**A combination of three ML models for prediction of preterm and early preterm birth**

Felipe-Ruiz A.1, Cremonesi F.2

1. Instituto de Biomedicina de Valencia, Spain.

2. Inria, Antibes, Provenza-Alpi-Costa Azzurra, Francia.

**Abstract**

Our model combines two K-mediods models for clustering of bacterial species and genus data and one RandomForestRegressor for the prediction of either preterm or early preterm birth.

**Introduction**

We made use of almost all the data provided for the synapse challenge. We took the values for alpha diversity, taxonomy relative species abundance, community state types, phylotype relative abundance and genus relative abundance. Because of the redundancy of the species and genus data, we only took the values of the genus *Lactobacillus* and *Prevotella*, as they are described as the main markers of risk of early delivery in previous studies (1,2).

Due to the huge multiplicity of data, we tried to reduce the complexity of the input, maximizing the amount of information delivered to the model for training. With such purpose we trained two K-mediods models for clustering using the data of relative species abundance and phylotype relative abundance, generating a prediction of clusters (between 1 and 10) for each data. The prediction of cluster was used as a variable for the training of Random Forest regressor model. The use of K-mediods models for clustering reduced the dimensionality of data and improved the values and time of training of the Random Forest model. The information extracted from these models was relevant in the training of the Random Forest Models, as some of the groups predicted in the K-mediods models showed enrichement of either non-preterm or preterm births (*Figure 1*). Once the K-mediods models were performed, we optimized the parameters for training and prediction of the Random Forest models, selecting those one obtaining a better performance in the test set for either preterm and early preterm birth prediction.



Figure 1: Clustering performed by K-mediods models.

Relative abundance of preterm birth or non-preterm birth in the clusters predicted by K-mediods on species data (A) or phylogenetic diversity (B).

**Methods**

*Data preprocessing*

The alpha diversity set, taxonomy relative species abundance set, community state types set, phylotype relative abundance 1e0 and *Lactobacillus* and *Prevotella* genus relative abundance were preprocessed as described below. Collect\_week and delivery\_week columns were eliminated for preventing data leakage. We used a OneHotEncoder for codification of all non-numerical variables in the dataset (including the output of the K-mediods models for clustering of phylotype and species abundance, SimpleImputer from sckit-learn package was used completing as “missing” all missing values, we used this strategy of imputation due to the presence of demographic and ethnological data in the variables, were missing values could be important to analyze by separate. The numerical variables were transformed using a StandardScaler, making all the variables to have media equal to 0 a standard deviation equal to 1, making that all different variables has the same effect on the model prediction.

*K-mediods clustering*

The data corresponding to phylogenetic relative abundance and species relative abundance was used as an imput for K-mediods models (we used the KMediods from sklearn\_extra.cluster), selecting 10 clusters, k-mediods++ as initialization method and random state equal to 0. Both models were used for the prediction of a variable ("Kmediods\_cluster" and "Kmediods\_phylo") used for training in Random Forest models.

*Random Forest Regressor*

For prediction of preterm and early preterm birth from thee data, we used two independent RandomForestRegressor models (from sckitlearn.ensemble). The parameters for training were optimized using GridSearch selecting multiple grid of parameters and selecting in each case the parameters obtaining a better performance in GridSearch. For the preterm model the parameters used were (bootstrap=False, max\_depth=20, max\_features='sqrt',min\_samples\_leaf=1, min\_samples\_split=2 and n\_estimators=1800) and (bootstrap=True, max\_depth=50, max\_features='sqrt',min\_samples\_leaf=2, min\_samples\_split=5 and n\_estimators=800) for the early preterm model.

**Discusion**

The models performed well using the test and train data used during the building of the model, obtaining a AUC of 0.8210878031510002 for preterm birth upon test data and 0.826840866914104 in early preterm prediction. The models were diploid as a docker container and submitted in the subchallenge 1, obtaining a bootstrap AUC value of 0.6088 obtaining the position number 32 in the dashboard.

**Authors Statement**

F-R.A, purified the datasets and contributed to establishing the prediction model. C.F., performed the docker deploiment. F-R.A, wrote the manuscript.

**References**

[1]Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med. 2019 Jun;25(6):1012–21.

[2] Kosti I, Lyalina S, Pollard KS, Butte AJ, Sirota M. Meta-Analysis of Vaginal Microbiome Data Provides New Insights Into Preterm Birth. Front Microbiol [Internet]. 2020 [cited 2022 Apr 17];11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.00476