
# Aplicación de técnicas de Aprendizaje Automático sobre datos de ahogamiento en España

## (Application of Machine Learning techniques on drowning data in Spain)

**Trabajo de Fin de Máster**
**para acceder al**

**MÁSTER EN CIENCIA DE DATOS**

**Autor: Nova Pagés, Adrià**

**Director\es: Garcia Saiz, Diego**

**Sal Sarria, Brian**

**Julio - 2023**

## Resumen

El accidente acuático es una de las causas de muerte por traumatismo no intencional más importantes en todo el mundo, representando aproximadamente el 7 % de todas las muertes relacionadas con traumatismos.

En la actual era digital en la que estamos inmersos, es fácil acceder a información sobre una amplia gama de temas de interés. Sin embargo, este no es el caso cuando se trata de obtener datos precisos sobre los accidentes acuáticos. Aunque muchos países hacen esfuerzos para recopilar información sobre el número de incidentes en su territorio, los resultados obtenidos son insuficientes debido a la falta de recursos y representación de estas comunidades. A pesar de los esfuerzos de la Organización Mundial de la Salud (OMS) para abordar esta problemática, aún queda mucho por hacer, ya que las organizaciones gubernamentales de cada país no siempre están comprometidas en recopilar datos de manera adecuada y asegurar su confiabilidad.

El campo de los accidentes acuáticos, tanto en España como en general, ha sido poco explorado, limitándose en su mayoría a análisis exploratorios básicos para comprender la situación de los incidentes acuáticos en un período y país específicos. Por ello, en este estudio, nuestro objetivo es adentrarnos en mayor detalle con el fin de descubrir patrones en los datos mediante la aplicación de técnicas de aprendizaje automático. Esto nos permitirá comprender mejor los motivos de los accidentes acuáticos en España y, a su vez, mostrar a los expertos en salvamento los resultados para que puedan tomar las decisiones oportunas que contribuyan a reducir la cantidad de víctimas mortales.



**Palabras clave:** accidente acuático, ahogamiento, grupo de edad, test de hipótesis, aprendizaje automático, análisis de datos.

## Abstract

The aquatic accident is one of the most important causes of unintentional trauma-related deaths worldwide, representing approximately 7 % of all trauma-related deaths.

In the current digital era we are immersed in, it is easy to access information on a wide range of topics of interest. However, this is not the case when it comes to obtaining accurate data on drownings. Although many countries make efforts to collect information on the number of drownings in their territory, the results obtained are insufficient due to the lack of resources and representation of these communities. Despite the efforts of the World Health Organization (WHO) to address this issue, much remains to be done, as government organizations in each country are not always committed to collecting data properly and ensuring its reliability.

The field of aquatic accidents, both in Spain and in general, has been poorly explored, mostly limited to basic exploratory analyses to understand the situation of drownings in a specific period and country. Therefore, in this study, our aim is to delve into more detail in order to discover patterns in the data through the application of machine learning techniques. This will allow us to better understand the causes of aquatic accidents in Spain and, in turn, present the findings to rescue experts so that they can make the appropriate decisions to help reduce the number of fatalities.

**Keywords:** aquatic accident, drowning, age group, hypothesis testing, machine learning, data analysis.

# Índice general

# Capítulo 1

# Introducción

Se calcula que en el mundo mueren cada año 236.000 personas por accidentes acuáticos. Esto nos lleva a pensar que las estimaciones mundiales subestiman notablemente la magnitud real del problema de salud pública que suponen los accidentes acuáticos. Es una realidad que el riesgo de ahogamientos es mayor en niños, varones y con fácil acceso al agua, según los informes de la Organización Mundial de la Salud (OMS)[**1**].

Los informes de accidentes acuáticos de la OMS[**2**] se basan en datos obtenidos a través de la colaboración con los estados miembros, sistemas de salud nacionales, redes de vigilancia, estudios de investigación y organizaciones asociadas. Sin embargo, las conclusiones e investigaciones realizadas sobre los accidentes acuáticos se limitan a simples análisis exploratorios de los datos. Esto significa que los datos tienden a mostrar los mismos patrones una y otra vez, lo que resulta en la creación de estrategias similares para prevenir y reducir el número de accidentes acuáticos.

Es común encontrar informes anuales a nivel mundial[**3**] sobre accidentes acuáticos en los que el único factor que cambia es el número de ahogamientos, mientras que las tendencias relacionadas con la edad, el lugar, los factores y el género se mantienen constantes. La cuestión de los accidentes acuáticos ha recibido poca atención en términos de análisis e investigación. Se han realizado varias investigaciones por parte de gobiernos y organizaciones con el objetivo de encontrar formas efectivas de salvar vidas. Algunas de estas medidas incluyen proporcionar información sobre los peligros de los accidentes acuáticos, promover el cercado y drenaje de estanques de jardín y piscinas domésticas, y aumentar la supervisión en piscinas, ríos, lagos y playas para reducir el número de accidentes acuáticos. Desafortunadamente, estas soluciones no son suficientes y pueden considerarse rudimentarias.

Es por ello que este trabajo busca abordar el tema de los ahogamientos desde una perspectiva innovadora, mediante la aplicación de nuevas metodologías de análisis sobre este campo. Uno de los objetivos es examinar si existen diferencias significativas en los casos de accidentes acuáticos según los distintos grupos de edad. Para ello, se emplearán técnicas de aprendizaje automático que permitan identificar y caracterizar los diferentes grupos de edad en relación con los incidentes de ahogamiento, utilizando el método del clustering. Además, se pretende desarrollar un modelo predictivo capaz de determinar si un accidente acuático será fatal o no. Esto permitirá obtener conocimientos valiosos sobre las causas determinantes de la mortalidad en los casos de accidente acuático, utilizando un modelo de Regresión Logística Binaria. Estos hallazgos serán de gran importancia para la prevención y la intervención adecuada en situaciones de riesgo. Por último, se propone la aplicación de un modelo de serie temporal con el fin de predecir la tendencia futura de los accidentes acuáticos. Estas herramientas proporcionarán a las autoridades competentes la información necesaria para tomar decisiones oportunas en materia de seguridad y prevención.

# Capítulo 2

# Análisis del estado del arte y metodología de investigación

En el ámbito de los accidentes acuáticos, tanto a nivel mundial como en España, se observa que existe una falta de avances significativos en la investigación de este campo. A pesar de esta limitación, los países recopilan datos sobre los ahogamientos y realizan análisis sencillos al respecto. Aunque se requiere más investigación en esta área, estos informes proporcionan una base importante para comprender y abordar la problemática de los accidentes acuáticos.

## 2.1. Revisión de trabajos sobre accidentes acuáticos a nivel mundial

A lo largo del apartado 2.1 se presenta una revisión de los trabajos sobre ahogamientos a nivel mundial, enfocándose en la calidad y disponibilidad de la información para distintos países. Se destaca la variabilidad en la calidad de los datos recopilados, siendo algunos países más exhaustivos que otros en sus investigaciones. La importancia de este apartado radica en promover la recopilación precisa y completa de datos en todos los países para mejorar la comprensión y prevención de los ahogamientos a nivel global en futuras investigaciones.

### 2.1.1. Australia y Nueva Zelanda

**Royal Life Saving National Drowning**[**3**] es un informe anual que lanza Australia cada año con el fin de conocer la distribución de los accidentes acuáticos en el país. El período de análisis abarca desde el 1 de julio de 2021 hasta el 30 de junio de 2022, y durante este último año ha habido un total de 339 muertes por accidentes acuáticos. Si nos adentramos un poco más, se han registrado un total de 686 accidentes

acuáticos no mortales. Esta cifra ha aumentado un 15

En Australia, las inundaciones causaron 43 muertes por ahogamiento, principalmente en la costa este. Los hombres representan el 82 % de las víctimas, con mayores riesgos para las edades de 65 a 74 años y de 35 a 44 años. Los ríos y las playas son los lugares más comunes de fallecimiento, con un aumento del 48 % respecto al año anterior. Aproximadamente el 20 % de los accidentes acuáticos ocurren en piscinas.

**Water Safety New Zealand Drowning Prevention Report**[**4**] es un informe que Nueva Zelanda realiza anualmente, mostrando una visión general de las muertes por ahogamiento evitables en 2022, comparada con una visión a más largo plazo de las muertes por ahogamiento según la actividad, el entorno y la región.

El último año se registraron 94 muertes por accidentes acuáticos, el aumento más alto en una década, con el 85 % de las víctimas siendo hombres. Las actividades más peligrosas estuvieron relacionadas con barcos (31 %), caídas (22 %) y nadar (20 %). Las playas (26 %), piscinas (23 %) y ríos/canales (22 %) fueron los entornos más comunes de los accidentes. Los grupos de edad más afectados fueron el de 35 a 44 años y el de 65 a 74 años, representando el 20 % de las muertes en 2022.

Ambos informes presentan resultados de accidentes acuáticos de manera informativa, mostrando la evolución a través de análisis exploratorios. Aunque las cifras de muertes varían ligeramente entre países, los patrones de edad, género y entorno son similares.

### 2.1.2. Estados Unidos y Canadá

El grupo de investigación **Centers for Disease Control and Prevention**[**5**] de Estados Unidos dedica un espacio para recopilar los datos de los accidentes acuáticos anualmente y cómo prevenirlos. Los datos incluyen accidentes acuáticos como resultado de nadar, jugar en el agua o caídas. También se incluyen los ahogamientos por inundaciones y navegación. En 2022, se presentan datos provisionales sobre los accidentes acuáticos en Estados Unidos, con un pico de incidencia en los meses de verano (junio a septiembre) cercano a 900 casos. Durante los otros meses, el número de ahogamientos es más bajo, con un promedio de 200 al mes en un país con casi 332 millones de habitantes.

Canadá carece de información actualizada y no dedica suficiente tiempo ni recursos para realizar análisis exhaustivos sobre los ahogamientos en el país. El informe más reciente abarca hasta el año 2020 y revela que hubo un promedio de 30 muertes y 24 hospitalizaciones por año entre 2000 y 2019. En el año 2020, debido a la pandemia, solo hubo 13 ingresos hospitalarios y 154 visitas a urgencias, cifras por debajo de la

media. Los grupos de edad más afectados son los menores de 19 años, mientras que los hombres presentan una mayor incidencia en ambos casos[**6**].

Estados Unidos y Canadá dedican pocos recursos al estudio de los accidentes acuáticos, especialmente en el caso de Canadá. Esto puede deberse a que el número de ahogamientos en estos países es relativamente bajo en comparación y no es una prioridad para ellos.

### 2.1.3. Finlandia, Suecia y Reino Unido

La **Safety Investigation Authority**[**7**] promueve la investigación sobre los ahogamientos accidentales en Finlandia. Siguen un enfoque de investigación tradicional, analizando los datos por años y tomando medidas de prevención con el fin de reducir el número de accidentes, basándose en análisis exploratorios de datos.

En Finlandia, el número de ahogamientos ha disminuido con el tiempo. En promedio, solían ocurrir alrededor de 360 ahogamientos por año en la década de 1970, luego se estabilizaron en alrededor de 240 en las décadas de 1980 y 1990, y desde el año 2000 se ha reducido a un promedio de 200 por año. En los últimos diez años, ha habido un promedio de 147 ahogamientos anuales. La mayoría de las víctimas son hombres (88 %), la mitad de los casos ocurren en personas mayores de 67 años, y alrededor del 50 % de los accidentes son causados por intoxicación. La mayoría de los ahogamientos ocurren en aguas interiores (80 %). En 2021, la mayoría de los casos ocurrieron durante actividades de ocio cerca del hogar, y solo tres casos estaban relacionados con el trabajo, como la limpieza de un río mediante buceo.

Suecia carece de un informe anual sobre accidentes acuáticos, pero han establecido una sociedad de salvamento desde 1898 para prevenir ahogamientos mediante instrucción y medidas preventivas. Estudios pasados han revelado que el número de ahogamientos fatales supera a los no fatales, con un promedio de 200 casos anuales entre 2003 y 2017. Los hombres (67 %) tienen mayor riesgo que las mujeres (33 %), y el grupo de edad más afectado es el de 36 a 65 años. Los cuerpos de agua naturales, como mares y ríos, son los lugares más comunes para los incidentes, seguidos de las piscinas[**8**].

El Reino Unido ha desarrollado una red voluntaria llamada **National Water Safety Forum (NWSF)**[**9**] en colaboración con los Principios Guía para la Gestión de la Seguridad en el Agua, la Base de Datos de Incidentes en el Agua y la Estrategia de Prevención de Ahogamientos del Reino Unido (2016-2026). Estas iniciativas se basan en conocimientos especializados y tienen como objetivo reducir los ahogamientos y los daños relacionados con el agua en el Reino Unido a través de un enfoque coordinado y colaborativo.

El National Water Safety Forum presenta informes anuales y una visualización interactiva en Power BI para facilitar la comprensión de los datos. En el último período de 2022, se registraron 226 accidentes acuáticos, lo que representa una reducción del 18,41 % en comparación con 2021. Los hombres representaron el 82 % de los afectados, mientras que las mujeres representaron el 17 %. El mes de julio fue el período con mayor cantidad de ahogamientos, y alrededor del 50 % de los casos ocurrieron durante actividades recreativas. Los grupos de edad más propensos fueron los de 21 a 25 años y los de 56 a 60 años. La mayoría de los accidentes ocurrieron mientras se caminaba, se nadaba o se realizaban actividades cerca del agua.

De todos los países, el Reino Unido es el que más información dispone, mientras que encontrar estudios que analicen los accidentes acuáticos anualmente en otros países puede resultar una tarea complicada.

### 2.1.4. Portugal, Grecia e Italia

La **Federación Portuguesa de Socorrismo**[**10**] informa que en 2021 hubo 101 muertes por ahogamiento, lo que representa un 17,2 % menos que en 2020. El 68,3 % de las víctimas eran hombres y el grupo de edad más afectado fue el de 70 a 74 años. Los ríos y mares son las zonas más comunes de incidentes.

**Safe Water Sports**[**11**] es una organización sin ánimo de lucro que trabaja para mejorar la seguridad en el agua y reducir los ahogamientos en Grecia. Según su informe de 2020, hubo 259 accidentes acuáticos, principalmente en el mar. El 70 % de los ahogamientos correspondieron a hombres, mientras que el 30 % fueron mujeres. Julio registró el mayor número de muertes, y el grupo más afectado fue el de mayores de 60 años.

Los datos sobre accidentes acuáticos en Italia están desactualizados, pero el **Instituto Superior de Salud** se encarga de recopilar la información.

## 2.2. Revisión de trabajos sobre accidentes acuáticos a nivel español

Una vez analizado el estado del arte a nivel global, podemos centrarnos en el estado del arte a nivel español. El objetivo de España es reducir el número de accidentes acuáticos mediante el diseño de medidas preventivas y eficaces a nivel nacional.

La **Real Federación Española de Salvamento y Socorrismo (RFESS)**[**12**] es la entidad dedicada a la promoción, práctica y desarrollo del Salvamento y Socorrismo, dentro de España, integrando a las Federaciones Autonómicas de salvamento y socorrismo, clubes y asociaciones deportivas, deportistas, técnicos y árbitros. Como organismo nacional promueve la formación a través de jornadas técnicas, congresos, congresos internacionales de prevención de ahogamientos, etc. También promueve la prevención a través de campañas, informes nacionales de ahogamientos, así como comparativas y estudios. En este último punto se ha podido observar que los estudios que se realizan a través de los datos no van más allá de realizar un análisis exploratorio de los mismos para sacar conclusiones de cómo ha ido el año. Para llevar a cabo los análisis se apoyan del Informe Nacional de Ahogamientos (INA) que es un informe que elabora mensualmente desde el año 2015 la RFESS con las personas ahogadas en el medio acuático a través del Sistema Integrado de Gestión de Datos de Incidencias en el Medio Acuático (SIFA) como se muestra en la figura 2.1.

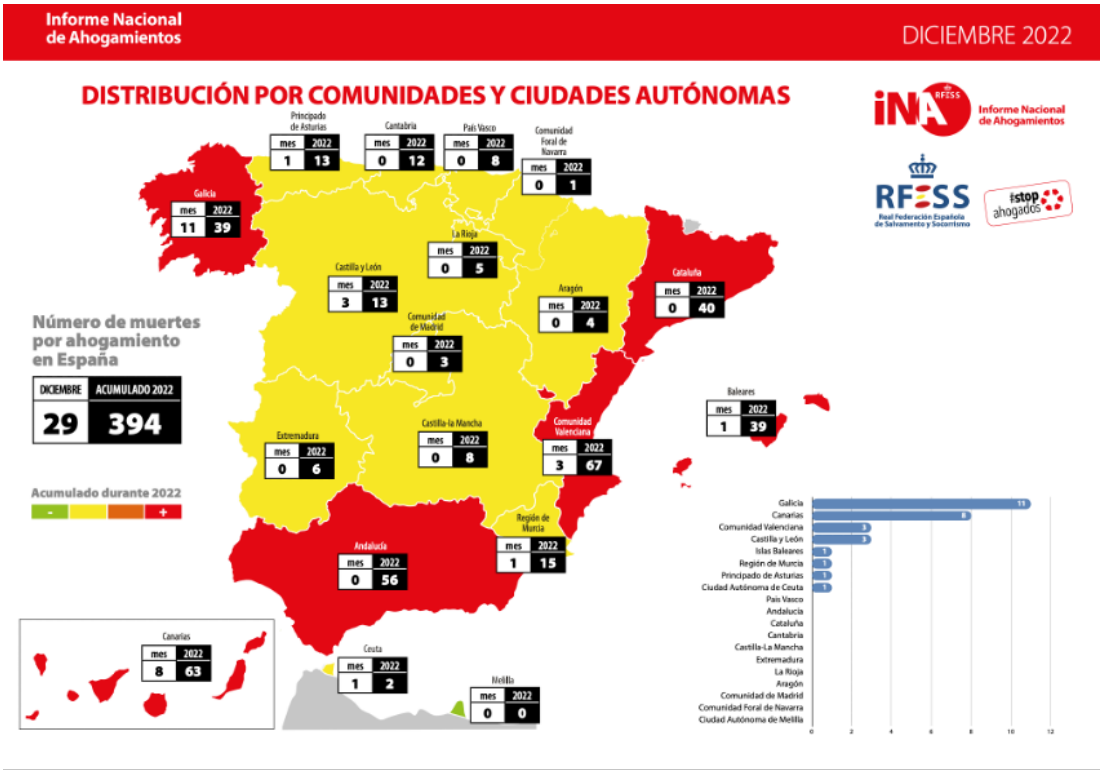



Figura 2.1: Distribución del número de muertos por comunidades y ciudades autónomas (2022). [**12**]

Por otro lado, en los últimos años se ha llevado a cabo una iniciativa conjunta en-

tre la Escuela Segoviana de Socorrismo[**13**] y la Asociación Española de Técnicos en Salvamento Acuático (AETSAS), con el objetivo de concienciar sobre la importancia de la prevención del ahogamiento tanto en España como en el resto del mundo. Esta iniciativa busca proporcionar un espacio donde se pueda compartir información, recursos y propuestas comunes, con el fin de resaltar la relevancia de este problema, brindar recursos e información útil a los profesionales en socorrismo y salvamento acuático, promover la educación de las personas y crear conciencia entre los responsables gubernamentales sobre la necesidad de implementar políticas de prevención y educación que reduzcan el riesgo de ahogamiento. Además, en nuestro trabajo utilizaremos los datos sobre ahogamientos en España correspondientes al período comprendido entre 2013 y 2020, proporcionados por la Escuela Segoviana de Socorrismo. Estos datos han sido recopilados manualmente a partir de noticias y se detallarán con mayor precisión en el capítulo 3.

Esta idea se encuentra actualmente en desarrollo con el objetivo de obtener conocimientos más profundos y aplicar estrategias de prevención eficientes basadas en datos de ahogamientos. El Instituto Nacional de Estadística no proporciona datos exhaustivos en España, por lo que es importante explorar este campo para reducir las muertes. El perfil de los accidentes acuáticos en España está principalmente relacionado con piscinas y playas, afectando especialmente a un grupo de edad mayor durante los meses de verano. España y otros países se encuentran rezagados en la exploración y aplicación de técnicas avanzadas para extraer conocimientos de los accidentes acuáticos. Actualmente, se están llevando a cabo investigaciones para utilizar el aprendizaje automático en la detección temprana de accidentes acuáticos.

En los últimos años, y especialmente a partir del año 2020, se han llevado a cabo estudios que aplican técnicas de aprendizaje automático en la detección de accidentes acuáticos. Un ejemplo destacado es la empresa estadounidense Lynxight, que se especializa en la detección de accidentes acuáticos en piscinas mediante el uso de cámaras con modelos de aprendizaje automático incorporados[**14**].

## 2.3. Enfoque del trabajo como científico de datos

El objetivo principal de este estudio es recopilar datos sobre incidentes en el medio acuático en los que exista la posibilidad real o potencial de ahogamiento, con el fin de obtener información cuantitativa y cualitativa. A partir de estos datos, se llevará a cabo un análisis estadístico y técnico exhaustivo sobre los ahogamientos en España. Además, se utilizarán técnicas de aprendizaje automático para analizar los datos en detalle y extraer conclusiones relevantes. El propósito de este análisis detallado es proporcionar a las autoridades competentes información contrastada que les permita diseñar campañas

efectivas de prevención, formación y educación dirigidas a la población.

A lo largo de este trabajo, se seguirá el flujo de trabajo característico de un científico de datos, que abarca desde la recopilación de datos hasta la aplicación de técnicas de aprendizaje automático, con el objetivo de extraer conocimiento. Este proceso incluye la realización de pruebas de hipótesis, análisis exploratorio de los datos y técnicas de Machine Learning.

Para llevar a cabo este proceso, el primer paso consistirá en obtener los datos necesarios. Para ello, se realizará un proceso ETL para extraer los datos meteorológicos de la AEMET, los cuales indican los factores meteorológicos presentes en el momento del accidente acuático. Estos datos se combinarán con los datos de ahogamientos proporcionados por la Escuela Segoviana de Socorrismo y la AETSAS. Una vez recopilada la base de datos, se procederá a realizar tareas de limpieza para prepararla adecuadamente y poder comenzar el análisis.

En el siguiente paso, realizaremos pruebas de hipótesis para investigar si el número de incidentes se ve afectado de manera similar al dividir los datos por grupos de edad. Esto nos permitirá determinar si hay diferencias significativas en el número total de incidentes entre estos grupos, así como comprender cómo se distribuyen los incidentes por edad y qué factores de riesgo están asociados a cada grupo. Seguiremos un enfoque escalonado: primero, analizaremos si existen diferencias significativas en el número total de incidentes entre los grupos de edad. A continuación, evaluaremos la importancia del ahogamiento en cada grupo. Por último, consideraremos la relación entre el número total de ahogamientos y la variable temporal (días, meses y años) por grupo de edad. A través de estas pruebas de hipótesis, obtendremos información detallada sobre la relación entre el número de accidentes acuáticos y los grupos de edad, teniendo en cuenta tanto aspectos cuantitativos como temporales. Es importante aclarar que en este contexto, se considera un ahogamiento como un evento fatal, mientras que un incidente implica que la persona estuvo involucrada en un accidente acuático, el cual puede o no haber resultado en muerte.

Después de verificar la significancia del número de incidentes en relación con los grupos de edad, procederemos a realizar una visualización de los datos, considerando también estos grupos. El objetivo es presentar la información de manera más visual y contrastarla con diferentes variables relevantes del conjunto de datos. En la visualización, exploraremos aspectos como la distribución de los incidentes por grupos de edad en relación con variables como el sexo, la provincia, la presencia de vigilancia, las técnicas de reanimación, entre otras. Al utilizar esta aproximación visual, podremos analizar de forma más efectiva las diferentes variables de interés y obtener una comprensión más

completa de la información contenida en el conjunto de datos.

Por último, nuestro objetivo es aplicar técnicas de aprendizaje automático para obtener un mayor conocimiento sobre los factores que desencadenan accidentes acuáticos. Utilizaremos la técnica de clustering para identificar grupos de edad con patrones similares, lo cual nos permitirá comprender mejor los motivos detrás de estos incidentes. Además, realizaremos la predicción de la mortalidad en casos de ahogamiento, basándonos en variables explicativas relevantes. Esto nos ayudará a determinar los factores que influyen en si un incidente resulta en muerte o no. Asimismo, aplicaremos técnicas de predicción con series temporales para anticipar las tendencias futuras en de los accidentes acuáticos. Mediante el análisis de datos pasados, podremos identificar patrones de comportamiento en los accidentes acuáticos en España. Al proporcionar esta información valiosa a las autoridades competentes, podremos colaborar en la adopción de medidas preventivas adecuadas.

Todo el código y análisis utilizado para este trabajo no se ha podido mostrar en el trabajo. Para ello se adjunta el repositorio de Github[**21**]

# Capítulo 3

# Obtención y transformación de los datos

## 3.1. Obtención de los datos

Los datos relacionados con los accidentes acuáticos han sido obtenidos a partir del proyecto de investigación **"Ahogamiento en España"**, desarrollado por la Escuela Segoviana de Socorrismo Escuela Segoviana de Socorrismo y AETSAS, donde se recopilan desde 2013 datos procedentes de las noticias de prensa publicadas en los medios impresos y digitales, redes sociales y comunicaciones de los servicios de emergencia. Estos datos se clasifican e indexan de acuerdo con los criterios metodológicos predefinidos para proceder a su análisis e interpretación.

Para ello, inicialmente se diseñó una base de datos de Microsoft Access 2010 con dos tablas maestras como se muetra en la tabla 3.1, una para recoger los datos relacionados con los incidentes y otra para las víctimas. Actualmente disponemos de una única tabla Excel unificada con la recopilación de todos los datos de accidentes acuáticos para estos campos para un periodo de 8 años que abarca del año 2013 al año 2020, con un total de 8015 observaciones.

| INCIDENTE | VICTIMA |
|---|---|
| Fecha | Sexo |
| Hora | Edad |
| Vigilancia | Nacionalidad |
| Localidad | CCAA |
| Provincia | Origen |
| CCAA | Riesgo |
| Titular | Factor |
| Deteccion | Pronostico |
| Riesgo | Extraccion |
| Localizacion | Causa |
| Intervencion | TipoAhogamiento |
| Latitud | Reanimacion |
| Longitud | Actividad |
| Enlace1 | IdPersona |
| IdAhogado | |

Figura 3.1: Campos de las tablas de los accidentes acuáticos [**15**].

**Descripción de las variables:**

- **Fecha:** Esta variable representa la fecha en la que ocurrió el accidente acuático y se compone de tres partes: el día, el mes y el año. Es una variable temporal que permite identificar el momento en que ocurrió el incidente.
- **IdAhogado:** Es un identificador único asignado a cada incidente de ahogamiento registrado en el conjunto de datos. Esta variable ayuda a distinguir cada incidente de ahogamiento registrado y a realizar análisis específicos para cada caso.
- **IdPersona:** Es un identificador único asignado a cada persona involucrada en un incidente de ahogamiento. Esta variable permite distinguir cada persona involucrada en el incidente y llevar a cabo análisis específicos para cada caso.
- **Localidad:** Es el nombre de la ciudad o pueblo donde se produjo el accidente acuático. Esta variable indica la ubicación geográfica del incidente y permite realizar análisis geográficos y comparar los incidentes en diferentes lugares.
- **Provincia:** Es la demarcación administrativa dentro del país donde se produjo el accidente acuático. Esta variable ayuda a realizar análisis a nivel regional y comparar los incidentes entre diferentes provincias.
- **Comunidad Autónoma (CCAA):** Es la entidad territorial dotada de autonomía dentro del actual ordenamiento jurídico constitucional, donde se produjo el accidente acuático. Esta variable permite realizar análisis a nivel regional y comparar los incidentes entre diferentes comunidades autónomas.

- **Hora:** sta variable indica la hora en la que ocurrió el accidente acuático y se expresa en horas y minutos. Es una variable temporal que permite identificar la hora del día en que se produjo el incidente.
- **Latitud:** Es la distancia en grados, minutos y segundos con respecto al paralelo principal, que es el ecuador ($0^{o}$). Esta variable indica la ubicación geográfica del incidente en el hemisferio norte o sur.
- **Longitud:** Es la distancia en grados, minutos y segundos con respecto al meridiano principal, que es el meridiano de Greenwich ($0^{o}$). Esta variable indica la ubicación geográfica del incidente en el este o oeste del meridiano principal.
- **Sexo:** Esta variable indica el género de la persona involucrada en el incidente de ahogamiento. Toma el valor de "M"si es masculino y "F"si es femenino.
- **Edad:** Esta variable indica la edad de la persona involucrada en el incidente de ahogamiento. Es una variable numérica que permite realizar análisis sobre la edad de los afectados.
- **Nacionalidad:** Esta variable indica la nación o territorio en el que vive la persona involucrada en el incidente de ahogamiento. Permite realizar análisis sobre los incidentes que involucran a personas de diferentes nacionalidades.
- **Origen:** Esta variable indica si la persona involucrada en el incidente de ahogamiento era local, extranjera, residente, limítrofe o de otra comunidad autónoma. Permite realizar análisis sobre la procedencia de los afectados.
- **Titular:** El titular se refiere al encabezado de la noticia que busca captar la atención de los lectores y resumir el contenido de esta.
- **Causa:** Esta variable se refiere a las circunstancias que llevaron al individuo a sufrir un accidente acuático, como pueden ser errores de juicio, falta de habilidad para nadar, la presencia de obstáculos o peligros en el agua, entre otros factores.
- **TipoAhogamiento:** El tipo de ahogamiento hace referencia a la forma en que el individuo se ahogó, como puede ser por sumersión, aspiración de agua, hipotermia, entre otras causas.
- **Factor:** Los factores se refieren a las condiciones o situaciones que aumentan la probabilidad de que un individuo sufra un accidente acuático, como pueden ser la falta de conocimiento o formación en natación, la ausencia de medidas de seguridad adecuadas, entre otros.
- **Intervención:** La intervención hace referencia a las acciones que se llevaron a cabo tras el accidente acuático para asistir al individuo, como pueden ser la aplicación de primeros auxilios, la llamada a servicios de emergencia, entre otros.

- **Pronóstico:** El pronóstico se refiere a la predicción de la evolución del individuo tras sufrir el accidente acuático, como puede ser su recuperación, secuelas o incluso fallecimiento.
- **Localización:** La localización se refiere al lugar donde fue encontrado el individuo que sufrió el accidente acuático, como puede ser una piscina, un río, el mar, entre otros.
- **Riesgo:** El riesgo hace referencia a las condiciones del agua para poder realizar actividades acuáticas, considerando banderas rojas, amarillas o verdes, que indican el nivel de peligro y seguridad para los bañistas.
- **Reanimación:** La reanimación se refiere al tipo de técnicas de primeros auxilios que se utilizaron para intentar revivir al individuo tras sufrir el accidente acuático, como pueden ser la respiración boca a boca o la aplicación de desfibriladores automáticos externos.
- **Vigilancia:** La vigilancia se refiere a si había presencia de personal especializado en el momento de producirse el accidente acuático, como pueden ser socorristas, bomberos o personal médico.
- **Actividad:** La actividad se refiere a la acción que estaba realizando el individuo en el momento de sufrir el accidente acuático, como pueden ser nadar, bucear, pescar, entre otras.
- **Detección:** La detección se refiere a quién fue el primero en notar o darse cuenta del accidente acuático, como pueden ser testigos, familiares o amigos del individuo, entre otros.
- **Enlace1:** Enlace1 se refiere al titular de noticia en la que se confirma el accidente acuático.

Además, hemos introducido datos meteorológicos referentes al accidente acuático como se muestra en la tabla 3.2, de modo que nos permita conocer las condiciones climatológicas que había a lo largo del día que se produjo el ahogamiento, ya que nos puede brindar información adicional valiosa para comprender y prevenir los accidentes acuáticos, mejorando así la seguridad en dichos entornos.

Para llevar a cabo la introducción de los datos meteorológicos a la base de datos de ahogamientos se ha seguido un proceso de extracción, transformación y carga (ETL) que se explicará con mayor detalle en el siguiente apartado.

| DATOS METEOROLOGICOS |
|---|
| Indicador |
| Estacion |
| Altitud |
| TemMed |
| Precip |
| TempMin |
| TempMax |
| TempMin |
| DirViento |
| VelMedViento |
| RachaViento |
| TiempoSol |
| PresionMin |
| PresionMax |

Figura 3.2: Campos de las tablas de los datos meteorológicos (Elaboración Propia).

**Descripción de las variables:**

- **Indicador:** El indicador se refiere al número asociado que se le otorga a cada estación meteorológica.
- **Estacion:** Estación se refiere al nombre que recibe la estación meteorológica por provincia.
- **Altitud:** La altitud se refiere la distancia entre un punto de la superficie terrestre respecto el nivel del mar en el que se produjo el accidente acuático.
- **TempMed:** La temperatura media se refiere a la temperatura que hizo a lo largo del día, de media, en que se produjo el accidente acuático.
- **Precip:** La precipitación se refiere a la cantidad de litros por m2 que hubo en el día del accidente acuático (l/m2).
- **TempMin:** La temperatura mínima se refiere a la temperatura más baja que se registró a lo largo del día en que se produjo el accidente acuático.
- **TempMax:** La temperatura máxima se refiere a la temperatura más alta que se registró a lo largo del día en que se produjo el accidente acuático.
- **DirViento:** La dirección del viento se refiere en la dirección en la que sopla el viento, ya sea norte, sur, este u oeste en el momento del accidente acuático.

- **VelMedViento:** La velocidad media del viento se refiere a la velocidad que se registró a lo largo del día, de media, en que se produjo el accidente acuático en metros por segundo(m/s).
- **RachaViento:** La racha de viento se refiere al aumento repentino del viento que excede el viento promedio de 18km/h para el día en que se registró el accidente acuático.
- **TiempoSol:** El tiempo de sol se refiere al número de horas que tuvo el dia del accidente acuático.
- **PresionMax:** La presión máxima se refiere a la fuerza máxima que ejerce el aire sobre la superficie terrestre, cuantos más accidentes acuáticos cerca del mar mayor presión.
- **PresionMin:** La presión mínima se refiere a la fuerza mínima que ejerce el aire sobre la superficie terrestre, cuantos menos accidentes acuáticos cerca del mar menor presión.

## 3.2. Proceso ETL de los datos meterológicos

Un proceso ETL consiste en extraer datos de diversas fuentes, transformarlos en un formato adecuado y cargarlos en un sistema de destino. Este proceso es el que hemos seguido para extraer los datos meteorológicos de la página oficial de la Agencia Estatal de Meteorología (AEMET).

Nuestra base de datos de ahogamientos almacena la variable Fecha y Provincia, de modo que hemos querido extraer todos los datos meteorológicos para cada fecha para cada provincia, pero solo seleccionando la estación meteorológica más representativa de la provincia que normalmente es la que presenta menos valores faltantes y suele ser la estación de la capital de provincia.

Una vez obtenidos los datos de todas las estaciones para los años 2013-2020, procedemos a eliminar los duplicados. Al especificar en el código períodos como del 01-01-20XX al 31-01-20XX, el sistema considera siempre 31 días, incluso en meses que no tienen esa cantidad. Por ejemplo, si estamos en febrero y especificamos 31 días, se considerarán los primeros 3 días de marzo. Por tanto, es necesario ajustar los datos para evitar inconsistencias.

Después de este preprocesamiento de los datos meteorológicos, obtenemos un total de 160,674 observaciones y 16 columnas. A continuación, realizamos un proceso de transformación en el que fusionamos la base de datos de accidentes acuáticos (con 8,016

observaciones y 27 columnas) con los datos meteorológicos. Esto se logra mediante la unión de tablas utilizando las variables Fecha y Provincia como referencia. Al finalizar, obtenemos una base de datos con 8,015 observaciones, una menos que la base original. Esto se debe a que un dato pertenecía al año 2012, el cual no fue incluido en el proceso de extracción. La base de datos final cuenta con 41 columnas, ya que hemos excluido las variables duplicadas Fecha y Provincia presentes en ambas bases de datos para mantener una única instancia de cada una en nuestra base de datos final.

# Capítulo 4

# Test de hipótesis

Los test de hipótesis son pruebas que permiten evaluar afirmaciones o suposiciones acerca de una población y determinar si existen evidencias estadísticas para respaldarlas. La finalidad de estas es tomar decisiones objetivas basadas en el análisis riguroso de los datos.

El interés ha sido analizar la distribución de los accidentes acuáticos según grupos de edad. Para llevar a cabo el análisis, se dividió el conjunto de datos en 5 grupos de edad:

- **Grupo 1 (adolescentes):** 0-15 años
- **Grupo 2 (jovenes):** 16-30 años
- **Grupo 3 (adultos jóvenes):** 31-45 años
- **Grupo 4 (adultos mediana edad):** 46-60 años
- **Grupo 5 (mayores):** >60 años

Antes de realizar cualquier test de hipótesis hemos querido conocer la distribución de los datos para cada grupo de edad, así como la homogeneidad de la varianza entre grupos de edad. Todo el código usado para este apartado se puede encotrar en Github[**21**].

## 4.1. Número total de incidentes por grupos de edad

En este primer test de hipótesis tenemos como objetivo inferir si los grupos de edad difieren significativamente en cuanto al número total de accidentes acuáticos. Para ello hemos definido una hipótesis nula y una hipótesis alternativa.

$H_0$: No hay diferencias significativas en el número total de incidentes entre los grupos de edad

$H_A$: Existen diferencias significativas en el número total de incidentes entre al menos dos de los grupos de edad

Para ello primero hemos comprobado la distribución del número de accidentes acuáticos por grupos de edad y se observa como ninguno de los grupos de edad siguen una distribución normal en los datos. Analizando la varianza entre los grupos de edad se observa como no es homogénea entre los grupos de edad. Como se concluye que no se cumple el supuesto de normalidad ni tampoco el supuesto de homogeneidad en la varianza entre los grupos, usaremos el test de hipótesis **Kruskal-Wallis** que cumple con estos principios.

El resultado obtenido es que hay diferencias significativas en el número total de incidentes entre los grupos y también entre todos los grupos en global, donde el p-valor obtenido para cada uno de los grupos (0.00) es inferior al nivel de significancia del 0.05.

## 4.2. Número total de ahogamientos por grupos de edad

En este segundo test de hipótesis tenemos como objetivo inferir si los grupos de edad difieren significativamente en cuanto al número total de ahogamientos, pero esta vez únicamente analizando los accidentes acuáticos que acabaron en óbito. Para ello hemos definido una hipótesis nula y una hipótesis alternativa.

$H_0$: No hay asociación entre el grupo de edad y el número de incidentes mortales.
$H_A$: Existe asociación entre el grupo de edad y el número de incidentes mortales.

Para ello hemos creado una variable fictícia a partir de la variable prónostico del dataset que toma el valor 1 en caso de que el incidente fuese mortal y 0 en caso contrario. Seguidamente realizamos la división del número de muertes por grupos de edad y creamos una tabla de contingencia de las frecuencias de ahogamiento para posteriormente realizar el test de **chi-cuadrado**.

El resultado obtenido es que efectivamente si hay evidencia estadística para afirmar que el grupo de edad influye en la mortalidad del accidente acuático, obteniendo un p-valor del 9.21304813991829e-37, muy inferior al nivel de significancia 0.05.

## 4.3. Número total de ahogamientos por grupos de edad por temporalidad

Con este test de hipotesis queremos ver si hay diferencias significativas en el número de ahogamientos totales por grupos de edad valorando la temporalidad, ya sea por años, meses y dias.

Para llevar a cabo este análisis hemos creado 3 apartados diviendo los datos por año, mes y dia. La hipótesis nula y alternativa del modelo es la siguiente:

$H_0$: No hay diferencias significativas en el número total de ahogamientos entre los grupos de edad por año/mes/dia
$H_A$: Existen diferencias significativas en el número total de ahogamientos entre al menos dos de los grupos de edad por año/mes/dia

Para cada uno de los análisis temporales se ha evaluado la distribución de los datos, así como la homogeneidad de la varianza entre los grupos de edad.

- **Variable temporal año:** Analizando la variable temporal del año observamos como la distribución de los datos por grupos de edad no siguen un distribución normal. En cambio, analizando la varianza entre los grupos, esta vez observamos que todos ellos comparten homogeneïdad en ella. De todos modos, consideramos que lo más sensato es seguir aplicando **Kruskal-Wallis**, ya que los datos por grupos de edad no siguen una distribución normal.

  La prueba de Kruskal-Wallis indica que hay diferencias significativas en el número total de ahogamientos entre al menos dos de los grupos de edad cuando se considera la temporalidad por año. Esto significa que las distribuciones de ahogamientos varían de manera significativa entre los diferentes grupos de edad cuando se analiza el año. Al examinar los resultados de la prueba para cada combinación de grupos de edad, se puede observar que en algunos casos no se encontraron diferencias significativas. Sin embargo, en otros casos sí se encontraron diferencias significativas, lo que indica que al menos dos grupos de edad presentan un número de ahogamientos significativamente diferente.

  Además, al realizar la prueba a nivel global (considerando todos los grupos de edad juntos), también se encontraron diferencias significativas, lo que implica que existe al menos un grupo que difiere en el número de ahogamientos de los demás grupos.

- **Variable temporal mes:** Analizando la variable temporal del mes observamos como la distribución de los datos por grupos de edad no siguen un distribución

normal. Asimismo, observamos que la varianza entre los grupos es heterogénea, excepto para el caso del grupo de edad jovenes - mayores y adultos jóvenes - adultos de mediana edad. Como se concluye que no se cumple el supuesto de normalidad ni tampoco el supuesto de homogéneidad en la varianza entre los grupos, usaremos el test de hipótesis **Kruskal-Wallis** que cumple con estos principios.

La prueba de Kruskal-Wallis indica que hay diferencias significativas en el número total de ahogamientos entre al menos dos de los grupos de edad cuando se considera la temporalidad por mes. Esto significa que las distribuciones de ahogamientos varían de manera significativa entre los diferentes grupos de edad cuando se analiza el mes. Al examinar los resultados de la prueba para cada combinación de grupos de edad, se puede observar que en algunos casos no se encontraron diferencias significativas. Sin embargo, en otros casos sí se encontraron diferencias significativas.

Además, al realizar la prueba a nivel global (considerando todos los grupos de edad juntos), también se encontraron diferencias significativas, lo que implica que existe al menos un grupo que difiere en el número de ahogamientos de los demás grupos.

- **Variable temporal dia:** Analizando la variable temporal del mes observamos como la distribución de los datos por grupos de edad no siguen un distribución normal. En cambio, analizando la varianza entre los grupos, esta vez observamos que todos ellos comparten homogeneïdad en ella. De todos modos, consideramos que lo más sensato es seguir aplicando **Kruskal-Wallis**, ya que los datos por grupos de edad no siguen una distribución normal.

  La prueba de Kruskal-Wallis indica que hay diferencias significativas en el número total de ahogamientos entre al menos dos de los grupos de edad cuando se considera la temporalidad por dias. Esto significa que las distribuciones de ahogamientos varían de manera significativa entre los diferentes grupos de edad cuando se analiza el dia. Al examinar los resultados de la prueba para cada combinación de grupos de edad, se puede observar que para todos los casos no se encontraron diferencias significativas.

  Además, al realizar la prueba a nivel global (considerando todos los grupos de edad juntos), tampoco se encontraron diferencias significativas.

En este capítulo 4, hemos llevado a cabo diferentes pruebas de hipótesis para analizar los accidentes acuáticos según grupos de edad. Dividimos los datos en cinco grupos distintos y evaluamos la distribución de los accidentes en cada grupo, así como la similitud de la variabilidad entre ellos. Encontramos diferencias significativas en el número total de incidentes entre los grupos de edad, lo que indica que la distribución de los accidentes varía en función de la edad. Además, descubrimos que el grupo de edad está

asociado con la mortalidad de los accidentes acuáticos, es decir, ciertos grupos tienen un mayor riesgo de sufrir incidentes mortales. También investigamos la influencia de la temporalidad en los accidentes y encontramos diferencias significativas en el número de incidentes según el año y el mes, pero no en el análisis diario. Estos hallazgos son fundamentales para comprender los factores de riesgo relacionados con la edad y tomar medidas preventivas adecuadas para garantizar la seguridad en los entornos acuáticos.

# Capítulo 5

# Análisis exploratorio de los datos

El análisis exploratorio de los datos es una etapa importante en la investigación científica, ya que proporciona una visión general y una comprensión profunda de los datos recopilados antes de aplicar técnicas estadísticas más avanzadas, como lo son las técnicas de aprendizaje automático. Este enfoque inicial permite identificar patrones, tendencias y posibles relaciones entre las variables.

Para este capitulo 5 nos focalizaremos en dividir el dataset en 5 grupos de edad con el objetivo de conocer si presentan diferencias significativas entre los grupos. Para llevar a cabo esta tarea, se ha dividido la muestra en cinco grupos: adolescentes, jóvenes, adultos, jóvenes, adultos de mediana edad y personas mayores. Cada grupo se caracteriza por una distribución específica de edades, y nuestro propósito es explorar si existen disparidades significativas entre ellos. Además, queremos conocer la distribución de los accidentes acuáticos por toda la península a lo largo de los años, así como la distribución de las condiciones climatológicas cuando se produjeron los incidentes a lo largo del periodo de análisis.

- **Grupo 1 (adolescentes):** 0-15 años
- **Grupo 2 (jovenes):** 16-30 años
- **Grupo 3 (adultos jóvenes):** 31-45 años
- **Grupo 4 (adultos mediana edad):** 46-60 años
- **Grupo 5 (mayores):** >60 años

Todo el código usado para este apartado se puede encotrar en el siguiente repositorio de Github[**21**].

## 5.1. Análisis exploratorio de los datos por grupos de edad

Una primera aproximación es conocer la distribución de los accidentes acuáticos por sexo para cada grupo de edad, donde se observa claramente en la figura 5.1 como la tendencia es que los hombres suelen ser más propensos o estar más expuestos a los incidentes acuáticos que las mujeres, independientemente de la edad.

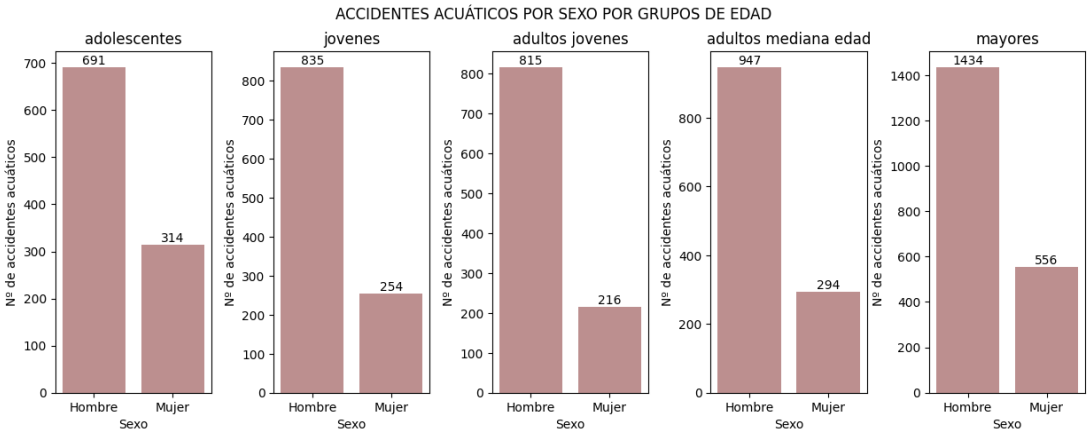



Figura 5.1: Distribución del número de accidentes acuáticos por sexo entre grupos de edad (Elaboración Propia)

Un segundo gráfico de interés es el número de accidentes acuáticos por meses por grupos de edad. La tendencia es muy clara en todos los grupos, como se muestra en la figura 5.2, donde el gran número de accidentes acuáticos se concentran en los meses de verano para todos los grupos. Si nos fijamos en la escala, los meses de verano los adolescentes y mayores suelen ser más propensos a los incidentes respecto al resto de grupos.

Otro aspecto notable que observamos es la relativa ausencia de adolescentes como protagonistas durante los meses fuera de la temporada de verano, a diferencia del resto de grupos. Esta particularidad puede atribuirse a que, durante el período escolar, los adolescentes no suelen frecuentar las playas o piscinas con la misma intensidad que el resto de grupos de edad.

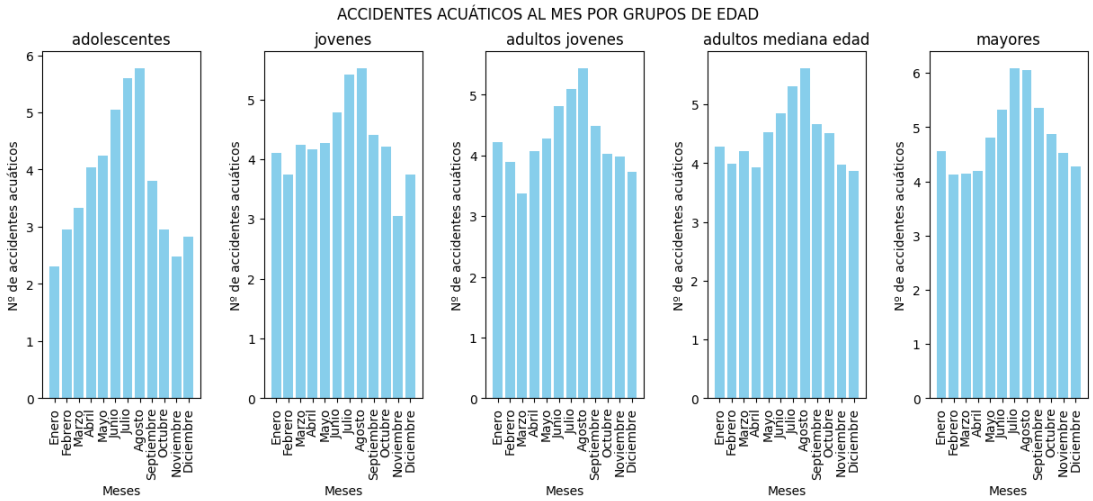



Figura 5.2: Distribución del número de accidentes acuáticos por meses entre grupos de edad (Elaboración Propia)

Otro punto interesante es conocer la distribución de los accidentes acuáticos mortales por grupo de edad. Analizando la salida de los accidentes acuáticos mortales por grupos de edad en la figura 5.3, se puede percibir a simple vista como a medida que va incrementando la edad, la probabilidad de que un accidente acuático acabe en fallecimiento es mayor, y el caso contrario en un accidente acuático en jóvenes que no acabe en fallecimiento.

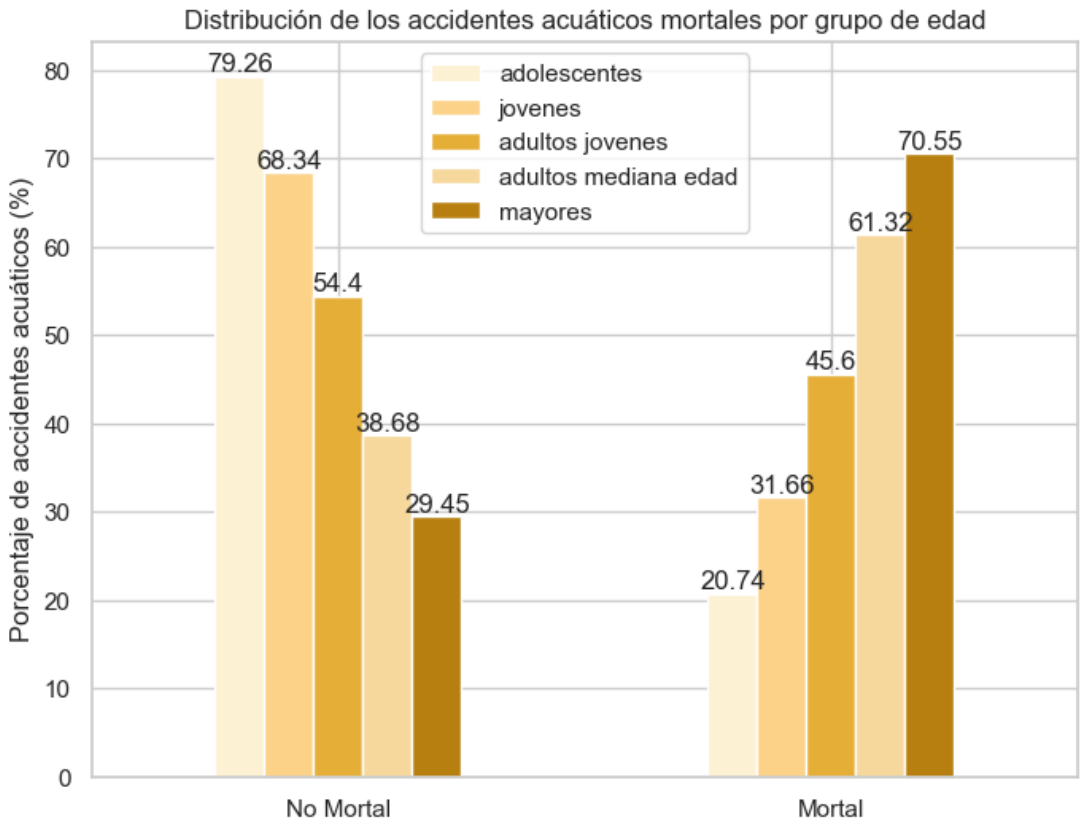



Figura 5.3: Distribución del porcentaje de accidentes acuáticos mortales entre grupos de edad (Elaboración Propia)

La ubicación geográfica de los accidentes acuáticos por provincia y grupos de edad revela patrones interesantes, como se muestra en la figura 5.4. Se identifican zonas estratégicas comunes a todos los grupos, generalmente en áreas costeras extensas. Las provincias más frecuentadas por turistas, como Islas Baleares, Cádiz, Alicante, Las Palmas y Santa Cruz de Tenerife, presentan un mayor número de incidentes acuáticos. Aunque provincias más frías como A Coruña o Pontevedra podrían asociarse inicialmente con inundaciones o riadas, la figura 5.5 demuestra que también experimentan un alto número de accidentes acuáticos en playas con vigilancia, destacando la importancia del verano en estas áreas.

Un dato curioso a comentar es el gran número de incidentes que se produce en el grupo de mayores en la provincia de Murcia, ya que a pesar de ser una zona costera, no se caracteriza por ser una zona altamente turista en comparación con el resto.

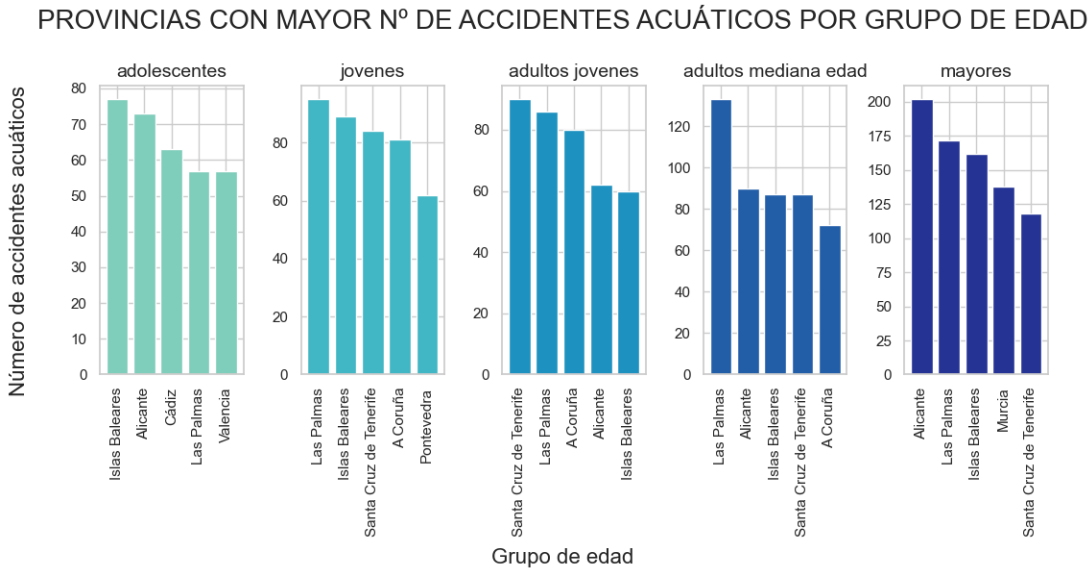



Figura 5.4: Distribución del número de accidentes acuáticos por provincia entre grupos de edad (Elaboración Propia)

La distribución de la localización del accidente acuático por comunidad autónoma nos permite saber si realmente el gran número de incidentes que se producen en el norte de España se deben a inundaciones y/o riadas, como se puede presuponer en un primer momento o también cumplen las particularidades de las zonas costeras de la península.

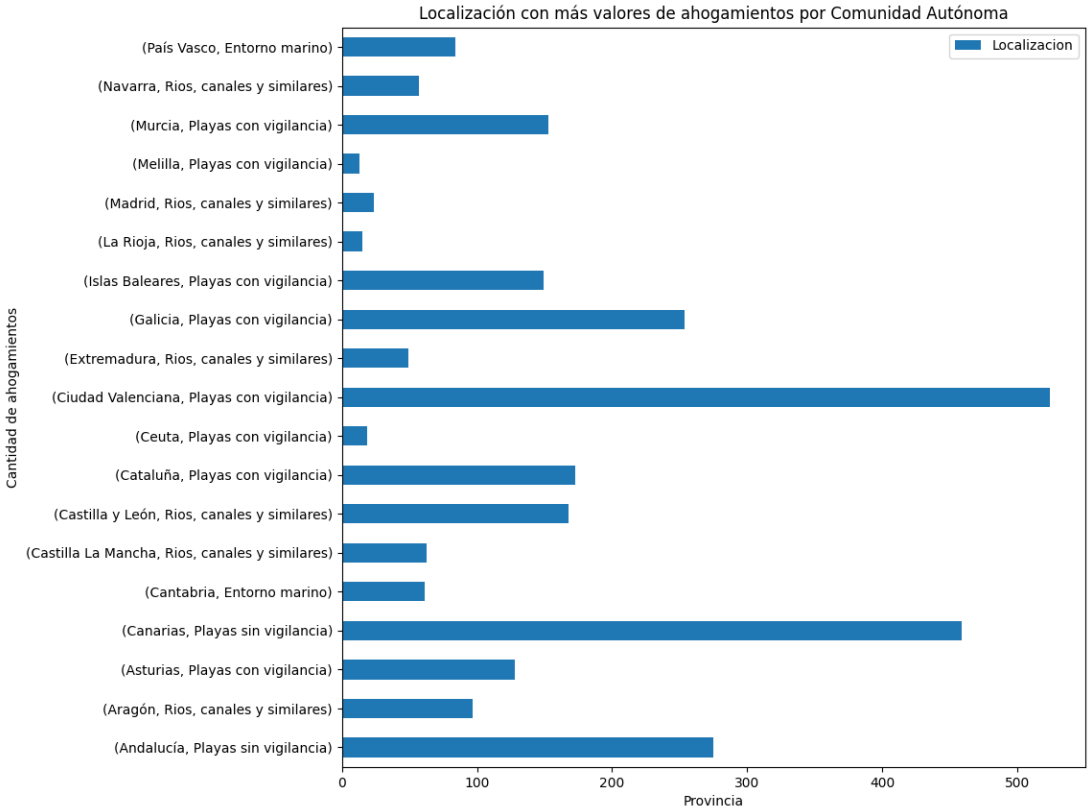
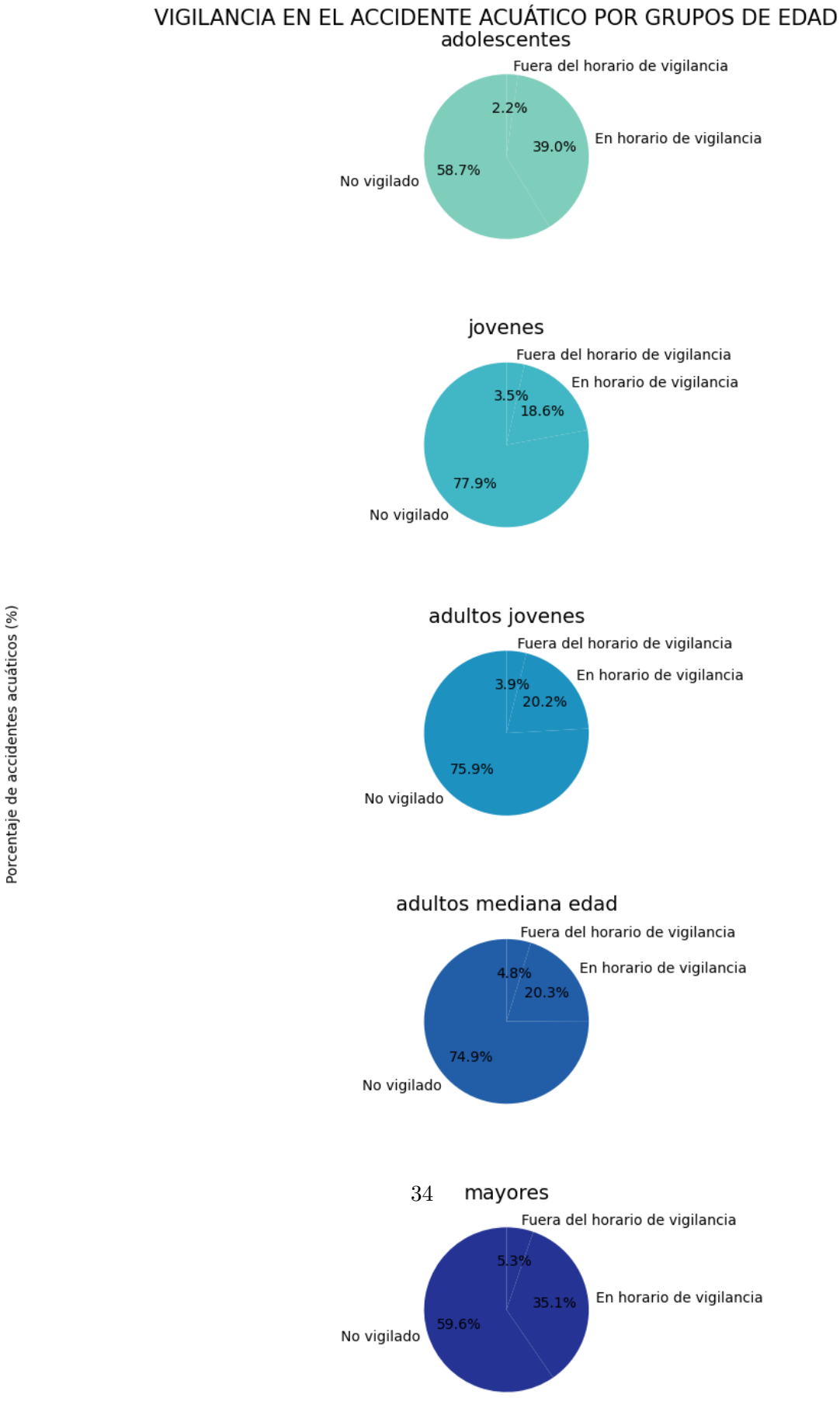



Figura 5.5: Distribución de la localización del accidente acuático por comunidad autónoma (Elaboración Propia)

Las zonas donde suelen ocurrir los accidentes acuáticos si analizamos a nivel de comunidad autónoma observamos a partir de la figura 5.5 como las áreas más costeras de la península suelen producirse accidentes acuáticos en playas con vigilancia, excepto Canarias que se producen más accidentes acuáticos cuando las playas no están vigiladas. El norte de España tenemos que la gran mayoría se producen debido a entornos marinos, más relacionados con el mar, Galicia, en cambio, se producen incidentes en las playas con vigilancia debido a sus aguas altamente agresivas. Las comunidades autónomas de interior mayormente se producen aincidentes debido a ríos, canales y similares.

Una cuestión que es interesante conocer es la vigilancia de las zonas cuando ocurre un accidente acuático como se muestra en la figura 5.6, de modo que se pueda conocer la importancia que tienen los socorristas y cuerpos de seguridad y/o emergencia en el rescate.


VIGILANCIA EN EL ACCIDENTE ACUÁTICO POR GRUPOS DE EDAD
adolescentes
Fuera del horario de vigilancia
2.2%
En horario de vigilancia
39.0%
58.7%
No vigilado
jovenes
Fuera del horario de vigilancia
En horario de vigilancia
3.5%
18.6%
77.9%
No vigilado
adultos jovenes
Fuera del horario de vigilancia
En horario de vigilancia
3.9%
20.2%
75.9%
No vigilado
Porcentaje de accidentes acuáticos (%)
adultos mediana edad
Fuera del horario de vigilancia
En horario de vigilancia
4.8%
20.3%
74.9%
No vigilado
mayores
Fuera del horario de vigilancia
5.3%
35.1%
En horario de vigilancia
59.6%
No vigilado

Se observa como más del 60 % de los incidentes se producen cuando no hay supervisión/vigilancia, independientemente del grupo de edad. La gran mayoría de los accidentes acuáticos ocurren cuando no hay socorristas o agentes que supervisen a los usuarios cuando se encuentran en contacto con el medio acuático. Este hecho habla muy bien de este colectivo, ya que cuando sí que están se consigue prevenir los accidentes acuáticos, aun así hay un gran recorrido de mejora, puesto que incluso manteniendo la supervisión obtenemos unas ratios de incidentes alrededor del 20 % de media para los grupos de edad.

Asimismo, podemos ver quién realizó la extracción del incidente acuático y corroborar que los socorristas junto con los cuerpos de seguridad y/o emergencias son los que más intervenciones realizan, como se muestra en la figura 5.7. Aunque parezca obvio, no necesariamente siempre se da el caso que las extracciones las realicen este colectivo, ya que hay extracciones realizadas por ciudadanos, acompañantes, familiares, amigos, fuerzas de orden público, helicóptero, etc.

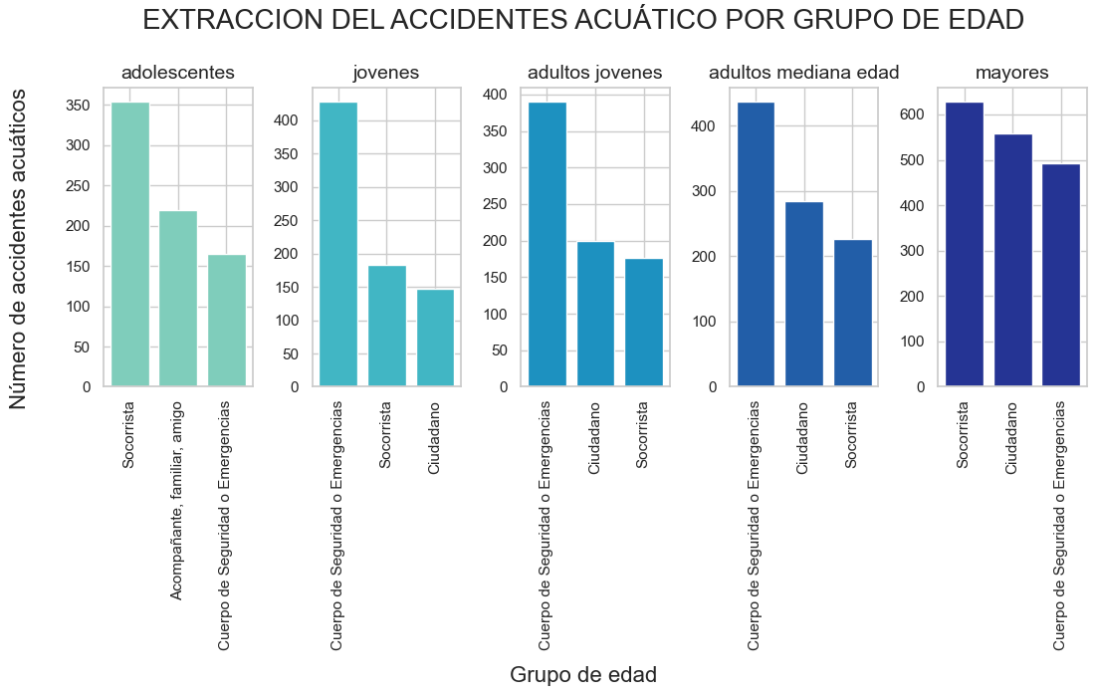



Figura 5.7: Distribución de la extracción de un cuerpo por grupo de edad (Elaboración Propia)

Se puede observar como el papel del socorrista y el cupero de seguridad y/o emergencias es fundamental para la salvación de un accidente acuático, pues aparece en

mayor o menor proporción en todos los grupos de edad.

Estos son algunos de los análisis exploratorios de los datos que se han realizado, pero para obtener el análisis completo se puede encontrar en el siguiente apartado del repositorio de Github[**21**].

## 5.2. Análisis exploratorio de los incidentes y datos climatológicos en España

La sección 5.2 tiene como objetivo analizar la distribución de los accidentes acuáticos en la península a lo largo de los años, así como la correlación existente entre dichos incidentes y las condiciones climatológicas en el momento en que se produjeron los accidentes acuáticos. El estudio de esta relación resulta importante, ya que proporciona información relevante para comprender la influencia de factores ambientales en la seguridad de los espacios acuáticos.

Me gustaría realizar una breve inspección del número de accidentes acuáticos durante el periodo de estudio 2013-2020 para identificar las zonas con mayor y menor incidencia.

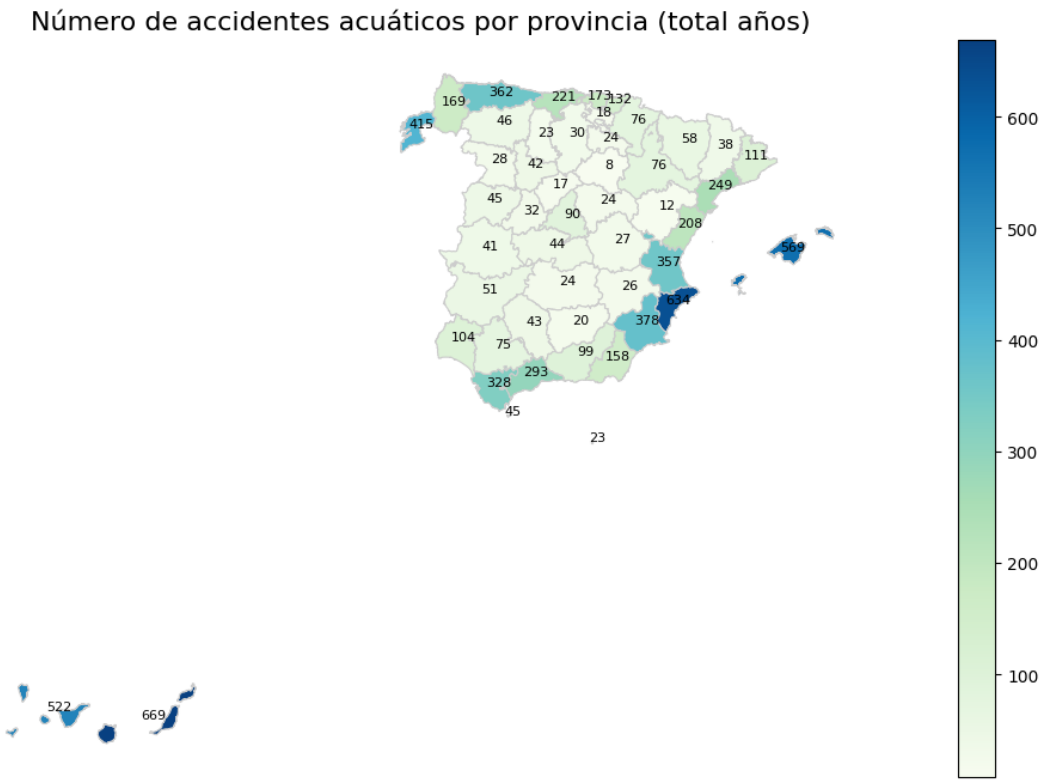



Figura 5.8: Distribución de los accidentes acuáticos por la península a lo largo de todo el periodo 2013-2020 (Elaboración Propia)

Al analizar el período de datos disponible de 2013 a 2020, como se muestra en la figura 5.8, se observa que la mayoría de los accidentes acuáticos ocurren en las Islas Baleares, Islas Canarias, la costa andaluza, la costa valenciana y catalana, y el norte de España. Esto indica claramente que los incidentes se concentran en los meses de verano, cuando las personas suelen frecuentar playas y piscinas. Aunque el resto de provincias no alcanzan valores tan significativos, también se producen incidentes relacionados con piscinas, ríos, lagos, entre otros.

La relación entre la temperatura media y el número de incidentes por áreas, así como su evolución a lo largo de los años, es relevante como se muestra en la figura 5.9.

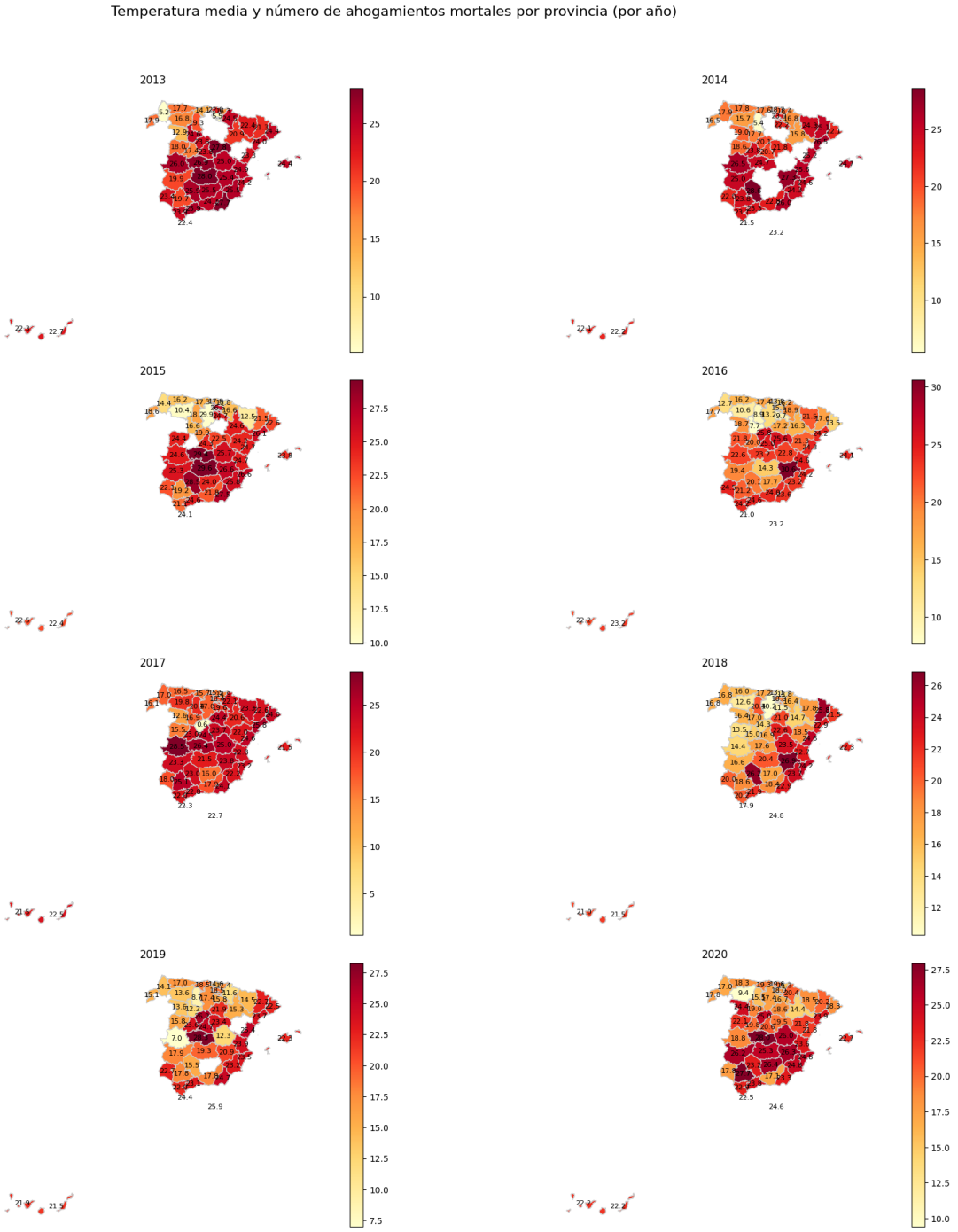



Figura 5.9: Distribución de la temperatura media y los accidentes acuáticos por la península a lo largo de todo el periodo 2013-2020 (Elaboración Propia)

Parece que hay años más calurosos que otros según los gráficos observados por año. Está claro que el norte de España se caracteriza por tener unas temperaturas medias mucho más suaves que el resto de provincias. Las costas en la península, además de las islas canarias y baleares, se observa un mayor grado de temperaturas. Al final, los incidentes en las costas se producen en épocas de verano, como se observa en los gráficos, ya que las temperaturas suelen ser más cálidas de media, mientras que en el norte los accidentes acuáticos tienen lugar en épocas del año que pueden ser veraniegas, pero las temperaturas se mantienen estables, porque suelen ser más frescas. Por otro lado, en el centro de España ocurren accidentes acuáticos en épocas cálidas, debido a actividades más relacionadas con las piscinas, ríos, lagos, embalses, etc.

En este punto hay que tener en cuenta que las temperaturas medias están recogidas para todos aquellos momentos del año que ha habido accidentes acuáticos, de modo que puede haber habido un accidente acuático en Sevilla en enero donde hiciese 18 grados y un accidente acuático en agosto donde hiciese 40 grados. En general, vemos que los accidentes acuáticos se producen cuando las temperaturas son óptimas para bañarse.

Me gustaría comentar un último gráfico realizando una comparativa entre la precipitación en el momento de producirse un incidente y el número de accidentes acuáticos por provincia para el periodo entre 2013-2020, ya que no todos los accidentes acuáticos están asociados a playas o piscinas, sino que se producen muchos otros accidentes debido a condiciones climatológicas como son las precipitaciones que se muestra en la figura 5.10.

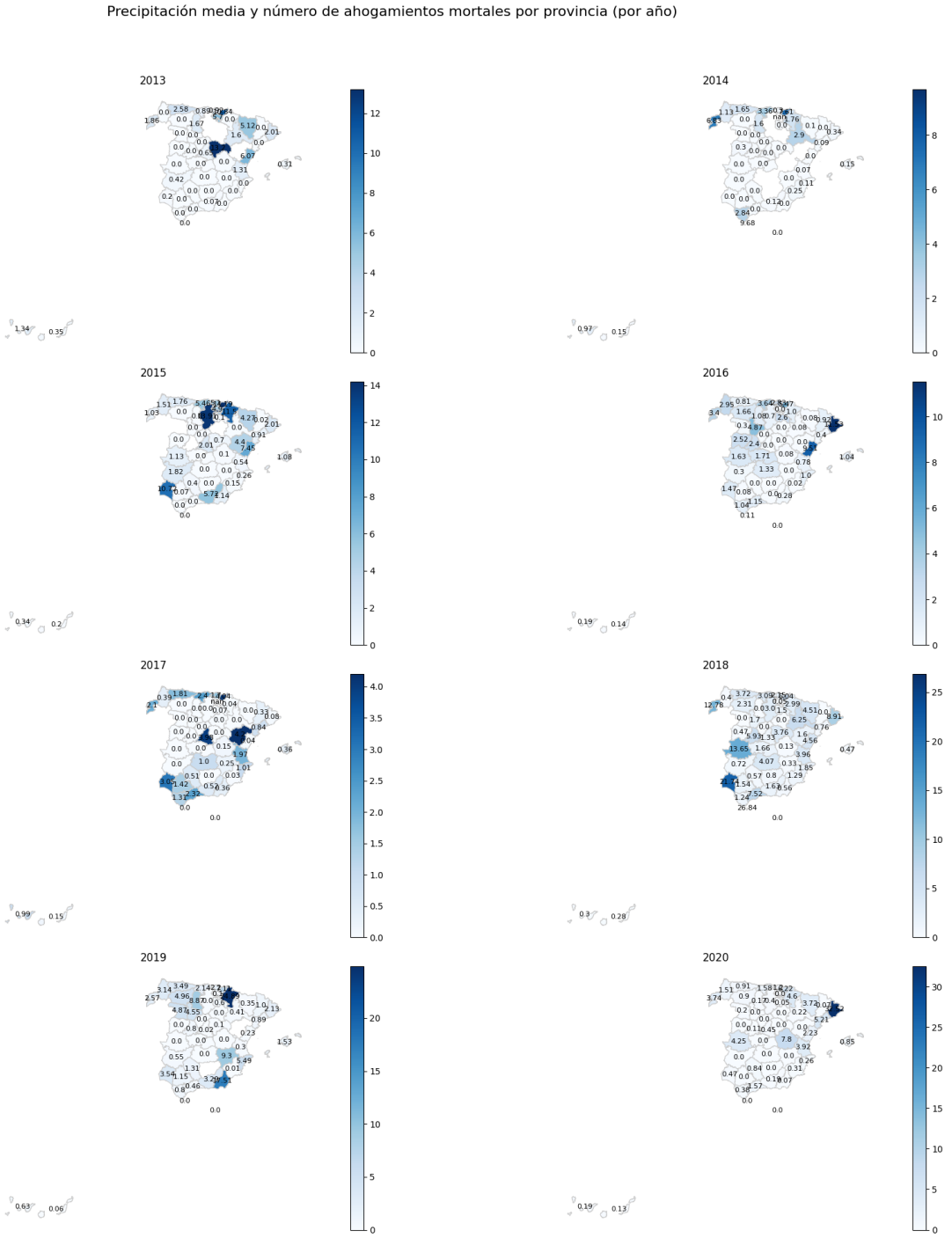



Figura 5.10: Distribución de la precipitación y los accidentes acuáticos por la península a lo largo de todo el periodo 2013-2020 (Elaboración Propia)

Las precipitaciones medias a lo largo del periodo analizado 2013 - 2020 se distribuyen de forma dispersa. Por ejemplo en 2013 las precipitaciones medias más intensas se concentran en el norte y centro de España, sobre todo en Guadalajara, pero en cambio en 2014 las precipitaciones medias fueron muy desapercibidas. En el año 2017 se observa como andalucía junto con madrid y teruel sufre precipitaciones medias superiores al resto de provincias, donde se producen muchos accidentes acuáticos por motivos climatológicos. En 2018 las precipitaciones medias se encuentran bastante repartidas por todo el territorio español, mientras que 2019 tiene fuerte relevancia Navarra y Almeria y 2020 Girona.

Todo este análisis se tiene que ver desde el punto de vista de los accidentes acuáticos. De modo que si nos encontramos que una provincia tiene altos niveles de precipitación de media en su comunidad autónoma, es indicador que el accidente acuático se deba por motivos de inundación y/o riada y no tanto por motivos veraniegos.

En resumen, el análisis exploratorio de los datos revela que los hombres tienen más accidentes acuáticos que las mujeres, sin importar la edad. La mayoría de los accidentes ocurren en verano, en áreas costeras y turísticas. La supervisión es crucial para prevenir los incidentes, y los socorristas desempeñan un papel importante en el rescate. Los accidentes están relacionados con temperaturas más cálidas y, a veces, con lluvias intensas. Este análisis proporciona información valiosa para desarrollar estrategias de prevención y promover la seguridad en espacios acuáticos durante los meses de mayor riesgo.

# Capítulo 6

# Técnicas de aprendizaje automático

En el presente capítulo 6 se aborda el concepto de aprendizaje automático, una disciplina en el campo de la inteligencia artificial que permite extraer conocimientos valiosos a partir de los datos. Se plantea la aplicación de tres modelos diferentes de aprendizaje automático para analizar los datos de incidentes en toda España durante el período comprendido entre 2013 y 2020. El objetivo principal es comprender con profundidad los perfiles de los individuos que experimentan accidentes acuáticos y su relación con diversos factores ambientales, así como investigar las causas de mortalidad asociadas a estos incidentes y detectar posibles tendencias futuras. Esta aproximación académica tiene como propósito proporcionar información relevante que pueda ser aprovechada por las autoridades competentes con el fin de mejorar la seguridad en entornos acuáticos.

Todo el código usado para este apartado se puede encotrar en el siguiente repositorio de Github[**21**].

## 6.1. Clustering - KPrototypes

En esta sección 6.1, la técnica de clustering, y en particular el algoritmo K-Prototypes, se utiliza con el objetivo de identificar perfiles o grupos similares dentro de un conjunto de datos. En el contexto de este estudio sobre accidentes acuáticos, aplicamos el clustering para descubrir diferentes perfiles de personas que han sufrido este tipo de incidentes. El objetivo es comprender las características y factores que están asociados con cada perfil, lo que nos permite tomar decisiones informadas. Al agrupar los datos en perfiles distintos, podemos obtener información valiosa sobre los factores de riesgo, las circunstancias específicas y las características comunes entre las víctimas de los

accidentes acuáticos.

En primer lugar, se han eliminado las variables que no aportan información al modelo. Aunque variables como Origen, Causa, Factor o Riesgo podrían revelar patrones de comportamiento, se decidió solo eliminar la variable Riesgo debido a la alta cantidad de valores faltantes y a su falta de mejora en modelos anteriores. En cuanto a la variable edad, con aproximadamente un 20 % de valores faltantes, se optó por utilizar únicamente las observaciones completas, basándonos en el análisis exploratorio y pruebas de hipótesis previo que mostraron un buen balance de datos en cada grupo de edad. Además, se encontró que imputar los valores faltantes no mejoraba el rendimiento del modelo. Para las variables categóricas con un bajo número de valores faltantes, se realizó una imputación utilizando la moda. Para las variables meteorológicas, se optó por imputar los valores faltantes utilizando la mediana, considerándola la mejor opción.

El siguiente paso consiste en seleccionar las variables más relevantes para aplicar la técnica de agrupamiento[**16**]. Para lograrlo, nos basamos en el análisis de selección de características para las variables categóricas y en la matriz de correlaciones para las variables numéricas. Las variables categóricas más relevantes fueron **Pronóstico y Reanimación, Causa y Factor**, mientras que las variables numéricas más relevantes fueron **Edad, Altitud, Temperatura Media, Precipitación, Dirección del viento, Racha del Viento y Tiempo de Sol**.

Una vez seleccionadas las variables, aplicamos el modelo de agrupamiento **KPrototypes**. El algoritmo de clustering KPrototypes es una extensión del algoritmo KMeans que permite agrupar datos que contienen tanto variables numéricas como categóricas. Combina la minimización de la varianza intra-cluster para las variables numéricas y la minimización de la disimilitud de Gower para las variables categóricas, con el objetivo de asignar objetos similares a los mismos grupos en función de su perfil de características mixtas[**17**]. Antes de introducir los datos al modelo, escalamos las variables numéricas para asegurar que estén en la misma escala. Como se muestra en la figura 6.1 el número óptimo de grupos para nuestro modelo es de $\mathbf{k = 4}$.

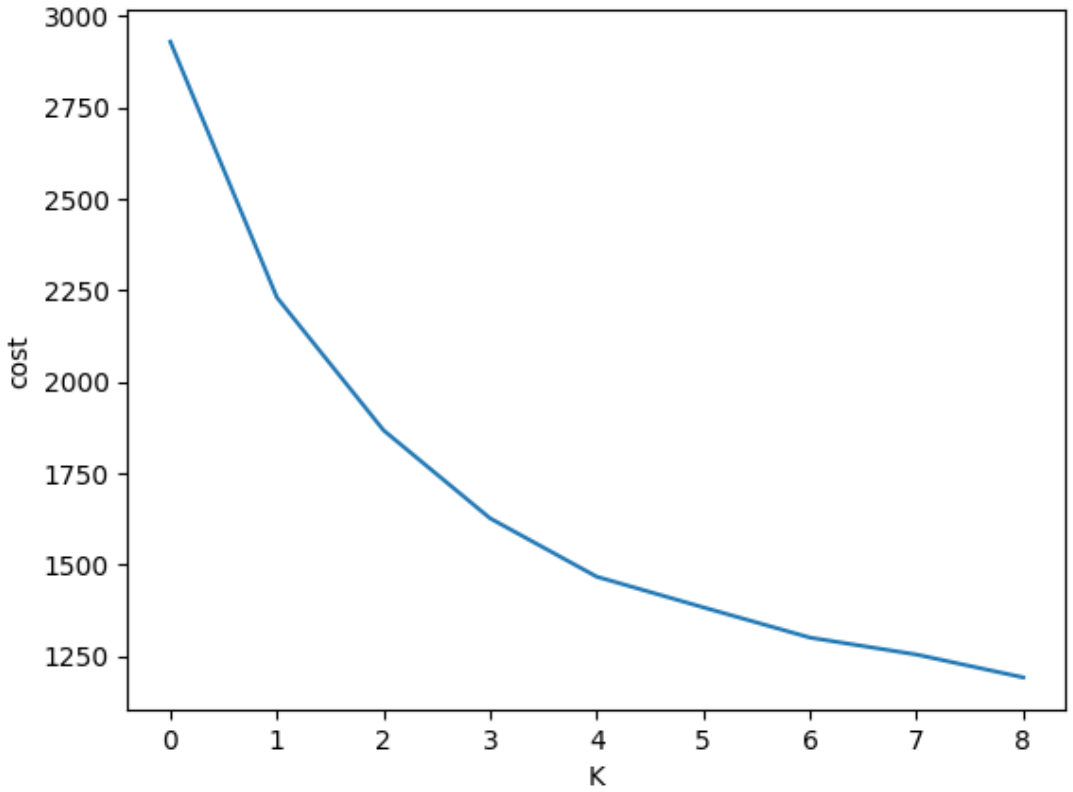



Figura 6.1: Número óptimo de grupos a aplicar al modelo KPrototype (Elaboración Propia)

Los resultados de los 4 perfiles de accidentes acuáticos se pueden observar en la figura 6.2.

| Clusters | Edad | Altitud | TempMed | Precip | DirViento | RachaViento | TiempoSol | Pronostico | Reanimacion | Causa | Factor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 22.93952 | 174.33837 | 22.88525 | 0.881648 | 17.466632 | 10.21074 | 10.189416 | Rescate sin consecuencias | Sin Reanimacion | Condiciones del medio acuático | Edad, Enfermedad, Enfermedad Mental |
| 1 | 44.70970 | 82.556536 | 22.29896 | 0.377225 | 98.823577 | 9.96263 | 9.317562 | Ahogamiento mortal | Sin Reanimacion | Condiciones del medio acuático | Edad, Enfermedad, Enfermedad Mental |
| 2 | 64.47258 | 124.72957 | 23.43651 | 0.164839 | 16.493011 | 9.988495 | 10.837366 | Ahogamiento mortal | RCP basica SOS y SVA por SEM | Condiciones del medio acuático | Edad, Enfermedad, Enfermedad Mental |
| 3 | 49.19773 | 227.21191 | 15.04429 | 5.304890 | 19.942594 | 11.46860 | 3.753863 | Ahogamiento mortal | Recuperacion de cadaver | Condiciones del medio acuático | Edad, Enfermedad, Enfermedad Mental |

Figura 6.2: Distribución de los clusters del modelo KPrototypes(Elaboración Propia)

En general, los resultados obtenidos son los siguientes:

- Perfil joven, en el que el rescate no tiene consecuencias y carece de la necesidad de reanimación, donde la causa principal es el medio acuático debido a la pérdida de control.
- Perfil adulto, en el que sufre un ahogamiento sin reanimación, donde la causa principal también es el medio acuático y debido a la pérdida de control.
- Perfil adulto, en el que también sufre un ahogamiento pero con la particularidad que no se produce tanto en playas o piscinas sino que es más en riadas o inundaciones, ya que se acaba recuperando el cadáver, siendo la causa el medio acuático y las condiciones climatológicas son el mal tiempo y la precipitación.
- Perfil mayor, en el que sufre un ahogamiento en el cual se intenta reanimar a la víctima, también debido al medio acuático pero que acaba muriendo

## 6.2. Clasificación - Bianry Logistic Regresion

En esta sección 6.2, la técnica de clasificación con el algoritmo de Regresión Logística Binaria se emplea con el objetivo de investigar las causas de mortalidad asociadas a los accidentes acuáticos. El propósito principal es identificar las variables más relevantes que influyen en la probabilidad de accidente acuático y comprender su impacto. Al aplicar este modelo, buscamos predecir y comprender mejor los factores de riesgo y las circunstancias relacionadas con los casos de accidentes acuáticos. Esto nos permite mejorar la atención y los recursos destinados a la atención de emergencias acuáticas. El objetivo final es reducir la tasa de mortalidad por accidentes acuáticos y mejorar la

seguridad y el bienestar de las personas en entornos acuáticos.

En un primer análisis, se procedió a la exclusión de variables que no aportan relevancia al modelo. A pesar de la potencial importancia de variables como Origen, Causa, Factor o Riesgo en la explicación de los incidentes se tomó la decisión de eliminarlas debido a la elevada presencia de valores faltantes, ya que la imputación de dichos valores podría tener un impacto significativo en los resultados. En relación a la variable edad, que presenta aproximadamente un 20 % de valores faltantes, se seleccionó, tras una evaluación exhaustiva de diversas opciones, la imputación mediante la media como la más adecuada. Para aquellas variables categóricas con una baja proporción de valores faltantes, se empleó la imputación basada en la moda. En cuanto a las variables meteorológicas, se optó por imputar los valores faltantes utilizando la mediana, considerándola como la estrategia más apropiada. Además, se introdujo una nueva variable denominada Ahogamiento Mortal, la cual toma el valor 1 en caso de producirse un incidente y 0 en caso contrario, siendo derivada de la variable Pronóstico.

Como paso previo hemos analizado la distribución de la variable Ahogamiento Mortal y se observa que se encuentra suficientemente balanceada, aún así a la hora de partir los datos en train y test, se han balanceado para que el modelo sea lo más robusto posible.

Como etapa preliminar, se ha llevado a cabo un análisis de la distribución de la variable Ahogamiento Mortal, evidenciando un nivel adecuado de balance. No obstante, al dividir los datos en conjuntos de entrenamiento y prueba, se ha realizado un balanceo adicional para fortalecer la robustez del modelo a través de aplicar cross validation con estratificación para balancear los datos de la variable objetivo.

Se ha empleado el algoritmo de Regresión Logística Binaria, el cual es un modelo estadístico utilizado para predecir la probabilidad de pertenencia a una clase binaria (ahogamiento o no ahogamiento). Basado en la regresión logística, este algoritmo utiliza una función logística para modelar la relación entre las variables independientes y la variable dependiente. Estima los coeficientes que maximizan la verosimilitud de los datos y permite realizar predicciones de clasificación. Dado que el algoritmo solo opera con variables numéricas, se ha aplicado el método de codificación de variables categóricas conocido como "dummies", donde las categorías se representan como variables binarias (1 si la categoría está presente, 0 en caso contrario).

Se ha empleado el método Recursive Feature Elimination (RFE) de la librería sklearn para seleccionar las variables más relevantes en la predicción de ahogamientos. Se realizó una criba inicial de las variables que no aportan información directa

y luego se implementó el modelo Logit. Se utilizaron herramientas como el VIF, que mide la multicolinealidad y evalúa la influencia de cada variable independiente en la varianza de las demás, y la matriz de correlaciones, que identifica correlaciones altas entre variables como posibles indicadores de multicolinealidad.

Una vez seleccionadas las variables más importantes, **Edad, Reanimación, Intervención y Tipo de ahogamiento**, se dividió el conjunto de datos en conjuntos de entrenamiento y prueba. Luego se aplicó la validación cruzada con estratificación para mantener el equilibrio.

Los resultados del modelo son altamente satisfactorios, con una precisión media del 87 % en los datos de prueba. En la figura 6.3 se presentan en detalle las diferentes métricas alcanzadas por el modelo.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.83 | 0.88 | 894 |
| 1 | 0.81 | 0.93 | 0.87 | 709 |
| accuracy | | | 0.88 | 1603 |
| macro avg | 0.88 | 0.88 | 0.88 | 1603 |
| weighted avg | 0.88 | 0.88 | 0.88 | 1603 |

Figura 6.3: Métricas obtenidas de la aplicación del algoritmo de Regresión Logística Binaria(Elaboración Propia)

- Podemos ver como la precisión para detectar un no ahogamiento es del 94 % (743 / (743 + 47)), mientras que la precisión para detectar un ahogamiento es del 81 % (662 / (662 + 151)).
- El recall es la capacidad del clasificador para encontrar todas las muestras positivas o negativas. Para el recuento de todos los no ahogamientos, el algoritmo es capaz de clasificar correctamente el 83 % de las obaservaciones, mientras que para el recuento de todos los ahogamientos es capaz de clasificar correctamente el 93 % de las veces.
- La puntuación F-score es una media armónica ponderada de la precisión y el recall, donde cuánto más próxima a 1 mejor. Y podemos observar que se encuentra en ambos casos cerca de 1.

- Podemos concluir que es un buen algoritmo para predecir un ahogamiento de uno que no.

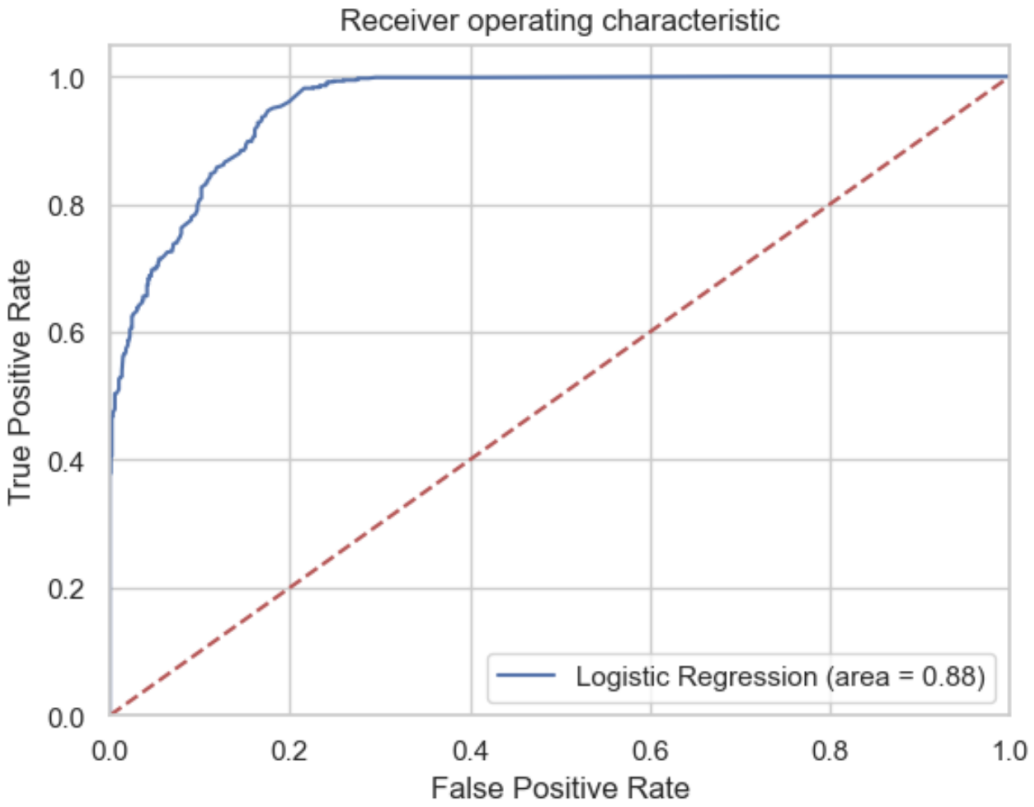



Figura 6.4: Curva ROC del alogirtmo de Regresión Logística Binaria (Elaboración Propia)

- La interpretación de la curva ROC de la figura 6.4 se basa en la relación entre la TPR y el FPR. La curva ROC muestra cómo varía esta relación a través de diferentes niveles de umbral.
- Una curva ROC ideal se acerca al vértice superior izquierdo del gráfico, lo que indica una alta TPR y un bajo FPR en todos los niveles de umbral como es nuestro caso. Cuanto más alejada esté la curva ROC de la línea de referencia diagonal (que representa el desempeño aleatorio), mejor será el desempeño del modelo y para nuestro caso se encuentra muy cerca del vértice.

En conclusión, en esta sección utilizamos la técnica de Regresión Logística Binaria

para investigar las causas de mortalidad asociadas a los accidentes acuáticos. Identificamos las variables más relevantes que influyen en la probabilidad de ahogamiento y comprender su impacto. A través de este modelo, logramos predecir y comprender mejor los factores de riesgo y las circunstancias relacionadas con los casos de accidentes acuáticos. Esto nos permite mejorar la atención y los recursos destinados a la atención de emergencias acuáticas, con el objetivo final de reducir la tasa de mortalidad por accidentes acuáticos y mejorar la seguridad y el bienestar de las personas en entornos acuáticos. Los resultados del modelo fueron altamente satisfactorios, con una precisión media del 87 % en los datos de prueba. Además, las métricas de precisión, recall y puntuación F-score demostraron un buen desempeño del modelo para predecir ahogamientos. La curva ROC también mostró un desempeño cercano al ideal, lo que indica la capacidad del modelo para distinguir entre casos de ahogamiento y no ahogamiento. Estas conclusiones respaldan la utilidad de la Regresión Logística Binaria como una herramienta efectiva para comprender y predecir los incidentes de ahogamiento en entornos acuáticos.

## 6.3. Series Temporales - SARIMA

En esta sección 6.3, la técnica de series temporales con el modelo SARIMA se utiliza con el objetivo de predecir el número de accidentes acuáticos en el futuro y comprender los patrones estacionales y tendencias asociados a estos incidentes. El modelo SARIMA permite capturar las fluctuaciones estacionales y las variaciones a lo largo del tiempo en los datos de accidentes acuáticos, lo que nos ayuda a tomar medidas preventivas y mejorar la seguridad en entornos acuáticos. Al analizar la serie temporal de incidentes, buscamos identificar los momentos de mayor incidencia, como los picos en los meses de verano, y utilizar esta información para implementar estrategias de prevención más efectivas en períodos de mayor riesgo. La predicción precisa de los accidentes acuáticos futuros nos permite estar preparados y tomar decisiones informadas para proteger a las personas y reducir la tasa de mortalidad por accidentes acuáticos[**19**][**20**].

Hemos trabajado con la variable "Fechaz el número de accidentes acuáticos registrados mensualmente. La variable Fecha abarca el periodo de 2013 a 2020, aunque no se registra un incidente todos los días, lo que genera observaciones con datos faltantes. Para abordar este problema, se realizó un proceso de relleno de datos. Primero, se completó la variable "Fechaçon todos los días desde 2013 hasta 2020. Luego, se asignó el valor 0 a aquellos días sin registros, asumiendo que la ausencia de registros indicaba la falta de accidentes acuáticos en esos días. Con esta estrategia, logramos tener una serie de tiempo completa y consistente para el análisis de incidentes.

Seguidamente, a partir de una exploración inicial de los datos, se ha observado,

como se muestra en la figura 6.5, que existen picos en los meses de verano, lo cual indica un patrón estacional consistente a lo largo del tiempo. Este hallazgo respalda la idea de que la mayor incidencia de accidentes acuáticos ocurre durante esta época del año.

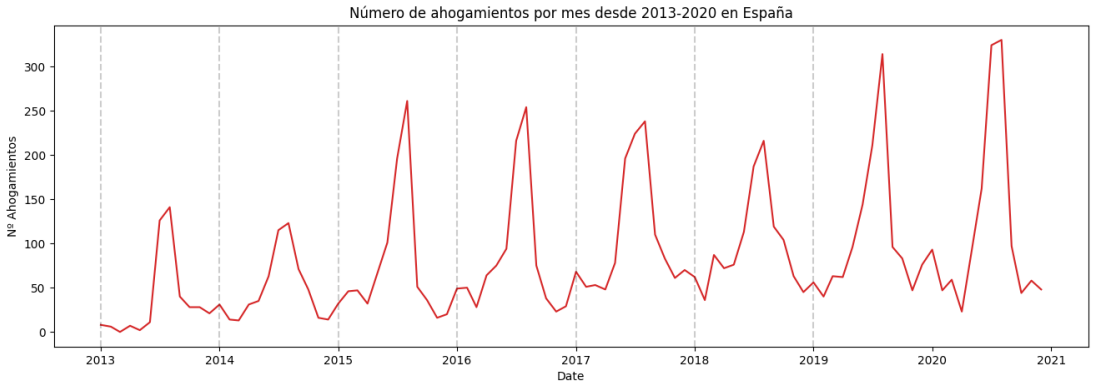



Figura 6.5: Distribución de los accidentes acuáticos a lo largo del tiempo 2013-2020 (Elaboración Propia)

A continuación, se realizó una descomposición de la serie temporal en la figura 6.6, donde se pudo observar una similitud entre la tendencia y la estacionalidad en la descomposición. Se identificó una tendencia ascendente durante los meses de verano y una estacionalidad anual asociada a los periodos estivales.

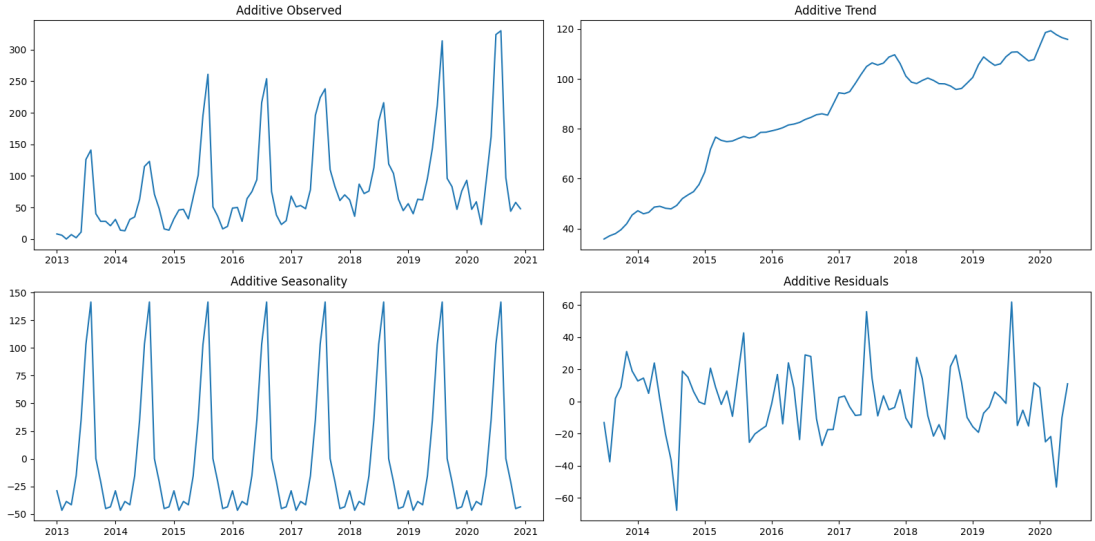



Figura 6.6: Descomposición de la serie temporal (Elaboración Propia)

Otro aspecto crucial a considerar al analizar una serie temporal es la estacionariedad, donde los valores de la serie no dependen del tiempo. La mayoría de los métodos de pronóstico estadístico están diseñados para trabajar con series de tiempo estacionarias. En este estudio, se aplicó la prueba de ADF (Dickey-Fuller Aumentada) para verificar la estacionariedad de la serie. El resultado obtenido, con un estadístico ADF significativo de -1.7803872563129084 y un valor p de 0.39024142644326826, sugiere que la serie temporal no es estacionaria, por lo que hay que diferenciar para conseguir la estacionariedad óptima en la serie.

Una vez nuestro modelo presenta estacionalidad, debemos aplicar SARIMA en vez de ARIMA, ya que el modelo SARIMA tiene en cuenta la estacionalidad y nos permitirá realizar pronósticos más precisos y acordes a los patrones cíclicos de los datos.

Un modelo SARIMA se compone de 3 términos:

- **p:** corresponde al orden del término autorregresivo (AR) en el modelo SARIMA. Indica cuántos retrasos anteriores de la serie temporal se deben tener en cuenta para predecir el valor actual.
- **q:** se refiere al orden del término de promedio móvil (MA) en el modelo SARIMA. Representa cuántos errores de pronóstico pasados se deben considerar para predecir el valor actual.

- **d:** indica el número de diferenciaciones necesarias para convertir la serie temporal en estacionaria. En nuestro caso, al haber realizado previamente la prueba de Dickey-Fuller y haber obtenido evidencia de estacionariedad, no es necesario aplicar diferenciación a nuestros datos.

El modelo SARIMA se representa como SARIMA(p,d,q)x(P,D,Q), donde P, D y Q son los órdenes de los términos SAR, la diferenciación estacional y SMA respectivamente, y 'x' es la frecuencia de la serie de tiempo. Para nuestro caso, hemos construido el mejor modelo al analizar el valor más bajo del criterio de información de Akaike (AIC), y encontramos que el modelo óptimo es SARIMA(0,0,1)(0,1,1)[12].

Una vez determinado el modelo, procedemos a realizar pronósticos para visualizar la evolución futura de los accidentes acuáticos. En la figura 6.7 se muestra el resultado de los pronósticos para los próximos 24 meses.

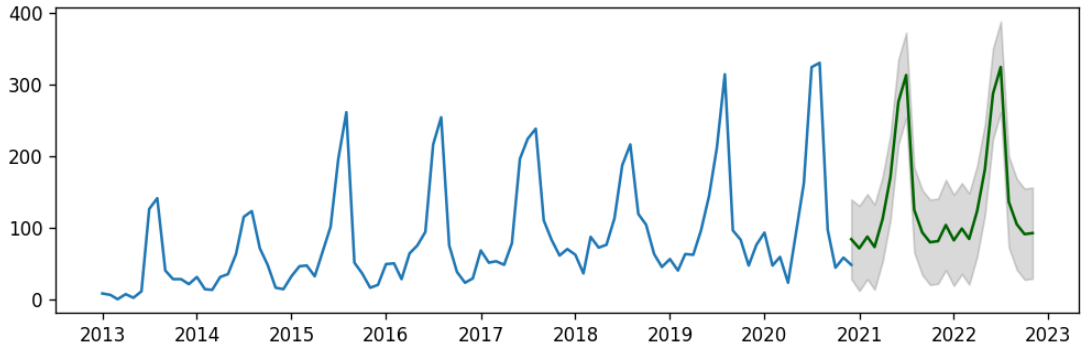



Figura 6.7: Pronóstico de la serie temporal para 24 meses (Elaboración Propia)

Se observa que en el pronóstico de la serie temporal se presentan picos que reflejan la estacionalidad de la serie en los meses de verano, mientras que en el resto de meses se mantiene una estabilidad en los valores. Esto indica la presencia de un patrón recurrente en los accidentes acuáticos, con incrementos en los meses estivales y una relativa estabilidad en el resto del año.

En conclusión, en este apartado utilizamos la técnica de series temporales con el modelo SARIMA para predecir el número de accidentes acuáticos en el futuro y comprender los patrones estacionales y tendencias asociados a estos incidentes. A través del análisis de la serie temporal de incidentes, identificamos un patrón estacional consistente con picos en los meses de verano, lo que nos permite implementar estrategias de prevención más efectivas en períodos de mayor riesgo. Además, realizamos una descomposición de la serie temporal que reveló una tendencia ascendente durante los meses

de verano y una estacionalidad anual asociada a los periodos estivales. Para abordar la estacionalidad de la serie, aplicamos el modelo SARIMA, que captura las fluctuaciones estacionales y las variaciones a lo largo del tiempo en los datos de accidentes acuáticos. El modelo SARIMA seleccionado, SARIMA(0,0,1)(0,1,1)[12], nos permitió realizar pronósticos precisos para los próximos 24 meses, mostrando un patrón recurrente de incrementos en los meses estivales y una relativa estabilidad en el resto del año. Estos resultados nos brindan información valiosa para tomar medidas preventivas, mejorar la seguridad en entornos acuáticos y reducir la tasa de mortalidad por accidentes acuáticos.

# Capítulo 7

# Conclusiones

En conclusión, este proyecto completo de data science ha abordado de manera integral el análisis de los accidentes acuáticos en España para un periodo que abarca desde 2013 hasta 2020, desde la recopilación de datos, test de hipótesis, visualización exploratoria de los datos hasta la implementación de modelos de machine learning y series temporales. A través de diversas fases y metodologías, se ha logrado alcanzar los objetivos establecidos de comprender los perfiles de las víctimas de accidentes acuáticos, identificar factores de riesgo y tendencias estacionales, y realizar predicciones de incidencia futura.

- Una área de mejora en el ámbito de los accidentes acuáticos se centra en la metodología utilizada para recopilar datos. Dado que tanto a nivel global como nacional se dedica relativamente poco esfuerzo a esta tarea, resulta difícil encontrar una recopilación de datos más precisa que la actual, como se mencionó en el Capítulo 2 de este trabajo. Esto tiene repercusión directa en cualquier análisis posterior, ya que si el dato no se recoge bien el resto tampoco funcionará bien.
- Se ha observado una distinción significativa en los ahogamientos según los diferentes grupos de edad, como se ha confirmado a través de los análisis de test de hipótesis, visualización de datos y la implementación del algoritmo de clustering KPrototypes. Estos enfoques han revelado la existencia de cuatro perfiles claramente diferenciados, tanto en términos de edad como en características específicas relacionadas con los ahogamientos.
- Además, se han utilizado dos algoritmos adicionales en este estudio. El primero se enfocó en predecir si un accidente acuático resulta en òbito o no, utilizando variables como edad, reanimación, intervención y tipo de ahogamiento como predictores significativos. El segundo algoritmo se utilizó para pronosticar la incidencia de ahogamientos en los próximos dos años, revelando la presencia de un componente estacional pronunciado durante los meses de verano. Estos análisis

complementarios han ampliado nuestra comprensión de los factores asociados con los ahogamientos y proporcionado información valiosa para la prevención y la toma de decisiones por parte de las autoridades competentes en el ámbito de la seguridad acuática.

- Aunque los resultados obtenidos hasta ahora en la serie temportal son prometedores, se requiere un análisis más exhaustivo y detallado para comprender completamente los patrones que caracterizan el comportamiento de los accidentes acuáticos a lo largo del tiempo. Estas investigaciones futuras podrían proporcionar información importante para la prevención de incidentes al permitirnos anticipar y tomar medidas proactivas basadas en los patrones identificados.
- Este estudio ha sentado las bases para animar a futuras investigaciones a utilizar técnicas de aprendizaje automático en el campo de los accidentes acuáticos.

En resumen, este proyecto ha sido un paso importante en mi formación como científico de datos, permitiéndome aplicar mis conocimientos en un contexto real y adquirir experiencia en la solución de problemas complejos utilizando técnicas de machine learning y series temporales. Ha sido un desafío apasionante y ha reforzado mi interés en continuar explorando y contribuyendo en el campo de la ciencia de datos y su aplicación en otros ámbitos de la vida.

# Bibliografía


[1] World Health Organization, *Ahogamientos*, https://www.who.int/es/news-room/fact-sheets/detail/drowning, 27 de abril de 2021.

[2] World Health Organization, *Global report on drowning: preventing a leading killer*, https://www.who.int/publications/i/item/global-report-on-drowning-preventing-a-leading-killer, 17 November 2014.

[3] Royal Life Saving Australia, *Royal Life Saving National Drowning Report 2022*, https://www.royallifesaving.com.au/research-and-policy/drowning-research/national-drowning-reports, 2022.

[4] Water Safety New Zealand, *Water Safety New Zealand Report 2022*, https://www.watersafetynz.org/, 2022.

[5] Centers for Disease Control and Prevention, *Fatal Injury Trends*, https://www.cdc.gov/injury/wisqars/fatal/trends.html, 2022.

[6] Injury Prevention Centre, *Alberta Non-fatal Drownings:2022 Report*, https://www.lifesaving.org/public/download/files/219442, September 2022.

[7] Safety Investigation Authority, *Y2021-S1 Accidental drownings 2021*, https://www.turvallisuustutkinta.fi/material/collections/20220822111005/Hlrj5o06t/Y2021-S1_Accidental_drownings_2021.pdf, 2021.

[8] ScienceDirect, *Incidence and characteristics of drowning in Sweden during a 15-year period*, https://www.sciencedirect.com/science/article/abs/pii/S030095722100037X#:~:textOver%2015%20years%2C%20a%20total,%E2%80%9317%20years%20of%20age, May 2021.

[9] National Water Safety Forum, *F2022 Annual Fatal Incident Report*, https://nationalwatersafety.org.uk/waid/annual-reports-and-data/, 2022.

[10] Federación Portuguesa de Socorrismo, *Observatório do Afogamento (Portuguese Drowning Observatory)*, http://observatoriodoafogamento.blogspot.com/, 2022.

[11] Safe Water Sports, *ANNUAL REPORT OBSERVATORY FOR WATER ACCIDENTS IN GREECE*, https://safewatersports.com/images/Documents/REPORT___FINAL_eng.pdf, 2021.

[12] Real Federación Española de Salvamento y Socorrismo, https://rfess.es/, 2023.

[13] Escuela Segoviana de Socorrismo, *Prevención de Ahogamientos*, http://www.ahogamiento.com/, última vez visitado el 7 de Julio de 2023.

[14] Lynxight, *Deep vision learning for swimmer safety and analytics*, https://cordis.europa.eu/project/id/854423/reporting, 1 February 2019 - 31 Julio 2019.

[15] Luis Miguel Pascual-Gomez, Diego García-Saiz *Cinco años de ahogamiento en España*,https://www.researchgate.net/publication/326972847_Cinco_anos_de_ahogamiento_en_Espana, ResearchGate, 4-8, 2018.

[16] Brian Roepke *4 Methods that Power Feature Selection in a Machine Learning Model*,https://www.dataknowsall.com/featureselection.html, 8 April 2022.

[17] Audhi Aprilliant *The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)*, https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type, 17 de enero de 2021.

[18] Susan Li *Building A Logistic Regression in Python, Step by Step*, https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8, 29 September 2017.

[19] Selva Prabhakaran *Time Series Analysis in Python – A Comprehensive Guide with Examples*, https://www.machinelearningplus.com/time-series/time-series-analysis-python/, 13 February 2019.

[20] Selva Prabhakaran *ARIMA Model – Complete Guide to Time Series Forecasting in Python*, https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/, 22 August 2022.

[21] Adrià Nova Pagés *Trabajo Final de Máster (TFM) - Aplicación de técnicas de Machine Learning sobre datos de ahogamiento en España*, https://github.com/adrianova8/Datos_Ahogamientos_Espana, 7 Julio 2023.