, nutrition experts' assessments have not yet significantly improved WNB accuracy. In contrast to NB, expert nutrition assessment shows potential for improving accuracy. Previous studies also utilized machine learning with feature selection to develop nutritional status prediction models [6], [5], [27]. Consistent with prior findings, our research identifies RF as a classifier that consistently outperforms other classifiers.

The method proposed in this paper identified the availability of clean water as the most significant risk factor for stunting. Based on research data sources, we will compare these findings with interventions implemented in Surabaya City. Since 2022, governmental interventions have prioritized providing nutrition and health services, alongside support for young women, schoolchildren, and prospective couples preparing for marriage [28]. As a result, ensuring access to clean water is not deemed a primary concern in Surabaya City. The findings of this study online with earlier research [29], [30], highlighting inadequate sanitation facilities and water management as contributors to stunting. Therefore, Indonesian policies and programs must prioritize water, sanitation, and hygiene to effectively address stunting [29]. Access to clean water can enhance family hygiene practices, thereby fostering long-term improvements in children's health [30].

This research investigates stunting risk factors based on available data. The findings can serve as a reference for addressing stunting in Surabaya. However, the study's sample size is relatively small, limiting the interpretation of the results. Model validation is crucial to mitigate the impact of small sample sizes. Further comprehensive studies are necessary to validate the data, particularly concerning family-related factors such as total monthly household income, family smoking habits, and paternal occupation, which are significantly influenced by respondents' psychological conditions. Our research demonstrates the utility of external information in developing predictive models. Future studies could explore applying the weighting criteria to enhance model performance. The risk factor prediction method developed in this research could also be adapted to address other nutritional issues model development, such as obesity and wasting.

4. CONCLUSION

This study proposes a NB feature-weighting technique based on MI and MOORA. The MI score provides the degree of feature relevance to the target class and the level of feature redundancy with the other features. The addition of MOORA in this research is to re-rank features based on relevance level and redundancy, which makes it possible to involve external factors as feature weights. With MI-MOORA, the results of ranking the ten best features that affect stunting are dominated by family variables. According to the test results, applying the MI-MOORA feature weighting method significantly enhances the standard NB's accuracy, precision, and F1-score by 88%, 88%, and 37.5%, respectively. The increase in classification performance also occurred in six other classification methods when the MI-MOORA approach was applied to obtain a stunting prediction model.

ACKNOWLEDGMENTS

This research was funded by the Ministry of Education, Culture, Research, and Technology, Directorate of Research, Technology, and Community Service (DRTPM) under the "*Penelitian Kerjasama Antar Perguruan Tinggi*" scheme with No. 0165/E5.4/DT.05.00/2023. The author would also like to thank the Surabaya City Health Service, Indonesia, for providing research recommendations.

REFERENCES

[1] A. Bose, "Let us talk about stunting," *Journal of Tropical Pediatrics*, vol. 64, no. 3, pp. 174–175, Jun. 2018, doi: 10.1093/tropej/fmx104.

[2] "Central and regional stakeholders gather in Jakarta to discuss accelerating the reduction of stunting prevalence rate to 14% by 2024 (in Indonesian: bahas percepatan penurunan angka prevalensi stunting 14% pada 2024, para pemangku kepentingan pusat dan daerah berkumpul di Jakarta)," *Sekretariat Wakil Presiden*, Jakarta, Indonesia, Oct. 05, 2023.

[3] A. A. Adepoju and S. Allen, "Malnutrition in developing countries: nutrition disorders, a leading cause of ill health in the world today," *Paediatrics and Child Health*, vol. 29, no. 9, pp. 394–400, Sep. 2019, doi: 10.1016/j.paed.2019.06.005.

[4] K. G. Dewey and K. Begum, "Long-term consequences of stunting in early life," *Maternal & Child Nutrition*, vol. 7, no. s3, pp. 5–18, Oct. 2011, doi: 10.1111/j.1740-8709.2011.00349.x.

[5] S. Ndagijimana, I. H. Kabano, E. Masabo, and J. M. Ntaganda, "Prediction of Stunting among under-5 children in Rwanda using machine learning techniques," *Journal of Preventive Medicine and Public Health*, vol. 56, no. 1, pp. 41–49, Jan. 2023, doi: 10.3961/jpmph.22.388.

[6] O. N. Chilyabanyama *et al.*, "Performance of machine learning classifiers in classifying stunting among under-five children in Zambia," *Children*, vol. 9, no. 7, p. 1082, Jul. 2022, doi: 10.3390/children9071082.

[7] H. Zhang, L. Jiang, W. Zhang, and C. Li, "Multi-view attribute weighted Naive Bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 7291–7302, 2023, doi: 10.1109/TKDE.2022.3177634.

[8] G. Ou, Y. He, P. Fournier-Viger, and J. Z. Huang, "A novel mixed-attribute fusion-based Naive Bayesian classifier," *Applied Sciences*, vol. 12, no. 20, p. 10443, Oct. 2022, doi: 10.3390/app122010443.

[9] H. Ji, S. Huang, Y. Wu, Z. Hui, and C. Zheng, "A new weighted Naive Bayes method based on information diffusion for software defect prediction," *Software Quality Journal*, vol. 27, no. 3, pp. 923–968, Sep. 2019, doi: 10.1007/s11219-018-9436-4.

[10] A. Chaudhuri and T. P. Sahu, "Feature weighting for Naïve Bayes using multi objective artificial bee colony algorithm," *International Journal of Computational Science and Engineering*, vol. 24, no. 1, pp. 74–88, 2021, doi: 10.1504/IJCSE.2021.113655.

[11] L. Jiang, L. Zhang, C. Li, and J. Wu, "A correlation-based feature weighting filter for Naive Bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 201–213, Feb. 2019, doi: 10.1109/TKDE.2018.2836440.

[12] G. Verma and T. P. Sahu, "A correlation-based feature weighting filter for multi-label Naive Bayes," *International Journal of Information Technology*, vol. 16, no. 1, pp. 611–619, Jan. 2024, doi: 10.1007/s41870-023-01555-6.

[13] W. Huang, H. Yan, R. Liu, L. Zhu, H. Zhang, and H. Chen, "F-score feature selection based Bayesian reconstruction of visual image from human brain activity," *Neurocomputing*, vol. 316, pp. 202–209, Nov. 2018, doi: 10.1016/j.neucom.2018.07.068.

[14] S. Ruan, B. Chen, K. Song, and H. Li, "Weighted Naïve Bayes text classification algorithm based on improved distance correlation coefficient," *Neural Computing and Applications*, vol. 34, no. 4, pp. 2729–2738, Feb. 2022, doi: 10.1007/s00521-021-05989-6.

[15] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 15–23, Feb. 2010, doi: 10.1016/j.jbi.2009.07.008.

[16] X. Wang, B. Guo, Y. Shen, C. Zhou, and X. Duan, "Input feature selection method based on feature set equivalence and mutual information gain maximization," *IEEE Access*, vol. 7, pp. 151525–151538, 2019, doi: 10.1109/ACCESS.2019.2948095.

[17] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[18] S. Chakraborty, H. N. Datta, K. Kalita, and S. Chakraborty, "A narrative review of multi-objective optimization on the basis of ratio analysis (MOORA) method in decision making," *Opsearch*, vol. 60, no. 4, pp. 1844–1887, 2023, doi: 10.1007/s12597-023-00676-7.

[19] H. Zhang and S. Sheng, "Learning weighted naive bayes with accurate ranking," in *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004*, 2004, pp. 567–570, doi: 10.1109/ICDM.2004.10030.

[20] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection," *Computers and Security*, vol. 24, no. 8, pp. 662–674, 2005, doi: 10.1016/j.cose.2005.05.003.

[21] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803–855, Aug. 2019, doi: 10.1007/s10462-018-9614-6.

[22] N. Uzir, S. Raman, S. Banerjee, and R. S. Nishant Uzir Sunil R, "Experimenting XGBoost algorithm for prediction and classification of different datasets experimenting XGBoost algorithm for prediction and classifi cation of different datasets," *International Journal of Control Theory and Applications*, vol. 9, 2016, [Online]. Available: https://www.researchgate.net/publication/318132203.
[23] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
[24] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2888, 2003, pp. 986–996.
[25] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
[26] M. M. El Sherbiny, E. Abdelhalim, H. E.-D. Mostafa, and M. M. El-Seddik, "Classification of chronic kidney disease based on machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 32, no. 2, pp. 945–955, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp945-955.
[27] W. T. Bekele, "Machine learning algorithms for predicting low birth weight in Ethiopia," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 232, Sep. 2022, doi: 10.1186/s12911-022-01981-9.
[28] "Surabaya's secret to lowest stunting rate in Indonesia (in Indonesian: rahasia Surabaya turunkan stunting terendah se-Indonesia)," *Sekretariat Wakil Presiden*, Sep. 27, 2023.
[29] E. Yunitasari, B. O. Lee, I. Krisnana, R. Lugina, F. K. Solikhah, and R. S. Aditya, "Determining the Factors that influence stunting during pandemic in Rural Indonesia: a mixed method," *Children*, vol. 9, no. 8, p. 1189, Aug. 2022, doi: 10.3390/children9081189.
[30] H. Torlesse, A. A. Cronin, S. K. Sebayang, and R. Nandy, "Determinants of stunting in Indonesian children: evidence from a cross-sectional survey indicate a prominent role for the water, sanitation and hygiene sector in stunting reduction," *BMC Public Health*, vol. 16, no. 1, p. 669, Dec. 2016, doi: 10.1186/s12889-016-3339-8.

BIOGRAPHIES OF AUTHORS

Citra Nurina Prabiantissa received an S.ST. degree in Politeknik Negeri Malang, Indonesia, and an M.Tr.Kom. degree in Politeknik Elektronika Negeri Surabaya, Indonesia. She is a lecturer in the Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. Her research areas are artificial intelligence, image processing, and computer vision. She can be contacted at email: citranurina@itats.ac.id.

Maftahatul Hakimah received the S.Si. and M.Si degrees in the Department of Mathematics, from the Institut Teknologi Sepuluh Nopember, Indonesia. She is a lecturer in the Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. Her research areas are data mining, optimization, and artificial intelligence. She can be contacted at email: hakimah.mafta@itats.ac.id.

Nanang Fakhrur Rozi received the S.ST. degree in Politeknik Elektronika Negeri Surabaya, Indonesia, and M.Kom. in Informatics Engineering from Institut Teknologi Sepuluh Nopember, Indonesia. He is a lecturer in the Department of Informatics Engineering, Institut Teknologi Adhi Tama Surabaya, Indonesia. His research areas are information retrieval systems and software engineering. He can be contacted at email: nanang@itats.ac.id.

Ira Puspitasari received the S.T. and M.T. degrees in the Department of Informatics Engineering from Institut Teknologi Bandung, Indonesia, and received a Ph.D. degree in Information and Physical Sciences from Osaka University, Japan. She is a lecturer in the Department of Information Systems at Airlangga University, Indonesia. Her research areas are consumer health informatics, data analytics, specialized enterprise architecture, and information systems theory. She can be contacted at email: ira-p@fst.unair.ac.id.

Laura Navika Yamani received her S.Si. and M.Si. degrees in the Department of Chemistry from Airlangga University, Indonesia. She received a Ph.D. in the Division of Molecular Medicine and Medical Genetics from Kobe University, Japan. She is a lecturer in the Department of Epidemiology, Population Biostatistics, and Health Promotion at Airlangga University, Indonesia; a researcher at the Institute for Tropical Diseases, Airlangga University with research experience related to biotechnology, molecular epidemiology, and bio-molecular tropical or infectious diseases in Indonesia from 2009 until now; and Chair of the Research Center on Global Emerging and Re-emerging Infectious Diseases (RC¬GERID), Institute of Tropical Disease, Airlangga University. Her research areas are topics related to health. She can be contacted at email: laura.navika@fkm.unair.ac.id.

Victoria Lucky Mahendra is a student of Informatics Engineering at the Institut Teknologi Adhi Tama Surabaya, Indonesia. He is interested in studying information retrieval systems, data mining, and artificial intelligence. He can be contacted at email: victoriamahendra@gmail.com.