UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

TFG del Grado en Ingeniería Informática

Jellyfish Forecast

Presentado por Pablo Santidrián Tudanca
en Universidad de Burgos — 25 de abril
de 2020
Tutor: José Francisco Díez Pastor y Álvar
Arnaiz González

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática

D. José Francisco Díez Pastor y D. Álvar Arnaiz González, profesores del departamento de Ingeniería Informática, área de Lenguajes y Sistemas Informáticos.

Expone:

Que el alumno D. Pablo Santidrian Tudanca, con DNI 71362353T, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado Jellyfish Forecast.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 25 de abril de 2020

Vº. Bº. del Tutor:	Vº. Bº. del co-tutor:
D. José Francisco Díez Pastor	D. Álvar Arnaiz González

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice de figuras

Índice de tablas

Introducción

El cambio climático está provocando que los fenómenos naturales extremos sean cada vez más habituales afectando a la poblaciones locales. Una de estas poblaciones locales son la medusas.

Las medusas tienen periodos de aparición estacionales y se alimentan de plancton, por lo que su densidad es mayor en zonas donde este abunda. Estas zonas suelen ser lugares cercanos al talud continental donde además se reproducen [4].

La aparición de medusas cerca de las costas, es un fenómeno que se da cada vez con mayor frecuencia. Estas floraciones tiene efectos perjudiciales en ámbitos como el turismo o la pesca, así como los daños que pueden provocar a la salud de las personas llegando en algunos casos a causar enfermedades graves [22, 23].

Alguno de los factores que están provocando el aumento de los acercamientos de las medusas a las playas son [4, 19]:

- La **climatología**, influye principalmente el cambio climático con el descenso del nivel de lluvias y el aumento de las temperaturas, que favorecen el aumento de la salinidad y de la temperatura del agua.
- La **contaminación** provocada por la modificación de las zonas costeras o los vertidos cercanos a los costas provocan la proliferación de bacterias o plancton que sirve de alimento para las medusas.
- La **sobrepesca** causa un descenso de depredadores así como de otras especies con las que las medusas compiten por el alimento.

Con este proyecto se pretende predecir el comportamiento de las poblaciones de medusas en las costas de Chile en función de datos meteorológicos

y marítimos obtenidos mediante el programa europeo *Copernicus*[1]. Estos datos se preprocesarán, es decir se eliminará la información que no es útil y delimitar la zona geográfica de estudio. A partir de ahí se entrenará una serie de modelos para predecir la llegada a las costas de las medusas.
Extender tema del modelo, que se eniende por modelo, machine learning...

[1]Programa de observación terrestre de la Unión Europea. https://marine.copernicus.eu/

Objetivos del proyecto

A continuación, se detallarán los objetivos que han motivado la realización de este proyecto.

2.1. Objetivos generales

- Recopilar y filtrar los datos necesarios para el modelo predictivo.
- Utilizar técnicas de aprendizaje automático para predecir la llegada de medusas a las costas.
- Desarrollo de una aplicación web permitiendo la consulta de las predicciones a los usuarios.

2.2. Objetivos técnicos

- Generar documentación en LaTeX, aprendiendo dicho lenguaje de marcado para la edición de textos con acabado profesional.
- Utilizar un sistema de control de versiones con la plataforma GitHub junto a la extensión ZenHub para facilitar la gestión del proyecto.
- Generar script para recopilar y filtrar los datos necesarios para la realización del proyecto.
- Generar una estructura de datos sobre la que se obtendrán los modelos, utilizando los datos de avistamientos de medusas y los datos oceánicos obtenidos de *Copernicus*.

- Comparar los resultados de los diferentes modelos obtenidos.
- Realizar una web en la que mostrar los resultados del modelo de una manera fácil e intuitiva.

2.3. Objetivos personales

- Investigar diferentes técnicas y herramientas utilizadas para la minería de datos.
- Adquirir conocimientos sobre el desarrollo web.
- Aprender a generar documentación en LaTeX.

Conceptos teóricos

3.1. Medusas

Las medusas son animales marinos formados por un cuerpo gelatinoso del que cuelga un manubrio tubular, encontrando la boca en la parte inferior de este. Algunas especies, tienen tentáculos con celular urticantes denominadas cnidocitos. Las medusas se desplazan mediante contracciones de su cuerpo absorbiendo agua para luego ser expulsada de manera brusca provocando el movimiento [6].

Su reproducción es asexual, siendo esta, una fecundación externa mediante óvulos y espermatozoides que son liberados por los machos y las hembras respectivamente. Esto da lugar a la fecundación de gametos que se convertirán en larvas denominadas plánulas. Más adelante estas larvas se adhieren a alguna superficie donde se transformarán en pólipos para finalmente desprenderse la medusa adulta [15].

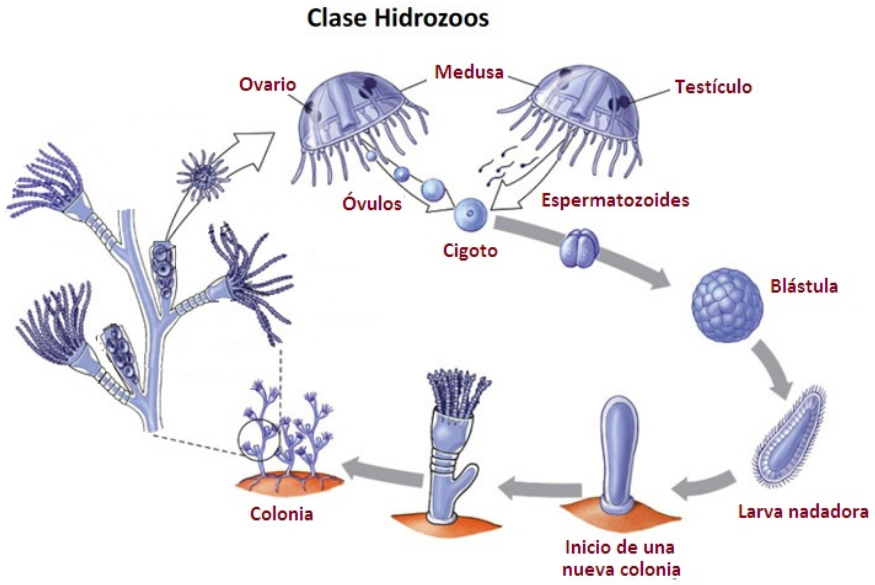

Figura 3.1: Fases del proceso de reproducción de las medusas.[7]

Su alimentación se basa se basa principalmente en plancton aunque también son capaces de comer crustáceos, huevos o peces pequeños.

Comportamiento de las medusas

Los colonias de medusas están muy influenciadas por las condiciones climatológicas y marítimas. La temperatura, salinidad, el viento y las corrientes son los principales factores a tener en cuenta.

El peso de estos factores en los desplazamientos de las colonias varía en función de la etapa de desarrollo en la que se encuentren. Las que tienen un tamaño pequeño o mediano, están más condicionadas a la dirección de los vientos y las corrientes ya que, debido a su pequeño tamaño, no son capaces de contrarrestar estas fuerzas. Por otro lado, en las medusas de un tamaño superior, estos factores tienen menos relevancia mientras que la salinidad del agua y la temperatura de la misma adquieren un mayor protagonismo. Hasta unos 25 grados el numero de medusas va en aumento. A partir de ese punto, la concentración de las mismas decrece [20].

La reproducción de estas también se ve influenciada por la temperatura del agua pues según diferentes experimentos se ha demostrado que existe una relación entre el aumento de las temperaturas y una mayor reproducción asexual en varias especies gelatinosas.¿Citas literales?

Los factores humanos también tienen su influencia en los movimientos de las poblaciones de medusas y en su reproducción. El aumento de materia orgánica provocado por vertidos como podrían ser los de una EDAR (Estación Depuradora de Aguas Residuales) o de una explotación agrícola, que provoca la eutrofización del medio haciendo que las medusas puedan desarrollarse de una manera más rápida. Del mismo modo, la construcción de estructuras costeras, proporcionan lugares donde pueden proliferar con mayor facilidad.

Teniendo en cuenta todo esto,diferentes estudios concluyen que estas alteraciones aleatorias del medio, tiene poca influencia en el desarrollo de las colonias de medusas, mientras que las condiciones ambientales que se repiten anualmente tiene una mayor importancia en comparación. Esto remarca la importancia de un análisis de las condiciones ambientales que provocan estos brotes para poder anticiparse a ellos.[19]

3.2. Knowledge Discovery in Databases (KDD)

Explicación

Preprocesamiento de los datos

Los datos obtenidos inicialmente no pueden ser utilizados directamente para la minería de datos. Estos datos deben ser preprocesados para eliminar posibles variables que no son necesarias o datos incorrectos y así conseguir transformar nuestro conjunto de datos iniciales, en un conjunto más útil y menos pesado por lo que el algoritmo de análisis posterior, tendrá una menor carga de trabajo. Reescribir con más puntos

Minería de datos

Actualmente se recopila una gran cantidad de información de todos los ámbitos y es necesario darla un uso práctico. La **minería de datos** es un campo de las ciencias de la computación por el cual se tratan de descubrir nuevos patrones o relaciones en conjuntos de datos y así, conseguir un conocimiento obtenido de manera automática (*Machine Learning*). Con estas nuevas relaciones se trata de explicar comportamientos actuales o predecir resultados futuros [21].

Sin embargo, la minería de datos es solo una fase del proceso de descubrimiento del conocimiento (KDD) pues es necesario tratar los datos antes

de analizarlos así como validarlos posteriormente. Este proceso podemos obtener las siguientes fases:

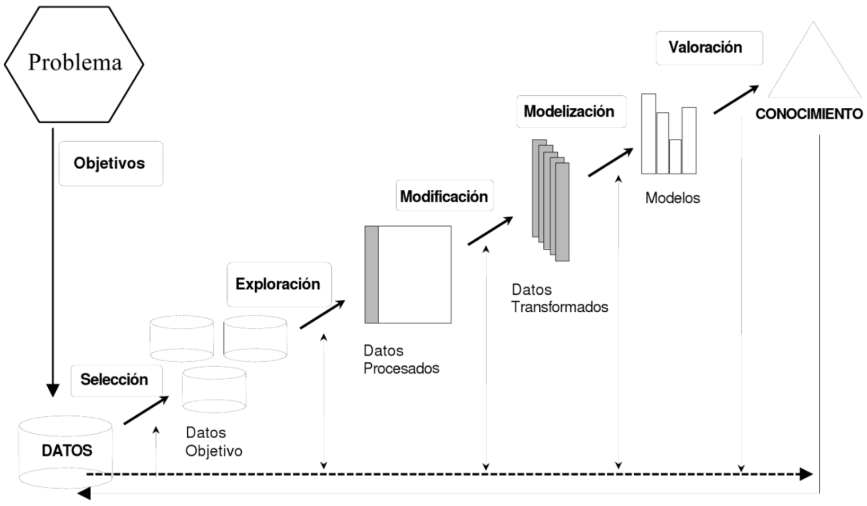

Figura 3.2: Fases del proceso de minería de datos [21].

Machine learning

Técnicas y herramientas

Este apartado se plasman las diferentes técnicas metodológicas y herramientas de desarrollo que se han utilizado en la realización del proyecto, así como las posibles alternativas que se han tenido en cuenta y el motivo de haberlas desechado.

4.1. Gestión del proyecto

Scrum

Scrum se trata de un marco de trabajo ágil destinado al manejo de proyectos de desarrollo *software*. Está destinado para equipos pequeños dividiendo el trabajo en objetivos que se van logrando de manera incremental a través de iteraciones denominadas *sprints* [24].

Git

Git se trata de un sistema de control de versiones gratuito y de código abierto para el manejo de proyecto. Se trata de un software de control de versiones de manera que se puede llevar un registro de cambios en los archivos y posibilita la coordinación de trabajos entre diferentes personas [3, 17].

GitHub

GitHub es un plataforma destinada a alojar proyectos, qu se basa en el software de control de versiones Git. Esta plataforma utiliza un interfaz web desde la que se nos permite realizar control de código, documentación,

gestión de tareas y otros muchas funcionalidades además de integración con otros servicios. GitHub es gratuito para proyectos *open source* [18, 11].

Alternativas

Otras alternativas a GitHub fueron, Gitlab y Bitbucket. Ambos servicios son bastante similares a GitHub en funcionalidades y basados en Git.

- Bitbucket fue rechazado rápidamente por la falta de familiaridad en su uso ya que no lo había usado nunca, únicamente había visto repositorios en la web.
- Gitlab es un entorno más conocido ya que es el software que he utilizado en las prácticas de empresa para el control de código dentro del equipo de trabajo del que formo parte.

Finalmente se decidió usar GitHub por haber sido utilizado en clase de gestión de proyectos por lo que se tenia un mayor conocimiento de su funcionamiento, así como por su integración con ZenHub del que se hablará a continuación.

ZenHub

ZenHub es una herramienta para gestión de proyectos que se integra con GitHub. Este complemento añade a GitHub un tablero canvas en el cual se representan las *issues*. Es posible estimar tareas, así como darlas prioridades y visualizar gráficos como el gráfico *burndown* [13].

Alternativas

Se plantearon otras alternativas como Trello o GitHub projects, pero finalmente se eligió ZenHub por su facilidad de uso y el haber sido usado anteriormente, requisito que las otras alternativas no cumplían.

4.2. Herramientas

Python

Para le desarrollo en Python se eligió Visual Studio Code. Esta es una aplicación totalmente gratuita basada en el framework Electron y posee gran cantidad de extensiones para facilitar la tarea a la hora de la programación

como auto completado con IntelliSense. Además tiene integración con Git para el control de versiones [10, 9].

Alternativas

Se estudiaron otras alternativas como PyCharm o Jupyter Notebook pero finalmente se eligió Visual Studio Code por la familiaridad con la misma y gran compatibilidad con diferentes lenguajes.

Jupyter Notebook

También se ha utilizado Jupyter Noteebook para el desarrollo con Python. Esta herramienta se ha utilizado para realizar todas las pruebas debido a su facilidad para documentar el código y la segmentación del mismo en apartados separados.

tmux

Se trata de un herramienta que permite lanzar múltiples terminales en una misma ventana, cada uno de manera independiente y corriendo en segundo plano. Esta herramienta se ha utilizado para la descarga de los datos obtenidos de *Copernicus*. En un principio se utilizó unicamente una conexión ssh con un equipo de cómputo de Universidad. Esto daba varios errores, por una parte, la conexión se cerraba perdiendo el proceso de descarga. También era necesario que un equipo personal estuviese encendido constantemente durante la descarga para que no se detuviera la conexión. Para solucionar estos problemas se abrió una sesión de tmux corriendo en segundo plano, pudiendo así cerrar a conexión ssh y abrirla para consultar el proceso.
Añadir si se usa para la ejecucion del modelo.

Pencil

Software gratuito para la realización de prototipos de interfaces gráficas. Permite la instalación de paquetes para crear maquetas de múltiples plataformas.

Creately

Aplicación web utilizada en la fase de modelado para la creación de los diagramas de uso. Dispone de una versión gratuita con opciones reducidas.

4.3. Documentación

LaTeX

LaTeX es un sistema de composición de texto plano destinado a la composición de textos con una calidad tipográfica alta. Incluye características diseñadas para la elaboración de documentación técnica y científica siendo un estándar en la publicación de documentos de investigación. LaTeX es totalmente gratuito [5].

Alternativas

Otras opciones valoradas fueron Microsoft Word y OpenOffice. La Universidad proporciona una plantilla para LaTeX y OpenOffice por lo que Microsoft Word fue la primera descartada. Finalmente debido al enfoque más técnico que proporciona LaTeX frente a OpenOffice, se decidió utilizarlo.

TeXstudio

TeXstudio se trata de un editor de textos gratuito, que ofrece diversas herramientas para elaborar documentación con LaTeX haciendo la escritura más confortable e intuitiva [8].

Alternativas

Las alternativas a este editor que se estudiaron fueron TeXmaker y Overleaf. Se terminó eligiendo TeXstudio por ser un editor instalado en local, permitiendo el trabajo aunque no se tuviera conexión a Internet así como por ser más intuitivo y fácil de utilizar que TeXmaker.

Zotero

La herramienta Zotero es un software gratuito para la gestión de referencias pudiendo recoger, organizar y citar creando referencias bibliográficas para cualquier editor. También cuenta con integración en le navegador. Una vez recopilados todas las citas, se puede exportar a un fichero BibTeX para la utilizarse con LaTeX [14].

4.4. Bibliotecas

Flask

Flask es un framework ligero para el desarrollo de aplicaciones web en Python bajo el modelo Modelo-Vista-Controlador (MVC). Está diseñado para desarrollar aplicaciones de manera rápida y sencilla y con capacidad de escalar a aplicaciones más complejas [2].

Xarray

Xarray se trata de un proyecto de código abierto desarrollado para Python que facilita el trabajo con matrices multidimensionales etiquetadas utilizando la librería NumPy. Consta de una gran variedad de funciones para el análisis y la visualización de estructuras de datos. Está inspirado en el funcionamiento de la librería pandas y diseñado para funcionar con archivos de tipo netCDF [12].

Pandas

Esta biblioteca se trata de una extension de NumPy y está destinada a la manipulación y análisis de datos en lenguaje Python. Permite trabajar con estructuras de datos y operaciones para su trasformación pudiendo estas ser tablas temporales o series numéricas [16].

FtpLib

Biblioteca destinada a la implementación de la parte del cliente en el protocolo FTP. Desarrollada para el lenguaje Python, nos permite automatizar accesos a servidores FTP [1].

tqdm

Pequeña librería utilizada para mostrar una barra de progreso a la hora de realizar la descarga de los datos del FTP.

Folium

Se trata de una biblioteca que nos permite la visualización de mapas, pudiendo superponer elementos en los mismos. Folium utiliza a su vez una biblioteca de javaScript llamada *leaflet*.

4.5. Bootstrap

Bootstrap es un conjunto de herramientas para facilitar el desarrollo en HTML, CSS y JavaScript

Aspectos relevantes del desarrollo del proyecto

5.1. Recogida de datos

La obtención de los datos necesarios para el posterior análisis, se obtuvieron a través de la organización europea *Copernicus*. Esta cuenta con una serie de datos recopilados por satélites de todo el mundo.

Para la descarga de estos datos, se encontraban disponibles dos alternativas. Por un lado, podían ser descargados a través de un servidor FTP y por el otro, había disponible una API de reciente lanzamiento.

En un principio se priorizó la opción de la descarga a través de la API ya que podían descargarse los datos ya tratados reduciendo el trabajo previo al análisis. Finalmente se descartó esta idea por la lentitud de respuesta del servicio, así como diferentes errores que se producían, haciendo que la descarga de la totalidad de los datos no estuviese asegurada.Por este motivo se descargaron los datos a través del servidor FTP y posteriormente, eran tratados eliminando las variables y zonas geográficas que no eran necesarias.

5.2. Ejecución remota

Para la ejecución de los scripts utilizados para la descarga de los datos, la ejecución del modelo y la realización de las diferentes pruebas, se ha utilizado un equipo de computo de la Universidad mediante una conexión ssh y una VPN. Todo esto ha hecho que surjan diferentes problemas.

En primer lugar no se disponía de permisos de administrador para instalar las diferentes bibliotecas necesarias. Esto se solucionó mediante la instalación de la herramienta Anaconda que nos permite ser instalada para un único usuario. Con esto, se creó un entorno virtual en el que poder instalar todas las bibliotecas necesarias.

tmux

Como se ha explicado anteriormente al hablar de tmux (4.2), se ha tenido problemas a la hora de ejecutar los diferentes scripts en el equipo de cómputo.

Para la ejecución de las diferentes pruebas, se utilizado la conexión ssh lanzado un proceso de Jupyter Notebook sin interfaz gráfica mediante el comando *"jupyter notebook –no-browser"*. Este servidor se inicializa por defecto en el puerto local 8888 por lo que se utiliza el comando *"ssh -p 22 -N -f -L localhost:9006:localhost:8888 pst0004@10.168.168.11"* para conectar este puerto del equipo de computo con un puerto (en este caso el 9006) de un equipo personal.

Existian casos en los que la conexión ssh se cerraba o se caía la conexión VPN, por lo que se perdía el proceso de ejecución o los últimos cambios realizados en los notebooks no se guardaban. Para eso se utilizó el la herramienta tmux permitiendo, que aunque la conexión se perdiera, los procesos en segundo plano no se perdían y así poder continuar con el trabajo.

5.3. Preparación de los datos

Tras haber descargado los datos oceánicos, disponemos de dos fuentes de datos. Por una lado, estos datos del estado de los océanos en las diferentes fechas y coordenadas, y por otro, un registro de avistamientos de medusas.

Los datos oceánicos están agrupados por cuadrantes separados cada 0,0833 grados. Por esto se redondearon las coordenadas de los avistamientos para coincidir con estos pasos.

A continuación se enlazan en un solo DataFrame los avistamientos con la variables oceánicas recogidas de ese cuadrante y en la fecha del avistamiento quedando la siguiente estructura de datos:
Foto estructura 1

Esta estructura inicial, contiene poca información pues no se puede predecir con exactitud la aparición de medusas observando unicamente las zonas más próximas a las playas. Por ello se añadieron más lecturas de los cuadrantes adyacentes mar adentro.

CrossCorelation

Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

[1] 20.8. ftplib — FTP protocol client — Python 2.7.17 documentation.

[2] Flask. Library Catalog: palletsprojects.com.

[3] Git.

[4] Las proliferaciones de medusas. Library Catalog: www.miteco.gob.es.

[5] LaTeX - A document preparation system.

[6] Medusozoa. Page Version ID: 124010357.

[7] Reproducción en hidrozoa. Library Catalog: Blogger.

[8] TeXstudio.

[9] Visual Studio Code - Code Editing. Redefined. Library Catalog: code.visualstudio.com.

[10] Visual Studio Code - Wikipedia, la enciclopedia libre.

[11] The world's leading software development platform · GitHub.

[12] xarray: N-D labeled arrays and datasets in Python — xarray 0.14.1 documentation.

[13] ZenHub - Agile Project Management for GitHub. Library Catalog: www.zenhub.com.

[14] Zotero | Your personal research assistant.

[15] Reproducción de las medusas, sus etapas, pólipos y larvas plánulas., October 2016. Library Catalog: medusas.wiki Section: Información general.

[16] Pandas, November 2019. Page Version ID: 121498623.

[17] Git, February 2020. Page Version ID: 123779424.

[18] GitHub, March 2020. Page Version ID: 124035823.

[19] Lisandro Benedetti-Cecchi, Antonio Canepa, Veronica Fuentes, Laura Tamburello, Jennifer E. Purcell, Stefano Piraino, Jason Roberts, Ferdinando Boero, and Patrick Halpin. Deterministic Factors Overwhelm Stochastic Environmental Fluctuations as Drivers of Jellyfish Outbreaks. *PLOS ONE*, 10(10):e0141060, October 2015. Publisher: Public Library of Science.

[20] Antonio Canepa, Verónica Fuentes, Mar Bosch-Belmar, Melissa Acevedo, Kilian Toledo-Guedes, Antonio Ortiz, Elia Durá, César Bordehore, and Josep-Maria Gili. Environmental factors influencing the spatio-temporal distribution of carybdea marsupialis (lineo, 1978, cubozoa) in south-western mediterranean coasts. 12(7):e0181611. Publisher: Public Library of Science.

[21] Daniel Santín González César Pérez López. Minería de datos: técnicas y herramientas - César Pérez López - Google Libros.

[22] Lisa-ann Gershwin, Anthony J. Richardson, Kenneth D. Winkel, Peter J. Fenner, John Lippmann, Russell Hore, Griselda Avila-Soria, David Brewer, Rudy J. Kloser, Andy Steven, and Scott Condie. Biology and ecology of Irukandji jellyfish (Cnidaria: Cubozoa). *Adv. Mar. Biol.*, 66:1–85, 2013.

[23] James Tibballs, Ran Li, Heath A. Tibballs, Lisa-Ann Gershwin, and Ken D. Winkel. Australian carybdeid jellyfish causing "Irukandji syndrome". *Toxicon*, 59(6):617–625, May 2012.

[24] Wikipedia contributors. Scrum (software development) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Scrum_(software_development)&oldid=943108748, 2020. [Online; accessed 10-March-2020].