Journal of Data Science 0 (0), 1–6 DOI: 0000
2025

Supplemental Material

This document holds supplemental material including variable summaries, correlations, and additional results that were explored.

1 Data Descriptions

Table S1: Categorical variables' counts and percentages.

Variable	Count (%)	Variable	Count (%)
Rotarix Vaccine	203 (51.92)	No Maternal Education	127 (32.48)
Male	196 (50.13)	Treated Water	241 (61.64)
Septic Tank/Toilet	212 (54.22)	No Open Drain Outside Home	62 (15.86)
No Shared Toilet	234 (59.85)	CMV Positive	327 (83.63)
IL-1b at week 18:		IL-4 at week 18:	
Lower 50th	226 (57.80)	Lower 50th	186 (47.57)
50th to 75th	78 (19.95)	50th to 75th	98 (25.06)
Upper 25th	87 (22.25)	Upper 25th	107 (27.37)
IL-5 at week 18:		IL-6 at week 18:	
Lower 50th	165 (42.20)	Lower 50th	193 (49.36)
50th to 75th	127 (32.48)	50th to 75th	95 (24.30)
Upper 25th	99 (25.32)	Upper 25th	103 (26.34)
IL-7 at week 18:		IL-10 at week 18:	
Lower 50th	204 (52.17)	Lower 50th	225 (57.54)
50th to 75th	96 (24.55)	50th to 75th	85 (21.74)
Upper 25th	91 (23.27)	Upper 25th	81 (20.72)
MIP1b at week 18:		TNFA at week 18:	
Lower 50th	221 (56.52)	Lower 50th	174 (44.50)
50th to 75th	92 (23.53)	50th to 75th	113 (28.90)
Upper 25th	78 (19.95)	Upper 25th	104 (26.60)

Table S2: Quantitative variables' mean ± standard deviation.

Variable	Mean ± SD
HAZ at 1 year	-1.42 ± 1.03
Days Exclusive Breastfeeding at 1 year	124.91 ± 57.76
Days of Diarrhea at 1 year	7.01 ± 4.93
Expenditure	11980.87 ± 7767.52
Income	13295.58 ± 9999.41
Maternal Weight at Enrollment (kg)	49.71 ± 9.35
Maternal Height at Enrollment (cm)	150.49 ± 5.73
Vitamin D at week 6	25480.26 ± 11873.14
Vitamin D at week 18	31238.75 ± 15410.77
Zinc at week 6	722.94 ± 104.74
Zinc at week 18	772.15 ± 144.54
Mannitol Recovery at week 12	0.02 ± 0.02
Mannitol Recovery at week 24	0.02 ± 0.02
Anti-LPS Ab at week 6	28.52 ± 43.82
Anti-LPS Ab at week 18	13.07 ± 44.69
RBP at week 6	25480.26 ± 11873.14
RBP at week 18	31238.75 ± 15410.77
Ferritin at week 6	227.19 ± 160.58
Ferritin at week 18	41.87 ± 36.81
sCD14 at week 6	1666.07 ± 647.27
sCD14 at week 18	1996.48 ± 729.75
Reg 1b week 6	55.89 ± 91.77
Reg 1b week 12	74.25 ± 105.63
Myeloperoxidase at week 12	11706.77 ± 11734.65
Alpha-1 Antitrypsin at week 12	0.91 ± 0.74
Calprotectin at week 12	750.32 ± 699.39
Neopterin at week 12	2691.80 ± 2079.14
Activin at week 6	1094.29 ± 1473.44
CRP Index at 1 year	2.00 ± 1.23

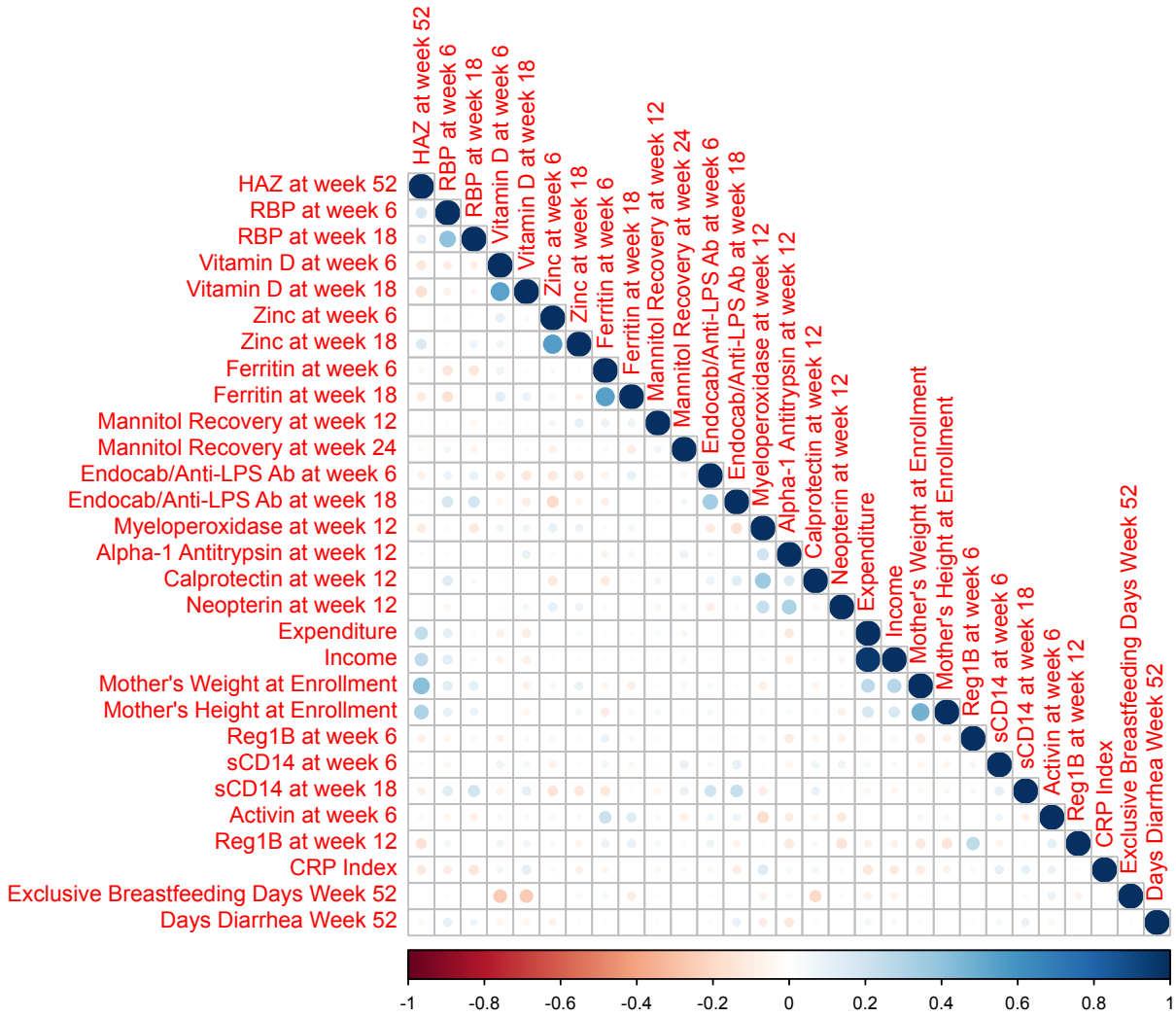

Figure S1: Spearman correlations for all continuous variables and subjects included for the analysis.

2 Methods Summary

Table S3: Summary of machine learning algorithms used. Note that other R packages may be available for the listed methods.

Method	R Package	Tuning Parameter
	Method Description	
Generalized Linear Model (GLM)	`stats`	None
	Generalized linear model with stepwise selection where AIC is the selection criteria.	
Bagged Trees (Bagg)	`ipred`	None
	Tree based method, which uses bootstrapping to aggregate trees. Each tree is built to the maximal depth. All predictors are used to find the best predictor at each split for each tree. Trees are not independent of each other.	
Random Forests (RF)	`randomForest`	mtry: Number of randomly selected predictors
	Tree based method which uses bootstrapping to aggregated trees but selects a random subset of predictors to use for split on each tree. Trees are independent of each other and each tree is built to the maximal depth.	
Conditional Random Forest (CRF)	`party`	mtry: Number of randomly selected predictors
	Extension of RF but takes into account the correlations between predictor variables to properly assess variable importance.	
Stochastic Gradient Boosting (GBM)	`gbm`	n.trees: Number of boosting iterations interaction.depth: Max tree depth shrinkage: Shrinkage/regularization n.minobsinnode: Min terminal node size
	Tree based method which uses random sampling to train a minial depth tree on the random subset using residuals (gradient) from the previous iteration as the outcome. The shrinkage parameter controls the parameter estimates and in turn helps reduce variance.	
Bayesian Additive Regression Trees (BART)	`bartMachine`	num_trees: Number of trees k: Prior boundary alpha: Base terminal node hyperparameter beta: Power terminal node hyperparameter nu: Degrees of freedom
	Tree based method where trees are added together for a sum of trees model which is defined by a prior (used for model regularization) and a likelihood in which MCMC is applied to gain predictions. For classification (a categorical response), the nu tuning parameter is no longer applicable.	
Neural Network (NNet)	`nnet`	size: Number of hidden units decay: Weight decay
	Nonlinear transformation is used on the combination of hidden units (which are linear combinations of the original predictors) to predict the outcome where weight decay controls the classification boundaries. This method has a tendency to overfit the data.	
Model Averaged NNet (MANN)	`nnet`	size: Number of hidden units decay: Weight decay bag: Indicator for bagging to be used
	Averages across many neural networks to obtain a model which reduces overfitting and creates more stable predictions than a single NNet.	

3 NNLS Optimization Results for HAZ

Here, the NNLS optimization results are presented for comparison to the Rank Loss optimization results in the main text. The NNLS optimization may be used for either a categorical or continuous response, however it is typically recommended to be used in regression settings. Hence, NNLS is explored here with HAZ at two years of age as the continuous response. For HAZ, the evaluation measures include root mean squared error (RMSE) and the correlation between the predicted values and the observed response values.

The same algorithms were used in the SL library for the continuous response as for the binary response with the needed adjustments to update the algorithms. Overall, only GLM was used in all 10 CV folds while Bagg and CRF were used in seven folds, BART in four, RF and GBM in three, MANN in two, and NNet in zero folds. The average coefficients for GLM was 0.577, RF 0.112, CRF 0.107, and MANN 0.106 while the average coefficients are below 0.1 for other algorithms as shown in Table S4. This tells us that even though RF was only used three times and MANN twice, when they were used, they carried a heavy weight. GLM was obviously the most used algorithm by the coefficients (ranging 0.408-0.814 except for one fold where the weight was 0.0002).

Looking at the discrete SL confirms that GLM was the method that could be chosen. Thus, with the high coefficients on GLM and the discrete SL being the GLM 9 out of 10 CV folds, if we wanted to pick a single model, GLM would be the one to choose.

Table S4: Summary of weight/coefficient values over all 10 CV folds for the algorithms in the SL library for the NNLS optimization methods of the height for age z-score (HAZ) at two years of age.

CV Fold	GLM	Bagg	RF	CRF	GBM	BART	NNet	MANN
1	0.592	0	0	0.146	0.058	0.148	0	0.055
2	0.408	0	0.508	0.060	0	0.023	0	0
3	0.711	0.021	0.268	0	0	0	0	0
4	0.535	0.041	0	0.425	0	0	0	0
5	0.573	0.008	0.339	0	0.080	0	0	0
6	0.548	0.227	0	0.044	0	0.182	0	0
7	0.0002	0.000002	0	0.00001	0.00005	0	0	0.9997
8	0.814	0.008	0	0.177	0	0	0	0
9	0.805	0.135	0	0	0	0.060	0	0
10	0.780	0	0	0.220	0	0	0	0
Mean	0.577	0.044	0.112	0.107	0.014	0.041	0	0.106
SD	0.243	0.076	0.189	0.138	0.030	0.068	0	0.315
Number Nonzero	10	7	3	7	3	4	0	2

GLM had the lowest RMSE with other algorithms' RMSE values just a bit higher excluding NNet and MANN which had the highest RMSE. The reason NNet and MANN had the same values was because the predictions from these two across the CV folds were always the same, a predicted value of zero. With the average Pearson correlation between the predicted values and the observed values measure, the GLM and SL had the highest, however, when considering the standard deviations, all these algorithms were compatible except for MANN and NNet which

had all the predictions the same and thus the correlations were not calculated. Across all the evaluations, GLM performed the best and could again be the one algorithm that is chosen.

Table S5: Evaluation measures for the NNLS optimization with a continuous response, HAZ at two years, averaged across CV folds for each algorithm and the SL ordered based on their average RMSE.

Algorithm	RMSE Mean ± SD	Correlation Mean ± SD
GLM	0.489±0.052	0.877±0.023
CRF	0.503±0.074	0.871±0.033
Bagg	0.508±0.069	0.868±0.025
RF	0.508±0.072	0.866±0.035
BART	0.516±0.068	0.858±0.039
GBM	0.527±0.073	0.859±0.026
SL	0.667±0.559	0.877±0.024
MANN	1.841±0.286	NA
NNet	1.841±0.286	NA

Since the consensus is that GLM would be the single model used, the variables selected can be explored. In a separate run of GLM predicting HAZ at two years of age (y), HAZ at one year (x_1), mother's weight (x_2), IL-7 at week 18 (x_3 for 50th to 75th percentile and x_4 for upper 25th percentile), Days Excl. BF (x_5), CMV result (x_6), zinc at week 18 (x_7), and CRP index (x_8) were selected in that order leading to the model

$$y = -1.710 + 0.908x_1 + 0.067x_2 + 0.140x_3 + 0.122x_4 - 0.055x_5 + 0.129x_6 - 0.045x_7 - 0.041x_8$$

.

As we can see, Days Excl BF, zinc, and CRP index individually have negative effects while HAZ at one year, mother's weight, IL-7, and CMV result have positive effects on HAZ at two years of age while considering all other variables in the model. Most of these directions make sense, however, Days Excl. BF, IL-7, zinc, and CMV result seem to have the opposite effect. Potentially, interactions need to be explored between the variables included or we have an example of a paradox as multicollinearity does not seem to be the issue (low VIF and no high correlations between these predictors).

The assumptions were checked and the residuals are approximately normal with constant variance. The RMSE was 0.459 with a correlation between the predicted values and HAZ at two years of 0.899. Additionally, 80.4% of the variation in HAZ at two years was explained. Overall, this is a simple yet informative model yielding informative interpretations.