BST-432 Final Project
Predict Babies' Respiratory Illness

Zhirou Zhou

Dec 13, 2019

Contents

1 Data Description

There are 166 infants involved in the study.

Different modules of data regarding demographics, follow up surveys, virus and bacteria testing, vaccination records, etc. are summarized in 14 tables, nine of which are longitudinal data identified by unique subject identifier.

Classifiers are developed to predict `is_illness` and `precedes_illness` while the remaining tables could be used as features.

2 Method

2.1 Preprocessing

2.1.1 Variable Transformation

All 14 tables, except for immune profiling `flowcytometry` are reformed as new tables named by `_new`.

Some of the numeric variables, such as the birth weight, length and head circumference are scaled by $Z(x|\mathcal{T}) = (x - \bar{x}/\hat{\sigma})$. All the binary variables are set as dummy variables in 0/1 or `TRUE/FALSE`. And for some of the sequential questions in surveys, the results are combined in to a single numeric variable, such as the `smoker` and `pets` in table `fup2_new`.

Table `tlda` is separated by the name of target virus or bacteria and the values `TRUE/FALSE` stands for positive/negative testing results.

Table `vaccines` is separated by the key of vaccination. The features are set to 0/1, which means whether the baby had that dose of vaccine at specific pCGA (staying at 1 after vaccinated).

The results of transformation are summarized in Table 1, containing the description and dimension of each dataset. And the declaration of each variable is listed in Table 10 in Appendix.

All the tables except for `flowcytometry` are joined as one, by identifier `Alias`, `visit_id` and `pCGA` in target table `illness_control`, which could identify the unique subject with certain time for visit in certain corrected gestational age. None of the record in `illness_control` is removed, so there are 2339 observations in the training data.

2.1.2 Missing Values

For variables without longitudinal timepoint, the missing values are filled using means (a value between 0 and 1 for dummy variables).

While for longitudinal data, the missing values are filled by 'last observation carried forward/backward', grouping by `Alias`.

Table 1: Summary of Datasets

Table	Description	Example Features	#Observations	#Features
illness_new	Target variables and time series info	is_illness, precedes_illness	2339	5
base_new	Birth medical history	birth.weights, chest.compression	166	19
demo_new	Birth demographics	gender, birth.season	166	4
fam_new	Family demographics	education	166	2
oxy_new	Oxygen exposure at birth	auc	166	2
preg_new	Pregnancy medical history	asthma, alcohol	166	20
fup1_new	Follow up survey 1	receive.breast.milk, non.milk.foods	635	4
fup2_new	Follow up survey 2	exposed.to.smoke, pets	316	6
nas_new	Nasal microbiome	nas_microbiome	1242	4
rec_new	Rectal microbiome	rec_microbiome	1481	4
thr_new	Throat microbiome	thr_microbiome	334	4
tlda_new	Virus and bacteria testing	Corona 1, Flu A	1846	25
vaccines_new	Vaccination record	DTaP (2 months), HepB (4th dose)	719	28

As for the `vaccines` data, the missing values are set by most frequent category grouping by `pCGA` (whether most of other babies with same pCGA received that dose of vaccine).

2.1.3 Feature Selection

The correlation between numeric variables are calculated to check the redundance and the results are shown in Table 2. Since the first four features are highly correlated with each other, the first three of them are removed from the dataset.

For some of the categorical variables, their infrequent classes are collapsed. For example, in Table 3, the last five category of `Pets` are combined as one.

Table 2: Correlation between Numeric Variables

	Weight	Head Circumference	Length	Age	Temp	BMI
Weight	1	0.938	0.947	0.910	0.040	-0.025
Head Circumference		1	0.944	0.920	0.057	-0.036
Length			1	0.927	0.048	-0.068
Age				1	0.041	-0.035
Temp					1	-0.013
BMI						1

Table 3: Example of Combining Infrequent Classes

#Pets	#Obs
0	2020
1	155
2	97
3	43
4	11
5	2
6	2
8	2
10	6
28	1

#Pets	#Obs
0	2020
1	155
2	97
3	43
4	11
5	13

2.2 Predictor A: Decision Tree

2.2.1 Model Description

Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. A popular method for tree-based regression and classification is called CART.

Consider a problem with continuous response Y and inputs X_1 and X_2 as an example. The left panel of Figure 1 shows that the two-dimensional feature space is partitioned by recursive binary splitting defined by different values of X_1 and X_2, as used in CART, while the right panel shows the corresponding tree. The terminal nodes or leaves of the tree correspond to the regions $R_1, R_2, ..., R_5$.

Suppose the data consists of p inputs and one response, for each of N observations: (x_i, y_i) for $i = 1, 2, ..., N$, with $x_i = (x_{i1}, x_{i2}..., x_{ip})$. Suppose we have a partition into M regions $R_1, R_2, ..., R_M$, and we model the response as a constant c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \tag{1}$$

Define $N_m = \#\{x_i \in R_m\}$, for regression trees,

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i \tag{2}$$

and for classification trees,

$$\hat{c}_m = argmax_k \hat{p}_{mk} = argmax_k \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \tag{3}$$

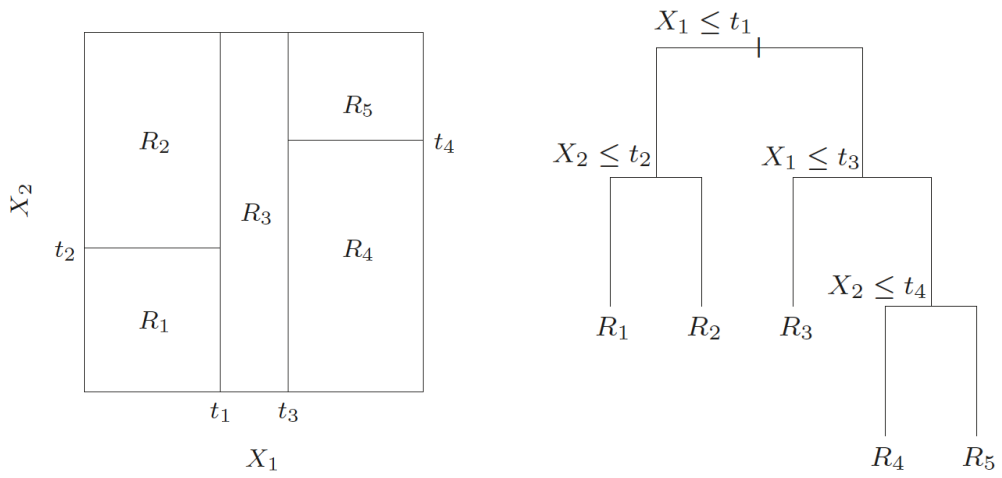

Figure 1: CART Partition and Tree

Starting with all of the data, consider a splitting variable j and split point s, define the pair of half-planes

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j > s\}. \tag{4}$$

We need to find out the splitting variable j and split point s that minimize the criterion. For regression trees, we usually adopt sum of squares

$$R_1(j,s) = X|X_j \leq s \text{ and } R_2(j,s) = X|X_j > s \tag{5}$$

where $Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$ is defined as a measure of node impurity.

While for classification trees, different measures $Q_m(T)$ of node impurity include the following:

$$\text{Misclassification error} : \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) \tag{6}$$

$$\text{Gini index} : \sum_{k \neq k`} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{7}$$

$$\text{Cross-entropy or deviance} : \sum_{k=1}^{k} \hat{p}_{mk} log \hat{p}_{mk} \tag{8}$$

2.2.2 Training Algorithm

Under cross validation, grow a large tree T_0, stopping the splitting process only when some minimum node size (30) is reached. Then this large tree is pruned using *cost-complexity pruning*: define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0 and let

$|T|$ denote the number of terminal nodes in T. The cost complexity criterion will be

$$C_\alpha(T) = \sum_{m=1}^{t} N_m Q_m(T) + \alpha|T| \tag{9}$$

The estimation of tuning parameter $\alpha \geq 0$ is chosen by the value $\hat{\alpha}$ that minimize the cross-validated sum of squares. The final tree is the subtree $T_{\hat{\alpha}}$ that minimize $C_{\hat{\alpha}(T)}$.

2.3 Predictor B: Random Forest

2.3.1 Model Description

Random forests is a substantial modification of bagging that builds a large collection of *de-correlated* trees, which is achieved in the tree-growing process through random selection of the input variables, and then averages them.

2.3.2 Training Algorithm

For $b = 1$ to B:

- Draw a bootstrap sample $\boldsymbol{Z}$ of size N from the training data
- Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached
  - Select m variables at random from the p variables
  - Pick the best variable/split-point among the m
  - Split the node into two child nodes
- Output the ensemble of trees $\{T_b\}_1^B$
- Classification prediction: $\hat{C}_b(x)$ = the prediction of the bth random-forest tree, then

$$\hat{C}_{rf}^B(x) = majority\, vote\{\hat{C}_b(x)\}_1^B$$

2.4 Validation

Both of the predictors are validated using 10-fold cross validation, block by `Alias`. The training data are also over-sampled to balance the target variables, either `is_illness` or `precedes_illness`.

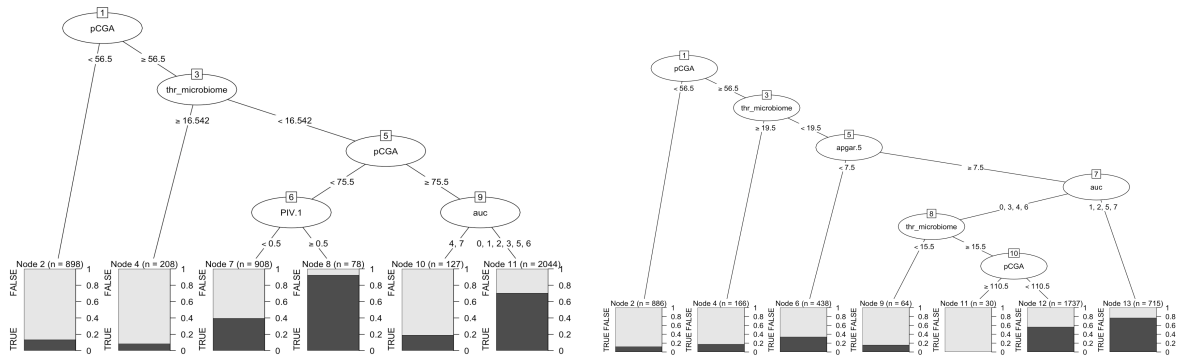

Figure 2: Pruned Classification Trees for is_illness and precedes_illness

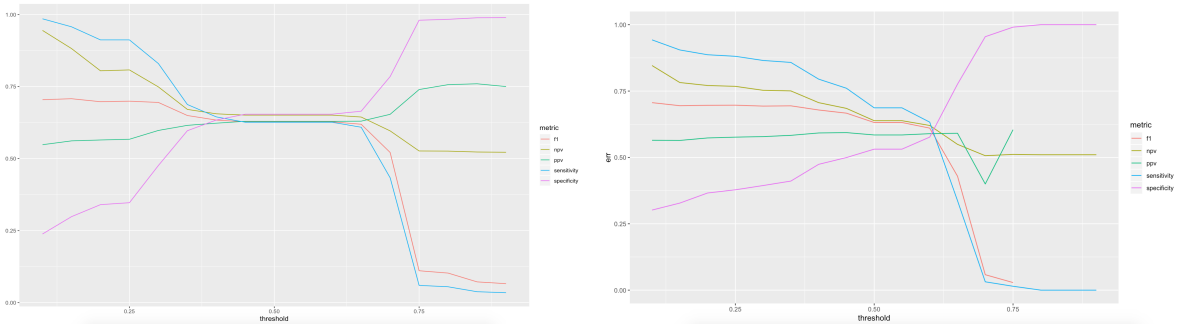

Figure 3: Comparison of Different threshold under Classification Trees

3 Results

3.1 Predictor A: Classification Tree

The training data is preprocessed and stored in `data_illness`, including 2339 observations and 97 features (excluding `Alias`, `visit_id`, `is_illness` and `precedes_illness`).

Two tuning parameters, minimum node size and α are tuned by 10-fold cross validation, blocked by `Alias`. The results are 50 and 0.03 for predicting `is_illness`, 30 and 0.005 for `precedes_illness`. Plots for the two classification trees are shown in Figure 2.

The default threshold for predictions in each terminal node is 0.5. Figure 3 shows the comparison of using different threshold. The value of 0.5 and 0.65 seems reasonable for the two trees.

The results of prediction under cross validation are shown in Table 4, Table 5 and Figure 4. And the performance of two trees are summarized in Table 6. The comparison of two trees shows that predicting `is_illness` is easier than predicting `precedes_illness`.

3.2 Predictor B: Random Forest

Two tuning parameters, minimum node size and the number of features that are randomly chosen when growing each tree are tuned by 10-fold cross validation, blocked by

Table 4: Prediction of is_illness using Classification Tree

	is_illness	
Prediction	FALSE	TRUE
FALSE	1458	782
TRUE	771	1309

Table 5: Prediction of precedes_illness using Classification Tree

	precedes_illness	
Prediction	FALSE	TRUE
FALSE	2081	1179
TRUE	1837	2585

Table 6: Performance of Classification Trees

	is_illness	precedes_illness
Accuracy	0.6405	0.6074
PPV	0.6293	0.5846
NPV	0.6509	0.6383
F1	0.6277	0.6316
AUC	0.7034	0.6435

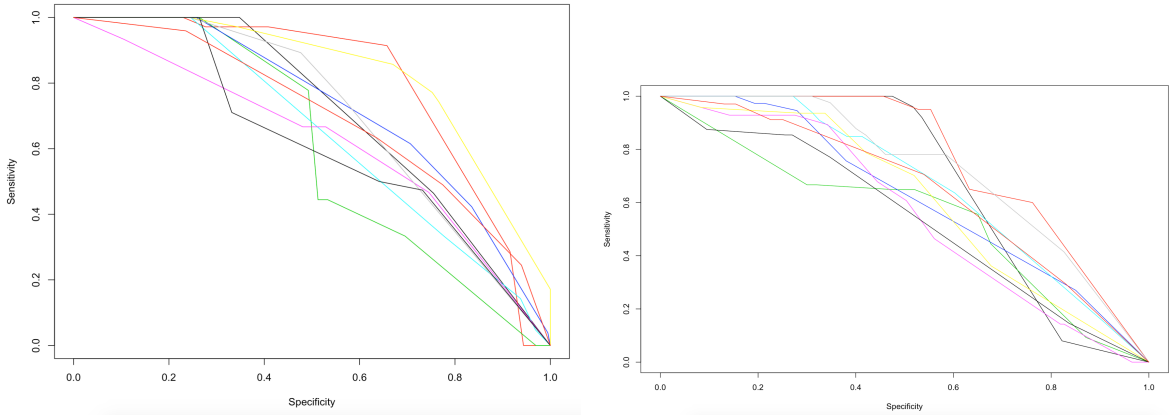

Figure 4: ROC Curves of 10-fold cross validation under Classification Trees

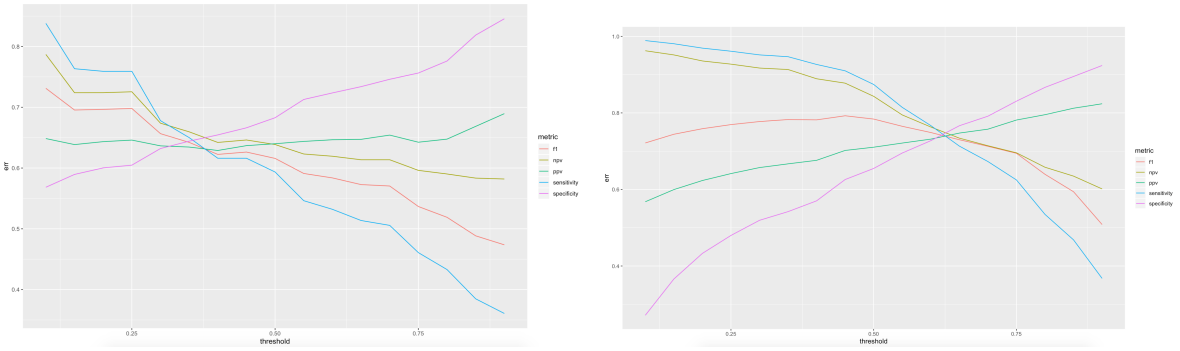

Figure 5: Comparison for Different threshold under Random Forest

Table 7: Prediction of is_illness using Random Forest

	is_illness	
Prediction	FALSE	TRUE
FALSE	1609	828
TRUE	620	1280

`Alias`. The results are 100 and 1000 for predicting `is_illness`, 20 and 100 for predicting `precedes_illness`.

Figure 5 shows the comparison of using different threshold. The value of 0.375 and 0.625 seems reasonable for the two random forests.

The results of prediction under cross validation are shown in Table 7, Table 8 and Figure 6. The performance of all classifiers are summarized in Table 9. The comparison shows that random forest does a greater job when predicting `precedes_illness`, and the overall performance of random forest is better than classification trees.

4 Discussion

Since decision tree is a method of very high variance, as the sample seeds changed when doing cross validation, the result will be quite different from each other, as shown in figure 7. While random forest can reduce the variance by increasing the number of trees, so that make the prediction more stable.

Table 8: Prediction of precedes_illness using Random Forest

	precedes_illness	
Prediction	FALSE	TRUE
FALSE	2600	477
TRUE	1370	3367

Table 9: Performance of All Classifiers

	Tree: is_illness	Tree: precedes_illness	Forest: is_illness	Forest: precedes_illness
Accuracy	0.6405	0.6074	0.6661	0.7636
PPV	0.6293	0.5846	0.6737	0.7108
NPV	0.6509	0.6383	0.6602	0.8450
F1	0.6277	0.6316	0.6387	0.7848
AUC	0.7034	0.6435	0.7352	0.8335

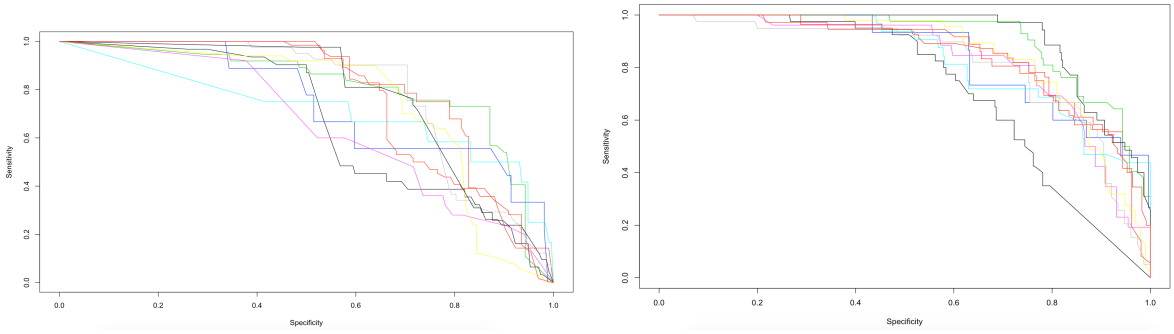

Figure 6: ROC Curve of 10-fold cross validation under Random Forest

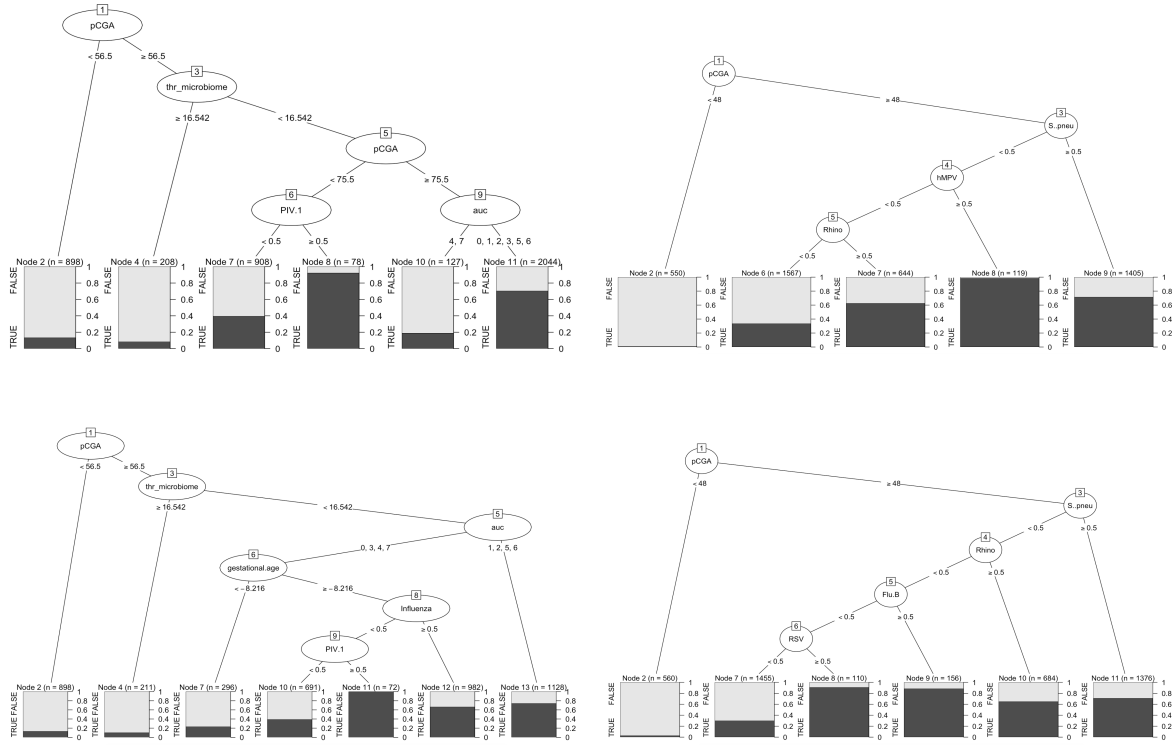

Figure 7: Variation of Decision Trees

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*, 3nd Edition. United States: Springer

[2] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. United States: MIT Press.

Appendix

Table 10: Declaration of Variables

Table	Name of Features	Feature Declaration
illness_new	Alias	Subject identifier
	visit_id	Longitudinal record for each subject
	pCGA	Corrected gestational age
	is_illness	TRUE / FALSE
	precedes_illness	TRUE / FALSE
base_new	birth.weights	Scaled birth weight
	head.circumference	Scaled head circumference
	birth.lengths	Scaled birth length
	temp	Scaled temperature at first NICU admission
	apgar.1	APGAR at 1 min
	apgar.5	APGAR at 5 min
	multiple	0 for single birth, >0 for birth order/total # of birth
	birth.location	0 for outside study center, 1 for inside study center
	stablization	0 for no, 1 for yes
	supo2	0 for no, 1 for yes
	cpap	0 for no, 1 for yes
	ventilation	0 for no, 1 for yes
	tpiece	0 for no, 1 for yes
	intubation	0 for no, 1 for yes
	chest.compression	0 for no, 1 for yes
	cardiac.drugs	0 for no, 1 for yes
	surfactant.admin	0 for no, 1 for yes
	prophylactic.indomethacin	0 for no, 1 for yes

demo_new	gender	0 for female, 1 for male
	birth.season	0 for July-Sep, 1 for Jan-Mar, 2 for Oct-Dec, 3 for Apr-Jun
	gestational.age	gestational age -39
fam_new	education	0-5 for different levels
oxy_new	auc	0-7 for different levels, +1 if 14d is more than two times larger than 7d
preg_new	diabetes	0 for no diabetes, 1 for have diabetes and receive insulin, 2 for not receive insulin
	hypertension	0 for no hypertension, 1 for have diabetes and receive medication, 2 for not receive medication
	asthma	0 for no asthma, 1 for have diabetes and receive medication, 2 for not receive medication
	rupture	0 for no membrane rupture 18 hours before delivery, 1 for membrane rupture 18 hours before delivery and no 7 days before delivery, 2 for both
	placental.pathology	0 for no placental pathology, 1 for placental pathology obtained and no histologic evidence of chorioamnionitis, 2 for both
	other.respiratory.illness	0 for no, 1 for yes
	prolong.pregnancy	0 for no, 1 for yes
	mother.smoke	0 for no, 1 for yes
	other.smoke	0 for no, 1 for yes
	alcohol	0 for no, 1 for yes
	placental.aruption	0 for no, 1 for yes
	chorioamnionitis	0 for no, 1 for yes
	antibiotics	0 for no, 1 for yes
	corticosteroids	0 for no, 1 for yes
	magnesium.sulfate	0 for no, 1 for yes
	onset	0 for no, 1 for yes
	delivery	0 for caesarean section, 1 for vaginal vertex
	preeclampsia	0 for no, 1 for yes
	bmi	Scaled mother BMI at time for delivery
fup1_new	receive.breast.milk	0 for no, 1 for yes

	non.milk.foods	0 for no, 1 for yes
fup2_new	smoker	Number of smokers at home, +1 if mother smokes
	exposed.to.smoke	0 for no, 1 for yes
	fire	0 for no, 1 for yes
	pets	Number of pets at home, -1 if care outside the home once or more per week
nas_new	nas_microbiome	Number of nasal microbiome detected
rec_new	rec_microbiome	Number of rectal microbiome detected
thr_new	thr_microbiome	Number of throat microbiome detected
tlda_new	(Various kinds of virus and bacteria)	TRUE / FALSE
vaccines_new	(Various kinds of vaccination)	0 for not taken, 1 for taken