

Which children are the most likely to have poor fiber intake?

Final Project: Biostat 202 Opportunities and Challenges of “Big Data”

Background

The American Academy of Pediatric recommend that children 12 years and under eat at least their age +5g of fiber every day. Dietary fiber is **important** for digestive health and regular bowel movements. Fiber also helps feel fuller for longer, can improve cholesterol and blood sugar levels and can assist in preventing some diseases such as diabetes, heart disease and bowel cancer. The NHANES is a program of studies run by the National Center for Health Statistics (NCHS) and designed to assess the health and nutritional status of adults and children in the United States.

Using the National Health and Nutrition Examination Survey (NHANES) Dataset, and focusing on children less than 12 years old, we will try to predict children with low fiber intake per the AAP recommendation (Age of the child in year +5g)¹ based on their demographics and socio-economic status. We will then study if asking about the consumption of fruit, vegetables and cereals help improve our predictive ability and particularly our sensitivity.

Research question

Does demographic characteristics predict low fiber intake among children aged 2-12yo and does having information about fruit, vegetables and cereals intake improve our prediction?

Methods:

Predictors used:

In the first models: Demographic only (Gender, Age, Household size, Household income: This was recoded to be usable as a continuous variable see methods. Ratio of income to poverty Race recoded as Mexican, White or Black non-hispanic (see methods)

In the subsequent models: Demographics + Consumption of fruits and vegetables + Consumption of cereals

In the final model: Age+Consumption of fruits and vegetables + Consumption of cereals

Outcome predicted: Low fiber intake.

A new variable was created that resulted in 1 if fiber consumption (predfib) was < Age + 5 and 0 if predfib >= Age + 5. This was done using a derive node.

Methods

¹ Kranz S, Brauchla M, Slavin JL, Miller KB. What do we know about dietary fiber intake in children and health? The effects of fiber intake on constipation, obesity, and diabetes in children. *Adv Nutr.* 2012;3(1):47–53. doi:10.3945/an.111.001362

A supervised learning was performed.

Importation of data: Source stat file and importing data. No issues.

Cleaning of the data: Performed in Modeler

- Type node inserted and variable type defined.
 - o SeqN defined as Record ID.
- Select Node from record Obs to select patients ≤ 12 yo = Population of interest
- Create a new variable (target): Low fiber intake = 1 if $\text{predfib} < \text{Age} + 5$ and 0 if $\text{predfib} \geq \text{Age} + 5$ using a Derive node
- Data audit node
 - o Confirm that I have selected only children ≤ 12 yo
 - o Filter some variable:
 - Non relevant: Serve in the army, Education level, pregnancy status and coffee/tea consumption.
 - Inconsistent data: Yr in the US (4% complete), citizenship status
 - o Recoding variables income to be continuous using the average in the range for each category and 10000 for the one who said < 20000 and 30000 for 20000 and above. 77 and 99 coded as missing
 - o Indhhin2 set as no predictive value and new value set as continuous
 - o 2 children have fiber intake > 80 seems impossible so marked as missing
 - o Select only population having fiber intake data available (2337/2393)

Analysis in Modeler

To answer our question, we used the autoclassifier with -C5, Logistic regression, KNN, SVM, XG boost Linear, Random trees, XGBoost Tree, CHAID and Neural Networks.

The analyses were performed first using only demographics as predictors and then using demographics and nutritional data.

Each time, the 2 best models based on AUC were selected and further analysed.

The results with and without the nutritional predictor were compared together.

Then a third model using only 3 predictors (Age, Fruit and vegetables as well as cereal intake) was assessed to see how a very basic model would perform at answering our research question.

Results

Table 1: Demographics depending on adequate fiber intake

	Total	<u>Adequate fiber intake</u>	<u>Inadequate fiber intake</u>	
Male sex	50.9%	54%	46%	$p < 0.01$
Age (mean-SD)	6.6 (+/-3.2)	5.4 (+/-2.9)	9.0 (+/-2.2)	$P < 0.01$
Number of people in the household	4.8	4.8	4.7	$P = 0.12$
Annual household income (mean in \$/y)	45 000	43 000	48 000	$P < 0.01$
Ratio of family income to poverty	1.9	1.84	2.09	$P < 0.01$
Non-Hispanic black	19%	17%	21%	$P = 0.02$

Non-Hispanic White	33%	33%	33%	P=0.90
Mexican	28%	29%	26%	P=0.07

This table is derived from Mean table.

According to this dataset and using the AAP guideline for sufficient fiber intake, as age advance, the likelihood of having low fiber intake increases. Low fiber intake is also more common in girls and in non-Hispanic black population.

According to this table children living in family with higher income seems be at higher risk of inadequate fiber intake. When graphing this information, the relationship is not as clear.

Figure 1: Prevalence of low fiber intake depending on household income.

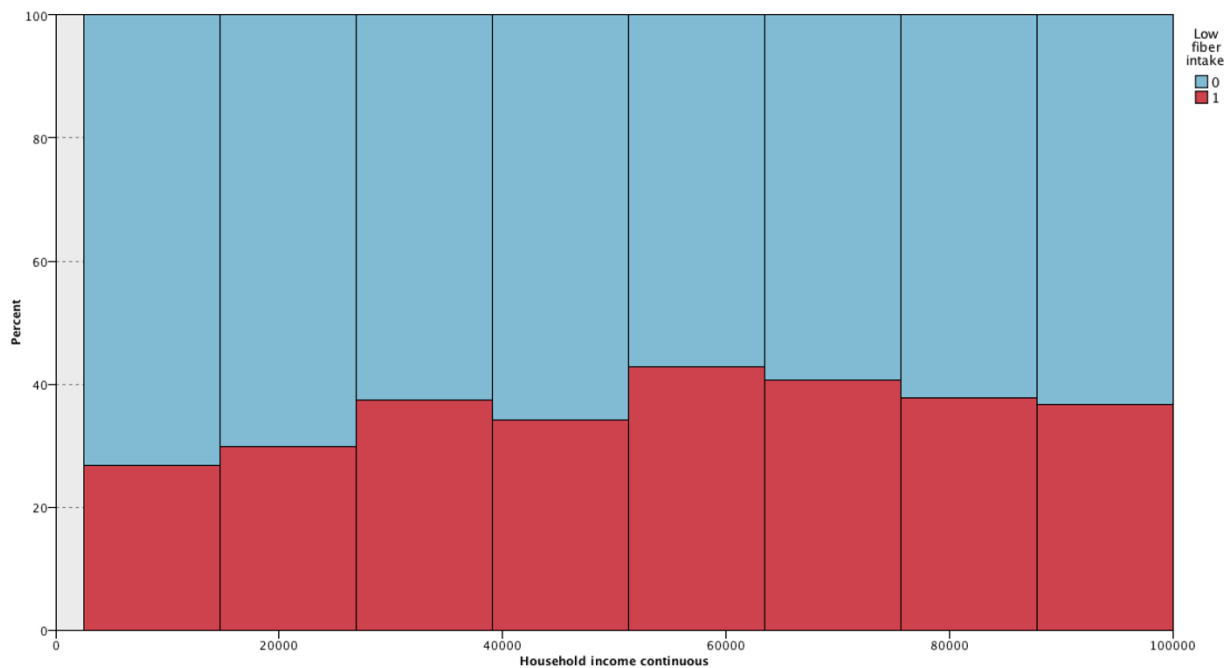
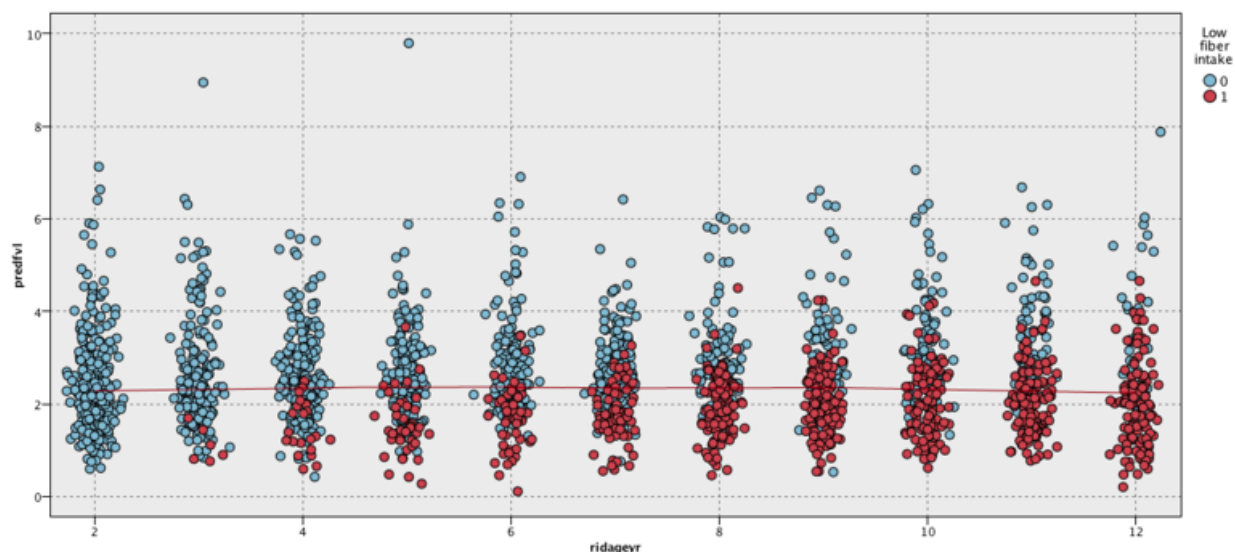


Figure 2: Consumption of fruits and vegetables in cup according to age



1. Low fiber intake
0. Adequate fiber intake

Children should eat at least their age in year +5 g of fiber per day. However, it seems that as they grow, their fiber intake does not increase hence they do not meet their nutritional needs.

Table 2: Performance of the best models using different types of predictors

PREDICTORS USED	Model selected	AUC		Accuracy		Sensitivity		Specificity	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
Demographic only	LSVM	0.84	0.81	76%	74%	62%	63%	83%	80%
	XGBoost Linear	0.84	0.81	75%	73%	55%	55%	86%	83%
Demographic + Dietary	LSVM	0.94	0.94	87%	87%	79%	79%	92%	92%
	XGBoost tree	0.98	0.94	93%	86%	96%	90%	86%	83%
Age+ Dietary	Random tree	0.92	0.91	85%	84%	88%	91%	83%	82%

Discussion

The performance of the first 2 models with the demographic variables only are quite poor. The AUC ranges in the low 80 which correspond to B-. More importantly, the sensitivity is very poor (63 and 55%). Because this predictive tool would be used as a screening test, we need to have

a high sensitivity. Indeed, we want to pick-up the majority of the cases with low intake. These models with only demographic predictors do not meet our needs.

We then added to the demographic predictors the nutritional questionnaire info of daily intake of fruit, vegetables and legume as well as cereals. The goal was to assess if that would improve our predictive ability. Unsurprisingly, the results are much improved by adding these 2 predictors. Compared to XGBoost tree, LSVM does slightly worse in term of sensitivity. However, the results are much more consistent between training and testing in LSVM. This consistency is reassuring that the model is not likely to be overfitting.

For XGBoost Tree, the training set does significantly better than the testing hence this model is probably overfitting and we prefer not to use it. Interestingly in XGBoost, the most important predictors are Age, Cereals, fruit and vegetables intake and ratio of family income to poverty. We find again this weak correlation between income and low fiber intake.

Lastly, we wanted to compare these complex models using many predictors to a simple model with only 3 predictors (1. age 2. fruit and vegetable intake and 3. cereals intake). Here we decided to use a very simple model: Random tree.

This very simple model with only 3 predictors has good AUC 0.91 correspond to A-. The sensitivity is around 90% meaning that only 1 in ten infants would be missed by this model. Furthermore, the results are consistent between testing and training.

Conclusions

- The older the child, the higher his risk is to be fiber deficient. This is especially true if he eats little fruit vegetables and cereals. Socioeconomic factors have little impact on fiber intake and pediatrician should concentrate on asking question about fruits vegetables and cereals consumption, especially in older children to screen for poor fiber intake.
- With a fairly simple model and only 3 predictors it is possible to predict low fiber intake with a reasonable accuracy and sensitivity. Sometimes simple models are the best!