Karaelmas Science and Engineering Journal
Jorunal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1555212

Research Article
Received / Geliş tarihi : 24.09.2024
Accepted / Kabul tarihi : 04.11.2024

Performance Comparison of Machine Learning Algorithms Using Oversampling Methods to Predict Childhood Anemia

Çocukluk Çağı Anemisinin Tahmininde Aşırı Örnekleme Yöntemlerini Kullanan Makine Öğrenmesi Algoritmalarının Performans Karşılaştırması

Kadriye Filiz Balbal*

Dokuz Eylul University, Faculty of Science, Department of Computer Science, Izmir, Türkiye

Abstract

Childhood anemia is a significant health problem. Anemia, which is common among preschool-aged children, causes physical and mental developmental delays in this age group. Therefore, this study employs machine learning techniques, a subfield of artificial intelligence, to predict the anemia levels in children aged 0–59 months in Nigeria. To address the issue of data imbalance, which can cause problems in estimating childhood anemia levels, the SMOTE and ADASYN oversampling techniques were employed in this study. The analyses performed with the newly obtained balanced data, showed that the SMOTE and ADASYN methods performed significantly better than those obtained with imbalanced data for all machine learning models. When the average results of all machine learning algorithms used in this study in terms of accuracy, precision, recall, and F1 score metrics are compared to the oversampling methods, the most successful result in terms of all metrics was obtained with the SMOTE method.

Keywords: ADASYN, artificial intelligence, childhood anemia, machine learning, SMOTE.

Öz

Çocukluk çağı anemisi önemli bir sağlık sorunudur. Okul öncesi çağdaki çocuklarda sık görülen anemi, bu yaş grubunda fiziksel ve zihinsel gelişimsel gecikmelere neden olur. Bu nedenle, bu çalışmada Nijerya'da 0-59 aylık çocuklarda anemi düzeylerini tahmin etmek için yapay zekanın bir alt alanı olan makine öğrenmesi teknikleri kullanılmıştır. Çocukluk çağı anemi düzeylerini tahmin etmede sorunlara neden olabilen veri dengesizliği sorununu ele almak için çalışmada, SMOTE ve ADASYN aşırı örnekleme teknikleri kullanılmıştır. Yeni elde edilen dengeli verilerle yapılan analizlerde, SMOTE ve ADASYN yöntemlerinin tüm ML modelleri için dengesiz verilerle elde edilen sonuçlardan önemli ölçüde daha iyi performans gösterdiği görülmüştür. Çalışmada kullanılan tüm ML algoritmalarının doğruluk, kesinlik, duyarlılık ve F1 puanı metrikleri açısından ortalama sonuçları, aşırı örnekleme yöntemleri ile karşılaştırıldığında, tüm metrikler açısından en başarılı sonuç SMOTE yöntemi ile elde edilmiştir.

Anahtar Kelimeler: ADASYN, yapay zeka, çocukluk çağı anemisi, makine öğrenmesi, SMOTE.

1. Introduction

Artificial intelligence improves its ability to perform complex tasks day by day. Machine Learning (ML), a fundamental field of artificial intelligence, is successfully used in making decisions by learning from data, detecting relationships between variables, and predicting results based on inputs. As in every area, ML applications are becoming widespread in medical sciences (Tesfaye et al. 2024). ML methods are frequently used in the medical field, especially in disease diagnosis and prediction, and successful results are obtained (Rahmani et al. 2021).

Childhood anemia is a major global health problem. According to the World Health Organization (WHO), approximately 269 million children under the age of five are affected by anemia. Children under the age of 5 are among the most vulnerable to anemia. Anemia, which is treatable and preventable, can cause cognitive and developmental disorders in children, especially in severe cases (WHO 2023).

Although it is a common and important problem, few studies in the literature have attempted to predict childhood

*Corresponding author: kadriyefiliz.balbal@deu.edu.tr

Kadriye Filiz Balbal orcid.org/0000-0002-7215-9964

anemia using machine learning methods (Ajakaye and Ibukunoluwa 2020, Anand et al. 2019, Aynalem et al. 2022, Bitew et al. 2022, Getawa et al. 2020, Kebede Kassaw et al. 2023, Marcos Valdez et al. 2023, Tesfaye et al. 2024). Tesfaye et al. (2024) attempted to predict childhood anemia using sociodemographic and mother-child variables with machine learning algorithms on 2016 Ethiopia Demographic Health Survey (EDHS) data. In their study, which employed logistic regression, random forest, decision tree, and K-nearest neighbor algorithms, the researchers observed that the prediction performance of the machine learning algorithms ranged from 60% to 66%. The findings revealed that logistic regression represented the most accurate prediction model, with a 66% success rate. Zemariam et al. (2024) aimed to determine the most effective machine learning model for predicting anemia by utilizing data from 5,642 young girls within the same dataset. A study using the Boruta algorithm for feature selection stated that the Random forest classifier was the best for predicting anemia. Khan et al. (2021), who tried six machine learning algorithms to predict the anemia status of children under five years of age in Bangladesh, the k-nearest neighbor algorithm gave the lowest accuracy rate, while the Random Forest algorithm gave the highest rate of 68.53%. It achieved high classification accuracy.

In addition to the studies mentioned above, some studies use image processing and deep learning methods to diagnose and predict anemia in children. When all these studies are examined, it is seen that the algorithms can achieve limited success. One of the most essential problems that negatively affect the results obtained in machine learning studies is the imbalanced data problem. It was observed that no study was conducted on imbalanced data in the mentioned studies. For this reason, in our study, ML analyses were applied after the imbalance problem between classes was resolved using the SMOTE and ADASYN oversampling methods. The main contributions of this study are as follows:

- Unlike previous studies that overlooked the imbalance problem in childhood anemia data sets, this study is the first to address this crucial issue. Our novel approach to tackling the data imbalance problem in childhood anemia sets us apart and makes our research particularly intriguing.
- In our study, we rigorously evaluate the effects of both SMOTE and ADASYN methods on childhood anemia datasets. This comprehensive evaluation is a crucial aspect of our research.
- Our study surpasses previous research by achieving more successful results, particularly addressing the class imbalance issue. This advancement is a significant contribution to the field of childhood anemia prediction.

This study is structured as follows: Section 2 describes the material and methods, including details about the dataset, the method used, explanations about ML algorithms, sampling techniques, and performance metrics. Section 3 presents the results and discussion, and Section 4 presents the conclusion and suggestions.

2. Materials and Methods

2.1. Dataset Description

The 2018 Nigeria Demographic and Health Survey (NDHS) data were used in this study. The dataset is publicly available on the Kaggle platform under "Factors Affecting Children Anemia Level". The dataset, collected from 36 states in Nigeria, includes data on children aged 0-59 months, their mothers aged 15-49 years, and different socioeconomic factors. The target variable, the anemia level of children aged 0-59 months, consists of four classes ('Not anemic', 'Mild', 'Moderate', and 'Severe').

Initially, a large amount of missing data was detected in the dataset, consisting of 33924 rows and 17 columns. After selecting and applying appropriate data preprocessing methods (completion with average, elimination of columns with large amounts of missing data and low impact on the target variable) for the missing data, 10182 rows and 14 columns remained in the dataset. All data preprocessing tasks were conducted using the Python programming language.

2.2. Method

ML methods were applied in the Google Colab environment using the Python programming language to estimate the level of childhood anemia. The dataset was divided into train (80%) and test (20%). Categorical data were converted to numerical data using one hot encoding. Feature Scaling was applied with StandardScaler. Since there was a class imbalance in the target variable, the classes were balanced using the SMOTE and ADASYN methods. ML methods were applied to the original imbalanced dataset, and the balanced datasets were obtained using SMOTE and ADASYN methods. The results were evaluated in terms of various metrics. The general structure of the Children Anemia Prediction System developed within the scope of this study is shown in Figure 1.

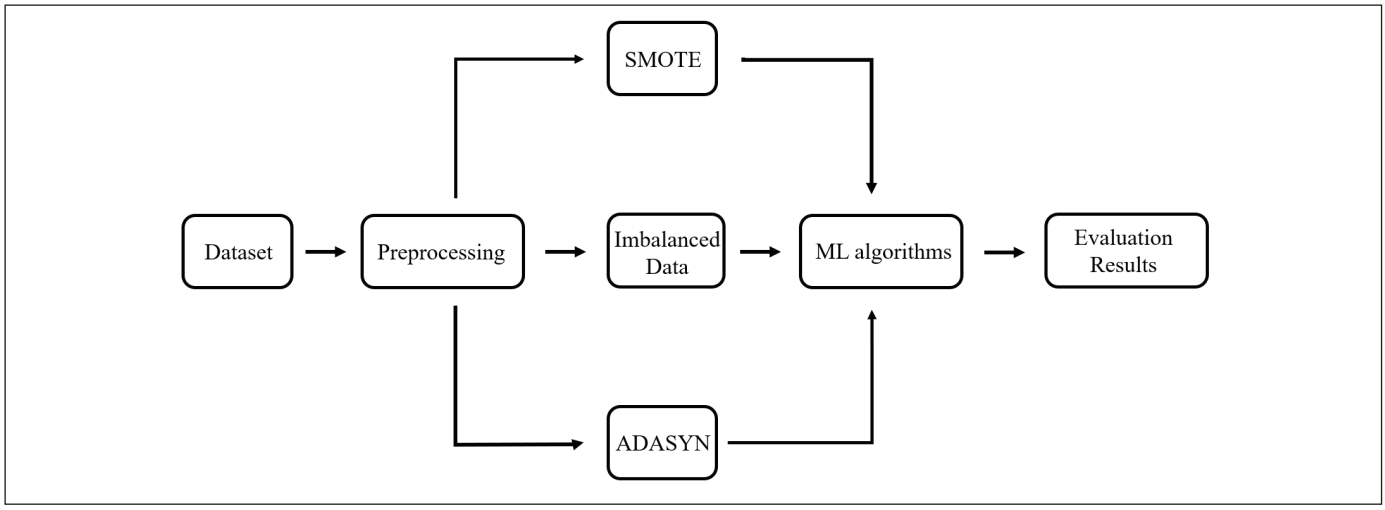

Figure 1. An overview of the children anemia prediction system.

2.3. Machine Learning Algorithms

In this study, we evaluated the performance of different ML algorithms on balanced and imbalanced data for predicting childhood anemia. The ML algorithms compared in this study are summarized below.

2.3.1. Logistic Regression (LR)

LR is a probability-based statistical method widely used in regression and classification problems. It estimates the probabilities of the relationship between a given outcome and inputs using the logistic function (Pedregosa et al. 2011). It is instrumental in diagnosing diseases based on symptoms in the medical field (Harrington 2012).

2.3.2. K-Nearest Neighbors (KNN)

KNN is an ML algorithm focusing on instance-based learning rather than building a general model (Aha et al. 1991). The algorithm calculates the k nearest neighbors of the training data in N-dimensional space according to pre-defined similarity criteria. The selection of the optimum number of neighbors and the data quality significantly impacts the algorithm's accuracy. The KNN method, which can be applied to regression and classification tasks, is robust against noisy training data (Sarker 2021).

2.3.3. Multi-layer Perceptron (MLP)

MLP is a supervised learning method with no loops, and the flow of information is one-way. The method aims to match the inputs with the desired output with a non-linear function. It is frequently used for regression and classification tasks (Naskath et al. 2023).

2.3.4. Gaussian Naïve Bayes (GNB)

It is one of the Naïve Bayes techniques that perform classification based on Bayes theorem. Based on the assumption that the data have a Gaussian distribution, the GNB algorithm performs very successfully on continuous numerical data (Pajila et al. 2023).

2.3.5. Random Forest (RF)

RF is a widely used ensemble classification algorithm. It has many applications in data science and machine learning. The prediction accuracy is obtained using the average or majority vote of the results obtained by multiple decision trees in parallel. The RF algorithm is suitable for regression and classification tasks and is robust to the overfitting problem (Sarker 2021, Sarker et al. 2019).

2.3.6. Gradient Boost (GB)

GB combines a series of decision trees sequentially to reduce the loss and strengthen the model performance (Davagdorj et al. 2020). It is also effective when training data are scarce by iteratively improving the learning abilities of weak learners (Mason et al. 1999). The GB algorithm is an ensemble learning algorithm that successfully solves regression and classification problems.

2.4. Sampling Techniques

SMOTE (Chawla et al. 2002) is a successful technique for minority classes. It takes samples from each minority class and creates new synthetic samples based on the neighbors of these samples. In this way, it is used to eliminate the imbalance problem between classes and increase the performance of the model. The initial step involves the selection of ran-

dom data from minority samples. Subsequently, the KNN is randomly chosen. In the SMOTE method, a new synthetic minority class, , is generated, which lies on the line segment between and , as illustrated by equation (1).

$$x_{new} = (x_i - x_k) \times \delta \quad (1)$$

where

x_i: minority class random data;

K: hyper-parameter of KNN;

x_k: KNN of x_i;

δ: random value between 0 and 1.

Although ADASYN (He et al., n.d.) is different from SMOTE in terms of the number of synthetic samples produced, it is an approach for minority classes like SMOTE. In the ADASYN method, the degree of class imbalance is calculated using a weighted distribution. Thus, it is determined how much synthetic data will be produced (Halim et al. 2023). Firstly, the algorithm calculates the requisite size of data samples for the minority class, as defined by equation (2). Subsequently, the number of synthetic data samples to be generated for each belonging to the minority class is determined by equations (3-5).

$$G = (S_{maj} - S_{\min}) \times \beta \quad (2)$$

$$\gamma_i = N_k / k \quad (3)$$

$$\hat{\gamma}_i = \gamma_i / \sum_{i=1}^{S_{\min}} \gamma_i \quad (4)$$

$$g_i = \hat{\gamma}_i \times G \quad (5)$$

where

S_{maj}: majority class sample size;

S_{min}: minority class sample size;

β: a value between 0 and 1 to ensure balance after the generation of synthetic data;

γ_i: a ratio of majority class sample size;

$\hat{\gamma}_i$: normalized γ_i;

g_i: the dimension of the synthetic data to be generated.

The dataset used in this study also had an imbalanced distribution between classes. The class distributions of the target variable "Anemia level.1", obtained after data preprocessing and consisting of 4 classes of object type, are presented in Figure 2. The class imbalance is clearly seen in Figure 2 (a). This study applied SMOTE and ADASYN methods to the class imbalance problem. Figures 2(b) and (c) present the class distributions obtained after applying the SMOTE and ADASYN methods.

Table 1 shows the number of samples in imbalanced and balanced classes of the target variable. Accordingly, while there were 322 samples in the class labeled 'Severe' in the imbalanced data set, which is much less than the others, this number increased to 2576 after the SMOTE method was applied, and 3111 after the ADASYN method was applied. Thus, balance was achieved between the classes of the target variable.

2.5. Performance Metrics

The performance of the methods presented in this study were evaluated using accuracy, precision, recall, and F1 score metrics which are defined in Equations (1) - (4), respectively. These metrics, which are widely used in evaluating the performance of machine learning models, take into account the correct and incorrect predictions made by the models. Here, the term TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative.

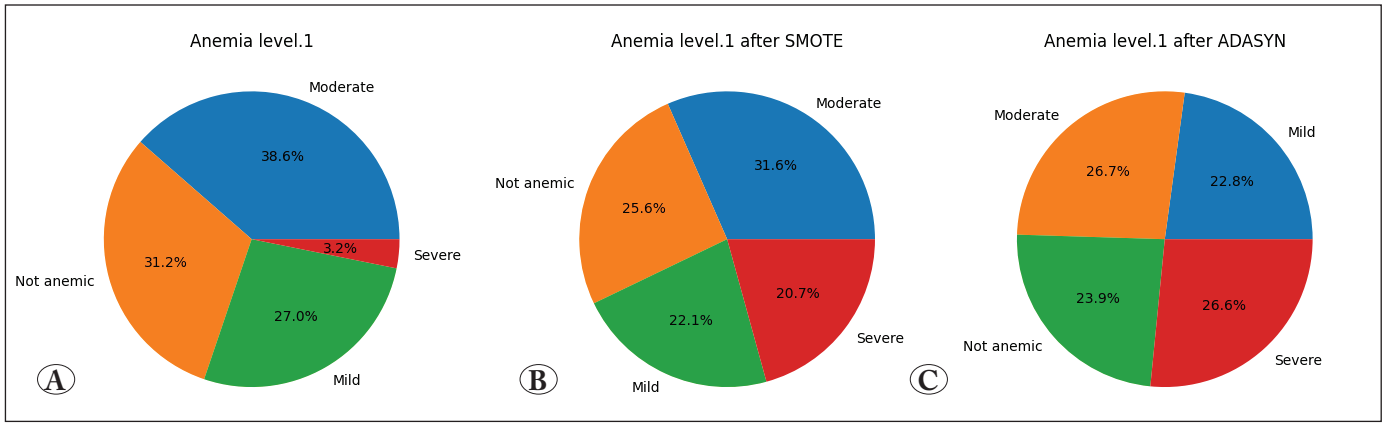

Figure 2. Class distributions of target variable. **A)** Imbalanced, **B)** after SMOTE, **C)** after ADASYN.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

3. Results and Discussion

In this study, machine learning models were applied to both the imbalanced and balanced datasets generated using SMOTE and ADASYN oversampling techniques. The results obtained by applying the LR, KNN, MLP, GNB, RF, and GB machine learning models, which are widely used in the literature, are evaluated using the accuracy, precision, recall, and F1 score metrics. The results obtained by applying the machine learning models to the initial dataset where the target variable has imbalanced classes are shown in Figure 3.

When Figure 3 is examined, it can be seen that the ML algorithm with the highest accuracy rate on imbalanced data is the RF (73.74) algorithm. This is followed by the KNN (71.95), GB (70.01), LR (69.93), MLP (69.17), and GNB (63.05) algorithms, respectively. Performance results of the SMOTE oversampling method are presented in Figure 4.

According to Figure 4, as a result of the analyses performed with the balanced data set due to the application of the SMOTE oversampling technique, it was determined that the most successful algorithm in terms of accuracy was LR (100). This was followed by the MLP, GB, KNN, RF, and GNB algorithms with accuracy rates of 99.68, 98.51, 97.87, 97.15, and 82.48, respectively. Figure 5 shows accuracy, precision, recall, and F1-Score performance results with the ADASYN oversampling method.

When Figure 5 is examined, it can be seen that the most successful algorithm in terms of accuracy is MLP (100) as a result of the analysis performed with the balanced data set due to the application of the ADASYN oversampling technique. This was followed by GB (99.85), RF (99.71), LR (99.02), GNB (75.85), and KNN (73.15) algorithms, respectively.

Table 1. Number of samples in imbalanced and balanced classes of the target variable.

Sampling Method	Not anemic	Mild	Moderate	Severe
Imbalanced	3179	2754	3927	322
SMOTE	3179	2754	3927	2576
ADASYN	2803	2676	3123	3111

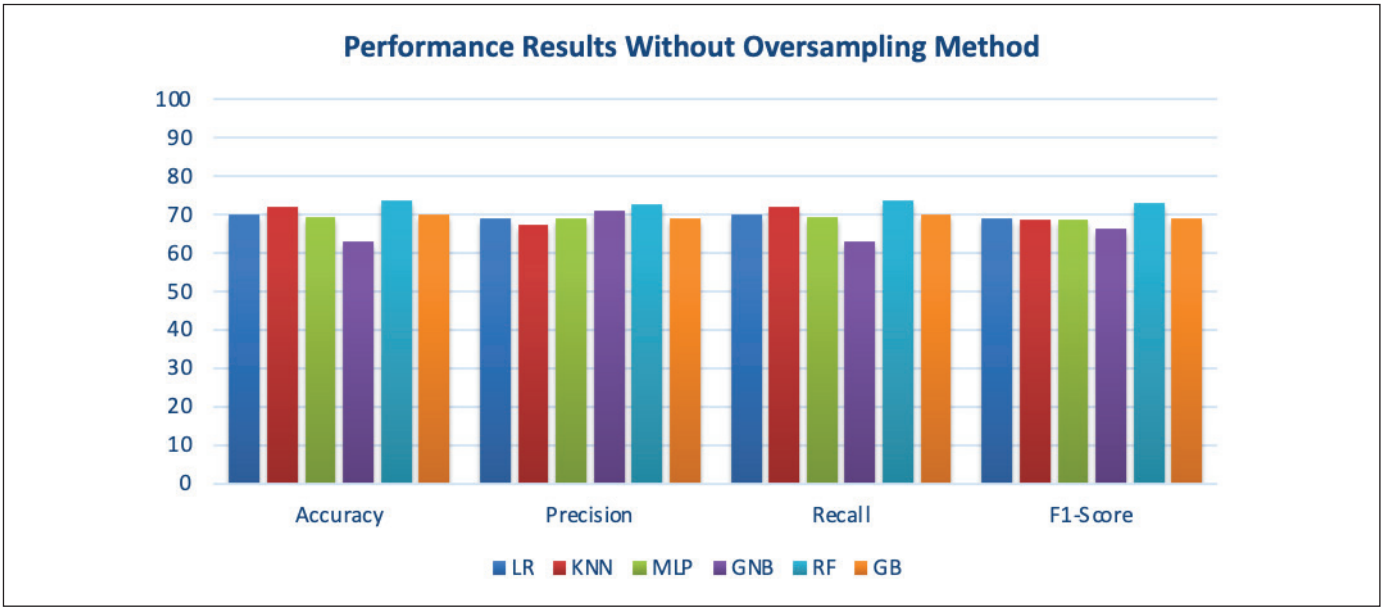

Figure 3. Performance results without oversampling.

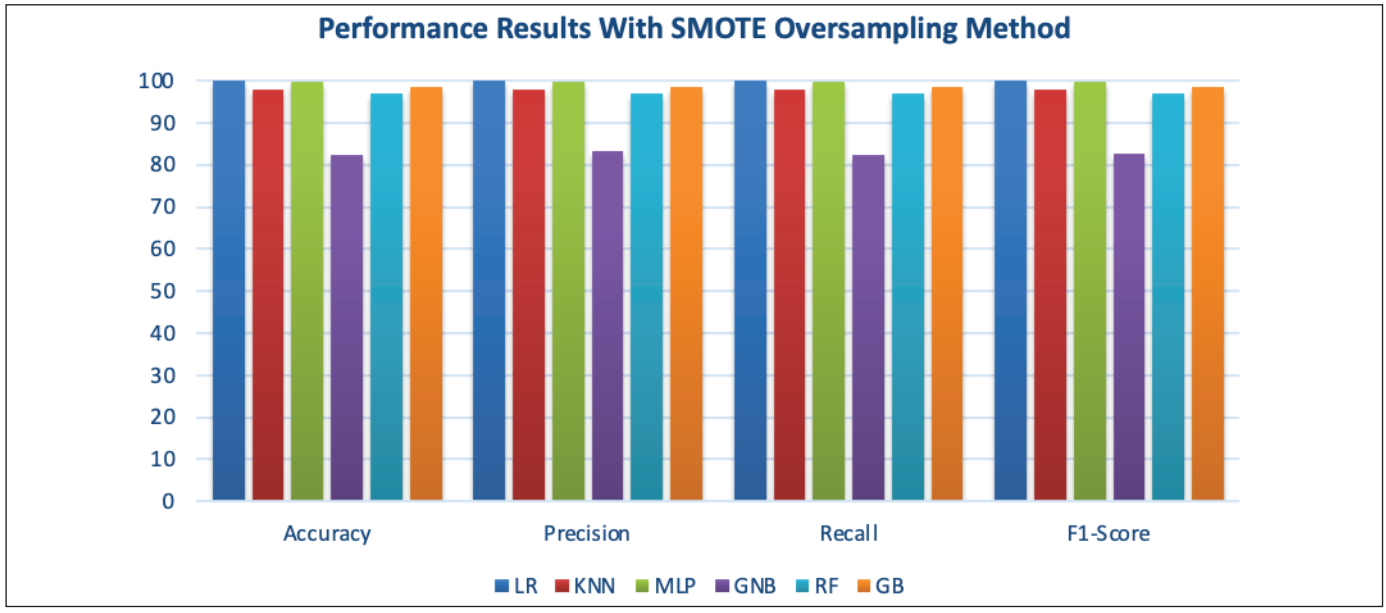

Figure 4. Performance results with SMOTE oversampling method.

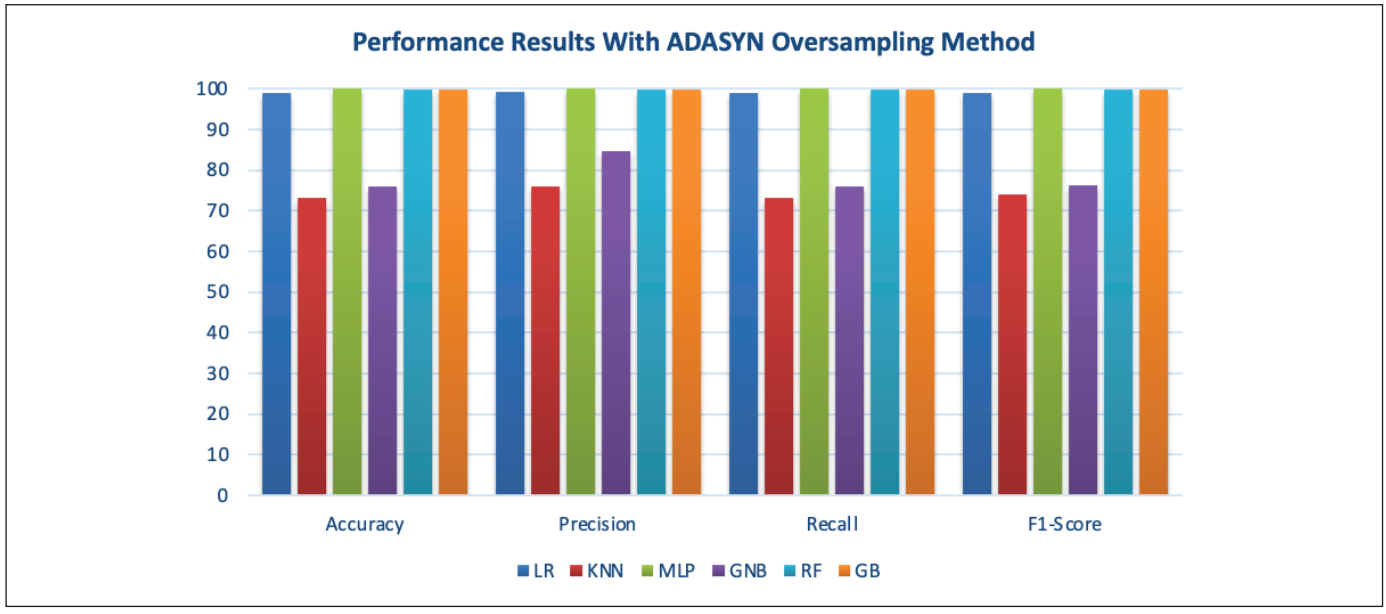

Figure 5. Performance results with ADASYN oversampling method.

The results of all machine learning methods used in the study on balanced and imbalanced data are shown in Table 2. The highest performance results obtained based on the model according to the evaluation metrics are marked in bold.

When Table 2 is examined, it can be seen that for all ML models, SMOTE and ADASYN methods demonstrate significantly higher performance than the results obtained with imbalanced data. According to Table 2, in the SMOTE oversampling method, the LR algorithm obtained the highest accuracy rate of 100%. It is seen that the success of the LR algorithm in terms of other metrics is also 100%. According to the results shown in Table 2, in the case of applying the ADASYN oversampling method, the MLP algorithm achieved the highest accuracy rate of 100%. The success of the MLP algorithm in terms of other metrics was also 100%. The comparison of balanced and imbalanced data in terms of accuracy only for all ML models is shown in Figure 6.

Table 2. Results of sampling methods by model.

ML Model	Sampling Method	Accuracy	Precision	Recall	F1-Score
LR	**Imbalanced**	69.93	69.12	69.93	69.11
	SMOTE	**100**	**100**	**100**	**100**
	ADASYN	99.02	99.16	99.02	99.05
KNN	**Imbalanced**	71.95	67.43	71.95	68.61
	SMOTE	**97.87**	**97.89**	**97.87**	**97.87**
	ADASYN	73.15	76.06	73.15	73.93
MLP	**Imbalanced**	69.17	69.12	69.17	68.72
	SMOTE	99.68	99.68	99.68	99.68
	ADASYN	**100**	**100**	**100**	**100**
GNB	**Imbalanced**	63.05	71.08	63.05	66.28
	SMOTE	**82.48**	83.34	**82.48**	**82.68**
	ADASYN	75.85	**84.7**	75.85	76.11
RF	**Imbalanced**	73.74	72.71	73.74	72.93
	SMOTE	97.15	97.14	97.15	97.14
	ADASYN	**99.71**	**99.71**	**99.71**	**99.71**
GB	**Imbalanced**	70.01	69.12	70.01	69.15
	SMOTE	98.51	98.53	98.51	98.51
	ADASYN	**99.85**	**99.86**	**99.85**	**99.85**

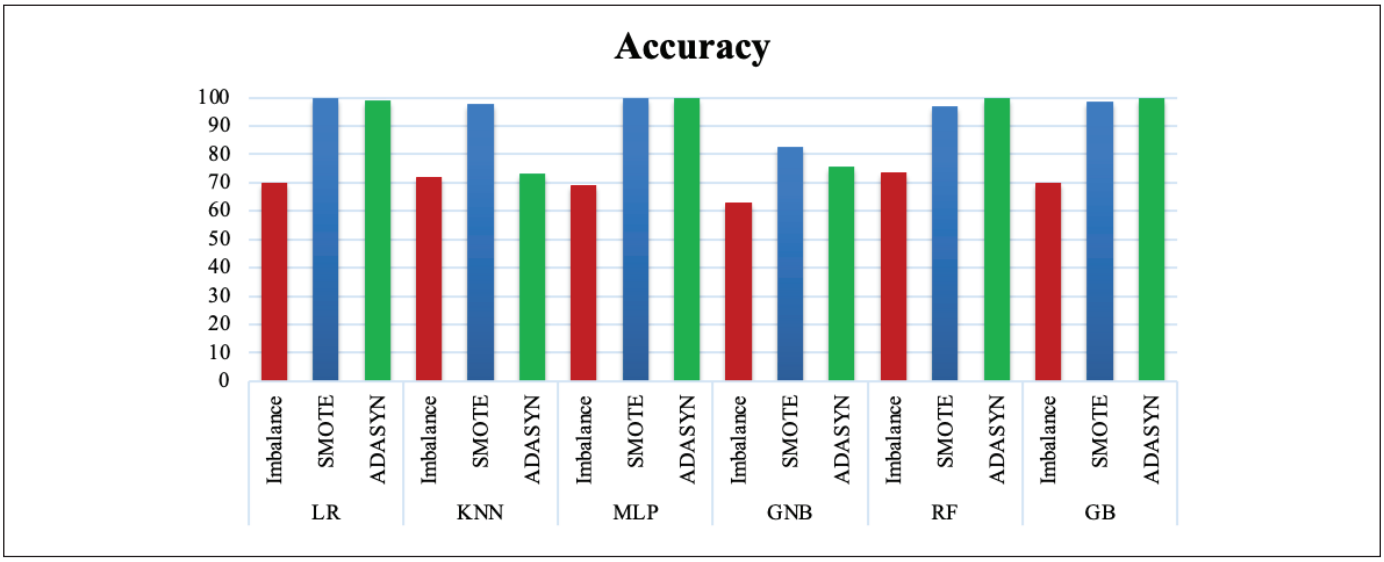

Figure 6. Comparison of accuracy results.

The averages of the results of all ML algorithms applied in this study in terms of accuracy, precision, recall, and F1-Score metrics are given in Table 3 according to sampling methods. When Table 3 is examined, it is seen that SMOTE and ADASYN oversampling methods increase the prediction success in predicting childhood anemia. The most successful result in terms of all metrics used was obtained with the SMOTE method.

The comparison of the results obtained in this study with other studies predicting childhood anemia using machine

Table 3. Average results of sampling methods.

Sampling Method	Accuracy (Mean)	Precision (Mean)	Recall (Mean)	F1-Score (Mean)
Imbalanced	69.64	69.76	69.64	69.13
SMOTE	**95.95**	**96.10**	**95.95**	**95.98**
ADASYN	91.26	93.25	91.26	91.44

learning methods is presented in Table 4. When Table 4 is examined, it is seen that SMOTE and ADASYN oversampling methods significantly increase the prediction success in all algorithms in the comparison in terms of accuracy metric.

Table 5 illustrates the confusion matrix of the RF algorithm that gives the best result for the imbalanced target variable with class labels 'Not anemic', 'Mild', 'Moderate', and 'Severe'. The diagonal elements of the matrix represent the number of correctly classified examples for each label. According to Table 5, 'Not anemic' was correctly classified 544 times, 'Mild' 316 times, 'Moderate' 603 times, and 'Severe' 38 times. The non-diagonal elements in the matrix show the number of incorrectly classified examples. For example, 138 examples that should have been classified as 'Not anemic' were classified as 'Mild'.

Table 6 shows the confusion matrix of the LR algorithm, which gives the most successful result for the SMOTE oversampling method of the target variable with class labels 'Not anemic', 'Mild', 'Moderate', and 'Severe', and Table 7 shows the confusion matrix of the MLP algorithm, which gives the most successful result for the ADASYN oversampling method. According to Table 6 and Table 7, all samples of the target variable are classified correctly.

4. Conclusion and Suggestions

This research has comprehensively investigated the effectiveness of machine learning algorithms in diagnosing of childhood anemia. The study evaluated SMOTE and ADASYN oversampling techniques for the imbalanced dataset problem. The findings show that these techniques significantly increased the performance of ML models and achieved higher accuracy rates in anemia diagnosis. This approach may play a critical role in the early diagnosis and treatment of childhood anemia.

In future studies, the performance of different machine learning algorithms can be compared using more extensive and diverse datasets. In addition, the difficulties encountered in integrating into clinical applications and strategies

Table 4. Comparison of the proposed methods and previous studies regarding accuracy.

References	Method	Accuracy (%)
(Anand et al. 2019)	LDA	62.10
	CART	61.40
	K-NN	60.36
	RF	67.18
	LR	61.67
(Khan et al. 2021)	LDA	63.84
	CART	62.14
	k-NN	67.73
	SVM (linear)	66.73
	RF	96.16
	LR	63.99
(Meitei et al. 2022)	Elastic	64.17
	Ridge	64.19
	LASSO	64.29
(Marcos Valdez et al. 2023)	NB	70.00
(Tesfaye et al. 2024)	LR	64.00
	DT	68.00
	RF	68.00
	KNN	66.00
This Study	SMOTE-LR	**100**
	ADASYN-LR	**99.02**
	SMOTE-KNN	**97.87**
	ADASYN-KNN	**73.15**
	SMOTE-MLP	**99.68**
	ADASYN-MLP	**100**
	SMOTE-GNB	**82.48**
	ADASYN-GNB	**75.85**
	SMOTE-RF	**97.15**
	ADASYN-RF	**99.71**
	SMOTE-GB	**98.51**
	ADASYN-GB	**99.85**

Table 5. Confusion matrix for RF with imbalanced data.

		Predicted			
		Not anemic	Mild	Moderate	Severe
Actual	Not anemic	544	138	0	0
	Mild	35	316	189	0
	Moderate	0	141	603	0
	Severe	0	0	32	38

Table 6. Confusion matrix for SMOTE-LR.

		Predicted			
		Not anemic	Mild	Moderate	Severe
Actual	Not anemic	636	0	0	0
	Mild	0	551	0	0
	Moderate	0	0	785	0
	Severe	0	0	0	515

Table 7. Confusion matrix for ADASYN-MLP.

		Predicted			
		Not anemic	Mild	Moderate	Severe
Actual	Not anemic	560	0	0	0
	Mild	0	535	0	0
	Moderate	0	0	624	0
	Severe	0	0	0	622

that can be developed to overcome these difficulties can be emphasized. In this context, multidisciplinary approaches and collaboration with healthcare professionals can ensure the effective use of machine learning models in clinical settings. In conclusion, this study demonstrates that machine learning and oversampling techniques have significant potential in diagnosing childhood anemia and provide a strong basis for future research.

Abbreviations

The following abbreviations were used in this paper.

ADASYN: Adaptive Synthetic Sampling Approach

CART: Classification and Regression Trees

DT: Decision Tree

Elastic: Elastic-net Regression

GB: Gradient Boost

GNB: Gaussian Naïve Bayes

k-NN: k-Nearest Neighbors

LASSO: Least Absolute Shrinkage and Selection Operator Regression

LDA: Linear Discriminant Analysis

LR: Logistic Regression

ML: Machine Learning

MLP: Multi-layer Perceptron

RF: Random Forest

Ridge: Ridge Regression

SMOTE: Synthetic Minority Oversampling Technique

SVM: Support Vector Machines

WHO: World Health Organization

Acknowledgment: This article is an original work; all results have not been published. Also, this study did not receive any funding or research grants during the study, research or assembly of the manuscript.

5. References

Aha, DW., Kibler, D., Albert, MK., Quinian, JR. 1991. Instance-based learning algorithms. Machine Learning 1991 6:1, 6(1), 37–66. Doi: 10.1007/BF00153759

Ajakaye, OG., Ibukunoluwa, MR. 2020. Prevalence and risk of malaria, anemia and malnutrition among children in IDPs camp in Edo State, Nigeria. Parasite Epidemiology and Control, 8: e00127. Doi: 10.1016/j.parepi.2019.e00127

Anand, P., Gupta, R., Sharma, A. 2019. Prediction of Anaemia among children using Machine Learning Algorithms. 11(2), 469–480.

Aynalem, M., Shiferaw, E., Adane, T., Gelaw, Y., Enawgaw, B. 2022. Anemia in African malnourished pre-school children: A systematic review and meta-analysis. SAGE open medicine, 10. Doi: 10.1177/20503121221088433

Bitew, FH., Sparks, CS., Nyarko, SH. 2022. Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. Public Health Nutrition, 25(2), 269-280. Doi: 10.1017/S1368980021004262

Chawla, NV., Bowyer, KW., Hall, LO., Kegelmeyer, WP. 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. Doi: 10.1613/JAIR.953

Davagdorj, K., Lee, JS., Pham, VH., Ryu, KH. 2020. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. Applied Sciences, 10(9), 3307. Doi: 10.3390/app10093307

Getawa, S., Getaneh, Z., Melku, M. 2020. Hematological abnormalities and associated factors among undernourished under-five children attending University of Gondar Specialized Referral Hospital, Northwest Ethiopia. Journal of Blood Medicine, 465-478. https://doi.org/10.2147/JBM.S284572

Halim, AM., Dwifebri, M., Nhita, F. 2023. Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets. Building of Informatics, Technology and Science (BITS), 5(1). Doi: 10.47065/BITS.V5I1.3647

Harrington, P. 2012. Machine Learning in Action (1st Edition). Manning Publications.

He, H., Bai, Y., Garcia, EA., Li S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks (IEEE, world congress on computational intelligence),1322-1328.

Kebede Kassaw, A., Yimer, A., Abey, W., Molla, TL., Zemariam, AB. 2023. The application of machine learning approaches to determine the predictors of anemia among under five children in Ethiopia. Scientific Reports, 13(1). Doi: 10.1038/s41598-023-50128-x

Khan, JR., Chowdhury, S., Islam, H., Raheem, E. 2021. Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh. Journal of Data Science, 17(1), 195–218. Doi: 10.6339/jds.201901_17(1).0009

Marcos Valdez, AJ., Navarro Ortiz, EG., Quinteros Peralta, RE., Tirado Julca, JJ., Valentin Ricaldi, DF., Calderon-Vilca, HD. 2023. Machine Learning for the Prediction of Anemia in Children Under 5 Years of Age by Analyzing their Nutritional Status Using Data Mining. Computacion y Sistemas, 27(3), 749–768. Doi: 10.13053/CyS-27-3-4315

Mason, L., Baxter, J., Bartlett, P., Frean, M. 1999. Boosting algorithms as gradient descent. Advances in Neural Information Processing Systems, 12.

Meitei, AJ., Saini, A., Mohapatra, BB., Singh, KJ. 2022. Predicting child anaemia in the North-Eastern states of India: a machine learning approach. International Journal of System Assurance Engineering and Management, 13(6), 2949-2962. Doi: 10.1007/s13198-022-01765-4

Naskath, J., Sivakamasundari, G., Begum, AAS. 2023. A Study on Different Deep Learning Algorithms Used in Deep Neural Nets: MLP SOM and DBN. Wireless Personal Communications, 128(4), 2913–2936.Doi: 10.1007/s11277-022-10079-4

Pajila, PJB., Sheena, BG., Gayathri, A., Aswini, J., Nalini, M., Siva Subramanian, R. 2023. A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications. Proceedings of the 4th International Conference on Smart Electronics and Communication, ICOSEC 2023, 1228–1234. Doi: 10.1109/ICOSEC58147.2023.10276274

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., … Fré. 2011. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research, 12, 2825–2830. Doi: 10.5555/1953048.2078195

Rahmani, AM., Yousefpoor, E., Yousefpoor, MS., Mehmood, Z., Haider, A., Hosseinzadeh, M., Ali Naqvi, R. 2021. Machine Learning (ML) in Medicine: Review, Applications, and Challenges. Mathematics, 9(22), 2970. Doi: 10.3390/math9222970

Rajula, HSR., Verlato, G., Manchia, M., Antonucci, N., Fanos, V. 2020. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 56(9), 455. Doi: 10.3390/medicina56090455

Sarker, IH. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 1–21. Doi: 10.1007/s42979-021-00592-x

Sarker, IH., Kayes, ASM., Watters, P. 2019. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, 6(1). Doi: 10.1186/S40537-019-0219-Y

Tesfaye, SH., Seboka, BT., Sisay, D. 2024. Application of machine learning methods for predicting childhood anaemia: Analysis of Ethiopian Demographic Health Survey of 2016. Plos one, 19(4), e0300172. Doi: 10.1371/journal.pone.0300172

WHO 2023. Anaemia Factsheet. https://www.who.int/news-room/fact-sheets/detail/anaemia (accessed on 02 September 2024).

Zemariam, AB., Yimer, A., Abebe, GK., Wondie, WT., Abate, BB., Alamaw, AW., … Ngusie, HS. 2024. Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia. Scientific Reports 2024 14:1, 14(1), 1–17. Doi: 10.1038/s41598-024-60027-4