Polytechnic University of Catalonia

Barcelona School of Informatics

Computer Science

System that predicts the source of water fecal matter contamination

Author
Ricard Meyerhofer Parra

Supervisor
Dr. Luís Antonio
Belanche Muñoz

July 22, 2018

Abstract

Fecal matter contains bacteria that in contact with other organisms such as humans, can be harmful. Drinking water contaminated by fecal matter can cause stomach and intestinal illness including diarrhea and nausea, and even lead to death.

In order to prevent human exposition to tainted water and to detect the causative source, microbial source tracking (MST) describes a suite of methods and an investigative strategy for determination of fecal pollution sources in environmental waters that rely on the association of certain fecal microorganisms with a particular host. These values depend on the source, dissolution, time, localization and season. Retrieving them is an expensive process.

In this project, we keep with the dynamic of recent literature where machine learning methods have been used to successfully detect the causative source. To contribute to this task, an online platform is created. This platform allows the user to combine microbiological analysis of water with machine learning techniques to detect the source of contamination and allows the user to introduce microbiological information, visualize it, create and execute prediction models and interact with the obtained results.

This project will allow the scientist community applying machine learning algorithms into MST data isolating them from difficulties such as the environment, programming languages and the economic cost.

Acknowledgements

I would like to thank Dr. Luís Antonio Belanche Muñoz firstly, for being my teacher in the Machine Learning course, which made me realize of what I want to work as. I want also to thank his help and advice during the project and thank for providing me the chance of learning and working in a field that I love.

I would also like to highlight the unconditional support received from my family during all these years that allowed me to study what interests me the most.

I would like to thank inLab FIB which has always welcome me with open arms and also for giving me the opportunity of a first job. I want also to thank Dr. Ricard Gavaldà for his support, working flexibility and the opportunity to work as a undergraduate research assistant in data science.

Finally, I want to thank all my friends and colleagues for their support. Without all them, reaching this point would have been way harder.

To my girlfriend Yejin, thanks for your apprehension and unconditional support.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Ricard Meyerhofer

Index

List of Figures

List of Tables

1 Introduction

1.1 Project formulation

The human exposure to fecal matter contamination, can be dangerous and can lead to diarrhea, vomits and even to death. These effects may be more severe and possibly life threatening for babies, children, the elderly or people with immune deficiencies or other illnesses. Fecal matter can be more or less harmful depending on the source, which makes it important detecting which source is the causative to prevent humans exposure and know what occurred.

The detection of each source is done by microbiological analysis where values of some bacteria allow us to know the contaminant source. These values depend on the time, increment of dissolution, localization and season. Experiments to retrieve this data are very expensive and take time to be done.

Currently, the process of some given data detecting which source is the origin is a procedure that is not done automatically (scientists know which is the source because they infer it by looking at the data). This process is easier when fecal matter is pure (only one source). If the fecal matter is mixed, this starts to be more problematic because it could be a mixture of more than two origins.

For the reasons above, Universitat de Barcelona (UB) department of microbiology that has a good relationship with my director Dr. Luís Antonio Belanche, proposed him to create a system that predicts the source of water fecal matter contamination.

This system will allow the user to combine microbiological analysis of water with machine learning techniques to create a system to detect the source of contamination and create a website that allows the user to introduce microbiological information, visualize it, create and execute prediction models and interact with the results.

1.2 Personal motivation

Before entering to Polytechnic University of Catalonia. I have always thought that building a website would be awesome. Once I started my degree, I re-

alized that there are many fields of computer science and finding which one you like is not easy.
Luckily for me, I decided to choose Computing as specialization and to enroll a Machine Learning subject [17] taught by Dr. Luís Antonio Belanche Muñoz [15]. At that instant, I discovered what I want to become skillful.
When I had to choose my final degree project, there was no doubt: I wanted it to be Machine Learning related and I wanted Dr. Luís Antonio to be my thesis supervisor.
I talked with Dr. Luís and from several projects available, I chose this project for the following reasons:

- **I have already worked as web developer** for a year and half at inLab FIB [16] (until the beginning of this final degree project).
- **It generates a positive impact on society:** An improvement in the prediction on the source of fecal matter contamination, is given to the scientist community. This is relevant because the exposition to bacteria such as salmonella is dangerous [7] to humans.
- **UB microbiological department [6] is interested** in this system. Furthermore, this project will be accessible to the scientist community so this project will be used and useful. I am delighted that my work can be applied.
- **Data science is everyday more popular**. Businesses have realized that, nowadays, the huge amount of data that used to be stored can be used to do predictions, improve business decisions, and many other applications.

1.3 State-of-the-art

There is no literature about a system that allows to introduce microbiological information, choose variables, create a prediction model and interact with the data. However, there is literature regarding to the prediction of fecal matter pollution in water but due to its concreteness, there is few work.

Focusing now in literature related to fecal water source tracking, an evolution during the years is shown. The starting point and the problematic that motivate the use of machine learning are that indicators present in fecal matter are not good enough to be used in source detection. Many MST techniques have been used such as assemblage and ratios between different microorganisms, phenotyping of bacterial isolates, genotyping of microorganisms, chemicals methods, etc.

Studies done in 2004 conclude that at the moment, there was no MST approach to accurately identify the origins of fecal pollution in aquatic environments.

In 2006-2007, even that any source identifier by itself can reach 100% correct classification, by using a combination of MST and machine learning methods, two predictive models based on two variables gave 100% correct classification.

In 2014-2015, appear studies where MST techniques are applied successfully to water with diverse pollution loads. Almost all MST methods in this study determined correctly the origin of fecal at the point source and in moderate concentration samples. Therefore, inductive machine learning methods are shown as a promising tool although there is still margin to improve.

2 Context

In this section, the context of the project will be covered. Concretely, the stakeholders, means and technologies will be covered.

2.1 Stakeholders

Knowing our stakeholders facilitates building more adequate product as being aware of their requirements and letting them be involved affects positively in our final product [12].

2.1.1 Developer

The author of the project is the main responsible of the product development and wants the project to successfully finish at time to be able to expose it as final degree project.

2.1.2 Director

The director of the project has the responsibility of orientating, giving advice, detecting errors ,and generally helping the developer. His action is key considering that he has profound knowledge and experience in the field.

2.1.3 Direct beneficiaries

The beneficiaries of this project are an heterogeneous group as it benefits not only the scientific community, concretely the UB microbiological department, but it also benefits any scientist that is interested in the source prediction of water pollution field.

These users will be benefited because the analysis and results of this kind of data are economically and temporally expensive and thanks to this system this data can be artificially simulated.

2.1.4 Indirect beneficiaries

A scientific profile is not the only kind of user that can get advantage of our system. A non-scientist user indirectly, can get benefited due to a better analysis of the water that affects to his daily life and health such as improvements in detection of the pollution in a certain water area that can affect his health.

2.2 Means and technologies

2.2.1 Definition of the environment

Once presented the state of the art and the context, it is time to introduce which environment would suit best in this project.
The system will be developed for PC (personal computer). The reasons are the following:

- **Working purposes system:** Because of this, the user does not need a constant access that a mobile can offer us. A mobile version is not useful as a computer version since computer has become a common daily work tool for scientists.

- **Big data to treat and working comfort:** The data that is going to be displayed is too big to fill in the phone in a comfortable way. Similar issues occur with other processes because of the screen size. Finger occlusion in phone and the complexity of clicking small items affect small screens.

In summary, phone is not suitable in this case because it is not comfortable and does not fit in a working context. There is no argument to build this system for a phone. Nevertheless, the system that is going to be constructed will be responsive to be displayed in all kinds of screen sizes.

2.2.2 Means

Once set that the system will be available only for PC, the name of the media required to develop the system is the following:

- **PC:** A computer is required for developing the project, test and execute it.
- **Latex:** To document the project Latex [13] will be used. Concretely the online platform ShareLatex. [14]
- **Google Chrome:** Even that RStudio allows us to test in the IDE, to test the responsiveness of the project, it is better to use the phone mode of google chrome. Also to check the code that Shiny creates and improve the UI, inspector mode is necessary.

2.2.3 Technologies

The technologies that will be used to successfully build the project and the reasons to choose them are:

2.2.3.1 R

R [1] is a programming language and software environment for statistical analysis, graphics representation and reporting. R is freely available under the GNU General Public License, and various versions for systems like Linux, Windows and Mac are provided.

2.2.3.2 RStudio

RStudio [2] is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management, etc

2.2.3.3 Shiny

Shiny [3] is an open source R package developed by RStudio that provides an elegant and powerful web framework for building web applications using R. Shiny also allows to create interactive website of analyses without mandatory use of HTML, CSS, or JavaScript knowledge.

2.2.3.4 Science nodes

Science nodes [20] is a platform that allows us to share and collaborate in project without having to install any environment (e.g., R and RShiny). Additionally, it facilitates user management, security and remote executions.

2.2.3.5 Git

Git [19] is a free (under the terms of the GNU General Public License) version control system for tracking changes in computer files and coordinating work on those files among multiple people. It is mainly used for code management, but it can be also used to keep track of changes in any set of files. As a distributed revision control system, it is aimed at speed, data integrity, and support for distributed non-linear work-flows.

2.2.3.6 Github

GitHub [18] is a web-based Git and Internet hosting service. It offers all of Git as well as its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project.

GitHub offers both plans for private and free repositories on the same account, which are commonly used to host open-source software projects.

2.2.3.7 Justification

- One of the reasons that motivates me to use R + RStudio + Shiny is that Dr. Luís Antonio Belanche Muñoz has already done the prediction with pure fecal matter in R and this code will be incorporated in the website. However, the main reason is that the combination of R and Shiny allows me to do a beautiful website easily with nice graphics and it will be easy to integrate it with previous works.
- RStudio is an IDE that allows to execute R and Shiny directly and in general, makes the developing process easier.
- Github is a personal choice, other services instead of Github could have been used as Bitbucked, Gitlab, etc. I have a student pack in Github so I can make private my repository in case of being necessary because for example, there is private data.
- Science nodes, provides security and user management to the project which otherwise, wouldn't be possible considering the time limitation of the project. My director suggested it because he has previous experience with the platform.

3 Scope of the project

Firstly, I will get in touch with Shiny by looking at the tutorial and how it works. Once familiarized with the framework, I will start implementing the upload, download and microbiological information related (introduction of the information, modifying, etc) and the overall design. One one hand, to validate the design, I will show it to my director periodically. On the other hand, to allow the user to manipulate microbiological information, a study of the different libraries will be done (which will include proves of concept to check the real potential of each) and validated by checking if it fulfills the requirements and is intuitive enough.

Once the design and microbiological information are done, the system will be aggregated to science nodes. Science nodes leads us to more secure environment and abstracts us to create users, and allows remote and scheduled executions. To validate these works, tests will be applied.

When the website and science node aggregation are done, the next step will be to create a super-matrix to be able to predict considering two factors, the season that that each microbiological has and the dissolution. The season is roughly how the bacteria grow/fall given a certain clime and the dissolution, is the amount of water where the results are extracted.

Once the super-matrix is created, the following step will be to integrate my director's work (we can validate this easily by checking if the result on the website is same as in a normal execution). Once it works as expected and there is time, the mixed part will start where the objective will be to get as much as precision as possible (as in the pure part). The goal would be to reach a 100% precision rate but this is a pretty ambitious goal and unlikely to be reached.

When the project is completely developed, testing of all functionalities will be done again and, if possible, I will try to improve my director's work. To check the improvement, measurements of the prediction rate in different cases will be done. Lastly if there is time, instead of predicting the season with a simple generalized linear models (glm) a better approach will be implemented.

3.1 Project Objectives

The two main objectives of the project are:

- Combine microbiological analysis of water with machine learning techniques to create a system to detect the source of contamination.
- Create a website that allows the user to introduce microbiological information, visualize it, create and execute prediction models and interact with the results.

As it can be seen clearly, this project has two differentiated parts: One part is purely related with applying machine learning techniques and the other part consists in creating a website so that user can use these models, introduce data ,and interact with it.
For these two main objectives, we can specify their objectives:

- **Machine Learning part**
  - Aggregate the work of my director to the project to accomplish the machine learning prediction in the case of only one source.

  - If possible create the machine learning part in order to be able to predict with more than one source.
  - If possible, improve the part of my director and in both cases, have a high precision prediction.

- **Website part**
  - Allow the user to upload, load, modify, and download microbiological information.
  - Allow the user to choose which information to display, the microbiological behaviour of each variable, the dissolution, creation of a prediction model, loading a previous prediction model.
  - Visualize and interact with the results. Display relevant information in order to facilitate user decisions and understanding of the problematic.

3.2 Methodology

What best suits a project of this type and dimensions is an agile methodology that, in this case, will be Scrum [**?**]. Scrum provides great flexibility, allows a fast development and is result-oriented unlike other non-agile methodologies such as cascade. However, Scrum ,as agile methodologies, is team-oriented so it is difficult to apply all of its principles such as stand-up dailies. Nevertheless, being team-oriented does not mean that it cannot be reduced to individuals.

One of the reasons that made me choose Scrum instead of other agile methodologies is that I have experience in working with Scrum and I feel comfortable with most of its principles.

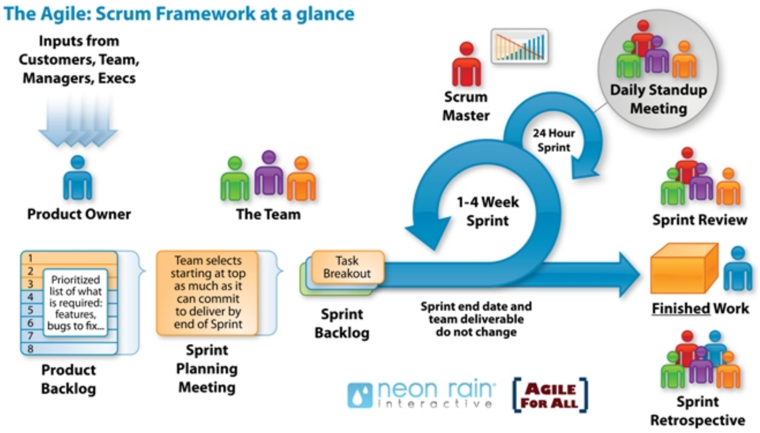

The Agile: Scrum Framework at a glance
Inputs from Customers, Team, Managers, Execs
Product Owner
The Team
Scrum Master
Daily Standup Meeting
24 Hour Sprint
1-4 Week Sprint
Sprint Review
1
2
3
4
5
6
7
8
Prioritized list of what is required: features, bugs to fix...
Product Backlog
Team selects starting at top as much as it can commit to deliver by end of Sprint
Sprint Planning Meeting
Task Breakout
Sprint Backlog
Sprint end date and team deliverable do not change
Finished Work
Sprint Retrospective
neon rain interactive
AGILE FOR ALL

Figure 1: Scrum phases

3.2.1 Sprints

As it can be seen in the diagram above, Scrum uses time unit sprints lasting from 1 to 4 weeks. Using an iterative approach with short cycles helps to:

- Keep the project on schedule.
- Receive feedback frequently.
- Be aware of the project state.
- Have constant workload, thus avoiding peaks.

3.2.2 Validation methods

In this project, there will be two clients: The director of the project and UB microbiological department. Since they have different roles, the opinions from both sides will be considered important and they will have positive impact on the final product.

As previously mentioned, my director's supervision and gatherings with UB, will help validating the results of the project. If tests are required, they will be included in the project even though the results are difficult to verify.

Depending on what needs to be verified, some tools or others will be used. In our case, currently are being considered:

- **RUnit:[23]**Allows unit testing in R.
- **RSelenium:[22]** To check if there are visual differences in comparison with what is expected.

3.2.3 Methodology changes during the project

It has been possible to follow Scrum but not in a consistent way. Eventually, I got stuck in some features. In addition with workload peaks, this made me unable to do meetings for more than a Scrum cycle. However, aside from the specific case, Scrum has been useful to keep the project on track and to get periodic feedback of my director and UB who have been highly positive for the project.

4 Project planning

4.1 Planning and scheduling

The estimated time for the proeject is 4 months. It will start on 11th September 2017 and will finish one day before the presentation date that will be between 22nd- 26th January 2018.
Even the planning is done, it is more than probable that the original schedule suffers of variations and modifications while the project advances due to the fact that an agile methodology is being used.

4.2 Task description

4.2.1 GEP course

In this task, I will get familiarized with the project and understand what has to be done. Moreover, I will analyze the requirements and difficulties of each part to have a global idea of how the project will be and do a study of what has been done so far. Lastly, I will look into business parts such as planning and budget.

This task will take 50 hours in total. As resources, only the amortized cost of computer, electricity and human cost will be required and to write, sharelatex that is free. From these hours, it will be decomposed as following:

- **Delivery 1:** State-of-the-art, scope (project objectives, methodology, validation), context, means and technologies, etc).
- **Delivery 2:** Temporal planning, task scheduling and estimation, Gantt chart.
- **Delivery 3:** Budget (cost identification and control) and sustainability matrix of the project (social, economical and environmental dimension).
- **Delivery 4:** Draft of a presentation.
- **Delivery 5:** Justification of the academic competences, subjects coursed in computing that help in the project, and basis why is from computing.
- **Delivery 6:** Whole project delivery, final presentation.

Task name	**Estimated Hours**
Delivery 1	10
Delivery 2	10
Delivery 3	10
Delivery 4	5
Delivery 5	5
Delivery 6	10
Total	**50 hours**

Table 1: GEP course hours planning

4.2.2 Familiarization with Shiny

Considering that I have no previous knowledge in Shiny, spending a short period of time getting familiarized with Shiny will not only be beneficial once I start doing the website itself, it will be also beneficial in its structure, quality of code ,and will reduce the time spent in the implementation.

In order to learn how Shiny works, I will follow the tutorials and check the gallery and codes that are on Shiny's website.

4.2.3 Website visual appearance

In this task, all the front end of the website (visual appearance) will be done. As resources, there is the cost of R, Shiny, RStudio, Github (all free) and amortized costs of computer, electricity and finally the human cost of the 50 hours that will be used to implement the front end.

It is very difficult to specify exactly all that has to be done and the time expected on each task. An approach would be the following:

- **Brainstorming design:** In this task, what we are going to do is to check sites like Pinterest [21] to see different designs that can be suitable for what we are looking for. Also some trials of different designs will be done in order to check which one suits best.
- **Table screen:** The screen where we have to choose a file to display a table and afterwards, allow for each column to select a season that will be used in the prediction part. For each of this columns, the season will be shown as an interactive plot, that allows to see how the data of each season is.
- **Model creation screen:** In this screen, we will provide the user all the options to allow the selection of the different options in the creation of this model.
- **Results screen:** This screen will show a serie of different graphics in order to help the user visualize and interact with the final results.
- **Front end review:** Once the whole front end is finished, we will check that everything is working as supposed and if possible, improve some UX mistakes.

Task name	**Estimated Hours**
Brainstorming design	5
Table screen	15
Model creation screen	10
Results screen	15
Front end review	5
Total	**50 hours**

Table 2: Distribution of hours for the front end task

4.2.4 Website back end

In this section, the back end of the website and the union with front end and back end will be done. During this task, checking that each functionality behaves correctly and executes fast will be a priority. This task will be done in parallel with the previous section as it has been mentioned. The only difference in resources with the previous task is that the human cost is 75 hours instead of 50.

In a similar way as in the front end, the planning of the back end is difficult. An overview would be the following:

- **Table library:** There are several libraries that allow us to represent and create an interactive table. Deciding which is the one that suits best for our requirements is the objective of this task.
- **Table screen:** In this screen, we will have to implement how to store the data selected, to display the selected data, to download the file, and to save the file, etc.
- **Model creation screen:** In this screen, we will have to link the selected options from front end that do the calculus in the back end in the following processes.
- **Results screen:** The implementation of the graphics and the plots that will be shown is going to be developed.
- **Bugs detection:** Once the whole back end is done, we will look for errors and if everything works appropriately.

Task name	**Estimated Hours**
Table library	5
Table screen	25
Model creation screen	15
Results screen	20
Bugs detection	5
Total	**75 hours**

Table 3: Distribution of hours for the back end task

4.2.5 Incorporate Machine Learning prediction on pure fecal matter prediction

Previous work has been done concretely on the pure fecal matter prediction. This work has to be understood and adapted, which will be possible due to my background acquired in the Machine Learning subject coursed one year ago and my knowledge in R from some other subjects.

Once understood this work, it will be incorporated to the website and if possible, will be improved.

The resources used for this task are going to be the code provided by my Director, R, RStudio, Shiny, Github, the computer amortized cost, electricity and the human cost specified in the budget.

4.2.6 Improve Machine Learning background

It is possible that this task runs in parallel with incorporating machine learning prediction. However, mainly in this task, my background of machine learning will be improved to be able to do a good approach on the mixed fecal matter prediction and, if possible, previous approach on pure fecal matter.

The main objective is learning in order to succeed in creating the machine learning part. The resources are not clear yet asides from the human cost.

4.2.7 Creation of a model on mixed fecal matter

For this task, all my knowledge in machine learning will be applied to create a model that predicts successfully the origin source in mixed fecal matter. What will be done in this task is seeking for the best model that we can create across different methods and approaches.

The resources used for this task are going to be R, RStudio, Shiny, Github, the computer amortized cost, electricity and the human cost specified in the budget.

4.2.8 Integration with Science Nodes

In this task, the integration of science nodes with the website will be done. To accomplish this task, science nodes documentation will be read and some help of Gerard Cegarra will be required considering that he has experience with the platform.

The resources used for this task are going to be R, RStudio, Shiny, Github, Science Nodes, the computer amortized cost, electricity and the human cost specified in the budget.

4.2.9 End of project

This last task will consist in roughly finishing all pending tasks such as preparing the exposition, finishing and revising the document, and polishing some errors in the website or machine learning parts.

The resources used for this task are going to be R, RStudio, Shiny, Github, ShareLatex, the computer amortized cost, electricity and the human cost specified in the budget.

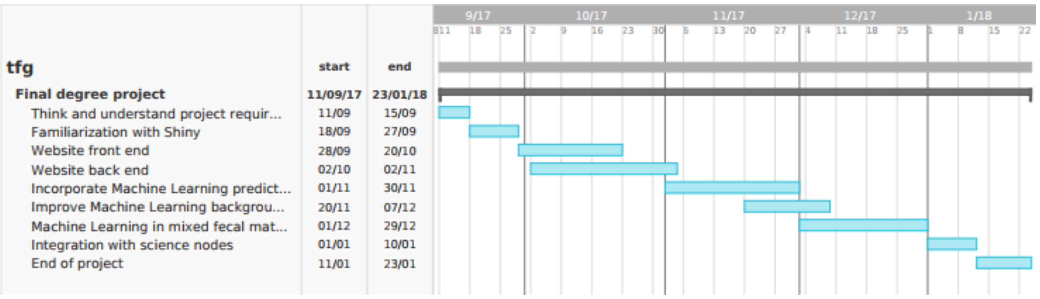
4.3 Estimated time

Task name	**Estimated Hours**
GEP course	50
Familiarization with Shiny	10
Website visual appearance	50
Website back end	75
Incorporate Machine Learning	50
Improve Machine Learning background	50
Creation of a model on mixed fecal matter	115
Integration with Science Nodes	25
End of project	25
Total	**450 hours**

Table 4: Project task distribution

4.4 Diagrams

4.4.1 GANTT



tfg	start	end
Final degree project	**11/09/17**	**23/01/18**
Think and understand project requir...	11/09	15/09
Familiarization with Shiny	18/09	27/09
Website front end	28/09	20/10
Website back end	02/10	02/11
Incorporate Machine Learning predict...	01/11	30/11
Improve Machine Learning backgrou...	20/11	07/12
Machine Learning in mixed fecal mat...	01/12	29/12
Integration with science nodes	01/01	10/01
End of project	11/01	23/01

Figure 2: Gantt chart of the project

4.5 Alternatives and action plan

The project resources are going to be the same on each task except the human cost that will be higher due to the problems that can appear during the project development, which are the following:

4.5.1 Website problems

- **Responsiveness:** Responsiveness occurs when a website is not suitable for all screen sizes and some problems like panel overlapping occur. This kind of problems diminish our user experience or can directly deny it. It is harder to detect these errors than to correct them. In order to detect them, what will be done is the test with Chrome[5] in different screen sizes with the inspector mode.
  Even I acknowledge the possibility of this problem, it should not be problem theoretically since Shiny creates fully responsive website. In case there are a lot of problems, some screen sizes will be specified and the project will guarantee a good user experience just for those, which will not cause a deviation from the project.
- **Unexpected behaviour:** A function or an action that does not perform as it should. To avoid this kind of problem, we will test each new feature once it is finished.

  However, if there is an unexpected behaviour that is relevant for the user experience, this will be solved. However, if it takes too much time to solve and is not significant, we will documentate it.

4.5.2 Prediction problems

- **Data quality:** Data is what our system is fed with. If data quality is not good enough, predictions will be poor. At the beginning of this project apparently the data quality is good because Dr. Luís Antonio Belanche Muñoz could do a successful prediction with pure fecal matter.

  In this case, the data is not good enough (even it is unlikely), we will just have to work with it because there is no alternative.
- **Over-fitting:** This happens when the model created to predict variables is not generic enough and is only valid for a reduced data set. To avoid this problem there will be always two subsets: training and test.

In the training dataset, the model will be created and will be tested in the test data to check the success rate of prediction. If a drastic difference is obtained between the successful rate of training set and test set, it will mean that there is over-fitting or non-significant training set.

In this case, we will check if the given training set is appropriate and complete enough. Also, we will check if the selection of this dataset is arbitrary.

4.5.3 Global problems

- **Scheduling:** Too optimistic scheduling can make our project foundations tremble. To create a realistic schedule is as important as the project itself. To avoid and solve problems that can happen with the schedule, from the first day of the project, there will be sprints and periodical gatherings with my director not to deviate from the planning.

  In case that the project deviates, the tasks of improving my machine learning background will be reduced and some help of my director in the creation of a model on mixed fecal matter, will be required.

- **Losing the project:** To avoid the loss of the project for problems such as hard drive disk not working or computer being stolen, the use of control version is a must. In this project, as it is explained in technologies section, Github will be used.

4.6 Changes from the original scheduling

During the development of the project, the importance of a consistent, scalable, secure, and reliable website has gained importance. In order to accomplish this requirement and due to the complexity of the website requirements, a major deviation has emerged.

This deviation implies changing the focus of the project to another that is more focused in the website part rather than in the machine learning part. This schedule change will guarantee UB's Microbiological department satisfaction in the final product.

In an economical point of view, the project does not suffer a significant deviation. The only factor that is modified is the number of hours of each role which is provided in the following table.

Role	**€/h**	**Hours**	**Salary(€)**
Web developer	15	250	3750
Data Scientist	25	150	3750
Project Manager	40	50	2000
Total			**9500 €**

Table 5: Different roles cost considering project's deviation

This also affects to the unexpected costs because it is related with extra hours of each role.

Role	**€/h**	**Hours**	**Salary(€)**
Web developer	15	30	450
Data Scientist	25	15	375
Project Manager	40	10	400
Total			**1225 €**

Table 6: Unexpected costs considering project's deviation

And the total cost is reduced to 12552.25€(a reduction from the initial expected cost).

Expense name	**Cost(€)**
Human	9500
Software	0
Hardware	122
Unexpected	1225
Indirect	68
Subtotal	10915
Contingency (15%)	1367.25
Total	**12552.25 €**

Table 7: Total cost

Regarding to the planning, the project has also suffered modifications mainly because as previously mentioned, now we are emphasizing more in the website part. In the following table, the new distribution can be seen where the order has also been modified.

Task name	**Estimated Hours**
GEP course	50
Familiarization with Shiny	10
Website visual appearance	80
Website back end	125
Integration with Science Nodes	35
Incorporate Machine Learning	95
Improve Machine Learning background	15
End of project	40
Total	**450 hours**

Table 8: Project task distribution with deviation

As mentioned previously in how to tackle global problems, in the deviation part, it was already mentioned that improving machine learning background could be reduced and help of my director in the creation of a mixed fecal matter model could be required. Finally, as it can be seen in the table, the machine learning part is reduced significantly in order to provide the website that the client needs.

I want to highlight that currently the website of the project is almost finished and that Sciences Nodes integration is also about to be finished. If possible, even though it is not planned, I will try to exceed the number of hours of TFG to do all the Machine Learning part that I proposed in my first schedule (personal motivation).

5 Budget

In this section, an estimated cost of the project is presented showing which hardware and software resources are used and their amortizations. Indirect costs of the project are also considered.

5.1 Cost identification

The following costs have been identified and in the coming sections will be detailed:

5.1.1 Human resources budget

The project is going to be developed by only one person during 450 hours that is the working time estimated for a final degree project.
In this 450 hours, different roles are required such as Project Manager, Web developer and Data Scientist. In the following table, an estimation of the human resources cost is provided.

Role	€/h	Hours	Salary(€)
Web developer	15	150	2250
Data Scientist	25	250	6250
Project Manager	40	50	2000
Total			**10500 €**

Table 9: Different roles cost

The next table, details the time that each role spends in the different tasks of the project:

Name	Hours	Web Dev (h)	Data Scientist (h)	P. Manager (h)
T1	50	22	11	17
T2	10	8	1	1
T3	50	35	10	5
T4	75	50	20	5
T5	50	10	35	5
T6	50	0	45	5
T7	115	10	100	5
T8	25	5	18	2
T9	25	10	10	5
Total	**450 h**	**150 h**	**250 h**	**50 h**

Table 10: Work distribution for each role

Task	**Name**
GEP course	T1
Familiarization with Shiny	T2
Website visual appearance	T3
Website back end	T4
Incorporate Machine Learning	T5
Improve Machine Learning background	T6
Creation of a model on mixed fecal matter	T7
Integration with science nodes	T8
End of project	T9
Total	**450 hours**

Table 11: Equivalences table

5.1.2 Software

Here is a detailed table of the costs of the project. Since the project uses license free tools, the cost of software is zero.

Product	**Units**	**Price (€)**	**Useful life**	**Amortization**
R	1	0	-	0
RStudio	1	0	-	0
Shiny	1	0	-	0
Science Nodes	1	0	-	0
Git	1	0	-	0
Github	1	0	-	0
Latex	1	0	-	0
Total		**0 €**		**0 €**

Table 12: Software cost

Softwares such as RStudio are only for free if the project is open-source. In this case, the website is open-source so the cost is 0€.

5.1.3 Hardware

As hardware costs, we will consider two computers that will be used during the project. One is a laptop and the other one is a customized computer.

Product	Units	Price (€)	Useful life	Amortization (€)
Lenovo Y580	1	1100	7 years	52
Custom Computer	1	1486	7 years	70
Total		**2586 €**		**122 €**

Table 13: Hardware cost

5.1.4 Unexpected costs

In this section, we show the number of extra hours as a result of suffering a deviation from the original planning. This gives a certain budget margin that will prevent getting run out of money in case of not being able to stick to the original planning of the different tasks.

Role	€/h	Hours	Salary(€)
Web developer	15	20	300
Data Scientist	25	20	500
Project Manager	40	10	400
Total			**1200 €**

Table 14: Unexpected costs

5.1.5 Indirect costs

Inside the total budget of the project, we have to take into consideration costs that are not related directly with the project such as water, internet, electricity, etc.

In the next table the price and the costs are detailed:

Product	**(%)**	**Price (€/month)**	**Months**	**Cost (€)**
Water	10	50	4	20
ADSL	30	40	4	48
Total				**68€**

Table 15: Indirect costs
4

5.1.6 Total expenses

The total expenses of the project is the sum of the costs that have been presented previously with a 15% for contingencies:

Expense name	**Cost(€)**
Human	10500
Software	0
Hardware	122
Unexpected	1200
Indirect	68
Subtotal	11890
Contingency (15%)	1783.5
Total	**13673.5 €**

Table 16: Total cost

5.2 Budget Control

To control budget deviations, at the end of each task the number of hours will be updated with the real amount of hours that have been accomplished, the cost of the resources used and unexpected cost. With the numbers obtained, a deviation from the first planning can be done with the succeeding formulas:

Cost deviation = (Estimated Cost - Real Cost) $\cdot$ Real Hours
Consumption deviation = (Estimated Hours - Real Hours) $\cdot$ Estimated Cost

If a high deviation is appreciated at the end of a task, the possible reasons for this will be analyzed and if there is any conclusion, it will be applied to the next tasks. However, some measures like unexpected costs and contingencies over the total cost have been taken which makes it highly improbable with a good planning. This is because main problem with the budget, is the human cost deviation which is regulated and checked previously in this document.

6 Sustainability and social commitment

6.1 Sustainability matrix

The purpose of this section is to analyze the impact that our project has on the economic, social and environmental dimension to evaluate the sustainability of the project. This analysis will be based on the next sustainability matrix:

Sustainability	**Economic**	**Social**	**Environmental**	**Range**
PPP	10/10	8/10	10/10	28/30
Useful life	18/20	18/20	20/20	56/60
Risks	-1/-20	-2/-20	0/-20	-3/-60
Total				**81/90**

Table 17: Sustainability matrix

From the total obtained, it can be concluded that the sustainability of the project is very high.

6.2 Economic dimension

The material cost is very low considering that only computer and electricity are the only non-human requirements that the system needs to work and even in the case that the project deviates, still being low.

The human cost is low considering that it is only developed by one person. If we consider outsourcing the project to other countries such as India, the project would be cheaper but it would not change dramatically (and communication could be harder).

Normally the data is generated in laboratories where real fecal matter is poured on the water and different variables are measured across the time.

The problem of these tests is that the media to do these experiments is required and there is also a human cost. Summarizing, the whole process is expensive. However, with this project, the number of results you can get is enormous in comparison with doing one by one, economically and also temporally since in this system you can have the results immediately. Also, all the money that used to be destined to the generation of this data set now can be used for other things that is required.

This project directly is not linked to any project but the studies from which we start are European projects.

6.3 Social dimension

This project wants to use the efforts done by different Microbiological departments in MST leaded by UB. Currently, obtaining microbiological data from fecal matter pollution is expensive so there is a lack of data.

Creating a system that allows the source prediction of the fecal matter is beneficial for the Microbiological departments in terms of costs and analysis and actually has the support of UB microbiological department.

This system not only makes life of scientists easier but also life of people in general because being exposed to water that is polluted with fecal matter can be dangerous.

If there is a system that can reduce the time to detect which is the origin of the fecal matter, people contaminated by fecal matter can be reduced. In case it is not dangerous, then users will be less affected by for example a beach closed due to fecal matter contamination when it is not threatening to human being.

The development of this project is not going to harm any collective. In fact, it will benefit society transversely and will help the society to have conscience about the problem that suppose pouring fecal matter in water. A problem that in first world countries is not considered.

6.4 Environmental dimension

The only footprint is the electric cost of developing the project and later, just the cost of computer to access the website and electric cost of hosting a website. In exchange, we save in car displacements, water used in future experiments and the electricity used for all the media and to light the laboratory so we are diminishing this footprint.
Furthermore, the code developed is going to be scalable and modularized which leads to reusing some generic parts of it.

The data used in the project can be reused to any project that requires the data. The data that our system predicts can lead to create artificial data.

For all the reasons above, we can say that it is sustainable.

7 Justification of the project specialty

7.1 Subjects useful for the project

I think that any subject from my mention helps to develop skills because they are crosscutting. Once said this, the most relevant are the following:

- **Algorithmics:** This subject is useful in terms of efficiency and being able to analyze the complexity of the algorithms that will have to be implemented in the project. The main objective of the subject is showing different programming techniques that can be suitable for a certain problem and because of this, is a very transversal knowledge that can be used and applied at any project.

- **Programming Languages:** As in the previous case, this subject gives overall knowledge that can be applied to any project. In this case, different programming languages and paradigms are taught in a way to cover what any language can potentially do and which language is the most suitable for the certain system. This allows me to explode the full potential of R with different techniques, to fully understand the language itself ,and to choose the best language to accomplish my objective.

- **Machine Learning:** This subject (guided by my director) which fully covers machine learning techniques and methods is completely suitable with what will be done in the project: Context, basic knowledge, different techniques and methods that otherwise I would have had to explore by myself.
- **Searching and Analysis of Massive Information:** Data mining and machine learning topics are covered in this subject. It does not directly deal with the topic that my project covers but it is related with some topics like clustering algorithms, data treatment and in general analysis of data.

7.2 Specialty of the project

This project fulfills the requirements of my specialty that in this case, is computing. Some of the reasons are the following:

- **Machine Learning based:** A significant part of this project is machine learning based and machine learning is a discipline inside computing. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data through building a model from a training set. Currently machine learning is applied to several fields. Among them, some are email filtering, computer vision, learning to rank, and etc.
- **Visualization of complex data:** Representation of complex data, require complex algorithms and data analysis to show to the final user something significant and understandable. Without this data representation, would be impossible to understand and interpret the data.
- **Efficiency and security are present on the website:** As most of online platforms, security of sensitive users and data, is present. Also, efficiency is mandatory because otherwise, anyone will want to use a slow platform and if the system faces with big data inputs, then it would be too slow due to a bad algorithms implementation.

In summary, this project is inside the computing category because treats with complex algorithms, has learning methods, interact and visualizes with the data given in a secure and efficient environment according to the requirements of the project.

7.3 Academic objectives

In this section I present the objectives that my project will cover followed by their importance. This academic objectives, can have four labels according to their importance: **None, A little bit, Enough** and **In depth**.
The ones that are not covered by my project (None), will not be presented for clarity:

- **CCO1.1:** To evaluate the computational complexity of a problem, know the algorithmic strategies which can solve it and recommend, develop and implement the solution which guarantees the best performance according to the established requirements. **[Enough]**
- **CCO1.3:** To define, evaluate and select platforms to develop and produce hardware and software for developing computer applications and services of different complexities. **[A little bit]**
- **CCO2.1:** To demonstrate knowledge about the fundamentals, paradigms and the own techniques of intelligent systems, and analyze, design and build computer systems, services and applications which use these techniques in any applicable field. **[A little bit]**
- **CCO2.2:** Capacity to acquire, obtain, formalize and represent human knowledge in a computable way to solve problems through a computer system in any applicable field, in particular in the fields related to computation, perception and operation in intelligent environments **[Enough]**
- **CCO2.3:** To develop and evaluate interactive systems and systems that show complex information, and its application to solve person-computer interaction problems. **[Enough]**
- **CCO2.4:** To demonstrate knowledge and develop techniques about computational learning; to design and implement applications and system that use them, including these ones dedicated to the automatic extraction of information and knowledge from large data volumes. **[In depth]**

- **CCO3.1:** To implement critical code following criteria like execution time, efficiency and security. **[Enough]**

Once introduced the importance of each objective and what each of them covers they will be accomplished in the following way:

- **CCO1.1:** This objective will be covered by taking into account in all algorithms how important it is and which is the algorithmic approach that will work according to them.
- **CCO1.3:** Choosing software and hardware technologies during the project such as R, Shiny require of this kind of analysis.
- **CCO2.1:** This can be accomplished because our website, includes an intelligent system that is the predictor so we are using an intelligent system.
- **CCO2.2:** All the interactive part where the user introduces information or selects options, and how the system treats this information to give back the results is related with the objective.
- **CCO2.3:** The system will allow to introduce microbiological data, visualize and interact with it. This is completely related with the competence.
- **CCO2.4:** All the machine learning part where we create a prediction system, completely corresponds to this part.
- **CCO3.1:** Because the website has to be secure and fast, efficient algorithms and a secure environment with science nodes will be provided.

8 Identification of laws and regulations

This project needs an authentication in order to be able to access to the datasets, created models and in general all kind of data that is necessary to use the website. However, Science Nodes is the platform that manages the users which implies that is not responsibility of this project how Science Nodes treats information and deals with privacy or sensitive data.

The website is an open source project since everyone can see the repository and access to the website so there is no need to worry about copyright problems. Regarding to the libraries used to build the website, we have ensured that all of them are open source therefore free for a non-commercial use.

In case it would be commercialized, RStudio is not free anymore since it is only free for open source purposes.

For all the reasons above, this project does not trespass any law or regulation.

9 Design and implementation of the website

The development of the website has been driven through some key aspects: user experience, scalability, performance and invested hours. In the next sections, a justification of the most relevant decisions will be done and an overall design will be provided

9.1 Requirements

Before introducing and reasoning about the design of the website, it is important to analyze which requirements the product has to cover and which are the most relevant. Knowing which requirements are the most important, allows a more efficient and useful development towards the final product that the stakeholder will get at the end of this final degree project.

In the project case, some requirements appeared during the development and others were present from the beginning.

Functional Requirements

- Provide a professional appearance and a good UX.
- Save/Load any given data.
- Being able to visualize/sort/modify, select which columns to display and in general, interact with the loaded data.
- Choose locations from a folder according to the column name and the possibility of saving them.

- Upload our own location in different formats (T90 or T99).
- Create/Load/Save the BIGMATRIX generated in order not to recalculate.
- Select which algorithm we want to apply to the BIGMATRIX.
- Visualize prediction.
- Integration with science nodes in order to have users and security in our platform.

Among many non-functional requirements that are appropriate for our website, a significant list is provided.

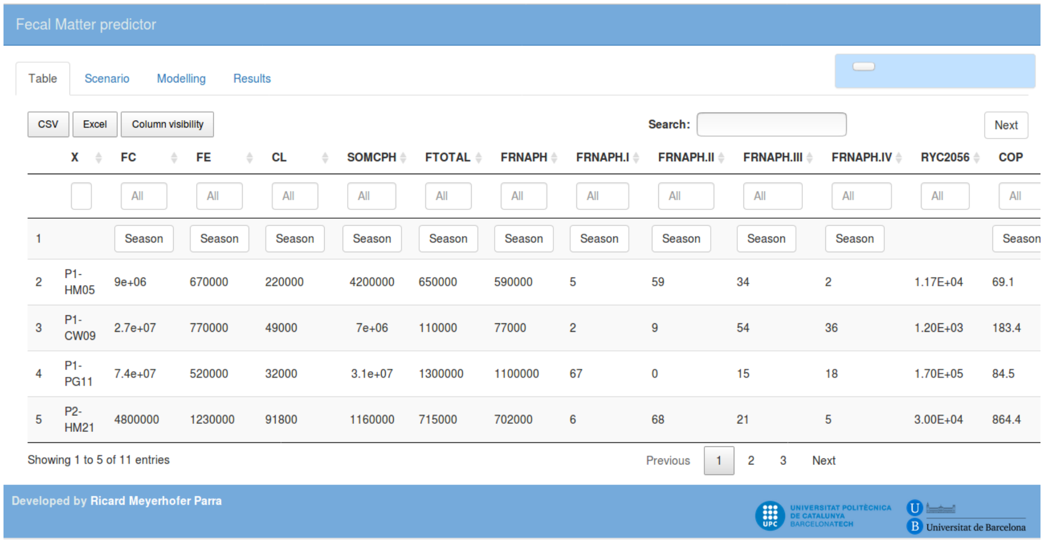
Non-Functional Requirements

- Accessibility
- Extensibility
- Performance
- Privacy
- Scalability
- Security
- Reliability
- Usability

9.2 General appearance and UX

One of the top priorities when building the website was the User Experience. Using the website has to be an easy task in order to facilitate users work.

During the design of the website, many ideas were evaluated. The design had to be something minimalist, easy to use, professional-looking and visual. In the first versions, the idea of a tabset on the top and a collapsible panel at the right side were the user could load the data was the best option considering the few options in the tabset. However, there was too much occlusion with the data because of its size and elements such as the footer, were not useful which motivated to discard this option.

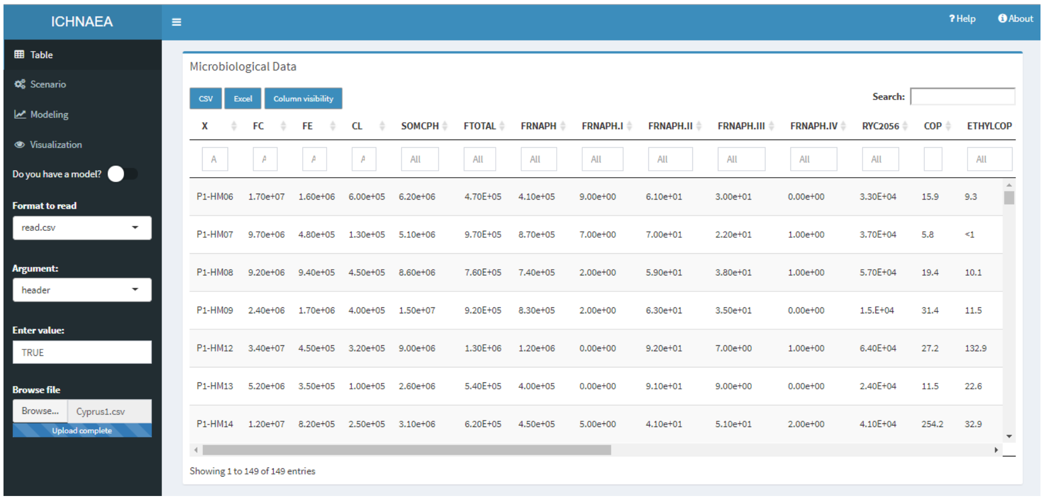


	X	FC	FE	CL	SOMCPH	FTOTAL	FRNAPH	FRNAPH.I	FRNAPH.II	FRNAPH.III	FRNAPH.IV	RYC2056	COP
		All	All	All	All	All	All	All	All	All	All	All	All
1		Season	Season	Season	Season	Season	Season	Season	Season	Season	Season		Season
2	P1-HM05	9e+06	670000	220000	4200000	650000	590000	5	59	34	2	1.17E+04	69.1
3	P1-CW09	2.7e+07	770000	49000	7e+06	110000	77000	2	9	54	36	1.20E+03	183.4
4	P1-PG11	7.4e+07	520000	32000	3.1e+07	1300000	1100000	67	0	15	18	1.70E+05	84.5
5	P2-HM21	4800000	1230000	91800	1160000	715000	702000	6	68	21	5	3.00E+04	864.4

Showing 1 to 5 of 11 entries

Figure 3: One of the first approaches of the Dashboard.

There was the need of a design that minimizes those occlusions, gives a more professional look, looks more responsive and in general that was more consistent and allowed a relatively fast development. What we thought that matches best with this ideas is a dashboard and RStudio, has a package named shinydashboard[25] that facilitates the creation of one. A fully custom dashboard is something that is out of the scope of the project due to the time restriction.



Format to read: read.csv

Argument: header

Enter value: TRUE

Browse file: Cyprus1.csv

Microbiological Data

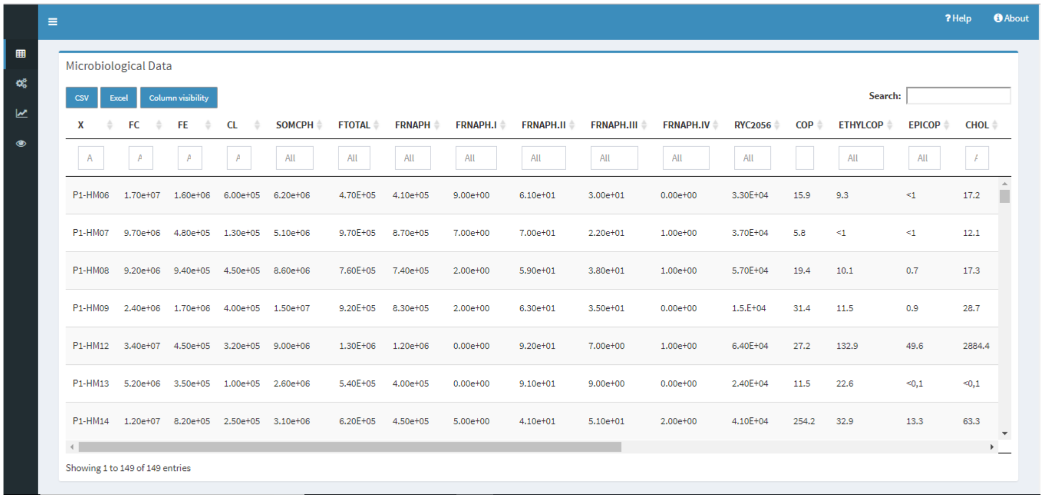
Search:

X	FC	FE	CL	SOMCPH	FTOTAL	FRNAPH	FRNAPH.I	FRNAPH.II	FRNAPH.III	FRNAPH.IV	RYC2056	COP	ETHYLCOP
A	A	A	A	All	All	All	All	All	All	All	All		All
P1-HM06	1.70e+07	1.60e+06	6.00e+05	6.20e+06	4.70E+05	4.10e+05	9.00e+00	6.10e+01	3.00e+01	0.00e+00	3.30E+04	15.9	9.3
P1-HM07	9.70e+06	4.80e+05	1.30e+05	5.10e+06	9.70E+05	8.70e+05	7.00e+00	7.00e+01	2.20e+01	1.00e+00	3.70E+04	5.8	<1
P1-HM08	9.20e+06	9.40e+05	4.50e+05	8.60e+06	7.60E+05	7.40e+05	2.00e+00	5.90e+01	3.80e+01	1.00e+00	5.70E+04	19.4	10.1
P1-HM09	2.40e+06	1.70e+06	4.00e+05	1.50e+07	9.20E+05	8.30e+05	2.00e+00	6.30e+01	3.50e+01	0.00e+00	1.5.E+04	31.4	11.5
P1-HM12	3.40e+07	4.50e+05	3.20e+05	9.00e+06	1.30E+06	1.20e+06	0.00e+00	9.20e+01	7.00e+00	1.00e+00	6.40E+04	27.2	132.9
P1-HM13	5.20e+06	3.50e+05	1.00e+05	2.60e+06	5.40E+05	4.00e+05	0.00e+00	9.10e+01	9.00e+00	0.00e+00	2.40E+04	11.5	22.6
P1-HM14	1.20e+07	8.20e+05	2.50e+05	3.10e+06	6.20E+05	4.50e+05	5.00e+00	4.10e+01	5.10e+01	2.00e+00	4.10E+04	254.2	32.9

Showing 1 to 149 of 149 entries

Figure 4: Final result of our dashboard.

Even though, in order to improve user experience, some modifications to exploit the capabilities of shinydashboard have been done. These modifications, allow a smoother user experience. For example, the creation of a semi-collapsible bar that instead of fully disappear when clicking the hamburger icon, still shows each tabitem in order to optimize screen space and user experience and including a help and an about button on the right side of the navbar would be the most relevant ones.



Microbiological Data

Search:

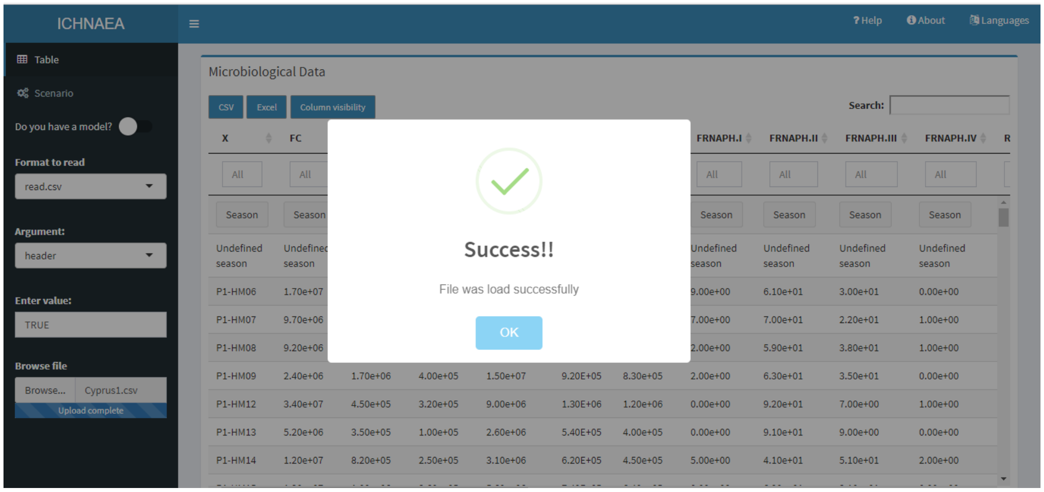
X	FC	FE	CL	SOMCPH	FTOTAL	FRNAPH	FRNAPH.I	FRNAPH.II	FRNAPH.III	FRNAPH.IV	RYC2056	COP	ETHYLCOP	EPICOP	CHOL
A	A	A	A	All	All	All	All	All	All	All	All		All	All	A
P1-HM06	1.70e+07	1.60e+06	6.00e+05	6.20e+06	4.70E+05	4.10e+05	9.00e+00	6.10e+01	3.00e+01	0.00e+00	3.30E+04	15.9	9.3	<1	17.2
P1-HM07	9.70e+06	4.80e+05	1.30e+05	5.10e+06	9.70E+05	8.70e+05	7.00e+00	7.00e+01	2.20e+01	1.00e+00	3.70E+04	5.8	<1	<1	12.1
P1-HM08	9.20e+06	9.40e+05	4.50e+05	8.60e+06	7.60E+05	7.40e+05	2.00e+00	5.90e+01	3.80e+01	1.00e+00	5.70E+04	19.4	10.1	0.7	17.3
P1-HM09	2.40e+06	1.70e+06	4.00e+05	1.50e+07	9.20E+05	8.30e+05	2.00e+00	6.30e+01	3.50e+01	0.00e+00	1.5.E+04	31.4	11.5	0.9	28.7
P1-HM12	3.40e+07	4.50e+05	3.20e+05	9.00e+06	1.30E+06	1.20e+06	0.00e+00	9.20e+01	7.00e+00	1.00e+00	6.40E+04	27.2	132.9	49.6	2884.4
P1-HM13	5.20e+06	3.50e+05	1.00e+05	2.60e+06	5.40E+05	4.00e+05	0.00e+00	9.10e+01	9.00e+00	0.00e+00	2.40E+04	11.5	22.6	<0,1	<0,1
P1-HM14	1.20e+07	8.20e+05	2.50e+05	3.10e+06	6.20E+05	4.50e+05	5.00e+00	4.10e+01	5.10e+01	2.00e+00	4.10E+04	254.2	32.9	13.3	63.3

Showing 1 to 149 of 149 entries

Figure 5: Dashboard improvement where the sidebar does not fully disappear

Besides the dashboard, many efforts have been done in order to polish small details and improve feedback. Some of this efforts include:

- Adding sweetalert[24] in case of loading successfully or when there is an error in order to provide a better feedback.



Do you have a model? ○

Format to read: read.csv

Argument: header

Enter value: TRUE

Browse file: Browse... Cyprus1.csv

Upload complete

Microbiological Data

Search:

Success!!

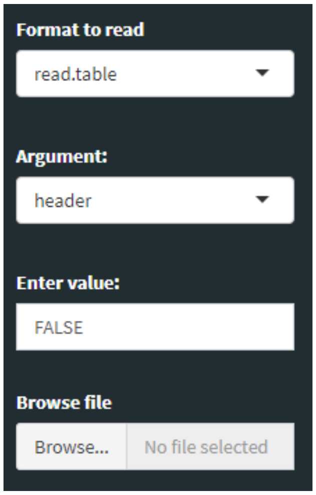
File was load successfully

X	FC						FRNAPH.I	FRNAPH.II	FRNAPH.III	FRNAPH.IV	R
All	All						All	All	All	All	
Season	Season						Season	Season	Season	Season	
Undefined season	Undefined season						Undefined season	Undefined season	Undefined season	Undefined season	
P1-HM06	1.70e+07						9.00e+00	6.10e+01	3.00e+01	0.00e+00	
P1-HM07	9.70e+06						7.00e+00	7.00e+01	2.20e+01	1.00e+00	
P1-HM08	9.20e+06						2.00e+00	5.90e+01	3.80e+01	1.00e+00	
P1-HM09	2.40e+06	1.70e+06	4.00e+05	1.50e+07	9.20E+05	8.30e+05	2.00e+00	6.30e+01	3.50e+01	0.00e+00	
P1-HM12	3.40e+07	4.50e+05	3.20e+05	9.00e+06	1.30E+06	1.20e+06	0.00e+00	9.20e+01	7.00e+00	1.00e+00	
P1-HM13	5.20e+06	3.50e+05	1.00e+05	2.60e+06	5.40E+05	4.00e+05	0.00e+00	9.10e+01	9.00e+00	0.00e+00	
P1-HM14	1.20e+07	8.20e+05	2.50e+05	3.10e+06	6.20E+05	4.50e+05	5.00e+00	4.10e+01	5.10e+01	2.00e+00	

Figure 6: Dashboard created

- Hiding the menuItems of Modelling and Visualization once the "Generate BIGMATRIX button" from Scenario menuItem is clicked.
- When loading a location, conditionally hiding persistence in case that we want to introduce our own.
- Providing help and information about the platform. The user will be able to download a manual in the help icon from the top right bar and will be able to read information about the participants of the platform in the "About" section placed next to help icon.
- Using libraries such as shinyWidgets[11] that make our website more aesthetic and intuitive.
- In order to predict, at the moment of choosing which variables, we create the taxonomy dynamically considering the input of the table. So if in the dataset introduced there are cats, dogs and cows, in the list will only appear cats, dogs and cows. This improves UX since only have to choose among relevant ones.

9.3 Load any given data



Format to read

read.table

Argument:

header

Enter value:

FALSE

Browse file

Browse... No file selected

Currently our platform supports all kind of data. However, is specially designed to support .csv and excel files. The reading function has many arguments in order to introduce all kind of data given. It can be complicated but is the only way to be able to offer such a genericity. The inputs are:

- **Format to read:** Allows choosing which R reading function to use.
- **Argument:** Read function has multiple parameters such as header, sep, quote, dec... Argument provides them to the user to be able to read any input data.
- **Enter value:** For each value, reactively shows its possible values.
- **Browse file:** To choose which file to upload

9.4 Table

Allowing the user to load a dataset, visualize, modify and in general, being able to interact with it, is one of the most important points in the website part.

There are some R packages that cover our requirements but each one differs in some aspects. Because of this, an analysis to detect which one is the most suitable and a proof of concept for each has been performed.

- **DT:[10]** Provides an R interface to the DataTables library in Javascript. DT is a library supported by RStudio.
- **RHandsontable:[9]** Provides an R interface to Handsontable in Javascript.
- **D3TableFilter:[8]** Based on D3 Javascript library and Max Guglielmi's "HTML Table Filter Generator".

A table comparing each of the packages is shown:

Feature	**DT**	**RHandsontable**	**D3TableFilter**
Coloring	✓	✓	✓
Comments	X	✓	X
Customization	✓	✓	✓
Display columns	✓	X	✓
Download	✓	X	X
Edit data	✓	✓	✓
Filter by column	✓	✓	✓
Formatting	✓	✓	✓
Last modification*	13th July 2017	1st September 2017	2nd December 2017
Popularity**	242	172	51
Scrolling data	✓	✓	✓
Search data	✓	✓	✓
Sorting	✓	✓	✓
Table footer	✓	X	✓
User experience***	✓	X	X
Validation	X	✓	✓

Table 18: Comparison among different table packages.

* Last modification on the github repository (Data taken on 15th December 2017).
** Github stars on the repository.
*** Personal opinion once implemented with each library an approach.

As can be seen by the table, these three libraries have a lot in common in a theoretical point of view. For this reason, choosing which one was the best, required prototypes in order to be able to decide.

On one hand, RHandsontable and D3TableFilter are packages that as DT, adapt very powerful javascript libraries into shiny but this task is done by a single author without any remuneration and are not as popular as DT. This implies that can turn into unmantained repositories very easily and that facing problems in RHandsontable or D3TableFilter can lead to that the problem has no solution or that there is not community that can help.

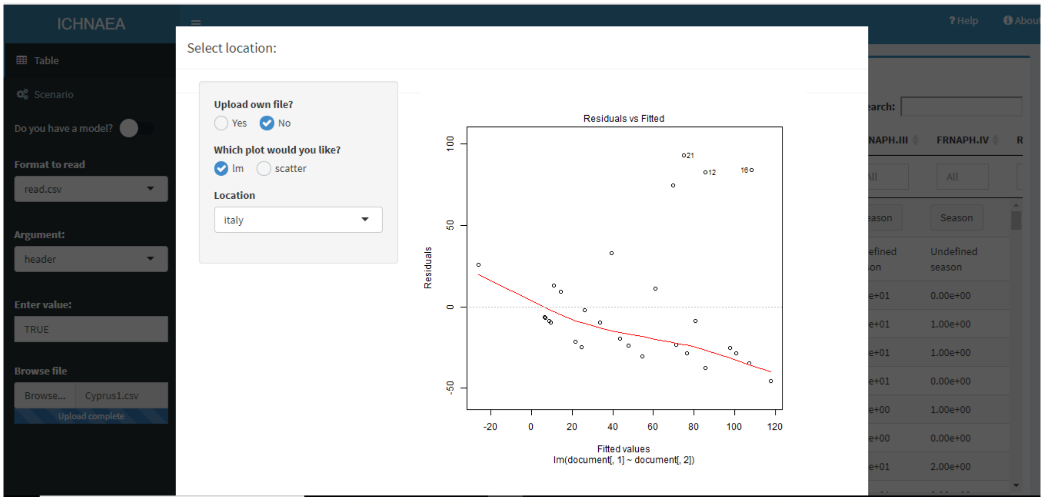
On the other hand DT is not as powerful as both previously mentioned but it has RStudio team behind, it has a bigger community and apparently, a better future scalability. However, DT allows us to build everything required and is the one with the best UX in my opinion (simpler visual appearance).

In conclusion, DT package was chosen because of a better UX, popularity and because the development team is bigger than in RHandsontable.Regarding to D3TableFilter it will always be better a library that can be permanently supported. However, the whole code does not rely on the use of DT and changing the library would only suppose to modify one file.

9.5 Selection of the location

One of the website's requirements is to be able to select a location for each column of our dataset. Generating a solution that could be extensible and independent to any dataset was complicated and hard to design.

To accomplish this task, what has been implemented is a row of buttons (each with a different ID) that open a dialog where the user, can see and select the locations available for the selected column and save the one (s)he wants. Once selected the location, automatically and re-actively Shiny will do the desired plot of all the seasons of the selected location (the user can choose which function to use among scatter and lm).



Select location:

Upload own file? ○ Yes ☑ No

Which plot would you like? ☑ lm ○ scatter

Location: italy

Do you have a model?

Format to read: read.csv

Argument: header

Enter value: TRUE

Browse file: Browse... Cyprus1.csv

Upload complete

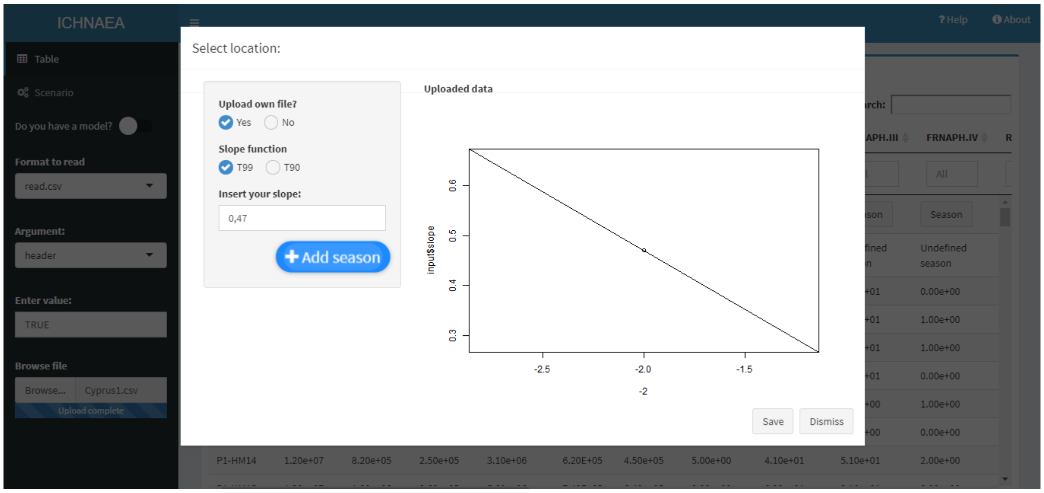
Figure 7: Loading location screen

In order to implement it, we had to know which row we were clicking, how to save the data and how to organize the information. We organized the information by having for each column a folder and inside this one, a folder for each location with many seasons inside.

9.5.1 Upload location

In case that the user wants to upload his/her own file instead of loading them, (s)he can upload a file in two different formats (T90 - time for 90% decay or T99 - time to decay two logarithms).

Since T90 and T99 require of only one number, instead of uploading a file, we have decided that would be more convenient to introduce the number and re-actively show the slope.



Select location:

Upload own file? ● Yes ○ No

Slope function ● T99 ○ T90

Insert your slope: 0,47

Format to read: read.csv

Argument: header

Enter value: TRUE

Browse file: Cyprus1.csv

P1-HM14 1.20e+07 8.20e+05 2.50e+05 3.10e+06 6.20E+05 4.50e+05 5.00e+00 4.10e+01 5.10e+01 2.00e+00

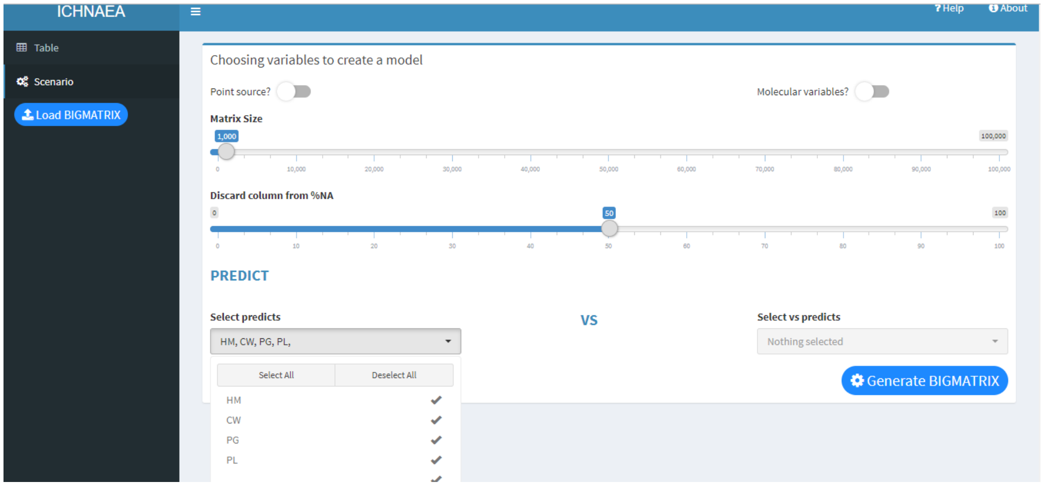
Figure 8: Upload location screen

9.6 Efforts in BIGMATRIX's creation

The generation of the BIGMATRIX is mostly machine learning based. Nevertheless, we have worked in creating a website that given any loaded dataset can generate this matrix and can apply machine learning techniques to it.

To generate this matrix we need to assign first a location to each variable in order to then create the matrix and apply aging (we have to see the value in the slope at a certain moment). These variables, are the ones that can variate across different datasets and because of this, an effort in creating this generic modeling has been done.

Aside from the creation of this matrix, there is all the environment required in order to select which machine learning technique will be applied and among many options, what to predict (e.g. human vs no human).



Choosing variables to create a model

Point source? ○

Molecular variables? ○

Matrix Size: 1,000 (100,000)

Discard column from %NA: 0 – 50 (100)

PREDICT

Select predicts: HM, CW, PG, PL,

VS

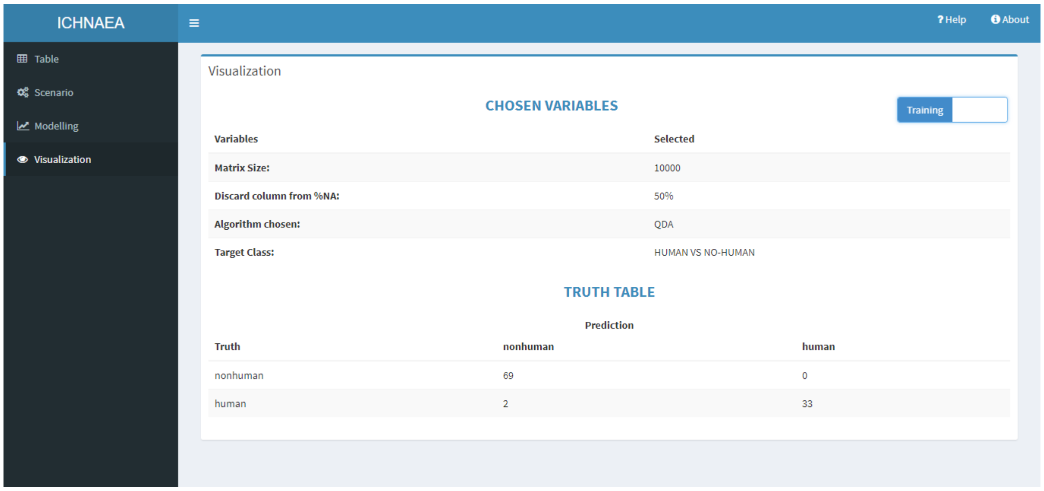
Select vs predicts: Nothing selected

HM ✓
CW ✓
PG ✓
PL ✓

Figure 9: Dashboard created

9.7 Visualization

In the last screen, the user can visualize the chosen filters and also can see a truth table showing the model predictions to the data. There is the option between seeing the truth table of training and also in the test case in order to see if both models are good enough.



Visualization

CHOSEN VARIABLES

Variables	Selected
Matrix Size:	10000
Discard column from %NA:	50%
Algorithm chosen:	QDA
Target Class:	HUMAN VS NO-HUMAN

TRUTH TABLE

	Prediction	
Truth	nonhuman	human
nonhuman	69	0
human	2	33

Figure 10: Visualization screen

9.8 Integration with Science Nodes

As previously mentioned, integrating with Science Nodes allows us to abstract the user from executing R and provides us from an environment where we can have users and security(things that otherwise, we would not be able to offer). These users can interact with others execution in this case, with our solution.

In order to integrate with Science Nodes, we had to mainly do the following steps:

- Modify local code in order to install the packages (install.packages(...)) that are being used. These instructions were not in my code, since my IDE is the one that asks for them.
- Change routes that we could have in the code in order now to fit with the new paths of each file once uploaded.
- Go to "Files" tab and upload all the files one by one and categorize them as: scripts, documents, media, warehouse, datasets or inputs.
- Go to "Solutions" tab where we choose the technology of our solution (Shiny), we select the files with the code and we create the solution.

- Run the solution. When running the solution a screen will prompt and we have to name the experiment, select input files and parameters and finally execute.
- Finally we can manage access to other users to our project from the "Projects" tab where we can choose between project manager and enrolled. In the "Nodes" tab we can choose if making our node accessible to everyone or to make it private.

Integration with this platform, has been specially tough since there is no documentation but could be possible thanks to Gerard Cegarra, student that developed the newest version of this platform as his final degree project (currently not available).

10 Design and implementation of the prediction

10.1 A brief introduction to machine learning

Machine learning is a computer science field that gives computers the ability to learn without being explicitly programmed. This is specially useful in cases where:

- We cannot directly write a computer program to solve our problem.
- The problem changes across the time or depends on a certain environment which would imply to explicitly write a solution per case.

In these cases, machine learning can provide a solution because uses data or past experience to optimize a performance criterion. We model the problem according to parameters and learning is the execution of a computer program to optimize the parameters (training). This predictive model is able to make predictions in the future with data or examples that has never seen and without the need of having to code a special case.

In training, we need efficient algorithms to solve the optimization problem, and to store and process big amounts of data. Once a model is learned, its representation and algorithmic solution for inference needs to be efficient as well. In some machine learning applications, the efficiency of the learning can be as important as its predictive accuracy.

I want to make note that previously to training a model, there is a significant part where we have to pre-process the data in order to extract from it useful information and also to eliminate data that is not good enough e.g significant number of NA's.

In machine learning we can establish multiple classifications depending on how the learning process is done:

- **Supervised learning:** The computer receives example inputs and their desired outputs. The problem then is learning a general rule that can map inputs to outputs.
- **Unsupervised learning:** The computer has no labels provided on the input, leaving it to the learning system to find the input structure in its own.

There are many algorithms to solve machine learning problems. Depending on the problem one or the other will have a more successful prediction rate. Some of this algorithms (among many others) are decision trees, neural networks, deep learning, support vector machines (SVM)...

10.2 Implementation

As could be seen in the previous section, machine learning is specially interesting for cases where we cannot write a solution that will prevail across the time.

In our case, the input data can be any kind of microbiological data that the user introduces. Having a solution that is adhoc per data, would mean having serious scalability problems.

10.2.1 Starting point

To implement our machine learning solution that comprehends all kinds of microbiological data, and that can give predictions with them, we have taken as a starting point a code from my director that has a 100% prediction rate in a certain scenario. Unfortunately, this solution is bounded to the data given and because of the nature of our problem, it can only be partially re-used since the user can introduce any data.

The machine learning solution given by my director, could be summarized in seven steps:

- **1 - Pre-processing:** In this first step what my director does is reading the dataset, eliminating columns with lack of data, substitutes values with the form "<X" for a 0, creates target classes (class for Human/-non Human) and creates seasons according to the Date value in the dataset.
- **2 - Handling detections:** When we dilute or age a value, if it goes below the Tau for the variable, then we make it 0. This Tau should be the same for all the observations for the same variable and it is obtained by calculating the slope from the T90 (T90 = -1/value) given by the scientists.
- **3 - Aging:** In case that not given T90, computing logistic regressions for aging essays on seasons (assuming no correction) to get slope and increment. Otherwise, calculate slope/intercepts for non-assays variables where T90 or T99 given.
- 4 - Remove animals with few occurrences, create ratios, apply the logarithm to the matrix.
- **5 - Scenarios 1 - 4:** Training scenarios where he tries different approaches in order to see which is the one with the best success predict rate with the minimum number of variables. He tries among:
  - Human/noHuman - Point Source - All variables.
  - Human/noHuman - Point Source - Mollecular variables.
  - 4 Sources (cow, human, pig and poultry) - Point source - All variables.
  - 4 Sources - Point source - Molleculars.
- **6 - BIGMATRIX:** BIGMATRIX's creation, age dilution, applies detectors, defines matrix's size, add ratios after aging and apply logs,

creates two versions: one with all variables and the second with molecular variables only.

- **7 - Scenarios 5 - 8:** Repeats scenarios 1 - 4 but with the BIGMATRIX.

10.2.2 Integration

In the following section we are going to explain how we could integrate the previous explained code in order to make it work in comparison with the provided code.

10.2.2.1 General overview

The code to integrate and my case, are in general terms the same but in code and purpose is very different. In my director's code, the general purpose is to test which is the best approach among many others that he executes indifferently if he has already one model that is perfect. In my code, the aim of the code is to allow the user do whatever (s)he wants in order to try models that (s)he may consider interesting.

Because of this, in the platform there are many variables that the user can choose in order to make it feel like (s)he is almost programming with the platform but without having to know or install R and its libraries.

Furthermore, the way to get the slopes and seasons is slightly different since we need to assign and upload seasons to each column before generating the BIGMATRIX. We can directly have an array with the slope of each column where each position is saved when we save the season to that column.

10.2.2.2 Pre-processing datasets

Previously, eliminating undesired columns was done by handpick e.g considering that a certain column has too few data to be relevant. In a generic approach, this can not be done by code since we do not know which data we are going to receive. In order to be able to eliminate undesired columns, we provide the user the possibility of eliminating columns with a threshold below a certain percentage of missing data (NA).

Creating target classes can be reused but would imply to standardize the way the user writes data the sample column (name of each row). In this case, we have assumed that the user will place the animal code in the 4th and 5th letter of the sample name i.e P1-**CW**10 where "CW" stands for Cow.

Data with values such as "<X" where X is a given number, are also eliminated in order to be able to operate with them. These values can be considered 0 since microbiologically talking what means "<X" is that in the sample used, there is less than a certain quantity but because of its size and that are not uniformly present in water, we can not assure the quantity.

Note that in order not to modify user's matrix, we are going to do the preprocessing part with a copy of it. This allows us to create as many columns as desired and apply any modification to it without affecting the user.

10.2.2.3 Generation of the BIGMATRIX

BIGMATRIX creation is done in a similar way than my director does. In order to generate the BIGMATRIX we create a function to age the matrix. This function what does is generating new rows with the value of each variable under a certain moment given a time. Once we have this aged matrix, we apply another function that dilutes this matrix (diluting is just dividing values by a factor). While this two methods are applied, if some value goes under a certain tau the value becomes 0. The tau is the same for all observations of the same variable. Next step consists in given the number of elements chosen by the user, generating a sample of that size. The resulting matrix, will be the BIGMATRIX in other words, the result of generating random samples and permutations of our dataset once applied the aging and dilution process.

This matrix, is the one that will verify how successful is the model we have built since that a model in train has a high successful rate doe not imply that will in test.

10.2.2.4 Training and test

When generating a predictive model, my director does an exhaustive search to find which is the best set of variables and then he applies the algorithm. In our case since is an interactive website, we execute what the user has chosen. What we do is allowing the user to choose which are the variables that (s)he wants to create a model of.

The current algorithms that the user can apply are QDA and LDA since are the ones that we know that work fine with the type of data.

10.3 Results

In order to verify that our platform works, we have executed the same matrix used in the integrated code. If the platform works appropriately, it has to give same results as the integrated code. Because of this, we have tested (among other tests done) that with the variables (SOMCPH/BTHPH and SOMCPH) has a 100% in the training part. Furthermore, tests with another matrix (with different variables and data) have been done and successfully worked. The prediction rate in this case is not relevant since the purpose of the system is predicting whatever the user wants which is very different than having an input data and once calculated which is the best approach give a prediction rate. Note that is a very demanding process that can take long.

11 Conclusions

11.1 Knowledge acquired

During the project I have learned about R, Shiny, machine learning and integration with platforms such as Science Nodes. Furthermore, I have seen that elaborating an economic plan, analyzing the stakeholders, defining which technologies to use and planning a project, is a tough task that suffers from many deviations and requires an exhaustive work in order to do it well.

I have learned that building a UI that provides the user of a easy interaction can be very complex and also, I have seen how hard it is designing a website flow when the client does not know exactly how (s)he wants it to be and it is complex to find an optimal representation.

Going more in depth with the technological part in the website, I have learned how building websites with Shiny is, how Shiny reactivity concept works, the potential and strong and weak points of Shiny, the many libraries that one can use and which ones are the best, and how hard is to build a solid product when it has to be very scalable and used in real life.

Related with machine learning I have seen the difficulty that implies working in a domain (microbiology) where you do not have knowledge. I have also experienced machine learning potential and I could overview concepts and techniques previously studied. I have also strengthen my R knowledge in machine learning by developing methods and functions in order to successfully predict and create a model.

11.2 Project results

To analyze our results, we have to see if our objectives have been successfully completed or not.

- **Machine Learning part**
  - Aggregate the work of my director to the project to accomplish the machine learning prediction in the case of only one source. (100%)

  - If possible create the machine learning part in order to be able to predict with more than one source. (30%)

- **Website part**
  - Allow the user to upload, load, modify, and download microbiological information. (100%)
  - Allow the user to choose which information to display, the microbiological behaviour of each variable, the dissolution, creation of a prediction model, loading a previous prediction model. (100%)
  - Visualize and interact with the results. Display relevant information in order to facilitate user decisions and understanding of the problematic.(100%)

As it can be seen in the project, most of the objectives have been successfully completed. However, I would like to mention that there has been a bigger emphasis on the website rather than the machine learning part unlike it was stated at first. This is something that happened while developing the project because of underestimation of the website complexity. Building such a solid website that provides the user all the functionalities in order to save and load in different states of the process, creating a solid UX and UI, being able to select each location while seeing in the plot of each different location, integrating with a platform, and designing how to provide the user with the possibility of doing all this process, has been something really tough. Nevertheless, since I am passionate about machine learning, I have adapted the code of my director to a functional code that can execute what the user wants and needs in a way that interaction between user and the platform is as user friendly as possible by diminishing the waiting time.

12 Future work

In order to keep progressing with the platform, there is still work to do:

- Maintain the website and correct possible bugs that may appear in the future.

- Apply automated tests into the platform in order to be able to scale solid and successfully
- If the number of users grows, it may be convenient to move into a shinyServer which would imply to control users, assure data, etc.
- Automatically deploy new versions to science nodes. Currently this cannot be done and it implies that whenever there is an upload, the user has to upload all the modified files to the science nodes server.
- Improve table functionalities in order to facilitate user's work. The more excel-like is the table, the better and more useful it will be. Some of these functionalities could be:
  - Allow the user to create sub-matrices.
  - Create ratios between columns.
- Parallelism in the matrix's creation in order not to make the user wait when generating the BIGMATRIX.
- Add more machine learning algorithms.
- Try to improve machine learning performance.
- Improve visualization part.

However, the platform is published via web and can be used without major issues and it has the client's approval.

A
R Packages used

The packages used and required to make the website work are the following:

- dplyr
- htmlwidgets
- scales
- shiny
- shinyBS
- shinyWidgets
- shinycssloaders
- shinydashboard
- shinyjs

References

[1] R webpage
https://www.r-project.org/about.html

[2] RStudio webpage
https://www.rstudio.com

[3] Shiny webpage
https://shiny.rstudio.com/

[4] Documentacion Github.
https://en.wikipedia.org/wiki/GitHubcite$_{n}ote-hugeinvestment-3$

[5] Chrome website
https://www.google.es/chrome/browser/desktop/index.html

[6] Universitat de Barcelona microbiological department website
http://www.ub.edu/microbiologia/

[7] Article talking about fecal matter contamination and the risks
http://www.lavanguardia.com/vida/20140702/54411472272/contaminacion-fecal-comun-playas.html

[8] D3TableFilter package
https://github.com/ThomasSiegmund/D3TableFilter

[9] RHandsontable package
https://jrowen.github.io/rhandsontable/

[10] DT package
https://rstudio.github.io/DT

[11] ShinyWidgets package
https://github.com/dreamRs/shinyWidgets

[12] Importance of stakeholders identifying
http://smallbusiness.chron.com/importance-identifying-stakeholders-project-74730.html

[13] Latex documentation
http://www.latex-project.org/

[14] Sharelatex documentation
https://es.sharelatex.com/

[15] Lluís A. Belanche Muñoz personal website
https://www.cs.upc.edu/ belanche/

[16] InLab FIB website
https://inlab.fib.upc.edu/

[17] Machine Learning subject
https://www.fib.upc.edu/en/studies/bachelors-degrees/bachelor-degree-informatics-engineering/curriculum/syllabus/APA

[18] Github webpage
https://github.com/

[19] Git webpage
https://git-scm.com/

[20] Science nodes webpage
http://science.cs.upc.edu/

[21] Pinterest webpage
https://www.pinterest.es

[22] RSelenium webpage
https://cran.r-project.org/web/packages/RSelenium/

[23] RUnit webpage
https://cran.rstudio.com/web/packages/RUnit/index.html

[24] SweetAlert documentation
https://sweetalert.js.org/guides/

[25] Shinydashboard documentation
https://rstudio.github.io/shinydashboard/

[26] R. Blanch, A., Belanche-Muñoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kühn, I., Moce, L., Muniesa, M., Schwartzbrod, J., Skraber, S., Papageorgiou, G., D. Taylor, H., Wallis, J. and Jofre, J. (2004). Tracking the origin of faecal pollution in surface water: an ongoing project within the European Union research programme. Journal of Water and Health, 02.4, pp.249-260.
http://jwh.iwaponline.com/content/ppiwajwh/2/4/249.full.pdf

[27] Blanch, A., Belanche-Munoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kuhn, I., Moce, L., Muniesa, M., Schwartzbrod, J., Skraber, S., Papageorgiou, G., Taylor, H., Wallis, J. and Jofre, J. (2006). Integrated Analysis of Established and Novel Microbial and Chemical Methods for Microbial Source Tracking. Applied and Environmental Microbiology, 72(9), pp.5915-5926.
http://aem.asm.org/content/72/9/5915.shortcited-by

[28] Belanche-Muñoz, L. and Blanch, A. (2008). Machine learning methods for microbial source tracking. Environmental Modelling Software, 23(6), pp.741-750.
http://www.sciencedirect.com/science/article/pii/S1364815207001818

[29] Balleste, E., Bonjoch, X., Belanche, L. and Blanch, A. (2010). Molecular Indicators Used in the Development of Predictive Models for Microbial Source Tracking. Applied and Environmental Microbiology, 76(6), pp.1789-1795.
http://aem.asm.org/content/76/6/1789.full.pdf+html

[30] Casanovas-Massana, A., Gómez-Doñate, M., Sánchez, D., Belanche-Muñoz, L., Muniesa, M. and Blanch, A. (2015). Predicting fecal sources in waters with diverse pollution loads using general and molecular host-specific indicators and applying machine learning methods. Journal of Environmental Management, 151, pp.317-325.
http://bit.ly/2A6GUzI

[31] Alpaydin, Ethem. Introduction to machine learning 2nd ed.

[32] Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn* C Bishop - Springer, New York, 2007