Implementation of the Naive Bayes Classifier Algorithm for Classifying Toddler Nutritional Status

Muhammad Insan Kamil [1]*, Adityo Permana Wibowo [2]**
* Informatics, University of Technology Yogyakarta
muhikamil@gmail.com [1], adityopw@uty.ac.id[2]

Article Info

Article history:

Received 2024-10-10
Revised 2024-10-23
Accepted 2024-11-04

Keyword:

Naive Bayes Classifier,
Nutritional Status,
Toddlers,
Malnutrition,
Class Imbalance.

ABSTRACT

This research addresses the pressing issue of malnutrition among toddlers in Indonesia, aiming to classify their nutritional status using the Naive Bayes Classifier (NBC). The study utilizes a dataset comprising 958 records from Puskesmas Cilandak and categorizes nutritional status into six class labels: good nutrition, at risk of excess nutrition, excess nutrition, obesity, undernutrition, and severe malnutrition. The methodology includes data preprocessing techniques such as class weighting to tackle class imbalance and Principal Component Analysis (PCA) for effective feature extraction. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1 score, achieving an impressive accuracy of 85.76% when class weighting is applied, which significantly enhances the recall and F1 scores for minority classes. The findings highlight the critical importance of robust preprocessing and evaluation metrics in improving machine learning models for public health applications. Furthermore, they suggest that further exploration of alternative algorithms and dataset expansion could yield more comprehensive insights into the classification of toddler nutritional status.

I. INTRODUCTION

Addressing the issue of malnutrition in children is crucial, as a significant number of Indonesian children continue to experience nutritional deficiencies, including severe malnutrition and stunting. Data from the Jakarta Health Department as of July 2023 indicate that 39,793 toddlers have been reported to suffer from nutritional disorders. Furthermore, research from the Center of Indonesian Studies highlights that approximately 21 million Indonesians, or about 7% of the total population, are affected by critical nutritional deficiencies. Malnutrition and stunting represent significant public health challenges, necessitating a coordinated response from both central and local governments, alongside increased public awareness regarding the importance of balanced nutrition, exclusive breastfeeding, healthy dietary practices, and the early identification of malnutrition in children [1] . Malnutrition in early childhood negatively impacts motor development, hinders behavioral and cognitive growth, and ultimately leads to a decline in learning and social skills, emphasizing the urgent need for effective interventions to combat these issues [2].

The assessment and determination of toddlers' nutritional status currently require a structured and systematic approach. According to the Indonesian Health Minister’s Regulation No. 2 of 2020 on child anthropometric standards, nutritional status in toddlers can be evaluated using anthropometric methods. Key indices include weight-for-age (W/A), weight-for-height (W/H), and height-for-age (H/A). Among children under 60 months, the weight-for-age (W/A) index is frequently used. However, this method can be time-consuming and may still present challenges in accurately determining nutritional status.[3].

Machine learning is a powerful technology that facilitates classification, clustering, and prediction by utilizing datasets[4]. Classification in machine learning is based on predefined attributes, and one widely used classification algorithm is the Naive Bayes Classifier (NBC).[5], [6]. Naïve Bayes is a classification method that relies on straightforward probability calculations [7].The Naive Bayes Classifier assists nutritionists in classifying toddler nutritional data, determining whether a child falls into categories such as good nutrition, moderate malnutrition, severe malnutrition, or stunting[8], [9]. This classification process is performed efficiently to guide subsequent interventions, which are later reassessed using anthropometric measurements. NBC functions by employing probabilistic methods to predict future outcomes based on historical data, making it a statistical tool for predicting class membership [5]. Naive Bayes has been shown to achieve high accuracy and speed when applied to large datasets[6], and it also has the advantage of maintaining high accuracy even with a limited amount of data [10]. One advantage of using Naive Bayes is that it requires only a small amount of training data to estimate the mean and variance of the variables needed for classification [6].

Several recent studies have explored the application of machine learning algorithms to classify the nutritional status of children, highlighting different approaches and findings. Yuliansyah et al. [7] NBC achieved a notable accuracy of 86.45%, yet its ability to handle class imbalance was inadequate compared to K-Nearest Neighbors (KNN), as evidenced by its lower F1 score of 32.53%. This suggests that while NBC performs adequately in terms of overall accuracy, it tends to misclassify minority classes, such as malnutrition, due to its simplistic probabilistic assumptions. Similarly, in Solok City [9], NBC achieved an accuracy of 90.94%, but KNN surpassed it in precision and recall metrics, indicating NBC's limited capability in managing complex feature interactions. Moreover, a web-based application for nutritional status classification in North Jakarta [10] reported that NBC achieved 80.60% accuracy, but once again KNN outperformed it, emphasizing NBC's struggles with datasets containing diverse feature sets. These studies collectively underscore NBC's primary drawback when dealing with imbalanced data, particularly in the context of public health, where accurate classification of minority classes is crucial for effective interventions. This research aims to address these limitations by employing NBC alongside advanced preprocessing techniques such as class weighting and oversampling to handle class imbalances. This improvement ensures better classification for underrepresented categories such as malnutrition, offering a more robust and precise tool for assessing toddler nutritional status. This research builds on the prior works by enhancing the model’s sensitivity to minority classes, thereby providing a more comprehensive solution to the challenges identified in earlier research. By addressing the critical issue of class imbalance, this research demonstrates how Naïve Bayes can be optimized for practical public health applications, offering a more balanced and effective approach for the classification of toddler nutritional health.

II. METHOD

This study follows a structured methodology consisting of several essential steps aimed at accurately classifying toddler nutritional status using the Naïve Bayes Classifier, with a focus on addressing data imbalance through the application of oversampling techniques. Method research flow as shown at Figure 1.

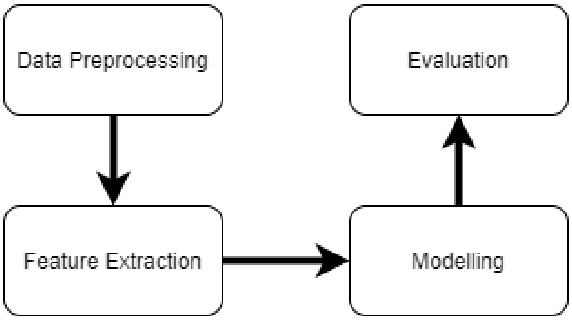


Figure 1. Method Research Flow

A. *Data Preprocessing*

The research process begins with data preprocessing, the initial dataset obtained from Puskesmas Cilandak, Kelurahan Cilandak Barat consisted of 958 records, with relevant variables such as height, weight, age, gender, and nutritional status. This dataset represents children from a specific region, and therefore, may not fully capture the diversity of the broader toddler population in Indonesia. As a result, the findings may have limited generalizability to the entire population of toddlers in Indonesia, the map as shown at Figure 2.

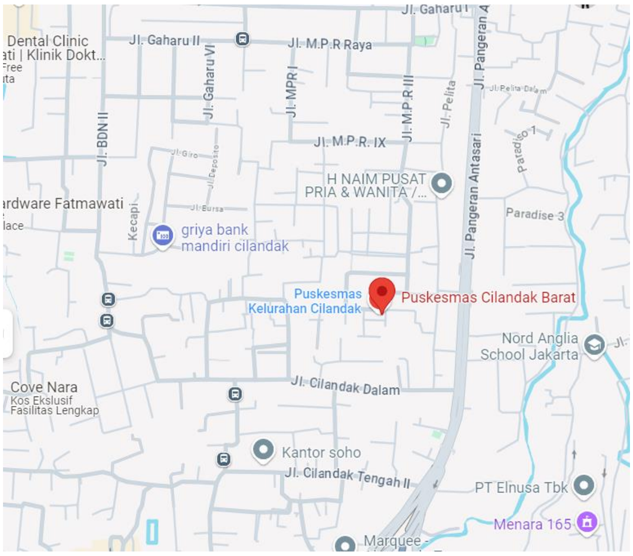


Figure 2. Location of Research Object

Ensuring the quality and consistency of this dataset is crucial, so several preprocessing steps were applied. Notably, this dataset did not contain any missing values, eliminating the need for imputation techniques. Instead, preprocessing focused on encoding categorical variables and scaling processes to ensure uniformity across numerical variables. A critical aspect of this stage involved handling the issue of class imbalance within the dataset, which was addressed using

Random Oversampling to increase the representation of minority classes [11], [12]. This technique improved the model's ability to enhance classification accuracy for underrepresented categories. Additionally, as an alternative to oversampling, class weighting could be employed to adjust the model's sensitivity to the minority class, thus offering another approach to mitigating class imbalance. The results from both methods, oversampling and class weighting, will be compared, and the method that yields the highest accuracy will be selected for subsequent processes.

B. Feature Extraction

Following data preprocessing, feature extraction was conducted to utilize all relevant attributes for the classification task. Key variables such as weight, height, and age were included due to their significance in predicting nutritional status. To further optimize the model's performance, Principal Component Analysis (PCA) was employed. Principal Component Analysis (PCA) is a statistical method used to identify patterns within data and highlight differences or similarities across a dataset. It is commonly applied as a dimensionality reduction tool, transforming high-dimensional data into a simpler form while retaining the most critical information. This is achieved by calculating the covariance matrix of the data, followed by determining its eigenvectors and eigenvalues. These components represent the principal directions in which the data varies the most. In addition to dimensionality reduction, PCA serves as a useful method for assessing whether variables in a dataset are correlated or independent of one another. By transforming data vectors into a simpler form, PCA allows for a more descriptive and interpretable representation of complex datasets, making it a valuable tool in various fields of research and analysis.[13], [14], [15] This method allowed for dimensionality reduction by transforming the original features into a smaller set of uncorrelated components, which retained most of the variance present in the data. The use of PCA enhanced the model's efficiency and accuracy by focusing on the most informative features while reducing noise and redundancy, ensuring a robust classification of toddler nutritional status.

C. Modelling

In the modeling phase, the Naïve Bayes Classifier is employed as the core algorithm. This model is chosen due to its simplicity and efficiency in handling classification tasks, particularly when dealing with relatively small training datasets. The model is trained using an appropriate data split, such as 70% for training and 30% for testing, with hyperparameter tuning and cross-validation applied to ensure optimal performance. Oversampling is integrated into this stage to address the imbalance within the dataset, improving the model's ability to classify minority classes accurately. Additionally, experiments will be conducted to evaluate the effectiveness of class weighting as an alternative method for handling data imbalance. This approach will allow for a comparative analysis of both techniques, providing insights into their respective impacts on the model's performance.

D. Evaluation

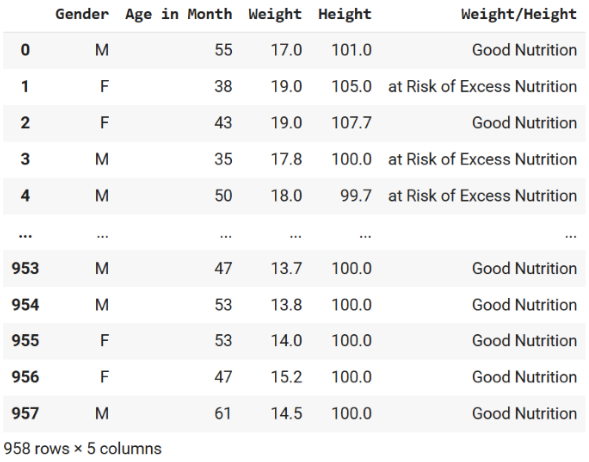
The evaluation of the model's performance is conducted using several key metrics, including accuracy, precision, recall, F1 score, and support to provide a comprehensive assessment. A comparison is made between the performance of the Naïve Bayes Classifier with and without oversampling, allowing for an analysis of the effectiveness of this technique in addressing class imbalance. Validation methods such as cross-validation are employed to ensure that the model's results are robust and generalizable to new data.

III. Results and Discussion

In this section, we present the findings of our research, detailing the outcomes of the data preprocessing, feature selection, and the application of the Naïve Bayes Classifier in classifying toddler nutritional status. The results are discussed in the context of their implications for improving classification accuracy and addressing class imbalance in public health data.

A. Data Preprocessing

The initial dataset obtained from Puskesmas Cilandak Kelurahan Cilandak Barat consisted of 958 records with relevant variables such as height, weight, age, gender, and nutritional status, as shown in Figure 3. It is important to note that the dataset did not contain any missing values, eliminating the need for imputation techniques such as mean, median, or mode filling. As a result, the data preprocessing phase focused on other tasks, such as encoding categorical variables and scaling numerical features to ensure consistency across the dataset.



	Gender	Age in Month	Weight	Height	Weight/Height
0	M	55	17.0	101.0	Good Nutrition
1	F	38	19.0	105.0	at Risk of Excess Nutrition
2	F	43	19.0	107.7	Good Nutrition
3	M	35	17.8	100.0	at Risk of Excess Nutrition
4	M	50	18.0	99.7	at Risk of Excess Nutrition
...	...	...	...	...	...
953	M	47	13.7	100.0	Good Nutrition
954	M	53	13.8	100.0	Good Nutrition
955	F	53	14.0	100.0	Good Nutrition
956	F	47	15.2	100.0	Good Nutrition
957	M	61	14.5	100.0	Good Nutrition

958 rows × 5 columns

Figure 3. Sample Data

In addressing the issue of class imbalance within the dataset, several preprocessing steps were taken to optimize the model's performance. Encoding categorical variables and scaling were applied to ensure consistency across the dataset, standardizing the numerical features and transforming categorical labels for better model interpretation. Additionally, random oversampling was used to handle the imbalance in minority classes, particularly malnutrition and stunting. By duplicating instances of the underrepresented categories, this technique helped mitigate the risk of biased predictions that could favor the majority class. The combination of these preprocessing steps improved the model's ability to accurately classify cases of malnutrition. Figure 5 shows the significant class imbalance in the initial dataset before oversampling, where the minority classes, such as malnutrition and stunting, were underrepresented. After applying random oversampling, Figure 6 demonstrates a more balanced distribution of classes, resulting in improved classification accuracy for the minority categories.

```
numeric_tr                                             =
dtrain.select_dtypes(include=np.number)
numeric_columns_tr = numeric_tr.columns
scaler = StandardScaler()
dtrain[numeric_columns_tr] =
scaler.fit_transform(dtrain[numeric_columns_tr])

object_columns_tr =
dtrain.select_dtypes(include='object').columns.to
list() #menyimpan kolom bertipe object
object_columns_tr.remove('Weight/Height')
label_encoder = LabelEncoder()
for kolom in object_columns_tr:
  dtrain[kolom] =
label_encoder.fit_transform(dtrain[kolom])

x = dtrain.drop('Weight/Height',axis=1)
y = dtrain['Weight/Height']

oversampler = RandomOverSampler(random_state=42)
X_resampled, y_resampled =
oversampler.fit_resample(x, y)
```

Figure 4. Encoding, scalling, and oversampling code

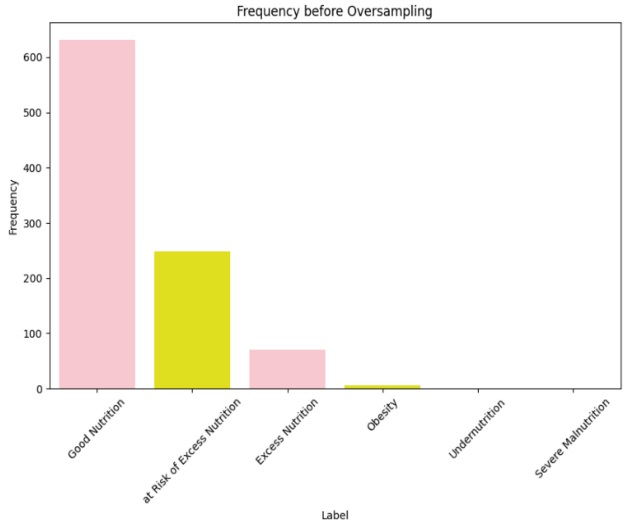


Figure 5. Imbalance Data Before Oversampling

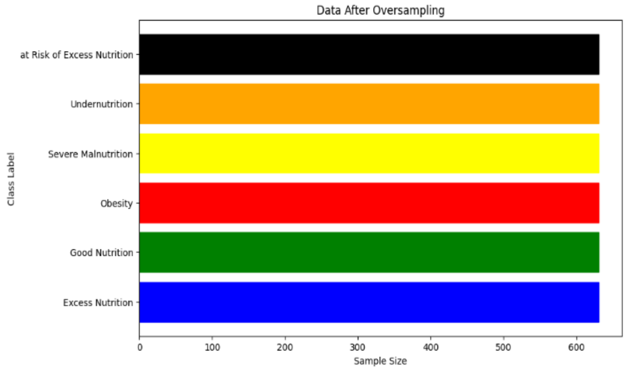
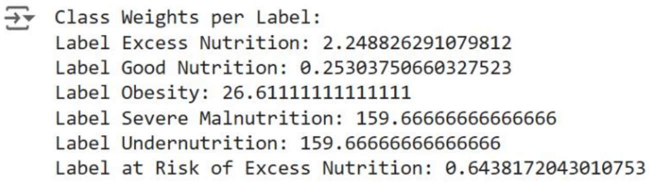


Figure 6. Result of Oversampling imbalance data

In addition to oversampling, class weighting was applied as an alternative method to address the class imbalance within the dataset. Class weighting adjusts the model to assign higher importance to underrepresented classes, such as malnutrition and stunting, by penalizing misclassifications more heavily in these categories. This approach allows the model to be more sensitive to minority classes without altering the size of the dataset. Figure 7 illustrates the source code used to implement class weighting, showing how the model was configured to handle imbalanced classes. The use of class weighting resulted in an improved ability to detect malnutrition cases, enhancing the model's recall for these categories. Figure 8 presents the results of the class weighting, demonstrating balanced improvements in classification performance across all classes, particularly by reducing bias towards the majority class. While class weighting did not significantly alter the overall accuracy compared to oversampling, it provided a more nuanced improvement in handling the minority classes.

```
class_weights =
compute_class_weight('balanced',
classes=np.unique(y), y=y)
class_weight_dict = dict(zip(np.unique(y),
class_weights))
print("Class Weights per Label:")
for label, weight in
class_weight_dict.items():
    print(f"Label {label}: {weight}")
```

Figure 7. Class Weight implementation code



```
Class Weights per Label:
Label Excess Nutrition: 2.248826291079812
Label Good Nutrition: 0.25303750660327523
Label Obesity: 26.611111111111111
Label Severe Malnutrition: 159.66666666666666
Label Undernutrition: 159.66666666666666
Label at Risk of Excess Nutrition: 0.6438172043010753
```

Figure 8. Result of Class Weighting for imbalance data

B. Feature Extraction

Feature extraction was conducted to utilize all relevant attributes for the classification task. Key variables such as weight, height, and age were included due to their significance in predicting nutritional status. To further enhance model accuracy and efficiency, Principal Component Analysis (PCA) was employed for feature extraction. PCA is a statistical technique that transforms the original correlated variables into a new set of uncorrelated variables known as principal components, which capture the maximum variance in the data. The process begins with standardizing the data to ensure each feature contributes equally[14], represented mathematically as shown at equation 1:

$$Zij = \frac{Xij-\mu j}{\sigma j} \ (1)$$

Where Zij is the standardized value, Xij is the original value, μj is the mean of the j-th variable, and σj is the standard deviation[16]. Following standardization, the covariance matrix is calculated to understand the relationships between the variables as shown at equation 2:

$$C = \frac{1}{n-1} Z^T Z \ (2)$$

Here, C denotes the covariance matrix, and Z is the matrix of standardized values[16]. Subsequently, eigenvalues and eigenvectors are derived from the covariance matrix through the equation 3:

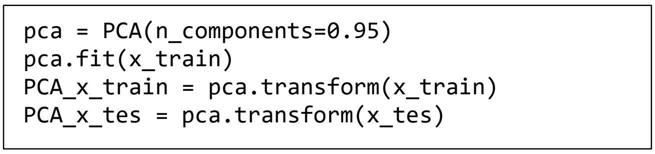
$$Cv = \lambda v \ (3)$$

Where v represents the eigenvector and λ the corresponding eigenvalue. These eigenvectors are then ordered by their eigenvalues to select the most significant principal components[16]. Finally, the original dataset is projected onto the selected components using the transformation as shown at equation 4:

$$Y = XW \ (4)$$

In this equation, Y is the transformed dataset, X is the original dataset, and W is the matrix of selected eigenvectors. This dimensionality reduction technique not only simplifies the dataset but also enhances the model's predictive capabilities by focusing on the most informative features[16].

The PCA reduced the dimensionality of the dataset, transforming it into a set of principal components while retaining 95% of the variance from the original data as shown at Figure 7. This process minimized noise and redundancy, allowing the Naïve Bayes Classifier to operate with reduced computational complexity while maintaining high accuracy.



```
pca = PCA(n_components=0.95)
pca.fit(x_train)
PCA_x_train = pca.transform(x_train)
PCA_x_tes = pca.transform(x_tes)
```

Figure 9. Feature Extraction Code

The application of PCA improved the model’s performance by focusing on the most informative features. This ensured that the classifier was not overwhelmed by irrelevant data, contributing to faster training and better predictive accuracy. Additionally, three different data split scenarios were implemented, as illustrated in Table 1, to evaluate the model’s performance under varying training and testing proportions. These split scenarios aided in assessing the robustness of the classifier and ensuring its generalizability across different data partitions..

TABLE I
SPLITTING DATASET

Scene	Percentage		Description	
	Training	Testing	Training	Testing
Scene 1	70%	30%	2650	1136
Scene 2	80%	20%	3028	758
Scene 3	90%	10%	3407	379

C. Modelling

The Naïve Bayes Classifier was applied to the preprocessed dataset, both before and after the oversampling technique was implemented. This classifier was chosen due to its simplicity and efficiency, particularly in handling classification tasks with relatively small datasets. Despite its assumption of feature independence which may not always hold true in health related data the Naïve Bayes Classifier has demonstrated robust performance in various applications. Additionally, its computational efficiency makes it well-suited for environments with limited resources. In this study, the application of Principal Component Analysis (PCA) helped to mitigate the impact of feature dependence by transforming correlated variables into uncorrelated principal components, further enhancing the suitability of the Naïve Bayes approach for our analysis. The general equation of Naïve Bayes is as shown at equation 5.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (5)$$

Where B is the Data with an unknown class, A is the Hypothesis that data B belongs to a specific class, P(A|B) is the Probability of hypothesis A given the condition B, P(B|A) is the Probability of hypothesis B given the condition A, P(B) is the Probability of hypothesis B[17].

The model was evaluated using a 70:30 train-test split, with cross-validation applied to validate the results. Without oversampling, the model showed an accuracy of 67.15%, but struggled with precision and recall for the minority classes. After applying imblearn, the classifier’s accuracy improved to 71.12%, with notable improvements in recall and F1 scores for underrepresented classes such as severe malnutrition and stunting. Subsequently, experiments using class weighting resulted in a significantly higher accuracy of 85.76%. This indicates that class weighting provided better performance than simply using random oversampling, demonstrating its effectiveness in addressing class imbalance and enhancing the model's ability to classify minority classes accurately. The code implementation of the Naïve Bayes model used in this study is illustrated in Figure 8, demonstrating the steps taken to train and evaluate the classifier.

```
model = GaussianNB()
model.fit(PCA_x_train,y_train)

y_predptb = model.predict(PCA_x_tes)
scores = cross_val_score(model, PCA_x_train,
y_train, cv=5)
mean_accuracy = scores.mean()
print(mean_accuracy)
print('Predict:',y_predptb[0])
print('Actual:',y_tes.values[0])
```

Figure 10. Modelling code

D. Evaluation

The results demonstrated that class weighting significantly impacted the classifier’s performance, particularly in addressing class imbalance. Experiments showed that using class weighting resulted in a remarkably higher accuracy of 85.76%. Class weighting works by assigning higher penalties to misclassifications of minority classes during the training process, allowing the model to become more sensitive to these underrepresented categories without altering the dataset's size. This approach highlighted the importance of addressing class imbalance in datasets related to public health, especially when using machine learning models to classify nutritional status. The improvement in recall for minority classes underscores the need for balancing datasets when working with imbalanced health data. After applying class weighting, the model achieved an F1 score of 85% and a recall of 80%, as shown in Figure 11, demonstrating a substantial improvement in its ability to correctly classify underrepresented categories.

To further elucidate the classification performance, several key metrics were employed: accuracy, recall, and precision[10]. Accuracy is defined as the ratio of correctly predicted instances to the total instances in the dataset, represented as shown at equation 6:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

where TP (True Positives) is the number of correct positive predictions, TN (True Negatives) is the number of correct negative predictions, FP (False Positives) is the number of incorrect positive predictions, and FN (False Negatives) is the number of incorrect negative predictions[10].

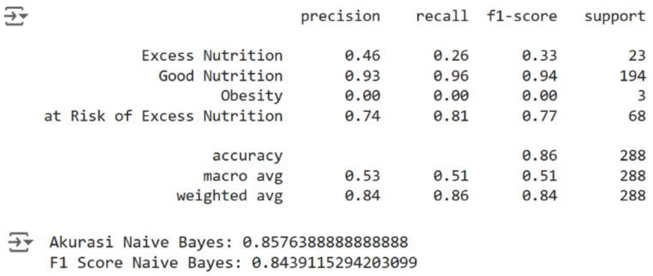
Recall, also known as sensitivity or the true positive rate, measures the model's ability to identify actual positive instances and is calculated as shown at equation 7[10]:

$$Recall = \frac{TP}{TP+FP} \quad (7)$$

Finally, precision, which evaluates the accuracy of positive predictions[10], is defined as shown at equation 8:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

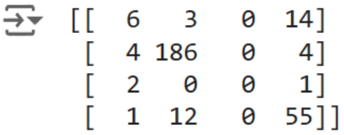
These metrics collectively provide a comprehensive view of the model's performance, especially in the context of class imbalance, where accuracy alone may not adequately reflect the model's ability to classify minority classes. Additionally, the application of PCA as a feature extraction method contributed to both the computational efficiency and the accuracy of the model, supporting the hypothesis that dimensionality reduction is essential in optimizing classification tasks.



```
                           precision    recall  f1-score   support

          Excess Nutrition      0.46      0.26      0.33        23
            Good Nutrition      0.93      0.96      0.94       194
                   Obesity      0.00      0.00      0.00         3
at Risk of Excess Nutrition     0.74      0.81      0.77        68

                  accuracy                          0.86       288
                 macro avg      0.53      0.51      0.51       288
              weighted avg      0.84      0.86      0.84       288

Akurasi Naive Bayes: 0.8576388888888888
F1 Score Naive Bayes: 0.8439115294203099
```

Figure 11. Evaluation Matrix and Accuracy

The confusion matrix, as shown in Figure 12, provides a comprehensive overview of the classifier's performance across different nutritional status categories. The first row indicates that out of 23 instances classified as "Excess Nutrition," 6 were correctly identified, while 3 were misclassified as "Good Nutrition," and 14 were incorrectly categorized as "at Risk of Excess Nutrition." In the second row, among the 194 "Good Nutrition" cases, 186 were accurately classified, with 4 misclassified as "Excess Nutrition" and none as "Obesity." The third row reveals that out of 3 instances labeled as "Obesity," none were correctly identified; instead, 2 were misclassified as "Excess Nutrition" and 1 as "at Risk of Excess Nutrition." Finally, the fourth row indicates that for the 68 instances categorized as "at Risk of Excess Nutrition," the model successfully classified 55, with 12 misclassified as "Good Nutrition" and 1 as "Excess Nutrition." Overall, the confusion matrix illustrates the model's strengths in accurately identifying "Good Nutrition" and "at Risk of Excess Nutrition," while also highlighting the challenges in classifying "Excess Nutrition" and "Obesity."



```
[[  6   3   0  14]
 [  4 186   0   4]
 [  2   0   0   1]
 [  1  12   0  55]]
```

Figure 12. Confusion Matrix

IV. CONCLUSION

Based on the results and discussion, it can be concluded that the implementation of the Naïve Bayes Classifier significantly enhanced the classification accuracy of toddler nutritional status. The study achieved an impressive accuracy rate of 85.76%, underscoring the effectiveness of addressing class imbalance through strategies such as class weighting. This research emphasizes the importance of employing robust preprocessing techniques and utilizing model evaluation metrics, including precision and recall, to improve the performance of machine learning models in public health applications.

The strengths of this study lie in its comprehensive approach to data preprocessing and the successful application of the Naïve Bayes Classifier, which effectively classified underrepresented categories within the dataset. These findings underscore the critical need to address class imbalance to ensure accurate nutritional assessments for toddlers, ultimately contributing to improved public health outcomes. For future research, it is anticipated that further exploration of alternative classification algorithms and additional balancing techniques will be pursued to enhance both the accuracy and generalizability of the findings across diverse populations. Moreover, expanding the dataset to encompass a broader demographic could yield valuable insights into the nutritional status of toddlers across various regions, thereby increasing the overall effectiveness of nutritional classification tools.

References

[1] Eva Safitri, “Soroti Gizi Buruk Anak, Kris Dayanti Dorong Pemerintah Fokus Kedaulatan Pangan,” detiknews. Accessed: Oct. 08, 2024. [Online]. Available: https://news.detik.com/berita/d-6979677/soroti-gizi-buruk-anak-kris-dayanti-dorong-pemerintah-fokus-kedaulatan-pangan

[2] E. Ramon, A. Nazir, and L. Oktavia, “Klasifikasi Status Gizi Bayi Posyandu Kecamatan Bangun Purba Menggunakan Algoritma Support Vector Machine (SVM),” *Jurnal Sistem Informasi dan Informatika (Simika) P-ISSN*, vol. 5, pp. 2622–6901, 2022.

[3] D. Gizi, M. Direktorat, J. Kesehatan, M. Kementerian, and K. 2018, “Hasil Pemantauan Status Gizi (PSG) Tahun 2017.”

[4] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python A Guide For Data Scientists Introduction to Machine Learning with Python*. Gravenstein Highway North, Sebastopol: O’Reilly Media, Inc., 2016.

[5] Harliana and D. Anggraini, “Penerapan Algoritma Naïve Bayes Pada Klasifikasi Status Gizi Balita di Posyandu Desa Kalitengah (Harliana, Dewi Anggraini),” *Jurnal Informatika Komputer, Bisnis, dan Manajemen*, vol. 21, no. 2, 2023.

[6] S. Alim, “Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes Orange Data Mining Implementation For Student Graduation Classification Using K-Nearest Neighbor, Decision Tree And Naive Bayes Models,” *Jurnal Ilmiah NERO*, vol. 6, no. 2, 2021.

[7] Moch. Rizky Yuliansyah, M. B, and A. Franz, “Perbandingan Metode K-Nearest Neighbors dan Naïve Bayes Classifier Pada Klasifikasi Status Gizi Balita di Puskesmas Muara Jawa Kota Samarinda,” *Adopsi Teknologi dan Sistem Informasi (ATASI)*, vol. 1, no. 1, pp. 08–20, Jun. 2022, doi: 10.30872/atasi.v1i1.25.

[8] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, “Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 257–265, Jan. 2024, doi: 10.57152/malcom.v4i1.1078.

[9] S. Kenia, P. Loka, and A. Marsal, “Comparison Algorithm of K-Nearest Neighbor and Naïve Bayes Classifier for Classifying Nutritional Status in Toddlers Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier Untuk Klasifikasi Status Gizi Pada Balita,” *MALCOM : Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, pp. 8–14, 2023.

[10] R. Setiawan and A. Triayudi, “Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web,” *Jurnal Media Informatika Budidarma*, vol. 6, no. 2, p. 777, Apr. 2022, doi: 10.30865/mib.v6i2.3566.

[11] G. Gumelar, Q. Ain, R. Marsuciati, S. Agustanti Bambang, A. Sunyoto, and M. Syukri Mustafa, “Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance,” in *Prosiding Seminar Nasional SISFOTEK*, 2021, pp. 250–255.

[12] H. Said, N. Matondang, H. Nurramdhani Irmanda, and S. Informasi, “Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi Application of K-Nearest Neighbor Algorithm to Predict Consumable Water Quality,” *Techno COM*, vol. 21, no. 2, pp. 256–267, 2022, [Online]. Available: www.kaggle.com

[13] D. Aditya Nugraha and A. Sartika Wiguna, “Seleksi Fitur Warna Citra Digital Biji Kopi Menggunakan Metode Principal Component Analysis Digital Image Selection of Coffee Seed Using Component Analysis Method,” 2020.

[14] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, “Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver,” *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 281–290, Feb. 2023, doi: 10.35746/jtim.v4i4.298.

[15] S. Raysyah, V. Arinal, and D. I. Mulyana, “Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode KNN dan PCA,” *Jurnal Sistem Informasi (JSiI)*, vol. 8, no. 2, pp. 88–95, 2021.

[16] Sumaiya Sande, “Principal Component Analysis (PCA) : Theory,” Analytics Vidhya, Medium. Accessed: Oct. 10, 2024. [Online]. Available: https://medium.com/analytics-vidhya/principal-component-analysis-theory-bc87ef8c31af

[17] T. Mitchell and M. Hill, *Chapter 3 Generative And Discriminative Classifiers: Naive Bayes And Logistic Regression Machine Learning 1 Learning Classifiers based on Bayes Rule*, vol. 3. CMU School of Computer Science, 2017. [Online]. Available: www.cs.cmu.edu/~tom/mlbook.html.