Machine Learning Algorithms to Predict Anemia in Children Under the Age of Five Years in Afghanistan: A Case of Kunduz Province

Abdullah Zahirzada [1], Naimatullah Zaheer [2], Mohammad Akbar Shahpoor [3]

[1] Department of Information Systems, Kunduz University, Afghanistan

[2] Department of Internal Medicine, Kunduz University, Afghanistan

[3] Department of Software Engineering, Kunduz University, Afghanistan

E-mail: [1] ab.zahirzada95@gmail.com, [2] naimatullahzaheer@yahoo.com

[3] akbarshshpoor@gmail.com

Abstract

Anemia has become an increasingly common problem, especially in developing and underdeveloped countries. Therefore, an ability to predict the anemia is beneficial and a good preventive measure. It is also a good indicator of future health risks of that infant. This study is concerned about the implementation of predictive anemia models for Afghanistan based on the data obtained from the hospitals of Kunduz province. The main objective of the study is to identify the most suitable machine learning techniques (i.e., classifiers) among the five popular ones. These are K-Nearest Neighbor (K-NN), Naïve Bayes, Multi-Layer Perceptron (MLP), Random Forest, and Support Vector Machine (SVM). Prior to implementing the predictive models, data preprocessing is carried out. This is done by means of data cleansing and feature selection. The well-known Correlation based Feature Selection algorithm (CFS) is employed to select the top fifteen attributes. The classifier in this study comprises two categories, Anemic and Non-Anemic. The preparation of the dataset is carefully done to ensure well-balanced samples in each category. The study reveals that Random Forest is the best classifier with an accuracy of 86.4% and with the Area Under the Curve (AUC) of 88.2%, respectively. The study has a direct benefit to the health and prevention policy making in Afghanistan.

Keywords— Afghanistan, Classifiers, Correlation-based Feature Selection (CFS), Anemia, Machine Learning.

I. Introduction

A disorder known as anemia is characterized by low levels of hemoglobin and a lack of functional red blood cells in the blood [1]. Anemia affects approximately 1.62 billion people worldwide, making it the second most common cause of disability in the world [2,3] and one of the most important global public health issues. Anemia affects 30.2% of women between the ages of 15 and 49, or 496 million people [4], as well as 43% of children, or 273 million people. Anemia causes disabilities for 68 million persons worldwide [5]. Anemia has detrimental health effects on the economics of human capital, costing billions of dollars annually [3].

Due to dietary deficiencies during pregnancy and infancy, anemia is more common, especially in low-income countries, and it has long-term negative effects on the brain's health [4–7]. Iron deficiency anemia is, in fact, the most common cause of nutritional anemia. Iron deficiency is the primary cause of around half of the worldwide anemia. That indicates that iron deficiency accounts for 50% of anemia cases worldwide [8, 9]. The two continents that are both most severely impacted by anemia and have the highest anemia prevalence are Asia and Africa [10]. The prevalence of anemia among children in Afghanistan was reported to be 46.4% in 2016. This is an increase from the previous figure for 2015, which was 45.8%. Data on Afghanistan's prevalence of anemia among children is updated yearly [11]. Therefore, if a predictive anemia model existed, it would be extremely beneficial to pinpoint the causes of anemia in Afghanistan. As a result, this study suggests a predictive model for anemia for a child who meets certain criteria in Afghanistan. Additionally, it looks for the best classifier for this task. K-Nearest Neighbor, Naive Bayes, Multi-Layer Perceptron, Random Forest, and Support Vector Machine are five well-known classifiers that are chosen for the study. The dataset used in this study was gathered using a self-structured and pretested questionnaire from Kunduz province hospitals, which comprises records of children whose age group is under five years.

II. Classification Selected

The main goal of classification, a supervised learning technique, is to use an applied predictive model based on previously known data to classify newly discovered or unknown data. During the use of the predictive model development in classification, a given dataset is split into training and test sets. 'Classifier' is a common term for a classification model. Finally, the effectiveness of the built classifier is verified using several metrics. The classifiers used for this research are briefly introduced in Subsections A through E, and previous related work is briefly described in Section III.

A. K-Nearest Neighbor (K-NN)

The K-Nearest-Neighbors (K-NN) approach does not make any assumptions about the elementary dataset because it is a non-parametric classification algorithm. It is renowned for being both straightforward and efficient. It is an algorithm for supervised learning. In order to predict the class of the unlabeled data, a labeled training dataset with data points divided into several classes is provided. Different criteria are used in classification to identify the class to which the unlabeled data belongs. Typically, KNN is employed as a classifier. It is used to categorize data based on nearby or close-by training examples in a certain area. This approach is employed due to its speedy computation and ease of operation. It computes its closest neighbors for continuous data using the Euclidean distance [12].

B. Multi-Layer Perceptron (MLP)

One or more hidden layers may be present between the input and output layers in MLPs, a particular type of feedforward neural network. A neuron in one layer of an MLP is connected to every neuron in the layer above it, which is referred to as being fully connected. There is a separate weight value for each connection. The backpropagation algorithm is used to train MLP. After sending a batch of data to the network, the backpropagation technique employs gradient descent to determine the error contribution of each neuron. The weights are then adjusted so that the network produces the desired result for a particular input [13].

C. Naïve Bayes

The Nave-Bayes family of probabilistic models uses the Bayes theorem to forecast the class label for a specific issue instance under the premise of conditional independence between the features. Statistical classifiers include Bayesian classifiers. They can forecast probabilities of class membership, such as the likelihood that a given sample belongs to a specific class. On Bayes' theorem, the Bayesian classifier is founded. The assumption made by naive Bayesian classifiers is that the impact of one attribute's value on a particular class depends only on that attribute's value. The term "class conditional independence" refers to this presumption. It is called "naive" in this sense because it is designed to make the computation involved simpler [14].

D. Random Forest

Similar to decision trees, Random Forest is an ensemble classifier that can be used to address classification and regression issues. With the help of the concept of creating multiple random trees, bootstrapping the training dataset, bagging on samples, voting scheme, and randomly selecting features for each decision split, the predictability is increased, and efficiency is increased. In most cases, it outperforms decision trees in terms of results. The random subspace approach, for instance, involves choosing a random subset of features. It was established in 2001 and is utilized in numerous applications, such as image processing and medical research. [15]

E. Support Vector Machine (SVM)

Support vector machines (SVM) are a frequently used kernel-based supervised machine learning approach for categorization issues. By maximizing the margin between the classes, the SVM algorithm builds a hyperplane that properly divides the training observations according to their class labels. Depending on which side of the hyperplane a test observation is located, SVM gives it to a class [16].

III. Previous Related Work

According to a study in [13], a machine learning algorithm for estimating hemoglobin level and categorizing anemia based on blood test factors has been proposed. Ahmed Shalaby lab's Mindray BC-5300 was used to acquire the dataset for training and testing. Before training the models, preprocessing is done in their suggested method, Hemoglobin Estimation and Anemia Classification (HEAC), to decrease and standardize the data. The output from this model is contrasted with output from other machine learning models. Hemoglobin estimation and anemia classification are produced with high accuracy by the suggested model. likewise, a study in [16], machine learning algorithms can be used to predict children under the age of five's anemia status using common risk indicators. Data were taken from the 2011 Bangladesh Demographic and Health Survey (BDHS), a nationally representative cross-sectional survey. In this study, a sample of 2013 kids were chosen whose data were available for all of the factors. To predict the presence of pediatric anemia, they employed a variety of machine learning (ML) algorithms, including logistic regression, random forest, support vector machines, linear discriminant analysis, and classification and regression trees (CART). The algorithms' accuracy, sensitivity, specificity, and area under the curve were all systematically assessed (AUC). They discovered that the Random Forest method had the best classification accuracy, scoring 68.53%, with 70.73% sensitivity, 66.41% specificity, and an AUC of 0.6857.

A cross-sectional investigation [17] was carried out in a hospital. Children between the ages of 6 and 36 months participated in this study. In order to predict the anemia status of children under 36 months old in Jammu, they developed a number of Machine Learning algorithms, including linear discriminant analysis (LDA), classification and regression trees (CART), k- nearest neighbors (K-NN), random forest (RF), and logistic regression (LR). Out of all the predictive models created using machine learning approaches, random forest demonstrated the best prediction accuracy of 67.18%. Similarly, research in [18] used the most recent National Family Health Survey (NFHS) data to apply Machine Learning algorithms in order to predict anemias among children in North East India. From a total of 29,312 eligible infants (6-59 months) in North-East India, 21,000 children with demographic characteristics and no missing observations are taken into consideration for this study, of whom 10,460 are anemic. Machine Learning algorithms are evaluated systematically for accuracy, sensitivity, specificity, F1-Score, and Cohen's k-Statistics. It is safe to say that factors like the mother's anemic status, the child's age, social status, mother's age, mother's education, and religion are significant in determining whether the child is anemic, having achieved the receiver operating characteristic value of over 70% in training and accuracy of above 64% during testing. Some health-related researches (low birth weight and Time Series Forecasting of Registered, Recovered, and Death Cases of COVID-19 for the Next Sixty Days in Afghanistan) which investigated medical and non-medical data were carried out in [19 and 20].

As mentioned in the previous works, many data mining and machine learning techniques have been used to infer meaningful knowledge for domain experts from healthcare data, particularly anemia, in developed and developing countries. Statistical methods were used in earlier studies on anemia in Afghanistan. They also restricted their work to information only pertaining to medicine. A predictive anemia model based on both non-medical variables (such as mother education level) and related medical data has not yet been attempted.

IV. Methodology

Machine learning techniques were applied to the prediction model that was created. Its job is to determine if a child's record, or information about a specific child, is more likely to fall into the anemic or non-anemic group. The study's goals, as mentioned in Section I, appear to be focused on selecting the best classifier among the most widely used machine learning approaches. In this study, there is a binary classification job where two categories are taken into account. The term "anemic" designates a group in which the sample falls and the possibility that the child is anemic and non-anemic indicates to a child with no anemia.

A. Data Source and Collection

The data used in this research were gathered using a self-structured and pretested questionnaire. To establish the validity of the contents, the questionnaire was initially written in English, then translated to Persian and back to English. To gather information, a face-to-face interview was conducted. Two clinical nurses and one internal medicine specialized participated as data collectors.

B. Data Preprocessing

The first phase of data preprocessing is required in this study, as it is in the majority of data mining processes. Data purification, filling in for missing values, and feature selection are the three primary stages of this process, which are as follows:

1) Data Cleansing

Initially the collected data comprises 402 records (i.e., samples). Data quality was checked for consistency, completeness, and accuracy, it is finding that 52 samples cannot be used and have to be eliminated due to no value is present in some attributes and these samples are too incomplete and beyond repair (i.e., by replacing any missing value techniques). At the end 350 useable samples remain. Table 1 reports number of samples on both categories.

Table 1. Number of samples in both categories (i.e., Anemic & Non-anemic)

No. in Anemic category	No. in Nonanemic category	No. of total Samples
175	175	350

2) Feature Selection

Each sample in the dataset consists of 26 attributes. Careful analysis reveals that some attributes are not related and inappropriate for the implementations. Therefore, an essential step in this initial stage is to identify relevant attributes related to Anemia. This is carried out by means of lengthy and vigorous consultations with doctors and experts in this area. Different doctors and experts had expressed different number of attributes which contribute to Anemia. The study identifies common attributes among these doctors and experts. Finally, it was found that 26 attributes are commonly express among them. Table II reports these 26 attributes as shown below.

Table 2. Selected attributes used in this study

No.	Features	Description
1	Age (year)	Age of the mother
2	Maternal Education	Education of the mother
3	Paternal education	Education of the father

4	Occupation	Occupation of the mother
5	Child age	Age of the child
6	Child stunting	Stunting status of the child
7	Breastfeeding	Child breastfeeding status
8	Maternal anemia	Is mother of the child anemic?
9	Gender	Sex of the child (Female or Male)
10	Maternal weight	Weight of the mother
11	Toilet facilities	Toilet facilities of the household
12	Water source	Household source of drinking water
13	Fever	Does child have fever?
14	Diarrhea	Does child have diarrhea?
15	Place of residence	Place of residence (Rural or Urban areas)
16	No. of children	Number of living children
17	Wealth status	Wealth status of the family
18	Size of child at birth	Size of child at birth (average, large, small, very large, very small)
19	Vitamin A within 6 months	Does child have Vitamin A within 6 months
20	Iron with 7 days	Does child have iron pills with days
21	Any drug	Any drug for parasites within 6 months
22	No. of household	Number of household members
23	No. of under 5 children	Number of children under five years
24	Blood donation	Blood donation in recent 3 months
25	Length of menstruation	Mother length of menstruation (<3 days 3-5 days >5 days)
26	Blood loss during menstruation	Mother blood loss during menstruation/day (low, moderate, high)

Any data mining model must be implemented successfully, and this requires having relevant features. In fact, including irrelevant features might hurt data mining because it makes it harder to learn from samples because they are stuffed with repetitive and useless information. The process of choosing crucial features for the task at hand is known as feature selection. In addition to reducing the data's dimensionality, it also enables better data visualization and interpretation. If an additional study were to be done, all 26 attributes' influences would need to be considered. Since the number of combinations to consider are as many as 26C1 + 26C2 + …... + 26C26. Doctors and domain experts suggested further that considering 15 important attributes ought to be sufficient for the classification so further analysis of the results will not be too overwhelming.

The most commonly known Principal Component Analysis (PCA) is inappropriate for the dataset used. PCA is best applied to numerical attributes which many of attributes used in this work are not. Converting categorical attributes to ranges and assigns numbers to them is not a good solution either as suitable discretization of these categorical data cannot easily be done. Discretization of categorical data is well discussed in [21].

In this study, the common feature selection, the correlation-based feature selection algorithm (CFS) [22], is adopted for its simplicity and its ability to handle both numerical and categorical attributes. CFS computes the correlation between all features and the output class and selects the best feature subset (i.e., the subset with features highly correlated with the class variable and has a low correlation with each other features) using a correlation-based heuristic evaluation function. CFS searches for highly correlated features with the target variable yet have minimal inter-correlation among the features themselves. It can be determined by the equation 1 below.

$$r_{zc} = \frac{k\bar{r}_{zi}}{\sqrt{k+k(k-1)\bar{r}_{ii}}} \qquad (1)$$

rzc: correlation between features and the class variable.

K: number of features.

rzi: average of the correlation between feature-class.

rii: average inter-correlation between features.

CFS is applied to the dataset to determine the top 15 attributes. Table 3 reveals the results of CFS application to datasets.

Table 3. Attributes selected by CFS

No.	Rural area	CFS value
1	Maternal anemia	0.37
2	Maternal weight	0.33
3	Place of residence	0.29
4	No of under 5 children	0.27
5	Having diarrhea	0.26
6	Wealth status	0.23
7	Child stunting status	0.21
8	No of family member	0.19

9	Drinking water source	0.18
10	Toilet facility	0.17
11	Child age	0.15
12	Mother occupation	0.14
13	Size of child at birth	0.13
14	Length of menstruation	0.12
15	Blood lose during menstruation	0.12

C. Evaluation Metrics

In order to evaluate the performance and effectiveness of the predictive model implemented, evaluation metrics are used. In this study, the commonly used four metrics are employed [23]. These are Accuracy, Precision, Recall, and Area Under the Curve (AUC). Their brief descriptions are given below:

1). Accuracy

This metric is the most used in classification as it is the first important indicator of how well the model performs. It is expressed in percentage (i.e., 0% to 100%). It can be determined by equation 2 below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

TP: the number of true positives (samples that are correctly classified in its correct class).

TN: the number of true negatives (samples that are correctly classified that they do not belong to the target class).

FP: the number of false positives (samples that are incorrectly labeled as the target class when they are not, in fact).

FN: the number of false negatives (samples that are incorrectly labeled as not the target class when they are, in fact).

2). Precision

The information retrieval field introduced this metric; however, it has an application in classification and is a useful addition in evaluating performance. It is also expressed in percentage. It can be determined by the equation 3 below:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

3). Recall

Recall is often used in conjunction with Precision in the information retrieval field. Hence, it can add useful information in evaluating performance. It is also expressed in percentages. It can be determined by equation 4 below:

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

4). Area Under the Curve (AUC)

The receiver operating characteristic (ROC) curve's area under the curve (AUC), by screening the range of threshold values for the decision-making, it offers a thorough evaluation of the accuracy of a model. The diagnostic test is more accurate the larger the region. Additionally, it is a percentage. It can be determined by equation 5 below:

$$AUC = \frac{s_p - n_p(n_n+1)/2}{n_p n_n} \tag{5}$$

sp: the sum of all positive examples ranked *np* and *nn*: the number of positive and negative examples.

V. Implementation of the Predictive Model

Predictive model implementation can start when preprocessing is complete. Finding the optimal tool to implement the predictive model is part of the study's goal, as stated in Section I. The widely used machine learning program WEKA [24] was selected for this study since it is the most well-liked program currently available. The five aforementioned data mining tools are used to create predictive models. The learning is most effective when there is an equal number of each type. According to Table I, the proportion of anemic samples to non-anemic samples is equal, and the ratio of the training and test sets is 80:20.

VI. Results and Discussion

Numerous attempts have been carried out in order to implement the most suitable predictive model from each tool by adjusting the parameters of that tool. The results are from the five implemented models. Table 4 reveals the results of the five implemented models.

Table 4. Results from five classifiers for the top 15 selected attributes

Classifier	***Accuracy***	***AUC***	***Precision***	***Recall***
K-NN	69.4%	85.4%	57.1%	72.7%
Naïve Bayes	79.6%	81.8%	67.9%	86.4%
MLP	81.3%	84.9%	73.9%	77.3%
Random Forest	**86.4%**	**88.2%**	**85.0%**	**77.3%**
SVM	72.8%	84.0%	63.6%	63.6%

As shown in table 4, Random Forest yields the best overall results when evaluated by metrics. This is certainly not proof that Random Forest is superior to other algorithms in all cases, as the nature of the data can greatly influence the algorithms used. It is arguable that for the predictive model using the dataset at hand, Random Forest is the most suitable tool, among the most popular tools, to implement predictive model with respect to the four popular classification metrics. The mathematics of K-Nearest Neighbor and Naïve Bayes may not be suitable for this task. At the same time, the Support Vector Machine is known to achieve very good performance for numerical attributes, in which this is not the case in this study.

VII. Conclusion and Future Work

Anemia affects approximately 1.62 billion people worldwide, making it the second most common cause of disability worldwide. Therefore, a better understanding of the critical factors positively associated with anemia and having the most suitable predictive model for anemia in Afghanistan is very beneficial to mitigate the problem mentioned. This study is the first attempt to determine the most suitable tool to implement a predictive anemia model for Afghanistan, where non-medical information is utilized. The study applied CFS in feature selection and identified crucial attributes which merit further analysis. Random Forest is found to be the most suitable tool with Accuracy, AUC, Precision, and Recall of 86.4%, 88.2%, 85.0%, and 77.3%. The results of this study ought to be beneficial to the healthcare authority and the policy-making administration of Afghanistan.

Future studies can be carried out from several aspects. This study adopts 15 attributes and relies on CFS to identify the importance of each selected attribute, more detail analysis can be carried out to ensure the order of influence of each attribute may have in anemia. In order to commence utilizing the models for practical decesion-making, we will deploy the models into a real production, this could be a web based application for anemia detection in children under the age of five.

Statements on ethics

The participants were protected by anonymizing their personal information in this study.

Declarations

Conflict of interest: The author has no conflicts of interest.

References

[1] World Health Organization. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. World Health Organization. (2011).

[2] Balarajan Y, Ramakrishnan U, Ozaltin E, et al. Anemia in low-income and middle-income countries. Lancet. 378, 2123-2135 (2011).

[3] World Health Organization. WHO recommendations on antenatal care for a positive pregnancy experience. World Health Organization. (2016).

[4] Hare D, Ayton S, Bush A, et al. A delicate balance: Iron metabolism and diseases of the brain. Front Aging Neurosci. 5, 34 (2013).

[5] Burke RM, Leon JS, Suchdev PS, et al. Identification, prevention, and treatment of iron deficiency during the first 1000 days. Nutrients. 6, 4093-4114 (2014).

[6] Pala K, Dundar N. Prevalence & risk factors of anemia among women of reproductive age in Bursa, Turkey. Indian J Med Res.128, 282-286 (2008).

[7] Al-Alimi AA, Bashanfer S, Morish MA, et al. Prevalence of iron deficiency anemia among university students in hodeida province, Yemen. Anemia. 2018, 1-7 (2018).

[8] World Health Organization. Iron deficiency anemia. assessment, prevention, and control. A guide for programme managers. 47-62 (2001).

[9] McLean E, Cogswell M, Egli I, et al. Worldwide prevalence of anemia, WHO vitamin and mineral nutrition information system, 1993-2005. Public Health Nutr. 12, 444-454 (2009).

[10] Pala K, Dundar N. Prevalence & risk factors of anemia among women of reproductive age in Bursa, Turkey. Indian J Med Res.128, 282-286 (2008).

[11] https://www.ceicdata.com/en/afghanistan/health-statistics/af-prevalence-of-anemia-among-children--of-children-under-5

[12] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest

Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[13] A Machine Learning Model for Hemoglobin Estimation and Anemia Classification
Journal of Computer Science IJCSIS

[14] Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

[15] Prediction of Low Birth Weight Infants and Its Risk Factors Using Data Mining Techniques
Senthilkumar D

[16] machine learning algorithms to predict childhood Anemia in bangladesh Jahidur Rahman Khan1 2, Srizan Chowdhury3, Humayera Islam3-4, Enayetur Raheem2

[17] Prediction of Anaemia among children using Machine Learning Algorithms Priyanka Anand1, Rahul Gupta2 and Ankita Sharma3

[18] Predicting child anaemia in the North-Eastern states of India: a machine learning approach A. Jiran Meitei1 · Akanksha Saini2 · Bibhuti Bhusan Mohapatra3 · Kh. Jitenkumar Singh4

[19] Zahirzada, A., & Lavangnananda, K. (2021, January). Implementing predictive model for Low Birth Weight in Afghanistan. In 2021 13th International Conference on Knowledge and Smart Technology (KST) (pp. 67-72). IEEE.

[20] Niazai, A. J., Zahirzada, A., Shahpoor, M. A., & Safi, A. R. (2020, December). Time series forecasting of registered, recovered, and death cases of covid-19 for the next sixty days in afghanistan. In 2020 IEEE international conference on advent trends in multidisciplinary research and innovation (ICATMRI) (pp. 1-6). IEEE.

[21] K. Lavangnananda and S. Chattanachot, “Study of Discretization Methods in Classification”, in Proc. of the 9th Int. Conf. on Knowledge and Smart Technology (KST-2017), 1st - 4th February, Pattaya, Thailand, 2017.

[22] K. B. Al Janabi, and R. Kadhim, “Data Reduction Techniques: A Comparative Study for Attribute Selection Methods,” IJACST. ISSN.

[23] M. Hossin and M. N. Sulaiman, “A review on evaluation metrics for data classification evaluations,” International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, p. 1, 2015.

[24] WEKA Machine Learning software to solve data mining problems [Online]. Available:
https://sourceforge.net/projects/weka
[Accessed: 30th-Mar-2020].