Predicting Gestational Age using Metabolite Profiles: A Machine Learning Approach

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Mahdi Baghbanzadeh[1], Adam J. Ciarleglio[1]

[1] Department of Biostatistics & Bioinformatics and Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052, USA

Abstract

This study aims to methodically analyze blood metabolites during pregnancy by using weekly collected maternal blood samples to predict gestational age based on metabolites. There are a total of 784 longitudinal samples from 30 subjects throughout their pregnancy and the postpartum period. To accomplish this task, we compare the performance of several machine learning algorithms. The algorithms are evaluated using Mean Absolute Error and Root Mean Squared Error, through cross-validation to select the best one for the task at hand. Out of all the models examined, the gradient boosting machine showed the highest performance. Further analysis of this model's feature importance revealed that THDOC, Estriol-16-glucuronide, and Progesterone are the top three metabolites that contribute the most towards predicting gestational age. The results of this study can have a positive impact on obstetric healthcare, improving the well-being of women and children.

Introduction

Pregnancy is a complex physiological process that involves significant changes and adaptations in the mother's metabolism throughout each week. Accurately estimating gestational age (GA) is critical for identifying potential risks to the pregnancy and for planning appropriate care. Traditional methods for estimating gestational age are based on ultrasound measurements, but these methods can be costly and resource-intensive. Alternative approaches that analyze blood metabolites have shown promise as a more affordable and accessible option[1].

Here, we use metabolite intensities as predictive variables to predict GA and evaluated the performance of 11 different algorithms, including, Linear regression, Decision tree, Decision tree with bagging, Random Forest, Boosting tree, LASSO, Ridge, Forward selection, Backward selection, Partial least squares regression, and Principal component regression.

Data Description

In this study, we systematically analyze 784 weekly collected maternal blood samples[1] (**Fig. 1**) from 30 subjects throughout their pregnancy to predict GA using machine learning (ML) algorithms. In the dataset, GA is a continuous variable showing the GA in weeks at sampling time and the independent variables are $Log_2(intensity)$ of the 264 metabolites[1].

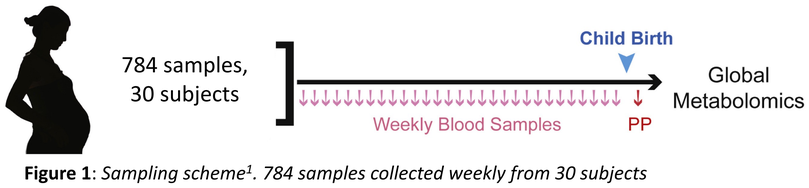

Figure 1: *Sampling scheme[1]. 784 samples collected weekly from 30 subjects*

Methods

Data Preparation:

- The analysis focused only on pregnancy data, consisting of 753 samples.
- Group 10-fold design (**Fig. 2**) was used for cross-validation, which involved dividing the data into 10 non-overlapping groups based on the subject ID.
- All the folds generated by the group 10-fold design were used consistently for evaluating all the models.
- A center and scale step was also performed to each fold to remove bias in the data and improve the performance of the ML models.

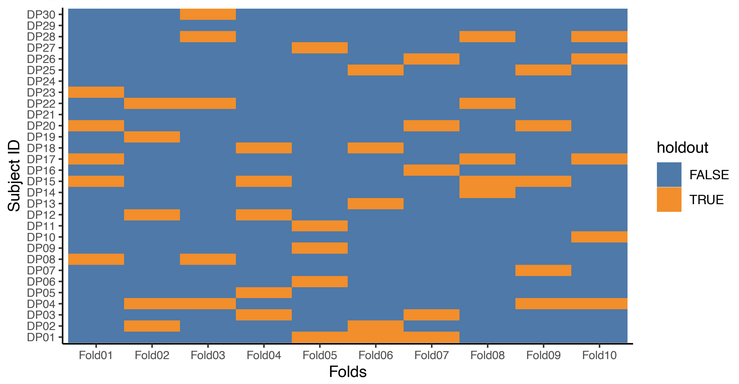

Figure 2: *Group 10-fold cross-validation design*

Principal Component Analysis

For exploratory data analysis, a PCA was performed to aid on understanding the relationships between samples and variables.

ML models and hyperparameter tuning:

We use the following models in this study optimize their performance by hyperparameter tuning using 10-fold cross-validation. Model that has the ***minimum RMSE*** is considered the best:

- **Linear regression**: a straightforward and easy-to-interpret algorithm widely used in regression tasks.
- **Forward and backward selection**: feature selection methods that choose a subset of features to use in the model. The number of variables in the final model was optimized using 10-fold cross-validation.
- **Partial least squares regression and principal component regression**: dimensionality reduction techniques that can help reduce the number of features in the model. The number of components (*ncomp*) was optimized by 10-fold cross-validation.
- **Lasso and ridge regression**: regularized linear regression methods that can handle a large number of features and prevent overfitting. The penalty term value (λ) was tuned using 10-fold cross-validation for both models.
- **Decision trees**: versatile and easy-to-interpret, but prone to overfitting. The complexity parameter (*cp*) of the tree was tuned based on 10-fold cross-validation.
- Decision tree with **bagging** and **random forest**: ensemble methods that combine multiple decision trees to improve accuracy and reduce overfitting. The number of features to be considered at each split was tuned for random forest using 10-fold cross-validation.
- **Boosting tree**: an ensemble method that focuses on improving the performance of weak learners. The maximum depth of each tree, number of trees, and learning rate were optimized with 10-fold cross-validation.

Results

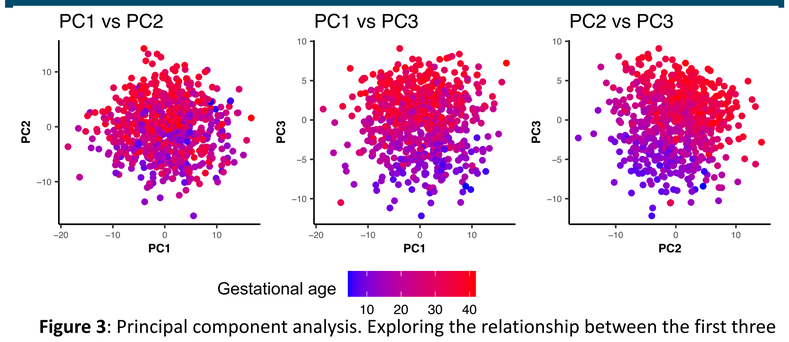

Figure 3: Principal component analysis. Exploring the relationship between the first three PCs and gestational age.

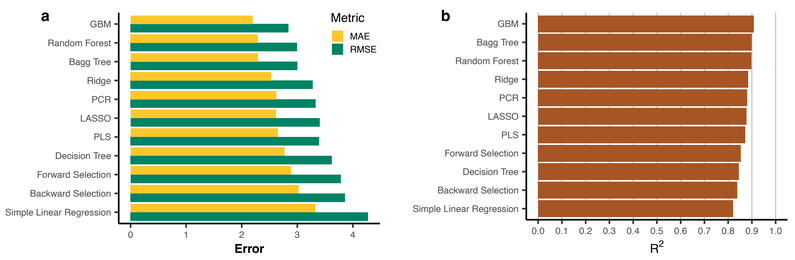

Figure 4: *Performance of the models on 10-fold cross validation study.* ***a****, MAE and RMSE scores,* ***b****,* R^2 score

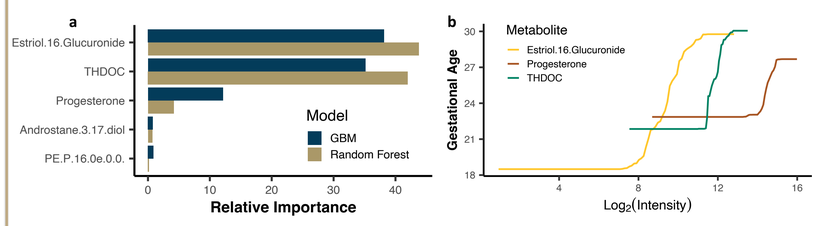

Figure 5: *Top important metabolites for predicting gestational age.* ***a****, top 5 metabolites based GBM and random forest models,* ***b****, marginal effect of top three metabolites on gestational age based on GBM model*

Discussion

The gradient boosting machine (GBM) outperformed all other 11 models in terms of RMSE, MAE, and R^2. GBMs are highly accurate and powerful for predictive modeling, as they can handle various data types and identify non-linear relationships between predictors and the target variable. The iterative tree-building process allows the model to continuously improve its performance, making it a useful tool in many real-world applications.

In the original paper[1], LASSO was used to fit the data, and the LASSO model developed in this study had comparable performance. However, other models such as random forest, bagged tree, and GBM performed better in this study. The metabolites reported here, could potentially be used as biomarkers for gestational age prediction.

One challenge in developing these models was hyperparameter tuning to optimize their performance, which required substantial computational resources and time. Additionally, the sample size of the dataset used in this study was relatively small, limiting the generalizability of the results. Future studies with larger sample sizes could help validate the findings of this study.

References

1. Liang, Liang, Marie-Louise Hee Rasmussen, Brian Piening, Xiaotao Shen, Songjie Chen, Hannes Röst, John K Snyder, et al. 2020. "Metabolic Dynamics and Prediction of Gestational Age and Time to Delivery in Pregnant Women." Cell 181 (7): 1680–92.