Bivalve's Growth Conditions in Coastal Ecosystems: A Decision Tree based Analysis

João Pedro Reis[1], António Pereira[1,2]
[1]University of Porto – Faculty of Engineering – DEI
[2]University of Porto - LIACC
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{ei07119, amcp}@fe.up.pt

Luís Paulo Reis[2,3]
[3]EEUM – School of Engineering, University of Minho – DSI
Campus de Azurém, 4800-058 Guimarães, Portugal
lpreis@dsi.uminho.pt

Abstract—**Simulation is a very useful tool to gather new information about an implemented model, because it can run artificial environments instead of putting in risk some entities that are influenced in the real process. The simulation of physical, chemical and biological processes in coastal ecosystems is used as a way to understand the system internal dynamics and to predict its evolution over time, in order to promote behaviors environmentally friendly and to induce effective and efficient management of the ecosystem as a whole. However, there are several ways of translating and interpreting the data provided by the simulation such as applying appropriate data mining models. This paper describes an approach that uses a Decision Tree model to produce intuitive information about the influence of several environmental variables on the growth conditions of bivalve species within an aquaculture exploration in a coastal ecosystem. This information is captured by relating the values of simulated variables, like water temperature or organic matter, with the length of the bivalve's shell, extrapolating information about the organic or physical conditions that increase or decrease the growth of the bivalve species, and guiding the stakeholders to locations for the best practice of the seeding process.**

Keywords – Decision Trees, Patterns, Data Preparation, Ecological Behavior.

I. INTRODUCTION

Decision Tree based methods are widely used in data mining and decision support applications [8]. The integration of Decision Tree approach with a well preprocessed dataset can extrapolate some relevant information from data, independently from the domain. This type of approach can reveal a wide spectrum of correlated variables that can produce rules about how a specific classification is formed. Applied to the specific domain of coastal ecosystems, it can be extrapolated from the tree-like graph the most significant variables that contribute to a good bivalve growth. The conditions of growth assessed are referent to Hydrodynamic and Biogeochemical modeling, which were provided by the results of EcoSimNet framework, in a dataset structure. EcoSimNet is a platform for simulation and support decision-making [9]. The evaluated dataset was generated taking into account the model of Sango Bay in China, which is a realist environment for simulation and posterior analysis [3]. The understanding of coastal ecosystems complexity is a strong reason to believe that this approach can change positively the way how scientific community sees the interacting organic variables in the simulation model.

The lagoon of the ecosystem is modeled as a two-dimensional vertically integrated, coupled hydrodynamic biogeochemical model, based on a finite difference bathymetric staggered grid with 1120 cells (35 lines by 32 columns) and a spatial resolution of 500m [3, 9]. The simulated processes in the model are built with partial differential equations and the dataset used is structured by attributes that describe different species, organic matter, position in the seeded zones, water temperature, etc. The area used for aquaculture is spread by 352 cells and the bivalve's growth cycle is simulated during one year and a half, generating a dataset with more than 800 thousands instances, forcing a preprocessing, making the data adequate to the main purpose of this project. It is expected that Decision Tree based approach, and a well preprocessed dataset, have the capability to provide a set of constraints that restricts the simulated environment to a set of conditions, in order to understand its influence on the bivalves' growth behavior.

This paper shows, in section II, the state of the art related with the approaches used in this project. Section III presents the necessary preprocessing to prepare the dataset, describing the different phases that compose it. Section IV refers the implementation phase, with the purpose to provide the final dataset used by C4.5 algorithm Decision Tree appliance. Section V explains the experiments made, seeking the best values to apply in its parameters. Section VI shows the obtained results, followed by section VII, in which the results are discussed reaching some conclusions and future work about the approach used and the attributes' relations.

II. STATE OF THE ART

Nowadays there is an increasing interest in scientific research to understand the biological environments, namely ecosystems. Most of the conventional ecological models are translated in partial differential equations, based on physical processes following the conservation laws of mass, momentum and energy [2]. Due to the limited understanding of ecosystem processes and availability of sufficient monitoring data, these models are rough simplifications of the reality. However they have been fundamental tools in the progress of ecological research. The widely recognition that many mechanisms of ecosystem dynamics are still unclear, non-linear, complex, and more qualitative than quantitative, makes it difficult to integrate those mechanisms in the traditional models [7].

During the last decade methods that combine incomplete knowledge and data were developed and applied to ecological

models to surpass the limited data available from *in-situ* measurements, and taking expert knowledge as references [4, 7]. Rule-base approaches, like decision trees, are built on cause-effect relationships, and are not based on mathematical description of the ecological processes. Decision tree is a model technique that splits the parameter domain into sub-domains, and learns the system output of each sub-domain through historical records. A split point is a node and the corresponding output is a leaf that can be split again, resulting in a tree-like modeling structure [2]. The key issue when building a decision tree model is to find the right attributes(s) and optimal splits of parameter domain.

One work that was developed in this field that seems to have consistent results is from Hahsler et al. [5]. The arules R-Package Ecosystem has implemented functionalities that provide a frequent item-sets, association rules and associative classification analysis. The main purpose of item-sets and association rules is to discover relations between dataset attributes, in datasets that have a large number of instances. Besides being a well-researched method, in this work the Decision Tree method is used trying to take advantage of its intuitive representation, making much easier to deduce information from this representation, instead of association rules and item-sets usage, and the flexibility of setting its parameters.

III. DATA PREPARATION

The need of datasets from several sources to be analyzed implies following some essential steps. In this specific problem, an important phase of CRISP-DM [11], which is a Data Mining Process, called *Data Preparation* was used. This phase aims to clean the original data and construct new instances and attributes, due to several issues like missing values, inconsistent data and not sufficient representative data to a specific problem.

Data Cleaning is the first phase of the process and it is important to treat outliers and missing data. The treatment of these specific cases will influence positively the final result, since the corrupted data don't represent the truth that needs to be analyzed. The simulation dataset, like previous said, is composed by a 32 lines by 35 columns, representing each cell a possible seed zone. The final result does not have all the 1120 cells ready to harvest, because there are cells that were not seed and others that are land and boundaries of the ecosystem. Hence, the outliers – data that is not common or expected to be different – could be easily found, and removed. All the not seeded cells were marked with an impossible value to differentiate from all the possible values from seeded cells. The option chosen was to remove the instance, because only cells that are seeded, cells in which occurs a variance of the parameters model, are relevant to the final result.

Relatively to the missing data, all the instances that have corrupted or empty information were removed. Like is known, the simulation was executed in 731 iterations, being this amount of instances per cells sufficient to justify their removal.

In order to select the most appropriate data for our domain, the Subset Selection Problem was resolved by an attribute variance analysis, which was based on choosing the attributes that have a number of sufficient different instances, being, only this way, possible to deduce some information from its variance. This selection promotes an improvement of C4.5 performance on domains with continuous variables, which is the case of Modeling and Simulation [6].

Entropy may be informally defined as the measure of impurity in a group example. It is maximum when we cannot predict nothing from the data – the probability of choosing an example in a group is the same – and it is minimum when we can say for sure that a certain data will be chosen – the probability of choosing an example is 1 (only one type of data in the group). This concept is important, because the several data regarding the dataset have a high level of entropy, taking into account all the attributes and useful instances.

A big slice of dataset descendant from EcoSimNet simulation platform was eliminated (results from cells that were not seeded). This reduces significantly the dataset size, improving the efficiency of the work, and avoids a higher value of entropy in the following steps.

After this step, we have to be aware of the attributes that are important to achieve the main final purpose. The initial dataset has the following attributes, excluding the time step, position, and species information: Box depth; Dynamic height; U Velocity; V Velocity; Salinity; DIN (Dissolved Inorganic Nitrogen); Phytoplankton biomass; POM (Particulate Organic Matter); TPM (Total Particulate Matter); Water temperature; Zooplankton biomass; Boundary NO3 concentration (Indicator of Water Quality); Boundary POM concentration; Boundary SPM (Suspended Particulate Matter) concentration; Boundary Zoo concentration.

After an analysis phase, in which we select the attributes that don't vary constantly, the final selection attributes are the following: U and V Velocity (North-South and East-West water velocity components), DIN, Phytoplankton biomass, Water temperature and Zooplankton biomass.

The last step of this phase separates the species information for a further independent treatment. In the EcoSimNet simulations, it was used three types of species, being Chlamys Farreri (Scallop), Crassostrea Gigas (Oyster) and Laminaria Japonica (Algae).

When the *Data Preparation* phase is concluded, there are some important questions to make, regarding the quality of the information and if the main goal of the problem could be achieved. To this specific problem, the main questions that should be made are: *Could the behavior growth be improved? Are there some strange behaviors? Have the bivalve growths the same pattern?* And the final question is: *What could be done?*

Fig. 1 is a representation sample of the growth behavior of Crassostrea Gigas (Oyster), from seed to harvest season, each line representing a single cell of the seeded grid simulated. As can be seen, the growth of this species has some irregularities that cannot be immediately understood. Irregularities like variance in growth behavior of different seeded cells provoke different times to harvest, and not a well-defined quantity per harvest. One of the purposes of this of work is to answer the questions above, with real information that can be understood

by a regular person, provided by very intuitive and easy representation methods.

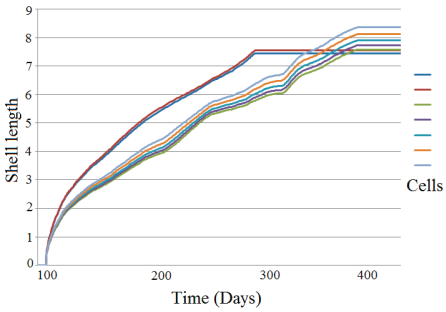

Figure 1. Crassostrea Gigas Growth Behavior

After the Growth Analysis phase, only Chlamys Farreri and Crassostrea Gigas will be used to generate its correspondent Decision Tree, due to the fact of Laminaria Japonica species doesn't have any unusual and specific conditions that promote positively or negatively its growth, being its growth behavior very similar between all the seeded cells.

The quality measure sub-section is intended to establish a value that should separate the good growth from the bad growth. This type of approach is needed in a way that could be implemented in the final phase of the work, telling the user if in certain attribute circumstances, the growth was good or not. This result describes an attribute that is capable to classify any instance, allowing the comparison between growth conditions.

To implement this method, it was calculated a derivative of the growth function of each species. These values are very important because they show the raise of growth in a single day of simulation. Fig. 2 presents a graph with the derivative calculation, from the whole production season in one seeded cell. The important question inherent in this approach is: *Which threshold represents better a quality measure?*

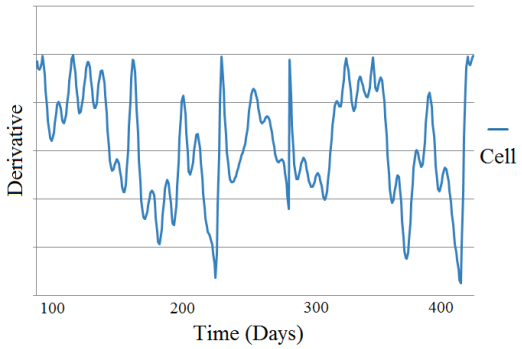

Figure 2. Crassostrea Gigas Derivative Growth Behavior

It is known that having fewer values in both measures - number of instances labeled as good or bad – couldn't ensure a high level of confidence in Decision Tree. This is due to the fact of not being relevant to the amount of counter-example. For example, in the dataset that is analyzed, if the number of good growth labeled instances are 3, and the number of bad growth labeled instances are 1000, it is natural that the Decision Tree couldn't induce an attribute condition that describes the good growth. So, the *Threshold* is a value that separates the derivative function of each growth species, into a good or bad label. This value will be discussed in the following sections.

IV. IMPLEMENTATION

In this section it will be discussed the developed application to provide a new dataset that was used to reach the final results, and the C4.5 algorithm used. This developed tool is essential because it filters some non-important instances from the initial dataset and organizes the information into the different species studied. To run the C4.5 algorithm efficiently, the final dataset should serve data in a way that facilitates the use of C4.5 algorithm parameters.

The application developed was written in Java Programming Language, implementing both *Clean*, *Format* and *Construct Data* phases, turning them into an automatic process, composed by three different phases. First, the application removes the instances that have missing values and outliers, making the data more viable and consistent for use. Like said before, this phase consists into only gather the information of cells that are seeded and the iterations that provide some attribute variation. Secondly, this filtered information was separated by species, creating different files for each species. Each file is composed by the two dimension cell location (e.g. 22, 10), and its growth during the simulation time. Those files could be used to analyze how each species grows, and have a specific view of its normal behavior, knowing when each cell was seeded and harvested.

The final phase is based on picking all the cells information of each species, and calculating the labels for every individual cell, iteration by iteration. This *Quality Measure* is associated for further use of Decision Trees, being the leafs this *Quality Measure* label, and the nodes, its correspondent attributes. The application has two parameters that are used to provide a simplest and equilibrated dataset, which are *Threshold* and *Delta*. The *Threshold* is the value that separates a good growth from a bad growth label. This value should be chosen taking into account the number of instances with different labels. In this specific problem, the option was a balanced separation between those two marks, to increase the confidence in the results. The *Delta* parameter is the distance between iterations in which the derivative was calculated. Hence, the derivative is calculated with the difference between the growth values divided by this *Delta* factor. The lower value chosen to the *Delta* parameter, the more reliable information is obtained, but, in other hand, more information is produced increasing the final size of dataset. For this problem, the minimal value for *Delta* was taken, being this value 1.

The C4.5 algorithm is used to reach the best possible results, being improved and developed with different versions along time. One of the motivations for the usage of this algorithm is its appropriate fit relatively to treatment of

continuous attributes [10], which is precisely the type of attributes generated from the EcoSimNet simulation.

Relatively to the C4.5 algorithm parameters, the criterion used was the Ratio Gain because this type of approach is more adequate, comparing to Information Gain, due to the fact of dealing with high different values per attribute, normally found in continuous variables [12]. In other words, if attribute have a large number of different instances, like numeric values, the *Information Gain* approach could be biased, and over-fitting could occur (selection of non-optimal attribute for prediction). The *Gain Ratio* is a based *Information Gain* method that takes into account the number of instances an attribute contains, reducing the bias on high-branch attributes. The *Information Gain* is a based *Entropy* method that takes into account the lower value of entropy, high information gain value, to choose the root of the calculated tree.

This type of criterion is used to calculate the root of a tree that maximizes the *Ratio Gain*. Hence, the final tree is the result of an iterative process that calculates the next attribute to use, taking into account the previous one. So, while the tree is constructed, the number of instances is reduced due to the fact of previous attributes limitations and leafs produced. Hence, with this parameters it is possible to modulate a consistent *Decision Tree* that fulfills the goals of this approach.

V. Experiments

To run these experiments, *RapidMiner 5* framework was used, because it has the possibility to create Data Mining models, and it includes an implementation of different algorithms, e.g., Decision Tree, Support Vector Machine, Artificial Neural Network based models, being the most used Data Mining framework tool [1].

The parameters tested are: the Minimal Gain (minimum value of Ratio Gain that should occur in an attribute to be chosen for tree expansion), and Minimal Leaf Size (minimum number of instances that a leaf should have in the Decision Tree). The variation of these parameters results in a different number of nodes and leafs. The number of nodes could not be too high due to the difficult understanding of the tree, and also could not be too low due to the difficulty of deduce the attributes relation.

Two different tests were made, one for each bivalve species: Chlamys Farreri and Crassostrea Gigas. Fig. 3 shows the relation between the number of nodes and the minimal leaf size for the Chlamys Farreri species. It can be seen that minimal leaf size values between 25 and 40 and minimal gain values equal to 0.01 or 0.001 provide the best results. Minimal gain value equal to 0.1 do not produce enough nodes to a good interpretation of the tree, and minimal leaf size values between 10 and 25 produce to many nodes.

Fig. 4 shows the relation between the number of nodes and the minimal leaf size for the Crassostrea Gigas species. The minimal gain value equal to 0.1 didn't produce any nodes, so only the values equal to 0.01 and 0.001 are shown. The minimal leaf size equal to 10, and minimal gain equal to 0.001 produce the best results. All the other tested values did not create enough nodes to a good Decision Tree interpretation.

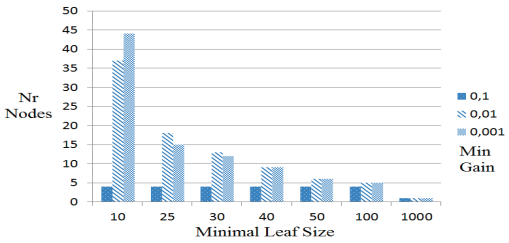

Figure 3. Chlamys Farreri: number of nodes vs. leaf size

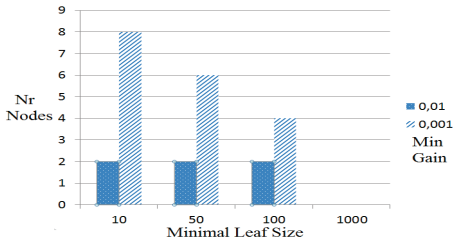

Figure 4. Crassostrea Gigas: number of nodes vs. leaf size

VI. Results

Each species have its own growth information in separated datasets. At each species analysis, it will be presented the relations of dataset labeled attributes, to confront a visual analysis with the computer analysis made, validating the coherence of results using the C4.5 algorithm.

An important observation that have to be made before the tables analysis (Tables I-IV), is that low values of instances per conditions (set of rules that satisfy a certain label: Good or Bad) do not discard the confidence inherent to it. The purpose of this paper is to find the circumstances that influence the bivalve's growth, independently of the number of instances. It is assumed that a good growth will not always occur from seed time to harvest period. Hence, we are looking for the specificity that organic conditions could provide, separating the two labels of associated growth. Another consideration that should be made is relatively to maximum and minimum values, generated by the simulation framework, from each attribute analyzed. The presented tables will not represent these values, so we assume that when a situation like *Zooplankton* < *55.96* and *Zooplankton* > *55.96*, the Zooplankton values should vary between its minimum possible value and 55.96, and between 55.96 and its maximum possible value, respectively.

The visual analysis made from the relation of attributes with labeled instances, as Good or Bad Growth, reveals that some conditions benefit or harm the bivalves' growth. This analysis has two different intentions: make a qualitative evaluation of the most significant attributes, and compare it with the Decision Tree results, confronting the computer results with the dataset visual analysis (dark dots represent Good Growth - light dots represent Bad Growth) – two totally independent analysis.

A. *Chlamys Farreri*

The number of instances that the dataset contains for this species is 15 042, being 7724 (51%) labeled as Bad Growth, and 7318 (49%) labeled as Good Growth. This dataset follows the Java implementation metrics, being: Threshold = 0.015 and the Derivative Factor = 1.

1) *VISUAL ANALYSIS*

From Fig. 5 it could be extrapolated that Water Temperature has a great importance in the bivalve growth, not being relevant the DIN value. We can almost see a line that separates good from bad growth, which is a very good indicator of the attribute importance in the growth conditions. The variation of DIN value have a minimum influence in the bivalve growth, which could not be totally discarded, but is not relevant to this very specific case.

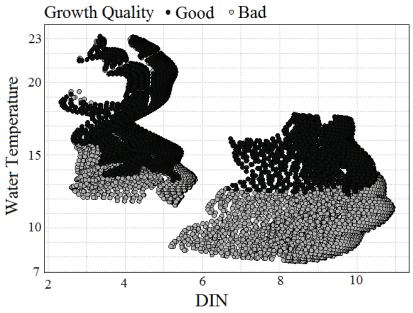

Figure 5. Chlamys Farreri – Water Temperature and DIN

From Fig. 6, we can say that high and very low values of Zooplankton are not good to a good growth. Relatively to U Velocity, it only can be relevant in some specific cases that will be explored in the Computer Analysis phase.

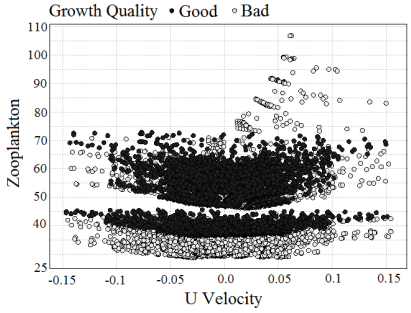

Figure 6. Chlamys Farreri - Zooplankton and U Velocity

2) *COMPUTER ANALYSIS*

Fig. 7 represents the Chlamys Farreri Decision Tree, obtained with the parameters: Minimal Size for Split: 2, Minimal Leaf Size: 40, Minimal Gain: 0.01, Maximal Depth: 20 and Confidence: 0.25.

The analysis of Fig. 7 originated two different tables. One of them tells about good growth conditions, Table I, and the other bad growth conditions, Table II.

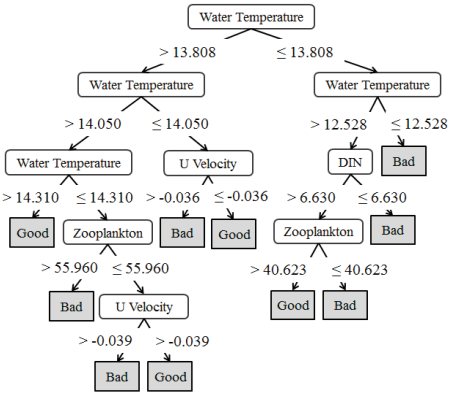

Figure 7. Chlamys Farreri Decision Tree

Regarding Table I, the following conclusions could be made, admitting that the natural organic conditions are satisfied: Chlamys Farreri will have a good growth if Water Temperature is above the 14 degrees Celsius. To ensure this growth, the Zooplankton values should be restricted between 40 and 55 and finally, the DIN value above 6.

TABLE I. GOOD GROWTH: CHLAMYS FARRERI

Leaf	Confidence	Instances	Conditions
Good	100.0	67	Temp ≈ 14 Zooplankton ≤ 55.96 U Velocity ≤ -0.04
Good	93.2	119	12.53 ≤ Temp ≤ 13.8 DIN > 6.630 Zooplankton > 40.62
Good	86.5	7504	Temp > 14.31
Good	78.3	83	13.80 < Temp ≤ 14.05 U Velocity ≤ -0.036

Regarding Table II, the following conclusions could be made, admitting that the natural organic conditions are satisfied: Chlamys Farreri will have a bad growth if Water Temperature is below 12 degrees Celsius. To induce this growth, the DIN value should be below 6 and Zooplankton value being above 55.

TABLE II. BAD GROWTH: CHLAMYS FARRERI

Leaf	Confidence	Instances	Conditions
Bad	100.0	45	Temp ≈ 14 Zooplankton > 55.96
Bad	99.8	4720	Temp ≤ 12.53
Bad	95.4	822	12.53 < Temp ≤ 13.80 DIN ≤ 6.63
Bad	69.5	358	13.80 < Temp ≤ 14.05 U Velocity > -0.036

B. *Crassostrea Gigas*

The number of instances that Crassostrea Gigas dataset contains is 19 991, being 10 397 (52%) labeled as Bad Growth, and 9594 (48%) labeled as Good Growth. This dataset follows the Java implementation metrics, being: Threshold = 0.02 and the Derivative Factor = 1.

1) VISUAL ANALYSIS

From Fig. 8 it could be said that high values of Water Temperature, and low and medium values for Phytoplankton, promote a good growth of bivalves. In the case of Crassostrea Gigas, we can say that the Phytoplankton doesn't have a huge direct impact on its growth, being the Water Temperature more responsible to promote it.

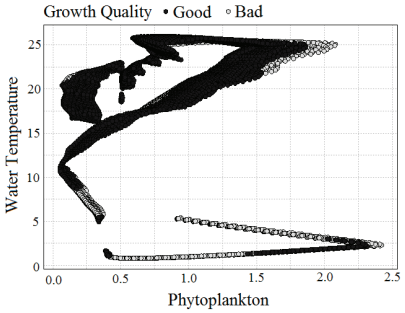

Figure 8. Crassostrea Gigas - Water Temperature and Phytoplankton

From Fig. 9 we can clearly see that high values of Zooplankton induce a bad growth. In some cases, both low values of Water Temperature and Zooplankton could promote a good growth, but is not certain to happen. Once again, high values of Water Temperature result, for sure, in a Crassostrea Gigas good growth.

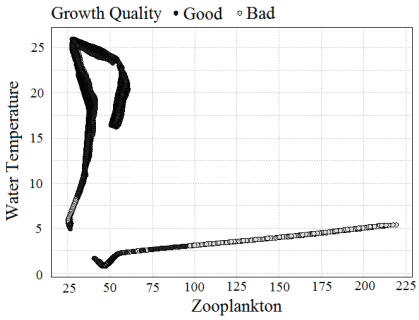

Figure 9. Crassostrea Gigas - Water Temperature and Zooplankton

2) COMPUTER ANALYSIS

Fig. 10 represents the Crassostrea Gigas Decision Tree, obtained with the parameters: Minimal Size for Split: 2, Minimal Leaf Size: 10, Minimal Gain: 0.001, Maximal Depth: 20 and Confidence: 0.25. The analysis of Fig. 10 originated two different tables. One of them tells about good growth conditions, Table III, and the other bad growth conditions, Table IV.

Regarding Table III, the following conclusions could be made, admitting that the natural organic conditions are satisfied: Crassostrea gigas will have a good growth if Water Temperature is above 26 degrees Celsius and the Zooplankton value above 27. To ensure this growth, the DIN value should be limited to 3.5 and Phytoplankton value below 2.3.

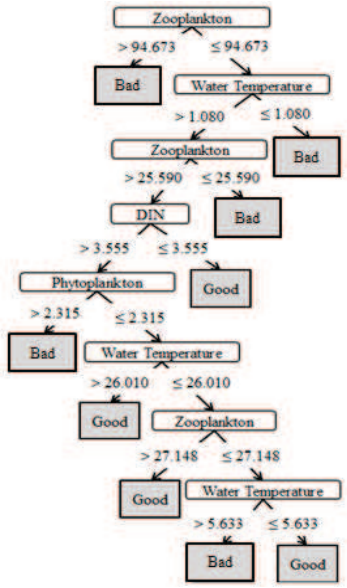

Figure 10. Crassostrea Gigas Decision Tree

TABLE III. GOOD GROWTH: CRASSOSTREA GIGAS

Leaf	Confidence	Instances	Conditions
Good	100.0	99	25.59 < Zooplankton ≤ 94.67 Temp > 1.08 DIN ≤ 3.55
Good	100.0	24	25.59 < Zooplankton ≤ 94.67 Temp > 26.10 Phytoplankton ≤ 2.32 DIN > 3.55
Good	85.7	35	25.59 < Zooplankton ≤ 27.15 Temp > 5.63 Phytoplankton ≤ 2.32 DIN > 3.55
Good	53.7	19079	27.15 < Zooplankton ≤ 94.67 Temp > 26.10 Phytoplankton ≤ 2.32 DIN > 3.55

Regarding Table IV, the following conclusions could be made, admitting that the natural organic conditions are satisfied: Crassostrea Gigas will have a bad growth if Water Temperature is below 26 degrees Celsius. To ensure this result, the Zooplankton value should be below 27 and Phytoplankton value below 2.

TABLE IV. BAD GROWTH: CRASSOSTREA GIGAS

Leaf	Confidence	Instances	Conditions
Bad	100.0	260	Zooplankton > 94.67
Bad	100.0	188	Zooplankton ≤ 94.67 Temp ≤ 1.08
Bad	100.0	176	25.59 < Zooplankton ≤ 27.15 5.63 < Temp ≤ 26.1 Phytoplankton ≤ 2.32 DIN > 3.55
Bad	100.0	109	Zooplankton ≤ 25.59 Temp > 1.08

VII. Conclusions And Future Work

Our analysis enables us to conclude what are the main attributes that influence the growth of bivalves. As bivalves are marine and freshwater mollusks, it is obvious that the Water Temperature had a great impact on their life quality, and this study proves exactly that. In the different species analyzed, the Water Temperature is an attribute that becomes evident in the separation between good and bad growth. Other organic conditions that have a huge impact on bivalve growth are Zooplankton and Phytoplankton, and it is known that all living species need to feed on some kind of matter to make its normal life cycle. Those two attributes showed a great influence in bivalve growth too, complementing the Water Temperature influence, ensuring the separation between good and bad growth. The other attributes like DIN and U/V Velocity have a minor influence on bivalve growth. Its impact could not be sufficient to determine if a growth behavior is good or bad, but they are important rules that benefit even more a good growth, or worsen a bad growth.

This type of conclusions could be obvious on a human analysis, but they are quite difficult for a machine. The Data Preparation phase and interpretations of its resulting data is not, sometimes, easy to understand for humans, and also the meaning of connecting all this steps reaching an intuitive and clean representation of results. Hence, this approach is a powerful tool for human analysis, and humans are essential to decide the value that could be given to the information it achieves, and its use for their own benefits.

This implementation of Decision Tree provides all the necessary parameters to obtain the best results and transformation of data into powerful information. Also, the Decision Tree is a very intuitive way to deduce the attribute behavior, and representation tool due to its simple and direct presentation. The C4.5 algorithm has demonstrated to be a powerful tool in datasets of continuous attributes, having different flexible parameters that can provide a better solution comparing with other approaches.

To complete the purpose of this approach, it would be necessary a full integration, by a module creation, with the EcoSimNet. This simulator and decision support system could be complemented in a way that data from the best simulations, used to advise the user, can get a higher significance. The main purpose of this integration is the identification of the environmental organic conditions that induce a good bivalves' growth, and not only the return of a specific areas or regions where the bivalves' production is better. This type of integration could provide more information about the behavior of organic matter, and how it relates with the environment that surrounds it. One of the final goals is to provide more and better information from the ecosystems, helping the scientific community and the stakeholders to understand nature, and making its work easier and with much more quality.

References

[1] "Data Mining / Analytic Tools Used Pol", May 2010. [Online]. Available: http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html.

[2] Q. Chen and A. Mynett, "Rule-Based Ecological Model", in Handbook of Ecological Modelling and Informatics, Jorgensen, S.E., Chen, T.-S. and Recknagel, F. (eds.), 2009.

[3] F. Cruz, A. Pereira, P. Valente, P. Duarte and L. P. Reis, "Intelligent farmer agent for multi-agent ecological simulations optimization", EPIA 2007, LNAI 4874, pp. 593-604, Springer-Verlag, 2007.

[4] M. Debeljak and S. Džeroski, "Decision Trees in Ecological Modelling", in Modelling Complex Ecological Dynamics: an Introduction into Ecological Modelling, Jopp, F., Reuter, H. and Brecking, B. (eds.), pp. 197-209, 2011.

[5] M. Hahsler, S. Chelluboina, K. Hornik and C. Buchta, "The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets", Journal of Machine Learning Research, vol.12, pp. 2021-2025, 2011.

[6] G. H. John, R. Kohavi and K. Pfleger, "Irrelevant Features and the Subset Selection Problem", in International Conference on Machine Learning, pp. 121-129, Morgan Kaufmann, 1994,.

[7] S. E. Jørgensen and G. Bendoricchio, "Fundamentals of Ecological Modelling", Elsevier, Third edition, Elsevier, 2001.

[8] F. P. Pach and J. Abonyi, "Association Rule and Decision Tree based Methods for Fuzzy Rule Base Generation", World Academy of Science, Engineering and Technology, vol. 13, pp. 45-50, 2006.

[9] A. Pereira, L. P. Reis and P. Duarte, "EcoSimNet: A Multi-Agent System for Ecological Simulation and Optimization", EPIA 2009, LNAI 5816, pp. 473-484, Springer-Verlag, 2009.

[10] J. R. Quilan, "Improved Use of Continuous Attributes in C4.5", Journal of Artificial Intelligence Research, vol. 4, pp. 77-90, 1996.

[11] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Warehousing, Vol. 5, Nr.4, 2000.

[12] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2005.