Plant Archives Vol. 21, No. 2, 2021 pp. 351–354 e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org
DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2021.v21.no2.054

DETERMINING RELATIVE IMPORTANCE OF HUMAN GUT MICROBIOTA, AGE, GENDER AND LIFESTYLE PATTERN AS A PREDICTOR FOR BMI USING LOGMPIE DATA

Komal Jani * and Shelly Gupta
Department of Microbiology, Lovely Professional University, Jalandhar, Punjab (India)
*E-mail: komaljani0108@gmail.com

(Date of Receiving : 01-04-2021; Date of Acceptance : 16-06-2021)

ABSTRACT

We use the '*Relative Abundance Table*' and '*LogMPIE Study Metadata'* from the "*Landscape of Gut Microbiome - Pan-India Exploration*", or *LogMPIE dataset* to find out the relative importance of human gut microbiota abundance (specifically *genus*), age, gender, and lifestyle pattern as a predictor for BMI (Body Mass Index). The LogMPIE data is taken from 1004 subjects and 993 unique microorganisms are reported along with BMI, age, and physical activity. We use *Random Forest Regressor* to find out the relative importance of the above-mentioned features (microorganism genus abundance, age, gender, and lifestyle pattern) in predicting the BMI of a subject. The objective here is not the prediction of BMI using the features but to find out the relative importance of these features as much as these affect the BMI.

Keywords **:** Gut Microbiota, BMI, Machine Learning, *LogMPIE*

INTRODUCTION

The BMI of an individual depends on several and very diverse factors. In this paper we use the '*Relative Abundance Table'* and '*LogMPIE Study Metadata*' from the "*Landscape Of Gut Microbiome - Pan-India Exploration*", or *LogMPIE (Dubey et al. 2018) data set (Dubey et al. fig share).* The LogMPIE data is taken from 1004 subjects and 993 unique microorganisms are reported along with BMI, age, and physical activity.

BMI depends on several and very diverse factors such as the genetics of the individual, energy intake and energy expenditure, amount of nutrients and energy density in the food consumed, lifestyle, various kinds of imbalances and diseases, etc. Therefore it would not be fruitful to try to predict the BMI of a subject using just the above-mentioned features (microorganism genus abundance, age, gender, and lifestyle pattern). Also, it is known that the microbiota in the human gut is unique to each individual (Costello *et al.,* 2009) which would lead to highly over-fitting of any kind of Machine Learning algorithm with such a small amount of data (1004 subjects), this is another reason why predicting BMI using this data would not be fruitful. Therefore, we only set the goal to figure out the relative importance of these features as much as they are a predictor for the BMI.

The LogMPIE data contains microbiota abundances up to species level and of 993 unique species. Since there are only 1004 subjects and a comparable number of features, it would lead to high over-fitting and so we sum the data for all species of a given genus and create an abundance dataset at the genus level and we find that there are 350 of them. A small sample is shown in Table 1. From the metadata table we just keep the age, gender, life_style_pattern and BMI columns. A small sample is shown in Table 2. We use Random Forest Regressor (Ho, Tin Kam 1995; Pedregosa *et al*., 2011), a machine learning model which fits several decision trees on various sub-samples of a data-set and then averages the predictions of the decision trees. The depth of a decision tree and the number of features included to construct a tree is in our control and can be used to improve predictive accuracy and to control over-fitting.

We use BMI as the dependent variable which the algorithm would learn to predict given all other variables (features). But as we mentioned above predicting the BMI is not our goal, we want to find the relative importance of the various features and determine the most important ones as a predictor for BMI. To find the feature importance we first use our model and find its accuracy in predicting BMI, then we take a feature (a column in the data) and randomly shuffle only that column and keep the rest of the data as it was. This should render that feature irrelevant for prediction. Then we use our model again for predicting and find its accuracy. We record the difference in the accuracy of the prediction that we get once with the original data and once with a given feature randomly shuffled. If a feature, say 'a', is more important than another, say 'b', then the reduction in accuracy by randomly shuffling the feature 'a' would be more in comparison to the reduction in accuracy by randomly shuffling the feature 'b'. We do this for each feature and record the reduction in accuracy and then scale those numbers appropriately to find the relative feature importance. After all the analysis the final top *10* important features among the ones we considered as a predictor for BMI, along with their relative importance is given in Table 5.

Table 1 : A small sample of first 5 rows of abundance data-set at the genus level obtained by summing the abundance data for all species of a given genus.

Subject_id / Genus	*Acidiphilium*	*Blautia*	*Campylobacter*	*Kluyvera*	*Odoribacter*
1001	0.0002812600450	0.0147058823529	0	0	0.0008839601414
1002	0	0.0071138211382	0	0	0
1003	0	0.0024247768103	0.0002204342554	0	0.0007347808516
1004	0.0049927102235	0.003996265956	0.0009335109451	0	0.0030207994629
1005	0	0.0012562182804	0	0	0.0006281091402

Table 2 : First 5 rows of the metadata table where we just keepthe age, gender, life_style_pattern and BMI columns.

Subject_id	gender	age	life_style_pattern	BMI
1001	Male	34	Non Sedentary	26.8122
1002	Male	64	Non Sedentary	23.2434
1003	Female	54	Non Sedentary	32
1004	Female	29	Non Sedentary	24.9199
1005	Female	24	Non Sedentary	28.3987

MATERIALS AND METHODS

Initially, we create a data set from the original by summing up the abundance of all the species of a given genus. This dataset has 1004 rows (as there are 1004 subjects) and 354 columns (350 genus of microorganisms, age, life_style_pattern, gender, and BMI). Then we randomly separate 200 rows from the above-mentioned dataset to create a dataset call it ‘test set’ and we put the remaining 804 rows in a separate dataset call it ‘analysis set’.

The reason behind keeping the test set is, that we do our analysis on one set of data and that means our model has seen all of that data during training and there is a possibility that our model is over-fitted to a high degree or there could be some error in the analysis because of some kind of peculiarity in our data. We can use our test set (which our model has never seen) to make predictions and make sure that the accuracy that we get on the test set is similar to the one we get during analysis.

Now out of the *354* features, we consider BMI as a dependent variable and all others as independent. We then fit a Random Forest Regressor using the scikit-learn library for Python programming language. Following is the description of the parameters that are important for our analysis in the Random Forest Regressor of scikit-learn.

RandomForestRegressor(n_estimators=100, criterion= 'mse', min_samples_leaf=1, max_features='auto', bootstrap= True, oob_score=False)

n_estimators - the number of decision tree estimators, predictions of which will be averaged for final prediction (the number of trees in the forest). Default value is 100 but we will tune it for maximum performance.

criterion - the criteria or performance metric on which the regressor learns (trains). Default is 'mse' which stands for mean squared error. This means that during the training process the regressor will try to minimize the mean squared error between the predicted BMI and the actual BMI. We keep the default value for this.

min_samples_leaf - the minimum number of samples required to be at a leaf node of a tree. The default value is 1 but we will use higher value for this because this has the effect of smoothing the model and reduce over-fitting. As we have already explained above that our model, because of the nature of our dataset is very highly prone to over-fitting.

max_features - the number of features that will be considered when looking for the best split at a node of a tree. We use the value max_features=0.5 which means that 50% of the features will be considered each time when looking for the best split. This is important because if we use all the features every time when looking for the best split and if there is a feature in our dataset that is highly dominant then most of the trees will have their first split based on that feature and that will reduce the diversity of our forest.

bootstrap - it stands for the same bootstrap procedure that is used in statistics. If bootstrap is 'False' then to build a tree all the rows of the dataset will be used. If bootstrap is 'True' then to build a tree same number of rows, as are present, will be picked at random with replacement. We will keep the default value 'True'.

oob_score - if this has a value 'True' then out-of-bag samples - which are the rows that are “left out” in the original data when taking bootstrap samples - are used to estimate the (explained below) on unseen data. (This is also why we can do away with a 'validation set' which is normally used in machine learning but reduces the amount of data in the training set which we have called analysis set above. The oob_score provides us with a way to estimate the model's performance on unseen data without actually using another unseen dataset.)

We will use R^2 (coefficient of determination) denoted "R squared"as a performance measure for our model. R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In the best case when the model predicts values which exactly match the observed value then we will have $R^2 = 1$. A model which always predicts a single value which is the average value of the dependent variable, will result in $R^2 = 0$.

RESULTS AND DISCUSSION

We do our analysis in three stages:

Stage 1

We use all the 354 features for creating the random forest regressor model using the analysis set. Once the model

is trained with some values of parameters, we can make predictions for the dependent variable. Using the predictions we calculate R^2 for the entire training set (analysis set) and compare it with the oob_score which is generated. As this model is highly prone to over-fitting, with default values of Random Forest Regressor parameters, R^2 will turn out to be far greater than the oob_score. We tune our parameters until the R^2 is close to the oob_score. (R^2 is generally slightly higher than the oob_score because it is the result of prediction on data which is seen by the model whereas oob_score is an estimate forR^2on unseen data and so we account for that). After tuning, following parameters were used to train the final stage 1 model:

[n_estimators=10000, criterion='mse', min_samples_leaf=50,max_features=0.5, bootstrap=True, oob_score= True]

The result we get is R^2=0.069 and oob_score = 0.045. Now we calculate the feature importance as described before by randomly shuffling a feature and record the difference in the accuracy (here high mean squared error correspond to less accuracy) of the prediction that we get once with the original data and once with a given feature randomly shuffled. We do this for each feature and record the reduction in accuracy and then scale those numbers appropriately to find the relative feature importance. The 20 most important features after stage 1 analysis are given in Table 3.

Stage 2

In this stage we only use these 20 most important features from stage 1 for the analysis. Rest is same as described in stage 1. After tuning, following parameters were used to train the model:

[n_estimators=10000, criterion='mse', min_samples_leaf=240, max_features=0.5, bootstrap=True, oob_score =True]

The result is R^2=0.044 and oob_score = 0.033. The 10 most important features after stage 2 is given in Table 4.

Stage 3

In this stage we only use these 10 most important features from stage 1 for the analysis. After tuning, following parameters were used to train the model:

[n_estimators=10000, criterion='mse', min_samples_leaf=240, max_features=0.5, bootstrap=True, oob_score =True]

The result is R^2=0.044 and oob_score = 0.034. The feature importance table we get after this final stage is given in Table 5.

This is our final result for the top 10 features as in importance as a predictor for BMI. We notice that gender has negligible importance as a predictor for BMI as it doesn't even appear in the top 10 features. Age has almost 50% importance and lifestyle pattern has around half the importance of age but is still significantly important as a predictor for BMI. Bifidobacterium, Blautia, Dorea, Dialister, Akkermansia, Veillonella, Roseburia, and Coprococcus are the top 8 genus of microorganisms that have some importance in predicting BMI.

Table 3: The 20 most important features as a predictor for the BMI after stage 1 of analysis.

Feature	**Relative Importance**	**Feature**	**Relative Importance**
age	0.44122	*Dialister*	0.009509
life_style_pattern	0.190641	*Veillonella*	0.009231
Bifidobacterium	0.105314	*Escherichia*	0.008757
Blautia	0.028642	*Ruminiclostridium*	0.008582
Akkermansia	0.01993	*Coprococcus*	0.007529
Alistipes	0.013654	*Romboutsia*	0.0075
Parabacteroides	0.013447	*Gemmiger*	0.006198
Dorea	0.012139	*Mitsuokella*	0.006009
Oscillibacter	0.011541	*Roseburia*	0.005911
Pseudomonas	0.010292	*Turicibacter*	0.005692

Table 4 : The 10 most important features as a predictor for the BMI after stage 2 of analysis.

Feature	**Relative Importance**	**Feature**	**Relative Importance**
age	0.45564	*Akkermansia*	0.02362
life_style_pattern	0.20774	*Dialister*	0.02286
Bifidobacterium	0.07038	*Veillonella*	0.02044
Blautia	0.0475	*Coprococcus*	0.02036
Dorea	0.02454	*Roseburia*	0.01918

Table 5 : Final result of the 10 most important features as a predictor for the BMI after stage 3 of analysis.

Feature	**Relative Importance**	**Feature**	**Relative Importance**
age	0.45848	*Dialister*	0.03038
life_style_pattern	0.22146	*Akkermansia*	0.03008
Bifidobacterium	0.08338	*Veillonella*	0.02786
Blautia	0.06156	*Roseburia*	0.02658
Dorea	0.03412	*Coprococcus*	0.0261

Finally we use our test set (which is completely unseen by our model) to calculate R^2 and check whether it is close to our final *oob_score.* We find $R^2$for test set to be R^2_{test}=0.015 which is reasonable and shows predictability.

REFERENCES

Dubey, A.K. *et al.* (2018). LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing. *Sci. Data*. 5:180232 doi: 10.1038/sdata. 2018.232.

Dubey, A.K. *et al*. *fig share* https://doi.org/10.6084/m9. figshare.c.4147079

Costello, E.K.; Lauber, C.L.; Hamady, M.; Fierer, N.; Gordon, J.I. and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science*, 326(5960): 1694–1697.

Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

Pedregosa *et al.* (2011) Scikit-learn: Machine Learning in Python. *JMLR 12,* pp. 2825-2830.