Mapping of whale shark sightings in continental waters between Isla Mujeres and Isla Contoy

Luis Eduardo Ramirez Padilla [1], Calvin Alberto López Álvarez [1],
Noé Vázquez Arzápalo [1] , Yarely Báez López [1]

[1] Universidad del Caribe - SM 78, Mza. 1 Lote 1, Esq. Fraccionamiento Tabachines, Cancún, Quintana Roo, 77528, México.
160300137@ucaribe.edu.mx 77508, 170300226@ucaribe.edu.mx 77510, 160300097@ucaribe.edu.mx,77533, ybaez@ucaribe.edu.mx 77528

Abstract. *The objective of this project is to create a web platform whose main characteristic is the mapping of whale shark sightings and marine fauna, through the CRISP-DM methodology, developing a web system in Flask that has the functions of data storage, data entry, login verification, representing an analysis of data in dashboards as well as the development of models for the prediction in order to observe in which places it is more common to find them and draw the most optimal routes for the company. In the same way, it seeks to carry out an analysis of the data provided by the company (such as geographical coordinates, sightings, types of fauna, time of sighting, among others...) in order to provide statistics and so that the company can make decisions to improve the quality of service.The results indicate that the proposal can be implemented in the current situation.*

Keywords: *Data analysis, Mapping, Data visualization, Predicting models.*

1. Introduction

The Universidad del Caribe [1] is a Mexican educational institution of higher education, founded on September 29, 2000, in the city of Cancun, Q. Roo. Currently, the Universidad del Caribe serves a total of 5,400 students, of which 4,787 are enrolled in 9 undergraduate degrees, 613 in 8 master's degrees. In the present pandemic, the current number of both people, administrative and vehicle personnel is between 100 and 150 people. Before the COVID-19 contingency, there were around 3,000 accesses between people and vehicles. In the current protocol of admission to the university, the records are annotated through a logbook in physical format, and the data that is recorded are the name, time of entry and exit, type of user, reason for visit and department to visit. This project focuses on the development of a virtual logbook and access control prototype, whose main feature is data analysis and the creation of data models with data mining and machine learning techniques, as well as providing a Dashboard with the results obtained, solving the problems of the null use of the data that is registered and the use of a physical format of a logbook.

2. State of the Art

In 2019, authors Gupte and Prins [2] presented "Fine-Scale Tracking of Ambient Temperature and Movement Reveals Shuttling Behavior of Elephants to Water". In it, they sought to provide a solution to the display of elephant movement in the thermal landscape of Kruger National Park, South Africa. Here, the authors made use of different data collected, such as: elephant neck temperature, environmental temperature, elephant speed, geospatial data, among others. They create a model that allows them to predict the elephant's movement behavior towards a specific area (mainly water).

In 2021, author Emilio Berti together with Davoli, Buitenwerf, Dyer, Hansen, Myriam, Svenning, Terlau, Brose and Vollrath [3] presented "The R package enerscape: A general energy landscape framework for terrestrial movement ecology". In it they explain that ecological processes and biodiversity patterns are strongly affected by the way animals move across the landscape. However, it remains a challenge to predict animal movement and space use. Here they present a new statistics-oriented programming language package, Project R.

Geoforge [4], presented "GeoTriple for Animal Tracking". It is a software sub-platform dedicated to animal movement tracking and allows compatibility with WMS (Web Map Services) and with them to manage geographic data on the ground, such as placemarks, routes, and areas (Geoforge Project, About & Animal Tracking page).

3. Methodology used

The methodology that will be used is CRISP-DM (Cross Industry Standard Process for Data Mining), which, as indicated by Villena Román [5]: Provides a standardized description of the life cycle of a standard data analysis project, in an analogous way to how it is done in software engineering with software development life cycle models.

3.1. Business Understanding

A couple of interviews were conducted with the beneficiary company to understand where the registered information comes from and what its objectives were with it. In the same way to know how they stored, handled and what use they had for them. With the information obtained, the following objectives were created:

Business Objectives

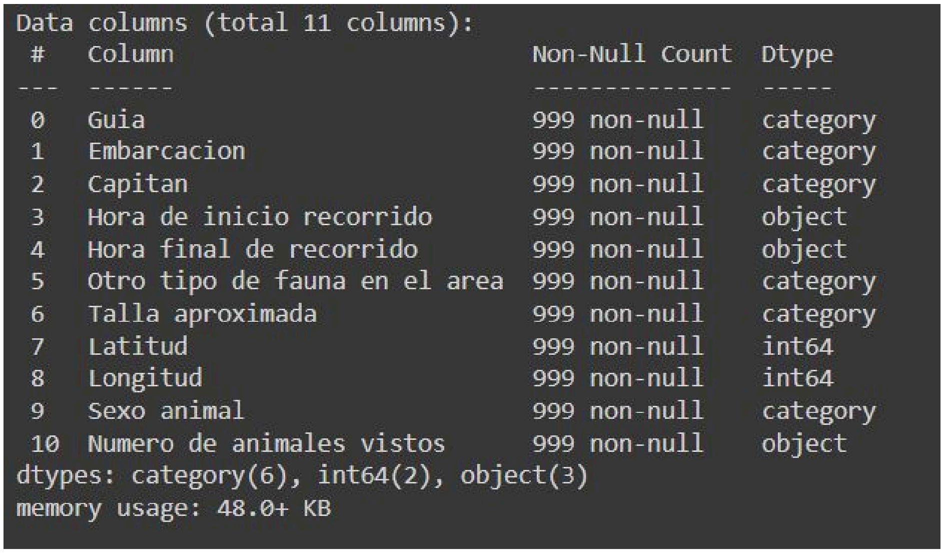
- Develop a virtual dashboard system.
- Store records digitally.
- Design access by username and password.
- Build models for data analysis of input and output records.
- View indicators from Dashboards.

Data Analysis Objectives

- Identify the total sightings that a shipment can have.
- Standardize the information obtained from the tours.
- Mapping of successful sightings of the whale shark and marine fauna in the area.
- Predict the coordinates that will have the biggest probability of sighting.

3.2. Data Understanding

The types of variables that were counted were identified and from that step, begin to manipulate, structure and distribute the type of modeling to use. Data was collected between June and September of the year 2021 and 2020, where the amount of 1608 records was obtained. We had to adjust some variables due some of them were duplicated. The team considered this amount of data a bit insufficient to be able to use in the model, this is because some amount was needed for model training. Also, much of this information obtained was not standardized and a robust cleaning had to be carried out.



```
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Guia                          999 non-null    category
 1   Embarcacion                   999 non-null    category
 2   Capitan                       999 non-null    category
 3   Hora de inicio recorrido      999 non-null    object
 4   Hora final de recorrido       999 non-null    object
 5   Otro tipo de fauna en el area  999 non-null    category
 6   Talla aproximada              999 non-null    category
 7   Latitud                       999 non-null    int64
 8   Longitud                      999 non-null    int64
 9   Sexo animal                   999 non-null    category
 10  Numero de animales vistos     999 non-null    object
dtypes: category(6), int64(2), object(3)
memory usage: 48.0+ KB
```

Table 1. Types of variables that belong to the data that were obtained.

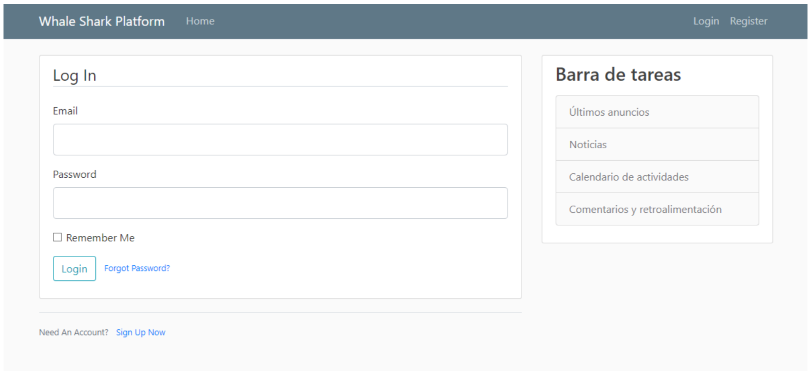
With the data already available, the analysis of the variables begins to identify the uses that can be given to them. The most important variables for the model were “Latitud, Longitud, Número de animales vistos” and for the general information “Guia, Embarcación, Hora Inicio, Hora Final, Otro tipo de Fauna”

3.3. Data Preparation

The raw data that the company provided us was contained in an endless number of .xlsx files with different types of formats, they collect the information in a somewhat peculiar way since they used spreadsheets as daily information logs, in theory it may be somewhat coherent to carry out this type of use for spreadsheets, however already in practice or in a scenario where an analysis of this information is required, which is where we as a team find ourselves, this way of storing the data makes it very difficult to extract and clean the data, since there are more than 700 files with specific information in each one, obtaining the data we required was quite a challenge, we opted to extract the remaining 30% of the information using a technique manually with the use of collaborative work tools from Google with which we obtained an information standard. This already allowed us to have visible control at the time of carrying out the first exploratory analysis of the data.

3.4. Modeling

To carry out the control of daily sightings during the tour season, this platform was created with the aim of unifying both all users who will have access to the form and an administrator who will be monitoring those who fill out the form and also You will have the option of publishing announcements and/or communications on the same platform, this will have the objective that the posts published within the platform have absolute relevance and notoriety beyond a simple comment in the WhatsApp group that they currently use within the company.



Log In

Email

Password

□ Remember Me

Need An Account?

Barra de tareas

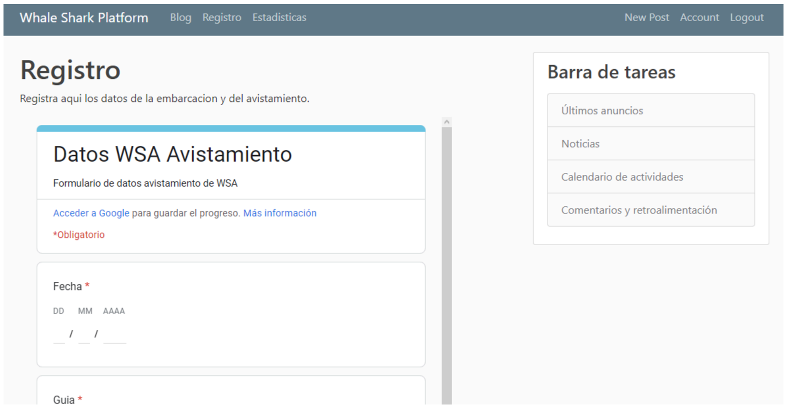
Últimos anuncios

Noticias

Calendario de actividades

Comentarios y retroalimentación

Figure 1. Virtual login system



Registro

Registra aqui los datos de la embarcacion y del avistamiento.

Datos WSA Avistamiento

Formulario de datos avistamiento de WSA

Acceder a Google para guardar el progreso. Más información

*Obligatorio

Fecha *

DD MM AAAA

__ / __ / ____

Guia *

Barra de tareas

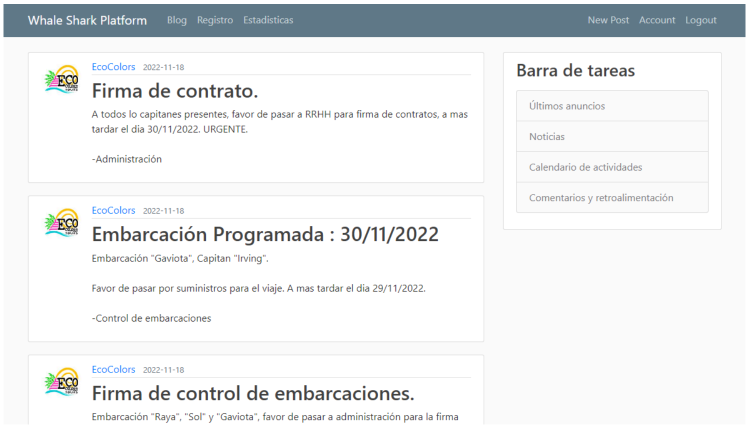
Últimos anuncios

Noticias

Calendario de actividades

Comentarios y retroalimentación

Figure 2. Form prototype view



EcoColors 2022-11-18

Firma de contrato.

A todos lo capitanes presentes, favor de pasar a RRHH para firma de contratos, a mas tardar el dia 30/11/2022. URGENTE.

-Administración

EcoColors 2022-11-18

Embarcación Programada : 30/11/2022

Embarcación "Gaviota", Capitan "Irving".

Favor de pasar por suministros para el viaje. A mas tardar el dia 29/11/2022.

-Control de embarcaciones

EcoColors 2022-11-18

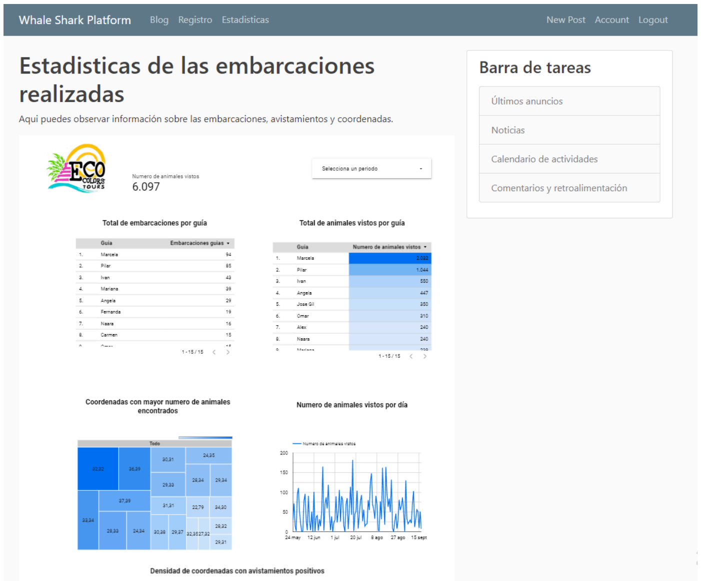
Firma de control de embarcaciones.

Embarcación "Raya", "Sol" y "Gaviota", favor de pasar a administración para la firma

Barra de tareas

- Últimos anuncios
- Noticias
- Calendario de actividades
- Comentarios y retroalimentación

Figure 3. Main blog post view



Estadisticas de las embarcaciones realizadas

Aqui puedes observar información sobre las embarcaciones, avistamientos y coordenadas.

Numero de animales vistos
6.097

Total de embarcaciones por guía

	Guia	Embarcaciones guias ▾
1.	Marcela	94
2.	Pilar	85
3.	Ivan	43
4.	Mariana	39
5.	Angela	29
6.	Fernanda	19
7.	Naara	16
8.	Carmen	15
9.	Omar	15

1 - 15 / 15

Total de animales vistos por guía

	Guia	Numero de animales vistos ▾
1.	Marcela	2.032
2.	Pilar	1.044
3.	Ivan	550
4.	Angela	447
5.	Jose Gil	350
6.	Omar	310
7.	Alex	240
8.	Naara	240
9.	Mariana	[illegible]

1 - 15 / 15

Coordenadas con mayor numero de animales encontrados

Numero de animales vistos por día

Densidad de coordenadas con avistamientos positivos

Barra de tareas

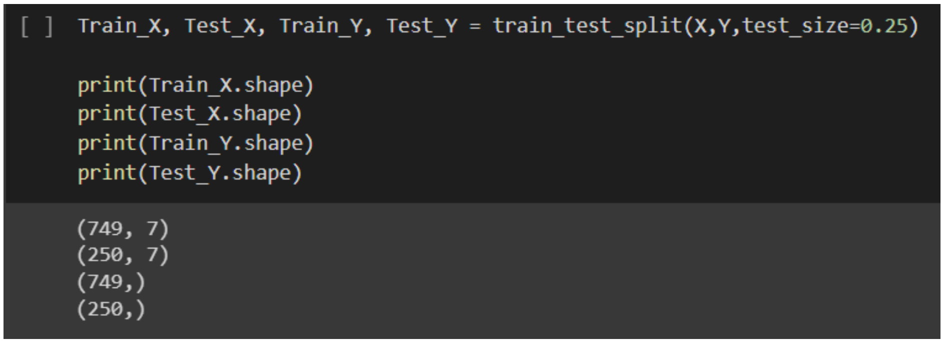
- Últimos anuncios
- Noticias
- Calendario de actividades
- Comentarios y retroalimentación

Figure 4. Statics and data view

Since the focus of this project is to make a model which takes advantage of times, locations (in discrete format) along with categorical data, we use tree-based methods and logistic regression.

Note that if you want to use other types of models, you may need to scale, normalize, or convert to one-hot encoding. In this case, we code the categorical variables and then standardize all the independent variables that we will use for the model.

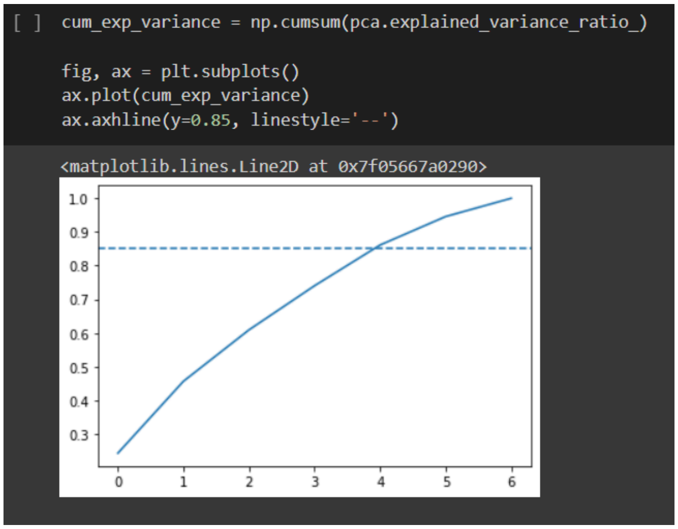
Following that, we'll use the Train-Test-Split methodology, which is a model validation procedure that reveals how your model behaves with new data. Where we will choose to use 75% of our data as training and 25% as testing.



```
[ ]  Train_X, Test_X, Train_Y, Test_Y = train_test_split(X,Y,test_size=0.25)

     print(Train_X.shape)
     print(Test_X.shape)
     print(Train_Y.shape)
     print(Test_Y.shape)

     (749, 7)
     (250, 7)
     (749,)
     (250,)
```

Figure 5. Exploratory data analysis.

Following this, we will proceed to apply the principal component analysis (PCA) methodology with which we will transform the correlated variables into a new set of uncorrelated variables. The objective of the analysis is to reduce the dimensionality in which the original set of variables is expressed. For which of the 7 components that make it up, we will choose 5 (index 0) to preserve 85% of what the original set of variables represents.



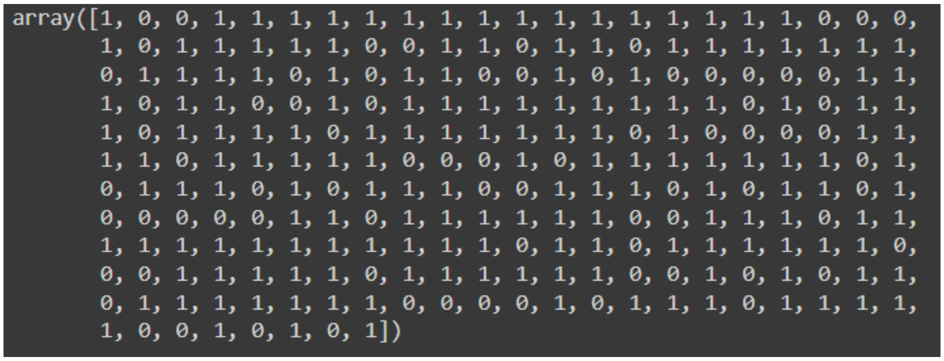
```
[ ]  cum_exp_variance = np.cumsum(pca.explained_variance_ratio_)

     fig, ax = plt.subplots()
     ax.plot(cum_exp_variance)
     ax.axhline(y=0.85, linestyle='--')

     <matplotlib.lines.Line2D at 0x7f05667a0290>
```

Figure 6. Exploratory data analysis.

Finally, we will proceed to apply three classifying models with the training data (80%) and evaluate their performance by comparing the results predicted by the model with the test values with the actual output values (20%), which are: Logistic regression multiple, Decision Trees and Random Forests.

4. Experimental results

The actual test values are as follows:

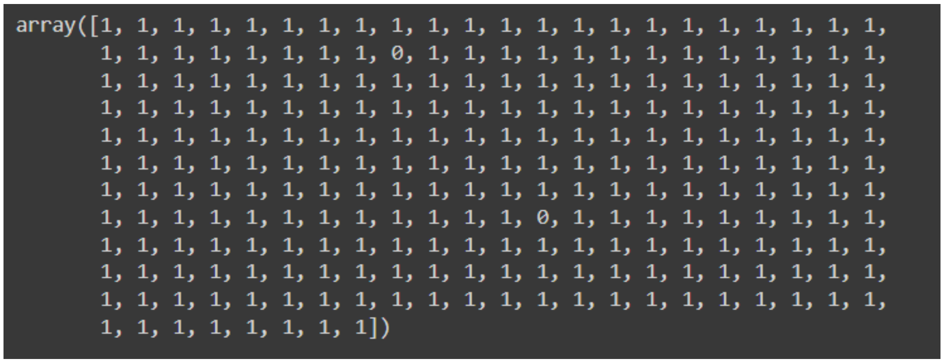


```
array([1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
       1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1,
       1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
       1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1,
       0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1,
       1, 0, 0, 1, 0, 1, 0, 1])
```

Figure 7. Actual values with which we will compare the predicted values

Multiple logistic regression

The values predicted by the model are the following:

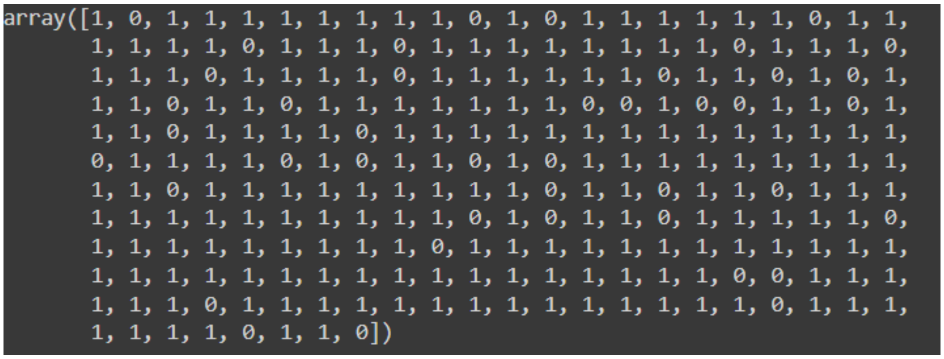


```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1])
```

Figure 8. Values predicted by the logistic regression model taking the test set

Decision trees

The values predicted by the model are the following:

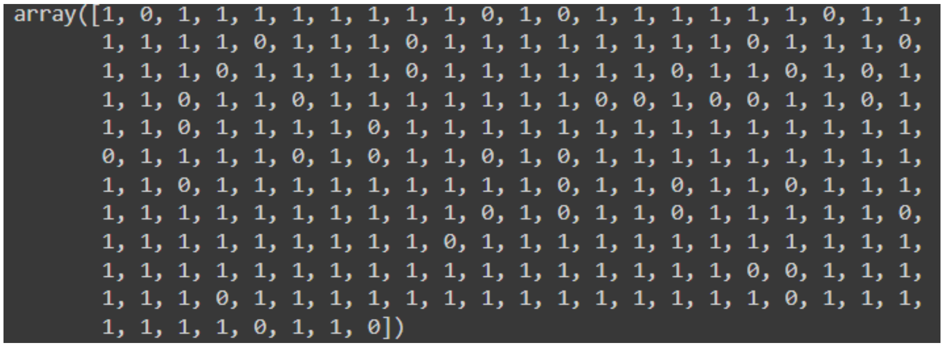


```
array([1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 0])
```

Figure 9. Values predicted by the decision trees model taking the test set

Random forests

The values predicted by the model are the following:



```
array([1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 0])
```

Figure 10. Values predicted by the random forests model taking the test set

4.1. Data collection and digitization

The collection and storage of data on the existing records was generally satisfactory, considering that they are samples from a specific period. We had some complications in this step, due to the standard and duplicate information they had, but we were able to complete the objective.

4.2. Model evaluation phase

As can be seen in figures 7, 8 and 9, the results (Precision) of each model were different:

Multiple logistic regression, which gave us an accuracy of 0.688 (68.8%).

Decision trees, which gave us an accuracy of 0.548 (54.8%).

Random forests, which gave us an accuracy of 0.624 (62.4%).

4.3. Dashboards and KPIs

The established KPIs are the Names of Guides, Place of sighting and Quantity of whale sharks sighted in the area, these KPIs were chosen because they help decision making, in cases such as identifying the guides with the greatest number of animals sighted during the season will allow the company to know which guides make better synergy with the crew of the boat because at the end of all this is an activity that requires excellent teamwork, on the other hand; the place of the sighting and the amount of animals seen in the area is also crucial to obtain the best results in the future on the tour, this will reduce the number of days where there are no animals seen and with this the tourists have a discontent and This could bring the company monetary losses due to bad criticism and refund requests because the tour did not meet the main expectation, seeing these animals in the water.

Another result that we can obtain with the conglomerate of information that we have from the company is to know how many times the same boat was used and for how many days in total throughout the season, this in order to be able to assign the best captains and boats. for the following season, taking the average number of animals seen in relation to the number of days of activity by the boats.

5. Conclusions and future research work

Due to everything mentioned in this document, we can develop predictive classification models for the problem of predicting whale shark sightings in inland waters and with this, the company can improve its service provision. However, due to the small amount of data on the part of the company, it is currently not possible to develop a model that can perform optimally and accurately that can support the company to improve the routing of its routes in a Data-Driven way. Therefore, the models developed will be a basis for future models that can be trained and developed taking advantage of the infrastructure developed in this project for the collection of company information and data standardization.

References

[1] Universidad del Caribe, «Entre el mar y la historia,» Universidad del Caribe, 2018. [En línea]. Available: https://www.unicaribe.mx/historia. [Último acceso: 30 Abril 2021]

[2] Thaker, M., Gupte, P. R., Prins, H. H., Slotow, R., & Vanak, A. T. (2019).Fine-scale tracking of ambient temperature and movement reveals shuttling behavior of elephants to water. Frontiers in Ecology and Evolution, 7, 4.

[3] Berti, E., Davoli, M., Buitenwerf, R., Dyer, A., Hansen, O. L. P., Hirt, M., Svenning, J.-C., Terlau, J. F., Brose, U., & Vollrath, F. (2022). The r package

enerscape : A general energy landscape framework for terrestrial movement ecology. Methods in Ecology and Evolution, 13(1), 60–67.https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13734

[4] Geoforge project. (s/f). Geoforge - home. Geoforge.org. Recuperado el 12 de agosto de 2022, de http://www.geoforge.org/index.php

[5] J. Villena Román, «CRISP-DM: La metodología para poner orden en los proyectos,» Sngular, 2 agosto 2016. [En línea]. Available: https://www.sngular.com/es/data-science-crisp-dm-metodologia/. [Último acceso: 23 3 2021].