



# Predicting child anaemia in the North-Eastern states of India: a machine learning approach

A. Jiran Meitei<sup>1</sup> · Akanksha Saini<sup>2</sup> · Bibhuti Bhusan Mohapatra<sup>3</sup> · Kh. Jitenkumar Singh<sup>4</sup>

Received: 2 March 2022 / Revised: 14 July 2022 / Accepted: 8 August 2022 / Published online: 4 September 2022

© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2022

**Abstract** Child anaemia is a serious global health issue and India is one of the highest contributors among the developing nations. Researchers identify many harmful effects of anaemia, which include psychomotor retardation, which in turn decreases the learning ability and causes low intelligence among pre-school children. The effects also include behavioural delays, low immunity, and susceptibility to frequent infections, increased mortality, and disability. The present study aims to predict anaemia among children in North-East India by applying Machine Learning (ML) algorithms to latest available National Family Health Survey (NFHS)-4 data. Out of the total 29,312 eligible children (6–59 months) in North-East India, a total of 21,000 children with demographic variables without any missing observations, wherein 10,460 are anaemic, is considered for this study. Machine learning (ML) algorithms have been applied through 3 different types of penalized regression methods—ridge, least absolute shrinkage and selection operator, and elastic net

for predicting anaemia. A systematic assessment of algorithms is performed in terms of accuracy, sensitivity, specificity, F1-Score, and Cohen's *k*-Statistics. Having achieved the receiver operating characteristic value of over 70% in training and accuracy of above 64% while testing, it can be safely asserted that factors like mother's anaemic status, age of the child, social status, mother's age, mother's education, religion are important in identifying the child as anaemic.

**Keywords** Anaemia · Elastic net · LASSO · Machine learning · Penalized regression · Ridge

## 1 Introduction

Anaemia is a serious global health problem that specifically affects young children and pregnant women (Jiahong et al. 2021; Chaparro and Suchdev 2019; WHO 2015; Gutema et al. 2014). It may result from several factors, the iron deficiency is the main contributor, but the proportion probably varies among population groups in different areas or according to the local conditions (Steven et al. 2013; Stoltzfus 2004). Other causes of anaemia include micronutrient deficiencies (e.g., folate, riboflavin, vitamin A and B12), acute and chronic infections (e.g., malaria, cancer, tuberculosis, and HIV), and inherited or acquired disorders that affect haemoglobin synthesis, red blood cell production or red blood cell survival (e.g., hemoglobinopathies) (Dey et al. 2013; Balarajan et al. 2011; Tolentino and Friedman 2007).

Iron deficiency anaemia is considered one of the 'Top Ten Risk Factors' causing death (Dubey 1994). People suffering from anaemia lack red blood cells that are responsible to carry oxygen to the body's tissues. Its symptoms may include fatigue, skin pallor, shortness of breath, light-headedness, dizziness or fast heartbeat, low body weight, etc. (Cho et al.

✉ Bibhuti Bhusan Mohapatra  
bbmohapatra@mac.du.ac.in

A. Jiran Meitei  
ajmeitei@mac.du.ac.in

Akanksha Saini  
akankshasaini.du.or@gmail.com

Kh. Jitenkumar Singh  
jitensingkh@gmail.com

<sup>1</sup> Department of Mathematics, Maharaja Agrasen College, University of Delhi, New Delhi, India

<sup>2</sup> Department of Operational Research, University of Delhi, New Delhi, Delhi 110007, India

<sup>3</sup> Department of Commerce, Maharaja Agrasen College, University of Delhi, New Delhi, India

<sup>4</sup> ICMR-NIMS, New Delhi, India

2021). If untreated it can lead to severe complications among children and adults, either in the short or long term. As far as the children are concerned, especially children below 5 years of age, anaemia affects them with behavioural delay, low cognitive development and intelligence, difficulty with concentration, low learning outcomes, low immunity, and susceptibility to frequent infections, increased mortality, and disability (Agaoglu et al. 2007; Katzman et al. 1972; Zou and Hastie 2005). It also affects adults, especially women, in the form of pregnancy complications such as premature birth, abortion, and low birth weight of babies, and the same is for adolescents too (Lokeshwar et al. 2011; Qiaoyi et al. 2009). Finally, untreated anaemia also results in heart attack and heart failure (Shah et al. 2013). Recent studies suggest that anaemia is associated with poor outcomes for patients hospitalized with COVID-19 infections (Faghhih et al. 2021; Rajanna et al. 2021).

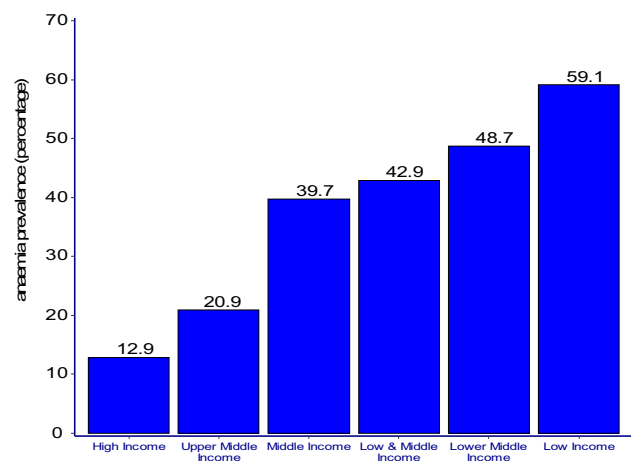
According to the World Health Organization (WHO), children of 6–59 months of age, and women in the child-bearing age group, especially those who are pregnant, are prone to anaemia. WHO (2021) further states that the prevalence of anaemia among 6–59-month-old children across the world in 2019 was around 40% while the same in the case of pregnant women was around 36.5%. Thus, anaemia among the 6–59 age group of children is a matter of serious concern in the world.

However, the prevalence of anaemia<sup>1</sup> varies widely across the regions and countries in the world. It is as high as 60% in Africa and as low as 7% in North America. And within Africa, it is as high as 69% in western and central Africa, the poorest regions of the continent. As far as the individual countries are concerned, Yemen records the highest rate (around 79.5%), and the lowest in the USA (around 6%). Owing to internal conflict and infighting, Yemen suffers from the worst humanitarian crisis—nearly three fourth of its population, especially women, suffer from extreme poverty (World Bank 2021), whereas, the USA is one of the richest countries in the world. In fact, going by strict economic criteria, the prevalence of anaemia in low-income countries is as high as 74% whereas the same in the case of high-income countries is around 13% (Fig. 1).

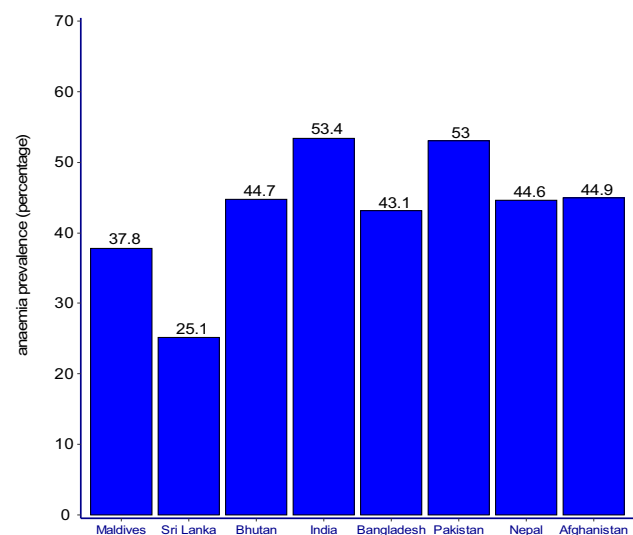
Clearly, the pattern of anaemia prevalence in the world indicates its possible negative correlation with income. However, impressively no such pattern is apparent in the context of South Asia in general and India in particular.

Despite being one of the poorest regions of Asia, South Asia has anaemia prevalence to an extent of 52%. Most of the countries in the region do not show a similar pattern.

<sup>1</sup> Henceforth throughout the paper, anaemia will refer to prevalence among children in 6–59 months age group.



**Fig. 1** Prevalence of anaemia in children (6–59 months) in the world. Source: Authors' calculations based on data published by Our World in Data (<https://ourworldindata.org>)

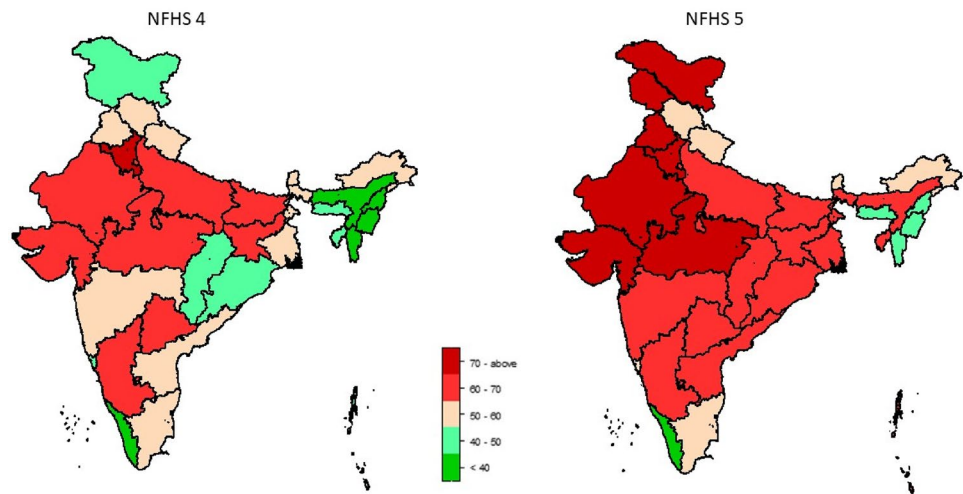


**Fig. 2** Prevalence of anaemia in children (6–59 months) in SAARC countries. (Arranged in descending order of GDP (PPP)). Source: Same as Fig. 1

Countries like Afghanistan, Bangladesh, Nepal, and Pakistan exhibit a lower prevalence of anaemia than India (Fig. 2).

As per the National Family Health Survey (NFHS)-4 (2015–16) survey report, anaemia prevalence rate is 58.5%, which of course is higher than the WHO figures for 2019. It also suggests that the economically developed regions, states, and UTs in the country have a higher prevalence of anaemia than their poorer counterparts, both in terms of its severity and proportion. Economically sound regions in North and South India have a higher and more severe prevalence of anaemia than economically weak regions like North-East India. Similarly, the proportion and severity of anaemia are higher in the richer states like Punjab, Haryana,

**Fig. 3** Heat map of child anaemia as per NFHS-4 and NFHS-5. Source: Authors' Compilations Based on NFHS-4 and NFHS-5, Published by MoHFW, Govt. of India



Anemia is showing increased from NFHS 4 to NFHS 5

Delhi, Gujarat, Maharashtra, Andhra Pradesh, Telangana, etc. whereas, it is the opposite in the case of economically less well-off states like Arunachal Pradesh, Assam, Manipur, Meghalaya, etc., All the states in North-East India are having a lesser prevalence than the national average, with Sikkim (55.1%) having the highest followed by Arunachal Pradesh (54.2%), while the rest of the states in the region have reported a prevalence of less than 50%. The available factsheets of the NHFS-5 (2019–21) survey report also suggests a similar pattern in a more alarming situation (Fig. 3).

The national prevalence rate has gone up drastically from 58.6 in NFHS-4 to 67.1% in NHFS-5 and across the states and UTs barring the state of Kerala, there has been rise in child anaemia during the same period. However, on the one hand, high prevalence of anaemia in India compared to its South Asian neighbours, and on the other hand within India, relatively low prevalence of anaemia in some North-Eastern states, Himachal Pradesh, Uttarakhand, and Sikkim, more or less counter the global scenario of anaemia that depicts its negative correlation with the income.

### 1.1 Broad research question and review of literature

The major research question that arises is 'Do factors other than income matter in the prevalence of anaemia among the children in the 6–59 months age group?' The present study seeks to explore an answer to the aforesaid research question in the context of North-East Indian states. This is because, as mentioned earlier, the region is relatively less well-off in terms of economy, child and woman welfare, and has a lower prevalence of anaemia. Besides, North-East India is home to a large chunk of tribal population whose living environment, social norms, culture and consumption pattern, etc., vary markedly from the economically well-off states in India. Further, the region is usually considered an

economically less developed region of India and hence since 2014, it has been the focus of the central government. The gross budgetary support of the central government for the North-East has gone up from Rs. 36,108 Crores in 2014–15 to Rs. 76,040 Crores in 2022–23, an increase of 110%. Pertinently, there can't be economic development without human development.

As far as the available literature is concerned, although there are a plethora of studies on the prevalence of child anaemia in India, there is hardly any specific study conducted in the context of North-East India. In this respect, however, we could trace one direct study and a couple of related studies.

Dey S et al. (2013), examined the factors that influence the occurrence of anaemia among children of 0–6 years in North-East India. They used a data set of 10,317 children from the Reproductive and Child Health-II (RCH-II) survey conducted between 2002 and 2004. Using the chosen data, the authors attempted to predict the probability of anaemia occurrence among the target group by fitting a multinomial logistic model. They found that the geographical location (rural or urban), religion, fertility and literacy of the mother, and age of the mother at marriage are significant determinants for the prevalence of anaemia among children of 0–6 years in North-East India. Meshram I. et al. (2020) assessed the prevalence of anaemia and vitamin A deficiency (VAD) among women and pre-school children in North-East India through a small sample survey. The study found the prevalence of anaemia to be low among pre-school children. The authors however suggest that Anaemia and VAD are important public health problems among the tribal population of North-East India, despite their rich biodiversity. Bezboruah et al. (2021) conducted a cross-sectional study of 104 HIV-positive children in one of the tertiary care centers in North-East India. According to their study, compared to the

older age group preschool children had a higher prevalence of anaemia. Further, the authors put forth that those rural children are more affected both in terms of prevalence and severity of anaemia. And in their study malnutrition is found to be an important risk factor for anaemia. Accordingly, the authors prescribe nutritional programs for improving the quality of life among HIV-infected children, especially those belonging to rural India.

De M et al. (2006) studied the prevalence of anaemia and hemoglobinopathies in the tribal population of North-East India through a sample survey of 1726 cases. Out of the total number of cases, approximately 73% were from the tribal population, collected from three states in the region, namely, Arunachal Pradesh, Assam, and Tripura, and the rest were non-tribal populations, collected from the state of West Bengal, as a control group. The incidence of anaemia among the tribal population of North-East states was significantly different from that of West Bengal. In particular, the study found the incidence of anaemia in Arunachal Pradesh, Assam, and Tripura were around 54%, 60%, and 57% respectively. Finally, the authors opined that the presence of hemoglobinopathies and thalassemia accounted for anaemia in a sizeable population of certain tribes in North-East India and urgent public health programmes were needed to address the issue.

Thus, the available literature suggests that the demographic variables like tribal and rural inhabitations, nutritional deficiency, literacy of mother, age of the mother during the marriage, fertility of women, etc., are responsible for anaemia in children of 6–59 months age group in North-East India. However, over the years there has been a rapid stride in the collection of family health statistics in India—we have moved from RCH to NHFS. Compared to RCH data used in Dey S et al. (2013), the NHFS-4 is a large database both in terms of sample size and variables. Moreover, the statistical techniques applied to analyze the survey has gone up tremendously—from multinomial logistics used in Dey S et al. (2013) to machine learning (ML) techniques like penalized regression. Of course, there has been a number of studies on child anaemia using NHFS-4 data and applying ML techniques but not in the context of North-East India. A few important of them are worth highlighting in this context.

Meena et al. (2019) applied data mining techniques such as decision tree and association rule mining to NHFS-4 to predict child anaemia in India. Similarly, Jain et al. (2021) analyzed the NHFS-4 data for child malnutrition. Using multilevel analysis, the authors found households as an important source of clustering and variation in child malnutrition outcomes. In predicting anaemia and malnutrition in children, ML techniques have also been used outside India at a global level. Talukder and Ahammed (2020) also compared various ML algorithms to predict malnutrition status for children under the age of five using the Bangladesh

Demographic and Health Survey (BDHS). Similarly, Wallner et al. (2022) tried to predict the occurrence of anaemia among children in the UK through CART analysis and ML techniques. Qusay and Emrullah (2022) have compared the performance of different ML techniques for anaemia prediction among children using various social factors. The authors viewed that the ML techniques like Multilayer Perceptron (MLP) and Decision Tree (DT) better predict the prevalence of anaemia among children than the traditional statistical methods. Bitew et al. (2022) using the data from the Ethiopian Demographic and Health Survey of 2016 compared 5 different ML algorithms to predict the socio-demographic risk factors for undernutrition. Dukhi et al. (2021) have reviewed the studies pertaining to the use of artificial intelligence in analyzing the prevalence of anaemia among children and adolescents in India, South Africa, and Russia. The authors opined that although the use of ML approach for the study of child anaemia is at a nascent stage, it could be used as a potential tool in identifying the risk associated with child anaemia at preliminary levels.

Thus, in recent years ML techniques have been widely used, as an improvement over the traditional methods, in predicting child anaemia and related health issues of children. However, we could not find a single study that has used either NHFS-4 data or any of the ML techniques in predicting child anaemia in North-East India, given the central government's focus on it in recent years.

Nevertheless, taking the aforesaid gaps in literature into consideration, the present study intends to evaluate the prevalence of anaemia among children (6–59 months) and examine the role of a wide variety of factors in it, using a sufficiently large set of the latest available demographic data i.e., NHFS-4 data. For this purpose, we have used one of the important ML techniques known as penalized logistic regression methods with the help of classification and regression training (CARET) (Kuhn M 2008) package in R. Thus, the present study is distinguished from the previous studies in terms of a wide variety of factors/determinants, a large sample size from the latest available data and use of sophisticated analytical methods.

The study is organized as follows. Apart from the current introductory section which includes research question and literature review, there are 4 more sections. The Sect. 2 is devoted to a threadbare discussion of methodology. In Sect. 3, we outline the nuance of the data, select the independent variables and evaluate the appropriate model suitable for the analysis. The results are discussed in the Sect. 4. Finally, we summarise and conclude the entire discussion by highlighting the limitations of the study.



## 2 Methodology

Machine Learning (ML) is an emerging analytical technique in quantitative research (Donepudi 2017). Being a data-centric technique, ML follows a large number of algorithms and widely used among them is supervised ML. Some of the popular supervised ML algorithms are decision tree, random forest, support vector machine, k nearest neighbourhood, penalised regression, etc. Penalised regression is widely used in analysing health and demographic data.

### 2.1 Penalized regression: the supervised ML algorithm

Traditionally, logistic regression is one of the most popular linear classification methods in the field of healthcare, banking, and other related areas. It has been found useful for binary classification wherein the dependent variable is known to have only 2 classes. One of the major limitations of using logistic regression is the high dimensionality of the data under investigation, especially in those cases where the sample size is less than the number of variables, in areas such as genomics, fMRI data, etc. Further, the use of too many predictive variables makes the regression exercise complex, and it even affects the predictive accuracy of the model. Also, it encounters the problems of multi-collinearity and over-fitting. These problems are well addressed when we use penalized regression methods (Greenwood et al. 2020, Abram et al. 2016, Aitor and Juan 2011). Penalized regression is an improvement over binary logistic regression as it allows us to handle complex regression problems and attain higher predictive accuracy. Penalized regression models impose a penalty on the logistic regression coefficients for having many explanatory variables. This in turn results in the shrinking of the regression coefficients of less important variables towards zero. This process is also known as ‘Regularization’.

The prominent penalized regression methods are ridge, LASSO (Least Absolute Shrinkage and Selection Operator), and elastic-net. The ridge regression (Arthur and Robert 1970) also known as the ‘Quadratic Regularization’ approach is the oldest penalized regression method. The ridge regression shrinks the regression coefficients by imposing  $l_2$ -norm penalty for having a large value. It is also known to shrink the coefficients of the correlated explanatory variables towards each other by acquiring strength from each other (Friedman et al. 2010). However, one important drawback of ridge regression is its inability to select the important variables. An improvement over a ridge could be LASSO, purposed by Robert (1996) using the  $l_1$ -norm penalty. It eliminates the least important variables by forcing their coefficients to be exactly zero. Hence, LASSO is otherwise known as the variable selection method. In fact, it not only improves the accuracy of

the classification but also makes the interpretation of the model easier (Pourahmadi 2013). Although this method demonstrated encouraging results, Zou and Hastie (2005) pointed out some shortcomings. The LASSO method is known to have problems when the explanatory variables are correlated and variables more than the size of the sample. Hence, we use elastic-net proposed by Zou and Hastie as an improvement over LASSO. Basically, elastic-net is a combination of the ridge and LASSO that overcome their individual drawbacks.

Penalised regression technique, as an improvement over classical logistic regression can be seen though its various functional forms.

$$P(y|x) = \pi(x) = \text{logit}^{-1} \left( \beta_0 + \sum_{j=1}^m \beta_j x_j \right) \quad (1)$$

here  $\text{logit}^{-1}$  is the inverse of the logit transformation,  $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)$  are the regression coefficients and  $x = (x_1, x_2, x_3, \dots, x_m)$  are the explanatory variables. These coefficients are obtained by maximizing the log-likelihood function,  $l(\beta)$  over the total given observations  $n$ .

$$\ell(\beta) = \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} \quad (2)$$

Then the penalized logistic regression,  $l(\beta)^p$  can be defined as below:

$$\ell(\beta)^p = \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda P(\beta) \quad (3)$$

$\lambda \geq 0$ , is the tuning parameter and controls the shrinkage of the coefficients of the explanatory variables of the model. The larger the value of this tuning parameter the more the weight to the penalty term  $P(\beta)$ . When this penalty term is replaced by  $l_2$ -norm penalty, we have ridge regression;

$$\ell(\beta)_{\text{ridge}}^p = \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda \sum_{j=1}^m \beta_j^2 \quad (4)$$

And the solution to the likelihood Eq. (4) is

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \left\{ \ell(\beta) + \lambda \sum_{j=1}^m \beta_j^2 \right\} \quad (5)$$

In the ridge regression, the tuning parameter  $\lambda$  only controls the amount of shrinkage in the regression coefficient but never takes them exactly equal to zero. When  $P(\beta)$  is put equal to  $l_1$ -norm penalty we have the LASSO regression;

$$\ell(\beta)_{\text{lasso}}^p = \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} + \lambda \sum_{j=1}^m |\beta_j| \quad (6)$$

And the solution to the likelihood Eq. (6) is

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \ell(\beta) + \lambda \sum_{j=1}^m |\beta_j| \right\} \quad (7)$$

Unlike the ridge, in LASSO, the tuning parameter will make some of the regression coefficient equal to zero, and thus eliminating the least important variable from the model.

When  $l_1$ -norm and  $l_2$ -norm penalties are used simultaneously we have the elastic net regression;

$$\begin{aligned} \ell(\beta)_{\text{elastic}}^p &= \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\} \\ &+ \lambda \left[ \left( \frac{1 - \alpha}{2} \right) \sum_{j=1}^m \beta_j^2 + \alpha \sum_{j=1}^m |\beta_j| \right] \end{aligned} \quad (8)$$

here the  $l_1$ -norm is responsible for variable selection by setting coefficients some of the variables exactly zero and  $l_2$ -norm does the job of shrinking the coefficient of the correlated variables with each other. In this way, the elastic net automatically handles the problem of multicollinearity in the model.

In (8), if  $\alpha = 0$ , then the elastic net will give ridge, and if  $\alpha = 1$  then it will give LASSO regression. Thus  $\frac{\lambda}{2}$  is equivalent to the tuning parameter of the ridge and  $\lambda$  is the LASSO tuning parameter.

Packages like CARET can handle both classification and regression models.

### 3 Data

We have used the data from National Family Health Survey-4 conducted by the International Institute for Population Science during 2015–2016 under the Ministry of Health and Family Welfare, Government of India. In the NFHS-4, haemoglobin testing was conducted on the children (6–59 months) using the capillary blood for identification and categorisation of anaemia. The anaemia was categorised as: severely anaemic (i.e., > 7.0 g/dl Hb level), moderate anaemic (i.e., 7.0–9.9 g/dl Hb level), and mild anaemic (10–10.9 g/dl Hb level). In the survey, total number of eligible children in North-East India was 29,312, out of which 10,504 children were found to be anaemic.<sup>2</sup> A total

of 17 demographic variables were identified for our study and basis on it, we dropped all the rows with any missing observation. As a result, we were left with 10,460 anaemic and 18,725 non-anaemic children. Further, to make the data symmetric in terms of child anaemic status, 10,540 non-anaemic children are selected using the simple random sampling technique from 18,725 making the total dataset into 21,000 children.

### 3.1 Independent variables

The set of sixteen independent variables considered is based on child-mother, household and socio-economic characteristics. Child characteristics include Sex (male, female), Child Age (in months) (6–23, 24–59), Child's Size (large, average, small), Breastfeeding (no, yes). Mother's characteristics have Mother's Age (in years) (15–19, 20–29, 30–39, 40–49), Mother's Education (no education, primary, secondary, higher), Mother's Anaemic Status (severe, moderate, mild, no). Household characteristics include Place of Residence (urban, rural), Sanitation (Hygienic, No-Hygienic), Disposal of Youngest Child Stool (safe, unsafe), Safe Drinking Water (yes, no), Household Size (<=4, 5–7, >=8) and Number of Living Child (1, 2, 3, >=4). The socioeconomic information includes Wealth Quintile (poorest, poorer, middle, richer, richest), Religion (Hindu, Muslim, Others), Social Status (Schedule Caste, Schedule Tribe, OBC, and Others). Thus, in the aggregate 48 independent variables have been considered comprising aforesaid sixteen variables along with their subclasses. Table 1 gives the complete list of variables and their associated acronym.

### 3.2 Model evaluation

The performance of the models is evaluated based on receiver operating characteristic (ROC) curve, sensitivity, and specificity for the training outcomes whereas accuracy, sensitivity, specificity, precision (positive predictive value), negative predictive value F1 score, and Cohen's Kappa value will be used for testing evaluation. The ROC curve is one of the best methods of accessing the performance of a classification algorithm. The area under the ROC curve (AUROC) is used as a basis for checking the discriminative adeptness of the model. AUROC or simply ROC value of a test is categorized as: 0.5–0.6 (fail), 0.6–0.7 (poor), 0.7–0.8 (fair), 0.8–0.9 (good) and 0.9–1.0 (very good). Accuracy is the degree of closeness to the true value of the population parameters. It is used for evaluating classification models to measure the proportion of cases of reproducibility (i.e., repeating the same value) of the measure set. Sensitivity is the proportion of true positives predicted as true

<sup>2</sup> All the children with mild, moderate and severe anaemic conditions are identified as anaemic in this study.

**Table 1** List of independent variables. Source: Authors’ contribution

Original variables	Sub-division	Original variables	Sub-division
Child age (CA)	6–23 months (CA.1)	Sanitation (SNI)	Hygienic (SNI.0)
	24–59 months (CA.2)		Non-hygienic (SNI.1)
Child sex (SEX)	Male (Sex.1)	Disposal of youngest child stool (DYCS)	Safe (DYCS.0)
	Female (Sex.2)		Unsafe (DYCS.1)
Child size at birth (CS)	Large (CS.1)	Safe drinking water (SDW)	No (SDW.0)
	Average (CS.2)		Yes (SDW.1)
	Small (CS.3)		
No. of living child (LC)	1 (LC.1)	House hold size (HHSIZE)	≤ 4 (HHSIZE.1)
	2 (LC.2)		05–07 (HHSIZE.2)
	3 (LC.3)		≥ 8 (HHSIZE.3)
	> =4 (LC.4)	Wealth index (WI)	Poorest (WI.1)
Breastfeeding (BF)	No (BF.0)		Poorer (WI.2)
	Yes (BF.1)		Middle (WI.3)
Mother anaemic status (MAS)	Severe (MAS.1)		Religion (REL)
	Moderate (MAS.2)	Richest (WI.5)	
	Mild (MAS.3)	Hindu (REL.1)	
	No (MAS.4)	Muslim (REL.2)	
Mother education (ME)	No education (ME.0)	Social status (SS)	Others (REL.3)
	Primary (ME.1)		SC (SS.0)
	Secondary (ME.2)		ST (SS.1)
	Higher (ME.3)		OBC (SS.2)
Mother age (MAGE)	15–19 years (MAGE.1)	Place of residence (PR)	Others (SS.3)
	20–29 years (MAGE.2)		Urban (PR.1)
	30–39 years (MAGE.3)		Rural (PR.2)
	40–49 years (MAGE.4)		

**Table 2** Bivariate results of child characteristics. Source: Author’s Estimates

Child characteristics		Anaemic	Non-anaemic	% of Anaemic child	$\chi^2$
Child age	6–23 months	4514	2644	63.06	762.944***
	24–59 months	5946	7896	42.96	
Child sex	Male	5449	5411	50.17	1.201
	Female	5011	5129	49.42	
Child size at birth	Large	1882	1947	49.15	11.399**
	Average	7377	7238	50.48	
	Small	1201	1355	46.99	
Breastfeeding	No	3086	3569	46.37	46.072***
	Yes	7374	6971	51.4	

Level of significance: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

by the classifier whereas specificity is the proportion of true negative cases predicted as negative by the classifier. Precision, also known as the positive predictive value, gives the proportion of the true positive out of the total predicted positive while negative predictive

value is the proportion of true negative out of the total predicted negative. The harmonic mean of accuracy and precision is defined as the F1 score or F-value. It evaluates the overall model’s accuracy. F-measure affects the false positives and false negatives. A good F-measure

**Table 3** Bivariate results of mother's characteristics. Source: Author's Estimates

Mother's characteristics		Anaemic	Non-anaemic	% of Anaemic child	$\chi^2$
Mother's anaemic status	Severe	72	41	63.72	532.189***
	Moderate	1304	659	66.43	
	Mild	4090	3258	55.66	
	No	4994	6582	43.14	
Mother's education	No education	556	799	41.03	94.356***
	Primary	2484	2183	53.22	
	Secondary	1896	1642	53.59	
	Higher	5524	5916	48.29	
Mother's age	15–19 years	348	269	56.4	42.356***
	20–29 years	6252	5958	51.2	
	30–39 years	3356	3769	47.1	
	40–49 years	504	544	48.09	

Level of significance: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

means the model has low false positives and low false negatives. For a model, an F1 score of 1 is considered as a perfect model whereas a value of 0 is a total failure. And finally, Cohen's  $k$  Statistics measures are useful in those problems where the data is imbalanced or involves a multiclass classification or both. This metric gives information about the agreement between predicted and estimated values.

## 4 Results and discussions

### 4.1 Bivariate analysis

A bivariate analysis is conducted before applying the ML techniques. In child characteristics, out of 21,000 children, 7,158 have come under the age group of 6–23 months having 63.06% (4514) prevalence of anaemia, the remaining children are under 24–59 months with 42.96% (5946) prevalence and  $p$ -value  $< 0.001$  indicating that there is a highly significant relationship between CAS and CA. The data have a total of 10,860 male children with a prevalence rate of 50.17% (5449) and females have a prevalence of 48.88% (5,011) and the  $p$ -value, in this case, is greater than 5%, highlighting that there is no significant relationship between CAS and Sex. The average-sized children have the highest prevalence of 50.48% (7377) followed by large-sized children 49.15% (1882) and the small-sized children have 46.99% (1201). The prevalence rate of breastfeed children is 51.4% (7374) than the children without breastfeeding 46.37% (3086). The  $p$ -values for the CS and BF indicate that there are significant relationships of CAS with both CS and BF (Details can be seen from Table 2).

Mothers with moderate anaemic status have the highest prevalence of anaemic child 66.43% (1,304) followed by

severely anaemic mothers 63.72% (72) and then mothers with mild anaemia 55.66% (4,090). Non-anaemic mothers have the lowest prevalence of 43.14% (4994). The  $p$ -value for the chi-square test also indicates a strong significance between CAS and MAS. The prevalence rate of child anaemia is highest among secondary educated mothers, 53.59% (1896), and lowest among illiterate mothers, 41.03% (556). In the meanwhile, there was a prevalence of 53.22% (2484) and 48.29% (5524) among the mothers with primary and higher educational backgrounds. In the different age groups of mothers, 348 children were found to be anaemic with a prevalence rate of 56.4% for mothers under the age group of 15–19 years, followed by a prevalence rate of 51.20% (6252) among the mothers in the age group of 20–29 years, 48.09% (504) for the age group of 40–49 years and finally, 47.10% (3356) for mother in the age group of 30–39 years. The  $p$ -values of ME, and MAGE are all less than 0.001, indicating a strong relationship between CAS and ME, MAGE (Details can be seen from Table 3).

In household characteristics, households with non-hygienic sanitation facilities have a higher prevalence of 50.66% (7145) than hygienic sanitation facility households with 48.07% (3315). The prevalence is higher among the children with access to safe drinking water 50.66% (7145) than unsafe 48.07% (3315). In the disposal of the youngest child stool, unsafe disposal has 52.63% (7316) whereas safe disposal has 44.28% (3144). The household size less than 4 has the highest prevalence of 51.22% (3478) and the greater than 8 has the least 48.48% (1836). In the case of number of living children, families with more than 4 living children have the highest prevalence 53.64% (2240) and the least in those families with 2 living children, 48.38% (3302). Children who live in rural areas have a higher prevalence of 49.97% (8618) as compared to the children in urban areas 49.05% (1842). The chi-square values in Table 4 indicate



**Table 4** Bivariate results of household characteristics. Source: Author’s Estimates

Household characteristics		Anaemic	Non-Anaemic	% of Anaemic child	$\chi^2$
Sanitation	Hygienic	3315	3581	48.07	12.408***
	Non-hygienic	7145	6959	50.66	
Disposal of youngest child stool	Safe	3144	3956	44.28	131.111***
	Unsafe	7316	6584	52.63	
Safe drinking water	No	3315	3581	48.07	12.409***
	Yes	7145	6959	50.66	
House hold size	≤ 4	3478	3312	51.22	8.892**
	05–07	5146	5277	49.37	
	≥ 8	1836	1951	48.48	
No. of living child	1	3035	3189	48.76	32.814***
	2	3302	3523	48.38	
	3	1883	1892	49.88	
	> = 4	2240	1936	53.64	
Place of residence	Urban	1842	1913	49.05	1.042
	Rural	8618	8627	49.97	

Level of significance: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

**Table 5** Bivariate results of socio-economic characteristics. Source: Authors’ Estimates

Socio-economic characteristics		Anaemic	Non-anaemic	% of Anaemic child	$\chi^2$
Wealth index	Poorest	2411	2423	49.88	11.380**
	Poorer	3631	3702	49.52	
	Middle	2558	2420	51.39	
	Richer	1372	1427	49.02	
	Richest	488	568	46.21	
Religion	Hindu	3327	4564	42.16	954.080***
	Muslim	1419	2417	36.99	
	Others	5714	3559	61.62	
Social status	SC	2477	4051	37.94	1.0e + 03***
	ST	6212	3934	61.23	
	OBC	612	805	43.19	
	Others	1159	1750	39.84	

Level of Significance: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

that except for a place of residence, CAS is showing a strong relationship between sanitation, disposal of the youngest child stool, safe drinking water, household size and the number of living children.

In the wealth quintile, the middle-class has the highest prevalence 51.39% (2558), followed by the poorer and poorest class which have an almost similar prevalence of 49.52% (3631) and 49.88% (2411). Richer people have a prevalence of 49.02% (1372) while the richest people have the lowest prevalence of 46.21% (488). In religion, Hindu and Muslims have prevalence 42.16% (3327) and 36.99% (1419) respectively, while others have the highest prevalence of 61.62%

(5714). Among the social strata, ST has the highest prevalence of 61.23% (6212) followed by OBC, others and SC with 43.19% (612), 39.84% (1159) and 37.94% (24,477) respectively. The chi-square values in Table 5 indicate there is a good degree of association between CAS and WI, REL, SS.

### 4.2 Training reports

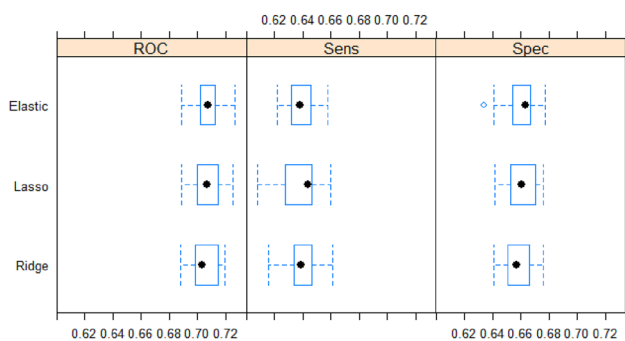
Before we develop the machine learning models, the hot encoding process is used to convert all the (categorical) variables into multiple variables, each with a value of 1 or 0. The whole data is randomly partitioned into 80:20 as commonly practiced in ML techniques (Gholamy et al. 2018). Eighty percent as training (16,800) and twenty percent as testing (4200) datasets. The method of repeated cross-validation is considered to avoid the problem of overfitting.

The training outputs are then compared based on ROC, sensitivity, and specificity. Table 6 presents these performance measures and the best values of alpha and lambda on the training dataset. We can see from the table that the ROC values of the three models are almost the same with LASSO having a slightly higher value of average ROC and closely followed by elastic net and ridge respectively. LASSO also has the highest average values of both sensitivity and specificity. When it comes to the best median values, elastic net has the best ROC and specificity. LASSO again has the best median value for sensitivity. Figures 4 and 5 also give the model comparison based on box plots and 95% confidence intervals (CI).

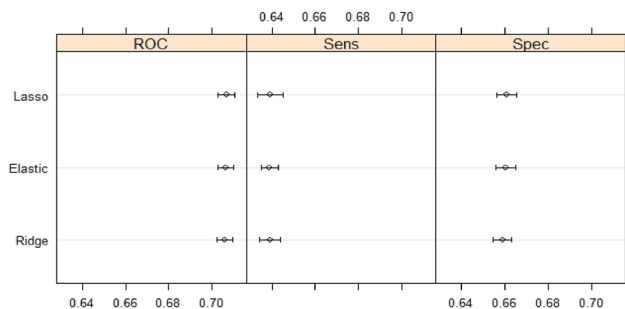
The models depicted in Figs. 6, 7 and 8 are ridge, LASSO and elastic net models respectively. These figures represent the plotting of penalized variables as a function of the

**Table 6** Comparison of training results. Source: Authors' Estimates

	Min	Q1	Q2	Mean	Q3	Max
<i>ROC</i>						
Ridge	0.6878	0.6985	0.7035	0.7059	0.7144	0.7195
LASSO	0.6884	0.6996	0.7065	0.7065	0.7147	0.7250
Elastic	0.6889	0.7018	0.7077	0.7062	0.7121	0.7263
<i>Sensitivity</i>						
Ridge	0.6153	0.6338	0.6390	0.6391	0.6464	0.6613
LASSO	0.6081	0.6278	0.6434	0.6393	0.6467	0.6601
Elastic	0.6219	0.6318	0.6384	0.6391	0.6455	0.6575
<i>Specificity</i>						
Ridge	0.6408	0.6509	0.6568	0.6587	0.6661	0.6763
LASSO	0.6414	0.6530	0.6609	0.6609	0.6702	0.6762
Elastic	0.6343	0.6542	0.6631	0.6603	0.6673	0.6773



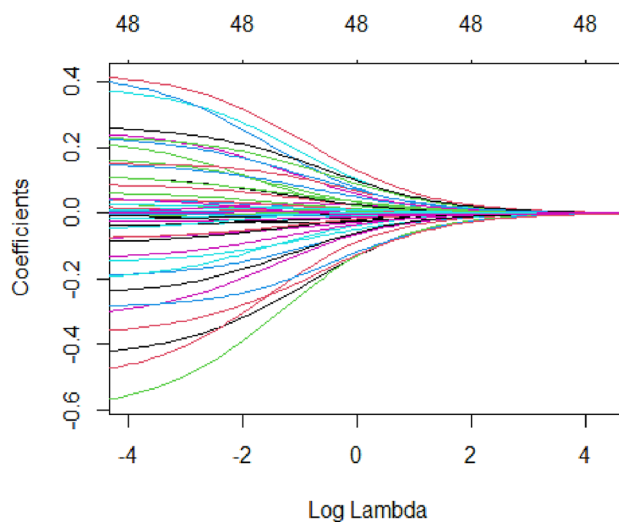
**Fig. 4** Accuracy comparison in terms of ROC, Sensitivity and Specificity



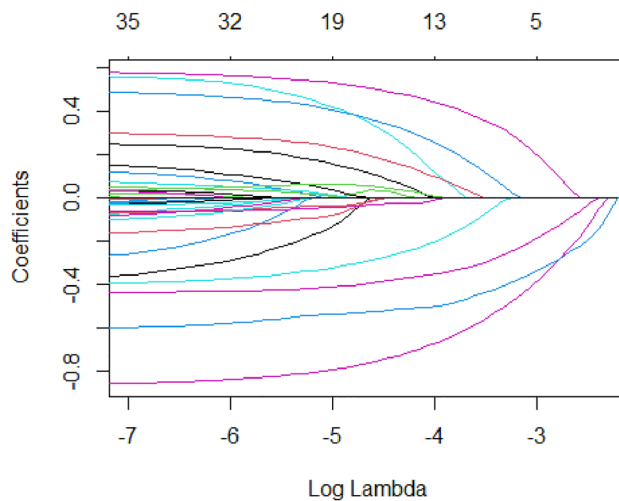
**Fig. 5** Accuracy comparison in terms of 95% confidence intervals

regularisation parameter. Each colour represents a different variable in the plots.

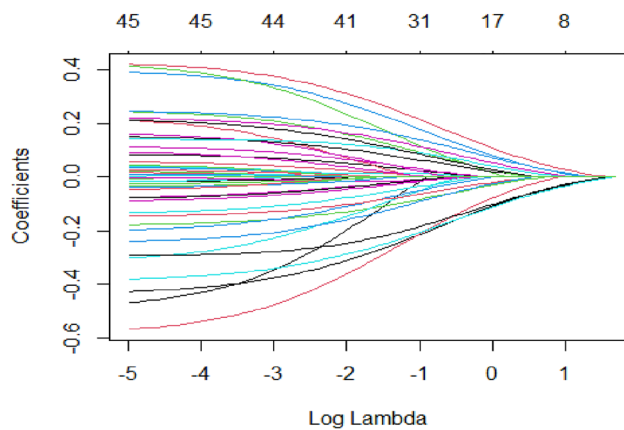
In Fig. 6, we can see that all variables are shrinking towards zero because of the property of ridge regression in which correlated variables shrink towards each other. The variables that shrink initially are the least important whereas the most important variables shrink at last. In Fig. 7, variables follow the LASSO property of selecting only one variable among the correlated variables and eliminating others



**Fig. 6** Final ridge model, lambda=0.0232



**Fig. 7** Final LASSO model, lambda=0.0020



**Fig. 8** Final elastic model,  $\lambda=0.030$  and  $\alpha=0.02$

by making them equal to zero. In Fig. 8, the plotting pattern follows both shrinking and selection of variables as it is the generalised case of ridge and LASSO regression.

### 4.3 Testing reports

The models are now tested with the test data. All the models are giving almost the same level of accuracy. LASSO with an accuracy of 0.6429 and Kappa 0.2856 is the best. LASSO is closely followed by ridge and elastic net with an accuracy of 0.6419 and 0.6417 respectively. LASSO also has the highest values of sensitivity, positive predictive value, negative predictive value and F1, whereas ridge has the highest value of specificity. Detailed values of the performance metric of the three models can be seen in Table 7.

After having a detailed investigation about the prevalence of anaemia through bivariate analysis and predicting it with acceptable levels of accuracy with the help of penalized regression models using ML techniques, we now turn to a brief discussion on major outcomes. Figure 9 gives VIP plots of the top 25 variables for ridge, LASSO and elastic net respectively.

The results also suggest that gender, wealth index, and place of residence are not the most important variables for the prevalence of anaemia which is in contrast to that of Dey S et al. (2013). The VIP plots (Fig. 9) across the models reveal that the variables such as mother anaemic status, age of the child, social status, mother's age, mother's education, and religion are important factors in predicting the prevalence of anaemia. However, there is a little variation in the degree of their respective importance across the models. For example, mother's anaemic status of moderate grade is the most important predictor in elastic and ridge models whereas in the case of LASSO, age of the child below 2 years is the most important one.

Now, as far as the social status is concerned, it is the ST category that dominates the prediction. As for maternal age, the mothers in the age group of 15–19 years are identified as the most important factor. In the mother's education category, mothers with low education contribute more to the prediction than those with high education. This corroborates the findings of NFHS-4 reports and other studies (Dey S et al. 2013). In religious categories, it is the non-Hindu and non-Muslim, levelled as others (Table 1) appears to be prominent across the models. This could be because of the fact that a large proportion of the population belongs to the ST category and many of them do not follow either Hindu or Muslim religion. Apart from important variables, the models have identified a group of variables that moderately contribute to the prediction of anaemia, viz. disposal of youngest child's stool, access to safe drinking water and wealth index.

Finally, the remaining variables like gender of the child, size of the child, number of living children, breastfeeding, household size, sanitation facility, and place of residence are identified by the models as least or unimportant variables for predicting anaemia.

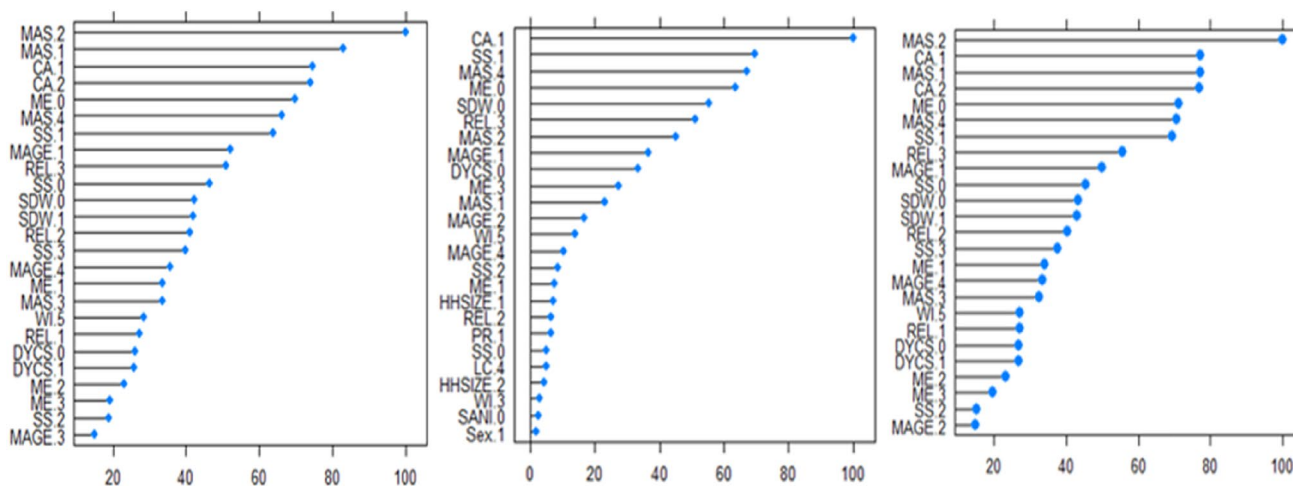
## 5 Conclusion

We analysed NFHS-4 data in context of North-East India through penalized regression with three different models, namely, ridge, LASSO and elastic net. Of the 3 models, LASSO is giving the best results but the difference is negligible. We have achieved a ROC value of above 70% with training data and accuracy of above 64% with testing data, which is a reasonably acceptable outcome when working with the survey data. This study demonstrates the efficiency of ML algorithms in analysing and drawing inferences from demographic data. The major finding suggests that the prevalence of anaemia depends on various factors such as mother's anaemic status, age of child, social status, mother's age, religion, etc. which are important in predicting the prevalence of anaemia. Hence, the aforesaid factors should be taken into consideration in designing any affirmative action program in controlling anaemia among children (6–59 months).

Certainly, there are some limitations of this study that need to be specified to bring the complete aspects of accuracy. Firstly, the data of the study, which is based on NFHS-4 (2015–16), can have limitations in the consistency of the survey questionnaire as per demographic requirements. Responses to the questionnaire may not be correct or truthful, because of the number of missing entries in the data, and the chances of biasness by respondents or interviewers. Secondly, it could be due to the lack of the availability of related literature. As per our knowledge, there is scant literature on the application of machine learning techniques in general

**Table 7** Predictive performance of the three classification methods. Source: Authors' Estimates

Metric	Elastic	Ridge	LASSO
Kappa	0.2832	0.2837	0.2856
95% CI	(0.254, 0.312)	(0.257, 0.313)	(0.257, 0.315)
Accuracy	0.6417	0.6419	0.6429
95% CI	(0.627, 0.6562)	(0.6272, 0.6564)	(0.6282, 0.6574)
Sensitivity	0.6190	0.6190	0.6224
95% CI	(0.598, 0.64)	(0.598, 0.64)	(0.601, 0.643)
Specificity	0.6641	0.6646	0.6632
95% CI	(0.644, 0.684)	(0.644, 0.685)	(0.643, 0.683)
Positive predictive value	0.6465	0.6469	0.6471
95% CI	(0.625, 0.667)	(0.62.5, 0.668)	(0.626, 0.668)
Negative predictive value	0.6372	0.6374	0.6389
95% CI	(0.617, 0.657)	(0.617, 0.658)	(0.618, 0.659)
F1 value	0.6325	0.6326	0.6345



**Fig. 9** VIP plots of ridge, LASSO and elastic net

and penalized logistic regression in particular for predicting anaemic children using demographic data. Thirdly, there might be some important variables that are missing in the analysis due to the unavailability of appropriate data sets. However, a detailed investigation is necessary for specific socio-ethnic communities who are more prone to anaemia using different datasets.

Future research should focus on applying alternate ML techniques and using different data sets in predicting child anaemia and assess their relative efficacies. As far as different data sets are concerned, predicting child anaemia using medical image processing data could be another potential research direction that can be explored. Further, since ML algorithms are capable of identifying trends and patterns easily, the future research can look at applying those in predicting other disease such as heart disease, lung cancer, etc. Nevertheless, in spite of its limited scope, the present study aims to draw the attention of the Indian policy makers

towards the various socio-economic factors in the fight against the child anaemia.

**Acknowledgements** The authors are thankful to editor and anonymous reviewers for their valuable suggestions and comments. Authors also acknowledge all the respondents for their active participation in nationally representative survey, NHFS-4 (2015–16).

**Funding** The authors did not receive any kind of fund or financial support to conduct the study.

**Declarations**

**Conflict of interest** There is no conflicting interest among the authors.

## References

- World Health Organization (WHO). Anaemia in women and children (2021) [https://www.who.int/data/gho/data/themes/topics/anaemia\\_in\\_women\\_and\\_children](https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children)
- Abram SV, Helwig NE, Moodie CA, DeYoung Colin G, MacDonald AW, Waller NG (2016) Bootstrap enhanced penalized regression for variable selection with neuroimaging data. *Front Neurosci*. <https://doi.org/10.3389/fnins.2016.00344>
- Agaoglu L, Torun O, Unuvar E, Sefil Y, Demir D (2007) Effects of iron deficiency anaemia on cognitive function in children. *Arzneimittelforschung* 57(6A):426–430
- Avinash R, Ramakrishnan S, Krishnegowda R (2021) Anaemia predicts poor outcomes of COVID-19 in hospitalized patients: a prospective study in a tertiary care hospital from South India. *Asian J Med Sci* 12:5–8. <https://doi.org/10.3126/ajms.v12i8.37795>
- Balarajan Y, Ramakrishnan U, Ozaltin E, Shankar AH, Subramanian SV (2011) Anaemia in low-income and middle-income countries. *Lancet* 378:2123–35. [https://doi.org/10.1016/S0140-6736\(10\)62304-5](https://doi.org/10.1016/S0140-6736(10)62304-5)
- Bezboruah G, Narayan Dev C, Chakraborty A et al (2021) Anemia in HIV-positive children in a tertiary care center in North-East India: prevalence and risk factors. *Indian J Pediatr* 88:952. <https://doi.org/10.1007/s12098-021-03847-w>
- Bitew FH, Sparks CS, Nyarko SH (2022) Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public Health Nutr Open Access* 25(2):269–2808. <https://doi.org/10.1017/S1368980021004262>
- Chaparro CM, Suchdev PS (2019) Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries. *Ann NY Acad Sci* 1450(1):15–31. <https://doi.org/10.1111/nyas.14092>
- Cho H, Lee S-R, Baek Y (2021) Anemia diagnostic system based on impedance measurement of red blood cells. *Sensors* 21(23):8043. <https://doi.org/10.3390/s21238043>
- De M, Halder A, Podder S, Sen R, Chakraborty S, Sengupta B, Chakraborty T, Das U, Talukder G (2006) Anemia and hemoglobinopathies in tribal population of Eastern and North-Eastern India. *Hematology* 11(5):371–3. <https://doi.org/10.1080/10245330600840180> (PMID: 17607589)
- Dey S, Goswami S, Dey T (2013) Identifying predictors of childhood anaemia in North-East India. *J Health Popul Nutr* 31(4):462–70. <https://doi.org/10.3329/jhpn.v31i4.20001> (PMID: 24592587; PMID: PMC3905640)
- Donepudi PK (2017) AI and machine learning in banking: a systematic literature review. *Asian J Appl Sci Eng* 6(3):157–162
- Dubey A (1994) Iron deficiency anaemia: epidemiology, diagnosis and clinical profile. In: Sachdev H, Choudhury P (eds) *Nutrition in children: developing country concerns*. B.I. Publications, New Delhi, pp 217–235
- Dukhi N, Sewpaul R, Sekgala MD, Awe OO (2021) Artificial intelligence approach for analyzing anaemia prevalence in children and adolescents in brics countries: a review (open access). *Curr Res Nutr Food Sci* 9(1):1–10. <https://doi.org/10.12944/CRNFSJ.9.1.01>
- Faghhi Dinevari M, Somi MH, Sadeghi Majd E, Abbasalizad Farhangi M, Nikniaz Z (2021) Anaemia predicts poor outcomes of COVID-19 in hospitalized patients: a prospective study in Iran. *BMC Infect Dis* 21(1):170. <https://doi.org/10.1186/s12879-021-05868-4> (PMID: 33568084; PMID: PMC7875447)
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22 (PMID: 20808728; PMID: PMC2929880)
- Gastón A, García-Viñas JI (2011) Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models. *Ecol Model* 222(13):2037–2041. <https://doi.org/10.1016/j.ecolmodel.2011.04.015>
- Gholamy A, Kreinovich V, Kosheleva O (2018) "Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation". Departmental Technical Reports (CS). 1209. [https://scholarworks.utep.edu/cs\\_techrep/1209](https://scholarworks.utep.edu/cs_techrep/1209)
- Greenwood CJ, Youssef GJ, Letcher P, Macdonald JA, Hagg LJ, Sanson A, Mcintosh J, Hutchinson DM, Toumbourou JW, Fuller-Tyszkiewicz M, Olsson CA (2020) A comparison of penalised regression methods for informing the selection of predictive markers. *PLoS ONE* 15(11):e0242730. <https://doi.org/10.1371/journal.pone.0242730>
- Gutema B, Adissu W, Asress Y, Gedefaw L (2014) Anaemia and associated factors among school-age children in Filtu town, Somali region. *Southeast Ethiop BMC Hematol* 14(1):13
- Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- Jain A, Rodgers J, Kim R (2021) The relative importance of households as a source of variation in child malnutrition: a multilevel analysis in India. *Int J Equity Health* 20:225. <https://doi.org/10.1186/s12939-021-01563-7>
- Katzman R, Novack A, Pearson H (1972) Nutritional anaemia in an inner-city community. Relationship to age and ethnic group. *J Am Med Assoc* 222:670–673
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28(5):1–26. <https://doi.org/10.18637/jss.v028.i05>
- Lokeshwar MR, Mehta M, Mehta N, Shelke P, Babar N (2011) Prevention of iron deficiency anaemia (IDA): how far have we reached? *Indian J Pediatr* 78:593–602
- Meena K, Tayal DK, Gupta V, Fatima A (2019) Using classification techniques for statistical analysis of anemia. *Artif Intell Med* 94:138–152. <https://doi.org/10.1016/j.artmed.2019.02.005> (PMID: 30871679)
- Meshram II, Kumar BN, Venkaiah K, Longvah T (2020) Subclinical vitamin A deficiency and anaemia among women and pre-school children from North-East India. *Indian J Community Med* 45:371–374
- Pourahmadi M (2013). High-dimensional covariance estimation: with high-dimensional data. Vol 882. John Wiley & Sons. <https://doi.org/10.1002/9781118573617>
- Qusay S, Emrullah S (2022) The efficiency of classification techniques in predicting anemia among children: a comparative study. *Commun Comput Inf Sci Emerg Technol Trends Internet Things Comput*. [https://doi.org/10.1007/978-3-030-97255-4\\_12](https://doi.org/10.1007/978-3-030-97255-4_12)
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soci Ser B (methodological)*. 58(1): 267–88. JSTOR 2346178
- Shah R, Agarwal AK (2013) Anaemia associated with chronic heart failure: current concepts. *Clin Interv Aging* 8:111–122. <https://doi.org/10.2147/CIA.S27105>
- Stevens GA, Finucane MM, De-Regil LM, Paciorek CJ, Flaxman SR, Branca F et al (2013) Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: a systematic analysis of population-representative data. *Lancet Glob Health* 1:E16–E25. [https://doi.org/10.1016/S2214-109X\(13\)70001-9](https://doi.org/10.1016/S2214-109X(13)70001-9)
- Stoltzfus RJ (2004) Iron deficiency: global prevalence and consequences. *Food Nutr Bull* 24(4\_suppl2):S99–S103. <https://doi.org/10.1177/15648265030244S206>
- Sun J, Han Wu, Zhao M, Magnussen CG, Xi Bo (2021) Prevalence and changes of anaemia among young children and women in 47 low- and middle-income countries, 2000–2018. *E Clin Med* 41:101136
- Talukder A, Ahammed B (2020) Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *J Nutr* 78:110861. <https://doi.org/10.1016/j.nut.2020.110861>



- Tolentino K, Friedman JF (2007) An update on anaemia in less developed countries. *Am J Trop Med Hyg* 77:44–51
- Wallner C, Hurst J, Behr B, Rony MAT, Barabás A, Smith G (2022) Fanconi anemia: examining guidelines for testing all patients with hand anomalies using a machine learning approach. *Children* 9(1):85. <https://doi.org/10.3390/children9010085>
- World Bank (2021) The World Bank In Yemen. Retrieved February 21, 2022, <https://www.worldbank.org/en/country/yemen/overview#:~:text=The%20UN%20estimated%20that%2024.1,more%20people%20into%20extreme%20poverty>
- World Health Organization (WHO) (2015) The global prevalence of anaemia in 2011. World Health Organization, Geneva, Switzerland
- Zhang Q, Ananth CV, Li Z, Smulian JC (2009) Maternal anaemia and preterm birth: a prospective cohort study. *Int J Epidemiol* 38(5):1380–1389. <https://doi.org/10.1093/ije/dyp243>
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67:301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.