NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Predicting Chelonia mydas nests survivability rates with use of Machine Learning techniques

Applying Machine Learning techniques on conservation data – case study

David Tiago Calçada

Dissertation submitted in partial fulfillment of the requirements for the degree of master's in information management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Predicting Chelonia mydas nests survivability rates with use of Machine Learning techniques

Applying Machine Learning techniques on conservation data – case study

David Tiago Calçada

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management, Specialization in Knowledge Management and Business Intelligence

Advisor: Mauro Castelli
Co-supervisor: Illya Bakurov

February 2020

Acknowledgements

I simply cannot thank the Príncipe Foundation enough for allowing me to work on such a project. Special thanks to Estrela Matilde, Vanessa Schmidt and Frazer Sinclair for their time and contribution to this thesis. I wish them continuous success and that we might continue to collaborate in conservation biology studies in the near future.

To my coordinators Professor Illya Bakurov and Professor Mauro Castelli, you have my eternal gratitude for your patience and support. Your guidance was more than enlightening.

To my family who kept pushing me to go beyond what I believed were my limitations and to my friends who kept me motivated throughout. You have my love and appreciation.

I would like to make a special dedication to my grandfather who taught me to never give up, no matter what obstacles get thrown in your way.

Abstract

It is the generalized goal of **knowledge discovery** techniques to help us find useful patterns in data whilst not subjecting us to the ambiguity and overcomplexity of models. In fact, it has become increasingly important to allow for a common language to exist between biologists and data scientists.

In my thesis I aim to make use of **Green Turtle** (*Chelonya mydas*) nesting data obtained in surveys conducted from 2015 to 2019 in Príncipe Island, in order to obtain two things: Firstly, to understand insights related to sea turtle survivability rates; Secondly, to develop prediction models on said rates via popular **Machine Learning algorithms**. For this purpose, I will detail how my collaboration with the sea turtle conservation team in Principe Island began, and work has been developed since.

I will describe all steps referring to the **dataset transformation, manipulation and exploration**, and detail how each step has allowed me to feed the sea turtle data into powerful Machine Learning algorithms that are to be evaluated against their ability to predict accurate nest survivability rates.

At the end of the contextual part of this document, I will explain my findings and present the limitations of this project; I hope to provide a solid example that will allow future students and researchers to keep in mind what challenges await them should they pursue this field.

Finally, a key aspect of this thesis that is very important that it's written in such a way that individuals with different backgrounds are able to understand its content and objectives.

Keywords: **knowledge discovery**, **green turtle**, **dataset transformation, manipulation and exploration**, **Machine Learning algorithms**;

INDEX

FIGURE INDEX

TABLE INDEX

1. Introduction

1.1 Context and importance

Ethology, the study of animal behaviour has been the focus of countless research studies carried out by both biologists and data scientists. By looking into animal data, we can attempt to learn and understand more about the reasons behind several natural phenomena and in turn, exercise human action when it comes to improving species abundance and welfare. This is where I believe I can make a contribution to conservation biology.

For such a thing to happen, it is crucial to work closely with conservation professionals in order to not only obtain the biological data, but to also acquire meaningful insights about the species and their behaviour. This constitutes the key approach to this thesis.

At this point I would like to introduce the Príncipe Foundation (PF) (ANNEX I), a non-governmental organisation (NGO) with the goal of protecting wildlife in Príncipe Island, including of course, the sea turtles. With their help, I have been able to obtain a dataset containing variables that include sea turtle nesting behaviours and sea turtle anatomical and biological description.

To validate the research and its title, I have, together with the conservation team and my academic coordinators, established the contents of this document. It falls under an umbrella that is best described as applying predictive algorithms in order to obtain sea turtle nest survivability rates (STNSR). Before this, I will demonstrate a thorough analysis on the data providing a step by step view of what I have learned.

1.2 Current research

After gaining a better understating of the context of the topic at hand, it would be pertinent to mention where current studies are lacking or have failed to meet expectations.

Through the years, many studies have focused on animal species with the aim of answering a specific question, be it through knowledge discovery or exemplification of an established hypothesis. Yet, there are considerably few studies that marry standard exploratory approaches on animal data with the powerful predictive capability of Machine Learning algorithms while maintaining a simple and organized structure on both sides.

In fact, it seems that in current times the goals of biological studies appear to be about procuring a balance between how much data understanding we can achieve versus keeping it grounded on a strong theoretical basis. Several studies have been built on this principle, and I will look to further explore some examples in subchapter 2.3.

1.3 My contribution

Referring to the above subtitle, my purpose is to combine an exploratory analysis on sea turtle data with the predictive capabilities of Machine Learning (ML) algorithms to achieve sea turtle nest survivability rates with feasible results.

Specifically, I aim to not only provide useful insights on sea turtle nesting status, but to also provide a basic framework from which future analytical projects can derive from (be it for the same topic or not). To achieve this, I have attempted to provide my findings in the clearest way possible while avoiding the ambiguity and abstraction that regression analysis typically leads us to.

1.4 Structure

To conclude my introduction to this thesis I will refer to the structure I will follow throughout:

In chapter 2, I present the theoretical background from which my research extends from. In the first subchapter, a short description of the country of São Tomé and Príncipe will be given with a focus on social, economic, meteorological and geographical contexts.
The next subchapter includes references to documents that approach conservation topics in Príncipe Island. I will be looking to discuss what approaches were used and what was discovered. I intend to explore possible limitations and where further developments could be made.
To contextualize ourselves in the regression problem that we are dealing with, the next titles will introduce concepts of Time series and Prediction and how these are relevant in the greater context.
To shore up this chapter, I will introduce Machine Learning algorithms and their use today. This includes a presentation of the algorithms I will be using in my thesis and the reason they were selected.

Chapter 3 will proceed with a description of the PF and their work on sea turtle conservation in Príncipe Island. I will describe the details of our collaboration and how together we established what topics I would focus my analysis on.

Chapter 4 will be divided between the pre-processing of the data and my exploratory approach to it. It will include the framework summary of what changes were needed for the data in order to have it in a structured form, as well as the establishment of the research cases that the NGO has asked to me to focus on.

Chapter 5 will then make use of the structured and clean data that was achieved previously and apply it to a benchmark with tuned parameters for the algorithms that were explained in Chapter 2. A presentation of the results and their meaning will follow.

Subsequently, in chapter 6 I will discuss the results and approaches accomplished thus far. I will present my overview over the whole document but focusing on the key learning aspects that were achieved. My assessment will focus not only on each step of the analysis, but on the big picture, allowing me to comment on the progress against my goals.

Finally, chapter 7 will contain both the discussion with my final remarks on the work done with limitations being presented in the context of what was set out to accomplish, as well as my response to the cases that the PF asked me to focus on, with recommended guidelines for the future and where to correct possible mistakes.

1.5 Tools

For the purpose of this thesis, I have made use of the Python programming language. After sourcing the excel files containing the turtle data, I have worked on the Anaconda distribution version 3 with Spyder[1] platform (4.0) to develop the coding steps for my analysis as well as to run my algorithms.

I have built my visual aids on Plotly and Seaborn Python libraries.

[1] Available for download at https://www.anaconda.com/distribution/

2. Theoretical revision

2.1 São Tomé and Príncipe

The islands of São Tomé and Príncipe are the two main islands that constitute the archipelagos of the Democratic Republic of São Tomé and Príncipe, an African country located of the western continental coast in the Gulf of Guinea [2, 14].

History

They were uninhabited until a Portuguese arrival on the 15th century gradually started colonizing the island and turning it into a commercial trade station. It remained under Portuguese authority until it obtained its independence in 1975, attaining a democratised form of government. Its culture is based on both African and European influences, as it can be seen in the country's customs and music.

Political, economic and social

Currently São Tomé and Principe is the second smallest sovereign nation in Africa with a 2018 study estimating around 201,800 individuals constituting its population, harbouring a mix of African natives and mestizo descent.

Economically the country harbours a high dependence on the exportation of cocoa (representing 95% of all agriculture exports), with a reasonably small fishing sector followed by an even smaller industrial sector. The countries government has nonetheless attempted to integrate tourism as an economical sector, but high restriction on nature conservation and subpar logistics have made this undertaking arduous. Also noticeable is the country's parallel economy that up until more recent times was heavily based on poaching for several native birds, land animals and sea species (including sea turtles).

Geography and climate

São Tomé is 50 km (30 mi) long and 30 km (20 mi) wide and the more mountainous of the two islands.

In comparison, Príncipe is about 30 km (19 mi) long and 6 km (4 mi) wide, making it the smallest of the two.

The climate is tropical with rain occurring mostly during October to May. High temperatures are at sea level, while there are more mild temperatures as one treads inland and into higher altitude grounds.

Due to the Príncipe's volcanic constitution, its soil is rich in sustenance for plants, which led to a prominent domestic plantation agriculture that is mainly used for exportation. Its land area is mostly covered with rich and varied flora that retains large amounts of its endemic background, although this has suffered significant changes due the transformation of the islands ground into plantation fields.

Beaches

Given we will be only looking into the beaches where a certain sea turtle species make their nests, the focus will be on detailing the geographical data of those same ones. All the 8 beaches that can be found on the dataset can be described as small white sand beaches with the tropical forest line a few meters from the tide line. A map showing the overall geography of Príncipe is given below, where markers pinpoint the location of the beaches we will be looking at:

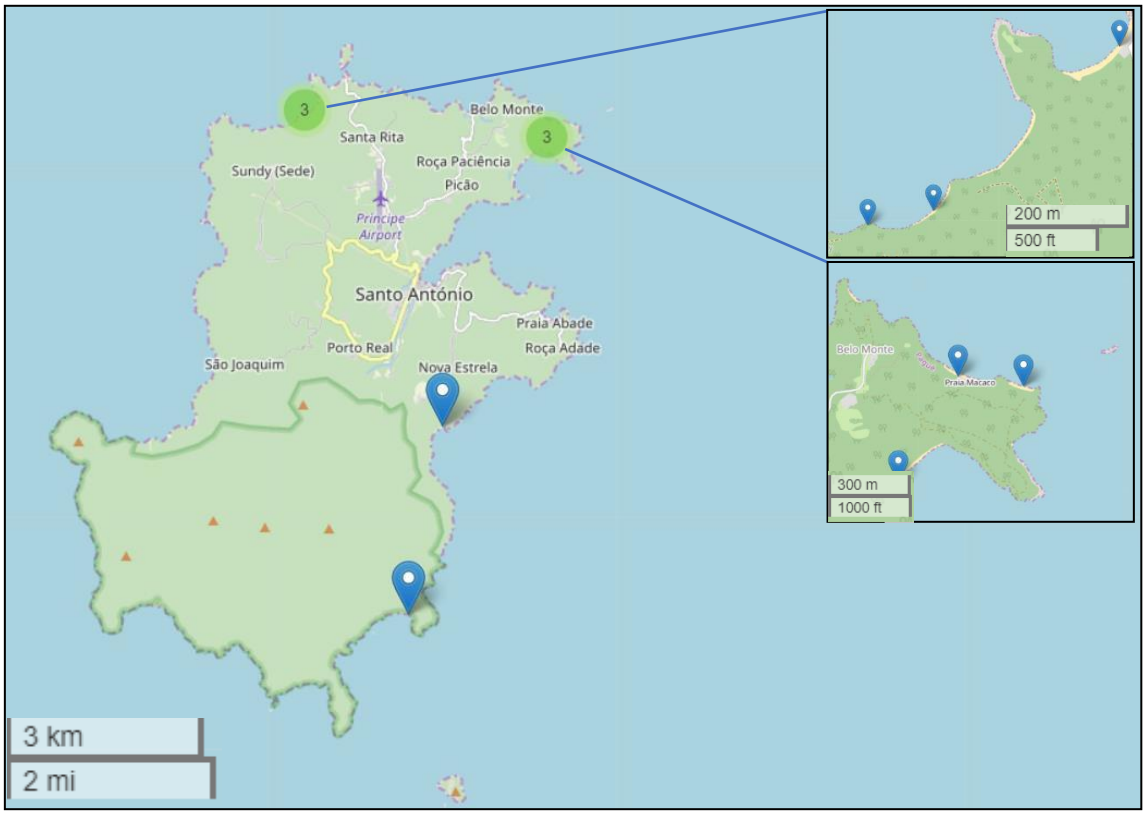


Figure 1 representation of Príncipe Island with 2 submaps

At the bottom we have INFANTE and BUMBO, on the top right corner box we have MICOTO, RIBEIRA IZÉ and BOMBOM with GRANDE, MACACO and BOI shown on the middle box at the right

Human activity

As we can see on the above figure, there is human presence in the island with Santo António being the main hub of the island served with an airport and harbour that allows for travel between the two islands and to mainland Africa. There is a designated protected area that aims to limit human activity at important natural landmarks. The line that establishes the frontier of this park is seen with the green light that crosses the middle of the island. Important to notice as well is that most beaches are located outside the natural park area, showing that the delimitation of the natural park area was not particularly driven by the sea turtle conservation effort. For this purpose, several volunteers conduct patrols on the several beaches, with the aim of preventing poaching of nesting turtles and their nests.

2.2 Turtle species in Príncipe

As the title suggests, we are only looking to study one species of sea turtle and their nests in the island of Príncipe, as one PF have concentrated their conservation efforts on, and the sites where the data was collected. That said, I can introduce the 3 turtle species the NGO focuses most of its analytical and conservation work: Green turtle (*Chelonia mydas*) (ANNEX II.I), Leatherback turtle (*Dermochelys coriacea*) (ANNEX II.III) and Hawksbill turtle (*Eretmochelys imbricata*) (ANNEX II.II).

The Green turtle (*Chelonia mydas*)

The *Chelonia mydas* (CM) is known as green turtle due to the characteristic greenish hue on the back and the colouring of its flesh. It measures on average about 83-114 cm long, weighing anywhere between 110-190 kg. It is not the biggest sea turtle species nesting in Príncipe[2], or the one that lays the most eggs. The interesting aspect of this species is that it is by far the most numerous, as being attracted to tropical climates has made the CM a frequent visitor to the beaches of Príncipe and one of the PF's main contributors to turtle data collection. This resulted in a more complete set of data on this species and its behaviour in the island. From the pie chart below we can see the percentage on the total number of nests for 3 turtle species:

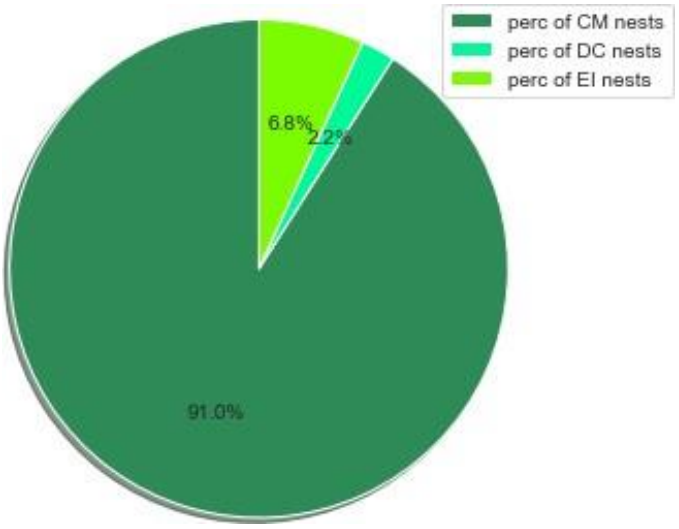


Figure 2 pie chart on the percentage of nests belonging to each species

Supported by what we can see in figure 2, the decision to focus on CM data comes mostly from the cheer dominance on the volume of data that relates to this species, and due to the fact that it is not possible to assess the other 2 species at this time due to highly incomplete data in the source files.

[2] The DC is quite larger, yet significantly less frequent sightings of them at the beaches of Príncipe have left data gathering in a complicated state.

Nesting

The CM has its nesting season in intervals of two years; after mating, the female will approach the beach and use its back flippers to dig a nesting chamber that can on average measure 22-54 centimetres (8 – 21 inches) deep. It will then lay within a range of100 to 126 eggs and proceed to cover the nest with sand, again by making use of its back flippers before getting back to sea. These eggs will then take up to 60 days to incubate inside the nest.

To give an idea of the nesting frequency and presence of this species in Príncipe, there have been over 300[3] CM nests in 2019 alone. The team has assured me, along with a quick analysis on the data, that the same turtle can nest several times per season. So even when looking at the grand total of nests made during the period of 4 years that is assessed, we must take the latter fact to consideration.

Looking at the data from the last 4 calendar years, it becomes increasingly obvious that the CM are the most numerous and frequent nesters. Even considering the 2-year interval between nesting seasons, a CM can nest several times during this period, with an average of 3-5 nests per season [17]. The reason one must insist in mentioning the existence of different nests being made by the same turtle is to clarify that each nest will be considered independent of each other, meaning I will not consider at this stage that each nest formation is dependent on the previous one.

Understanding what beaches the CM tends to nest on more often comes next.

Beach count and percentage		
Beach name	**Count**	**Percentage**
GRANDE	644	82,04%
INFANTE	97	12,36%
BOI	34	4,33%
BOMBOM	3	0,38%
RIBEIRAIZE	2	0,25%
MICOTO	2	0,25%
BUMBO	2	0,25%
MACACO	1	0,13%

Table 1 summary for nest counts and percentage for all beaches

As is it noticeable in table 1, out of 806[4] nests, the vast majority of them are lain in the beach of Grande. Through Príncipe Foundation's feedback, it is understandable that this is indeed the most widely used beach by the CM due to its size and location. Although we can see that there is only one nest in the beach of Macaco, we must take into account that the dataset that is used will only keep rows with complete information on each individual nest. Although we can assume there are more nests on that particular beach, we just do not have the data to properly verify it. This means that if it came to be that the beach where the nest is made is pertinent for STNSR, a careful but limited assessment must be made.

[3] This number accounts for the number of rows on the 2018/2019 nesting season.
[4] Total number of rows in the dataset for the 4 nesting seasons.

2.3 Previous studies on the area

2.3.1 Research on species survival

Typically, the classic approach to this research topic has been made via using classical models that test hypothesis based on new or past existing theories constructed on grounded literary work. Unfortunately, this approach rarely allows for the implementation of more interesting pattern discovery techniques that might create the foundations for new ideas altogether.

Having said that, current research papers tend to approach the topic of animal behavioural analysis in a fashion similar to what is being done in this thesis. Two recent studies conducted in Príncipe were a good basis to understand what type of approach one needs to take in order to further expand this ideology.

Rapid decline of the endemic giant land snail *Archachatina bicarinata* on the island of Príncipe, Gulf of Guinea by Dallimerand M. and Melo M. [10]:

The first example is a paper made on the population decline of the endemic Giant Land Snail. It goes through the effort of identifying key human factors that lead to fewer individuals of this species existing in the island of São Tomé by applying surveying actions through the island's different geographical landscape and then conduction analysis on the results.

Due to the purpose of this paper, there isn't much exploration on advanced methods for knowledge discovery, but instead a bigger focus on establishing a basis for good explanatory variables. The writers claim:

"We used multiple regression, with both the abundance of live snails (snailabundance) and the occurrence of live or dead individuals (snail presence) as the response variables, and measures of habitat type, productivity and location as explanatory variables."

Keeping in mind that survey and surveillance effort was conducted by these same individuals, it would have been a very time-consuming undertaking to then conduct extensive knowledge discovery using more complex methods than linear regression. Yet, it is indeed pertinent to consider how the use of different feature selection techniques or application of advanced predictive algorithms such as the ones in Machine Learning (ML) literature might have improved on model accuracy for the Giant Snail presence.

Modelling the distribution of São Tomé bird species: Ecological determinants and conservation prioritization THESIS by Filipe Soares [33]:

Filipa looks into endemic vs non-endemic bird species in the island of São Tomé and makes an interesting case for the use of classical regression techniques as she sets out to discover interesting patterns in data that are visually detectable with the use of maps and clustering techniques. With this methodology, it is possible to identify several dozens of different bird species and serves as a good stepping point to what I wished to achieve with my research. In this case, an understanding approach to undiscovered or unsuspecting causes that lead to STNSR.

This thesis conducts a very complete view on data with a very detailed approach to variable importance both relative to natural habitats, as well as the human threat factor. Indeed, it is a good basis to any study on groups of animals of one or several species and a first of its kind in São Tomé and Príncipe.

The methodology is based on:

- the calculation of Relative Variable importance in order to identify behavioural differences between different bird guilds;
- applying multi linear regression on data;
- conducting logistical regression in order to analyse the response of each species to continuous variables, followed by ranking of the relative importance of those said variables;

Again, it would be possible to extend this research further on by advancing it to ML techniques. Having said that, it is understandable that such an endeavour is neglectable for the overall purpose of the thesis and the great results it achieved.

Estimating carrying capacity at the green turtle nesting beach of the East island French Frigate by Shoals G. H. S. Tiwari, M. Baladz, [36]

This paper is also based on a prediction effort made on top of a green turtle nesting habits dataset. In this specific case the author aims to discover if enough nests are being made (carrying capacity) at the different beaches of the archipelagos of East Island Frigate Shoals and understand the causes that might lead to a positive or negative answer. Available to the author is a dataset that contains data from 37 nesting seasons made by green sea turtles in 10 different islands.

For this intent, and much like my thesis, the data is based on a surveying effort and is to be ran through a model that is robust in face of time series data so as to help establish factors that lead to the estimation of the carrying capacity.

Throughout the paper, the author details what are the impacts of factors and if they can be considered important or neglectable for the carrying capacity whilst establishing the framework for the calculation of important variables. Time series analysis is conducted in order to understand the existence of a trend that has led to more or less nests being formed. This is an important consideration especially given the vast time period that the study covers.

In the end it was not only concluded that there is no predominant trend, but also that nesting was carried out mainly in one beach, while the others are considered to be below their carrying capacity. Very important considerations are made for the quality of the data in the surveys and how more information on both geography and climate can help improve the analysis.

I find this paper particularly interesting as it helped me understand what issues tend to arise when analysing historical data based on surveys. It also makes an interesting case for the possibility of added value on one's analysis by adding more variables that relate to different themes that (in theory) might affect sea turtle nesting habits.

2.3.2 Prediction

In statistics, Prediction is a statement for a future event for a whole population based on the knowledge of a sample of the population.

Objectively, in this research topic it is not enough to have an interesting dataset, it is perhaps even more important to be supported by a team of conservationists with a great understanding of how the data was collected and its meaning. To make full use of this potential, it is wise to attempt to go beyond the application of standard[5] regression analysis on prediction and to explore the output of more developed algorithms.

The use of ML techniques in prediction has existed for many decades in the form of models of many types and purposes. We have continued to extend the boundaries on the work of ML algorithms in attempts to acquire better predictions that seek to beat standard models e.g. linear or logistical regression. The goal is to apply exhaustive search for patterns inside the data that may lead to more accurate results, while operating on higher volumes of data at higher speeds without having to recur to overcomplex specifications of data.

2.3.3 Machine Learning in biological data

When one searches for examples on the use of ML in the area of biology, it is quite common to come across several examples of its use in health [41]. Its contribution to it is undeniable and it seems it will continue to expand into a wider range of fields. Just to give a few examples:

- Atomwise: Builds 3D modules out of molecules through a powerful algorithm;
- DeppVariant: A statement on the use of Deep Learning for genome studies;
- Cell Profiller: Being able to identify thousands of features in cell groups;

When it comes to health, it is possible to understand that most advancements derive from the need to understanding and visualizing several aspects of an individual's health. This in turn allows us to develop powerful algorithms that provide accurate results and higher speeds in order to help doctors evaluate someone's status whilst greatly reducing room for error.

It is regrettably noticeable however, how lacking we are in terms of applying these techniques on other animal species. A recent study article on animal species classification using Machine Learning techniques gives us the following abstract:

"created by collecting the features of ears and eyes from 10 animals and an experiment was conducted using Machine Learning techniques such as SVM and MLP to classify them as predators or pets."

This has been the most frequent example nowadays of how we can use ML techniques in order to understand the animal kingdom and create useful tools that helps us identify species in a quick and feasible manner. It seems as though we are lacking vison on the possibility of applying new solutions to old problems, e.g specie survivability prediction or behavioural pattern analysis.

So, in this thesis I will attempt to show that even though I am using ML for a classical regression problem, its use and relevance today are not in any way diminished and I hope to show that learning from such endeavours as these ones will keep pushing us to strengthen and diversify ML to other biological areas.

[5] By this, I mean the application of OLS squares to test pre-established hypothesis.

2.3.4 Pooled data

I am using data collected from 2015 to 2019. In detail, I have nesting data for each season starting from February 2015 to July 2019: The different series are detailed as:

- Censo temporada reprodutiva 2015-2016
- Censo temporada reprodutiva 2016-2017
- Censo temporada reprodutiva 2017-2018
- Censo temporada reprodutiva 2018-2019

Having such a structure of data split in different periods, means we are dealing with Time Series [38]. In other words, we have a set of observations on the values that a variable takes at different times.

Yet it is also true that we are analysing a nest and its characteristics when registered at a single point in time, meaning that although a turtle might have nested in 2016 and then came back 2 years later to nest again, these are considered independent happenings from one another and as such show that the data is partly cross-sectional.

The conclusion is that the whole data set can be considered a combination of both Time Series and Cross-sectional, which in this case is named Pooled data.

Figure 3 line plot for trend

When dealing with this type of data, it is important to assess if our target variable has any peculiar trends that can induce seasonality[6]. If this phenomenon is observed, certain techniques can be used to fix it and reduce the impact that it may have on statistical analysis.

I can inform that we do not need to worry about this, although in the graph of picture 3 there is drop right at the end of the series, it is taking an STNSR average of only 3 observations out of the 806 total. This, along a quick assessment for trend analysis, showed that I had not any variation that might lead me to believe we are dealing with seasonality. Thus, I do not consider it necessary to apply trend harmonization techniques.

[6] Seasonality - the presence of variations at specific regular intervals less than a year.

2.3.5 Machine Learning algorithm selection

As I have articulated several times above, I will use popular Machine Learning algorithms to achieve STNSR. Before I state the reasons for my selection, I consider that it is important to introduce the main fields we can categorize them in.

Deep Learning

It's a subfield of Machine Learning that refers to the algorithms based on the functioning of human cells that we call Artificial neural networks. Studies [5,31] have given a good explanation on why the use of Neural networks has taken centre stage. High volumes of data and higher performance being the most accepted ones. For the purpose of testing how an ANN performed in comparison with the other models, I have decided to insert it in my study.

Ensemble methods

In a broad term, it is a concept that states that combining a set of weak learners can create a stronger one (by decreasing its variance) [27]. In short, what we are doing is using several learning algorithms that can provide us with better prediction power whilst creating a robust model that is strong in terms or variability and resistance to outliers.

The form that these specific methods assume are that of several Decision Trees that can be provided with voting criteria that help them learn how to form the best tree structure. This is achieved by sampling and splitting the data at different points in hope of achieving the lowest possible error deviation from the actual value we are trying to predict.

Going into more detail, we can differentiate the structures of algorithms that fall in this main category as either baggers (bagging type algorithms) or boosters (boosting type algorithms).
Given the high amount of documentation that supports the use of these methods in most generalised regression problems, I have decided to use a combination of different algorithms that relay on either boosting or bagging.

Xgboost

Although also an ensemble method based on boosting, it is important to specify why I have selected this algorithm.
For the last 5 years, its popularity has increased as its performance has been documented to have improved drastically, ever since its stable version in 2019. There will be a more theoretical review on Xgboost in the following chapter, but I wanted to clarify that its integration in this research is because I believe that based on its recent popularity, it can outperform all other algorithms [34].

2.4 Machine Learning Algorithms and benchmark creation

2.4.1 Benchmarking

Predicting the STNSR rates will be achieved by using a selected set of Machine Learning algorithms that will be tuned based on a search for the most suitable parameters. This is called a benchmark. An iterative process that selects the best set of parameters for each algorithm based on an evaluation metric. The selection of what parameters the benchmark will build on and how it will evaluate its performance is based on my theoretical study on the area, supported by my understanding on the problem of prediction at hand.

Having said that, it is pertinent to refer to the algorithms[7] that are used and to give a brief explanation on how they fall on the Machine Learning scope as well as their workings based on their literature.

2.4.2 Supervised learning

Unlike unsupervised learning where we are unaware of patterns in data and make no use of pre-existing labels, here not only do we have historical data for our variables, we also aim to create mappings of input to output based on "what is known". To denote a more obvious intricacy to supervised learning, the Training data is run by an algorithm in order to infer a function that lowers the error of our prediction [28].

As members of the Supervised learning family and used in this thesis we have:

Artificial Neural Network

Neural Networks have taken centre stage in the ML world for the last few years much thanks to their high programmability and computational power. It is particularly due to the former that it was decided to make use of this algorithm. For example, Neural Networks have been known to outperform the traditional algorithms when the amount of data that we have increases. My scenario does not necessarily call for the use of a neural network, given that we do not have a high volume of data. There is however an interest in assessing and comparing its performance for the dataset provided. Thus, my inclusion of the ANN on my benchmark [7].

How it works

It is commonly compared to the workings of the human brain, specifically the signal transmission between axons and synapsis. However, the name artificial network comes from the fact that it is not of biological nature, it is human defined.

A basic Neural Network (NN) is constituted by 2 layers alone, the input and output layer (no existence of a hidden layer). This is typically described as the linear form of the neural network, as weights are used in order to calculate the importance of a certain input when computing the output.
Once we start inserting hidden layers into the architecture of the NN, we need to consider that we are dealing with two different types of weights: the first class being the ones that connect the inputs to the hidden layer; the other one being the ones that connect the hidden layer to

[7] All algorithms except the Xgboost Regressor come from the SKLEARN python library. Their parameters will be defined based on the description given on the library contents as to diminish ambiguity. Xgboost and its parameters are instead provided by the library of its own name.

the output. For the purpose of achieving proper outputs that are interpretable, Neural Networks make use of an activation function between the latter's connections.

Several parameters are necessary in order to establish the NN, and although the construction of their architecture depends on numerous factors, it is quite complicated to assert what is the best initialisation.

Given this is a supervised learning problem, the calculation of weights is based on the back-propagation algorithm. Since we have the correct answer for the output at every iteration, weights are recalculated by use of errors (the lower the better) in order to achieve the best possible predictions [39].

Summary

Strength	**Weaknesses**
Capable of solving complex problems	Prone to overfitting[8]
Versatility of use	Prone to topology issues
Tends to excel in high volumes of data	Might require extensive parameter tunning

Table 2 summary table for artificial neural network strengths and weaknesses

Parameters

- Number of hidden layers – number of hidden layers to use;
- Hidden layer size – number of neurons in each hidden layer;
- Maximum iterations - Maximum number of iterations. The solver iterates until convergence (determined by 'tol') or this number of iterations;
- Learning rate – Learning rate schedule for weight updates;
- Initial learning rate - The initial learning rate used. It controls the step-size in updating the weights;
- Alpha - L2 penalty (regularization term) parameter;
- Tolerance - Tolerance for the optimization. When the loss or score is not improving by at least tol for a number of consecutive iterations, convergence is considered to be reached and training stops;
- Beta1 - Exponential decay rate for estimates of first moment vector in adam, should be in [0, 1). Only used when solver='adam';
- Beta2 -Exponential decay rate for estimates of second moment vector in adam, should be in [0, 1). Only used when solver='adam';

[8] Overfitting – Occurs when a model can be particularly good at predicting values for one or few data sets but bodes poorly when applied to a larger share of others. In essence the model has become too good at predicting that/those data set(s) and lacks the necessary variability to be applied on others.

Bagging or bootstrap aggregation

As stated in article [5]:

"*Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor".*

In a sort of specialized way, Bagging is particularly good at preventing overfitting which is typically caused by the known belief that a Decision Tree makes use of if-else statements in its construction. Consequently, this may cause an underlying favour for certain features or a too thorough spread of said features inside the Decision Tree, bordering on the redundant.

Bagging calculates all the variances for the dataset (by sampling and replacing data). Thus, multiple models are tested in this context.
The follow up and what gives this method its "aggregation" name, is that each singular hypothesis is attributed to a weight. Its particular workings are what led me to select it for my study.

How it works

The first of the ensemble methods to be introduced, so naturally it is based on the creation of Decision Trees. It is also known as bootstrap aggregation, as there is several resampling of data as well as sub-selection of features with the goal of finding the feature that provides the best split. Each tree is running in parallel with the others, in such a way that leads to the establishment of models with high variances but low bias. In the end, an aggregation of all the predictions of each models is made by calculating the highest possible variance with constant weights in order to establish the final value through voting [5, 11].

Summary

Strength	**Weaknesses**
Strong performance on outliers	As the ANN, might lead to lesser interpretability the more complex the model becomes
Tends to efficiently reduce variance to avoid overfitting	Requires proper parameter tuning

Table 3 summary table for bagging strengths and weaknesses

Parameters

- Base estimator - The number of base estimators in the ensemble;
- Max features - The number of features to draw from X to train each base estimator;
- Max samples - The number of samples to draw from X to train each base estimator;
- Number of estimators - The base estimator from which the bagging ensemble is built. In this case I use 3 types of Decision Trees, all with minimum split of 25 samples but with 3, 4 and 5 maximum depth;

Random Forest

A Random forest (RF) is a collection of Decision Trees that makes use of bagging (again with the use of replacing and resampling) to provide each decision tree with a certain set of features built on random data points. The models that are generated through this method are supposed to be robust in face of outlier and capable of providing good estimates for the data.

The final stage of the algorithm in terms of regression is to take an average of each individuals Decision Tree's estimates.

One of the biggest issues with Random Forest is that it might lead to low interpretability on data given the high number of trees and the variability of data that runs through the model. It is somewhat possible to overcome this issue if we know how to detect and collect what group of trees has split the most relevant data. It mostly depends on the backstage complexity of the tree.

How it works

An RF is a collection of Decision Trees built though the bagging method that was mentioned above. The key difference is that each tree is built on top of the information of the previous one. In visual terms this will lead to highly diverse sets of trees with several levels of depth, which leads to a general lack of interpretability on data from the algorithm at a superficial level. What we seek to obtain are several models with high variance and low bias caused by the exhaustive search for the best splits in data. In the end we merely obtain an average on the estimates of all models [6].

Summary

Strength	**Weaknesses**
Widely used algorithm	They are not easy to interpret
Not sensible to overfitting	Might have drastically varying results depending on splits
No need for data normalization	-

Table 4 summary table for random forest strengths and weaknesses

Parameters

- Number of estimators - The number of trees in the forest;
- Max number of samples - The maximum depth of the tree;
- Minimal split point - The minimum number of samples required to split an internal node;

Adaptive Boosting

Commonly known as Adaboost [27], it is the first of the "Boosting family. Just like the previous and following two methods, it is an ensemble method that can also be used for regression. Unlike bagging though, it has a more "horizontal" development as each model is dependent on what the previous model has selected for it.

It is the base from which several boosting algorithms were created, and although typically outperformed by other modern boosters it still holds prevalence thanks to its framework and parameter tuning being well documented and explained [5, 11].

How it works

As mentioned, being an ensemble method means it is built on top of several decision trees. The key difference to bagging is in its bootstrapping phase. Each sample of data is weighted differently, meaning some samples are run more frequently than others. This is because Boosting will select the highest error outputs and give them heavier weights, causing these to go through different iterations of the algorithm in order to better train the model [11].

In the final stage of the algorithm, it will select the best output based on the weights. There is a high change of the models having learnt "incorrectly", as outliers might provoke the model into considering some features more important than what they really are.

Summary

Strength	**Weaknesses**
Does not penalize weak features	Prone to outliers
Does not require higher parameterization	Not the most powerful algorithm when compared with the other algorithms here
Less prone to overfitting than an ANN	-

Table 5 summary table for adaptive boosting strengths and weaknesses

Parameters

- Base estimator - The maximum number of estimators at which boosting is terminated. In case of perfect fit, the learning procedure is stopped early;

- Number of estimators - Learning rate shrinks the contribution of each regressor by *learning_rate*. There is a trade-off between *learning_rat*e and *n_estimators;*

- Learning rate - The base estimator from which the boosted ensemble is built. In this case I use 3 types of Decision Trees, all with minimum split of 25 samples but with 3, 4 and 5 max depth;

Gradient boosting

Part of the Boosting family, it is a popular and widely known ensemble method [37]. My reasoning for its use can be given to the other methods. The main intent was to use algorithms that allowed confidence on their feature selection process and that typically lead to accurate predictions [5].

How it works

Gradient Boosting aims to reduce the "Loss" in the Loss function. The Loss is given by the residual difference of the actual vs predicted value and iteratively aims to reduce said difference. A first model is calculated trying to fit the model, at each iteration the residual difference is calculated, and weak learners are added to the model in order to shore up the areas where we have highest variance [5, 11].

The loss function is given as:

$$y=ax+b+e$$, with e being the error term.

Summary

Strength	Weaknesses
Can outperform a Random Forest	Requires more complex parameter tuning
Boosting based approach	Can overfit with a high number of tress

Table 6 summary table for gradient boosting strengths and weaknesses

Parameters

- Number of estimators - The number of boosting stages to perform. Gradient boosting is fairly robust to overfitting, so a large number usually results in better performance.

- Learning rate - learning rate shrinks the contribution of each tree by *learning_rate.* There is a trade-off between *learning_rate* and *n_estimators*.

- Subsample - The fraction of samples to be used for fitting the individual base learners. If smaller than 1.0 this results in Stochastic Gradient Boosting. *subsample* interacts with the parameter n_*estimators*. Choosing *subsample < 1.0* leads to a reduction of variance and an increase in bias.

Xgboost

The full name being Extreme Gradient Boosting, Xgboost is a powerful algorithm known for its speed and performance in both classification and regression problems.

Famous after its success in Higgs Machine Learning Challenge it follows the same base logic of gradient boosting. The difference is that it has been tweaked in order to maximise computing performance to achieve higher accuracy in less execution time.

There are several parameters that can be modified for the Xgboost. Yet, given there is scarce documentarian on the scalability of the algorithm, I have decided not to overcomplicate the parameter tuning of this algorithm, as there is no particular reason that could lead one to believe this will affect its potential in any sort of way [15, 34].

How it works

Xgboost base performance is built on the same logic as its predecessors as it is in fact a tree booster (boosting based on the creation of trees), it distinguishes itself by its use of two useful techniques called shrinkage and column subsampling, the algorithm builds on top of an exact greedy approach algorithm that exhaustively looks for the best possible split on the data. This allows for particularly low bias, while not necessarily endangering higher variances [34].

Summary

Strength	Weaknesses
Efficient and fast	Not that many parameters to tune
Exhaustive feature splitting	-
Avoids overfitting	-

Table 7 summary table for Xgboost strengths and weaknesses

Parameters

- Number of estimators - The number of boosting stages to perform. Xgboost is fairly robust to overfitting so a large number usually results in better performance;

- Eta - Step size shrinkage used in update to prevent overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative;

- Subsample - *Subsample* ratio of the training instances. Setting it to 0.5 means that Xgboost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting. Subsampling will occur once in every boosting iteration;

3. Príncipe Foundation (PF) and analysis focus points

3.1 The organisation

As it was mentioned in the beginning, this thesis was made possible by a collaboration with the Príncipe Foundation.

I began my contact with the Director of the organisation Estrela Matilde in January 2019. Afterwards I was introduced to Vanessa Schmidt, in charge of the sea turtle conservation project. We then started laying the groundwork for what became this thesis. It was in July 2019 that I received all the data that constitutes my whole dataset and that forms the basis of my analysis. I have since been in contact with both Vanessa Schmidt and Estrela Matilde in order to successfully collaborate with them in answering questions related to STNSR.

Translating this to the work at hand, the first interactions were based on both sides attempting to understand what common ground is to be found in an analytically based project with biological theory support. As such, a list of headlines was established between me and the NGO in order to agree on a solution that benefits both sides. An output of this thesis is an initially theoretical topic that is converted to a data driven solution that begins with historical collective surveying work and finishes in estimation via Machine Learning algorithms.

3.2 The cases

As mentioned above, it is important to have a compromise between two different perspectives of work. The following cases are important headliners for the exploratory chapter of this document:

1 – Are there any particular sets of variables that allow the conservation team to better understand STNSR at this stage?

2 – How well do the predictions obtained in this project represent the reality, keeping in mind that it is important to maintain clarity and simplicity when interpreting results?

- What structure would be needed to be followed in order to scale up or maintain the quality of data?

3 – Is it possible to map STNSR at this stage?

- To analyse if the natural conservation area is efficient in preventing high death rates
- If a solution that allows the NGO to demonstrate the results to the communities in Príncipe exists, so as to better engage them on the topic.

In the discussion paragraphs in chapter 6, I will be giving my input on each of these cases and assess how well we can answer each issue at this stage. The goal is to at the very least provide a good idea as to where the NGO is standing at the moment and where it could focus its efforts in the future.

3.3 Data collection and presentation

As it was mentioned before, the data is collected by the PF. More specifically, it originates from a constant surveying effort made by the NGO's led group of volunteers that gather data on different nests from the different beaches in Príncipe.

It was in 2015 that the current state of Censos [9] was started in order to better understand sea turtle's nesting behaviours, aimed at improving the understanding of species endangerment. Nowadays, what started as an action with aims to reduce poaching of sea turtles and their eggs, grew to be an analytically driven approach that seeks to better understand STNSR.

As mentioned above, most of the data is collected by individuals that receive training in dealing with turtle nests. They are instructed on what to write down when observing a nest once it is carefully opened and to take notes on a series of observations, e.g. number of eggshells, malformed eggs, predation signs, distance to shoreline, to name a few.

Figure 4 an example of a CM nest dud open for analysis. Acquired in: https://nicolemclachlan.wordpress.com/2012/01/20/conservation-on-the-reef/

This of course means we are mostly dealing with data from October to May, the months corresponding to the nesting season. Having said that, it is important to note that there are observations in the data that refer to the summer years that are already out of the expected season, yet these are scarce.

It is also important to mention that although one nest can be reviewed several times during the year, I will only be contemplating the data relevant to the last nest survey entry, which refers to the last day where the nest was open after the expected hatching period. This will in turn prevent me from analysing the same nest twice and incurring on some sort of dynamic time series analysis where other techniques and analysis would be necessary (subchapter 2.3.4).

A total of 78 variables are considered when during the nesting season (ANNEX III). This totals just about 6600 rows of data and showcases the ambition of this project. Yet, after preliminary data pre-processing is done for the 4 data files, only around 1000 rows can be considered for the purpose of my research, as large amounts of missing data, noise and irrelevant variables are dismissed from the analysis.

In conclusion, although there is a significant drop in the volume and variability of data, there is still a reasonable amount of observations that assures we can proceed with STNSR analysis throughout this thesis.

[9] Censos – Short name for *Censos temporada reprodutiva,* the surveying effort conducted by the teams from 2015 to 2019.

4. Data Pre-processing

4.1 Data set transformation

At the start, it was mentioned that my raw data is based in a series of excel files containing data on 3 sea turtles nest's biology and behaviour from 2015 to 2019. The first thing that needs to be said, is that the PF and its conservation team have indeed made different changes to the *Censos temporada reprodutiva* datasets as time went on, taking in new data and discarding useless one. Nevertheless, the excel files maintain a basic template throughout and that the focus was on standardizing the information contained in said files and preparing them to be used as my sources.

Below the reader will see the data transformation process detailing what was the initial data and what we finished with before applying exploratory analysis:

1st stage

I was given copies of 4 datasets. Each file contains the same overall formatting template that allows for quick consulting of data and for basic metric calculations. It is, however, necessary to perform several column renaming efforts to assure that the 4 sets have the exact format. Also, it is necessary to have in account at that this stage not all columns will be used and, as such, a descending reorder of the columns is made so that the ones containing more missing values are easily identified:

In summary, we have:

- Censo temporada reprodutiva 2015-2016
- Censo temporada reprodutiva 2016-2017
- Censo temporada reprodutiva 2017-2018
- Censo temporada reprodutiva 2018-2019

2nd stage

Given that each Censos excel file contains two spreadsheets that we will make use of (*Eclosões* and *Masterfile*), we must merge the two based on one unique key, which in this case is the code for the nest (unique for each nest). This gives us one dataset for each nesting season containing all the matched observations, listed as *tartarugas_v2, tartarugas_v3, tartarugas_v4, tartarugas_v5.*

3rd stage

To enrich the data set, I have made use of historical weather[10] data for each month contained on each dataset. This data is joined witch the respective dataset using the date column.

The same effort is made for data referring to the longitude and latitude coordinates of each beach, that in hand provides me with a way to explore maps and to further pattern discoveries.

At this stage, I also took to filling in for missing values at a cut-off of 5.0% on the total number of observations in each column. Given the python-based approach of this thesis, either the mean and mode functions of the Pandas library for DataFrame columns were used or it was resorted to create a logical sequence through the select function from the Numpy library.

[10] All weather data was collected on www.timeanddate.com/weather while beach coordinates came from google maps.

4th stage

The penultimate task is to join all 4 datasets and concatenate them into one. This data set will contain all rows from the 4 sets created in stage 3.

5th stage

To achieve the final set, we need only to filter out each turtle specie to its own data set. Thus, we end up with tartarugas_CM, tartarugas_DC and tartarugas_EI. Just like it was explained at the beginning of the introduction, for this thesis, the filtered data set in usage on the CM sea turtle, tartarugas_CM.

A summary view is displayed below:

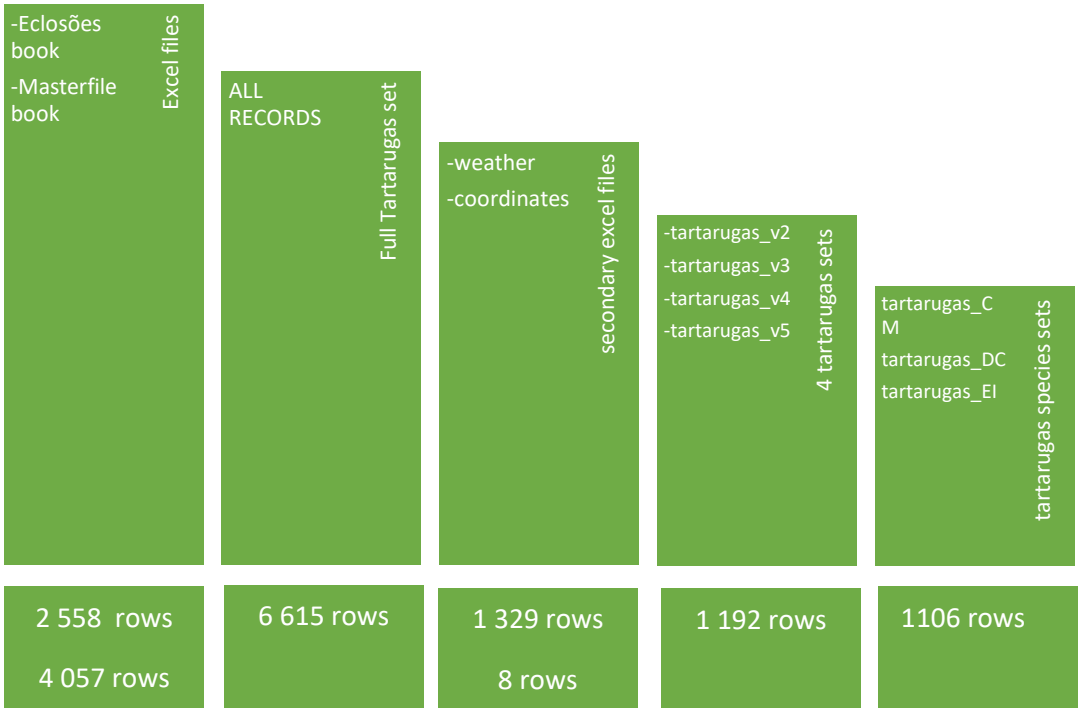


Figure 5 Summary framework for dataset transformation

In ANNEX IV the reader can see a more detailed view on the data manipulation effort with more explanation on what each step entails.

The final set that I will use contains 39 columns and 806 rows. All referring to the Green turtle specimen and its nests from the 2016[11] to 2019.

[11] The data rows referring to the year 2015 had to be dropped to the high amount of missing values.

4.2 Variable set

Now that we have obtained a fully structured dataset, it is necessary to understand what its contents are. Without going into too much detail, the next step will show what important variables are contained inside the dataset and in this case, how some of them relate to the calculation of the so important *Sobrevivencia* target variable (Survivability rate).

Variables referring to egg status:

Variable:	Definition:
$Total\_Ovos_i$	Total number of eggs belonging to one nest, obtained as $Cascas_i + Ovos\_Não\_Eclodidos_i$
$Cascas_i$	Number of eggshells belonging to one nest
Com_gema_i	Number of eggs with yolk in them belonging to one nest
$com\_embriao_i$	Number of eggs with embryo in them belonging to one nest
$Crias\_libertadas\_no\_mars_i$	Number of younglings who made it to sea from each nest obtained as $Cascas - Ovos\_Não\_Eclodidos_i$
$Crias\_mortas_i$	Number of dead younglings that didn't make it to sea
$Crias\_vivas_i$	Number of younglings who made it to sea from each nest (serves as comparison to released younglings)
$Ovos\_Não\_Eclodidos_i$	Number of unhatched eggs belonging to each nest
$Anomalias_i$	Number of eggs in each nest with anomaly

Table 8 referring to egg status

Variables referring to nest status:

Variable:	Definition:
$Profundidade_i$	Depth of the nest in centimetres
$Largura\_do\_ninho_i$	Width of the nest in centimetres
$ninho/tentativan/t/ml_i$	Nest was made or attempted
$Aberto\_antes\_de\_emergênc/N)_i$	Opened before hatching, yes or no
$zona\_maré/vegetação_i$	Nest is shore or vegetation area

Table 9 referring to nest status

Variables referring to turtle anatomy:

Variable:	Definition:
$comprimento\_carapaça_i$	Turtle length in centimetres
$largura\_carapaça_i$	Turtle width in centimetres

Table 10 referring to turtle anatomy

Variables referring to nest predation

Variable:	Definition:
$Predacao_i$	1, if the nest suffered predation, 0 otherwise
$Predador_i$	If $Predacao_i$is 1, what animal is responsible for the predation

Table 11 referring to nest predation

Variables referring to geographical and climate factors:

Variable:	Definition:
$praia_i$	Beach where nest was made
$zona\_ilha_i$	Area of the beach where nest was made, North or South
$storm_i$	If 1, storm happened during month of hatching, 0 otherwise
$Precipitation_i$	Average level of precipitation during the period of incubation in millimetres
$Wind_i$	Wind average during month of hatching in miles per hour
$Mintemp(Cº)_i$	Minimum temperature during month of hatching in Celsius
$Maxtemp(Cº)_i$	Minimum temperature during month of hatching in Celsius
$Avg\_weather'_i$	Average temperature during month of hatching in Celsius

Table 12 referring to geographical and climate factors

Variables referring to date and time:

Variable:	Definition:
$DATA_i$	Date variable of when nest was open
dia_i	Day variable of day when nest was open
$mês_i$	Month variable of month when nest was open
ano_i	Year variable of year when nest was open

Table 13 referring to date and time

Variables referring to extra factors or ambiguous in definition:

Variable:	Definitiaon:
$'\_Observações'_i$	Notes of information on nest opening

Table 14 referring to other variables

Target variable calculation

Sobrevivencia, which is the Sea turtle nest survivability rate is calculated by subtracting the *Mortalidade* variable (mortality rate) that is already existent in the dataset, to 100.

This of course means all survival values will exist between 0 and 100, such that:

Variable:	Calculation:
$Mortalidade_i$	(Crias_mortas $_i$+ Ovos_Não_Eclodidos$_i$)/ (Total_Ovos$_i$)*100
$Sobrevivencia_i$	$100 - Mortalidade_i$

Table 15 calculation of mortality and survivability

Final Preparation

Before starting, I ensured I remove all columns whose values serve as identifiers or labels for the filter that distinguishes either the specie or the nest types. I have also removed variables that were related too closely (or even used) to calculate the survivability variable. This is to, of course, avoid extreme correlation between variables and to also not jeopardize the stability of the future models by having variables contain too much explanatory power.

As such, the following columns are removed from consideration:

- *'Mortalidade';*
- *'codigo_ninho_x';*
- *'codigo_ninho_y';*
- *'espécie_key';*
- *'key';*
- *'_Observações';*
- *'ninho/tentativan/t/ml';*
- *'Crias_mortas';*
- *'Total_Ovos';*
- *'Cascas';*
- *'Ovos_Não_Eclodidos';*
- *'Crias_vivas';*
- *'Crias_libertadas_no_mar';*

4.2 Exploration

In this section I will be looking to not only explore the data that I have available with graphical and theoretical support, but also to discover what variables play an important role in predicting STNSR. This also implies that variable transformation was applied when necessary.
It will also be at this stage that I will proceed with outlier removal and feature selection before I introduce the final set to the algorithms.

4.2.1 Outlier removal

Given I am running the data set though a set of algorithms that will look to learn interesting patterns in data, it is important to consider the existence of extreme or nonsense values that may exist in our data. Specifically, I am referring to values that could induce our algorithms into learning uninteresting patterns that lead to poor feature selection. Failure to do so, could have as a direct consequence the establishment of poor predictors and undermine the whole analysis. That being said, below are the histogram and boxplot graphical views for the variables where outliers were removed. The reasoning behind each selection is given after the graphics:

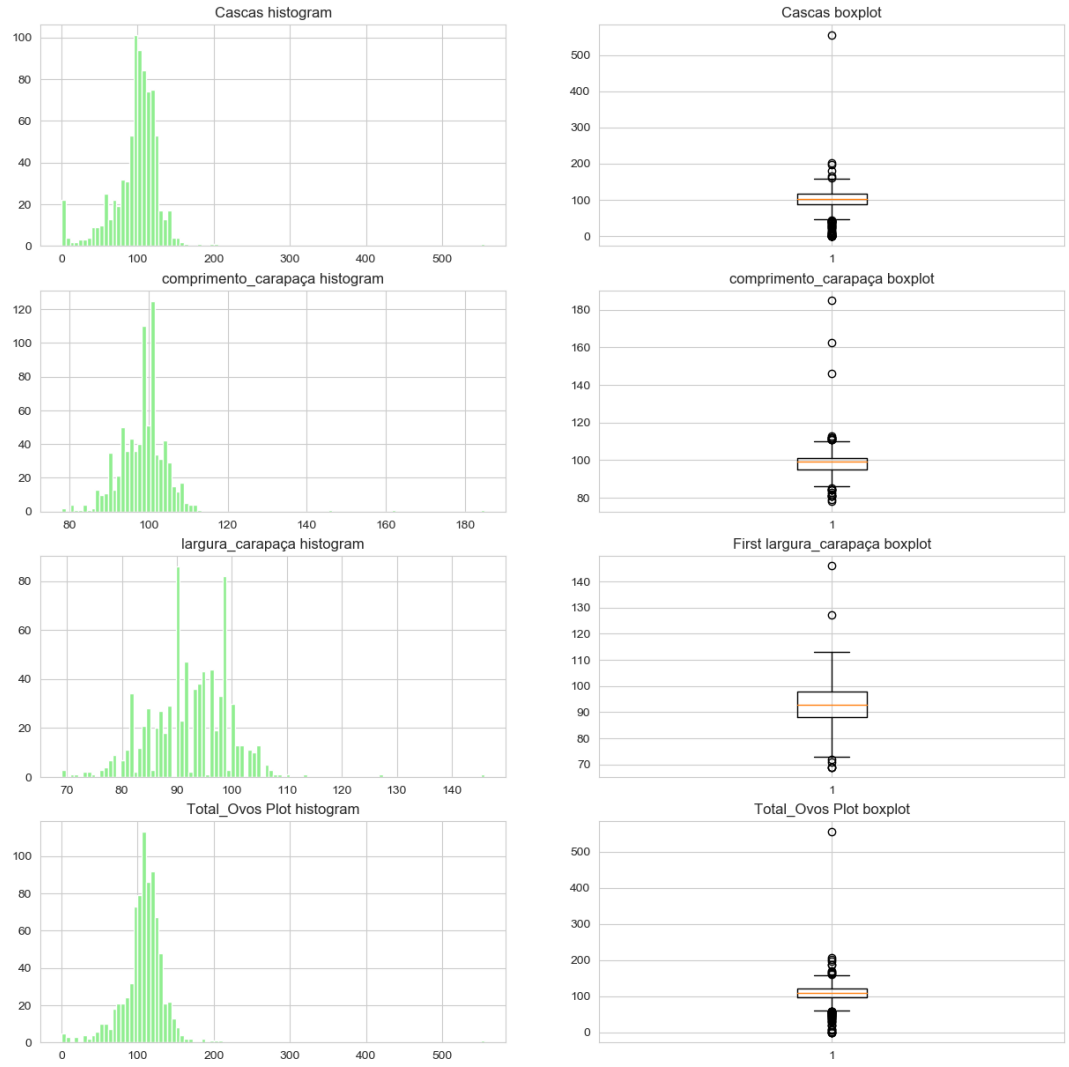


Figure 6 summary plot for outlier removal

As it can be seen, 4 variables have been selected for outlier removable:

Cascas (number of shells in nest) – I found values residing on the 500 plus number of shells. This is an odd occurrence given 500 shells in one nest is atypical in CM nesting behaviour. That would mean either a turtle has laid around 500 eggs or that 2 or more nests were made on top of each other. This is not an issue that needs further consideration and is admittedly removed from my analysis.

comprimento_carapaça (turtle length) – here one must take note on of any odd values on the anatomy of turtles. In this case I found total length values not befitting of a CM turtle anatomy and consider them as miscategorization of the species. Most likely this is a leatherback turtle, due to its bigger size.

largura_carapaça (turtle width) – Same reasoning behind *comprimento_carapaça*. I found values that would make more sense to belong to other turtle species than the CM, as such I remove these values from the dataset.

Total_Ovos (total egg count in nest) – It is important to provide special attention to this variable due to its importance in the calculation of survival and mortality rates. Extreme values found on this variable would have an understandable impact on the calculation of the two main rates (mortality and survivability), thus they are removed.

Conclusion

In the end I removed 5% of all observations in the dataset, going from a total of 806 variables to 785. I slightly exceeded the recommended rule of thumb for outlier removal that establishes that one should remove about 3%, but I considered that I should be less conservative, given that for my prediction efforts I use algorithms that are prone to overfitting.

4.2.2 Correlation matrix

Correlation arises as a major factor in model creation. Since our goal is to find what variables help us obtain the most accurate predictions, it is necessary to eliminate highly correlated [12]variables that will provide the same level of explanatory power. This in turn allows us to avoid redundancy and running into multicollinearity problems.

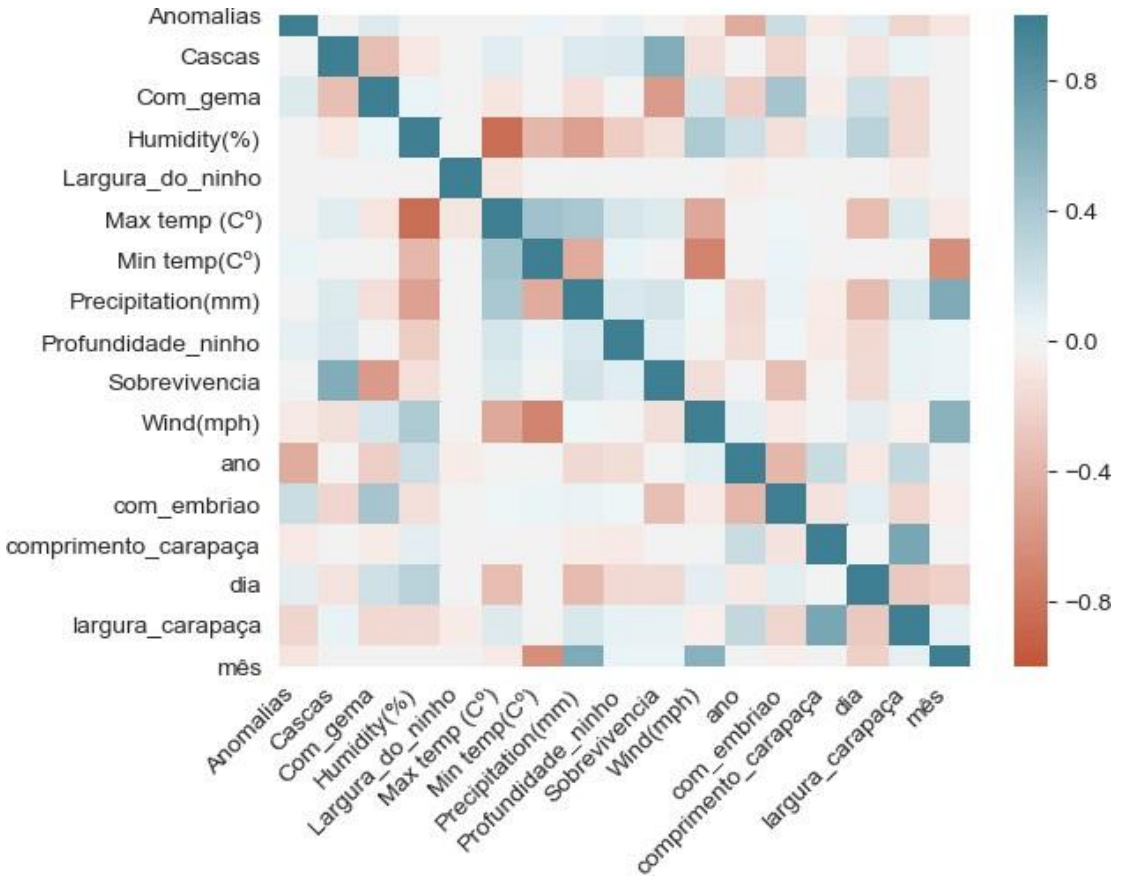


Figure 7 correlation heatmap

Apparently, the variable referring to the average level of humidity during the month of hatching is highly correlated with the maximum level of temperature obtained during the month of hatching. From this point onwards, I will proceed with caution and defer to use one variable or the other to build my models, depending on the context.

[12] In ANNEX V we can see the correlation scores to resolve ambiguity issues.

4.3 Feature extraction

When considering the execution of a benchmark, it is worthwhile to consider running feature extraction methods that aim to reduce the dimension of the original data whilst maintaining a proper amount of variance[13]. With this in mind, I ran a Principal Components Analysis (PCA) and a Support Vector Machine in order to understand if it was worth the effort. The results of the analysis (ANNEX VI) together with the previous understanding that there isn't a high number of features or observations to begin with, has led me to put aside this step and proceed with the feature selection process.

4.4 Feature selection

In literature, feature selection has brought mainly two important benefits to an analytical effort. One being the strengthened performance on predictions, due to having a subset of variables instead of all the main ones provides faster and cost-effective predictors. The other reason is connected to the improvement on our understanding of the data, as looking into what features prevail inside the grander selection pool may show us what factors play a bigger role in the formation of our predictions.

Training and testing set

For the purpose of training the models and later testing on them, I split the data in 80% for training and 20% for testing. I preferred to consider taking in more observations for training than the typical "rule of thumb", because I considered that given the size of the dataset I had to make sure that enough variability existed inside the training set for interesting features to be selected.

4.4.1 Linear regression

The first step I decided to take was to analyse the full set of variables available to me and run the Ordinary Least Squares (OLS) on a linear regression.

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{ip} + \varepsilon_i, \qquad i = 1, \dots, n,$$

Where β_0 is the intercept and ε_i corresponds to the error term.

Testing for Heteroscedasticity

The existence of Heteroscedasticity [38] is a concern for OLS, as it is established that statistical inference cannot be made because true variance and the covariance are underestimated. As direct consequence, coefficients and statistical test values are invalid.
For the purpose of statistical inference and to obtain a correct estimation on the explanatory power of a full set of variables I have decided to run a test for Heteroscedasticity concerning the target variables prediction. My logic is that if I am incurring on Heteroscedasticity, this would mean I might have an issue with omitted variables or poorly specified variable structure that could be related to an unforeseen trend in the data (subchapter 2.3.4). For the purpose of this test I will use the white test.

White test – a typically used statistical test that establishes whether the variance of the errors in a regression model is constant, which is the opposite of Heteroscedasticity.

[13] What this "proper" amount of variance means depends on the study

Metric	Definition	Value
F-test p-value	Tests individual impacts that might affect the explained variance	0.138
LM-test p-value	Score provided on the gradient likelihood function	0.178

Table 16 summary for heteroscedasticity test

Given the results are p > 0,05 It does not seem the model is incurring on Heteroscedasticity, allowing us to assume that the variance and covariance are correctly calculated. This means we do not incur on omitted variable bias, nor do we have to necessarily suspect any issue with the specification of the model. This concludes that the metrics that result out of the OLS estimation are correct and interpretable.

Variance explained

In statistics, variance is the expectation of the squared deviation of a random variable from its mean. In this context, it will refer to the proportion to which this model accounts to dispersion (variation) as explained variance.

In essence, the explained variance will provide me with an idea of how well the use of all variables is performing on the prediction.

R-squared (R^2)

The R^2 a statistical measure referring to how close the data points are to the fitted regression line. In essence a high R^2 means that our model fits the data well:

$$R^2 = \frac{\sum_i (f_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_{res}}{SS_{tot}}$$

$\sum_i (f_i - \bar{y})^2 = SS_{res}$ is the explained sum of squares residual and $\sum_i (y_i - \bar{y})^2 = SS_{tot}$ is the total sum of squares. The former refers to the optimality criterion that establishes the discrepancy between the data and the given estimation model. The latter is the squared differences between the observations and their overall mean

It is sufficient to say that higher R^2 values are preferable, meaning that a model with a score of 100% represents a model that perfectly explains the relation between the dependent variable and the independent ones.

Metric	Model estimator	Value
R-Squared	OLS	0.418

Table 17 summary for OLS r-squared

Having just about 42% of variance explained leads me to believe we are missing on significant amount of explanatory power. In fact, it shows how one cannot expect the current format of features to predict STNSR to a more complex degree. In other words, it means that the addition of more independent variables that could impact STNSR will be a very interesting case study for future developments.

My other important conclusion is that it is pertinent to consider that feature selection should be conducted by use of other methods that will look to better understand the variance inside the

data. My other conclusion is that I should continue to aim for as much data interpretability as possible, given there will be no significant specification changes.
Having this in mind, the coefficient signals obtained by the results of the OLS (ANNEX VII) will show me if a variable has a positive or negative impact on the target variable. The plan is to make further use of this information while focusing on assessing the most important features for the models.

4.3.2 Recursive feature selection with cross validation (RFSCV)

My first approach to feature selection is to understand what is right number of features the algorithms should be provided with. For this purpose, it is expected of the RFSCV to provide a view on what number of features I am able to maximize negative mean squared error of the model. The reason negative MSE is used is because the unified scoring API in this Cross Validation always maximizes the score, so scores which need to be minimized are negated in order for the unified scoring API to work correctly. The correct value of the MSE is simply the positive version of the value obtained.

The cross validation will run different samples of data through the models and provide me with an iterative score on how well a set of features performs on the subsample.

To implement my RFSCV it is decided to run the cross validation on 5 *Random Forest* with different parameters on splitting (2, 3, 4, 5, 6) being applied on all of them. My intent is to use algorithms that don't differ drastically from each other, but that do indeed show an underlying difference on how they split the data. Below is the graphical representation of the RFSCV results:

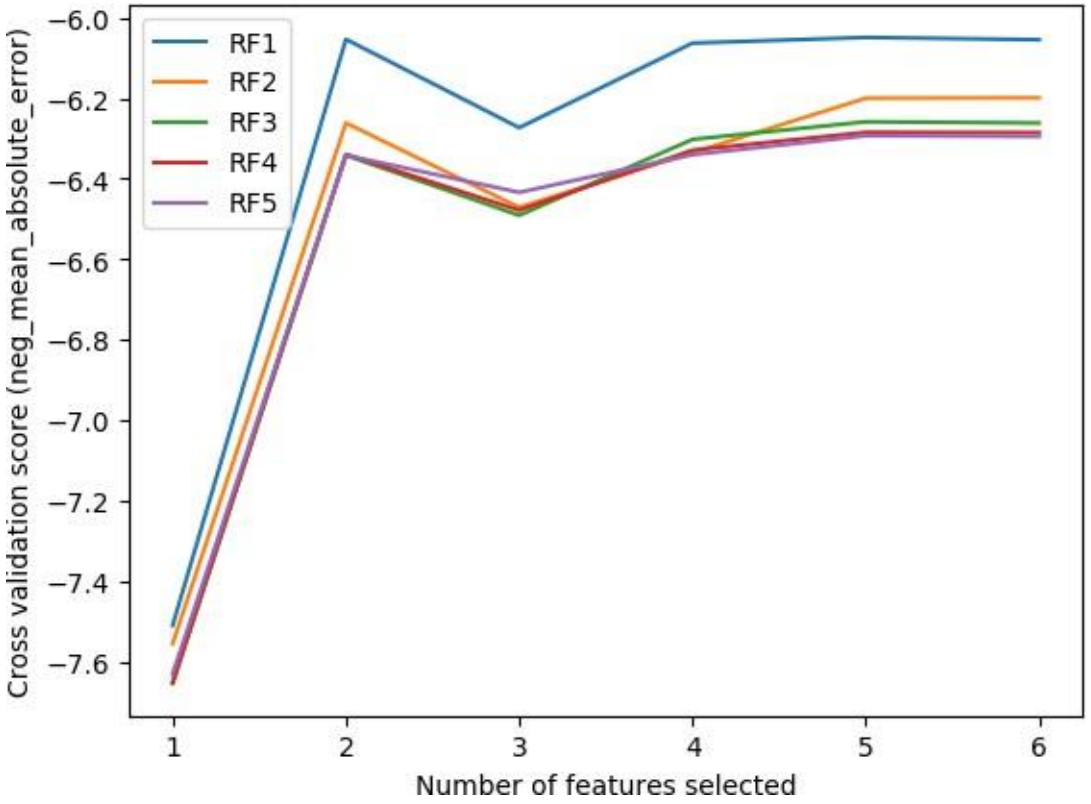


Figure 8 plot for recursive feature selection

From the graphical representation we can see that the higher values on split which are given by Random Forest 1, 2 and 3 that recommend the use of 5 features. This means I will need to consider carefully from this point onwards that I might not use the full set of variables, opting instead to use just 5 features to proceed with the estimation of STNSR. The main goal is assuring we have heterogeneous data so that the algorithms can perform accurate predictions. As such, more specific feature selection analysis is made from his point onwards in search of the best approach, but this time I will look to see what are the features that serve me best and rank them.

4.3.3 Boruta package

Having a variable worth gives us the understanding of how much variance is explained thanks to it. As such, while one looks to not only obtain a smaller set of features, we should also distinguish them based on their importance. For that intent I make use of the Boruta package for Python.

Originally a feature selection method used in R language, it was recently recoded into Python language and renamed as BorutaPy. Predictor values are shuffled and gathered with the original predictors. It then runs the merged dataset through a *RandomForest*. The calculated variable importance difference between the randomised variable with original variables is made. Finally, the original variables that hold higher importance than the randomised variables are selected and displayed.

In my case, I have created a feature selector with Boruta that uses a Random Forest regressor as a base estimator. Thus, the parameters[14] will look as the following:

BorutaPy		
Parameters	**Definition**	**Values**
estimator	A supervised learning estimator, with a 'fit' method that returns the feature_importances_ attribute.	Random Forest
verbose	Controls verbosity of output.	2
n_estimators	Sets the number of estimators in the chosen ensemble method. If 'auto' this is determined automatically based on the size of the dataset	2
random_state	The system state that will split the data. If integer value, it allows to replicate the experiment with the same split	1
max_iter	The number of maximum iterations to perform.	50
perc	Instead of the max, ituses the percentile defined by the user, to pick our threshold for comparison between shadow and real features.	90

Table 18 boruta parameter tuning

[14] The parameters were the recommended ones from the https://github.com/scikit-learn-contrib/boruta_py. Due to the package being fairly new to python, changing the parameters too much might induce in execution errors.

Random Forest		
Parameters	**Definition**	**Values**
n_jobs	The number of jobs to run in parallel. fit, predict, decision_path and apply are all parallelized over the trees.	1
min_samples_split	The minimum number of samples required to split an internal node:	2
max_depth	The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.	5

Table 19 boruta random forest tuning

After running the data set through the Boruta pipeline we are left with the following variables:

'Anomalias', 'Com_gema', 'Largura_do_ninho', 'Precipitation(mm)', 'Profundidade_ninho', 'com_embriao';

Again, it seems it is recommended to use 5 variables instead of the full set. This time though, we can see which ones they are. To fully understand why the RFSCV and the Boruta feature selection have guided me this way and will look to check the worth of each variable.

4.3.4 Random Forest variable importance

Given Boruta makes use of a *Random Forest* to establish its features. I decided to run my feature selection through a similar Random Forest but this time, to extract from it the variable importance for each of them.

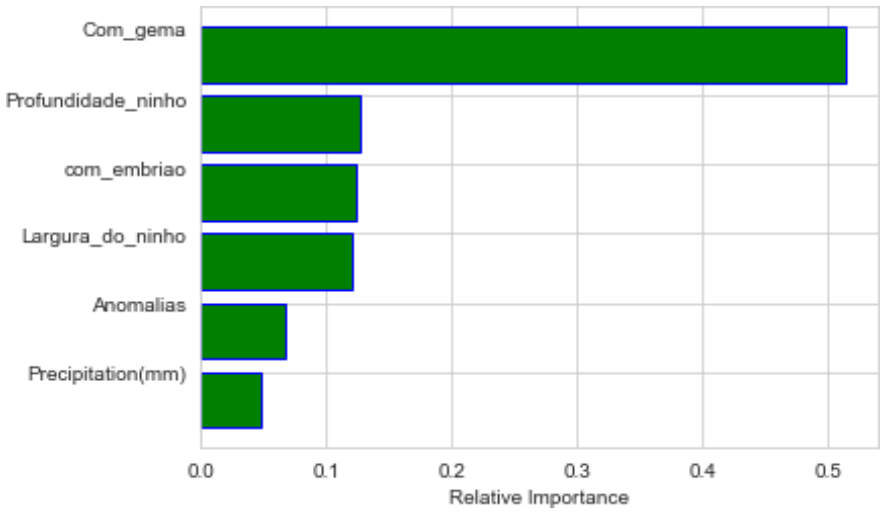


Figure 9 plot for relative variable importance

The reasoning behind the decision to check variable worth for 6 variables instead of the 5 recommended ones, is because it was noticed in another run with the full data, that *Largura_do_ninho* (nest width) was very close to *Profundidade_ninho* (nest depth). Without wanting to lose any important information on splitting it is best to run the two together for further analysis.

As seen in figure 8, one is able to retrieve a ranking of the most important variables. Apparently, the variables referring to eggs with yolk (*Com_gema*) and eggs with embryo (*com_embriao*) take centre stage. In fact, the two alone already provide a lot of explanatory power to the model.
Precipitation and *Anomalias* are considered the less important of the set with relative importance below 0.1%, but the two variables referring to nest size show me that *Profundidade_ninho* (nest depth) should be chosen over *Largura_do_ninho* (nest width) when considering reducing the model to 5 features. Given that the Boruta feature selection has shown me that nest depth provides better splitting of the data, only a deeper analysis on both can provide better understanding on which one will be used in the models.

4.3.4 Subset variable assessment

Anomalias

According to the explanation given by the surveying context, this variable refers to the number of eggs in a nest that present some sort of oddity. The causes can range from a number of factors, e.g damaged eggs removed from the nest by unknown causes, egg fall damage, egg color, to name a few. It is indeed a complicated variable to assess and would need for each case to be considered and debated together with the conservation team in order to establish the right causes for each anomaly. Therefore, it is best to instead focus in the impact this variable has on the survivability rate and approach it from that angle.

Looking into the histogram for this variable we can see what most nests do not suffer from an anomaly with over 400 eggs being clear of any condition vs the less than 200 total that present some sort of oddity. Numbers higher than 1 show that there was a focus from the surveying team to detail exactly how many eggs from each nest show the uncharacterized display.

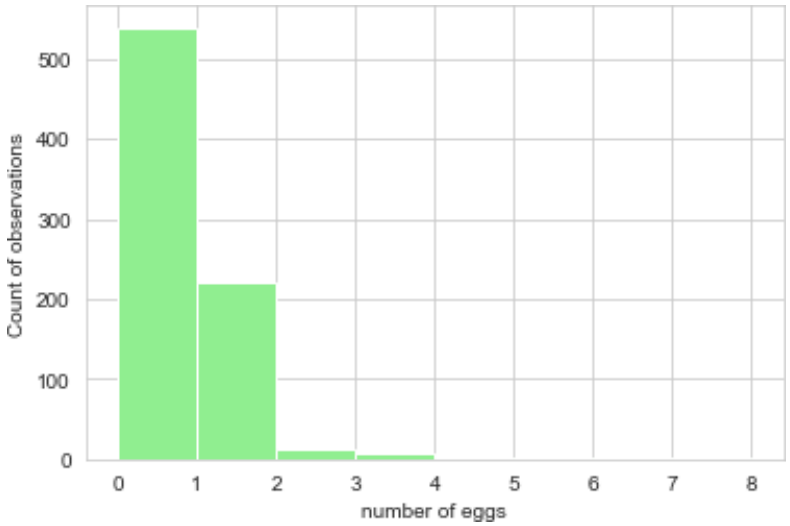


Figure 10 histogram for anomalies

By looking to the histogram, we can see that most nests do not tend to contain an abnormality. That being said, it is important to assess the volume of abnormalities in proportion to the number of nests. For this purpose, a pie chart follows:

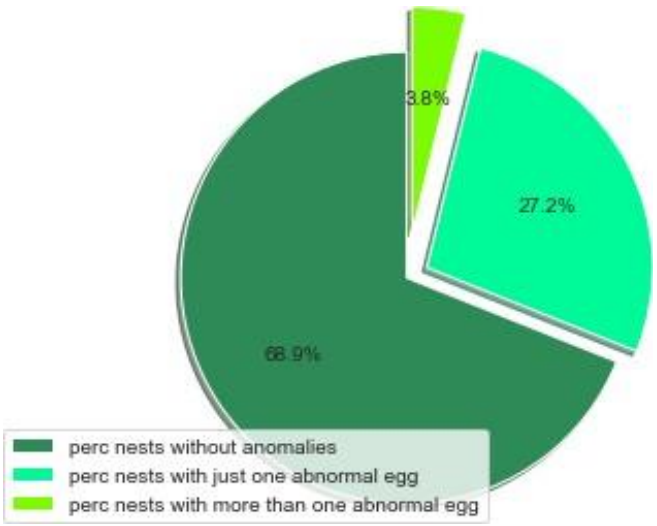


Figure 11 pie chart for nests with abnormalities

By looking into the pie chart, we can see that we indeed do have roughly 31.3% of nests with at least one egg demonstrating some sort of odd behaviour. Yet inside the grand scheme of things, it does not appear that, on average, a nest tends to have more than one abnormal egg.

Concluding, it is quite interesting to see that this variable has been selected as it supports the idea that we need more dimensions on STNSR. Given each case deserves to have its own analysis on what could be the causes and symptoms that constitute an abnormal egg, it would be too taxing to delve deeper into the behaviour of this variable. From my perspective, this situation requires a branching off from the main goal STNSR prediction. Having said that, *Anomalias* will be considered for use in the estimation effort.

Com_gema (egg with yolk) and Com_embriao (egg with embryo)

It should come with little surprise that these two variables have been selected. Both variables refer to an egg state that translates to egg malformation. The reasons that lead to this have to be explained by biological journals detailing the incubation period of a Green turtle egg. Some experiments have been made in order to understand it, but too many details would need to be discussed in order to even theorize a possible cause [40].

Regardless, it is more interesting to understand the connection that these variables have with the survival rate.

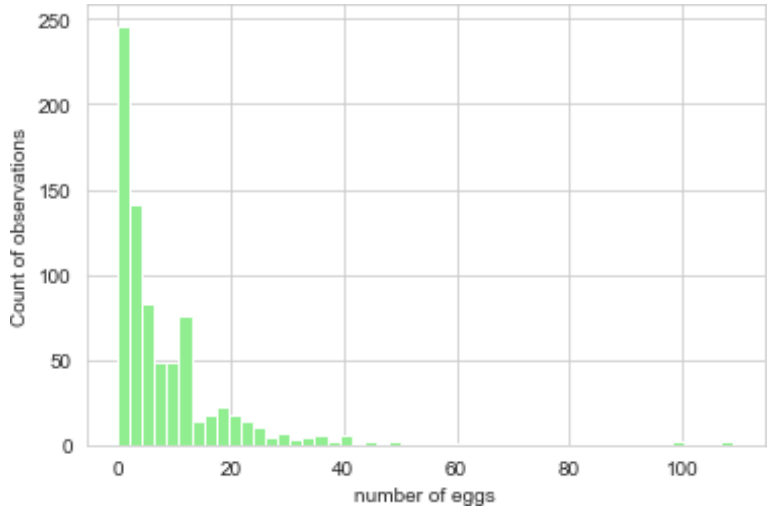


Figure 12 histogram for eggs with yolk

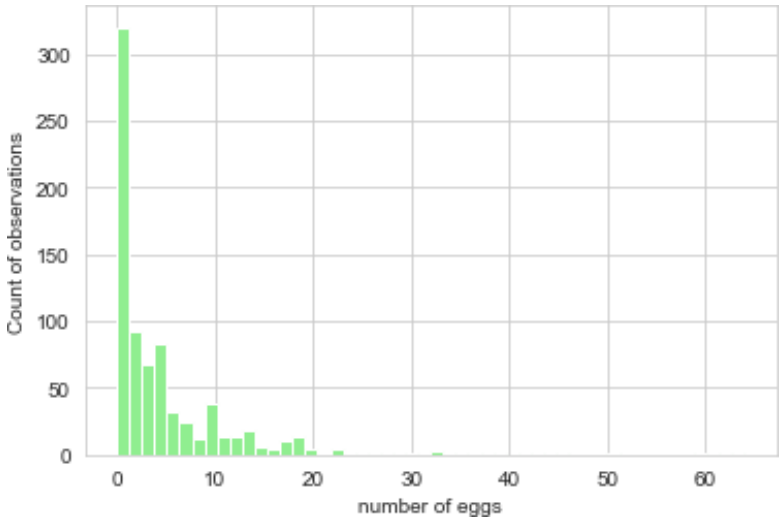


Figure 13 histogram for eggs with embryo

The logic here is simple, the more eggs left with yolk or embryo means the smaller the survivability ratio. In fact, it is understandable that these variables correlate strongly with the *Sobrevivencia* (Survivability) variable due to that. This would explain the high ranking of these two variables in the Random Forest from figure 8, having detailed records on how many eggs could possibly be malformed, is in itself, a great help for future predictions on STNSR.

Eggs with yolk		Eggs with embryo	
Categories	**Count**	**Categories**	**Count**
[0, 20.0]	560	[0, 10.0]	526
[20.0, 40.0]	54	[10.0, 20.0]	69
[40.0, 60.0]	6	[20.0, 30.0]	18
[80.0, 100.0]	4	[30.0, 40.0]	9
[60.0, 80.0]	2	[40.0, 50.0]	5
		[60.0, 70.0]	1

Table 20 summary for egg status

After converting these two variables to the categorical type, we can see that we do not contain very high frequencies on either egg status. Yet, it is interesting to see that there are more eggs with embryos than with yolk on the higher categories. This means that we have more occurrences where eggs have already started forming an embryo at some stage, only to become stillborn at some point. This further supports an existing suspicion that the factor that might lead to lower survivability rates is occurring somewhere between the first and second weeks of the incubation period.

Concluding, the variable worth assessment has considered these two variables good data splitters most likely due to their variability. Although, as mentioned, we do not seem to have extreme values, there is a certain behaviour that seems quite interesting to observe.

Precipitation (mm)

Here we have a variable that does not come from the original turtle data, but from the weather dataset that was created to supplement the former (as seen in sub chapter 4.2). We know from the description of the meteorological behaviour in Principe that there is a lot of rain falling though most of the year, but more so during the months of October to May. The question is if the data can tell us something about the relation of rain levels has with STNSR.

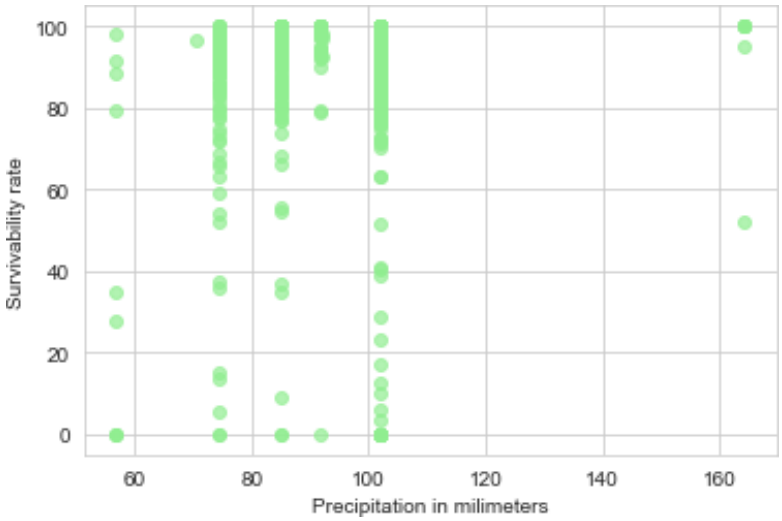


Figure 14 scatter plot for precipitation

The above scatter plot shows us the existence of extreme values on the level of precipitation, but they do not seem to negatively affect survivability rates. In fact, it seems that we cannot take any conclusions based on the observation of the points alone. A more in-depth view into the behaviour of rain levels during the year might give us a better idea on how this variable might impact our target.

Temperature and its relation to precipitation could perhaps give an idea of a pattern or causality:

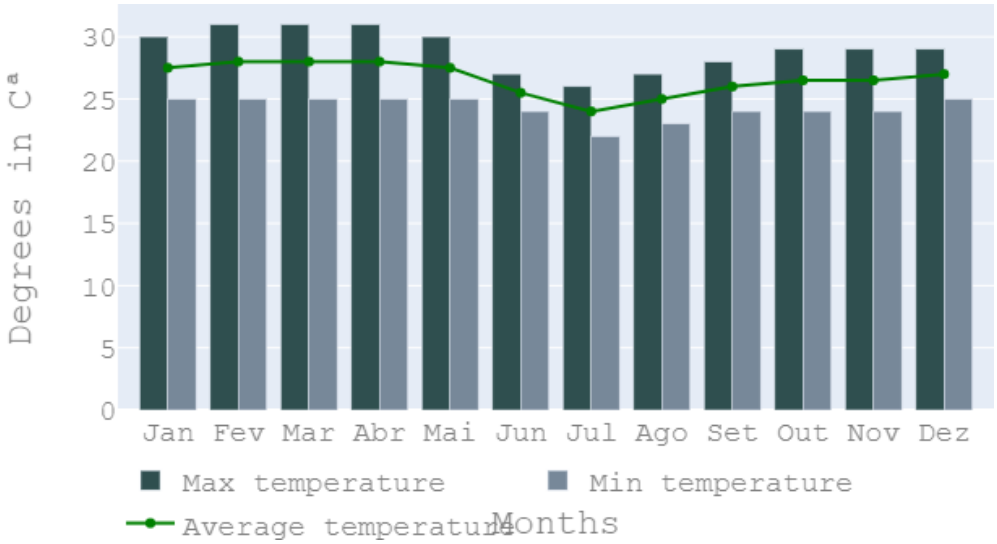


Figure 15 bar and scatter plot for temperature levels per month

Figure 16 scatter plot for precipitation levels per month

In figures 13 and 14 we can see an interesting trend regarding the relation between temperature and precipitation levels. Like in the meteorological description given in the introduction, we can see that there is more rain from October to May, just as Average temperatures drop. This leads me to hypothesize if we can expect different levels of nest survivability rates depending on the month of the year, as that would attest that higher intensities of rain might affect STNSR negatively. Indeed, if we look at the OLS coefficient value (-0.1312) for *Precipitation(mm)* in ANNEX VI, we can see that precipitation has a negative effect on STNSR.

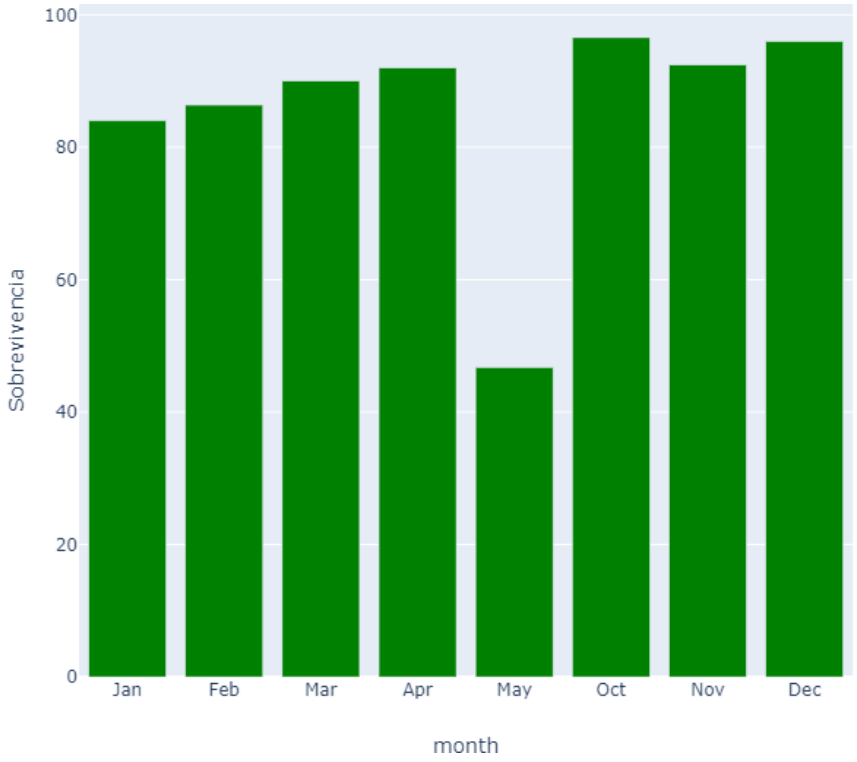


Figure 17 bar plot for survival rate per month

Having in mind that after pre-processing the dataset only contains data from 8 out of 12 months of the year, one can see a very obvious drop on survival rates in May, the final month where sea turtle eggs hatch. For the rest of the months, there doesn't appear to be a particular and unexpected trend. As the by far busiest nesting 4 first months of the year progress, survival rates tend to increase in linear fashion. It would seem that there is a proportional relation between how much rain falls on the nests and their survivability.

Profundidade do ninho (nest depth)

A turtle nest is made by the mother's use of her back legs. Their instinctive intent is to produce a hole that is both wide and deep enough to permit a healthy incubation for the eggs. It is however, seldom occurring that a sea turtle might dig too deep a hole making it more difficult for eggs to hatch and make their way to the surface. It is also possible for nests to be dug to an almost superficial level, leaving the eggs more exposed to outside factors like weather or predators. [9]

Given this occurrence, it wouldn't be very reasonable not to refer to *Largura_do_ninho* (nest width) at the same time as analysing this nest depth. Both serve to calculate the size of the nest where the eggs lay, and it is natural to assume that there is a correlation between the two. Yet this is not the exact case, in subchapter 4.4.2 where it is possible to see the correlation between the data set variables, the two only garner a positive correlation of 0.11. This doesn't mean in no way there isn't a connection between the two, but it is important to go into further detail as to why *Profundidade_do_ninho* prevailed as a selected feature.

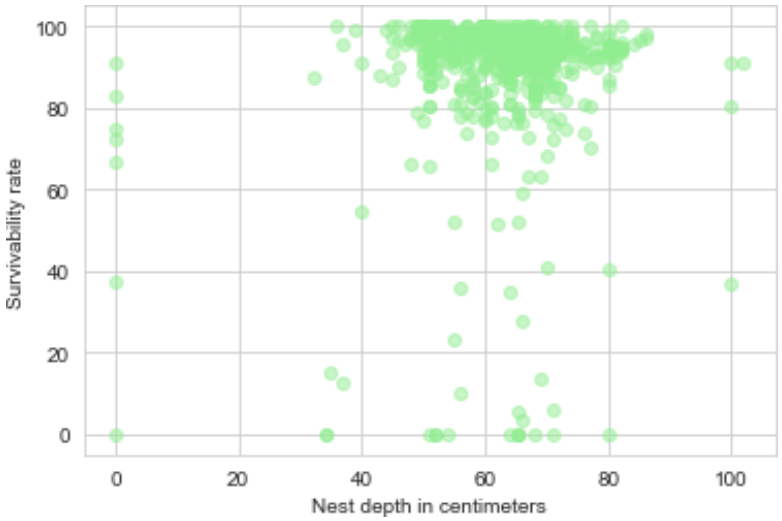


Figure 18 scatter plot for nest depth

Above, the reader can see the scatter plot of turtle nest depth, a representation of all points referring to the size in centimetres of the turtle nest. The values at 0 represent nests that have been unsurfaced and present a large level of exposure of eggs to outside factors.

Nest depth:	Rate of nest survival
Equal to 0 centimetres deep	62.42%
More than 0 centimetres deep	82.02%

Table 21 table summary on egg size

As we can see from table 12 there is a significant drop in nest survival rate if a nest does not present significant depth.

This leads me to believe that the feature selection model considered these values to be pertinent for the global assessment. Yet, by looking below to the nest width scatter plot, it is possible to notice another factor in play.

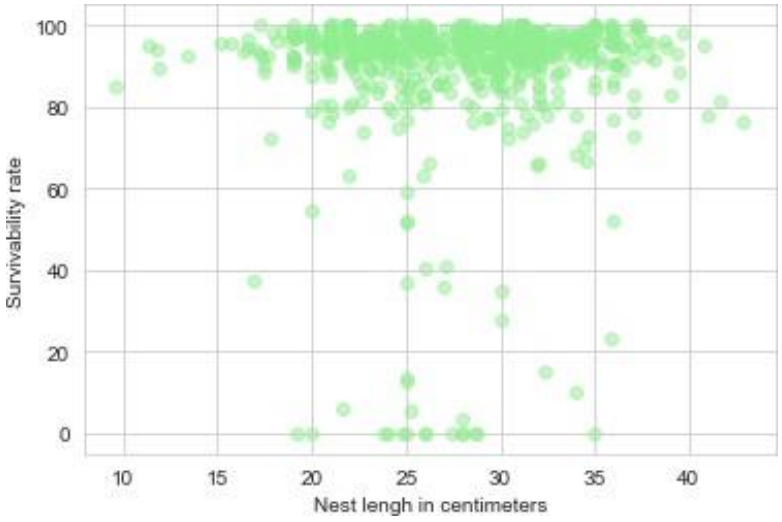


Figure 19 scatter plot for nest length

Here we can see that nest width doesn't seem to present any extreme values outside its larger cluster of points located on the middle top segment. Furthermore, it can be assumed that a sea turtle might be more inclined to neglect the depth of a nest, than to ignore its width, as it requires lesser effort from the mother.

Conclusively, nest depth is preferred instead of nest width, as it can be considered that it will provide better pattern discovery for the algorithms. This way, the suggested subset of 5 feature are as follows:

'Anomalias', 'Com_gema', 'Largura_do_ninho', 'Precipitation(mm)', 'Profundidade_ninho', 'com_embriao';

5. Experiment

5.1 Benchmark

At this stage, an exhaustive search is conducted for the best parameterization to use on the 6 ML algorithms that are applied in this thesis. For this intent one can make use of a Grid Search. Just like in Feature Selection this makes use of cross validation, yet here we are running several combinations of parameters and looking to rank them by the measure of their performance on a given metric.

Data standardization

Some ML algorithms require input standardization before training. This is the case of ANNs. It is known that input standardization allows gradient-based estimators to achieve better results as the search-space gets less rugged. In this case, it is best for the ANN to use normalized values, which means having the same range of values for each of the inputs so that the network may properly converge the weights and deal with bias. For the purpose of achieving normalization, we can make use of the SKLEARN Preprocessing package tool called Standardscaler that standardizes each input variable with centre equal to 0 and standard deviation equal to 1.

Evaluation metric

I made use of the Mean Squared Error (MSE) for this purpose. Since the MSE will be mentioned again right at the beginning of the next chapter, but with more detail as to what it is and how to calculate it, here we will only refer to how it was used. In short, the lower the value the better.

In this scenario we have obtained a base MSE for my model of 355.83. Iterations that beat this score are all to be considered moving forward, although ultimately, we are looking for the smallest MSE values overall.

Running the Benchmark

Next, the goal is to run several iterations on 6 different algorithms, where the parameters change at each cycle. Thus, a different set of MSE values will be obtained for each algorithm. In the end, for based on the lowest score obtained for each algorithm, the best set of parameters will be chosen.

The following tables (22 to 27) present the grid of explored parameters along with the parameter-set which achieves the highest expected generalization ability. For each algorithm, the first column, entitled as "Parameters", provides labels for each hyper-parameter[15]in the same exact order as it was presented in subchapter 2.4. The second column, entitled as "Values", represents the set of explored values for a given hyper-parameter. The third and the last column will contain the selected parameters that will be accepted as the best generalizers.:

[15] notice that the labelling follows Sklearn's nomenclature

5.1.1 Artificial Neural Network

Parameters	Values	Grid search
hidden_layer_sizes	[(50,), (100,), (10, 10),(50, 10), (50, 50), (10, 10, 10)]	(10,10,10)
max_iter	[500]	500
alpha	[0.0001, 0.001, 0.01, 0.1]	0.0001
tol	[0.01, 0.001]	0.01
beta_1	[0.1, 0.5, 0.90]	0.9
beta_2	[0.1, 0.5, 0.90]	0.1
learning_rate	[0.001, 0.01. 0.1]	0.001

Table 22 parameter tuning for ANN

5.1.2 Bagging or bootstrap aggregation

Parameters	Values	Grid search
n_estimators	[50, 100, 200, 300]	100
max_samples	[0.50, 0.75, 1.0]	1.0
max_features	[0.50, 0.75, 1.0]	0.5
base_estimator	DecisionTreeRegressor(max_depth=3, min_samples_split=25)	DecisionTreeRegressor(max_depth=3, min_samples_split=25)
	DecisionTreeRegressor(max_depth=4, min_samples_split=25)	
	DecisionTreeRegressor(max_depth=5, min_samples_split=25)	

Table 23 parameter tunning for bagging

5.1.3 Random Forest

Parameters	Values	Grid search
n_estimators	[50, 100, 200, 300]	300
max_depth	[3, 4, 5]	5
min_samples_split	[25]	25

Table 24 parameter tunning for Random Forest

5.1.4 Adaptive boosting

Parameters	Values	Grid search
n_estimators	[50, 100, 200, 300]	50
learning_rate	[0.01, 0.1, 0.5, 1.0]	1.0
base_estimator	DecisionTreeRegressor(max_depth=3, min_samples_split=25)	DecisionTreeRegressor(max_depth=3, min_samples_split=25)
	DecisionTreeRegressor(max_depth=4, min_samples_split=25)	
	DecisionTreeRegressor(max_depth=5, min_samples_split=25)	

Table 25 parameter tunning for Adaptive boosting

5.1.5 Gradient boosting

Parameters	Values	Grid search
n_estimators	[50, 100, 200, 300]	300
learning_rate	[0.01, 0.1, 0.5, 1.0]	1.0
subsample	[0.50, 0.75, 1.0]	0.5

Table 26 parameter tuning for Gboost

5.1.6 Xgboost

Parameters	Values	Grid search
n_estimators	[50, 100, 200, 300],	50
eta	[0.01, 0.1, 0.5, 1.0]	0.01
subsample	[0.50, 0.75, 1.0]	0.5

Table 27 parameter tuning for Xgboost

5.2 Evaluation of results

Having prepared all parameters for our models and established what features will be fed to them, we can finally proceed to the evaluation of the final results for STNSR and discuss what their accuracy and meaning.

To present the quality of the predictions of the models, the following two tables represent their performance on 2 evaluation metrics: Mean Absolute Errors and Mean Squared Errors.

5.2.1 Mean Absolute Error (MAE)

The MAE is a common global performance measure used for regression algorithms. It is the mean sum of absolute differences of the expected value with the actual value. A formal representation of the MAE is given as:

$$\text{MAE} = \frac{\sum_{i=0}^{n} |y_i - x_i|}{n}$$

Where y_i is the predicted value and x_i is the actual value.

Mean abosolute error	Training - mean (std)	Testing - mean (std)
ANN	11.37(6.82)	17.22(8.92)
RANDOMFOREST	6.66(1.55)	8.45(4.25)
BAGGING	7.91(1.46)	9.74(4.56)
ADABOOST	10.86(1.58)	11.74(4.71)
GBOOST	14.58(2.40)	21.08(6.37)
XGBOOST	6.57(1.54)	7.04(2.89)

Table 28 summary results for MAE

5.2.2 Mean Squared Error (MSE)

Used to measure the quality of an estimator, the lowers the MSE the better, as we are calculating the average squared difference between the estimated values and the actual values:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

With n being predictions generated from an n sample data points on all variables, Y_i the vectors for the observed values on the variable being predicted and $\hat{Y}_i$ being the predicted values.

Mean squared error	Training - mean (std)	Testing - mean (std)
ANN	218.53(88.73)	649.55(551.31)
RANDOMFOREST	214.74(113.55)	260.62(320.05)
BAGGING	239.00(117.98)	312.46(352.42)
ADABOOST	291.20(114.83)	377.71(340.74)
GBOOST	757.23(338.65)	1663.66(1542.73)
XGBOOST	194.01(114.66)	250.15(329.51)

Table 29 summary results for MSE

The reason the two measures are used, is while the MAE provides an easily interpretable value given the actual context of the problem, the MSE's larger values will give a better understanding on the difference of performance between training and testing sets.

5.3 Results

Overall assessment

As a reminder, STNSR is given as a value between 0 and 100. Using the values in table 27, we can say that the MAE of each model is on average how many points away our prediction is from the actual target value. Using the MAE result of 8.45 for the Random Forest as an example, if a nest's survivability rate is actually at 85%, our model has given us a prediction that establishes a confidence interval of [85 - 8.45 and 85 + 8.45] percentual points, represented as [76.55; 93.45]. This means that the model considers that the given nest's survivability rate lays, on average, anywhere in between 77% and 93%.

The MSE will give us an idea in how well the model performed on the overall model. In this case, if the features that were selected were well interpreted by the algorithm and if the calculated variances allow for good residual reduction. Again, the closer the value is to 0, the better. Overall, reasonably high values on MSE have been obtained, which indicate that there is a large number of errors in the prediction. However, in this context the focus in the MSE should be in the comparison of its value in the training and testing sets, in order to understand if the model occurred in overfitting or not.

Ann

In terms of MAE it did not perform as poorly as Gradient Boosting, it did however show overfitting given not only by the considerable difference on the training and testing set, as one can see from the difference in MSE values.

Bagging

A reasonable performance, given the high levels of variance found in the data. As expected, the algorithm did not outperform Random Forest and Xgboost. However, it is interesting to notice that id did not perform that badly while maintaining the third lowest level of overfitting from the 6 algorithms.

Random Forest

The second-best performer out of all the algorithms (only outdone by the powerful Xgboost). Given how the algorithm words, it was expected of it to avoid overfitting (smaller difference between training and testing set means) as well as performing adequate predictions on the data. This is indeed what happened, with its MSE being the second smallest out of all.

Adaboost

Another algorithm that has made it into the middle of the group. Possibly some more parameter tuning could've improved the prediction values. But given there are better results for 3 other algorithms, this is not something considered particularly concerning. We also say that some level of overfitting occurred in the model, although not as drastically as say the ANN for the Gradient boosting

Gradient Boosting

The difference in values on both testing MAE and MSE shows that gboost not only made poor predictions on the data, but that it also overfit. This was a possibility that was already considered when presenting its theoretical background, as it comes with little surprise.

Xgboost

The best model in performance for both mean MAE (7.04) and mean MSE (250.15) in testing. In subchapter 2.4 its tendency to outperform other algorithms had already been established and indeed it has done so again. Using the above example as a reference, for a given STNSR of 85%, Xgboost can predict STNSR to be anywhere between a confidence interval of [77.96 ; 92.04] (78% to 92% survival rate). It did however show some levels of overfitting, as it can be seen clearly by the MSE value. In fact, it was outperformed in that department by the Random Forest as it obtained a smaller difference between the values of MSE in the training and testing sets.

6. Conclusion

An entire prediction effort on sea turtle nest survivability rate has been conducted. It is possible for the reader to draw several important assessments from this thesis. Firstly, the importance on the existence of researches such as this one in order to bring better understating on biological conservation not only for the island of Príncipe, but for all faunas and floras. To add to this point, this document provides a good example on how to synergise advanced knowledge discovery techniques with periodical surveying.

The second important aspect to be collected, is the breakdown of the initial Censos source file. Following a step by step approach it was concluded that it was not possible to make use of the full data set for prediction given its high levels of missing values and existence of redundant data. Having finned out the dataset and following an OLS estimation, an R^2 value of 42% denotes that there is already a large number of unexplained variances even if we made use of all the existing variables in the data set.
The next logical step was to then reduce the dimensionality of the of the data to not only simplify the understanding of pattern discovery of the ML algorithms, but to also avoid overfitting issues.

Thus 5 variables and their worth were retrieved from this analysis, leading to a more detailed assessment of their pattern distribution and relation so STNSR. They were:

- *Com_gema* – We understood that this variable is the most important for STNSR prediction. It did not correlate highly with any other variable, and more interestingly even, it does not have any proportionate growth with the variable referring to eggs with embryo;

- *com_embrião* – There seems to be an interesting phenomenon occurring during the middle of the incubation period that leads to eggs not fully developing. Further analysis on possible factors that lead to the eggs being affected in such a way might bring very interesting insights on how to prevent this issue;

- *Precipitation(mm)* – The amount of rain has a negative impact on a nests survivability rate. To what degree depends on further analysis on climate conditions related to precifipation levels;

- *Anomalias* – An interesting feature with peculiar behaviour that deserves its own in-depth analysis to why it causes an impact on survivability rates. Its inclusion in the model is based on the belief that this factor might lead to better STNSR predictions;

- *Profundidade_ninho* – The most interesting conclusion we could take is that a nest's depth is more relevant to STNSR than its width. This becomes particularly relevant on lower survivability rates for higher surface level nests;

To transition to the next stage, it was necessary to conduct a benchmark that would allow for the selection of the best set of parameters for each of the 6 algorithms that are in use in this thesis. This same benchmark can be further scaled and fully deployed for other parameterization processes.
After achieving said selection, an iterative process was executed where each algorithm could choose from a pool of literature supported parameters with the aim of achieving the best possible combination based on the lowest possible Mean Squared Error value.
Finally, a final execution is made for the testing dataset where each algorithm would conduct STNSR using the best set of parameters. From the 6 algorithms the Xgboost algorithm stood out as the one with the most accurate predictions, having achieved an average of 7,04 error points from the actual predicted value.

This prediction will now allow the Príncipe Foundation's sea turtle conservation team to conduct survivability rate estimation using several sets of parameters. This serves as a strong starting point for future sea turtle nest survivability understanding and it also allows for possible scaling of the issue, as more factors that contribute to the rate can be included for assessment.

7. Final remarks

7.1 PF cases assessment (limitations)

7.1.1 Prediction of STNSR

We have achieved prediction on Sea turtle nest survivability rate and identified key variables that greatly impact it. The phenomena of eggs with yolk and embryo needs to be looked into, as understanding to why these eggs are malformed is key to further understanding STNSR. On the same note, anomalies in the eggs that require a case by case approach presents itself as a sort of "grey area" at the moment, but one that can be unveiled with the use of different surveying techniques and taking the necessary considerations to assess each case individually. Nest depth comes as a great insight as to why egg hatching might be affected, as deeper and shallower nests seem to suffer from lower survivability rates. Finally, precipitation not only shows us how weather analysis is important for the understanding of survival rates, it also shows us how searching for different factors that might influence the prediction of nest survivability is key to improving on it.

7.1.2 Dataset

My thesis made use of a complex dataset that observes a very specific set of variables. The broadness and exploration capability of said data is to some degree limited to what the Principe organization could provide me at this time. Based on the study at hand, we can identify that there is a significant amount of output variance that is unexplained. This can be attributed to a small volume of data as well as low variability on data. From subchapter 4.1 we can make pertinent point on how important quality over quantity surveying is necessary. This originates from the fact that from roughly 6600 rows we were left with only 1106 useable ones. The future is optimistic though, as more surveying with a renewed focus will provide us with better results.

7.1.3 Mapping

An important consideration for this thesis was the use of heatmaps detailing the STNSR along the different beaches in Príncipe. This was however made neglectable given the lack of proper

coordinate values on each nest. In the future, an emphasis on proper nest tracking will bring about further possibilities on this front.

7.2 Conclusive paragraph

Ultimately this was a very satisfactory research that provided me with valuable insights on how to deal with several challenges. For one, I was able to communicate and partner with a group of specialists and create a common language where each individual contributes with a different skill set. I have also been able to implement several statistical methods that were aimed at helping better understand the data and provide useful insights on patterns that might not have been so easy to discover had they not been under a strict review. Finally, the application of Machine Learning algorithms into a biological conservation themed research has created a framework on what future thesis in the same area might look like.

8.REFERENCES

[1] "Information about sea turtles: Hawksbill," Sea turtle conservation programmes, p. 1, 2019. [Online]. Available: https://conserveturtles.org/information-about-sea-turtles- hawksbill-sea-turtle/

[2] "Information about sea turtles: Leatherback," Sea turtle conservation programmes, p. 1, 2019. [Online]. Available: https://conserveturtles.org/information-about-sea-turtles- leatherback-sea-turtle/

[3] N. Acosta-Mendoza, A. Morales-Reyes, H. J. Escalante, and A. Gago- Alonso, "Learning to assemble classifiers via genetic programming." International Journal of Pattern Recognition and Artificial Intelligence, vol. 28, Dec 2014.

[4] R. Alves, "Natureza e ambiente," Flora de SÃo Tomé e Príncipe, pp. 1–6, 1999. [Online]. Available: http://naturlink.pt/article.aspx?menuid=2cid=5837bl=1viewall=true

[5] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," Machine Learning, pp. 105–139, Jul 1999.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct 2010

[7] V. Bukhtoyarov and O. Semenkina, "Comprehensive evolutionary approach for neural network ensemble automatic design." Evolutionary Computation (CEC, Jul 2010), pp. 1–6.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system." KDD '16 The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 28, p. 785–794, Aug 2016.

[9] S. T. Conservancy, "Information about sea turtles: Green sea turtle," Sea turtle conservation programms, p. 1, 2019. [Online]. Available: https://conserveturtles.org/information-sea-turtles-green-sea-turtle/

[10] M. Dallimer, M. Melo, "Rapid decline of the endemic giant land snail archachatina bicarinata on the island of príncipe, Gulf of Guinea," June 2009.

[11] T. Dietterich, "Experimental comparison of three methods for con- structing ensembles of decision trees: Bagging, boosting, and randomization. machine learning," Machine Learning, pp. 148–152, Aug 1998.

[12] P. Domingos, "A unified bias-variance decomposition and its applications," in In Proc. 17th International Conf. on Machine Learning. Morgan Kaufmann, 2000, pp. 231–238.

[13] P. Domingos, "A unified bias-variance decomposition for zero-one and squared loss," in AAAI/IAAI. AAAI Press, 2000, pp. 564–569.

[14] Fauna and F. International, "S˜ao tom´e e pr´ıncipe," 2019. [Online]. Available: https://www.fauna-flora.org/countries/sao-tome-principe

[15] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in In proceedings of the thirteenth international conference on machine learning. Morgan Kaufmann, 1996, pp. 148–156.

[16] J. Friedman, “On bias, variance, 0/1—loss, and the curse-of- dimensionality,” Data Min. Knowledge Discovery, vol. 1, pp. 55–77, Mar 1997.

[17] S. R. James, Sea Turtles - A Complete Guide to Their Biology, Behaviour, and Conservation. Baltimore, Maryland: The Johns Hopkins University Press and Oakwood Art, 2004.

[18] U. Johansson, T. L¨ofstr¨om, R. K¨onig, and L. Niklasson, “Building neural network ensembles using genetic programming.” The 2006 IEEE2001, Jan 2006, pp.1260 – 1265. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[19] M. Keijzer and V. Babovic, “Genetic programming, ensemble methods and the bias/variance trade off – introductory investigations,” 04 2000, pp. 76–90.

[20] W. Koehrsen, “Random forest,” An Implementation and Explanation of the Random Forest in Python Learning, vol. 24, no. 2, pp. 123–140, Aug 2018. [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b760

[21] R. Kohavi and D. Wolpert, “Bias plus variance decomposition for zero- one loss functions,” Proceedings of the Thirteenth International Conference, 09 1997.

[22] P. Kordík and J. Cerny, “Building predictive models in two stages with meta-learning templates optimized by genetic programming.” IEEE Symposium on Computational Intelligence in Ensemble Learning, Dec 2014.

[23] B. Leo, Bagging Predictors. Kluwer Academic Publishers. Boston, Aug 1996, vol. 24, no. 2.

[24] K. M. T. K. Li, W, “On the squared residual autocorrelations in non- linear time series with conditional heteroscedasticity,” vol. 15. Journal of Time Series Analysis, Jun 2008, pp. 627–636.

[25] A. C. Lorena and A. C. P. L. F. Carvalho, “Uma introdução a support vector machines,” RITA, vol. 14, pp. 44–52, Nov 2007.

[26] N. Loureiro, ”Tartarugas em São Tomé e Príncipe”, 2009. [Online]. Available: https://tartarugasstomeprincipe.wordpress.com/sao-tome/

[27] E. M. Kleinberg, “A mathematically rigorous foundation for supervised learning,” Multiple Classifier Systems, First International Workshop, vol. 1857, pp. 67–76, Jun 2000.

[28] R. Polikar, “Polikar, “Ensemble based systems in decision making”,21-45,” Circuits and Systems Magazine, IEEE, vol. 6, pp. 21 – 45, Oct 2006.

[29] M. Pourahmadi, “Foundations of time series analysis and prediction theory,” vol. 1. Wiley series in probability and statistics. Applied probability and statistics section, June 2001.

[30] A. Pretorius, S. Bierman, and S. Steel, “A bias-variance analysis of ensemble learning for classification,” vol. 58. 58th Annual Conference of the Statistical Association of South Africa, At Cape Town, Dec 2016.

[31] L. Rokach, “Ensemble-based classifiers,” Artificial Intelligence Revolution, vol. 33, pp. 1–39, Feb 2010.

[32] Y. F. R. Schapire, “A short introduction to boosting.” Machine Learning, 1999, pp. 1–3.

[33] J. Schmidhuber,“S˜ao tom´e,” vol. 61, Jan 2015, pp. 85–117. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014002135

[34] F. Soares, “Modelling the distribution of São Tomé bird species: Ecological determinants and conservation prioritization,” Universiade de Lisboa, January 2019.

[35] J. R. Spotila, Experiments with a New Boosting Algorithm. The Johns Hopkins University Press: Baltimore and London, 2004.

[36] S. Swati G.Anantwar, Rajeshri R. Shelke, “Simplified approach of ann: Strengths and weakness,” International Journal of Engineering and Innovative Technology (IJEIT), vol. 1, April 2012.

[37] G. H. S. Tiwari, M. Baladz, “Estimating carrying capacity at the green turtle nesting beach of east island, french frigate shoals.” Marine Ecology Progress Series, Nov 2010.

[38] T.Parr and H.Becker, “How to explain gradient boosting.” Machine Learning, 1996.

[39] D. H. Wolpert, Stacked Generalization. Elsevier Ltd, 1992, vol. 5.

[40] R. W. F. E. Wood, J, “Artificial incubation period of green sea turtle eggs (Chelonia mydas),” vol. 10, March 1979, pp. 215–22.

[41] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Chap- man and Hall/CRC, Jan 2012, vol. 14.

9. ANNEXES

ANNEX I - Fundação Príncipe organization information

Figure 20 PF logo. Acquired at: https://www.facebook.com/FundacaoPrincipe/photos/a.305011313023399/1018648151659708/?type=1&theater

Foundation: 2015

Per their Facebook page:

"Fundação Príncipe (Principe Foundation) is a local NGO of 47 staff members, 91% locals. We drive to give capacity and training to our local team so that they can grow in their responsibilities and position on the organization. Our Team is our Family."

Contact:
Facebook: m.me/FundacaoPrincipe
Mail: info@fundacaoprincipe.org
Instagram: fundacaoprincipe

Table 30 PF contact information

ANNEX II – Species description

II.I - *Chelonia mydas* (CM)

CM species summary table with most relevant information, with image representation of the species appearance both as an adult and a hatchling. It follows bellow:

Figure 21 CM adult turtle. Acquired at: https://www.shutterstock.com/pt/video/clip-10326566-green-sea-turtle-chelonia-mydas-eating-seaweed

Figure 22 CM younglings. Acquired at: https://www.shutterstock.com/es/image-photo/baby-green-turtles-chelonia-mydas-crawling-1203768316

	Labels		Details
1	Scientific Name		Chelonia mydas
2	Size		83-114 cm
3	Weight		110 - 190 kg
4	Nesting		
4.1		interval	2 years
4.2		eggs	100-126 eggs
4.3		incubation	60 days
5	Range		temperate and tropical climates
6	Numbers		85000-90000 nesting females

Table 31 summary on CM

II.II - *Eretmochelys imbricata* (EI)

EI species summary table with most relevant information, with image representation of the species appearance both as an adult and a hatchling. It follows bellow:

Figure 23 EI adult turtle. Acquired at: https://pt.m.wikipedia.org/wiki/Ficheiro:Eretmochelys_imbricata_01.jpg

Figure 24 EI younglings. Acquired at: https://www.shutterstock.com/es/image-photo/hatching-found-these-hawksbill-turtle-hatchlings-473426905

	Labels		Details
1	Scientific Name		Eretmochelys imbricata
2	Size		71-89 cm
3	Weight		40-76 kg
4	Nesting		
4.1		interval	2-4 years
4.2		eggs	160 eggs
4.3		incubation	60 days
5	Range		tropical and subtropical climates
6	Numbers		20000-23000 nesting females

Table 32 summary on EI

II.III - *Dermochelys coriacea* (DC)

DC species summary table with most relevant information, with image representation of the species appearance both as an adult and a hatchling. It follows bellow:

Figure 25 DC adult turtle. Acquired at: https://www.pinterest.pt/pin/630996597766367109/?lp=true

Figure 26 DC younglings. Acquired at: https://www.alamy.com/stock-photo-newborn-hatchling-leatherback-sea-turtles-dermochelys-coriacea-searching-18031032.html

	Labels		Details
1	Scientific Name		Dermochelys coriacea
2	Size		130-183 cm
3	Weight		300 - 500kg
4	Nesting		
4.1		interval	1,2 to 3 years
4.2		eggs	30 unfertilized and 80 fertilized
4.3		incubation	65 days
5	Range		wide distribution of climates
6	Numbers		34000-36000 nesting females

Table 33 summary on DC

ANNEX III – Data sets description

Master File	Definition
A. SEGUIMENTO ACTIVIDADE FÊMEAS	
Data inserção dados	Date in which data was inserted
Tipo Monitorização Noct, Diur, Mari	Nightime, Daytime, Dawn
Data	Date of observation
Zona Ilha	Area of the island where the observation occurred (North or South)
Praia	Beach where sighting ocurred
Zona da Praia (see comment)	Area of the beach where the observation occurred (North or South)
ID Femea	Female's ID tracking number
Fêmea avistada?	Female sighted
Espécie	Species
Nova/Recap? (N/R)	New or recaptured
Pitag	Pitag
Anilha ESQ	Left anill
Anilha DTA	Right anill
Comprimento carapaça	length in centimiters
Largura carapaça	width in centimiters
Ninho/Tentativa (N/T/ML)	Nest/attempt
Código Ninho	nest code
Início(horas)	Begining of observation (hour)
Fim(horas)	End of observation (hour)
Lat N	Latitude coordinate
Long E	Longitude coordinate
Zona (Maré/Vegetação)	Area of Sea or Vegetation
Dist. linha maré (m)	Distance to shoreline in meters
Dist. vegetação (m)	Distance to vegetation in meters
Ninho translocado (N/S)	Transported nest (Yes or No)
Zona de translocação	Area where nest was transported to
Lat ninho translocado (N)	Latitude coordinate of transported nest
Long do ninho translocado E	Longitude coordinate of transported nest
responsavel pela translocaçao	Responsible for transportation
Data de translocação	Date of transportation
Tempo de translocaçao	Time of transportation
COMP do ninho (cm)	Nest length
LARGURA do ninho (cm)	Nest width
Numero total de ovos translocados	Total number of eggs transported
Numero ovos translocados viaveis	Total number of viable eggs transported
Numero ovos translocados não viaveis	Total number of unviable eggs transported
Ovos predados	Predation number on eggs

Table 34 master file A.Seguimento actividade fêmeas

B. SEGUIMENTO DE NINHOS	**Definition**
Cascas	Total number of shells observed
Crias Vivas	Total number of younglings alive
Crias Mortas	Total number of dead younglings
Crias deformadas no ovo	Total number of younglings with deformation in the egg
Com gema	Total number of eggs with yolk
Com embrião	Total number of eggs with embryo
SubTotal	Sum of the two above occurrences
Ovos predados	Number of eggs predated on
Total Ovos	Total number of observed eggs
Data de Eclosão	Date of egg hatching
Data de Exumação	Date of removal of remains
Período Incubação (dias)	Number of days for eggs incubation
Taxa de Eclosão (%)	Hatching ratio
Taxa de Emergência (%)	Emergence from nest ratio
Taxa de Predação (%)	Egg predation ratio
Profundidade (cm)	Depth in centimeters
Observações	Observation
STATUS (perdido/activo/eclodido)	Status (lost/active/hatched)

Table 35 masterfile B.Seguimento de ninhos

Eclosões file	Definition
P_+	Hatching success ratio
P-	Hatching insucess ratio
Número Ovos	Total number of eggs in nest
Espécie	Turte species
dd-MMM	date of nest oppening
Praia	Beach where nest was located
Código Ninho	nest unique code
Cascas	Total number of shells observed in nest
Ovos Não eclodidos	Total number of unhatched eggs
Crias mortas	Total number of dead younglings
Com gema	Total number of eggs with yolk
Com embriao	Total number of eggs with embryo
Crias vivas	Total number of younglings alive
Anomalias	Total number of eggs with anomalies
Data da Desova	date of nest creation
Data da Eclosão	date of nest hatching
Período de Incubação	Period of incubation
Assinalado? (S/N)	Signaled
Aberto antes de emergência? (S/N)	Opened before emergency (Y/N)
Crias libertadas no mar	Total number of younglings who made it to sea
Largura do ninho	Nest length in centimeters
Profundidade ninho	Nest width in centimeters
Predacao	Predation
Quantidade predada	Predation quantity
Predação Total/Parcial	Total or partial predation
Predador	Predator
Observações	Observations

Table 36 Eclosões file

III.III – Coordinates

A summary table for the *Coordinates* sheet of the Coordinates excel file with a brief description of each of its fields. It follows bellow:

Variables	Description
name	beach name
opm lat	'openstreetmap' latitude
opm lon	'openstreetmap' longitude
lat/lon opm coordinates	'openstreetmap' latitude and longitude concatenated
beach location	index from 1 to 8 for beach location

Table 37 coordinates file

III.IV weather_stp

A summary table for the *weather_stp* excel file with a brief description of each of its fields. It follows bellow:

Variables	Description
Mes	Month for N = 12, from January to December,
Humidity(%)	average level of humidity observed for each month
Min temp(Cº)	average minimum temperature observed for each month
Max temp (Cº)	average maximum temperature observed for each month
Precipitation(mm)	average given in millimetres (mm) for each month
Wind(mph)	average given in Miles per hour (mph) for each month

Table 38 weather summary file

III.IV - weather_stp

The data in the *weather_stp* excel file collected on timeanddate.com follows bellow:

#	Mes	Humidity(%)	Min temp(Cº)	Max temp (Cº)	Precipitation(mm)	Wind(mph)	Avg_weather
1	Jan	84	25	30	74,6	7	27,5
2	Feb	83	25	31	85,2	7	28
3	Mar	82	25	31	101,9	7	28
4	Apr	84	25	31	91,7	7	28
5	May	84	25	30	56,8	8	27,5
6	Jun	83	24	27	1,2	9	25,5
7	Jul	82	22	26	0	9	24
8	Aug	81	23	27	0,8	10	25
9	Sep	81	24	28	9,5	10	26
10	Oct	84	24	29	70,7	9	26,5
11	Nov	86	24	29	164,3	8	26,5
12	Dec	85	25	29	92	7	27
AVG		83,25	24,25	29	62,4	8,2	26,625

Table 39 weather summary 2 file

ANNEX IV – Model of data manipulation phase

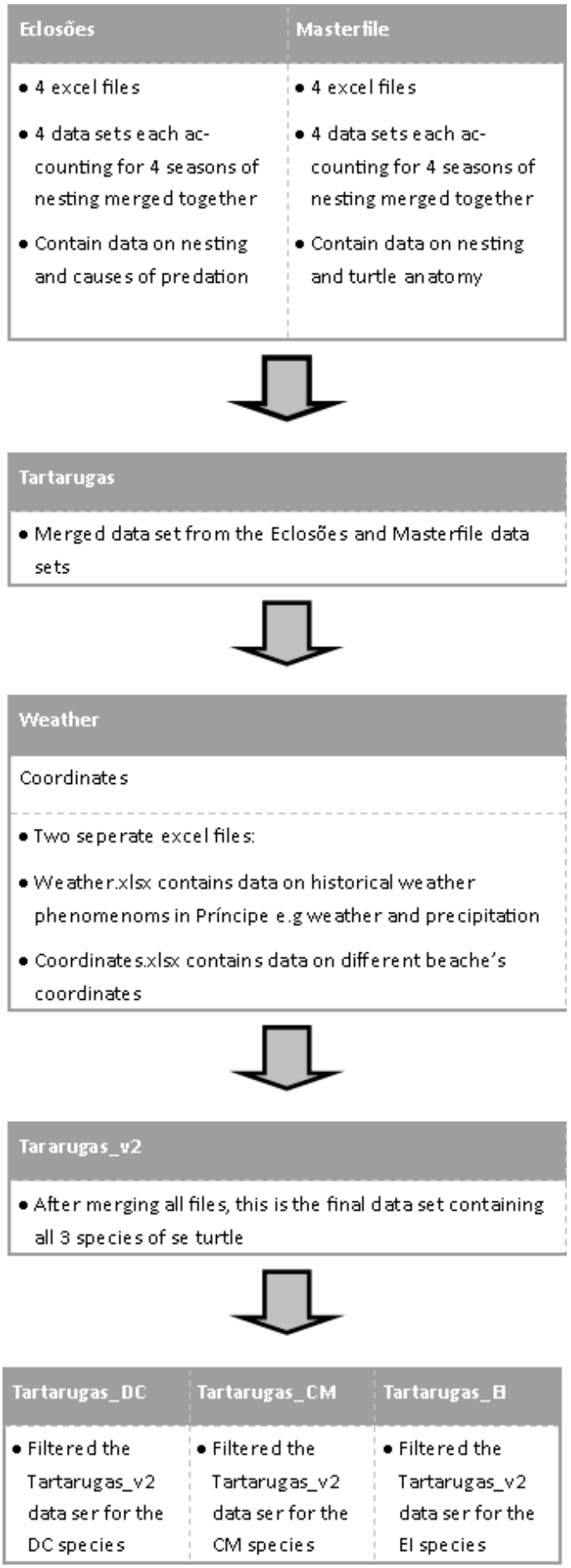
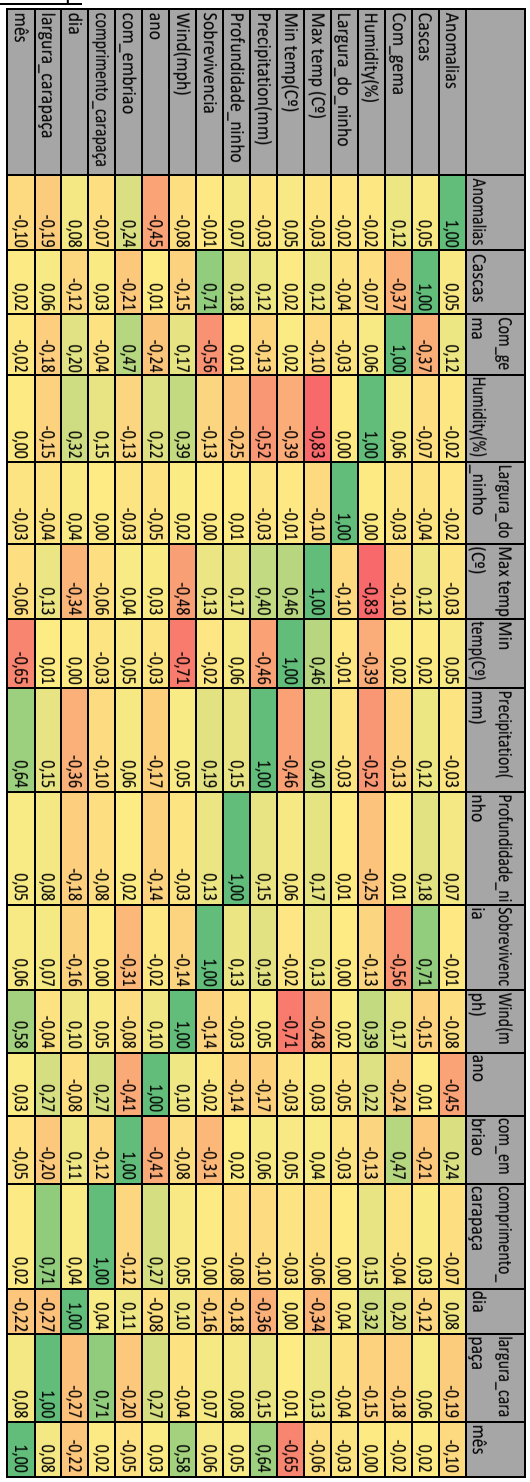

Eclosões
4 excel files
4 data sets each accounting for 4 seasons of nesting merged together
Contain data on nesting and causes of predation
Masterfile
4 excel files
4 data sets each accounting for 4 seasons of nesting merged together
Contain data on nesting and turtle anatomy
Tartarugas
Merged data set from the Eclosões and Masterfile data sets
Weather
Coordinates
Two seperate excel files:
Weather.xlsx contains data on historical weather phenomenoms in Príncipe e.g weather and precipitation
Coordinates.xlsx contains data on different beache's coordinates
Tararugas_v2
After merging all files, this is the final data set containing all 3 species of se turtle
Tartarugas_DC
Filtered the Tartarugas_v2 data ser for the DC species
Tartarugas_CM
Filtered the Tartarugas_v2 data ser for the CM species
Tartarugas_EI
Filtered the Tartarugas_v2 data ser for the EI species

Figure 27 data maniputation summary

ANNEX V – Correlation map



	Anomalias	Cascas	Com_gema	Humidity(%)	Largura_do_ninho	Max temp (Cº)	Min temp(Cº)	Precipitation(mm)	Profundidade_ninho	Sobrevivencia	Wind(mph)	ano	com_embriao	comprimento_carapaça	dia	largura_carapaça	mês
Anomalias	1,00	0,05	0,12	-0,02	-0,02	-0,03	0,05	-0,03	0,07	-0,01	-0,08	-0,45	0,24	-0,07	0,08	-0,19	-0,10
Cascas	0,05	1,00	-0,37	-0,07	-0,04	0,12	0,02	0,12	0,18	0,71	-0,15	0,01	-0,21	0,03	-0,12	0,06	0,02
Com_gema	0,12	-0,37	1,00	0,06	-0,03	-0,10	0,02	-0,13	0,01	-0,56	0,17	-0,24	0,47	-0,04	0,20	-0,18	-0,02
Humidity(%)	-0,02	-0,07	0,06	1,00	0,00	-0,83	-0,39	-0,52	-0,25	-0,13	0,39	0,22	-0,13	0,15	0,32	-0,15	0,00
Largura_do_ninho	-0,02	-0,04	-0,03	0,00	1,00	-0,10	-0,01	-0,03	0,01	0,00	0,02	-0,05	-0,03	0,00	0,04	-0,04	-0,03
Max temp (Cº)	-0,03	0,12	-0,10	-0,83	-0,10	1,00	0,46	0,40	0,17	0,13	-0,48	0,03	0,04	-0,06	-0,34	0,13	-0,06
Min temp(Cº)	0,05	0,02	0,02	-0,39	-0,01	0,46	1,00	-0,46	0,06	-0,02	-0,71	-0,03	0,05	-0,03	0,00	0,01	-0,65
Precipitation(mm)	-0,03	0,12	-0,13	-0,52	-0,03	0,40	-0,46	1,00	0,15	0,19	0,05	-0,17	0,06	-0,10	-0,36	0,15	0,64
Profundidade_ninho	0,07	0,18	0,01	-0,25	0,01	0,17	0,06	0,15	1,00	0,13	-0,03	-0,14	0,02	-0,08	-0,18	0,08	0,05
Sobrevivencia	-0,01	0,71	-0,56	-0,13	0,00	0,13	-0,02	0,19	0,13	1,00	-0,14	-0,02	-0,31	0,00	-0,16	0,07	0,06
Wind(mph)	-0,08	-0,15	0,17	0,39	0,02	-0,48	-0,71	0,05	-0,03	-0,14	1,00	0,10	-0,08	0,05	0,10	-0,04	0,58
ano	-0,45	0,01	-0,24	0,22	-0,05	0,03	-0,03	-0,17	-0,14	-0,02	0,10	1,00	-0,41	0,27	-0,08	0,27	0,03
com_embriao	0,24	-0,21	0,47	-0,13	-0,03	0,04	0,05	0,06	0,02	-0,31	-0,08	-0,41	1,00	-0,12	0,11	-0,20	-0,05
comprimento_carapaça	-0,07	0,03	-0,04	0,15	0,00	-0,06	-0,03	-0,10	-0,08	0,00	0,05	0,27	-0,12	1,00	0,04	0,71	0,02
dia	0,08	-0,12	0,20	0,32	0,04	-0,34	0,00	-0,36	-0,18	-0,16	0,10	-0,08	0,11	0,04	1,00	-0,27	-0,22
largura_carapaça	-0,19	0,06	-0,18	-0,15	-0,04	0,13	0,01	0,15	0,08	0,07	-0,04	0,27	-0,20	0,71	-0,27	1,00	0,08
mês	-0,10	0,02	-0,02	0,00	-0,03	-0,06	-0,65	0,64	0,05	0,06	0,58	0,03	-0,05	0,02	-0,22	0,08	1,00

Figure 28 correlation matrix with values

ANNEX VI - Feature extraction methods applied in the context of this thesis

VI.I Principal Component Analysis (PCA)

In PCA we look to reduce dimensionality [16] by creating new sets of independent variables from the original ones. With this we aim to achieve good explanatory power with feature elements for our models to consider. It does help reduce overfitting and improve on algorithm performance time. The downside to this, is that there is an understandable loss in variable interpretability, as the new variables that we create might not present an obvious relation.

For the purpose of understanding if PCA will help me improve on my prediction I have decided to plot a graph that shows variance explained as the number of features increases.

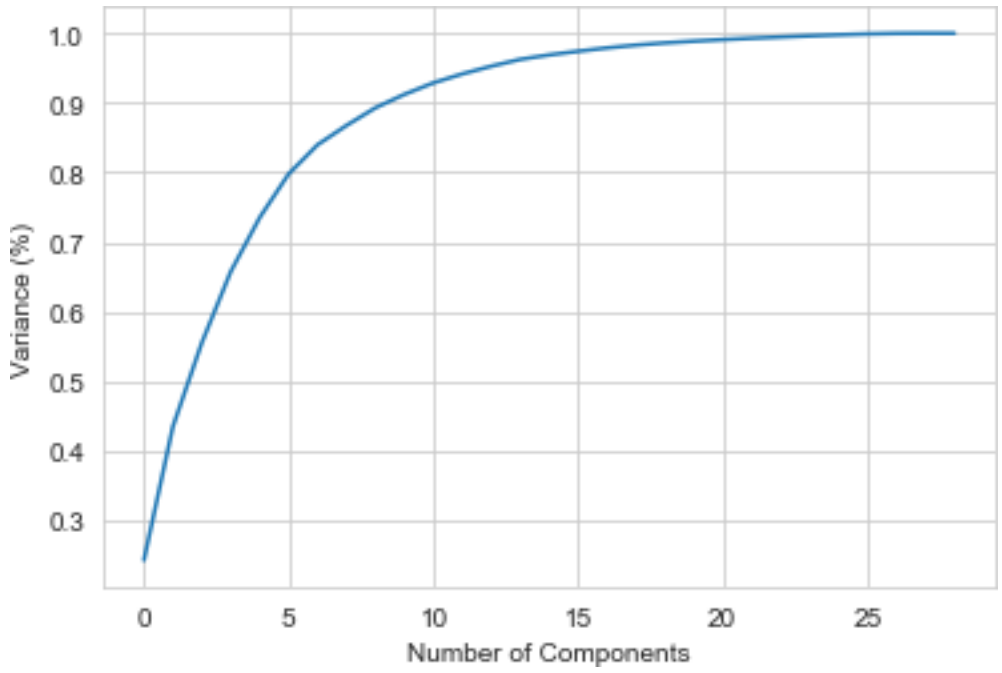


Figure 29 PCA elbow graph

From the above graph it is visible that the "elbow" from which we can take the number of PC's is located on 15 components. Even more from 5 to 15 components, variance lays between 80% and 90%. Given that we have a total of 18 variables in play at this point, the gain in feature extraction is scarce in face of such a loss in the interpretation of variables. Having said that, I opted to not proceed with this approach.

VI.II Support Vector Machine (SVD)

The SVM will look to reduce dimensionality and improve performance on models when we do contain a high number of variables and low number of observations. Although this is not such a case, SVM will show me how the relation between different vectors is presented. If I have high levels of explained variance at 1 or support vectors, it might be worth considering extracting these said vectors [29].

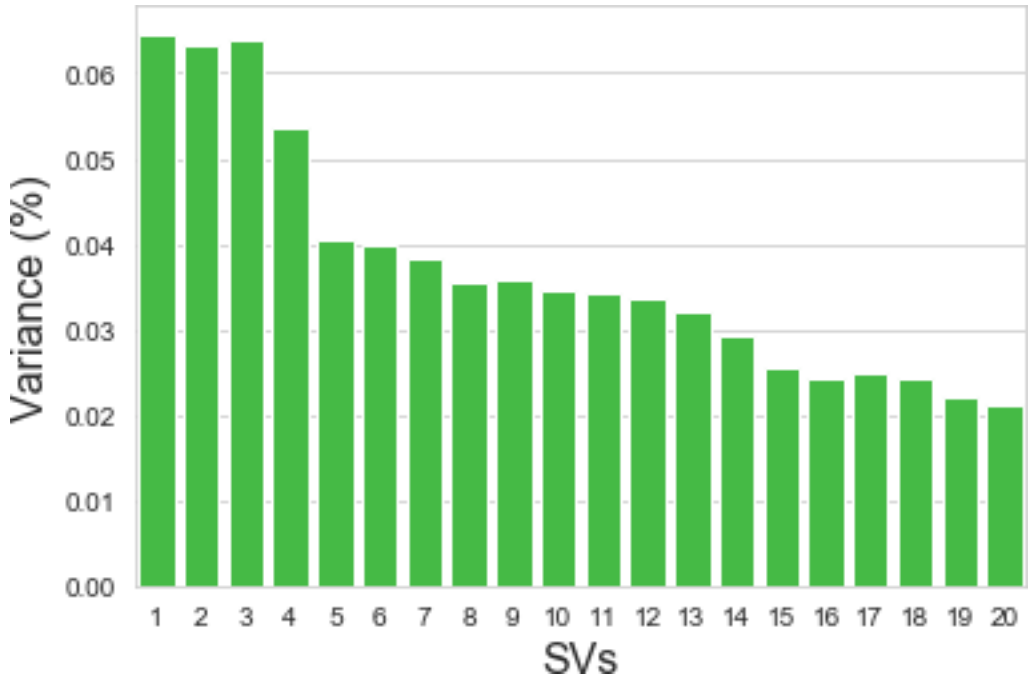


Figure 30 summary graph

In a similar reasoning to the Principal Components Analysis, SVM will not bring significant gains in prediction when faced with a big loss in interpretability of variables. If I were to simply feed a certain number of support vectors to the algorithms, I would very likely obtain improved results, but at the cost of losing the important reasoning factor for feature selection.

ANNEX VII – Ordinary Least Squares result summary

	coef	std err	t	P>|t|	[0.025	0.975]
const	7010.5828	1701.867	4.119	0.000	3668.398	1.04e+04
Anomalias	-0.3211	0.887	-0.362	0.717	-2.063	1.421
Avg_weather	-11.0297	7.867	-1.402	0.161	-26.479	4.419
Com_gema	-0.7317	0.059	-12.315	0.000	-0.848	-0.615
Humidity(%)	-1.7737	2.329	-0.762	0.447	-6.348	2.801
Largura_do_ninho	-0.1433	0.125	-1.145	0.253	-0.389	0.102
Maxtemp(Cº)	6.5144	5.490	1.187	0.236	-4.268	17.296
Mintemp(Cº)	-28.5739	20.189	-1.415	0.157	-68.221	11.074
Precipitation(mm)	-0.1312	0.207	-0.634	0.527	-0.538	0.276
Profundidade_ninho	0.0111	0.060	0.184	0.854	-0.107	0.130
Wind(mph)	-20.1456	11.937	-1.688	0.092	-43.589	3.297
ano	-2.8740	0.747	-3.848	0.000	-4.341	-1.407
com_embriao	-0.4322	0.089	-4.856	0.000	-0.607	-0.257
comprimento_carapaça	0.0758	0.141	0.538	0.591	-0.201	0.352
dia	0.0269	0.087	0.311	0.756	-0.143	0.197
largura_carapaça	0.0095	0.156	0.061	0.952	-0.298	0.317
mês	0.6531	1.210	0.540	0.590	-1.724	3.030
Aberto_antes_de_emergência?_(S/N)_S	-1.0089	2.139	-0.472	0.637	-5.209	3.192
Predador_0	3.1279	3.274	0.955	0.340	-3.302	9.558
Predador_CARANGUEJO	0.7889	4.457	0.177	0.860	-7.964	9.542
Predador_FORMIGA	4.9149	6.737	0.730	0.466	-8.316	18.146
praia_BOMBOM	5.8015	9.375	0.619	0.536	-12.609	24.212
praia_BUMBO	-2.4059	11.952	-0.201	0.841	-25.878	21.066
praia_GRANDE	1.9661	3.086	0.637	0.524	-4.094	8.026
praia_INFANTE	-9.8587	5.250	-1.878	0.061	-20.169	0.452
praia_MACACO	0.7845	9.505	0.083	0.934	-17.881	19.450
praia_MICOTO	-3.0031	15.900	-0.189	0.850	-34.228	28.222
praia_RIBEIRAIZE	4.8322	15.073	0.321	0.749	-24.769	34.433
storm?_storm	-0.9522	1.462	-0.651	0.515	-3.824	1.919
zona_ilha_NORTE	-2.8628	4.494	-0.637	0.524	-11.689	5.963
zona_maré/vegetação_M	-2.7779	1.857	-1.496	0.135	-6.426	0.870

Figure 31 OLS estimation summary