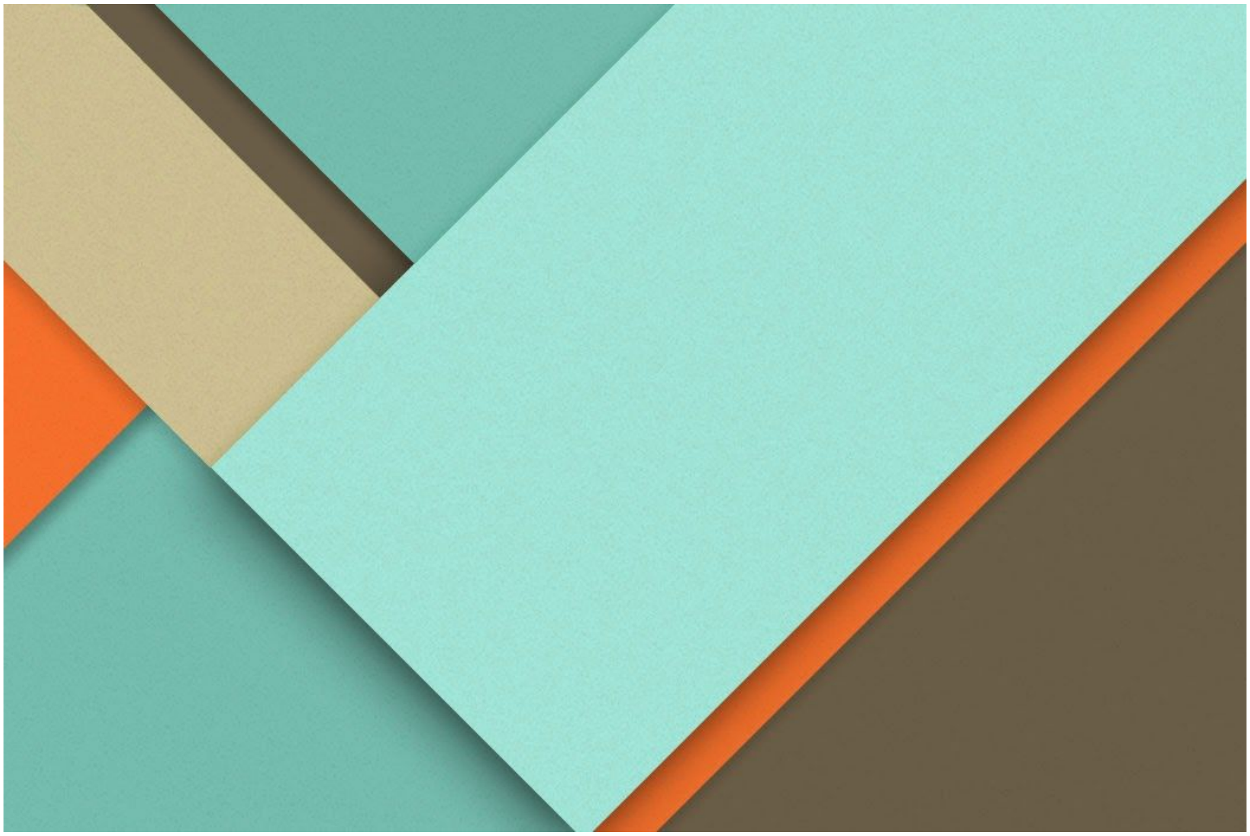

HeathCare Challenge Lab

Predicting the Gestational Age in Pregnant Women

Team: (Group - 2)

Hari Devaraj

Rakesh Siri

Sheaj Aneja

Shubham Goel

Shubham Saboo

Introduction

Multiple biological clocks govern a healthy pregnancy. These biological mechanisms produce immunologic, metabolomic, proteomic, genomic and microbiomic adaptations during the course of pregnancy. A particularly pressing issue is to identify the biological pathways and the converging pathological processes that lead to preterm birth. Preterm birth is the major cause of neonatal death, and the second leading cause of mortality in children under the age of 5 years. Modeling the chronology of these adaptations during full-term pregnancy provides the frameworks for future studies examining **deviations implicated in pregnancy-related pathologies including preterm birth and preeclampsia**.

Aim

The aim is to predict the "Gestational Age" of women based on the 7 multi-omics datasets. We have to build a Machine Learning/Deep Learning model that is capable of predicting the gestational age (GA) from temporal high-dimensional datasets including the data of **immunome, transcriptome, microbiome, proteome** and **metabolome**.

Challenges

1. Predict Gestational Age using **Immunome, SerumLuminex, plasmaLuminex** and **plasmaSomalogic** data.
2. Predict Gestational Age using **cell-free RNA, metabolome and microbiome** data.
3. Predict Gestational Age using **all of the above datasets**.

Input Data

The input data consists of the details about the Gestational Age of 14 womens. It is further splitted into train and test datasets that can be used for the development of Machine Learning/Deep Learning models. The input data has the following datasets **Immunome, SerumLuminex, plasmaLuminex**, **plasmaSomalogic**, **cell-free RNA, metabolome and microbiome** obtained during the pregnancy period.

Methodology

From a bioinformatics point of view, current multiomics efforts belong to two categories generally known as multi-staged and meta-dimensional. In multi-staged analyses, measurements of the same biological factors (e.g. genes) are integrated at various biological levels and using different technological platforms (e.g. DNA and RNA sequencing). Meta-dimensional multiomics approaches are now emerging that aim to combine heterogeneous datasets to identify key factors at various biological levels, their interactions with each other, and with clinical outcomes.

❖ Study Design

Fourteen pregnant women were invited to participate in a cohort study to prospectively examine environmental and biological factors associated with normal and pathological pregnancies. Multiple data points for these women were collected during their pregnancy timeline as per the requirement of the input datasets with aim to model the variations in the data points. The time points were chosen such that a peripheral blood sample (CyTOF analysis), a plasma sample, a serum sample and a series of culture swabs (microbiome analysis) were simultaneously collected from each woman during the first , second and third trimester of pregnancy and 6-week postpartum.

❖ Models Explored

We have explored several Machine Learning models in the exploratory process to come up with the best model for this specific use-case. The models that we have explored are as follows:

- **LINEAR REGRESSION**

  Linear Regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). This was the very first model used by us to understand the relationship between the different variables in data.

- **ELASTICNET**

  Elastic Net is a regularized regression model that combines the goodness of Ridge as well as Lasso Regression regularization techniques. It applies both L1-norm and L2-norm regularization to penalize the coefficients in a regression model. It provided us with the option of cross validation, and significantly reduced the MAE as compared to the Linear Regression Model.

- **RANDOM FOREST REGRESSION**

  Random Forest Regression is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees for coming up with a decision. We have used this model as it is an ensemble technique that is capable of combining several weak learners in the attempt to build a strong learner by randomly selecting subsets of features to work with, using the technique called Bootstrap Aggregation (also commonly known as Bagging). **The MAE in all the 3 sub-challenges was around 2.3-2.4 weeks, which was better than that of ElasticNetCV.**

- **XGBOOST**

  XGBoost is an implementation of gradient boosted decision trees designed for high performance. It provides higher execution speed as compared to any other tree based methods and also dominates the structured or tabular datasets on regression based predictive modeling problems. It is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

Results

After evaluating several models for the estimation of Gestational Age we have figured out that **XGBoost** is the most stable model in terms of Mean Absolute Error for all the three stages.

Although, **Random Forest** performs well in stage 1 and stage 3 but it shows very poor performance for stage 2, that's why it was abandoned for sub-challenge 2.

The MAE for **ElasticNet** was around 2.8 and 3.3 weeks for the data in sub-challenges 1 & 3. It was abandoned for sub-challenge 2 due to convergence issues / very poor performance.

The order of model performance as per our evaluation criteria are as follows:

XGBoost > Random Forest > ElasticNet > Linear Regression