




UNIVERSIDADE DA CORUÑA

# Lab 2: Critical comment on a proposed article

*Fundamentals of Bioinformatics*

Expert Systems With Applications 185 (2021) 115648

Contents lists available at [ScienceDirect](#)

 **Expert Systems With Applications**

journal homepage: [www.elsevier.com/locate/eswa](http://www.elsevier.com/locate/eswa)



**Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes**

Diego Fernández-Edreira<sup>a</sup>, Jose Liñares-Blanco<sup>a,b</sup>, Carlos Fernandez-Lozano<sup>a,b,\*</sup>

<sup>a</sup> Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, Campus Elviña s/n, A Coruña, 15071, Spain

<sup>b</sup> CITIC-Research Center of Information and Communication Technologies, University of A Coruña, A Coruña, 15071, Spain

Pedro Sánchez García  
Season 2021-22

## Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes

### Critical comment

We are currently situated in a context of broad technological advances, which translate into an improvement in knowledge and treatment of problems that constituted important challenges for science. An example is the clinical field, where the complexity of certain pathologies and the handling of their data by conventional analysis have slowed down advances in diagnosis and treatment. In the case of Type I diabetes, the keys of the progression are not fully understood. It was in recent years when several promising pieces came into play: the microbiota, Machine Learning techniques and mass sequencing. If we focus on each of them, with respect to the microbiota, we can highlight that it's fascinating to know the intimate relationship that we present with that high number of microorganisms, which acts in our life as a balance in which imbalances end in diseases such as Type I diabetes. Based on the above, numerous studies focused on the microbiota-disease relationship emerged, which improved notably thanks to the use of the information provided by mass sequencing techniques, another key piece that offers the profile of the microbiota for each individual. However, massive sequencing provides such a large amount of data that it's difficult to handle and search for relationships using conventional analysis. At this point, the last key pieces joins, which are Machine Learning techniques, whose consist in the introduction of data, in this case from sequencing, to an algorithm, by means of which a model is obtained that allows to carry out predictions for other data sets. If we look closely, the set of parts constitutes a robust and promising machine for the clinical field.

The present work is located in this set, where it starts from a cohort accessible to the scientific community with data on the microbiota of children and the characteristic workflow of these projects is followed, as explained in the classes, where three algorithms (Random Forest, Support Vector Machine and Generalized Linear Model) are evaluated using R, a very powerful and versatile tool, different variables that provide more information from the starting data are extracted and a model is reached, which is trained and compared with the others for subsequent project processes. In this case, like the result we reached in practice with Weka of the subject, the Random Forest algorithm is the best to achieve the model, taking into account the rigorous analysis of the accuracy and AUC to make the decision. On the other hand, the extraction of the variables, corresponding to species of bacteria, shows an aspect that I found very interesting: the selection of species that are not significant, but that appear and allow the model to show us as a "signal" of its relevance in the development of the disease, which contributes to another achievement of the project. Likewise, the possibility of identifying seroconverted patients by classifying the proposed model is of great help for those intermediate stages in which Type I diabetes is not detected and which represent a challenge for current medicine. Based on the results obtained in the work, we can appreciate the coherence with respect to the knowledge present in the scientific literature, as it happens with *Bacteroides uniformis*, with a role in obesity and metabolic pathologies and others of the same genus, which indicates the

robustness of the model in the interpretation of the relevant results for the diagnosis of Type I diabetes.

I would like to conclude this critical comment by highlighting a series of aspects. In the first place, despite the fact that in the Microbiology subject in the degree we had been commented on the relevance of the microbiota, I was fascinated by its magnitude when I saw the I International Conference and Workshop: Microbiota: What ' s cooking ? a few months ago, where prestigious scientists intervened, who showed with extensive and rigorous explanations, the complexity of the microbiota in different pathologies. This relevance was widely highlighted in the last class, with the advancement of massive sequencing and the need for bioinformatics to treat the corresponding data. All this, together with the Machine Learning techniques, of which I have acquired a good base in the classes, have also allowed me to follow the different concepts and decisions exposed in the article. Personally, I find this type of project very interesting and show the robustness and capacity of promising techniques to shed light on the complex network that constitutes diseases such as Type I diabetes. Without a doubt, I will consult in the next years the evolution of this project, subsequent transfer and validation to the clinical field.