Facultad de Ciencias

Análisis de datos aplicado a incidentes acuáticos utilizando técnicas de Machine Learning

(Data analysis applied to aquatic incidents using Machine Learning techniques)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Luis Miguel Garay Teja

Director: Diego García Saiz

Co-Director: Francisco Matorras Weinig

Octubre - 2020

Agradecimientos:

En primer lugar, quiero agradecer a Diego haberme dado la oportunidad de realizar este proyecto, ayudándome y guiándome en todo momento, siempre con predisposición y la mejor de las actitudes. Por supuesto también a Francisco por su apoyo y consejos.

A mis padres, mi hermano y mis abuelos, que han sido, son, y serán un pilar fundamental para mí, sin ellos no habría sido capaz de llegar hasta aquí.

A Patricia, por sostenerme y empujarme.

También me quiero acordar de mi compi de trabajo Lina, siempre dispuesta a darme ideas cuando le contaba a cerca del proyecto.

Y por supuesto a mis amigos… ya saben ellos el por qué ☺.

Tabla de contenido

Resumen

El presente trabajo de Fin de Grado (TFG) trata sobre la realización, de principio a fin, de un proyecto de *data science*, acometiendo todas las fases que este tipo de proyectos llevan asociadas.

El *data science* es un campo interdisciplinar que aglutina método científico, procesos y sistemas con el objetivo de extraer información útil de datos en sus diferentes formas. Para ello conjuga diferentes campos como la estadística, el *machine learning* y la analítica predictiva. Hoy en día es especialmente relevante no solo en el mundo de la empresa privada, sino también en el de la investigación, como podría ser la física. Gracias a que las herramientas y técnicas propias de este campo proponen formas alternativas de trabajar con los datos recopilados en los proyectos en los que se usen, pueden llegar a facilitar la resolución de los diferentes problemas que se aborden.

En el caso concreto que nos ocupa, tendremos dos fuentes de información principales, a saber, datos meteorológicos, y datos relativos a incidentes acuáticos. El objetivo principal de este proyecto es aplicar la metodología y técnicas propias del *data science* con el fin de caracterizar los incidentes en base a los datos disponibles, con especial atención a los datos climáticos. En el proyecto se aplicarán diferentes técnicas estadísticas y se profundizará en la extracción del conocimiento mediante la aplicación de técnicas de *machine learning*.

Palabras clave: Data science, machine learning, estadística, correlación, significancia, clustering, árboles de Decisión

Abstract

In this Final Degree's Project we develop a full Data Science project from beginning to end, undertaking all the phases associated with this kind of project.

Data Science is an interdisciplinary field that brings together scientific method, processes and systems with the aim of extracting useful information from data in its different forms, combining different fields such as statistics, machine learning, and predictive analytics. Today it is especially relevant not only in the private business field, but also in the research field, such as physics. Thanks to the fact that the tools and techniques of this field propose alternative ways of working with the data collected in the projects in which they are used, they can facilitate the resolution of the different problems that are addressed.

In the specific case that concern us, we will have two main sources of information, namely, meteorological data, and data related to aquatic incidents. The main objective of this project will be to apply the methodology and techniques of Data Science in order to (extract useful information?) determine if there are weather patterns, correlations between the different variables, and we will perform machine learning models in order to make predictions.

In the specific case at hand, we will have two main sources of information, namely, meteorological data, and data related to aquatic incidents. The main objective of this project is to apply the methodology and techniques of data science in order to characterize incidents based on the available data, with special attention to climate data. Different statistical techniques will be applied in the project and the extraction of knowledge will be deepened through the application of machine learning techniques.

Keywords: Data science, machine learning, statistics, correlation, significance, clustering, Decision trees

1 Introducción

En los últimos años ha habida una explosión en la cantidad de datos generados y recopilados por dispositivos, empresas, asociaciones, gobiernos y otras entidades y entornos. Cualquier dispositivo electrónico hoy en día está procesando y generando una cantidad ingente de datos continuamente. Eric Schmidt, entonces director ejecutivo de Google, afirmó en 2010 que generamos más datos en un día que todos los que produjimos hasta 2003. Un análisis más detallado mostró que la cantidad de datos grabados y replicados en 2002 era equivalente a una semana de colección y transmisión de datos en 2011 [1]. Con el objetivo de dar sentido a esos datos, entenderlos, poder trabajar con ellos, extraer información útil de lo que a priori solo parecen GB de información, a veces sin sentido, nace el *data science*.

La definición de esta nueva disciplina puede ser algo difusa. Unas veces se confunde con el *big data*, otras con *machine learning* y a veces con la estadística. El *data science* es un campo interdisciplinar que se encarga de aplicar métodos, procesos o técnicas a datos recopilados en bruto, con el objetivo final de darles un valor añadido y permitirnos tener un mejor entendimiento de los mismos. Estas técnicas o procesos engloban diferentes campos tales como las matemáticas y estadística, el *machine learning*, la analítica predictiva, o el *business intelligence*, entre otros.

Una de las grandes ventajas del *data science*, y más concretamente el *machine learning*, es la es la capacidad de modelar, ya sean distribuciones de datos enormes en vertical (muchos registros), en horizontal (muchas columnas), o ambas dos. La capacidad de realizar estimaciones a partir de un gran número de variables permite la inferencia probabilística en situaciones que antes eran impensables [2].

Más concretamente, en el campo de la Física, el *data science* se está utilizando, por ejemplo, en diferentes proyectos del CERN dado el volumen de datos que allí se generan. Las ventajas que ofrecen las técnicas mencionadas resultan de gran utilidad. Para hacernos una idea, cada segundo se está generando 1 PB de información que, tras ser filtrada y agregada, se almacena, reduciendo drásticamente la cantidad de datos generada por los experimentos. El 29 de junio de 2017 se superó la cifra de 200 PB almacenados en sus servidores [3].

Para evidenciar la transversalidad del *data science*, se van a señalar brevemente algunos ejemplos. Concretamente se puede hablar de la utilización de técnicas de *machine learning* para encontrar evidencias de La física más allá del Modelo Estándar, desarrollando algoritmos que permiten aumentar la sensibilidad de la búsqueda para incluir escenarios inesperados [4]. Otro ejemplo podría ser la aplicación de técnicas de *machine learning* para mejorar el valor de las correcciones ópticas en aceleradores de partículas, en este caso también en los detectores del LHC [5].

Pero no solo en el CERN se utilizan técnicas de este estilo, si no en otro tipo de proyectos. En los últimos años los telescopios también han visto un gran desarrollo tecnológico que les ha permitido tomar imágenes con una calidad muy superior, recopilando de esta manera una cantidad de datos mucho mayor. Debido a esto, son capaces de, en una sola noche, recopilar la misma cantidad de datos que en un estudio completo hace un par de décadas [6].

Con este trabajo se pretende realizar un proyecto de *data science* de principio a fin, esto es, transitando por todas las fases inherentes a este tipo de proyectos: Detección y análisis del problema, obtención de los datos, preprocesado y enriquecimiento de los datos, modelado, etc.

El problema a estudiar, en este caso, es el de los incidentes acuáticos de cualquier tipo, ya sea en playas, piscinas, lagos, etc.

El objetivo principal de este proyecto es, partiendo de dos fuentes de información diferentes, a saber, datos meteorológicos y datos relativos a incidentes acuáticos, aplicar la metodología y técnicas propias del *data science* con el fin de caracterizar los incidentes en base a los datos disponibles, con especial atención a los datos climáticos. En el proyecto se aplicarán diferentes técnicas estadísticas y se profundizará en la extracción del conocimiento mediante la aplicación de técnicas de *machine learning*.

En definitiva, el principal objetivo es el de aportar valor añadido a estos datos y ser capaces de generar predicciones además de extraer conclusiones.

2 Data Science

El *data science* es un campo interdisciplinar en el que se aplican técnicas o procesos engloban diferentes campos tales como las matemáticas y estadística, el *machine learning*, la analítica predictiva, etc.

2.1 Metodología

Existen múltiples metodologías a la hora de abordar y llevar a cabo un proyecto de *data science*, no existe una concreta que invalide el resto, ya que en general su filosofía es similar. En este caso se ha tratado de ser lo más fiel posible al modelo propuesto por John Rollins, reputado científico de datos que ha desarrollado gran parte de su carrera en *IBM Analytics*, ampliamente aceptado y utilizado para proyectos de *data science*, y que, en general, consta de las siguientes fases [7], que se muestran en la Ilustración 1:

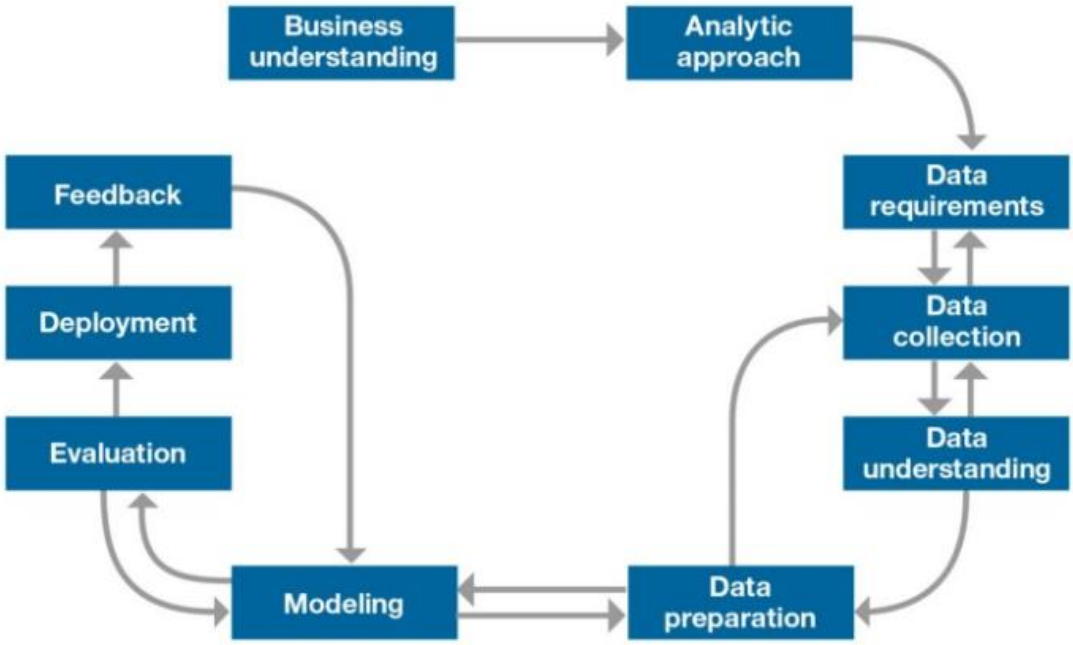

Ilustración 1 Fases de un proyecto de data science [7].

1. El proceso de comprensión del problema es crucial porque ayuda a aclarar el objetivo. Se analizan y aclaran las dudas, se definen los requisitos funcionales y técnicos manteniendo reuniones con los diferentes intervinientes.
2. El siguiente paso es el enfoque analítico, durante el cual, una vez que el problema se ha acotado, se definen los diferentes métodos y técnicas, utilizando los más apropiados en función de la problemática a estudiar.
3. Requerimientos de los datos: Esta es la etapa en la que identificamos el contenido, los formatos y las fuentes de datos necesarios para la recopilación de datos inicial.
4. Recopilación de los datos: Se identifican los recursos de datos disponibles relevantes para el dominio del problema, así como los métodos para llevar a cabo su recopilación o descarga.
5. Comprensión de los datos: Se intenta comprender más sobre los datos recopilados anteriormente. Se trata de comprobar el tipo de cada dato, y de conocer más sobre los atributos y sus nombres.
6. Preparación de los datos: Se preparan los datos para el modelado, que es uno de los pasos más cruciales ya que el dataset ha de estar limpio y sin errores. Hay que asegurarse de que los datos estén en el formato correcto para el algoritmo de aprendizaje automático que elegimos en la etapa de enfoque analítico. Los nombres de los diferentes campos del dataset han de tener un nombre de columna apropiado, un valor booleano unificado (sí, no o 1, 0), etc.
7. Modelado: Este apartado se centra en aplicar técnicas estadísticas y de *machine learning* con el objetivo de extraer información de la información en base a los requisitos establecidos en las primeras fases.
8. Evaluación del modelo: Se evalúan los resultados obtenidos fruto de la aplicación de las técnicas y algoritmos para determinar si responden a los requisitos y tienen calidad suficiente. Para ello se llevan a cabo pruebas sobre diferentes conjuntos de datos, calculando en el caso que nos ocupa, por ejemplo, la precisión en los métodos de clasificación.
9. Implementación: En este punto se puede llevar a cabo desde el simple despliegue y publicación de los resultados, hasta la implementación de una aplicación.
10. Retroalimentación: A estas alturas del proyecto, generalmente, se aprovecha al máximo del feedback del cliente o destinatario. Tras la etapa de implementación se puede determinar si el modelo funciona para los propósitos que ha sido desarrollado, o no. Se analiza este feedback y se decide si se debe mejorar todo el proceso. Esto se debe a que el proceso desde el modelado hasta la retroalimentación es muy iterativo.

2.2 Machine learning

A grandes rasgos, el aprendizaje automático, o *machine learning*, es una disciplina del ámbito de la Inteligencia Artificial que a través de los datos crea y usa modelos obtenidos o generados [8]. Se entiende por *modelo* la información generada fruto de aplicar técnicas y algoritmos de *machine learning*. Es una tecnología que permite automatizar una serie de operaciones reduciendo la necesidad de que intervengan los humanos. Esto puede resultar de gran ayuda a la hora de controlar una gran cantidad de datos de una forma mucho más efectiva.

La cuestión del aprendizaje hace referencia a la capacidad del sistema para identificar series de patrones complejos determinados por una gran cantidad de atributos o parámetros. En base a lo anterior, se puede afirmar que la máquina no aprende por sí misma, sino un algoritmo de su

programación, que, a medida que va consumiendo datos y realizando iteraciones, es capaz de predecir escenarios futuros o tomar acciones de manera automática según ciertas condiciones. Como estas acciones las realiza la máquina por si misma sin intervención humana, se dice que el aprendizaje es automático.

Tradicionalmente, como podría ser en un experimento clásico de física, la forma en que se enfoca el estudio comienza con el análisis de la problemática en cuestión. A partir de dicho análisis se deducen unas reglas que posteriormente se ponen a prueba. Los patrones detectados se evalúan para determinar si se adaptan al problema planteado y se puede dar por bueno o si, por contra, se han de analizar los errores detectados y volver a comenzar el proceso desde el principio.

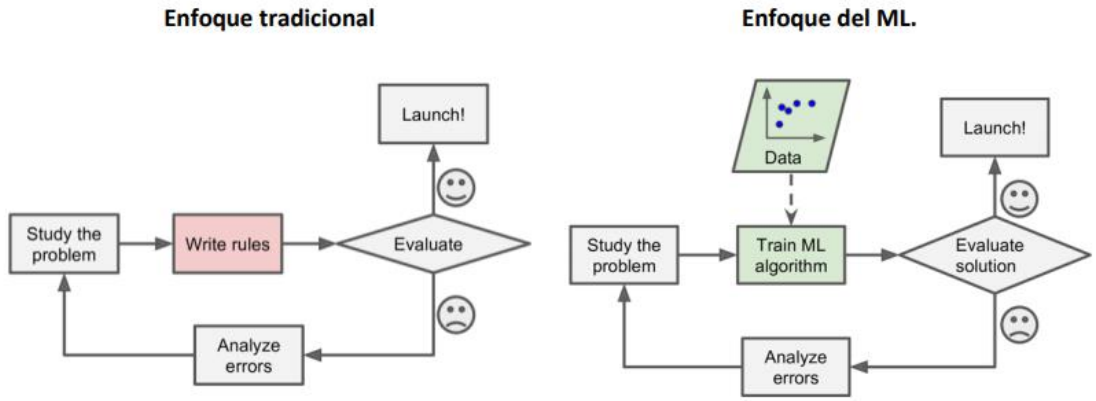

Ilustración 2. Diferencias en el enfoque tradicional y de ML [9].

El enfoque del *machine learning,* en cambio, tiene su inicio con el entrenamiento de un algoritmo a partir de los datos con los que se dispone para hacer el estudio. El algoritmo obtenido se pone a prueba y, como en el caso del enfoque tradicional, pueden detectarse o no errores. En el caso de no haberlos, o de ser inferiores a un umbral, el algoritmo puede darse por bueno. En caso contrario, se han de analizar los errores detectados y realizar cuantas iteraciones sean necesarias entrenando el algoritmo [9].

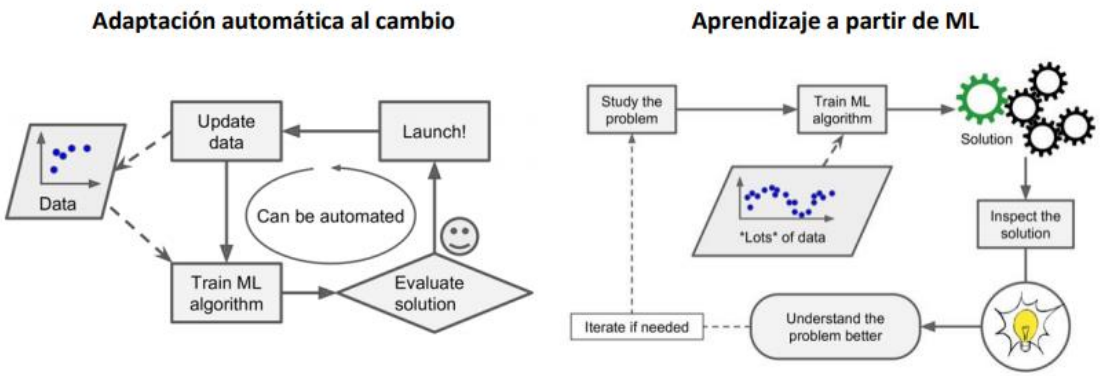

Ilustración 3. Funcionamiento del ML [9].

Existen diversos métodos de validación, como por ejemplo *cross validation, leave one out…* En este proyecto se ha utilizado el método conocido como *hold out*, consistente en:

1. Se entrena el algoritmo (70-80% de los datos).

2. Se evalúa su eficacia (30-20% de los datos).
3. Se repite el proceso cada vez que se actualiza la base de datos.

2.2.1 Clasificación de los sistemas de Machine learning

Los algoritmos de *machine learning* se clasifican en diferentes grupos. Estos pueden variar, por un lado, según el tipo de supervisión que reciben durante su aprendizaje/entrenamiento, si pueden aprender incrementalmente, cómo trabajan, etc. A continuación, se describen aquellos tipos de algoritmos que se han utilizado para el desarrollo del presente proyecto, los cuales pertenecen a las familias de supervisados/no supervisados.

Aprendizaje supervisado

Las técnicas que pertenecen a este campo se usan para predecir la clase o etiqueta de un dato basándose en un conjunto de entrenamiento. El conjunto de entrenamiento está compuesto por una serie de datos formado por un conjunto de atributos predictores, y un atributo de clase. En base al conjunto de entrenamiento, se genera un modelo que sirve para, o bien clasificar nuevos datos de los que se desconoce la clase, que puede tomar un conjunto finito de valores (problemas de clasificación), o bien realizar una predicción de un valor numérico (problemas de regresión).

Estas técnicas de aprendizaje se suelen utilizar, por ejemplo, en medicina para detectar enfermedades en base a datos de otros pacientes, detección de spam, reconocimiento de patrones, reconocimiento de objetos en visión por computador, y otros.

Algunos de los algoritmos o técnicas de aprendizaje supervisado son *K-Nearest Neighbors, Linear Regression, Logistic Regression, Decision Trees* o *Neural networks,* entre otros [10] [11].

Aprendizaje no supervisado

Este tipo de técnicas buscan patrones sin un objetivo concreto o clase asociada, es decir, el modelo se ajusta exclusivamente a las observaciones. Al contrario que con el aprendizaje supervisado, no utiliza un conjunto de entrenamiento que contenga la clase del dato, sino que los datos consisten en un vector de atributos. El objetivo de las técnicas no supervisadas consiste en buscar patrones no detectados previamente en estos *vectores de atributos*. Este método de aprendizaje se distingue del aprendizaje supervisado en que no se tiene un objetivo concreto sobre una clase o etiqueta, sino que se va generando en base a observaciones.

Ejemplos de uso de este tipo de técnicas pueden ser la agrupación de páginas web en base a etiquetas, para la compresión de datos, monitorización de contenido en internet, y otros.

Algunos de los algoritmos más utilizados en el aprendizaje no supervisado son, por ejemplo, *k-Means, DBSCAN o Apriori,* entre otros [10] [11].

2.3 Técnicas utilizadas

A continuación, se da una explicación más detallada de las técnicas concretas que se han utilizado en las fases de análisis de los datos, caracterización de los incidentes, y predicción.

2.3.1 Análisis estadístico

Con el objetivo de extraer las primeras inferencias o relaciones entre las diferentes variables del dataset, se han aplicado una serie de técnicas estadísticas, a saber, coeficientes de correlación y significancia entre variables.

Correlación

En probabilidad y estadística, la correlación indica la fuerza y dirección de una relación lineal y proporcionalidad entre dos variables estadísticas. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los de la otra: si tenemos dos variables (A y B) existe correlación entre ellas si al disminuir los valores de A lo hacen también los de B y viceversa [12].

El signo nos da cuenta de la dirección de la relación:

- Un valor positivo se refiere a una relación directa o positiva.
- Un valor negativo indica relación inversa o negativa.
- Un valor nulo significa que no se está dando una tendencia entre ambas variables (puede ocurrir que no exista relación o que la relación sea más compleja que una relación lineal, por ejemplo, una relación en forma de U).

La magnitud nos indica la intensidad de esta relación, y toma valores entre -1 a 1. Cuanto mayor sea el valor de la magnitud en valor absoluto, más fuerte será la relación de las variables, o bien, menor será la dispersión que existe en los puntos alrededor de dicha tendencia. Cuanto más cerca del cero esté el coeficiente de correlación, más débil será la relación entre las variables, es decir, la nube de puntos será más dispersa.

- Si la correlación toma valor 1 o -1 se dice que la correlación es perfecta.
- Si la correlación toma valor 0, no existe correlación entre las variables.

A pesar de no estar estandarizado y depender de la naturaleza de los datos, el criterio de Cohen está ampliamente aceptado para dar cuenta de cuán fuerte es la correlación:

- 0.1-0.3 representa una correlación débil.
- 0.3-0.5 representa una correlación intermedia.
- ≥0.5 representa una correlación alta.

Como se ha mencionado, son valores arbitrarios que pueden servir para dar una idea de la correlación que existe entre las variables, pero es recomendable contextualizar para poder interpretar la magnitud de la correlación. No es lo mismo analizar datos de un experimento físico controlado en un laboratorio donde habrá poco ruido en los datos, ya que imponen el control científico al probar una hipótesis en un entorno artificial y altamente controlado, que analizar datos sociales o biológicos donde se espera encontrar menores valores de correlación debido a la gran cantidad de dispersión o variabilidad de los datos, como podría ser el objeto de estudio de este trabajo.

En función de la distribución de los datos, tendremos dos tipos de pruebas estadísticas [13]:

- Paramétricas:
    - Pearson: Prueba paramétrica que mide una tendencia lineal entre dos variables numéricas. Se establece que en los datos:

        - La relación ha de ser de tipo lineal.
        - No se dan valores atípicos.
        - Las variables deben ser numéricas. Si las variables son de tipo ordinal no se podrá aplicar la correlación de Pearson.
        - Se tiene un juego de datos suficientemente grande (algunos autores recomiendan tener más de 30 puntos u observaciones).
- No paramétricas [14]:
    - Spearman: El coeficiente de correlación de Spearman mide una tendencia monótona (creciente o decreciente) entre dos variables. En los casos donde no se cumplen los requisitos del coeficiente de correlación lineal de Pearson, es conveniente utilizar la correlación de Spearman. Es una prueba no paramétrica (no asume una distribución previa de los datos) y es más robusta frente a la presencia de outliers que la prueba paramétrica de Pearson
    - Kendall: Cuando se estudia la relación entre variables cualitativas de tipo ordinal se debe utiliza el coeficiente de correlación de rangos de Kendall (1938), denominado τ (tau) de Kendall, del cual existen dos variantes tau-b y tau-c. Como este indicador está basado en rangos y no en los datos originales, su estimación requiere que los valores de la variable ordinal sean transformados en rangos, este coeficiente se ve poco afectado ante la presencia de un número pequeño de valores atípicos (extremos) en la muestra estudiada, adaptándose bien en aquellas variables que reportan moderadas asimetrías en torno a la relación general

Significancia entre variables

El test de significancia, o *T-Test*, es un método para evaluar los promedios de dos grupos de una variable continua, y establecer el grado de significancia que existe entre sí. El hecho de que los valores promedio de cada grupo no sean iguales no implica que haya evidencias de una diferencia significativa. Para estudiar si la diferencia observada entre las medias de dos grupos es significativa, se puede recurrir a métodos paramétricos como el basado en la distribución T-student.

La distribución T-Student, en honor a su desarrollador William Sealy Gosset (firmaba con el pseudónimo Student), es muy similar a la distribución normal. Tiene como parámetros la varianza y la media, y, además, incorpora una modificación a través de los grados de libertad que permite flexibilizar las colas en función del tamaño que tenga la muestra. El número de grados de libertad se define como número de registros de la muestra menos 1. A medida que se reduce el tamaño de la muestra, la probabilidad de que registros de la muestra caigan en las colas aumenta, siendo así menos estricta de lo cabría esperar en una distribución normal. Una distribución T-student con 30 o más grados de libertad es prácticamente igual a una distribución normal, como se puede apreciar en la ilustración 4 [15].

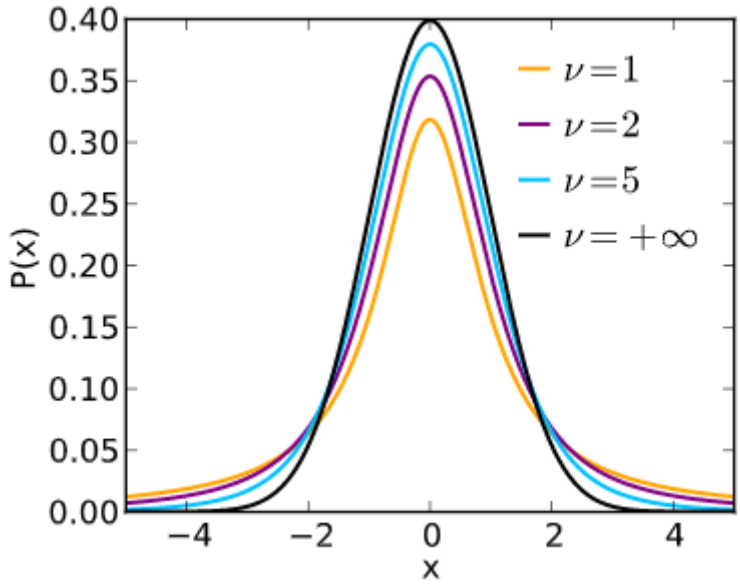

Ilustración 4. Distribución T-Student para diferentes grados de libertad. Se puede observar que a medida que aumentan estos grados de libertad, la distribución tiende a asemejarse a una distribución normal.

Las condiciones para aplicar un test de hipótesis basados en la distribución T-Student son:

- Independencia: Las observaciones han de ser únicas e independientes.
- Normalidad: Las poblaciones que se comparan tienen que distribuirse de forma similar a la distribución normal. En caso de cierta asimetría los t-test son considerablemente robustos cuando el tamaño de las muestras es mayor o igual a 30.
- Igualdad de varianza: La varianza de ambas poblaciones comparadas debe de ser igual.

En caso de no cumplirse esta última condición se puede emplear un *Welch Two Sample t-test*. Como el *T-Test* de Student, es un método para evaluar los promedios de dos grupos de una variable continua, y establecer el grado de significancia que existe entre sí, pero, en este caso, no se asume la igualdad de varianza. En este caso también se asume que la distribución de los datos se asemeja a la distribución normal, y que, por lo tanto, las medidas son únicas e independientes. Tanto el *T-Test* de Student como el de Welch arrojan como resultado el siguiente parámetro que da cuenta de cuan significantes son los grupos que se analizan:

- P-Value: Cada T-Value tiene un P-Value asociado. Un P-Value es la probabilidad de que los resultados de los datos de la muestra se produzcan por casualidad. Los valores p van del 0 al 1. Los valores p bajos son buenos; Indican que sus datos no se produjeron por casualidad. Por ejemplo, un P-Value de 0.01 significa que solo hay un 1% de probabilidad de que los resultados de un experimento hayan ocurrido por casualidad. En la mayoría de los casos, se acepta un P-Value de 0,05 (5%) para indicar que los datos son válidos [16] [17].

2.3.2 Clustering

K-Means es uno de los algoritmos de aprendizaje automático no supervisados más simples y populares. Normalmente, los algoritmos no supervisados hacen inferencias a partir de conjuntos de datos utilizando solo vectores de entrada sin hacer referencia a resultados conocidos o etiquetados.

K-Means es un método de agrupamiento, y, por lo tanto, su principal objetivo es el de la partición de un conjunto de n observaciones en K grupos, los cuales llamaremos clústeres, en el que cada observación pertenece al clúster cuyo valor medio es más cercano. La motivación para utilizar este algoritmo es la de descubrir patrones subyacentes para así poder caracterizar diferentes casuísticas que a simple vista no se apreciarían.

Una vez elegido el dataset que se utilizará como input, el primer paso será definir un número adecuado de centroides K. Un centroide es la ubicación imaginaria o real que representa el centro del clúster.

Para determinar un número adecuado de clústeres se ha utilizado el método conocido como *la regla del codo*. Este método consiste en ejecutar N veces el algoritmo k-Means, esto es, para K=1…N clústeres. Una vez hecho esto, se representa en una gráfica el número de clústeres K frente a la inercia obtenida en cada caso, obteniendo una gráfica del tipo a la representada en la ilustración 5:

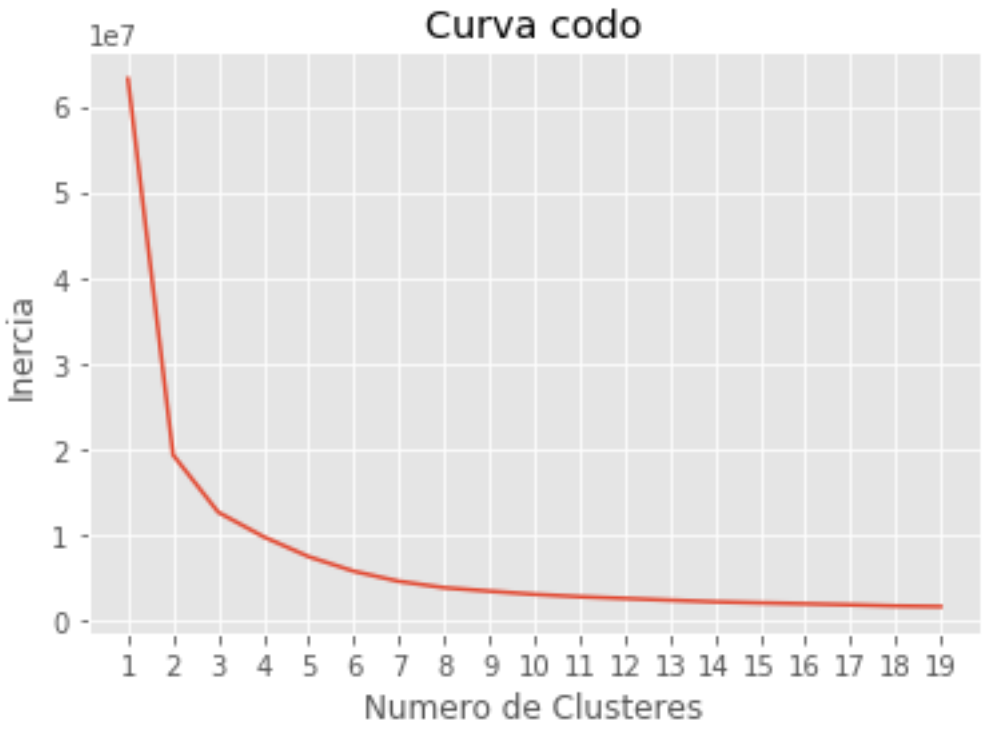

Ilustración 5. Ejemplo de curva codo.

Siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su centroide:

$$Inercia = \sum_{i=0}^{n} ||x_i - \mu||^2$$

El número óptimo de clústeres es el que proporciona el valor mínimo de inercia para un valor mínimo de clústeres, es decir, el primer clúster para el que la variación de inercia deja de ser significativa. Intuitivamente, la inercia dice cómo de lejos están los objetos dentro de un clúster del centroide, por lo tanto, se busca que estos objetos estén lo más cerca posible del centroide para un número mínimo de clústeres. Es interesante que el número de clústeres sea mínimo, ya que cuantos menos haya, más sencillo será de caracterizar el problema que se esté estudiando. Hay que tener en cuenta que este método es visual, y que el número de clústeres elegido puede

no ser el perfecto, debido al propio error humano, o a que el codo no caiga exactamente encima de un numero entero (por ejemplo 4.8 en lugar de 5) en la gráfica.

Esto se podrá observar más claramente en el apartado 4.6.2 Caracterización de los grupos en función de las variables input, en el que el dataset que consume el algoritmo varía su número de variables de un caso a otro, y el valor de K también puede variar.

2.3.3 Árboles de decisión

Los arboles de decisión son algoritmos de aprendizaje supervisado. Generan un diagrama que representa de forma secuencial condiciones y acciones. Destacan por su sencillez permitiendo que cualquier persona que no tenga grandes conocimientos en ciencias computacionales o estadística sea capaz de interpretarlos.

La estructura de los árboles de decisión es la siguiente: cuentan con un primer nodo, llamado nodo raíz, al cual se le plantea una condición booleana (verdadero o falso). Este nodo raíz se bifurca en dos caminos, y en función de la respuesta a la condición booleana planteada, se tomará un camino u otro. Cada uno de estos dos caminos o ramas conducen, respectivamente, a dos nodos diferentes. Posteriormente, estos dos nodos, realizan el mismo proceso y se vuelven a bifurcar. Este proceso se repite sucesivamente hasta llegar a los nodos finales, los cuales clasifican el caso, como se puede ver en el ejemplo de la ilustración 6:

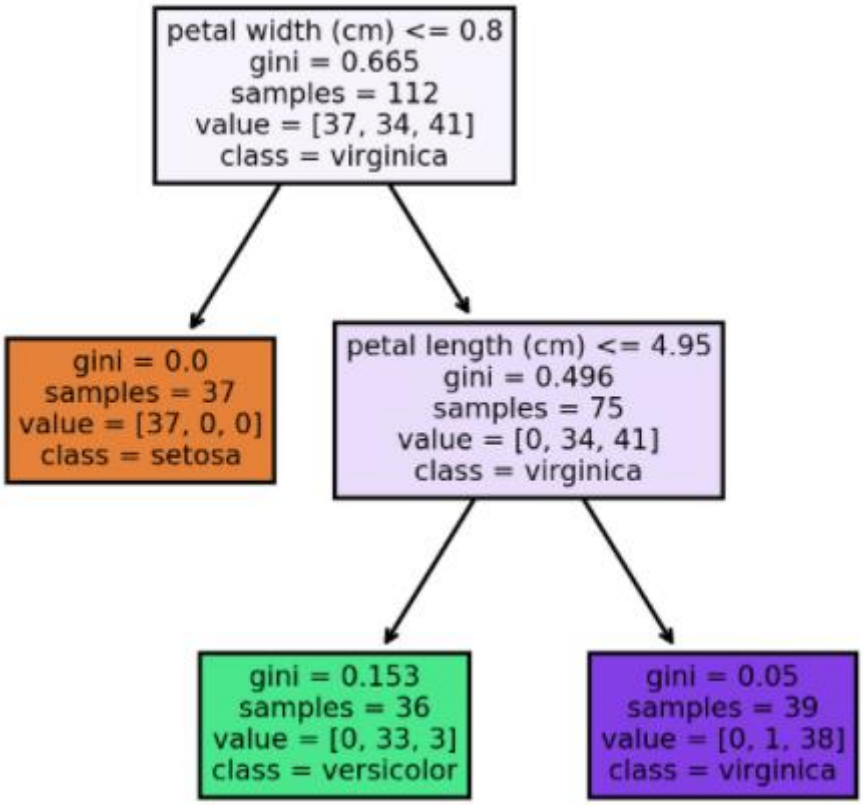

Ilustración 6. Ejemplo de árbol de decisión.

Para construir el árbol, el algoritmo va evaluando las predicciones conseguidas, y compara entre todas las combinaciones posibles para elegir la mejor. Para ello, se ayuda de las siguientes funciones: por una parte, el índice de Gini, que indica el grado en que los nodos están mezclados

una vez divididos, interesa que sea mínimo. Por otra, la ganancia de información, mediante la cual se estima la información que aporta cada característica, interesa que sea máxima [18].

El *DecisionTreeClasifier* es una función integrada en *Scikit-Learn* que consiste en un método de clasificación. Se trata de una de las técnicas supervisadas más utilizadas en *machine learning*, y está basado en árboles de decisión. Para construir arboles de decisiones existen diferentes algoritmos. El ID3, C4.5 (es el sucesor de ID3), C5.0, o CART, entre otros. *Scikit-Learn* usa una versión optimizada del algoritmo CART [19].

El algoritmo CART es el acrónimo de *Classification And Regression Trees* (Árboles de Clasificación y de Regresión) fue diseñado por Breiman et al. (1984). Con este algoritmo, se generan árboles de decisión binarios en base a las siguientes reglas:

- Las reglas basadas en los valores de las variables se seleccionan para obtener la mejor división para diferenciar las observaciones basadas en la variable dependiente
- Una vez que se selecciona una regla y se divide un nodo en dos, se aplica el mismo proceso a cada nodo "secundario" (es decir, es un procedimiento recursivo)
- La división se detiene cuando CART detecta que no se puede realizar más ganancia o se cumplen algunas reglas de parada preestablecidas.

Cada rama del árbol termina en un nodo final, con lo que cada observación cae en exactamente un nodo final, y cada nodo final está definido de forma única por un conjunto de reglas [20].

En el caso de interés de nuestro proyecto, uno de los objetivos a alcanzar es ser capaces de, dadas unas condiciones meteorológicas concretas que rodean a un incidente concreto, construir un modelo de clasificación binario (mortal o no mortal), y ser capaces de, a partir de este modelo, generar predicciones. De esta manera, el árbol se irá creando en base a los atributos meteorológicos asociados a cada incidente, y clasificará en cada caso con una etiqueta de 1 o 0 (mortal o no mortal). Así, cuando se le pase un caso nuevo que clasificar, este irá atravesando los sucesivos nodos en base a las características del mismo, y finalmente decidirá si es un incidente mortal o no.

Esta técnica presenta las siguientes ventajas:

- Permite analizar todas las posibles consecuencias antes de tomar una decisión.
- Clasifica la información con un bajo coste computacional.
- Es capaz de cuantificar el coste de un resultado y la probabilidad de que suceda.
- Toma las mejores decisiones en base a la información existente.

3 Herramientas

3.1 Python

Python es un lenguaje de programación de alto nivel, interpretado, flexible y con una sintaxis clara y concisa. Se trata de un lenguaje multiparadigma ya que soporta más de un paradigma de programación, a saber, orientación a objetos, programación imperativa y programación funcional. Es un lenguaje de propósito general y, a parte, también se puede utilizar para desarrollar páginas web. *Python* se desarrolla como un proyecto open source, es decir, distribuido y desarrollado libremente [21].

3.1.1 Elementos del lenguaje

A diferencia de la mayoría de los lenguajes de programación, *Python* nos provee de reglas de estilos, a fin de poder escribir código fuente más legible y de manera estandarizada [22].

- Variables: Las variables se definen de forma dinámica, lo que significa que no se tiene que especificar cuál es su tipo de antemano y puede tomar distintos valores en otro momento, incluso de un tipo diferente al que tenía previamente. Se usa el símbolo = para asignar valores.
- Operadores Aritméticos: Los operadores son símbolos reservados por el lenguaje que se utilizan para realizar operaciones sobre uno, dos o más elementos llamados operandos. Los operandos pueden ser números, variables, el valor devuelto por una expresión, o el valor devuelto por una función.
- Comentarios. Los comentarios en *Python* se utilizan para explicar el código. Estos comentarios son ignorados por la computadora cuando ejecutan el código, es decir, no es código fuente.
- Tuplas: Una tupla es un conjunto ordenado e inalterable de elementos que pueden ser del mismo o diferente tipo. Las tuplas se representan escribiendo los elementos entre paréntesis y separados por comas.
- Listas: Una lista en *Python* es una estructura de datos formada por una secuencia ordenada de objetos. El concepto es similar al de una tupla, con la diferencia de que permite modificar los datos que esta contiene.
- Diccionarios. Los diccionarios en *Python* son un tipo de estructuras de datos que permite almacenar un conjunto no ordenado de pares clave-valor, siendo las claves únicas dentro de un mismo diccionario. Esto significa que en un diccionario no existen dos elementos con una misma clave.

3.1.2 Módulos, paquetes y namespace

- Módulos: Un módulo es un objeto de *Python* con atributos con nombres arbitrarios que puede enlazar y hacer referencia. En definitiva, un módulo es un archivo con extensión .py., el cual puede definir funciones, clases y variables, también puede contener código ejecutable.
- Paquetes: Un paquete es una carpeta que contiene uno o varios módulos. Para que *Python* entienda que es un paquete y no una simple carpeta, debe contener siempre un archivo *__init__.py*. Este archivo, es utilizado para inicializar paquetes de *Python*.
- Namespace: En *Python*, un *namespace*, es el nombre que se ha indicado tras la palabra *import*, es decir la ruta del módulo. Para acceder desde el módulo donde se realizó la importación a cualquier elemento del módulo importado, se realiza mediante el *namespace*, seguido de un punto y el nombre del elemento que se desee obtener.:

import paquete.subpaquete.modulo1.CONSTANTE_1

También es posible asignar a los *namespace* un alias. Esto se realiza durante la importación, asignando el alias con el cual nos referiremos en el futuro a ese *namespace* tras un *as* [22]:

import paquete.subpaquete.modulo1 as psm

3.2 Bibliotecas

A continuación, se presenta una breve descripción de las principales bibliotecas utilizadas para desarrollar el proyecto.

3.2.1 Numpy

NumPy (Numeric Python) es una biblioteca de *Python* que da soporte para crear vectores y matrices multidimensionales, junto con una gran colección de funciones matemáticas para operar con ellas [23].

3.2.2 Matplotlib

Matplotlib es una biblioteca que permite generar gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación *Python* y su extensión matemática *NumPy*.

3.2.3 Pandas

Pandas es una biblioteca escrita como extensión de *NumPy* para manipulación y análisis de datos para el lenguaje de programación *Python*. En particular, ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales [25]:

- Series: Son arrays unidimensionales indexados capaces de contener datos de cualquier tipo (numéricos, alfanuméricos, booleanos, etc.).
- Dataframes: El *dataframe* permite almacenar y manipular datos tabulados en columnas y filas, al estilo de las tablas de bases de datos relacionales.

3.2.4 Scikit-Learn

Scikit-Learn es una biblioteca de *Python* que contiene una gran cantidad de modelos que evitan programar algoritmos desde cero [19].

3.3 Jupyter

El *Jupyter Notebook* es un entorno web interactivo, que permite generar y compartir documentos que contienen código, ecuaciones, visualizaciones y texto. Estos documentos generados se denominan *IPython Notebook* y poseen una extensión *.ipynb*. A dicho entorno se accede directamente desde un navegador, habiendo creado previamente el entorno virtual de desarrollo en local.

Suele emplearse para la limpieza y transformación de datos, tareas de simulación numérica, modelado estadístico y aprendizaje automático, entre otros [26].

4 Ejemplo de aplicación: Análisis de datos aplicado a incidentes acuáticos.

4.1 Introducción

Al inicio del proyecto se mantuvieron reuniones con los diferentes intervinientes para llevar a cabo un análisis del problema, en este caso el profesor Diego García Saiz, y el responsable de la Escuela Segoviana de Socorrismo, Luis Miguel Pascual Gómez. Una vez enfocado el objeto de estudio, se plantean más reuniones con el objetivo de establecer los requisitos funcionales y técnicos, siendo este el momento de decidir el tipo de herramientas a utilizar, las técnicas, y estudiar los diferentes métodos de abordarlo.

Posteriormente se obtuvieron, por un lado, series temporales de datos relativos a condiciones meteorológicas tomados en diferentes puntos del territorio peninsular español, y, por otro lado, datos relativos a incidentes acuáticos también ocurridos por toda la geografía española a lo largo del intervalo temporal 2013-2019, ambos inclusive. Como se explicará en detalle más adelante, en el caso de los datos meteorológicos la información se descargó de una web que almacena registros históricos de diferentes puntos de la geografía española, y en el caso de los incidentes acuáticos, se realizaron descargas de la base de datos que lleva a cabo su registro.

Los datos meteorológicos contienen información relativa a fecha, ID y nombre de la estación, latitud y longitud, temperaturas media, máxima y mínima del día, precipitaciones, presión atmosférica, dirección y velocidad del viento, nubosidad, profundidad de la nieve e insolación.

En el caso de los incidentes acusativos, la información extraída inicialmente contiene datos relativos a fecha y hora, ID del incidente, ID de la persona (en un incidente pueden verse involucradas n personas, localidad, provincia, comunidad autónoma, latitud y longitud del incidente, sexo, edad, nacionalidad, origen, método de extracción, titular de la noticia donde se recogió el incidente, causa del incidente, tipo de ahogamiento, factores externos, tipo de intervención, pronostico, riesgos agregados al incidente, si hubo reanimación o no, si el lugar tiene o no vigilancia de socorristas, actividad que se desarrollaba al producirse el incidente, y como se detectó el incidente.

Es importante aclarar que a la hora de llevar a cabo el análisis de datos no se van a utilizar todos estos parámetros, ya sea porque están poco informados (la columna posee un gran número de registros sin información, es decir, vacíos), y bien porque debido al tipo de análisis que se va a realizar, carecen de interés.

Posteriormente se realizará un preprocesado a los datos con el objetivo de unificar formatos, eliminar caracteres extraños, unión de ficheros de series temporales en uno solo, etc. De esta manera los datos brutos quedarán listos para poder trabajar con ellos.

Una vez hecho esto, y teniendo claro el problema y el tipo de análisis que queremos llevar a cabo, los cuales se expondrán más adelante en el apartado 4.2 Objetivo, a la información que hemos preprocesado hay que darle forma. Este proceso se conoce comúnmente enriquecimiento de la información. Esto se llevará a cabo mediante una ETL (acrónimo de Extract-Transform-Load) desarrollada en unos cuantos *notebooks* de *Jupyter* y utilizando para ello *Python*. A lo largo de los scripts que componen esta ETL unificaremos la información procedente de los diferentes datasources, se eliminarán variables que no resultan útiles, se calcularán otras a partir de los datos disponibles que son necesarias, se filtrará información para

evitar errores posteriores y se sustituirán campos descriptivos por campos de códigos, entre otros.

En el momento que la información está lista, se comienzan a aplicar las técnicas de análisis estadístico y *machine learning*, con el objetivo de descubrir patrones, comprobar si existen correlaciones entre las condiciones meteorológicas en el instante y lugar que se produjo el incidente acuático, y generar modelos de predicción.

4.2 Objetivo

En el contexto de este proyecto, el problema como tal son los incidentes acuáticos en sí, y el objetivo es aplicar técnicas de *data science* para:

- Realizar un estudio descriptivo de las diferentes variables que vamos a utilizar.
- Comprobar el nivel de correlación entre las diferentes variables calculando para ello el coeficiente de correlación de Spearman.
- Estudiar la significancia de estas variables aplicando T-Test.
- Caracterizar los incidentes aplicando técnicas de *machine learning* como el algoritmo *K-Means.*
- Predecir la mortalidad de un incidente con unas características dadas utilizando para ello árboles de decisión.
- Poner de manifiesto la transversalidad de estas técnicas independientemente de la naturaleza del dato.

4.3 Datos brutos

Se dispone de dos fuentes principales de información:

- Datos meteorológicos: http://meteomanz.com/ proporciona datos meteorológicos observados de lugares de todo el mundo obtenidos de los mensajes SYNOP y BUFR emitidos por estaciones meteorológicas oficiales. La base de datos contiene información desde el año 2000 hasta la actualidad. Están tomados de los mensajes alfanuméricos SYNOP terrestres (surface synoptic observations) y de los datos binarios BUFR (Binary Universal Form for the Representation of meteorological data). Ambos formatos están recogidos por la Organización Meteorológica Mundial (OMM) y son utilizados por casi todos los servicios meteorológicos del mundo.
- Datos sobre incidentes acuáticos: Datos procedentes del proyecto de investigación *Ahogamiento en España* desarrollado por la Escuela Segoviana de Socorrismo (ESS) y la Asociación Española de Técnicos en Salvamento Acuático y Socorrismo (AETSAS).

  Es importante señalar el esfuerzo llevado a cabo por ambas organizaciones de socorrismo, ya que estos datos se recogen de manera manual desde 2013 en base a noticias de prensa publicadas en medios escritos y digitales, redes sociales, y comunicaciones de los servicios de emergencias.

  De forma periódica se emiten informes, y estos se pueden consultar en http://www.ahogamiento.com/.

4.3.1 Obtención

Datos meteorológicos

Los datos meteorológicos se han obtenido realizando descargas de la página web http://meteomanz.com/. En este punto aparece uno de los primeros problemas: el administrador de la web, con el objetivo de no saturar su servidor solo permite realizar descargas de 2 días cada vez. Esto se traduce en que para descargar la información relativa a 1 mes habría que descargar 15 ficheros diferentes. Teniendo en cuenta que hay que seleccionar un combo de opciones en la web, introducir un código captcha, etc. para cada descarga, esto hace que el proceso sea muy ineficiente. Se contacta con el administrador y se consigue ampliar el intervalo de fechas de 2 días a 30, lo cual hace el trabajo mucho más ágil. Sabiendo que en http://www.ahogamiento.com/ la profundidad histórica de los datos es menor que en http://meteomanz.com/, con descargar la serie histórica 2013-2019 (ambos inclusive) será suficiente.

Datos sobre incidentes acuáticos

Los datos relativos a incidentes acuáticos se encuentran en tablas alojadas en una base de datos Microsoft Access administrada por http://www.ahogamiento.com/ de la cual se realizarán descargas relativas al intervalo 2013-2019 (ambos inclusive).

4.4 Análisis preliminar de los datos

4.4.1 Preprocesado de los datos

Datos meteorológicos

El primer punto a comentar es que los ficheros obtenidos de la web http://meteomanz.com/ tienen formato *.xlsx*, no siendo este el más adecuado para que sean consumidos. Para poder trabajar correctamente con estos ficheros, se transforman de *.xlsx* a *.csv* separados por *pipes* ("|"). También se codifican a UTF8 y se configura el fin de línea WINDOWS. Una vez que todos los ficheros están con el formato y codificación adecuados, se unen en un solo fichero. El siguiente paso será hacer una primera limpieza de los datos para tratar de tener el menor número de errores posible, a saber, eliminar acentos, caracteres extraños tales como *Ñ*, tildes y convertir los nombres de las estaciones a mayúsculas. A este fichero se le llamará *historico_estaciones.csv.*

Una vez hecho lo anterior, se construye otro fichero que contenga la relación entre el identificador único de la estación, y el nombre de la estación. Esto se consigue acudiendo al código fuente de la web http://meteomanz.com/ en el que podemos encontrar un listado de estos *Id's* con la estación meteorológica asociada. Se llevan a cabo las mismas modificaciones que al fichero anterior, tales como la codificación y el fin de línea, eliminación de caracteres susceptibles de ser extraños, y convertir el descriptivo del nombre de las estaciones a mayúsculas. A este fichero se le llamará *estaciones_codigo_fuente_def.csv.*

Por último, para en un futuro ser capaces de relacionar la posición geográfica de los incidentes con la posición geográfica de las estaciones meteorológicas, es necesario conocer las coordenadas de las mismas. En este caso, http://www.ahogamiento.com/ nos proporciona un fichero con la relación de estaciones meteorológicas y sus coordenadas que podremos utilizar. A este fichero se le llamará *OMM_Estaciones.csv*, y también se le hará un tratamiento de la codificación, final de línea, mayúsculas, tildes, etc. similar a los anteriores. En este caso, además,

se eliminará una serie de duplicados contenidos para 3 estaciones. El motivo de la existencia de estos duplicados se debe a que la estación asociada al id, en cada caso, ha tenido dos localizaciones diferentes. Lo que se hace será eliminar estos duplicados del fichero *OMM_Estaciones.csv* y se conservará la que aparece en el código fuente de la web, y que a su vez es la que está vigente. El nuevo fichero sin duplicados lo llamaremos *OMM_Estaciones_NoDupli.csv*. Es importante señalar que las coordenadas contenidas en este fichero tienen formato sexagesimal.

En resumen, como orígenes de datos relativos a estaciones meteorológicas, tendremos los siguientes ficheros:

- *historico_estaciones.csv*: Información a nivel de estación-datos meteorológicos.
- *estaciones_codigo_fuente_def.csv*: Información ID_ESTACION-NOM_ESTACION
- *OMM_Estaciones_NoDupli.csv*: Información ID_ESTACION-NOM_ESTACION-POS_ESTACION

Datos sobre incidentes acuáticos

Como ha sido el caso de los ficheros asociados a la meteorología, en primer lugar, se ha transformado el formato de *.xlsx* a *.csv* separados por pipes ("|"). También se codifican a UTF8 y se configura el fin de línea WINDOWS. Al igual que con el resto de ficheros, para intentar tener los datos de la manera más homogénea posible, se han eliminado los caracteres susceptibles de ser problemáticos (por ejemplo, la *Ñ*), todos los campos descriptivos se han transformado a mayúsculas, tildes, etc. A este histórico de incidentes se le llamará *Datos_2019-2013.csv*, y principalmente contiene información a nivel de persona (atributos de la persona tales como edad, sexo, nacionalidad, etc) y de incidente (atributos del incidente tales como lugar del mismo, fecha, y demás). La gran mayoría de estos atributos que aparecen en este fichero son descriptivos, lo cual a la hora de realizar cualquier tipo de análisis no resulta óptimo para trabajar. En este caso, en la BBDD Access tenemos disponibles tablas de códigos para la mayoría de atributos que caracterizan el incidente, y de esta manera a la hora de realizar un análisis podremos utilizar valores numéricos (códigos) en lugar de valores alfanuméricos (descriptivos).

En resumen, como orígenes de datos relativos a incidentes acuáticos, tendremos los siguientes ficheros:

- Datos_2019-2013.csv
- aux_actividad.csv
- aux_causa.csv
- aux_CCAA.csv
- aux_deteccion.csv
- aux_factores.csv
- aux_intervencion.csv
- aux_localizacion.csv
- aux_origen.csv
- aux_pronostico.csv
- aux_provincias.csv
- aux_reanimacion.csv
- aux_riesgo.csv
- aux_tipo.csv

- aux_vigilancia.csv

4.4.2 Enriquecimiento

A partir de aquí se explica script a script la ETL que se ha desarrollado para llevar a cabo el proceso de transformación de los datos origen con el fin de tener un juego de datos óptimo para podérsele aplicar técnicas más avanzadas. A continuación, se realiza una breve descripción del funcionamiento de cada uno de los scripts que han formado parte de la ETL.

LeftJoin_Meteo_Estacion_1.1.ipynb

En este script se va a unificar en un único fichero la triada ID_Estacion-Meteo-Coordenadas. Para conseguirlo se ha acudido al código fuente de la página web de la que obtenemos los datos de meteo, y se ha montado un fichero con el listado de las estaciones con su id asociado, montando el fichero *estaciones_codigo_fuente_def.csv*.

Posteriormente se ha generado un fichero que contiene ID, Estación y sus coordenadas. Para ello se ha realizado un cruce (LEFT Join) por el campo ID entre *OMM_Estaciones_NoDupli.csv* y *estacionces_codigo_fuente_def.csv.* De esta manera tenemos en un mismo fichero todas las estaciones que de las que nos ofrece información la web, con sus coordenadas asociadas. Este fichero se llamará *estaciones_coordenadas_2.csv*.

Una vez hecho esto, se realiza otro LEFT Join entre *historico_estaciones.csv* y *estaciones_coordenadas_2.csv* para incorporar los campos ID_Estacion y coordenadas, y de esta manera tener en el mismo registro los datos de ID_Estacion, coordenadas, y todos los datos meteorológicos recogidos por dicha estación. Al fichero resultante se le nombra *historico_estaciones_coordenadas_2.csv*.

- Inputs:
  - estaciones_codigo_fuente_def.csv
  - OMM_Estaciones NoDupli.csv.csv
  - historico_estaciones.csv

- Outputs:
  - estaciones_coordenadas_2.csv
  - historico_estaciones_coordenadas_2.csv

Decimal_coordinates_1.0.ipynb

Ahora el objetivo principal es transformar el formato en el que se reciben los datos de las columnas latitud y longitud. Los datos referentes a coordenadas se están recibiendo en formato sexagesimal. El formato que se tiene antes de realizar la transformación de sexagesimal a decimal para un par de coordenadas (latitud, longitud) es: *43-33-36N, 005-42-03W.*

Desde el punto de vista de procesamiento de los datos, este formato no es en absoluto óptimo ya que al fin y al cabo es una mezcla de caracteres alfanuméricos que una función/programa de cualquier lenguaje de programación va a ser incapaz de interpretar, realizar cálculos, etc. Lo que se hace en este caso es, para los dos ficheros input, transformar a formato decimal los datos

correspondientes a coordenadas para que, llegado el momento, sean más utilizables. El formato que se tiene para un par de coordenadas (latitud, longitud) tras haber transformado el formato de sexagesimal a decimal es: *43.560000, 5.700833*.

Se generan entonces 2 ficheros de salida, homólogos a los ficheros input, pero con un formato más adecuado.

- Inputs:
  - estaciones_coordenadas_2.csv
  - historico_estaciones_coordenadas_2.csv

- Outputs:
  - estaciones_coordenadas_decimal.csv
  - historico_estaciones_coordenadas_decimal.csv

asigna_estacion_incidentes_1.1.ipynb

Una vez se tiene lista la parte referente a los datos recogidos por las estaciones meteorológicas, es necesario asignar a cada incidente los datos recogidos por la estación más cercana al mismo en el día que se produjo.

Antes de comenzar con esto, se establecen una serie de restricciones para no empobrecer los juegos de datos.

Durante el preprocesado de los datos obtenidos por las estaciones meteorológicas se ha podido observar que, para numerosas estaciones y fechas diferentes, ciertos valores de los parámetros recogidos por las mismas están sin informar, si no todos (esto último es una situación algo más excepcional). Esto puede deberse a averías parciales o completas de las estaciones o periodos de mantenimiento, entre otros. En definitiva, que, para una fecha dada, alguno (o todos) de los dispositivos encargados de registrar los diferentes parámetros meteorológicos no se encontraban disponibles. Si esto no se tuviera en cuenta, podría llegar a darse el caso extremo de asociar a un incidente los datos meteorológicos de una estación (la más cercana) y que estos fuesen datos nulos/vacíos.

Dicho esto, las restricciones que se han convenido establecer son las siguientes:

- Temperatura media, máxima y mínimas siempre informadas.
- Precipitaciones siempre informadas.
- Presión atmosférica siempre informada.
- Velocidad del viento siempre informada.

Una vez aplicadas estas restricciones, se procede a asignar a cada incidente la estación meteorológica más cercana. El proceso, a grandes rasgos, es el siguiente:

1. Se transforman las coordenadas de grados decimales a radianes
2. Se cruzan ambos datasets por fecha. Con esto lo que conseguimos es asignar a cada incidente *n* registros meteorológicos, donde *n* es el número de estaciones con registros meteorológicos para la fecha del incidente. Evidentemente tendremos duplicados a nivel de incidente, pero con diferentes valores de estaciones meteorológicas.

3. Ahora se define la función *calcular_distancia*, la cual, dados dos pares de coordenadas, es capaz de calcular la distancia entre esos dos puntos.
4. Una vez hecho esto, para cada incidente, se inicializa una variable *distancia* que recoge el valor de la distancia calculado por la función anteriormente descrita. Calculada esta distancia, lo que se hace es definir una clave primaria, ordenar por distancia esta clave primaria de forma ascendente, y quedarnos con el primer registro.

De esta manera, para un incidente dado, se tiene asociada la estación meteorológica más cercana, teniendo en cuenta las restricciones anteriormente descritas.

Una vez conseguido el objetivo, se revisan los datos, se hacen una serie de operaciones de modificación de formato, tales como eliminar caracteres alfanuméricos de columnas como la de presiones (1035.4 HPa ---> 1035.4), dirección del viento (340º(N) ---> 340), etc. para que estos pasen por numéricos y sea más sencillo su análisis.

Finalmente se genera 1 fichero de salida en el que ya disponemos de todos los datos del incidente acuático con los datos meteorológicos de la estación más cercana que cumpla las restricciones establecidas.

- Inputs:
  - Datos_2019-2013.csv
  - historico_estaciones_coordenadas_decimal.csv

- Outputs:
  - historico_incidentes_estaciones_meteo.csv

asigna_codigos_incidentes_1.0.ipynb

De cara a acabar con el preprocesado de los datos para tenerlos listos para ser analizados, lo que se hace en este script es sustituir todos los valores de las columnas de atributos descriptivos que caracterizan el incidente acuático por sus valores de códigos asociados. De esta manera se podrán aplicar tales como clustering, clasificación, etc. más fácilmente.

Se genera un fichero homólogo al de entrada.

- Inputs:
  - historico_incidentes_estaciones_meteo.csv
  - aux_actividad.csv
  - aux_causa.csv
  - aux_CCAA.csv
  - aux_deteccion.csv
  - aux_factores.csv
  - aux_intervencion.csv
  - aux_localizacion.csv
  - aux_origen.csv
  - aux_pronostico.csv
  - aux_provincias.csv
  - aux_reanimacion.csv

- aux_riesgo.csv
- aux_tipo.csv
- aux_vigilancia.csv

- Outputs:
  - historico_inc_est_meteo_codigos.csv

4.5 Análisis estadístico

4.5.1 Descriptivo

En primera aproximación se hace un estudio descriptivo de los atributos meteorológicos asociados a cada incidente sin discernir entre incidentes mortales/no mortales, incidentes costeros/no costeros, etc. Posteriormente se realiza esta distinción y no se aprecian diferencias significativas. Cabe mencionar que se ha partido del dataset sin datos asociados a aquellos incidentes ocurridos en las Islas Canarias, precisamente, porque la página web http://meteomanz.com/ no dispone de información meteorológica del archipiélago.

Incidentes mortales y no mortales

Mes

Como se puede apreciar en la ilustración 7 presentada a continuación, la mayoría de incidentes se producen en meses veraniegos. Desde luego es obvio que durante estos meses es cuando se produce la mayor afluencia a zonas donde practicar actividades relacionadas con el agua, con lo cual, estos datos tienen todo el sentido.

Estadísticas

```
count    5468.000000
mean        6.974031
median      7.000000
std         2.483320
min         1.000000
25%         6.000000
50%         7.000000
75%         8.000000
max        12.000000
```

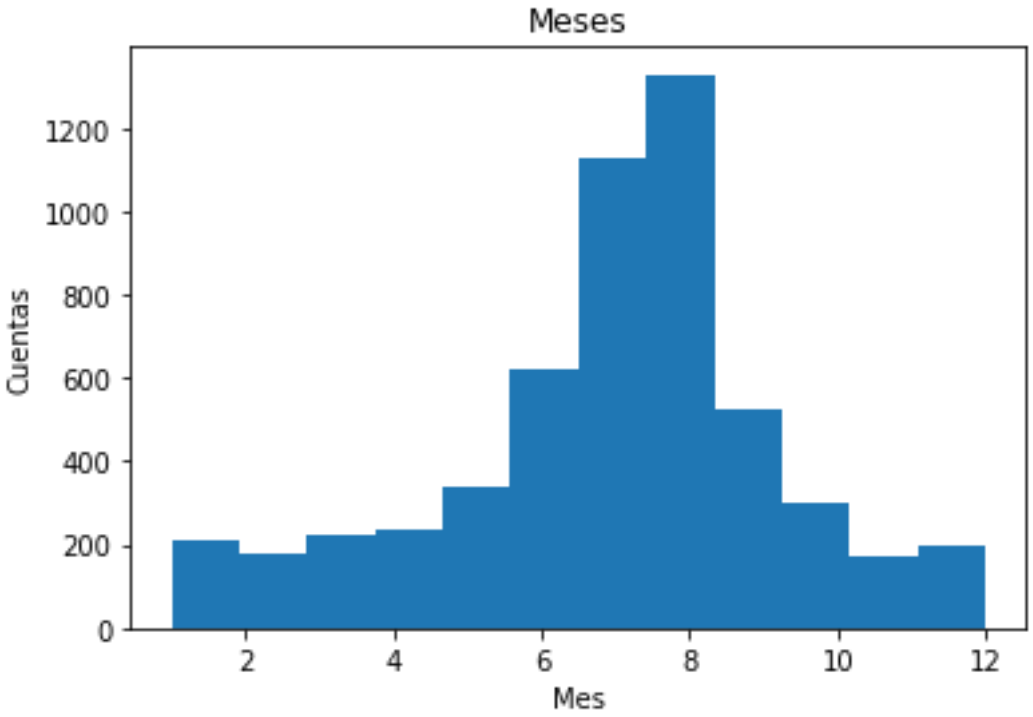

Ilustración 7. Representación del número de incidentes ocurrido a lo largo de los meses.

Temperatura Media

Se puede observar en la ilustración 8 que el máximo de cuentas se obtiene en la zona de temperaturas que podrían interpretarse como veraniegas.

Estadísticas

```
count    5468.000000
mean       20.689100
median     21.800000
std         6.356178
min         0.600000
25%        16.600000
50%        21.800000
75%        25.800000
max        38.000000
```

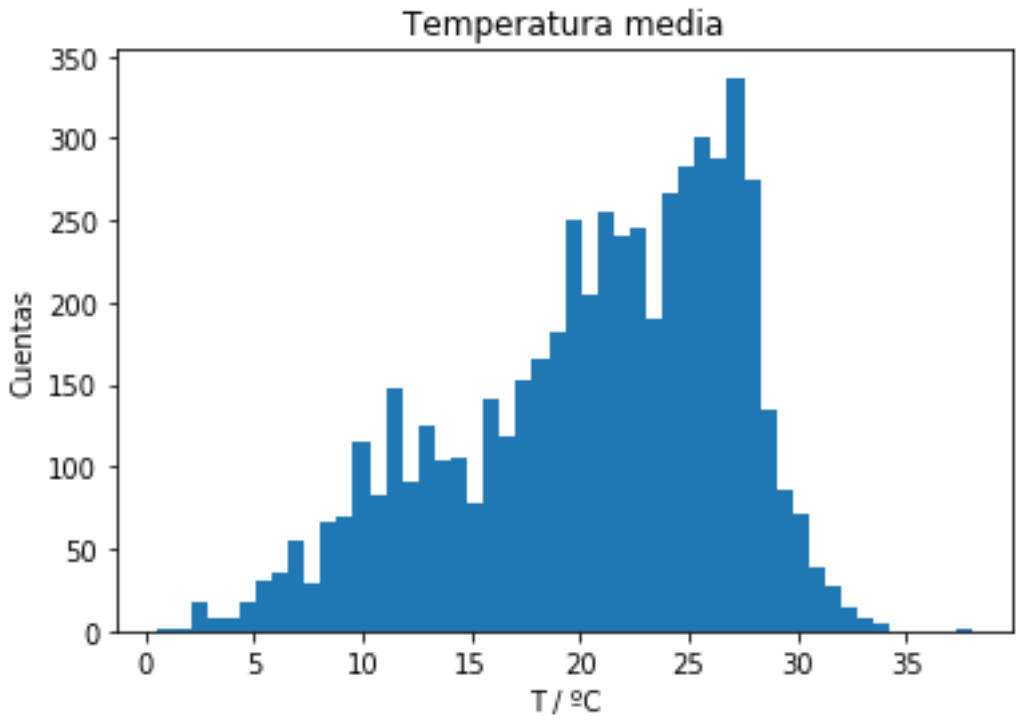

Ilustración 8. Representación del número de incidentes ocurridos para cada valor de temperatura media.

Precipitaciones

En la ilustración 9 se observa que el grueso de los incidentes se produce en condiciones principalmente sin precipitaciones. Esto es consistente con que la mayoría de incidentes se produzcan en los meses veraniegos.

Estadísticas

```
count    5468.000000
mean        2.496928
median      0.000000
std        10.590294
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max       171.600000
```

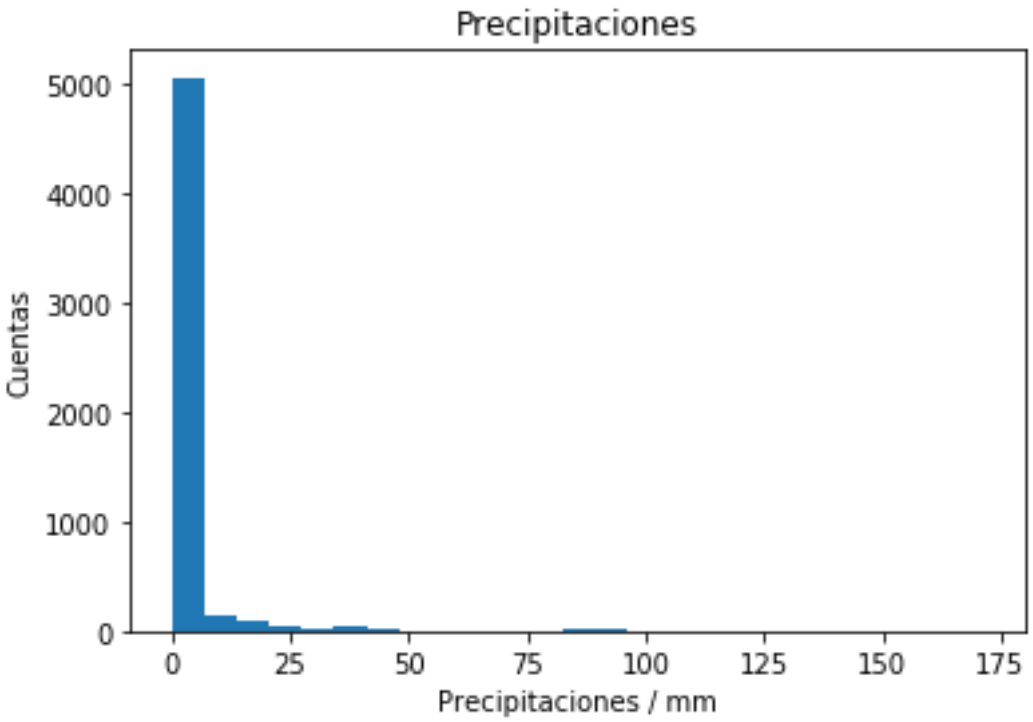

Ilustración 9. Representación del número de incidentes ocurridos para cada valor de precipitaciones.

Presión

En el caso de la presión se puede ver en la ilustración 10 que aparecen 2 outliers en torno a 850 HPa y 925 HPa. Estos se deben a que ciertas estaciones meteorológicas no se encuentran a nivel del mar, y los registros que se tienen de presión para estas estaciones son alturas geopotenciales. Estas alturas geopotenciales se expresan en *yyyym(xxx HPa)*, esto es, se da una presión a una altura dada. Por el motivo que fuere, en los casos que se utiliza esta forma de expresar la presión (altura geopotencial), siempre se expresa para 850 y 925 HPa.

Dejando de lado este detalle se observa que la presión se distribuye de forma bastante homogénea entorno al centro del pico, a unos 1015-1020 HPa imitando una distribución normal.

Estadísticas

```
count    5468.000000
mean     1009.365984
median   1016.100000
std        31.824045
min       850.000000
25%      1013.200000
50%      1016.100000
75%      1019.300000
max      1040.800000
```

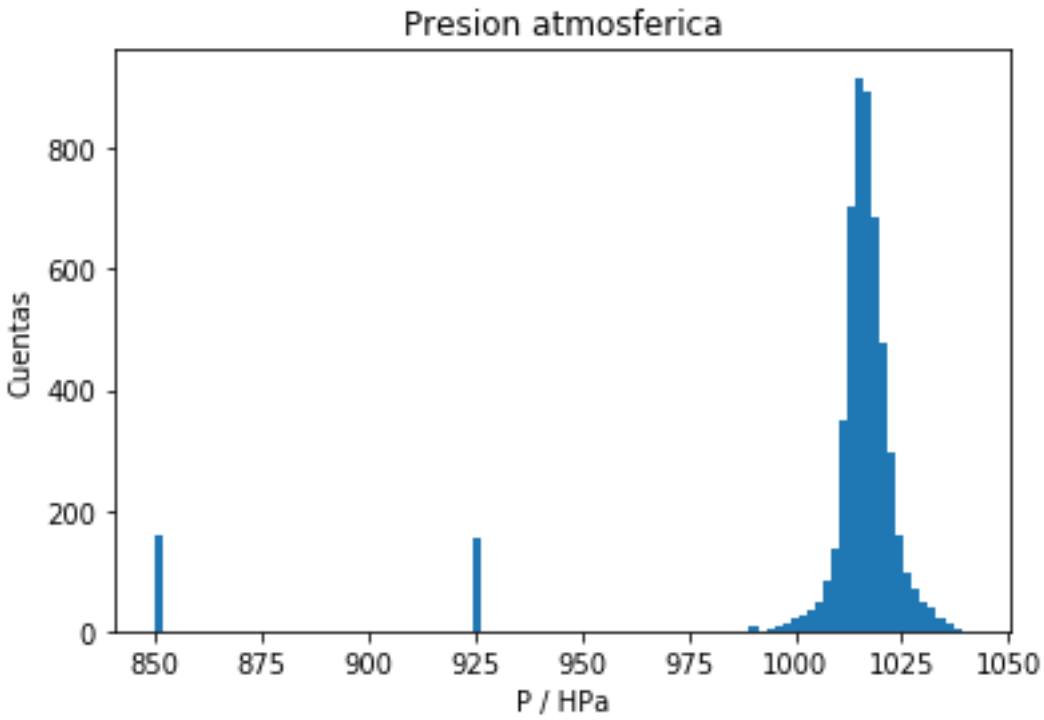

Ilustración 10. Representación del número de incidentes ocurridos para cada valor de presión atmosférica.

Dirección del viento

Para la dirección del viento se observa en la ilustración 11 que las cuentas se distribuyen de una forma bastante homogénea sin que destaque ningún rango de dirección, si acaso el nordeste.

Estadísticas

```
count    5463.000000
mean      170.457304
median    173.000000
std       101.839314
min         0.000000
25%        73.000000
50%       173.000000
75%       257.000000
max       360.000000
```

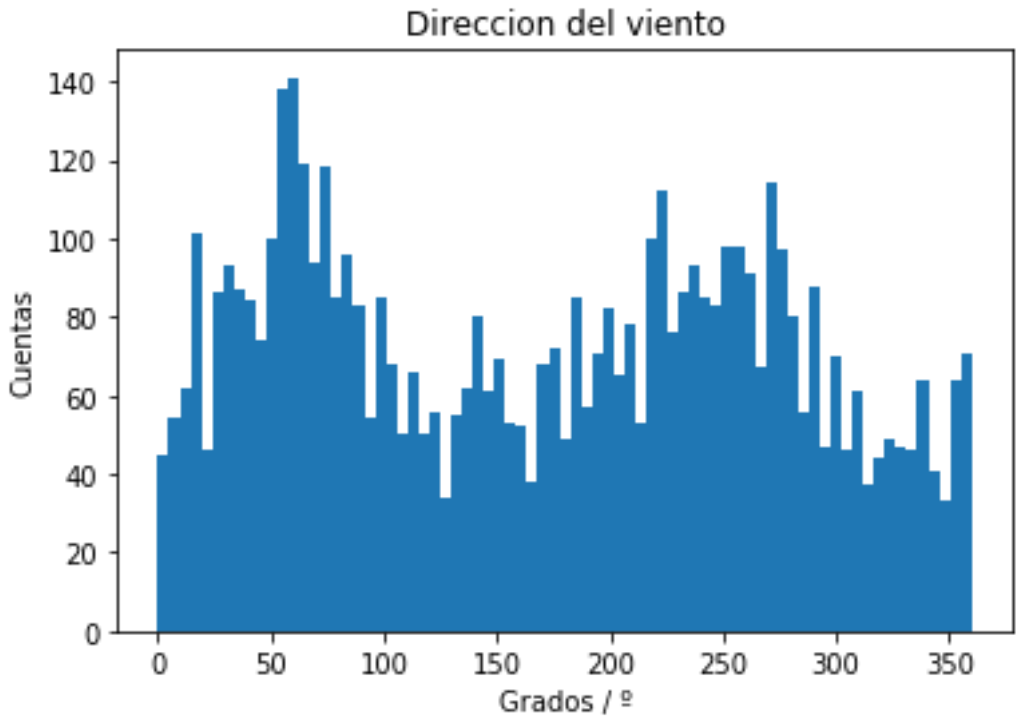

Ilustración 11. Representación del número de incidentes ocurridos para cada valor de dirección del viento.

Velocidad del viento

Se puede apreciar que la mayoría de cuentas de la ilustración 12 caen en el rango inferior a los 20 km/h de velocidad del viento. Esto indica que no es necesario que la intensidad del viento sea elevada para que se produzcan incidentes acuáticos.

Estadísticas

```
count    5468.000000
mean       11.159108
median     10.000000
std         5.993863
min         0.000000
25%         7.000000
50%        10.000000
75%        13.000000
max        52.000000
```

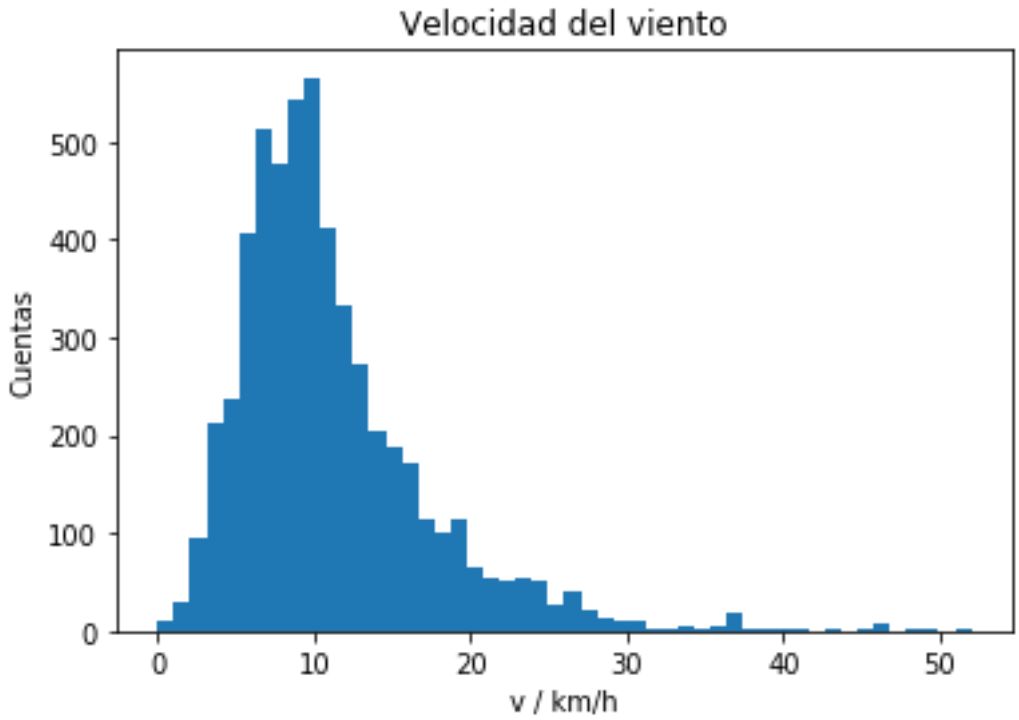

Ilustración 12. Representación del número de incidentes ocurridos para cada valor de velocidad del viento.

Nubosidad

En el caso de la nubosidad, como se puede observar en la ilustración 13, resulta curioso que las cuentas se distribuyen de forma homogénea hasta el 90% de nubosidad, y tenemos un outlier en el 100% de nubosidad.

Estadísticas

```
count    5468.000000
mean        0.549630
median      0.500000
std         0.367883
min         0.000000
25%         0.250000
50%         0.500000
75%         1.000000
max         1.000000
```

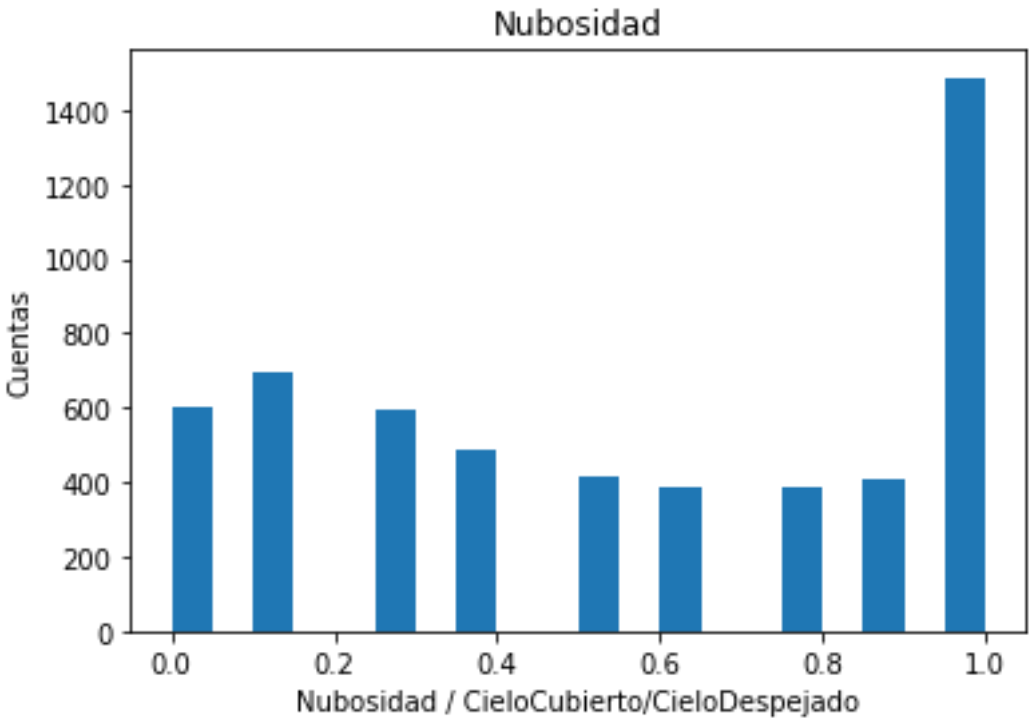

Ilustración 13. Representación del número de incidentes ocurridos para cada valor de nubosidad.

Incidentes meses de verano

Debido a las condiciones más favorables que se dan en los meses estivales (junio a septiembre, ambos inclusive), se da por hecho que la afluencia a lugares propicios a que ocurran este tipo de incidentes es mayor.

Es por esto que se ha querido comprobar si la distribución de los datos de las variables que en principio se van a ver más afectadas por la época del año en que se midan varían de forma significativa en el momento que se descartan aquellos incidentes que se salgan de este periodo.

A continuación, se revisa si en los meses de verano, que es cuando la incidencia es mayor, si existe una relación directa entre los días más calurosos y el aumento de casos. También se revisará si el hecho de ser localizaciones costeras tiene una incidencia real.

Temperaturas

En la ilustración 14 se observa que, para el caso de incidentes mortales, la distribución es más puntiaguda que en el caso de incidentes no mortales, ilustración 15, tanto para casos costeros como no costeros.

También, para ambos casos (más sutilmente en el caso de los incidentes no mortales), los picos de las distribuciones tienden a ir hacia la derecha, es decir, hacia temperaturas más altas. En ambos casos se observan outliers (principalmente costeros) pasados los 30º.

Incidentes mortales

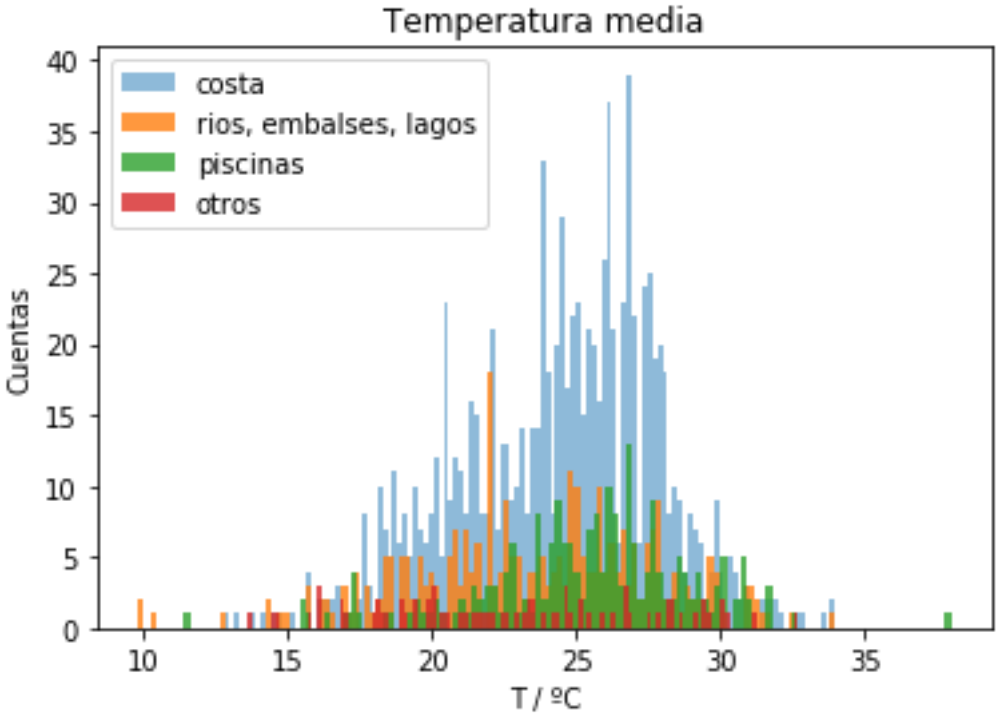

Ilustración 14 Representación del número de incidentes ocurridos para cada valor de temperatura media en el caso de incidentes mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

Incidentes no mortales

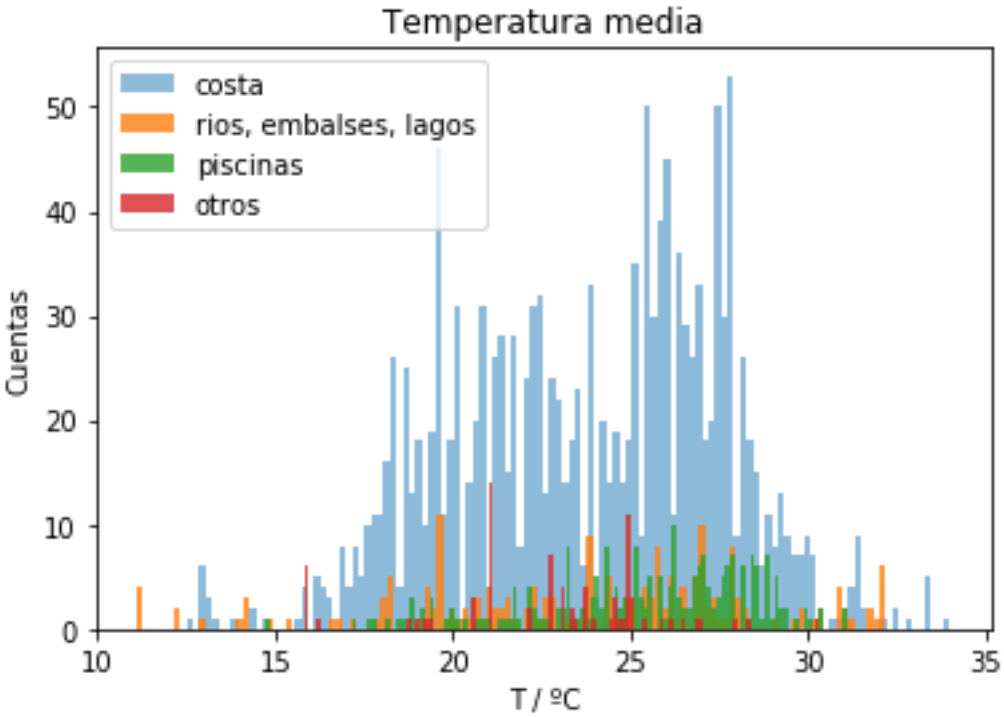

Ilustración 15 Representación del número de incidentes ocurridos para cada valor de temperatura media en el caso de incidentes no mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja

aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

Precipitaciones

A continuación, se revisa si en los meses de verano la incidencia de las precipitaciones es mayor, si existe una relación directa entre los días más lluviosos y el aumento de casos. También se revisará si el hecho de ser localizaciones costeras tiene una incidencia real.

A la vista de los resultados obtenidos, no parece que estas gráficas (ilustraciones 16 y 17) agreguen información demasiado interesante más allá de que la inmensa mayoría de incidentes se producen en días no lluviosos, y, sobre todo, en entornos costeros.

Incidentes mortales

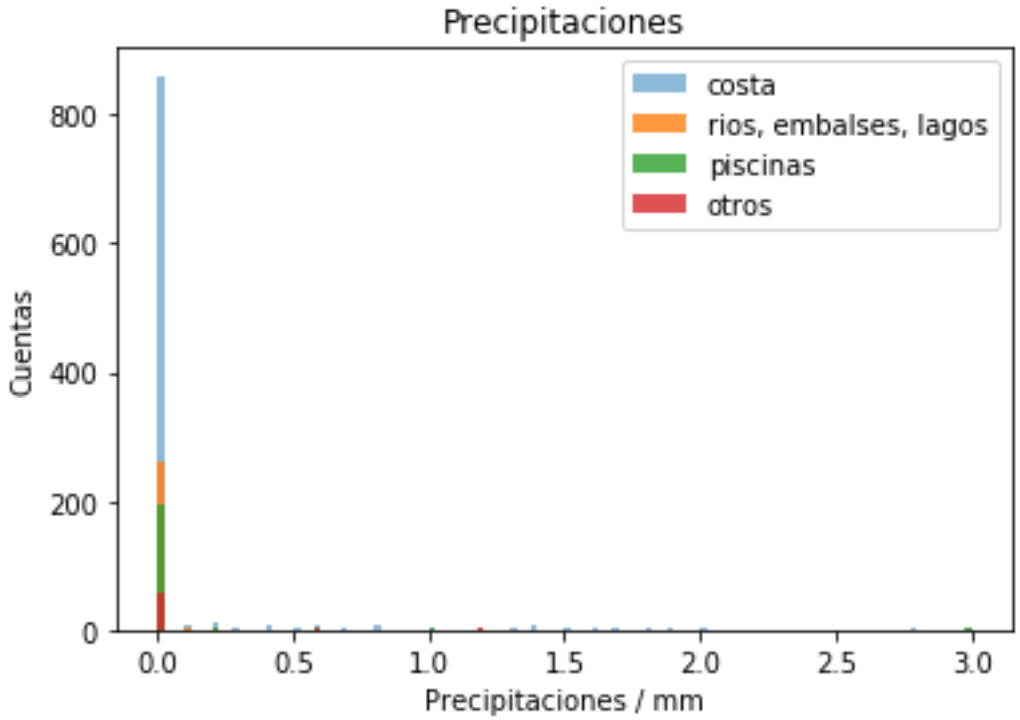

Ilustración 16. Representación del número de incidentes ocurridos para cada valor de precipitaciones en el caso de incidentes mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

Incidentes no mortales

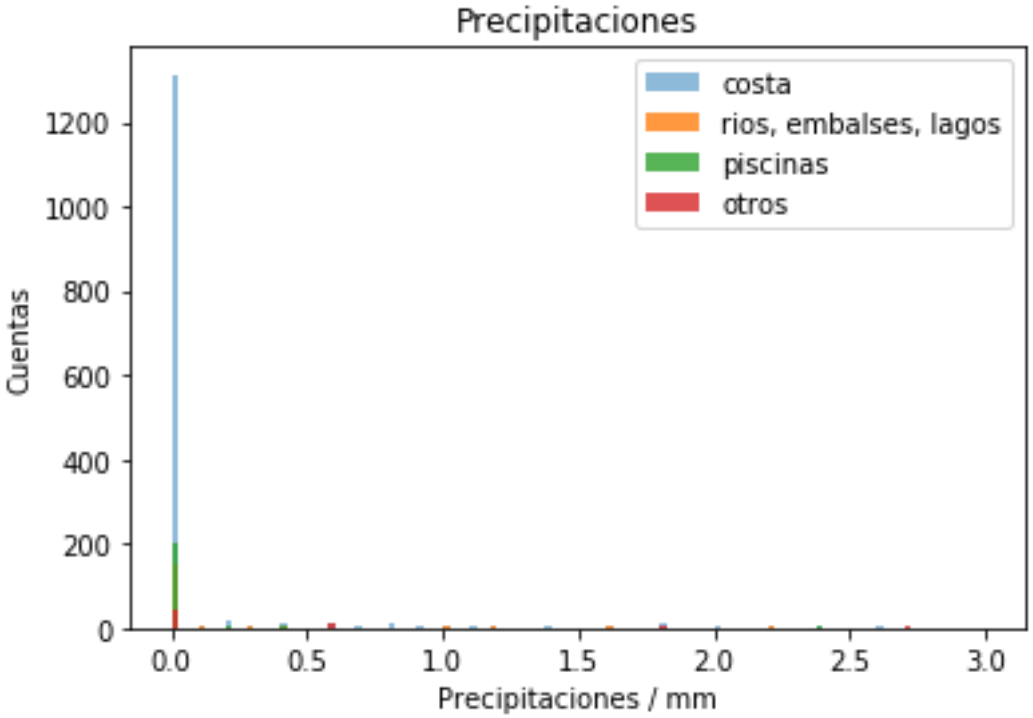

Ilustración 17. Representación del número de incidentes ocurridos para cada valor de precipitaciones en el caso de incidentes no mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

Velocidad del viento

A continuación, se revisa si en los meses de verano la incidencia del viento es mayor, diferenciando entre los diferentes entornos con el objetivo de analizar su incidencia.

A grandes rasgos ambas distribuciones son similares (ilustraciones 18 y 19). Si acaso cabría mencionar que la distribución de los fenómenos no costeros es más puntiaguda en el caso de los incidentes mortales que en los no mortales, estando el pico entorno a los 8km/h aprox. Según la escala de Beaufort, esta velocidad se corresponde con una brisa ligera. Esto haría pensar que los días con condiciones meteorológicas agradables se prestan más a la asistencia de piscinas y ríos, u otros lugares de interior donde se pueden llevar a cabo actividades acuáticas.

Por otro lado, cabe mencionar que para ambas gráficas el pico esta corrido hacia a la izquierda con respecto a la gráfica en la que no se distingue por meses. Se podría aplicar el mismo razonamiento que en el párrafo anterior.

Incidentes mortales

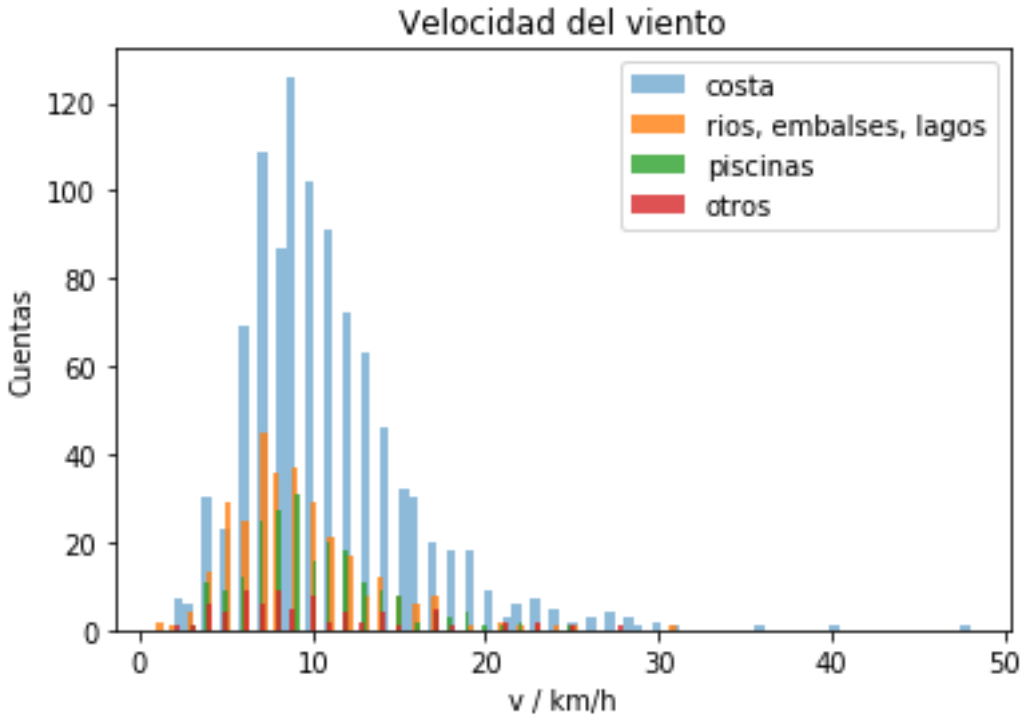

Ilustración 18. Representación del número de incidentes ocurridos para cada valor de velocidad del viento en el caso de incidentes mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

.

Incidentes no mortales

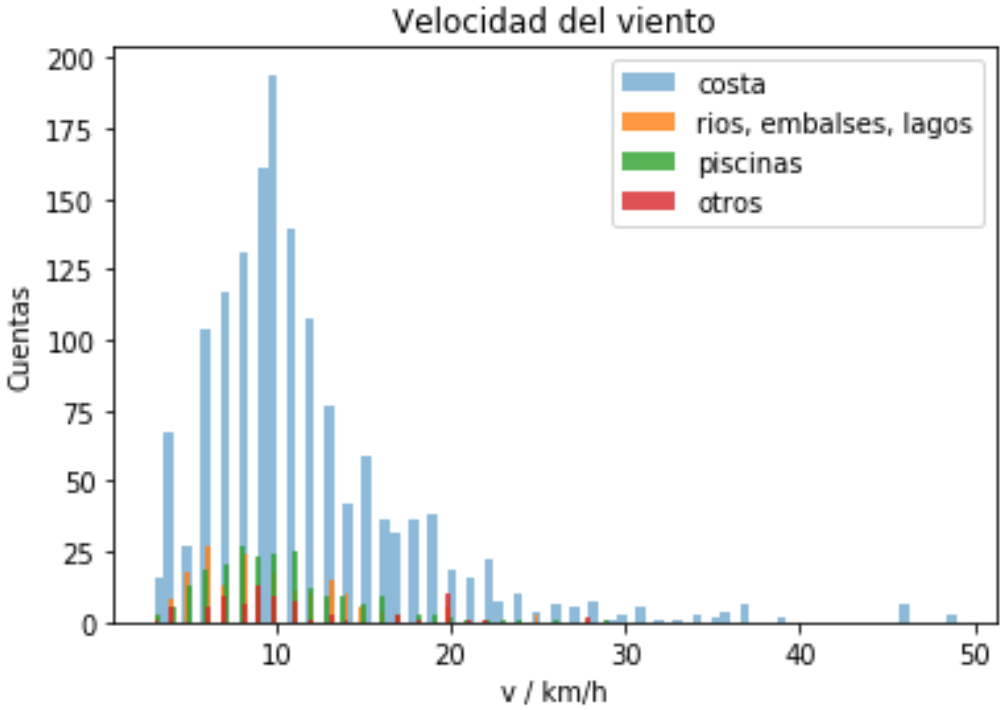

Ilustración 19. Representación del número de incidentes ocurridos para cada valor de velocidad del viento en el caso de incidentes no mortales. En azul se representan los incidentes ocurridos en localizaciones costeras, en naranja

aquellos ocurridos en aguas de interior, en verde los ocurridos en piscinas, y en rojo aquellos que no pertenecen a ninguna de las clasificaciones anteriores.

4.5.2 Correlación entre variables meteorológicas-incidentes acuáticos

En el dataset objeto del estudio tiene las siguientes características:

- Existen algunas (pocas) variables no numéricas en el dataset, a saber, mes, costa, incidente mortal, etc.
- No todas las variables se distribuyen con respecto a una distribución normal.
- No todas las variables guardan una relación perfectamente lineal entre sí.
- Los valores de las variables tampoco están basados en rangos de valores.

Por todo esto, y en virtud de lo descrito en el apartado 2.3.1, se determina que el coeficiente de correlación que mejor se ajusta al dataset es el coeficiente de correlación de Spearman.

En el anexo I se pueden encontrar los diferentes mapas de calor para trazar la correlación entre las diferentes variables que componen el dataset. Inicialmente se ha aplicado el método sin restricciones en la época del año, y posteriormente se ha vuelto a aplicar para cada una de las estaciones, con el objetivo de encontrar variaciones significativas de los valores.

Por otro lado, y para que resulte más sencillo el análisis, se han extraído, en cada caso, las 40 correlaciones más fuertes, 20 positivas en orden descendente, y 20 negativas en orden ascendente.

Los resultados obtenidos, que se pueden consultar en el *Anexo I: Correlaciones* no son concluyentes en tanto en cuanto no se encuentran correlaciones significantes a parte de las más obvias, como pudieran ser Mortalidad-Pronóstico, Precipitaciones-Nubosidad, Temperatura-Mes, etc. El resto de pares de atributos no alcanzan un valor relevante para el coeficiente de correlación. Es por esto que con el objetivo de continuar explorando los datos se recurren a otras técnicas expuestas a continuación.

4.5.3 Test de significancia

Inicialmente se ha calculado el valor del promedio y de la varianza de los pares de grupos en base a las medias de incidentes mortales y no mortales para cada atributo:

	Promedio(Mortales/NoMortales)	avg(mortales)/ avg(NoMortales)
Temperatura media	20.43753918495298 / 20.909259259259258	0.977439656
Precipitaciones	1.8186128526645768 / 3.0905692729766807	0.58843944
Presión atmosférica	1007.5450626959248 / 1010.9596021947874	0.996622477
Dirección del viento	175.2748322535415 / 166.24114012598034	1.054340894

Velocidad del viento	10.716300940438872 / 11.546639231824416	0.928088314
Nubosidad	0.5293397335423198 / 0.5673868312757202	0.932943284

Tabla 1. Valores referentes, por un lado, al promedio de los diferentes atributos para casos mortales/no mortales, y, por otro lado, cociente de estos promedios.

Se observa que aquellos atributos cuyos subgrupos son más homogéneos, es decir, el cociente entre sus medias tiene a 1, son la presión atmosférica, temperatura media y dirección del viento.

	Varianza(Mortales/NoMortales)	var(mortales)/ var(NoMortales)
Temperatura media	43.00889645222636 / 38.00093621399216	1.131785181
Precipitaciones	59.07984556798587 / 157.81003794752658	0.374373179
Presión atmosférica	1252.3451480368226 / 797.31153674161	1.57070993
Dirección del viento	10060.37241326284 / 10583.886055010347	0.950536727
Velocidad del viento	32.04333 / 38.990623	0.821821677
Nubosidad	32.04333929501477 / 38.990623135964306	0.935063928

Tabla 2. Valores referentes, por un lado, a la varianza de los diferentes atributos para casos mortales/no mortales, y, por otro lado, cociente de estas varianzas.

A la vista de los resultados se puede determinar que ningún par de subgrupos tienen asociadas varianzas iguales, a excepción de la dirección del viento y la nubosidad. Por esto, y en virtud de lo expuesto en el punto 2.3.1 Análisis estadístico (Significancia entre variables), se da por bueno el uso el *T-Test* de Welch para llevar a cabo el test de significancias, exceptuando estas dos magnitudes, para las cuales se aplicará el *T-Test* de Student.

Una vez obtenidos los valores de promedio y varianza de los pares de subgrupos de cada atributo, se han aplicado sendos t-test, obteniendo los siguientes resultados:

	Método	P-Value
Temperatura media	Welch	0.006389
Precipitaciones	Welch	4.87924E-06
Presión atmosférica	Welch	9.53662E-05
Dirección del viento	Student	0.00105632
Velocidad del viento	Welch	2.61461E-07
Nubosidad	Student	0.00013489

Tabla 3. Valores obtenidos para el P-Value.

Es importante señalar que el hecho de que los p-value sean bajos (<0.01) significa que las medias de los incidentes mortales son diferentes a las de los no mortales, y por tanto se puede concluir que las variables atmosféricas tienen influencia en la mortalidad de los accidentes.

4.6 Análisis de grupos

A la hora de aplicar las técnicas de clustering se ha buscado dar dos enfoques diferentes con el objetivo de observar cómo afecta a la caracterización de, en este caso, los incidentes, la elección por un lado de un valor adecuado de clústeres, y por otro, de uno u otro dataset, quitando unos u otros atributos del mismo. Lo que se pretende con esa caracterización es agrupar los incidentes según diferencias en las variables atmosféricas.

De acuerdo con lo anterior, sin descartar ningún atributo meteorológico del dataset, primero se ha querido encontrar cual es número de clústeres adecuado para el dataset completo, es decir, manteniendo todos los atributos del juego de datos. Para ello se ha ejecutado el algoritmo para diferentes valores de K, observando como varían los resultados. Posteriormente, se ha modificado el dataset original, eliminando atributos del mismo, con el objetivo de encontrar una potencial mejor caracterización de los incidentes. Hay que tener en cuenta que, para este segundo enfoque, al variar de una prueba a otra el número de variables, es necesario recalcular el valor de *K* en cada caso. Es por esto que se puede observar que el valor de *K* no es constante en este apartado.

4.6.1 Caracterización de los grupos en función del número de clústeres

En este apartado se han obtenido diferentes conjuntos de clústeres tomando valores diferentes para *K*.

Como se ha explicado en el apartado 2.3.2 Clustering, se ha representado la inercia de los datos frente al número de clústeres, obteniendo la siguiente gráfica:

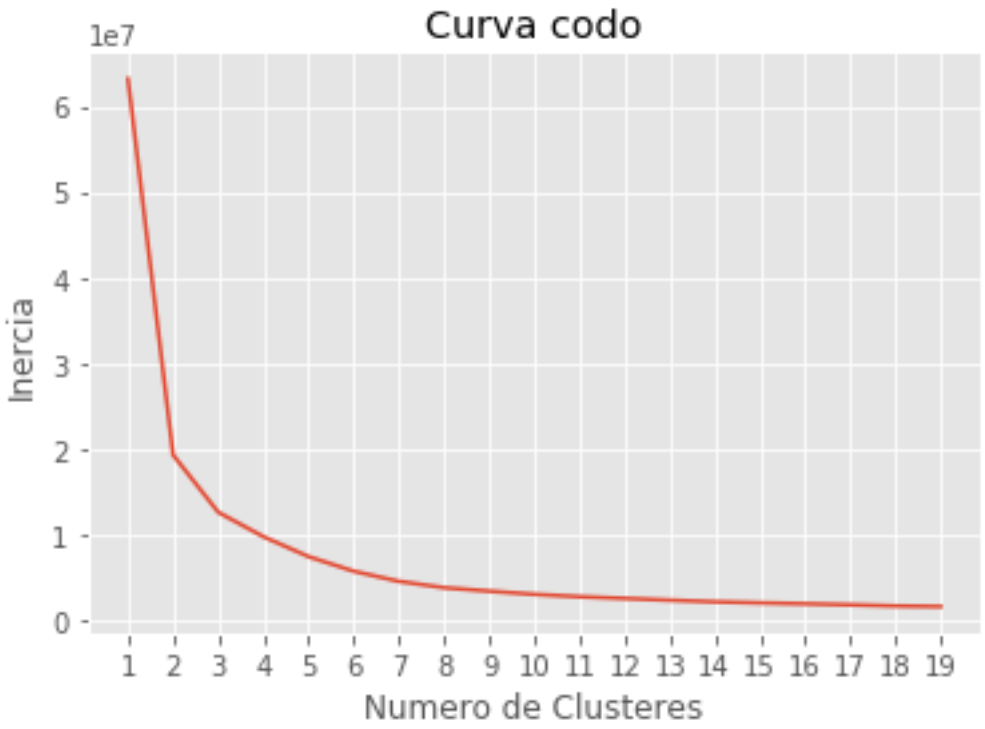

Ilustración 20. Curva codo para obtener el número de clústeres.

En ella se aprecia perfectamente como varía la curva a medida que aumenta el número de clústeres.

A continuación, se exponen los resultados obtenidos y se determinará qué *K* es la más adecuada:

K=3

Tmed	Precipitaciones	Presion	Dir_viento	V_Viento	Nubosidad
19.0038811	3.02704669	1007.68933	292.762432	11.4014554	0.57959369
20.8338489	2.5420583	1008.79607	188.955257	10.2248662	0.55294468
21.8750702	2.0525725	1011.10725	61.5816183	11.7067353	0.52391254

Tabla 4. Centroides para K=3.

Tomando K=3, se pueden caracterizar los incidentes con las siguientes propiedades principales:

1. Los incidentes caracterizados por el primer clúster se caracterizan por tener una temperatura ligeramente inferior a la de los otros dos clústeres, 19 ºC, y unas precipitaciones superiores, que en las escalas más habituales se corresponden con las de una lluvia ligera. En este caso el viento soplaba dirección oeste-noroeste.
2. Los incidentes asociados al segundo clúster destacan principalmente por tener una intensidad en la velocidad del viento inferior a la de los otros dos clústeres, clasificada en la escala de Beaufort como *Flojito* (Brisa muy débil), siendo en los otros casos *Flojo* (brisa ligera). En este caso el viento sopla dirección sur.
3. Los incidentes contenidos en el tercer clúster poseen la principal característica de ser aquel que reúne las condiciones meteorológicas más secas: temperatura más alta, y precipitaciones más bajas. En este caso la dirección del viento soplaba este-nordeste.

K=4

Tmed	Precipitaciones	Presion	Dir_viento	V_Viento	Nubosidad
16.6986425	6.08144796	882.918552	216.203234	9.69230769	0.75961538
19.3166667	3.04156939	1015.58923	293.39251	11.4779507	0.56882296
20.9429467	2.0339185	1015.78514	189.833149	10.2721003	0.54224138
21.9181517	2.07345972	1013.20962	61.1774882	11.750237	0.51919431

Tabla 5. Centroides para K=4.

Tomando K=4, se pueden caracterizar los incidentes con las siguientes propiedades principales:

1. Los incidentes del primer clúster se caracterizan principalmente por ser el que tiene un porcentaje de nubosidad sustancialmente más alto que el resto, aparte de por tener una temperatura ligeramente inferior a la de los otros dos clústeres, 16 ºC, y unas precipitaciones superiores, que en las escalas más habituales se corresponden con las de una lluvia ligera, casi moderada. En este caso el viento soplaba dirección sudoeste, siendo el menos intenso (brisa muy débil en la escala de Beaufort). Cabe mencionar que, por la presión atmosférica inusualmente baja, los incidentes integrados en este clúster podrían corresponderse con aquellos producidos a una cierta altitud.
2. En los incidentes asociados al segundo clúster el único atributo meteorológico que podría destacar por encima del resto sería la intensidad del viento, calificándose como moderada (al igual que en el 4º clúster) según la escala de Beaufort, y teniendo una dirección oeste-noroeste.
3. En el tercer clúster los incidentes se caracterizan por tener las menores precipitaciones de los 4 grupos, calificándose como lluvias débiles, y teniendo una dirección del viento sur.
4. El cuarto clúster se caracteriza por agrupar los incidentes con las temperaturas superiores, y también una intensidad del viento superior, llegando a calificarse como brisa moderada, y teniendo una dirección este de este-nordeste. Cabe destacar también que es el grupo en el que el porcentaje de nubes en el cielo es menor.

K=5

Tmed	Precipitaciones	Presion	Dir_viento	V_Viento	Nubosidad
17.9659218	2.33128492	870.949721	128.195054	9.89944134	0.73393855
18.4635163	2.86371951	1008.21596	315.410823	10.7581301	0.56770833
20.1505176	3.0310559	1012.97288	236.742236	11.3678399	0.57781228
21.761137	1.80643057	1016.19143	149.155985	10.0792171	0.53087139
21.9832866	2.29259675	1016.85788	53.6542344	11.9871004	0.50953449

Tabla 6. Centroides para K=5.

Tomando K=5, se pueden caracterizar los incidentes con las siguientes propiedades principales:

1. Los incidentes del primer clúster se caracterizan principalmente por ser el que tiene un porcentaje de nubosidad sustancialmente más alto que el resto, aparte de por tener una temperatura ligeramente inferior a la de los otros dos clústeres, 17.9 ºC. Cabe

mencionar que, por la presión atmosférica inusualmente baja, los incidentes integrados en este clúster podrían corresponderse con aquellos producidos a una cierta altitud.

2. En el caso del segundo clúster, los incidentes no tienen ningún atributo que destaque por encima de los demás. Si acaso podrían caracterizarse por la dirección del viento, siendo en este caso noroeste.
3. En los incidentes asociados al tercer clúster el único atributo meteorológico que podría destacar por encima del resto serían las precipitaciones, clasificándose según la escala de intensidades como *lluvia ligera*. La dirección del viento en este caso soplaría sudoeste.
4. Los incidentes pertenecientes al 4º clúster se caracterizan por tener unas temperaturas notablemente superiores al resto de clústeres (a excepción del 5º), y unas precipitaciones menores que el resto. En este caso la dirección del viento es sur-sudeste.
5. Por último, el quinto clúster agrupa a los incidentes con la temperatura más alta (ligeramente superior a la de los incidentes del 4º clúster), y una velocidad del viento superior, calificándose por su intensidad como brisa moderada, y soplando dirección nordeste. También cabe mencionar que en este grupo de incidentes el porcentaje de cielo cubierto por nubes es el menor.

También se ha aplicado para K=6 y K=7, pero no se han obtenido clústeres con elementos diferenciadores entre sí, con lo que no merece la pena comentarlos. Se pueden encontrar los resultados obtenidos para estos casos al final del documento, en el apéndice II.

4.6.2 Caracterización de los grupos en función de las variables input

En este caso se estudia como varía el valor de K en función de los dataset que se le pasa como input al algoritmo, y se analizan los grupos de clústeres resultantes.

Sin dirección del viento

En este caso se ha eliminado del dataset el atributo *dirección del viento*, obtenido una curva codo con las siguientes características:

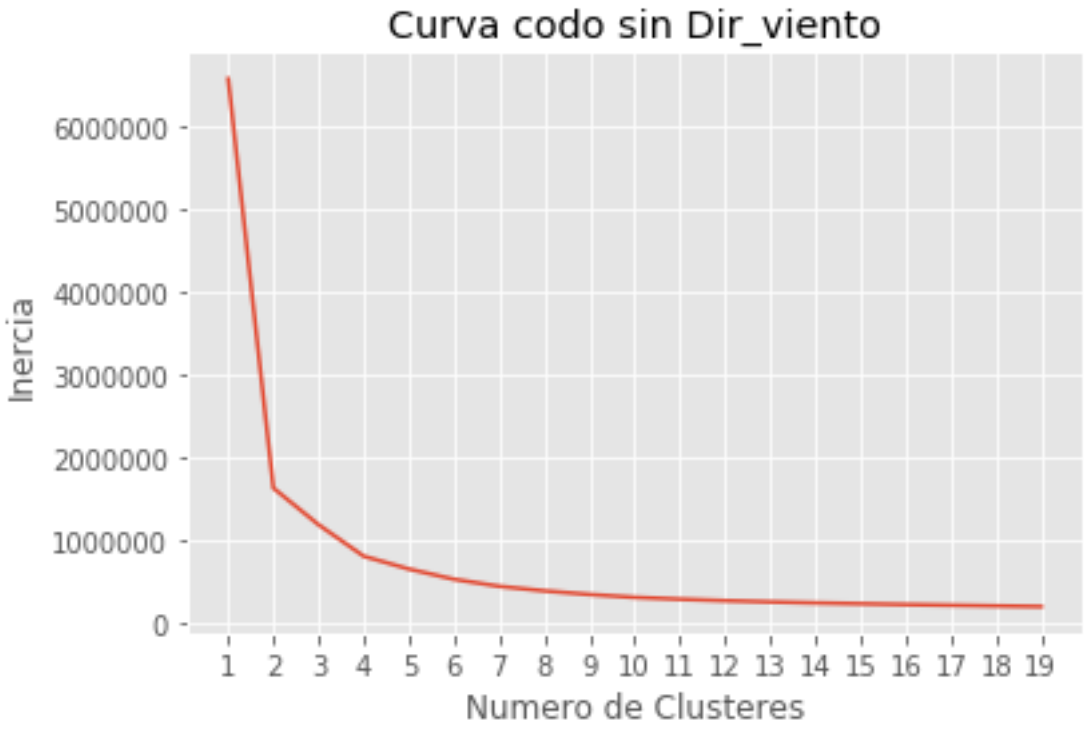

Ilustración 21. Curva codo para obtener el número de clústeres, habiendo eliminado del dataset la dirección del viento.

En este caso, dada la forma de la curva representada en la ilustración 21, y habiendo eliminado una de las variables del dataset, se ha determinado que el número adecuado de clústeres sea K=4. Como se puede apreciar, el cambio de curvatura se produce en algún punto del intervalo $K \in [2.5\,, 5]$, más cercano a 5 pero sin llegar.

Teniendo en cuenta esto, obtenemos los siguientes centroides:

Tmed	Precipitaciones	Presión	V_Viento	Nubosidad
13.5765217	60.7565217	1007.6687	16.373913	0.92173913
16.4757962	2.06305732	850	10.7834395	0.69187898
17.9822368	2.52434211	925	8.34868421	0.79605263
21.063977	1.18132435	1016.90747	11.1365979	0.52929223

Tabla 7. Centroides para K=4, resultantes de haber eliminado del dataset la dirección del viento.

1. En los incidentes asociados al primer clúster se observa que la temperatura es sustancialmente más baja en el resto de casos, teniendo unas precipitaciones moderadamente altas, calificadas de *lluvia muy fuerte a torrencial*, con un viento también moderadamente intenso, y una nubosidad alta.
2. En este clúster los incidentes se caracterizan principalmente por la baja magnitud de la presión atmosférica, lo cual daría a entender que se producen a una altura sustancialmente mayor de la habitual
3. En este clúster ocurre lo mismo que en el anterior, pero a una altura intermedia.
4. Por último, los incidentes asociados al 4º clúster tienen la principal característica de producirse a una temperatura superior al resto, menos precipitaciones, y menos nubosidad.

Sin dirección del viento ni precipitaciones

En este caso se ha eliminado del dataset el atributo *precipitaciones* además de *dirección del viento*, obtenido una curva codo con las siguientes características:

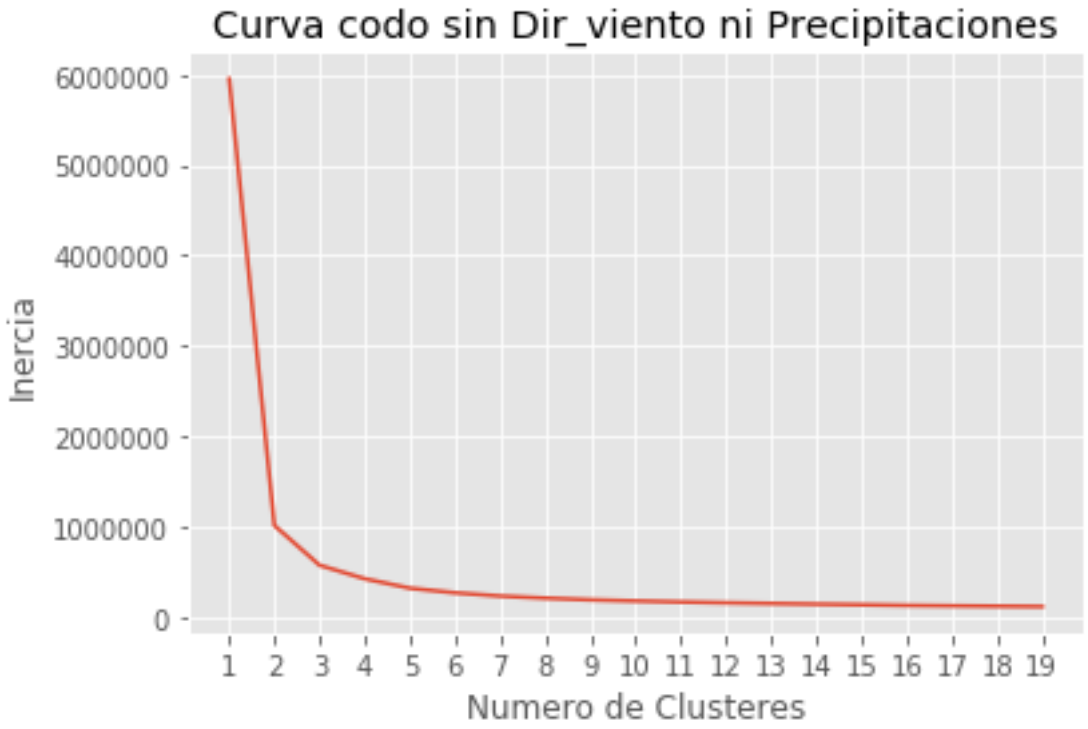

Ilustración 22. Curva codo para obtener el número de clústeres, habiendo eliminado del dataset la dirección del viento y las precipitaciones.

En este caso se puede apreciar en la ilustración 22 como el cambio de curvatura se ha corrido hacia la izquierda, haciéndose más brusco. El codo tiene lugar en el intervalo $K \in [2.5\,, 3.5]$ aproximadamente, siendo K más cercana a este último punto. Se toma, pues, K=3, obteniendo los siguientes centroides para este dataset:

Tmed	Presion	V_Viento	Nubosidad
16.4757962	850	10.7834395	0.69187898
17.7467949	925	8.33974359	0.80128205
20.9064597	1016.77269	11.2558681	0.53768186

Tabla 8. Centroides para K=3, resultantes de haber eliminado del dataset la dirección del viento y las precipitaciones.

1. Otra vez la presión vuelve a caracterizar los clústeres. En este caso, en el primero, los incidentes parece que se hayan producido a una altura sustancialmente mayor que la del nivel del mar.
2. En el caso del segundo clúster, los incidentes parece que se hayan producido a una altura intermedia.
3. En el tercer caso, la presión atmosférica parece más estándar con lo que indica que el incidente se hubiera producido a nivel del mar, y según el resto de atributos en días de temperaturas moderadamente cálidas y con una nubosidad no demasiado alta.

Sin presión

En este caso se ha eliminado del dataset el atributo *presión*, obtenido una curva codo con las siguientes características:

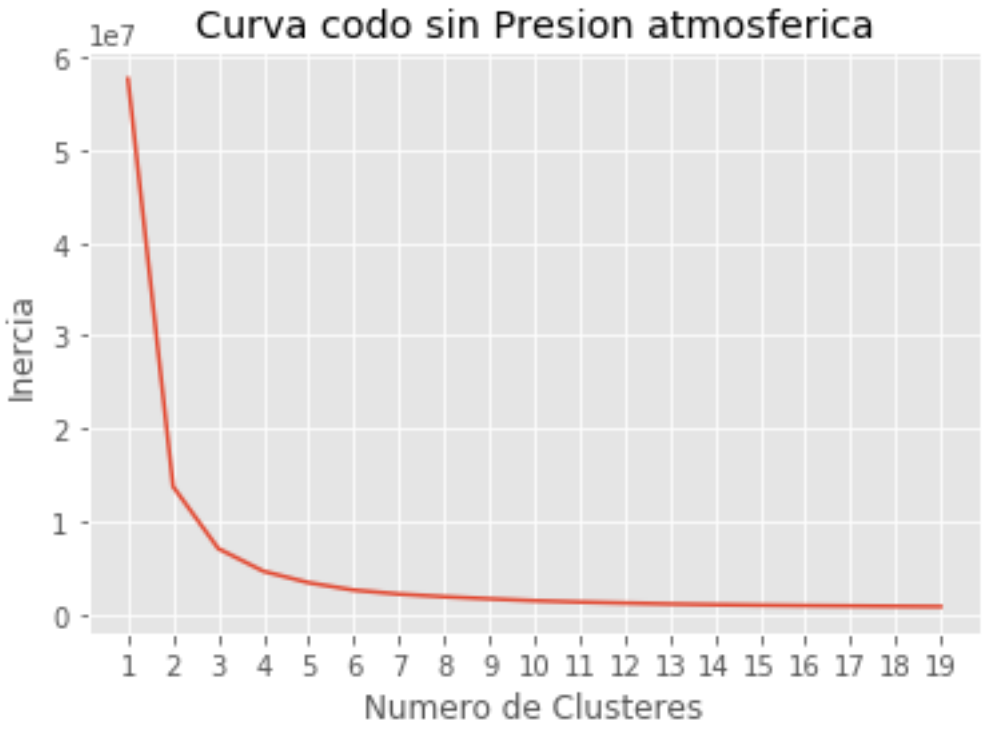

Ilustración 23. Curva codo para obtener el número de clústeres, habiendo eliminado del dataset la presión atmosférica.

En este caso se puede apreciar como el cambio de curvatura representada en la ilustración 23 tiene lugar en el intervalo $K \in [2, 3]$ aproximadamente, siendo K más cercana a este último punto. Se toma, pues, K=3, obteniendo los siguientes centroides para este dataset:

Tmed	Precipitaciones	Dir_viento	V_Viento	Nubosidad
19.0337553	3.00880048	292.450422	11.3936106	0.57798373
20.8093301	2.55502392	188.60693	10.2230861	0.55412679
21.8801123	2.05409453	61.5510061	11.7094057	0.5240992

Tabla 9. Centroides para K=3, resultantes de haber eliminado del dataset la presión atmosférica.

Eliminando la presión del dataset de entrada, el único atributo que hace que los clústeres sean diferentes entre sí es la dirección del viento, siendo oeste-noroeste en el primer clúster, sur en el segundo, y este-nordeste en el tercero.

4.7 Análisis predictivo

A continuación, se muestra el árbol de decisión creado en nuestro proyecto entrenado con 3827 incidentes:

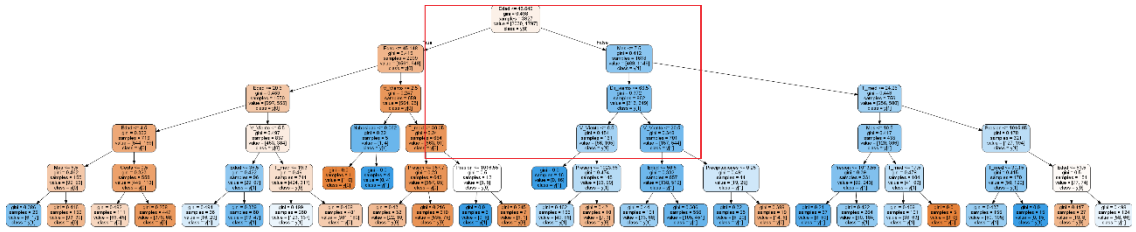

Ilustración 24. Árbol completo

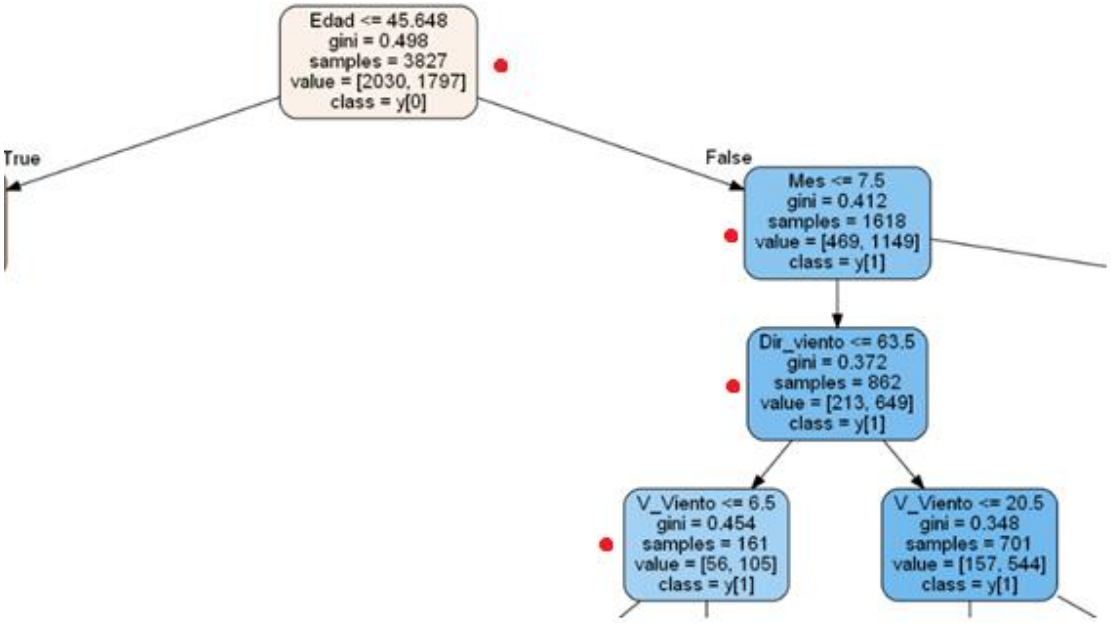

Ilustración 25. Fragmento del árbol

En la ilustración 25, la cual es un fragmento extraído de la ilustración 24 se puede observar el nodo raíz, el cual se ramifica en 2. El primer parámetro de los diferentes nodos representa diferentes atributos del incidente (edad, mes del año, dirección del viento, velocidad del viento, etc.). Siguiendo el camino indicado por los puntos rojos, una posible representación sería la siguiente: la primera caja dice que Tenemos 3827 casos. De esos, 1618 tienen una edad>=45.648. La segunda caja indica que de esos 1618, 862 invidentes ocurrieron antes del mes 7.5 (mediados de Julio). La siguiente caja dice que 862 incidentes se produjeron con una dirección del viento inferior a 63.5^0. y, por último, la siguiente indica que, de esos 862 incidentes, 161 se produjeron con una velocidad del viento inferior a 6.5km/h, y así sucesivamente hasta llegar abajo del árbol y quedar el incidente clasificado en mortal o no mortal.

Teniendo un dataset con 5468 registros, se ha destinado un 70% de los incidentes a entrenamiento, y el 30% restante a test.

Para determinar un valor adecuado para el parámetro *max_depth_range* (profundidad del árbol), se ha calculado el *DecisionTreeClasifier.Score* en función del *max_depth_range*, obteniéndose:

max_depth_range	*DecisionTreeClasifier.Score*
5	0.711152
1	0.706886

2	0.706886
3	0.706886
6	0.706277
7	0.702011
9	0.702011
4	0.698964
8	0.694698
11	0.689823
10	0.688605
12	0.680073
14	0.678854
13	0.669104

Tabla 10. Valores relativos a profundidad del árbol y DecisionTreeClasifier.Score*, que no es más que la precisión media del árbol.*

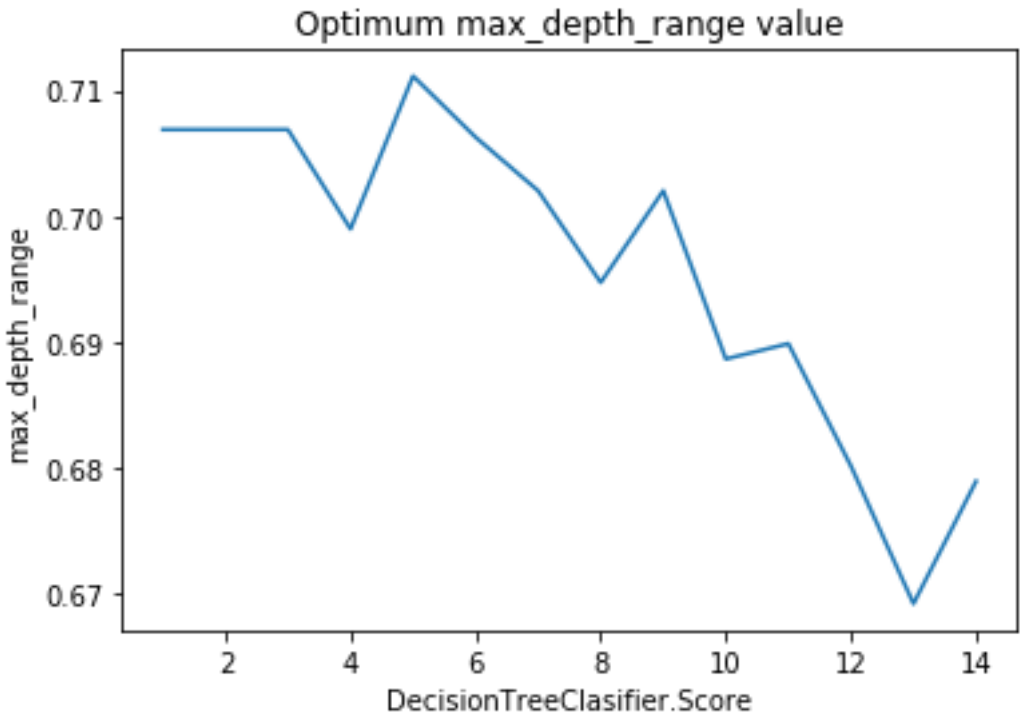

Ilustración 26. Representación gráfica de los valores de la tabla 10.

A la vista de los resultados que arroja la gráfica representada en la ilustración 26, es evidente que el valor óptimo para el parámetro es 5.

Para saber exactamente qué tan bien funciona el árbol que se ha generado, se construye una matriz de confusión, que recibe como inputs valores de test, y valores predichos por el árbol, obteniendo:

	Positivo	Negativo
Positivo	609	277
Negativo	197	558

Tabla 11. Matriz de confusión

Para evaluar los resultados, calculamos:

- Sensibilidad o Tasa de positivos verdaderos: Proporción de casos positivos que están bien detectadas por la prueba.

$$TPV = \frac{tp}{tp + fp}$$

  Donde tp es el número de verdaderos positivos y fp el número de falsos positivos.
- Especificidad, también llamada Tasa de verdaderos negativos: proporción de casos negativos que son bien detectadas por la prueba.

$$TVN = \frac{tp}{tp + fn}$$

  Donde tp es el número de verdaderos positivos y fn el número de falsos negativos.

Como se ha mencionado anteriormente, se trata de un modelo de clasificación binaria. Esto significa que los resultados se clasifican como 1 o 0. Se asumen como positivos aquellos incidentes cuya mortalidad toma valor 1, es decir, el desenlace es *mortal*. Los casos negativos serán aquellos incidentes que se hayan clasificado como 0, es decir, el desenlace es *no mortal*.

Se han obtenido los siguientes resultados:

$$TPV = \frac{tp}{tp + fp} = \frac{609}{609 + 277} = 0.6874$$

$$TNV = \frac{tp}{tp + fn} = \frac{609}{609 + 197} = 0.7558$$

A partir de los resultados anteriores, se puede establecer que, en este caso, la precisión (media ponderada de TPV y TNV) será de un 72.16%.

5 Discusión de resultados

En primer lugar, se ha observado que, tras haber calculado el coeficiente de correlación de Spearman entre los diferentes atributos, y haber representado los resultados en los diferentes mapas de calor, no se han encontrado correlaciones significantes más allá de las que a priori pudieran resultar obvias, como por ejemplo Mortalidad-Pronóstico, Precipitaciones-Nubosidad o Temperatura-Mes. El resto de pares de atributos no alcanzan un valor conclusivo para el coeficiente de correlación, no se llega a valores conclusivos.

Es por esto que para obtener información más concluyente ha sido necesario recurrir a otros métodos como puedan ser los de T-Test o técnicas de *machine learning*.

Tras haber aplicado el método *Welch Two Sample t-test* a los subgrupos de los diferentes atributos se pudo observar que los diferentes valores obtenidos para el P-Value en todos los casos es menor que el nivel de significancia de 0.05. Esto implica que se puede rechazar la hipótesis nula (promedio atributo casos mortales=promedio atributo casos no mortales) y se puede concluir que, para cada atributo, el promedio del mismo en los casos mortales es diferente del de los casos no mortales, es decir, es significante.

En cuanto al análisis de grupos, se han aplicado técnicas de clustering con dos enfoques diferentes. En el primer caso se observó que buscar elementos diferenciadores más allá de los 5 clústeres resulta bastante complicado, ya que la magnitud de los atributos para los diferentes

grupos que se crean acaban solapando, o evolucionando sin grandes saltos, y generando agrupaciones las cuales parecen similares entre sí, es decir, que las condiciones meteorológicas sí tienen influencia en la mortalidad de los accidentes.

Tras analizar todos los grupos, se determinó que el número ideal de clústeres para establecer unos grupos con características diferenciadas entre si es el de K=4 por los siguientes motivos:

- Tenemos un clúster (el 1º) que tiene un porcentaje de nubosidad notablemente superior que el resto, precipitaciones moderadas, además de poder inferir por la presión atmosférica que los incidentes son a una cierta altura.
- En el segundo clúster encontramos una media de precipitaciones a una presión atmosférica que indicaría una altura a nivel del mar superior al resto, con una intensidad del viento moderada.
- Tenemos otra agrupación de incidentes que se caracterizan por tener temperaturas moderadamente altas a la vez que un nivel de precipitaciones bajo.
- Por último, tenemos otro grupo en el que los incidentes se producen con un nivel de precipitaciones también bajo, pero con una intensidad del viento más moderada.

Si nos quedásemos con la K=3 se pudo observar que no llegan a aparecer aquellos valores notablemente más altos de nubosidad o precipitaciones, o el de la presión sustancialmente más baja.

Para el segundo caso se probó a repetir la operación de eliminar unos u otros atributos del dataset sin conseguir diferencias significativas aparentes entre los clústeres obtenidos. A la vista de los resultados, se puede determinar que eliminando la dirección del viento se obtienen clústeres que, teniendo elementos diferenciadores entre sí, no llegan a sufrir una pérdida de información significativa.

Tras abordar el clustering desde dos enfoques diferentes, se puede determinar que el primero es el que ha arrojado unos resultados mejores y más completos a la hora de caracterizar los incidentes acuáticos.

Por último, se ha aplicado el algoritmo de los árboles de decisión con el objetivo de clasificar los incidentes en función de las condiciones meteorológicas que rodean al mismo. Se ha observado que el árbol es más efectivo a la hora de clasificar casos con desenlace no mortal, que incidentes con desenlace mortal, obteniéndose una tasa de positivos verdaderos del 68.74%, y una tasa de negativos verdaderos del 75.58%. En términos globales, el árbol que se ha generado tiene una precisión del 72.16%. En cuanto a los arboles de decisión.

Se puede discutir que, en el caso del algoritmo de árboles de decisión, la precisión no es la más alta, pero el objetivo del presente TFG no era tanto buscar la mejor implementación de un algoritmo en concreto, si no realizar de principio a fin un proyecto de *data science*, desarrollando todas y cada una de las etapas que lo componen, así como utilizar técnicas de las diferentes familias de algoritmos de *machine learning* que existen.

6 Conclusiones

La realización de este TFG da cuenta del potencial que poseen las técnicas y metodología aplicadas. Esto es de gran interés para cualquier disciplina científica, y en especial para la Física, en tanto en cuanto es una ciencia en cuyos experimentos, con el desarrollo actual de la

tecnología, el volumen de datos recabados está en constante aumento. Esto se traduce en que el investigador, en el momento que es capaz de dominar estas disciplinas, puede llegar a alcanzar un nivel superior en la compresión de los datos, siendo de gran ayuda a la hora de afrontar futuros problemas.

En el caso concreto de este TFG me gustaría remarcar la idea de que, en un proyecto de estas características, siendo *el dato* el protagonista, lo menos importante es su naturaleza. Esto creo que es tremendamente importante porque, el hecho de conocer y saber aplicar la metodología, filosofía y técnicas pertenecientes al ámbito del *data science*, y más concretamente del *machine learning*, te abre un abanico de posibilidades inmenso en el mundo laboral, ya sea en el ámbito de la empresa o el académico.

En plano más personal este proyecto me ha servido para aprender desde cero y adquirir destrezas en el uso del lenguaje *Python*, algunas de sus bibliotecas más importantes, la utilización de *notebooks* de *Jupyter*, así como las diferentes técnicas de aprendizaje supervisado y no supervisado de *machine learning* que se exponen a lo largo de este TFG, y que se han aplicado en el caso de incidentes acuáticos.

Todas estas destrezas adquiridas, metodología, forma de abordar un problema en el cual se buscar llevar a cabo un análisis de datos desde diferentes enfoques, etc. estoy seguro que me completaran en el ámbito profesional en el cual ya me desarrollo.

6 Bibliografía

[1] Boletín informativo. Colegio Oficial de Físicos. Noviembre del 2015.

[2] D. Guest, K. Cranmer, and D. Whiteson. Deep learning and its application to LHC Physics. Annual Review of Nuclear and Particle Science, 2018.

[3] https://home.cern/news/news/computing/lhc-pushing-computing-limits Consultada el 24 de septiembre de 2020.

[4] J. Collins, K. Howe, B. Nachman. Anomaly detection for resonant new Physics with Machine Learning. Physical Review Letters, 2018.

[5] E. Fol, J. Coello de Portugal, G. Franchetti, R. Tomás. Optics corrections using Machine Learning in the LHC. 10th Int. Particle Accelerator Conf. Melbourne, 2019.

[6] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Steenstrup Pedersen, C. Igel. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. IEEE Intelligent Systems, 2017.

[7] J. B. Rollins, Metodología fundamental para la ciencia de datos. IBM Analytics, 2015.

[8] I. H. Witten, E. Frank, M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc, 2011.

[9] A. Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media, 2017.

[10] B. Sierra. Aprendizaje Automático: conceptos básicos y avanzados. Pearson Educación, 2006.

[11] J. H. Orallo; M. J. Ramírez Quintana, C. F. Ramírez. Introducción a la Minería de Datos. Pearson Educación, 2004.

[12] https://es.wikipedia.org/wiki/Correlaci%C3%B3n Consultada el 25 de septiembre de 2020.

[13] Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med, 2018.

[14] P. Morales, L. Rodriguez. Aplicación de los coeficientes correlación de Kendall y Spearman, 2016.

[15] https://www.cienciadedatos.net/documentos/12_t-test#T-test:_Comparaci%C3%B3n_de_medias_poblacionales_independientes

[16] G. E. Dallal. *The Little Handbook of Statistical Practice*, 2012.

[17] R. A Fisher. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd, 1925.

[18] http://www.aprendemachinelearning.com/arbol-de-decision-en-python-clasificacion-y-prediccion/ Consultada el 27 de septiembre de 2020.

[19] https://scikit-learn.org/stable/ Consultada el 27 de septiembre de 2020.

[20] https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees_.28CART.29 Consultada el 27 de septiembre de 2020.

[21] https://es.wikipedia.org/wiki/Python Consultada el 27 de septiembre de 2020.

[22] E. Bahit. Python para principiantes, Edición 2020.

[23] https://numpy.org/ Consultada el 27 de septiembre de 2020.

[24] https://matplotlib.org/ Consultada el 27 de septiembre de 2020.

[25] http://pandas.pydata.org/ Consultada el 27 de septiembre de 2020.

[26] https://jupyter.org/ Consultada el 27 de septiembre de 2020.

Anexo I: Correlaciones

Sin restricciones de meses:

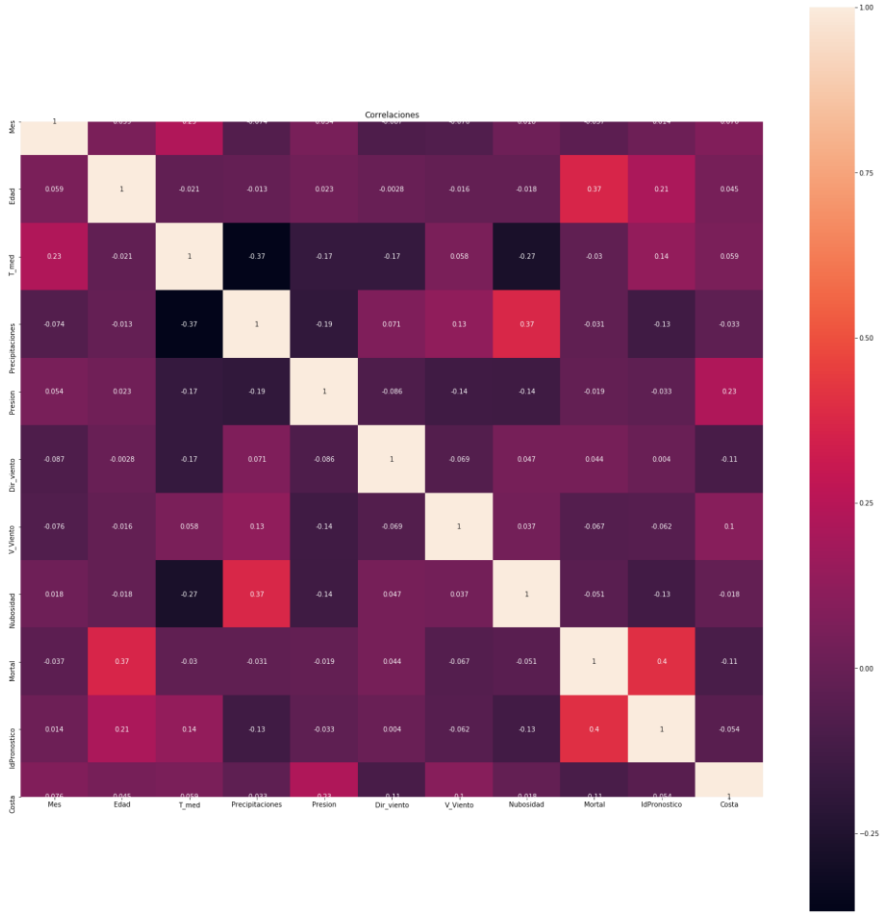

Ilustración 27. Mapa de calor de las correlaciones entre las diferentes variables sin restricciones de meses.

	Variable1	Variable2	correlacion
0	Mortal	IdPronostico	0.404376
1	Precipitaciones	Nubosidad	0.373877
2	Edad	Mortal	0.366838
3	T_med	Mes	0.230387
4	Presion	Costa	0.225109
5	Edad	IdPronostico	0.211477
6	T_med	IdPronostico	0.138071
7	Precipitaciones	V_Viento	0.126428
8	Costa	V_Viento	0.101601
9	Mes	Costa	0.076021
10	Precipitaciones	Dir_viento	0.070778
11	Costa	T_med	0.058797
12	Edad	Mes	0.058762
13	T_med	V_Viento	0.057680
14	Presion	Mes	0.054084
15	Nubosidad	Dir_viento	0.046975
16	Costa	Edad	0.044584
17	Mortal	Dir_viento	0.043715
18	Nubosidad	V_Viento	0.036501
19	Edad	Presion	0.023243
20	Mes	Nubosidad	0.017960

Tabla 7. Primeras 20 correlaciones ordenadas de manera descendente.

	Variable1	Variable2	correlacion
54	Precipitaciones	T_med	-0.367838
53	Nubosidad	T_med	-0.274266
52	Precipitaciones	Presion	-0.186483
51	Presion	T_med	-0.174684
50	T_med	Dir_viento	-0.173357
49	Nubosidad	Presion	-0.141009
48	V_Viento	Presion	-0.135917
47	Precipitaciones	IdPronostico	-0.132929
46	IdPronostico	Nubosidad	-0.129219
45	Costa	Dir_viento	-0.107437
44	Costa	Mortal	-0.105358
43	Mes	Dir_viento	-0.086605
42	Presion	Dir_viento	-0.085798
41	Mes	V_Viento	-0.076338
40	Precipitaciones	Mes	-0.073789
39	V_Viento	Dir_viento	-0.068960
38	V_Viento	Mortal	-0.067198
37	V_Viento	IdPronostico	-0.062370
36	IdPronostico	Costa	-0.054468
35	Nubosidad	Mortal	-0.050893
34	Mes	Mortal	-0.036680

Tabla 8. Primeras 20 correlaciones ordenadas de manera ascendente.

Meses de invierno:

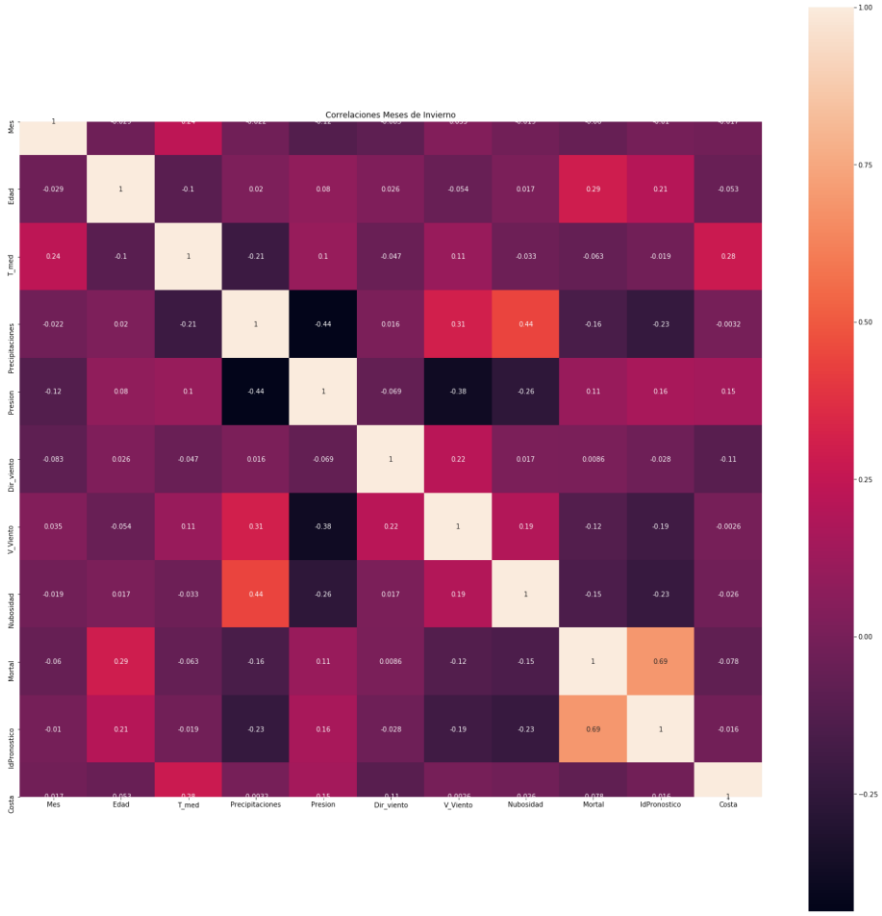

Ilustración 28. Mapa de calor de las correlaciones entre las diferentes variables asociadas a los meses de otoño.

	Variable1	Variable2	correlacion
0	Mortal	IdPronostico	0.693024
1	Precipitaciones	Nubosidad	0.441539
2	Precipitaciones	V_Viento	0.306835
3	Mortal	Edad	0.293387
4	T_med	Costa	0.279448
5	Mes	T_med	0.235181
6	V_Viento	Dir_viento	0.217256
7	Edad	IdPronostico	0.206018
8	V_Viento	Nubosidad	0.192834
9	Presion	IdPronostico	0.161206
10	Presion	Costa	0.151040
11	Presion	Mortal	0.110118
12	T_med	V_Viento	0.108560
13	Presion	T_med	0.102471
14	Edad	Presion	0.079779
15	V_Viento	Mes	0.035051
16	Dir_viento	Edad	0.026439
17	Precipitaciones	Edad	0.019652
18	Nubosidad	Edad	0.017156
19	Dir_viento	Nubosidad	0.016828
20	Dir_viento	Precipitaciones	0.016359

Tabla 9 Primeras 20 correlaciones ordenadas de manera descendente.

	Variable1	Variable2	correlacion
54	Presion	Precipitaciones	-0.436713
53	V_Viento	Presion	-0.379426
52	Presion	Nubosidad	-0.258707
51	IdPronostico	Precipitaciones	-0.231533
50	Nubosidad	IdPronostico	-0.229035
49	T_med	Precipitaciones	-0.208677
48	IdPronostico	V_Viento	-0.191241
47	Mortal	Precipitaciones	-0.160438
46	Mortal	Nubosidad	-0.148676
45	Mortal	V_Viento	-0.124558
44	Mes	Presion	-0.123293
43	Costa	Dir_viento	-0.106995
42	Edad	T_med	-0.100563
41	Dir_viento	Mes	-0.083434
40	Mortal	Costa	-0.077613
39	Dir_viento	Presion	-0.068876
38	Mortal	T_med	-0.063371
37	Mes	Mortal	-0.059876
36	V_Viento	Edad	-0.054271
35	Edad	Costa	-0.053228
34	Dir_viento	T_med	-0.046645

Tabla 10. Primeras 20 correlaciones ordenadas de manera ascendente.

Meses de primavera:

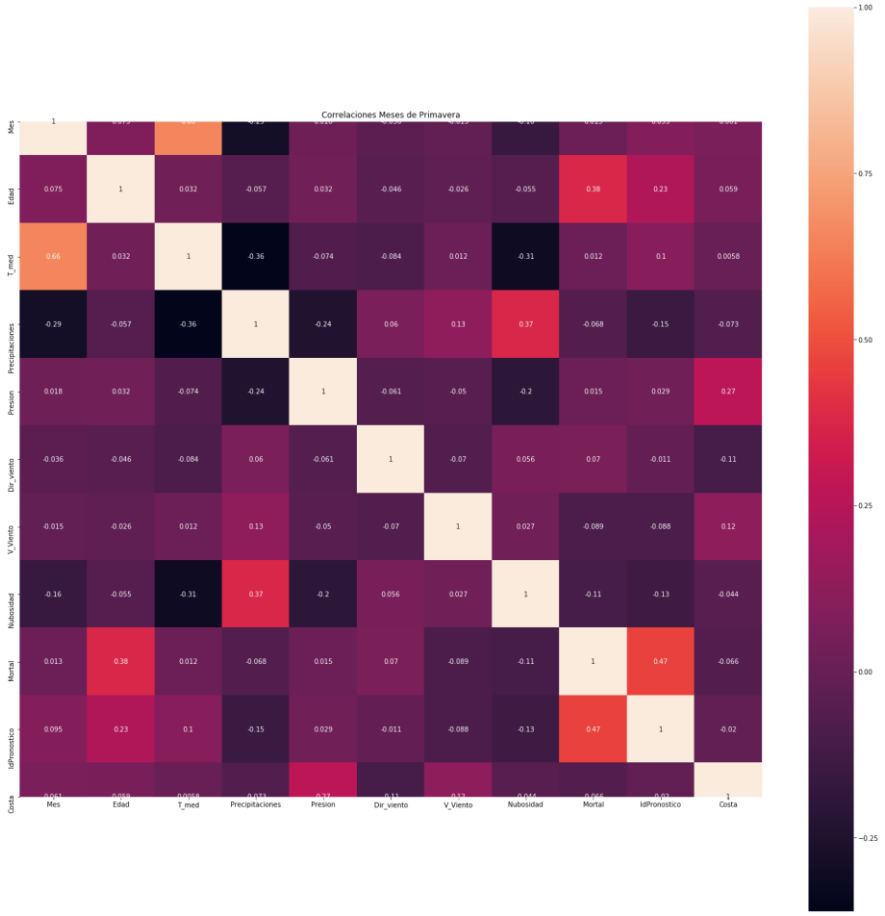

Ilustración 29. Mapa de calor de las correlaciones entre las diferentes variables asociadas a los meses de primavera.

	Variable1	Variable2	correlacion
0	Mes	T_med	0.655491
1	Mortal	IdPronostico	0.465086
2	Edad	Mortal	0.375840
3	Nubosidad	Precipitaciones	0.373468
4	Costa	Presion	0.268722
5	IdPronostico	Edad	0.226564
6	V_Viento	Precipitaciones	0.128649
7	Costa	V_Viento	0.124451
8	IdPronostico	T_med	0.102761
9	Mes	IdPronostico	0.095177
10	Mes	Edad	0.075311
11	Mortal	Dir_viento	0.069842
12	Mes	Costa	0.061279
13	Precipitaciones	Dir_viento	0.060212
14	Costa	Edad	0.058912
15	Nubosidad	Dir_viento	0.055872
16	Presion	Edad	0.031722
17	Edad	T_med	0.031696
18	Presion	IdPronostico	0.028814
19	Nubosidad	V_Viento	0.027399
20	Presion	Mes	0.018474

Tabla 11 Primeras 20 correlaciones ordenadas de manera descendente.

	Variable1	Variable2	correlacion
54	T_med	Precipitaciones	-0.360637
53	Nubosidad	T_med	-0.312607
52	Mes	Precipitaciones	-0.287222
51	Precipitaciones	Presion	-0.240652
50	Nubosidad	Presion	-0.199811
49	Nubosidad	Mes	-0.161669
48	IdPronostico	Precipitaciones	-0.151605
47	Nubosidad	IdPronostico	-0.134142
46	Mortal	Nubosidad	-0.109939
45	Dir_viento	Costa	-0.108081
44	Mortal	V_Viento	-0.088875
43	V_Viento	IdPronostico	-0.087781
42	Dir_viento	T_med	-0.084365
41	T_med	Presion	-0.073842
40	Precipitaciones	Costa	-0.073244
39	Dir_viento	V_Viento	-0.069666
38	Precipitaciones	Mortal	-0.067642
37	Mortal	Costa	-0.065870
36	Presion	Dir_viento	-0.060572
35	Edad	Precipitaciones	-0.057214
34	Edad	Nubosidad	-0.054922

Tabla 12. Primeras 20 correlaciones ordenadas de manera ascendente.

Meses de verano:

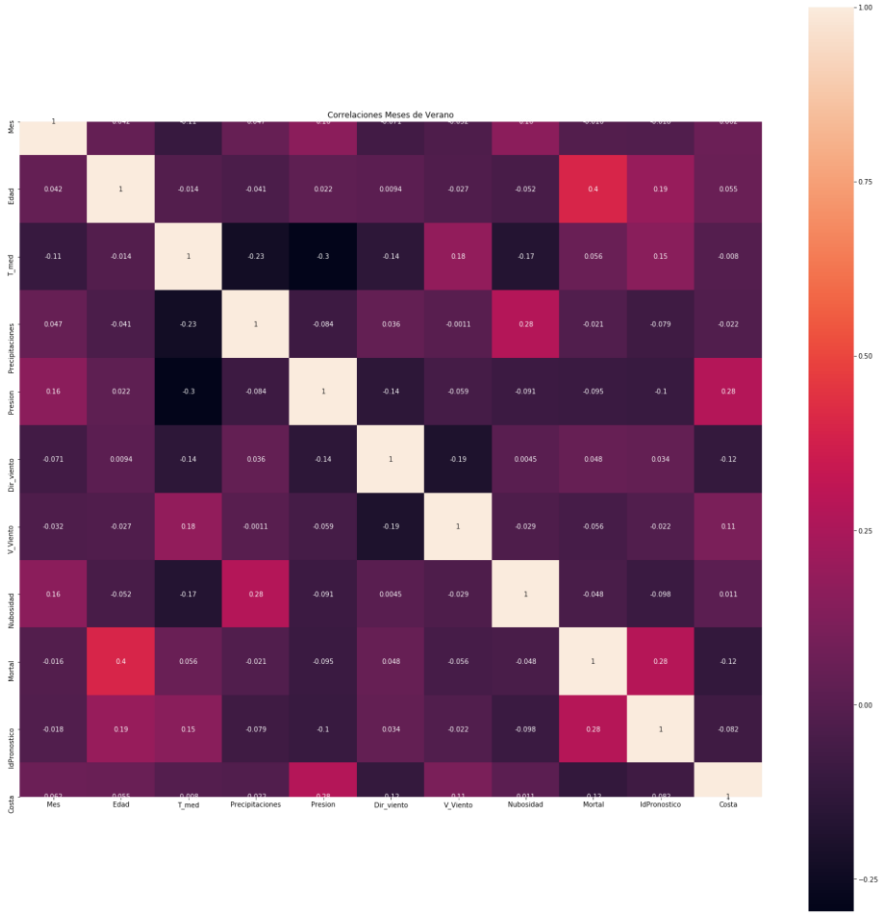

Ilustración 30. Mapa de calor de las correlaciones entre las diferentes variables asociadas a los meses de verano.

	Variable1	Variable2	correlacion
0	Edad	Mortal	0.400535
1	IdPronostico	Mortal	0.284922
2	Nubosidad	Precipitaciones	0.280699
3	Costa	Presion	0.277290
4	Edad	IdPronostico	0.191028
5	V_Viento	T_med	0.179076
6	Mes	Presion	0.158227
7	Nubosidad	Mes	0.157645
8	T_med	IdPronostico	0.150273
9	V_Viento	Costa	0.107893
10	Mes	Costa	0.062222
11	Mortal	T_med	0.056418
12	Costa	Edad	0.055443
13	Dir_viento	Mortal	0.048037
14	Mes	Precipitaciones	0.047354
15	Mes	Edad	0.042098
16	Dir_viento	Precipitaciones	0.035717
17	IdPronostico	Dir_viento	0.034288
18	Presion	Edad	0.021768
19	Nubosidad	Costa	0.010700
20	Edad	Dir_viento	0.009370

Tabla 13 Primeras 20 correlaciones ordenadas de manera descendente.

	Variable1	Variable2	correlacion
54	T_med	Presion	-0.296136
53	Precipitaciones	T_med	-0.231050
52	Dir_viento	V_Viento	-0.190859
51	Nubosidad	T_med	-0.170429
50	Presion	Dir_viento	-0.142737
49	Dir_viento	T_med	-0.139570
48	Mortal	Costa	-0.122836
47	Costa	Dir_viento	-0.116198
46	T_med	Mes	-0.107477
45	Presion	IdPronostico	-0.101316
44	IdPronostico	Nubosidad	-0.097537
43	Presion	Mortal	-0.095238
42	Presion	Nubosidad	-0.090571
41	Precipitaciones	Presion	-0.084038
40	Costa	IdPronostico	-0.082045
39	IdPronostico	Precipitaciones	-0.079111
38	Mes	Dir_viento	-0.070916
37	V_Viento	Presion	-0.059304
36	V_Viento	Mortal	-0.056414
35	Nubosidad	Edad	-0.051773
34	Nubosidad	Mortal	-0.047986

Tabla 14. Primeras 20 correlaciones ordenadas de manera ascendente.

Meses de otoño:

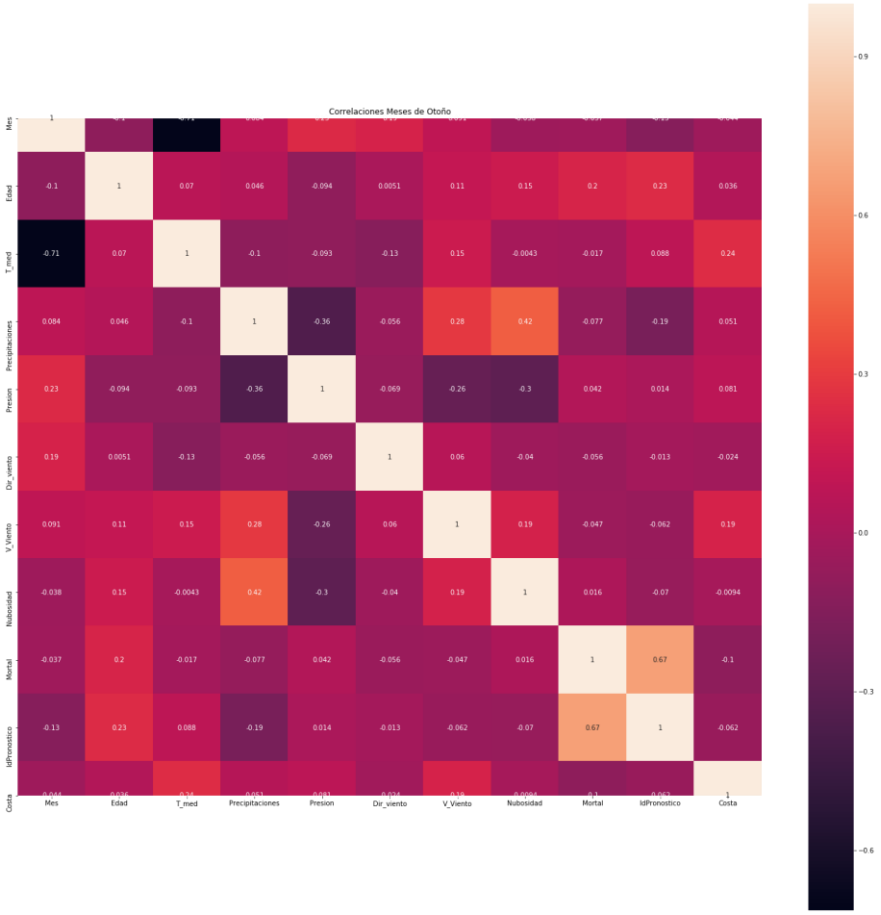

Ilustración 31. Mapa de calor de las correlaciones entre las diferentes variables asociadas a los meses de otoño.

	Variable1	Variable2	correlacion
0	IdPronostico	Mortal	0.670378
1	Precipitaciones	Nubosidad	0.423887
2	Precipitaciones	V_Viento	0.284999
3	T_med	Costa	0.241139
4	Edad	IdPronostico	0.233224
5	Mes	Presion	0.227230
6	Mortal	Edad	0.199542
7	Mes	Dir_viento	0.193610
8	Costa	V_Viento	0.191731
9	V_Viento	Nubosidad	0.187765
10	Nubosidad	Edad	0.148713
11	V_Viento	T_med	0.146440
12	V_Viento	Edad	0.114693
13	V_Viento	Mes	0.091118
14	T_med	IdPronostico	0.088401
15	Precipitaciones	Mes	0.083519
16	Presion	Costa	0.080982
17	Edad	T_med	0.070126
18	Dir_viento	V_Viento	0.059950
19	Costa	Precipitaciones	0.051298
20	Edad	Precipitaciones	0.045871

Tabla 16 Primeras 20 correlaciones ordenadas de manera descendente.

	Variable1	Variable2	correlacion
54	T_med	Mes	-0.713799
53	Presion	Precipitaciones	-0.356474
52	Presion	Nubosidad	-0.298536
51	Presion	V_Viento	-0.258653
50	Precipitaciones	IdPronostico	-0.187601
49	Mes	IdPronostico	-0.132687
48	Dir_viento	T_med	-0.129140
47	Precipitaciones	T_med	-0.101554
46	Costa	Mortal	-0.100742
45	Edad	Mes	-0.100245
44	Presion	Edad	-0.094362
43	Presion	T_med	-0.092669
42	Precipitaciones	Mortal	-0.077356
41	Nubosidad	IdPronostico	-0.069923
40	Dir_viento	Presion	-0.068903
39	V_Viento	IdPronostico	-0.062297
38	Costa	IdPronostico	-0.061796
37	Precipitaciones	Dir_viento	-0.056302
36	Mortal	Dir_viento	-0.055910
35	V_Viento	Mortal	-0.046537
34	Mes	Costa	-0.043645

Tabla 157. Primeras 20 correlaciones ordenadas de manera ascendente.

Anexo II: Clustering

K=6

Tmed	Precipitaciones	Presion	Dir_viento	V_Viento	Nubosidad
15.5373418	6.66075949	892.721519	272.493671	9.91772152	0.72863924
18.7116129	2.20387097	881.935484	87.5929975	9.20645161	0.76451613
18.836106	2.70219378	1017.05192	308.313757	11.2001828	0.56215722
20.4682979	2.76276596	1015.6173	225.385461	11.1262411	0.56134752
21.9531558	1.48402367	1016.33738	135.708454	10.0433925	0.53180473
22.0191203	2.3840562	1017.85089	50.5390959	12.1557728	0.50458155

Tabla 18. Centroides para K=6.

K=7

Tmed	Precipitaciones	Presion	Dir_viento	V_Viento	Nubosidad
15.5373418	6.66075949	892.721519	272.493671	9.91772152	0.72863924
18.6202614	2.23267974	881.372549	86.5098039	9.31372549	0.76143791
18.6875792	3.04296578	1017.4275	321.222117	10.7224335	0.5524398
20.0241002	2.59037559	1015.63795	247.420383	11.7503912	0.57609546
21.2346193	2.1986802	1015.54173	173.940646	9.92994924	0.53236041
21.6997391	2.95008696	1017.90557	38.6508696	12.2956522	0.51043478
22.6367539	1.03612565	1017.46346	94.6	11.1287958	0.51335079

Tabla 19. Centroides para K=7.