Tratamiento y Análisis de Datos del Registro de la Flota Pesquera

Julio Úbeda Quesada y Lucas Zamora Vera

Imagen creada con https://openart.ai/home.

1. Contexto

Este conjunto de datos recoge información técnica y administrativa de los buques registrados en el Registro de la Flota Pesquera Española, a partir de identificadores únicos (CFR) extraídos del registro europeo. Cada buque está identificado de forma unívoca, lo que garantiza la trazabilidad y evita duplicidades.

Los datos cubren un período comprendido entre el 31 de diciembre de 1987 y el 25 de marzo de 2025, lo que permite realizar un análisis detallado y evolutivo de la flota pesquera española durante casi cuatro décadas. Esta información es fundamental para comprender la composición, características técnicas, distribución geográfica y evolución operativa de la flota.

Es importante resaltar que con el fin de evitar redundancias con respecto a la práctica anterior, en la que ya se aplicó una preselección de los datos, se ha integrado el contenido del apartado 2 del enunciado de la práctica en este primer punto. En dicha práctica se filtraron los registros para incluir únicamente los buques españoles con código CFR válido, lo que constituye la base del conjunto de datos actual.

Para el análisis se ha realizado una selección de las variables más relevantes, en función de los objetivos planteados: analizar la estructura de la flota, identificar patrones de distribución del número de buques entre modalidades de pesca, y detectar posibles tendencias históricas en la baja o modernización de buques.

Las variables incluidas en el conjunto de datos son las siguientes:

- **CFR**: Código de Referencia Comunitario (identificador único del buque).
- **Nombre**: Nombre registrado del buque.
- **IMO**: Número de identificación internacional del buque (si aplica).
- **IRCS**: Código de llamada internacional (señal de radio).
- **Matrícula**: Código de identificación nacional del buque.
- **Alta en RGFP**: Fecha de alta en el Registro General de la Flota Pesquera.
- **Estado**: Situación administrativa (activo, baja definitiva, etc.).
- **Eslora total**: Longitud total en metros.
- **Arqueo GT**: Arqueo bruto (volumen interior).
- **Potencia**: Potencia del motor en kilovatios (y caballos de vapor entre paréntesis).
- **Material del casco**: Tipo de material constructivo (ej. madera, acero).
- **Puerto base**: Localidad y provincia de registro.
- **Administración responsable del Registro**: Organismo responsable del registro.
- **Censo por modalidad**: Tipo de arte o modalidad de pesca asignada (ej. cerco, palangre).
- **Capacidad del buque no aportable por**: Campo que indica si hay datos no disponibles.

- **Tipo de auxiliar**: Especifica si el buque es auxiliar y su tipo.

La selección de este subconjunto está justificada por su relevancia directa para los objetivos del análisis y por la calidad y disponibilidad de los datos. Estas variables permiten realizar distintos tipos de análisis: técnicos, geográficos, históricos o funcionales.

A continuación, se presenta una muestra de los datos seleccionados y limpiados para facilitar una visión general de las variables de interés, sus rangos y su tipología.

Column	Data Type	Null Count	Unique	mean	min	max	Date Min	Date Max
cfr	object	0	27364	NaN	NaN	NaN	NaT	NaT
nombre	object	0	19587	NaN	NaN	NaN	NaT	NaT
fc_alta_rgfp	date	0	2538	NaN	NaN	NaN	1987-12-31	2025-03-25
eslora_total	float	0	3922	11.50	1.00	116	NaT	NaT
arqueo_gt	float	0	6852	45.45	0.10	4406	NaT	NaT
material_casco	object	0	8	NaN	NaN	NaN	NaT	NaT
potencia_kw	float	0	3025	107.84	0.74	5851.63	NaT	NaT
estado_rgfp	object	0	4	NaN	NaN	NaN	NaT	NaT
fc_estado	date	0	5055	NaN	NaN	NaN	1987-12-31	2025-04-02
Puerto	object	0	287	NaN	NaN	NaN	NaT	NaT
Provincia	object	0	24	NaN	NaN	NaN	NaT	NaT
Comunidad Autónoma	object	0	12	NaN	NaN	NaN	NaT	NaT
Tipo de Arte	object	0	7	NaN	NaN	NaN	NaT	NaT
Edad_buque	int	0	38	17.31	0.00	37	NaT	NaT

Este análisis está diseñado para responder a preguntas como:

- ¿Cómo ha evolucionado la flota pesquera española en términos de número de buques desde los años 80 hasta hoy?
- ¿Cómo ha evolucionado la flota pesquera española en términos de eslora, arqueo y potencia desde los años 80 hasta hoy?
- ¿Existen patrones asociados a la baja de buques según sus características (tamaño, antigüedad, material, etc.)?
- ¿Existen patrones asociados a la pertenencia de un censo por modalidad según sus características (tamaño, antigüedad, material, comunidad autónoma en la que está registrado, etc.)?

El estudio es de gran utilidad tanto para investigadores como para responsables de políticas públicas, ya que permite identificar tendencias, evaluar la sostenibilidad del esfuerzo pesquero (expresado en tamaño, potencia y capacidad de los buques) y orientar decisiones en materia de ordenación y modernización del sector.

Además de esto, se realizarán modelos de aprendizaje automático con el objetivo de:

- Clasificar buques en un tipo de arte de pesca concreto según sus características técnicas y administrativas.
- Segmentar la flota en diferentes agrupaciones dependiendo de sus características técnicas.

2. Preparación y limpieza inicial de los datos

2.1. Revisión de duplicados

No se han encontrado valores duplicados en el dataset. Para verificarlo, se ha comprobado la aparición de los códigos CFR (Community Fishing Fleet Register), identificadores únicos de cada buque en el ámbito europeo. Dado que cada CFR es único, pueden utilizarse como clave primaria en nuestra tabla.

2.2. Identificación de caracteres especiales, valores vacíos y/o nulos

Las columnas con valores problemáticos son: ‘Tipo de auxiliar’, ‘Capacidad del buque no aportable por:’, ‘IMO’, ‘IRCS’, ‘Censo por modalidad’, ‘Material del casco’, ‘Potencia’, ‘Arqueo GT’ y ‘Eslora total’.

- La columna ‘Tipo de auxiliar’ está completamente vacía (valores nulos) o contiene caracteres especiales (como guiones), por lo que puede eliminarse directamente.

- ‘Capacidad del buque no aportable por:’ tiene más del 99% de valores nulos, por lo que también puede eliminarse.
- En ‘IMO’ y ‘IRCS’ son dos identificadores del buque que contienen casi en su totalidad (> 80 %) carácteres especiales (guiones principalmente) sin aportar información, por lo que podrían eliminarse. Nos quedaremos con el CFR como identificador único del buque.
- ‘Censo por modalidad’ contiene unos 2297 registros con carácteres especiales, en concreto (-) y un valor nulo. Puesto que esta columna será procesada para obtener el arte, dichos registros serán agrupados en una categoría de 'Otros', puesto que un buque debe estar registrado en un censo.
- ‘Material del casco’ muestra muy pocos registros con carácteres especiales, en concreto (-) y se procederá a su agrupación en una categoría 'Otros'.
- Las columnas ‘Potencia’, ‘Arqueo GT’ y ‘Eslora total’ muestran en torno a 2000 registros con valor 0, lo cual parece ilógico (un buque no puede tener una eslora total de 0 m, ni una capacidad o arqueo de 0 toneladas, ni una potencia del motor de 0 kw).

En los siguientes apartados, estos valores serán imputados según el método especificado.

2.3. Simplificación del nombre de las columnas

Se han simplificado y estandarizado los nombres de las columnas del dataset para facilitar la escritura y lectura del código al eliminar espacios, tildes y caracteres especiales, permitiéndonos trabajar más eficientemente con funciones automatizadas o en bucles. Además, con esto se mejora la legibilidad, reduce errores y hace que el código sea más limpio, profesional y fácil de mantener, especialmente en proyectos colaborativos como el presente.

La simplificación ha sido:

- 'CFR': 'cfr',
- 'Nombre': 'nombre',
- 'Matrícula': 'matrícula',
- 'Alta en RGFP': 'fc_alta_rgfp',
- 'Estado': 'estado',
- 'Eslora total': 'eslora_total',
- 'Arqueo GT': 'arqueo_gt',
- 'Potencia': 'potencia',
- 'Material del casco': 'material_casco',
- 'Puerto base': 'puerto_base',
- 'Administración responsable del Registro': 'admin_registro',
- 'Censo por modalidad': 'censo_modalidad',
- 'Tipo de auxiliar' : 'tipo_auxiliar',

- 'Capacidad del buque no aportable por:' : 'capacidad_no_aportable',
- 'Material del casco' : 'material_casco',
- 'IMO' : 'IMO',
- 'IRCS': 'IRCS'

2.4. Corrección de errores de tipo de dato (1)

La limpieza y conversión de las columnas ‘eslora_total’, ‘arqueo_gt’ y ‘potencia’ fueron necesarias para garantizar una correcta interpretación y análisis de los datos. Originalmente, estas columnas estaban almacenadas como texto (object) debido a varias inconsistencias en su formato. Por ejemplo, ‘eslora_total’ y ‘arqueo_gt’ incluían unidades de medida como 'm' o caracteres no numéricos, y usaban comas , como separador decimal en lugar del punto (.) que Python espera para tratar valores numéricos. Esto impedía realizar cálculos o análisis estadísticos directamente.

En el caso de la columna ‘potencia’, el problema era aún más complejo: contenía información duplicada en diferentes unidades, como '202,26 kW (275,0 CV)', lo que requería extraer y estandarizar los valores en una sola unidad (por ejemplo, kilovatios).

2.5. Corrección de errores de tipo de dato (2)

La conversión de la columna ‘alta_rgfp’ a tipo fecha es necesaria porque originalmente los valores estaban almacenados como texto (object), lo que impedía aprovechar las funcionalidades específicas que ofrece Python para trabajar con fechas. Este campo representa la fecha de alta en el Registro General de la Flota Pesquera, una información temporal clave para análisis cronológicos, cálculos de antigüedad, o filtrado por rangos de fechas.

3. Transformación estructural del dataset

3.1. Descomposicion de columnas

3.1.1. ‘estado’ en dos: ‘estado_rgfp’ y ‘fc_estado’

La separación de la columna ‘estado’ en dos columnas distintas, ‘estado_rgfp’ y ‘fc_estado’, es necesaria para mejorar la claridad, la estructura y la utilidad de la información contenida en esa variable.

Originalmente, la columna ‘estado’ combinaba dos tipos de datos diferentes: por un lado, una categoría del estado administrativo del buque (por ejemplo, “Baja Definitiva”) y, por otro, una fecha asociada a ese estado (por ejemplo, “(desde el 02/12/1998)”). Esta mezcla dificulta el análisis automatizado, ya que impide utilizar la fecha como un dato temporal y el estado como una etiqueta categórica independiente.

3.1.2. ‘puerto_base’ en cuatro: ‘codigopostal’, ‘puerto’, ‘provincia’ y ‘comunidad atutonoma’

En ‘puerto_base’, varios datos (código postal, puerto, provincia y comunidad autónoma) estaban combinados en una sola cadena. Esto impide realizar análisis geográficos detallados o agrupar los registros por nivel administrativo. Separar estos elementos en columnas individuales permite normalizar la información y facilitar operaciones como filtrado, agrupación o visualización por comunidad o provincia.

3.2. Estandarización de categorías

En Censo por modalidad, cada valor combinaba el tipo de arte de pesca con la zona de actividad, sin un formato homogéneo. Clasificar estos censos en categorías estandarizadas como "Arrastre", "Palangre", "Cerco", etc., permite agrupar modalidades distintas bajo criterios comunes, simplificando el análisis técnico y operativo de la flota.

4. Tratamiento de valores anómalos y calidad de datos

4.1. Identificación y gestión de valores extremos

4.1.1. Imputación de los valores 0 e ilógicos

Se ha realizado una imputación en las variables numéricas que contienen valores cero, reemplazándolos por valores nulos, ya que un buque no puede tener eslora 0 m, arqueo 0 toneladas o potencia 0 kW; estos ceros probablemente se deben a la antigüedad de los datos y a limitaciones en la recolección original. Para la imputación utilizamos el método de los k vecinos más cercanos con k=15, lo que nos brinda mayor robustez frente a valores atípicos.

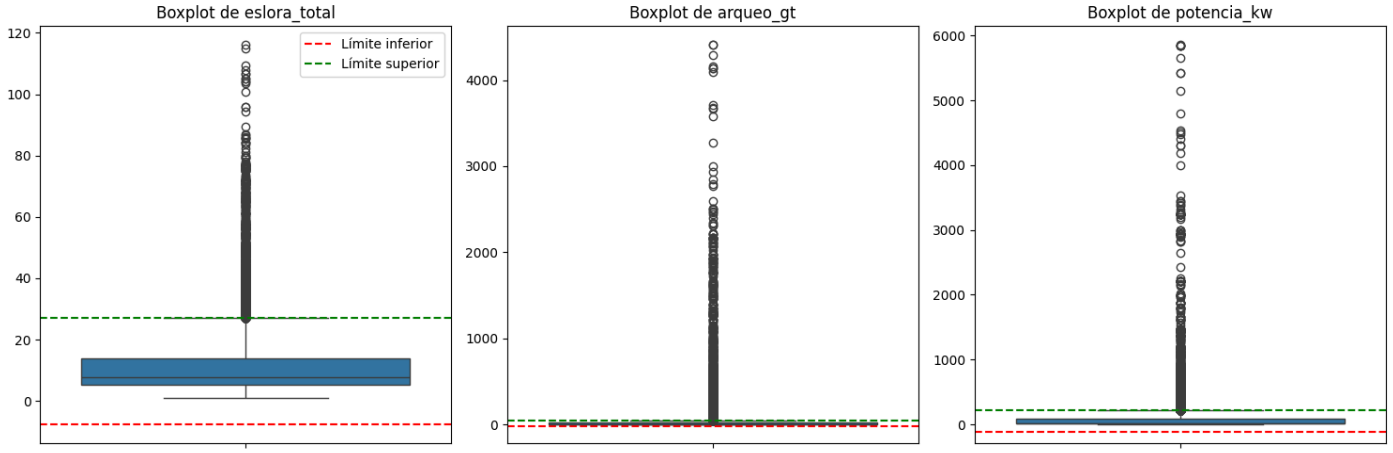
Optamos por k=15 porque con un número bajo de vecinos, como 3, la imputación puede distorsionarse si alguno de esos vecinos es un outlier. Al aumentar a 15 vecinos, el impacto de valores atípicos se diluye al promediar más observaciones, captando mejor la tendencia general y suavizando la variabilidad extrema. Además, dado que contamos con más de 27,000 registros, este valor de k es razonable y evita sesgos o sobre ajustes.

Tras la imputación, los valores cero y NaN desaparecieron por completo en las columnas tratadas, y las estadísticas descriptivas muestran que las medias, medianas y desviaciones estándar se mantienen muy similares antes y después (eslora_total: (Media original: 11.481; Media imputada: 11.501; Mediana original: 7.500; Mediana imputada: 7.600; Desviación estándar original: 9.986; Desviación estándar imputada: 9.856); arqueo_gt: (Media original: 44.236; Media imputada: 45.424; Mediana original: 2.100; Mediana imputada: 2.310; Desviación estándar original: 180.621; Desviación estándar

imputada: 179.309); potencia_kw: (Media original: 118.664; Media imputada: 107.847; Mediana original: 29.410; Mediana imputada: 22.060; Desviación estándar original: 274.204; Desviación estándar imputada: 262.285)), salvo en potencia_kw, donde la imputación suavizó la distribución. Esto confirma que el método preserva los patrones multidimensionales y la estructura latente del dataset, logrando imputaciones consistentes y realistas dentro del contexto general.

4.1.2. Detección de *outliers*

Los siguientes tres boxplots para las variables 'eslora_total', 'arqueo_gt' y 'potencia_kw', junto con líneas que marcan los límites para identificar outliers. En 'eslora_total', la mayoría de los buques miden menos de 25 metros, pero hay muchos valores por encima del límite superior, indicando embarcaciones atípicamente grandes. En 'arqueo_gt', la distribución está centrada en valores bajos, aunque se observan numerosos outliers, algunos por encima de 4000 GT, reflejando un pequeño grupo de buques con gran capacidad. En 'potencia_kw', el patrón se repite: predominan potencias bajas, pero existen muchos valores extremos, incluso mayores a 5000 kW. Estos outliers, aunque válidos, deben considerarse en futuros análisis, ya que pueden afectar la interpretación estadística de la flota.



4.2. Eliminación de columnas no relevantes

Se eliminaron las columnas 'IMO', 'IRCS', 'matricula', 'estado', 'potencia', 'puerto_base', 'admin_registro', 'censo_modalidad', 'capacidad_no_aportable', 'tipo_auxiliar' y 'Código Postal' por no aportar información relevante para el análisis actual.

5. Enriquecimiento y análisis exploratorio inicial

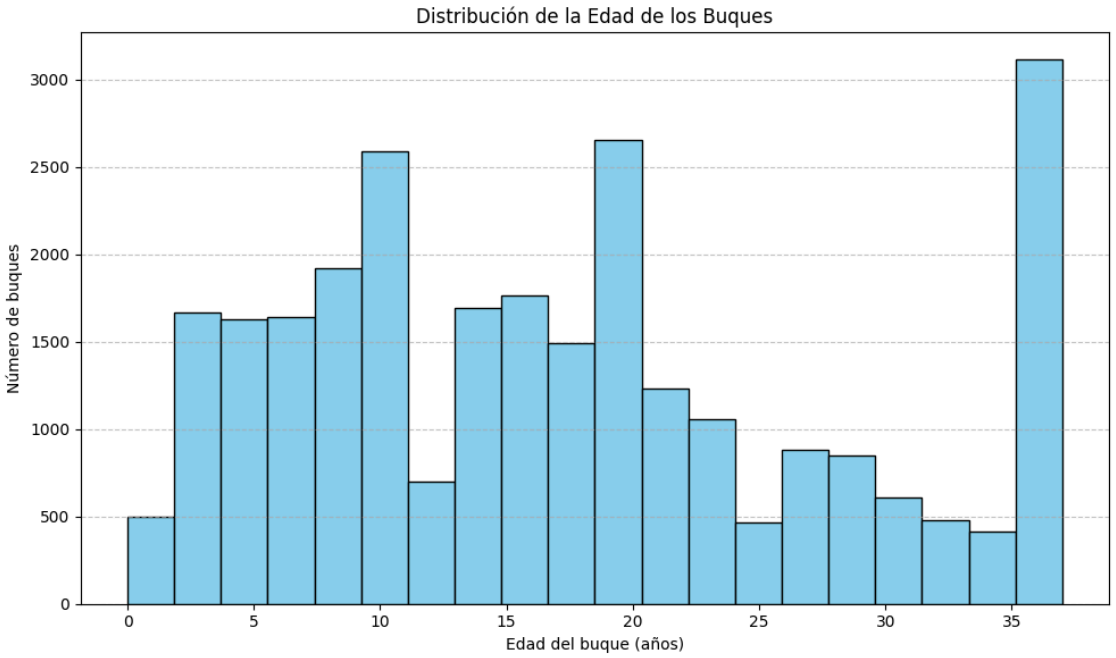
5.1. Creación de la variable 'edad_buque'

La variable ‘edad_buque’ ha sido creada con el objetivo de calcular la antigüedad de cada embarcación desde su fecha de alta en el registro hasta su fecha de baja definitiva (si aplica) o hasta la fecha actual, en caso de que siga activa. Esta variable permite analizar la edad operativa de los buques, clave para estudios del ciclo de vida o posibles reemplazos.

5.1.1. Distribución de los buques por su edad

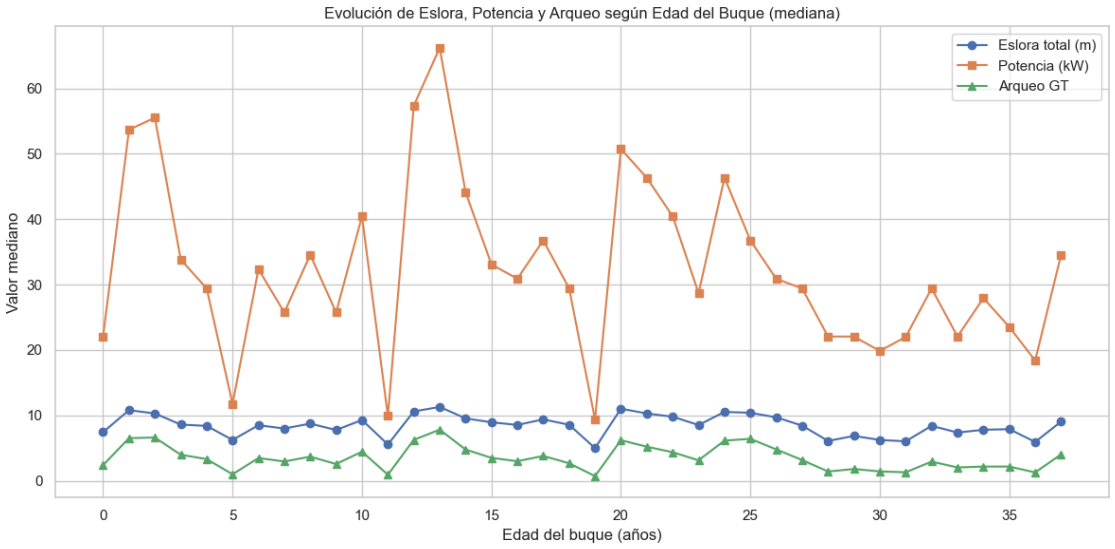
A partir del gráfico de distribución de la edad de los buques, se observa que una parte significativa de la flota pesquera española está compuesta por buques con más de 30 años, lo que indica que muchos fueron construidos en los años 80. Sin embargo, también se aprecia una presencia notable de buques más recientes, especialmente en tramos de 0 a 10 años, lo que sugiere procesos de renovación. Estos datos permiten inferir que, aunque la flota ha envejecido, ha habido cierta modernización en las últimas décadas.

Las variaciones en el número de nuevos buques pueden estar relacionadas con eventos como la adhesión de España a la Comunidad Económica Europea en 1986 y la consiguiente adaptación a la Política Pesquera Común, que impulsó planes de reducción de flota y subvenciones para el desguace de buques (Europa-azul). Además, las reformas de la PPC y las restricciones medioambientales y económicas posteriores, como la crisis de 2008, podrían haber limitado la incorporación de nuevas unidades.



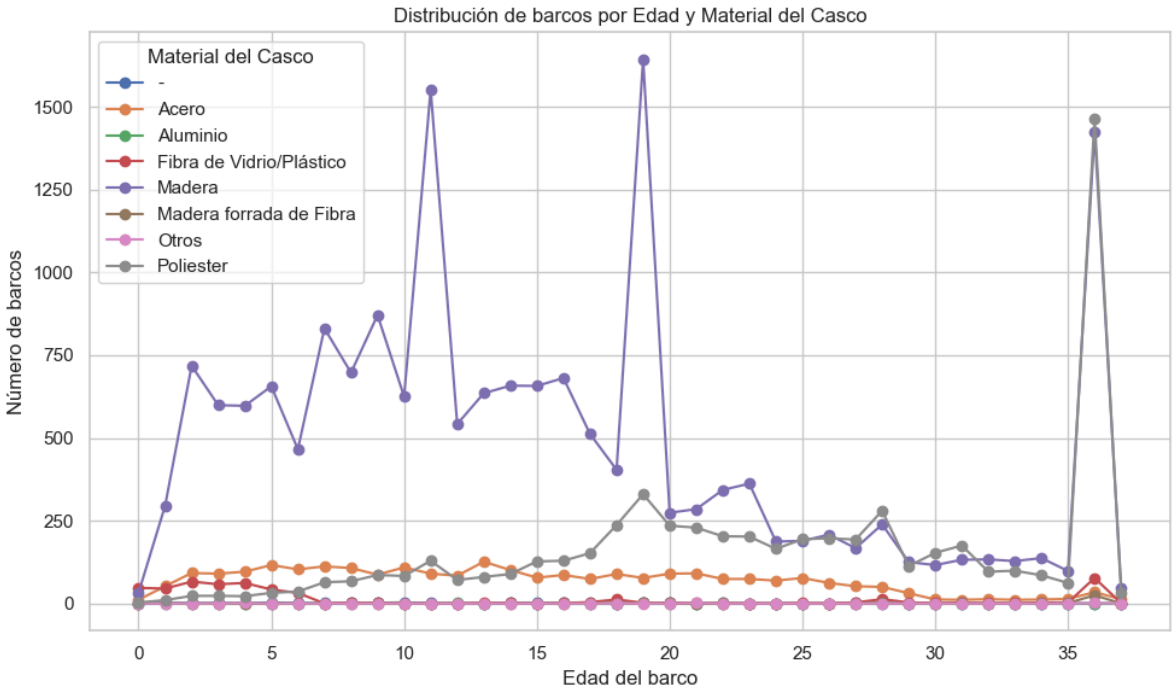
5.2. Evolución temporal de la eslora, arqueo, potencia y material del casco de los buques

El gráfico muestra la evolución de las características físicas de la flota pesquera española (eslora, arqueo, potencia y material del casco) a lo largo del tiempo, representadas en función de la edad de los buques. En términos generales, se observa que los buques más nuevos (menor edad) tienden a tener una eslora y arqueo ligeramente superiores, lo que indica una tendencia hacia embarcaciones algo más grandes y con mayor capacidad de carga. Sin embargo, es la potencia la que muestra más variabilidad, con picos elevados entre los buques más recientes, lo que sugiere una modernización orientada al rendimiento y la eficiencia operativa. Esta evolución puede estar relacionada con avances tecnológicos y con regulaciones que favorecen una flota más especializada y adaptada a las exigencias actuales del sector. Por ejemplo, el aumento de la potencia auxiliar permitió en 2008 el uso de redes más grandes, haciendo posible la pesca a grandes profundidades (Eur-lex-europa). Esta serie de subvenciones se podrían ver reflejadas en el pico más alto del gráfico para la potencia.



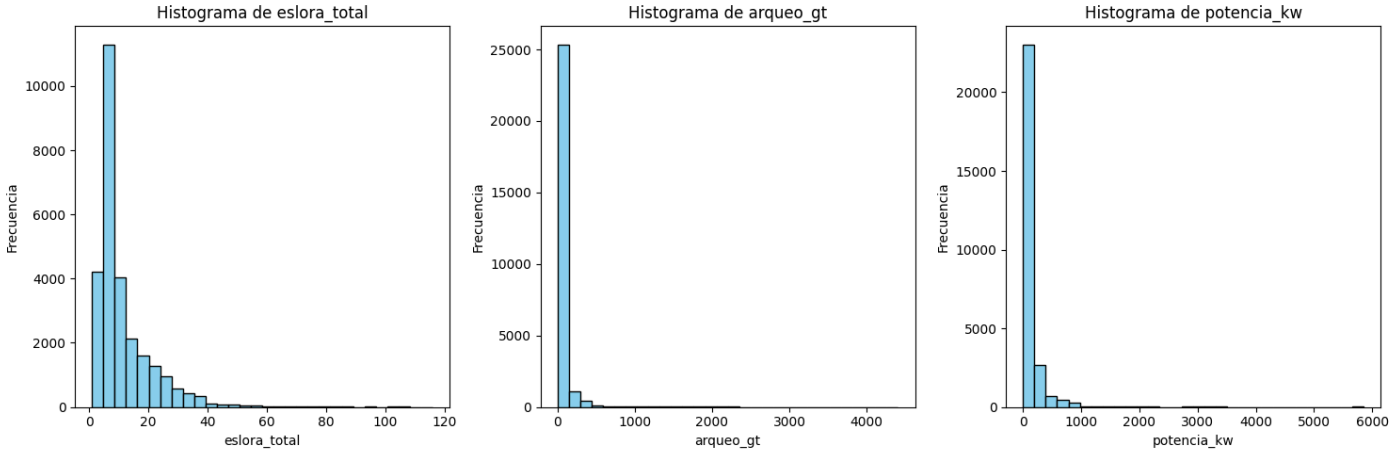
En cuanto al material del casco, podemos observar un aumento del número de buques con casco de madera en los últimos 20 años, si comparamos estos picos con las caidas en la potencia, arqueo y eslora del gráfico anterior vemos que estos picos de buques con casco de madera sugieren la entrada en la flota de una gran cantidad de buques de pequeño tamaño hace 18 y 12 años, siendo estos los que predominan a día de hoy.

El resto de materiales de casco utilizados en la flota se ha mantenido estable durante el tiempo a excepción de los buques con casco de poliester, que son los predominantes entre 35 y 40 años de antigüedad, y desde 20 años hasta ahora, han ido reduciendo en número, lo que implica la mejora en materiales utilizados y la modernización en la flota, además de la aparición de nuevos materiales como la fibra de vidrio, material que ha sufrido un incremento en número de buques en los últimos 5 años.



5.3. Distribución en las columnas numéricas

Los siguientes tres histogramas correspondientes a las variables 'eslora_total', 'arqueo_gt' y 'potencia_kw', muestran distribuciones fuertemente sesgadas hacia la derecha. En el primer gráfico, la mayoría de las embarcaciones tienen una 'eslora_total' entre 0 y 20 metros, con un pico entre 5 y 10 metros y una larga cola que alcanza los 120 metros. El 'arqueo_gt', que representa el volumen interno, también se concentra en valores bajos (principalmente debajo de 500 GT) con algunos valores atípicos que superan los 4000 GT. De forma similar, la 'potencia_kw' se agrupa mayoritariamente por debajo de los 500 kW, aunque existen casos que superan los 6000 kW. Estas distribuciones indican una predominancia de embarcaciones pequeñas o medianas, con algunos pocos casos extremos, posteriormente se aplicaran transformaciones como la logarítmica, útiles para los análisis.



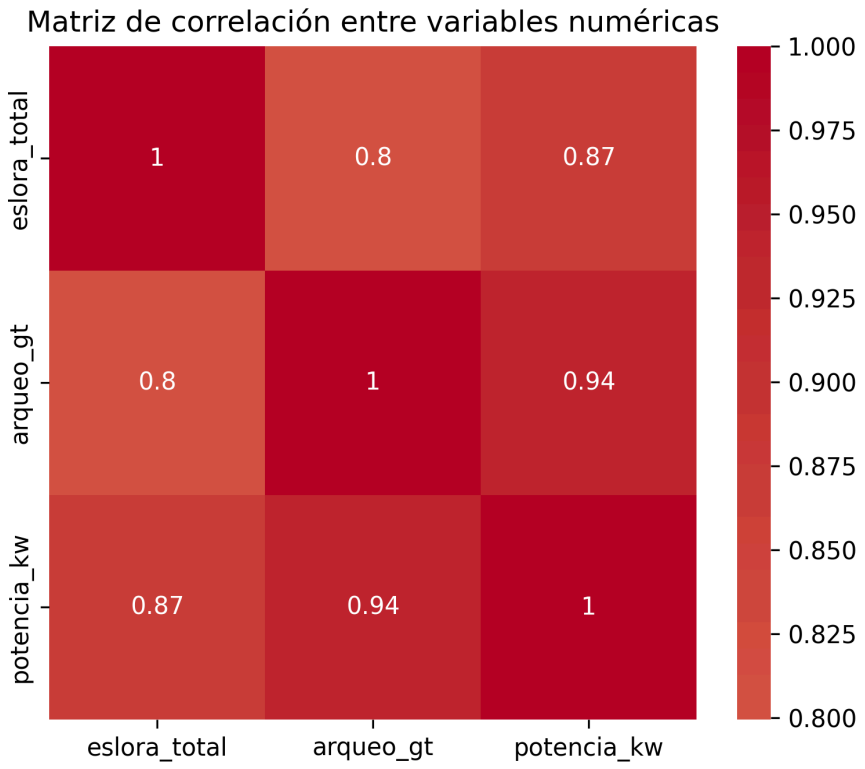
En el caso de la distribución de la edad de los buques representada anteriormente, ésta presenta picos en los 35, 20 y 10 años correspondientes a la renovación de la flota.

5.4. Correlaciones

La matriz de correlación muestra relaciones fuertes y positivas entre las tres variables numéricas analizadas:

- 'Eslora total' y 'arqueo_gt' tienen una correlación de aproximadamente 0.80, lo que indica que a medida que aumenta la eslora (longitud) del buque, también tiende a aumentar su arqueo (tonelaje). Esto es esperable porque barcos más largos suelen tener mayor capacidad o peso.
- 'Eslora total' y 'potencia_kw' presentan una correlación aún más alta, alrededor de 0.87, lo que sugiere que buques más largos también suelen contar con motores de mayor potencia, reflejando la necesidad de impulsar embarcaciones de mayor tamaño.
- La relación entre 'arqueo_gt' y 'potencia_kw' es la más fuerte, con un valor de 0.94, lo que indica que el peso o volumen del buque está muy ligado a la potencia instalada. Esto es coherente, pues un mayor arqueo implica barcos más pesados que requieren motores más potentes para su operación eficiente.

En conjunto, estas correlaciones confirman que las variables están estrechamente relacionadas y reflejan dimensiones físicas y de rendimiento del buque que crecen conjuntamente. Esta consistencia también respalda la calidad de la imputación, ya que mantiene las relaciones naturales esperadas entre estas características.



6. Análisis de los datos

6.1 Contraste de hipótesis:

Se realizaron distintos contrastes de hipótesis con el objetivo de responder a las siguientes preguntas:

- ¿Existe una relación significativa entre las comunidades autónomas y el tipo de arte de pesca de los buques?

- ¿Está relacionada la eslora del buque con el tipo de arte de pesca utilizado?

6.1.1 Relación entre comunidades autónomas y tipo de arte.

Dado que en el conjunto de datos predominan las embarcaciones pequeñas (alrededor del 90% de la flota), cuyo tipo de arte habitual son las artes menores, se decidió excluir esta categoría del análisis. Por tanto, la pregunta que se plantea es:

¿Existe una relación estadísticamente significativa entre la comunidad autónoma a la que pertenece el buque y su tipo de arte, excluyendo las artes menores?

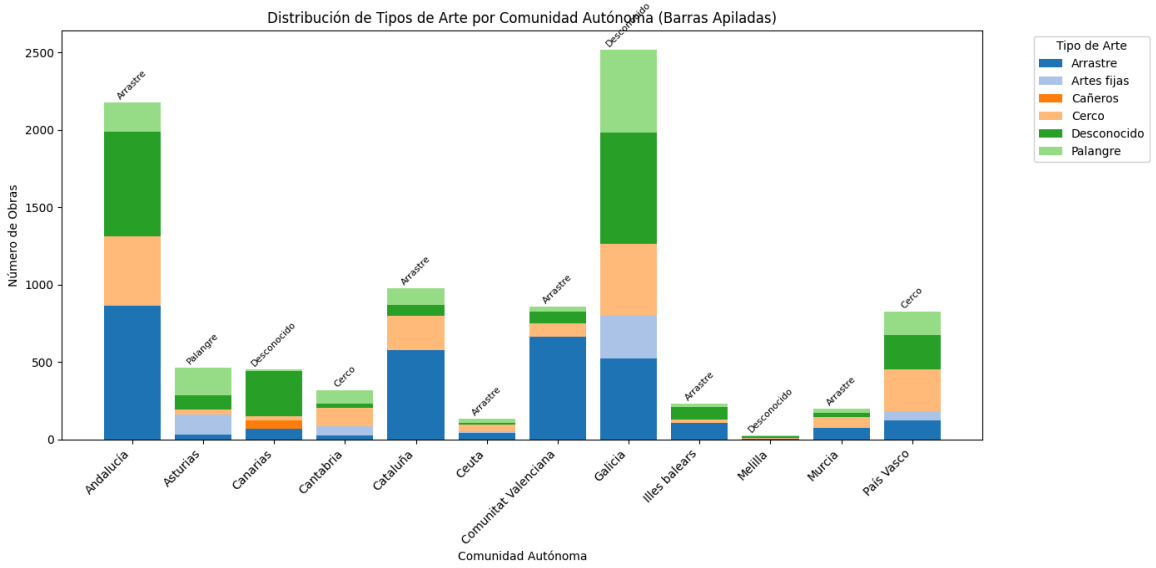
Para evaluar esta posible relación entre dos variables categóricas, se aplicó un test de independencia de chi-cuadrado. Las hipótesis formuladas fueron:

- $\mathbf{H}_0$ (hipótesis nula): El tipo de arte es independiente de la comunidad autónoma.

- $\mathbf{H}_1$ (hipótesis alternativa): El tipo de arte depende de la comunidad autónoma.

El resultado del test mostró un valor de $p < 0.05$, por lo que se rechaza la hipótesis nula.

Esto indica que, con un nivel de confianza del 95%, existe una relación significativa entre la comunidad autónoma del buque y el tipo de arte utilizado.



6.1.2 Relación entre el tamaño del buque y tipo de arte.

Ya que las variables eslora, arqueo y potencia están fuertemente relacionadas (a mayor eslora, mayor arqueo y mayor potencia), tomaremos la eslora como variable a contrastar en esta relación.

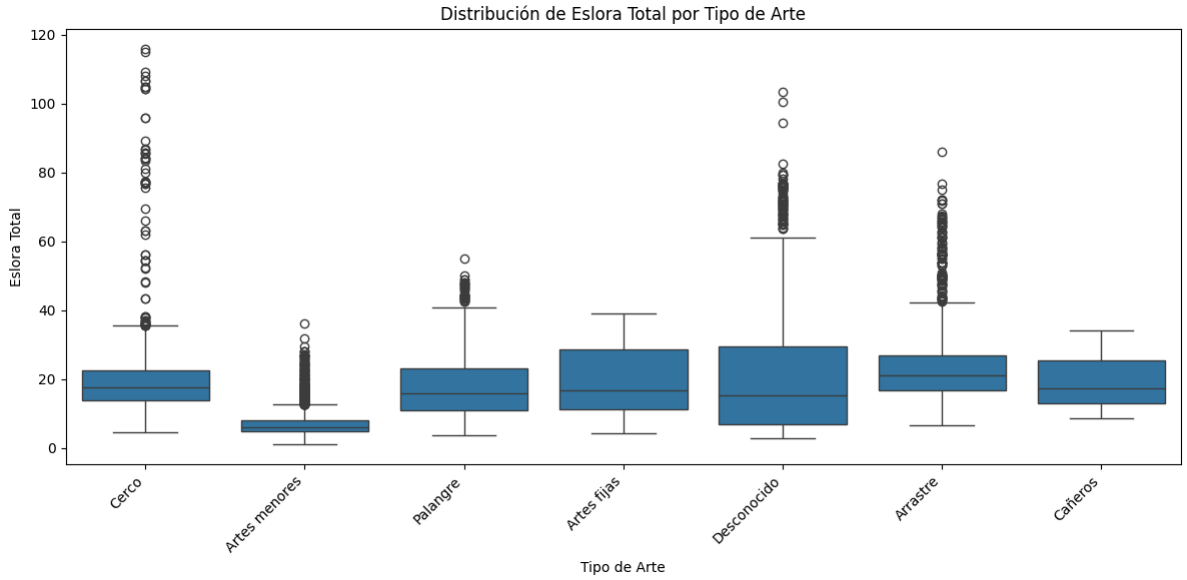
En este caso se analiza la relación entre una variable numérica (eslora) y una variable categórica (tipo de arte). Antes de aplicar un test de comparación de medias, es necesario comprobar si se cumple el supuesto de homocedasticidad (igualdad de varianzas entre grupos). Para ello, se aplicó el test de Levene, cuyo resultado fue un p-valor $< 10^{-16}$, indicando que no se cumple la homocedasticidad. Por tanto, se optó por utilizar el test no paramétrico de Kruskal-Wallis, que no requiere el supuesto de normalidad ni igualdad de varianzas.

Las hipótesis planteadas fueron:

- **H_0:** Todos los grupos (tipos de arte) tienen la misma distribución de eslora.
- **H_1:** Al menos un grupo presenta una distribución de eslora diferente.

El resultado del test de Kruskal-Wallis tuvo un p-valor $< 10^{-16}$, por lo que se rechaza la hipótesis nula. Esto sugiere que el tipo de arte está significativamente relacionado con la eslora del buque.

Esto se puede visualizar en el siguiente boxplot de distribución de eslora por tipo de arte:



Al analizar las distribuciones, se observan las siguientes particularidades:

- **Cerco**, **arrastre** y **arte desconocido** presentan las mayores esloras dentro de la flota.

- **Palangre** y **artes fijas** comparten distribuciones similares, aunque el palangre muestra una mayor presencia de embarcaciones de gran tamaño.

- **Artes fijas** y **cañeros** presentan distribuciones más compactas, sin presencia de valores atípicos (outliers).

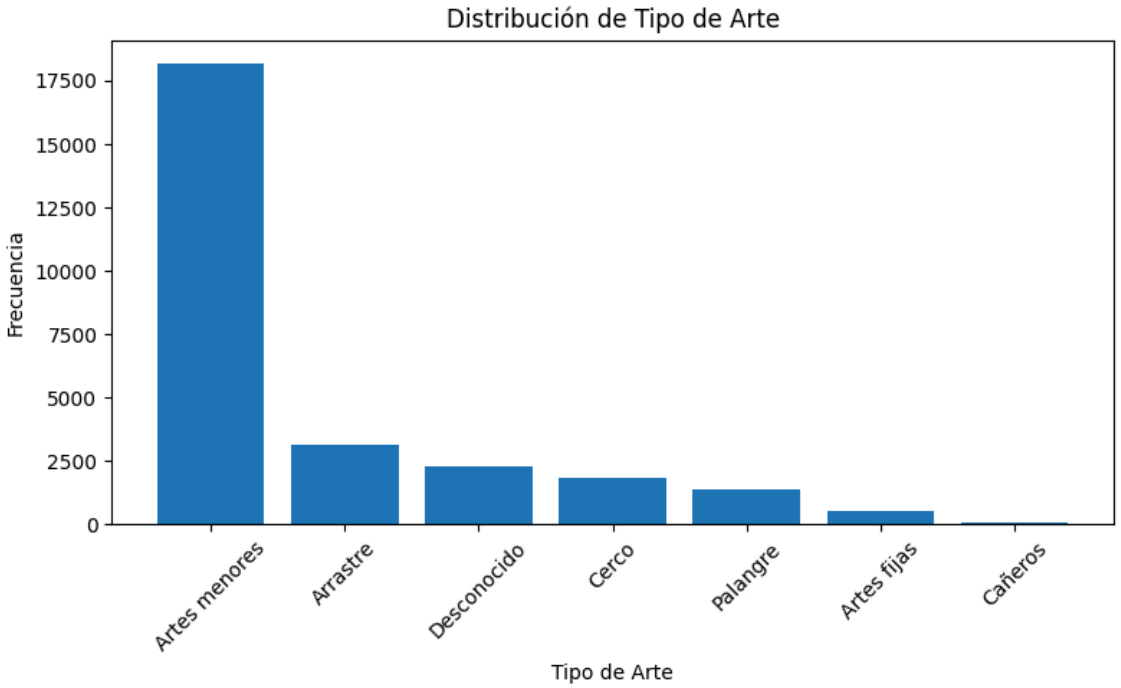
- En contraste, **artes menores** concentran los buques de menor tamaño de forma clara y diferenciada del resto de artes.
-

6.2 Modelado de los datos:

6.2.1 Modelo supervisado:

Ya que el dataset cuenta con los tipos de arte que utiliza cada buque, nuestro objetivo es crear un modelo de clasificación que pueda predecir el tipo de arte de pesca que utilizará un buque en base a sus características técnicas.

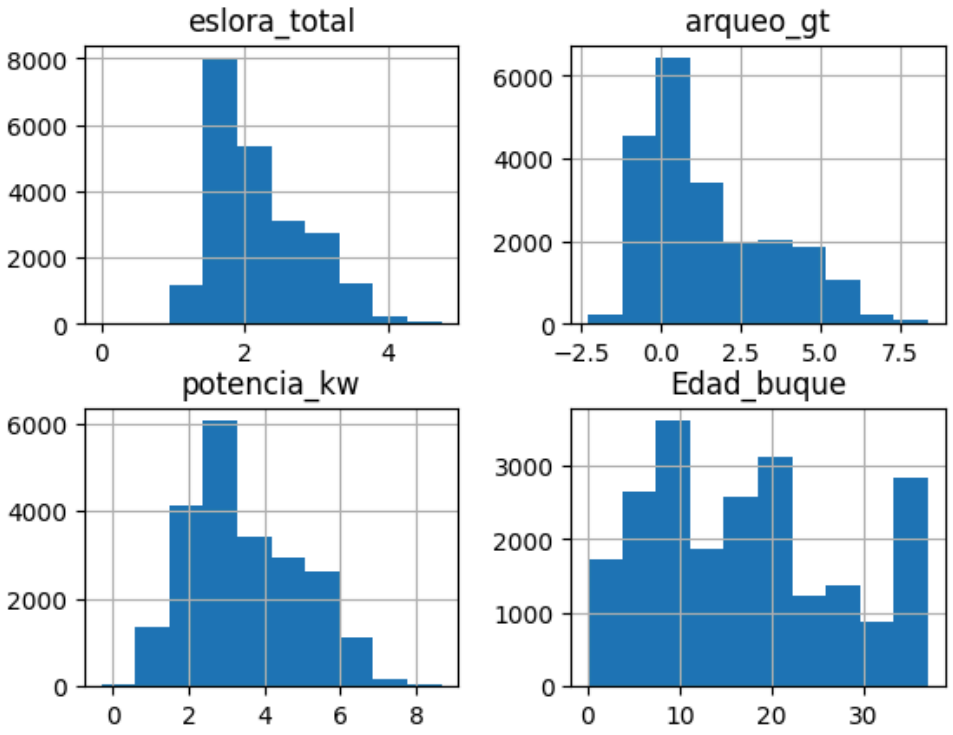
Para ello, tomamos “Tipo de Arte” como nuestra variable objetivo. Si la visualizamos podemos observar que se trata de un dataset muy desbalanceado ya que la mayor parte de la flota (aproximadamente un 90%) son buques pesqueros menores de 12 metros de eslora, los cuales están en la categoría de "Artes menores". Esto puede suponer un problema a la hora de entrenar los modelos por lo que ha de tenerse en cuenta a la hora de elegir las métricas a evaluar o aplicar técnicas de balanceo del target.



Una vez visualizada la variable objetivo, se divide el conjunto de datos en los subconjuntos de entrenamiento y testeo, siendo el conjunto de testeo el 20% del conjunto original y estando estratificado para mantener la misma distribución de la variable objetivo en el conjunto original.

Como hemos comprobado anteriormente, las variables numéricas correspondientes a eslora, arqueo y potencia presentan una distribución sesgada debido a la mayor cantidad de buques pequeños. Para normalizar estas distribuciones aplicamos una transformación logarítmica a estas variables, excluyendo la edad.

La distribución de las variables numéricas tras la transformación es la siguiente:



Ya que estas variables se presentan en diferentes unidades (metros, tonelaje bruto, kilovatios y años) se realiza un estandarizado de las variables para que la información de estas sea consistente, manteniendo una media de 0 y la proporción de la desviación estándar original, eliminando las unidades originales y convirtiéndose en medidas relativas.

En el caso de las variables categóricas, se transforma cada categoría a una subcategoría binaria con valor 0 y 1 o True y False.

	eslora_total	arqueo_gt	potencia_kw	Edad_buque	material_casco_-	material_casco_Acero	material_casco_Aluminio
1647	2.253812	2.350627	2.206875	-1.274535	False	True	False
16661	-0.437211	-0.295791	-0.467431	0.638309	False	False	False
13	1.813216	1.684972	1.484563	-0.605039	False	False	False
9166	1.692959	1.753181	1.757437	-1.465819	False	False	False
18564	-0.169584	-0.389223	-0.641094	-0.031186	False	False	False

Una vez realizadas estas transformaciones, se realiza un primer modelado con 4 algoritmos de clasificación basados en métodos de *bagging y boosting*. Estas técnicas reducen los sesgos y errores entrenando una combinación de modelos de árboles de

decisión, en paralelo en el caso del *bagging* o en serie en el caso del *boosting*. Estos modelos (Decision Tree, Random Forest, LightGBM y CatBoost) se entrenan con sus hiperparámetros por defecto y tomando como métrica de evaluación *balanced_accuracy* debido al desbalanceo de clases en la variable objetivo.

Tras este primer entrenamiento, y con las siguientes métricas de evaluación se toma el algoritmo LightGBM como modelo base para optimizar:

Modelo	**Métrica *(balanced_accuracy)***
Decision Tree	0.5940
Random Forest	0.6241
LightGBM	0.6657
Catboost	0.6379

Se realiza una optimización basada en *Grid Search Cross Validation,* en la que se prueban diferentes combinaciones de hiperparámetros para encontrar el mejor resultado, evaluando con validación cruzada. Tras esta optimización obtenemos los siguientes resultados medios:

Precision	Recall	Accuracy
0.72	0.64	0.88

Para evitar el sesgo hacia la clase mayoritaria, se realiza un segundo modelo utilizando la técnica de sobremuestreo SMOTE, que a diferencia de un sobremuestreo tradicional, en el que se duplican registros, este utiliza un algoritmo KNN para generar nuevos registros en base a las distancias de los vecinos más cercanos del registro original.

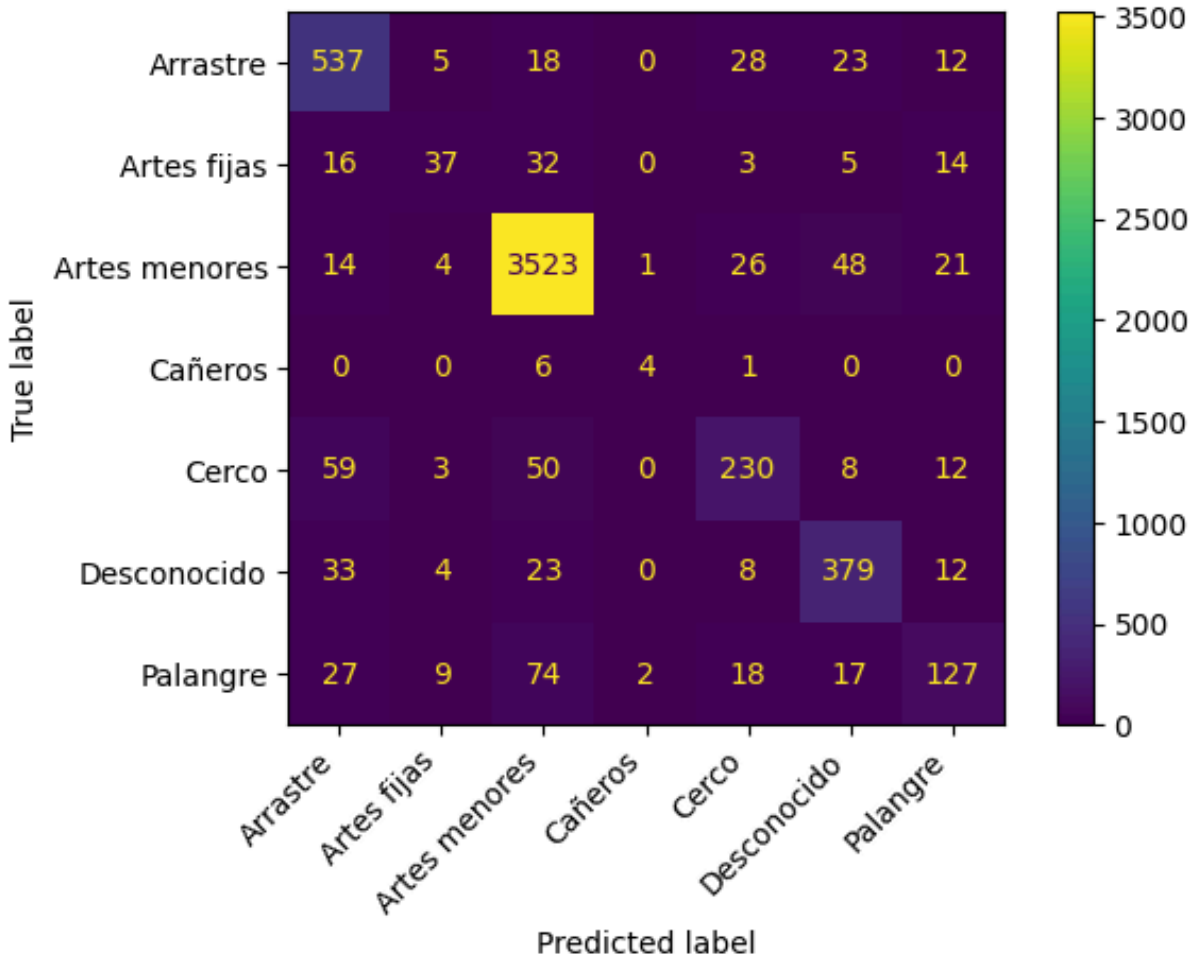
Tras realizar este sobremuestreo y volver a entrenar el modelo obtenemos las siguientes métricas:

Precision *(resampling)*	Recall *(resampling)*	Accuracy
0.67	0.71	0.87

Aunque ha habido una disminución en la precisión, la sensibilidad (recall) ha aumentado considerablemente, lo que supone una mejoría en la capacidad del modelo para identificar correctamente los casos positivos, reduciendo así la cantidad de falsos negativos.

Podemos ver el siguiente reporte de clasificación, y matriz de confusión para hacer una diagnosis del modelo y entender en qué casos funciona mejor o peor:

Arte	Precision	Recall	F1-Score
Arrastre	0.80	0.80	0.80
Artes fijas	0.50	0.52	0.51
Artes menores	0.96	0.95	0.95
Cañeros	0.41	0.64	0.5
Cerco	0.68	0.71	0.69
Desconocido	0.79	0.82	0.81
Palangre	0.53	0.52	0.53



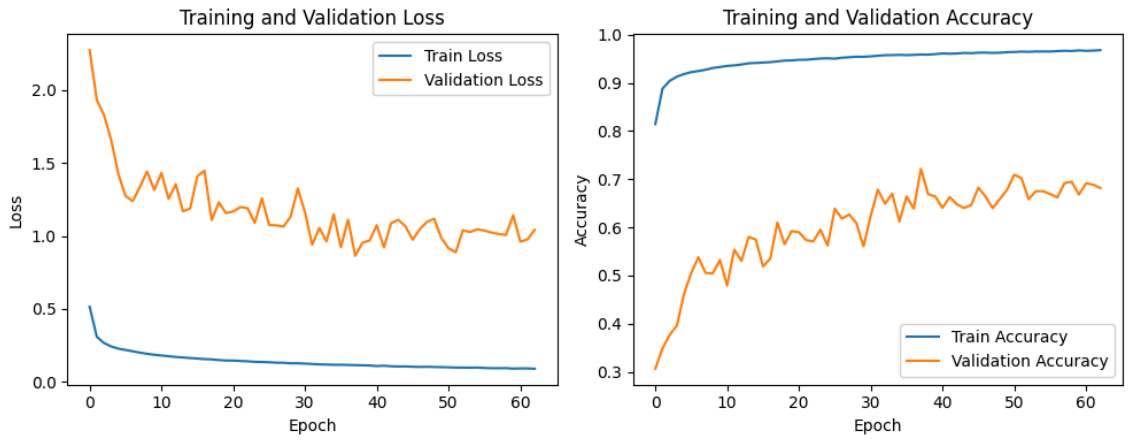
Tras ver el reporte y matriz de confusión, observamos que las clases en las que mejor clasifica el modelo son: arrastre, artes menores y desconocido (otros), y las clases en las que peor funciona son: cañeros, palangre y artes fijas.

Esto se debe a que los buques que realizan estos tipos de pesca cuentan con características técnicas muy similares, y un mismo buque podría realizar pesca con caña o palangre sin necesidad de modificación en sus características.

Para poder mejorar estas métricas, habría que incorporar nuevos datos, por ejemplo los datos de posición geográfica de las balizas GPS que poseen los buques, mediante los cuales se podrían obtener las trayectorias realizadas durante la pesca, las cuales sirven como "firma" del tipo de pesca que se está realizando. Lamentablemente estos datos no son públicos por lo que no podemos tener acceso a ellos.

Como última prueba de aprendizaje supervisado, realizamos un modelo de redes neuronales con capas densas *(deep learning)* utilizando una arquitectura en forma de pirámide invertida.

Arte	Precision	Recall	F1-Score
Arrastre	0.78	0.82	0.80
Artes fijas	0.38	0.59	0.46
Artes menores	0.95	0.93	0.94
Cañeros	0.43	0.55	0.48
Cerco	0.66	0.70	0.68
Desconocido	0.76	0.79	0.78
Palangre	0.44	0.35	0.39



Aunque hay una pequeña mejora en cuanto al *recall* en algunas clases, se pierde *precision* por lo que tomamos como modelo final el *LightGBM* entrenado con sobremuestreo de los datos.

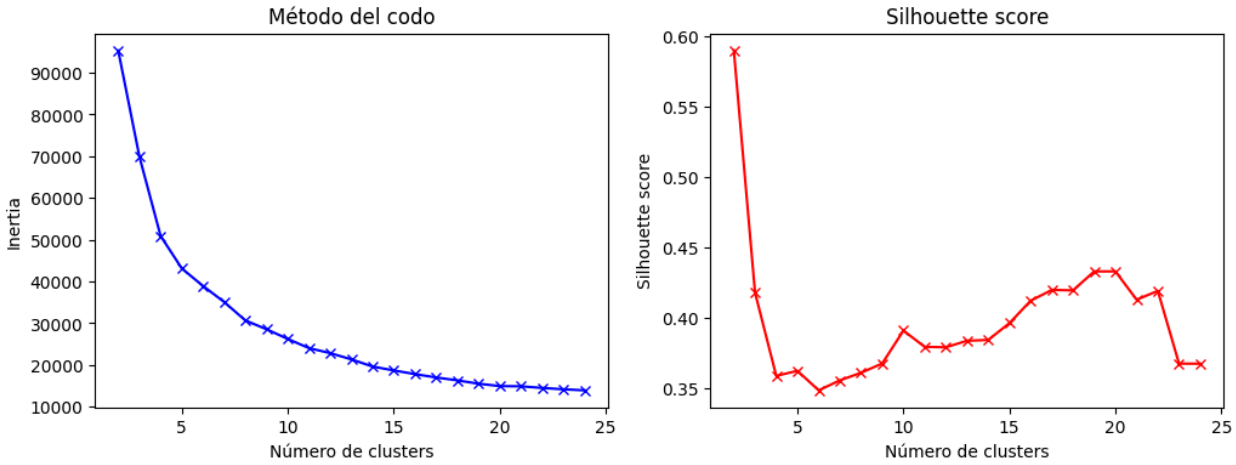
6.2.2 Modelo no supervisado:

El objetivo del modelado no supervisado en este conjunto de datos es la agrupación de buques con características técnicas similares. Esto puede ser de utilidad en cuanto a la optimización de recursos, políticas pesqueras, la asignación de ayudas o detección de buques que tengan unas características muy diferentes al resto de la flota.

Para ello no se tendrá en cuenta el tipo de arte de pesca ni el puerto, provincia y comunidad autónoma en la que se encuentra el buque, sino sus características técnicas.

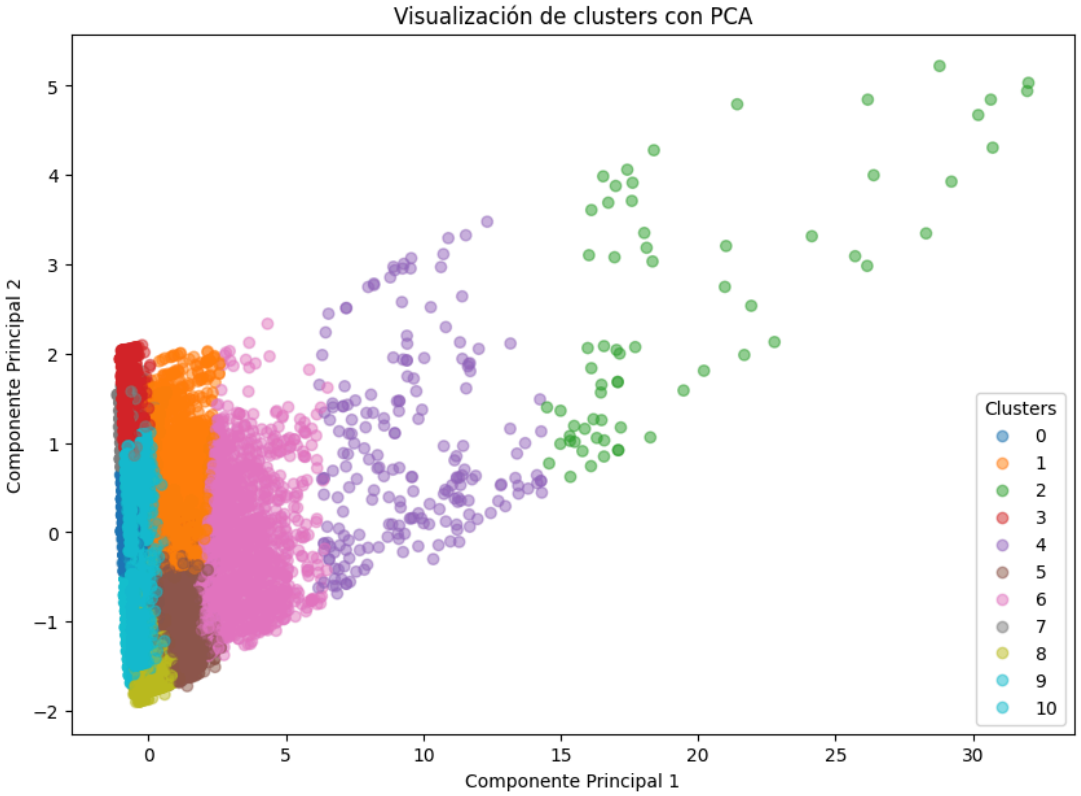
El algoritmo utilizado para la agrupación es un KNN de vecinos más cercanos, de esta forma se agruparán los buques dependiendo de la distancia a otros registros de buques con características similares.

Para encontrar el número óptimo de grupos en los que segmentar la flota, se realiza un entrenamiento en bucle en un rango desde 2 a 25 agrupaciones, buscando en qué número concreto de *clusters* disminuye la inercia y aumenta el *silhouette score*, que mide lo bien definidos que están los grupos.

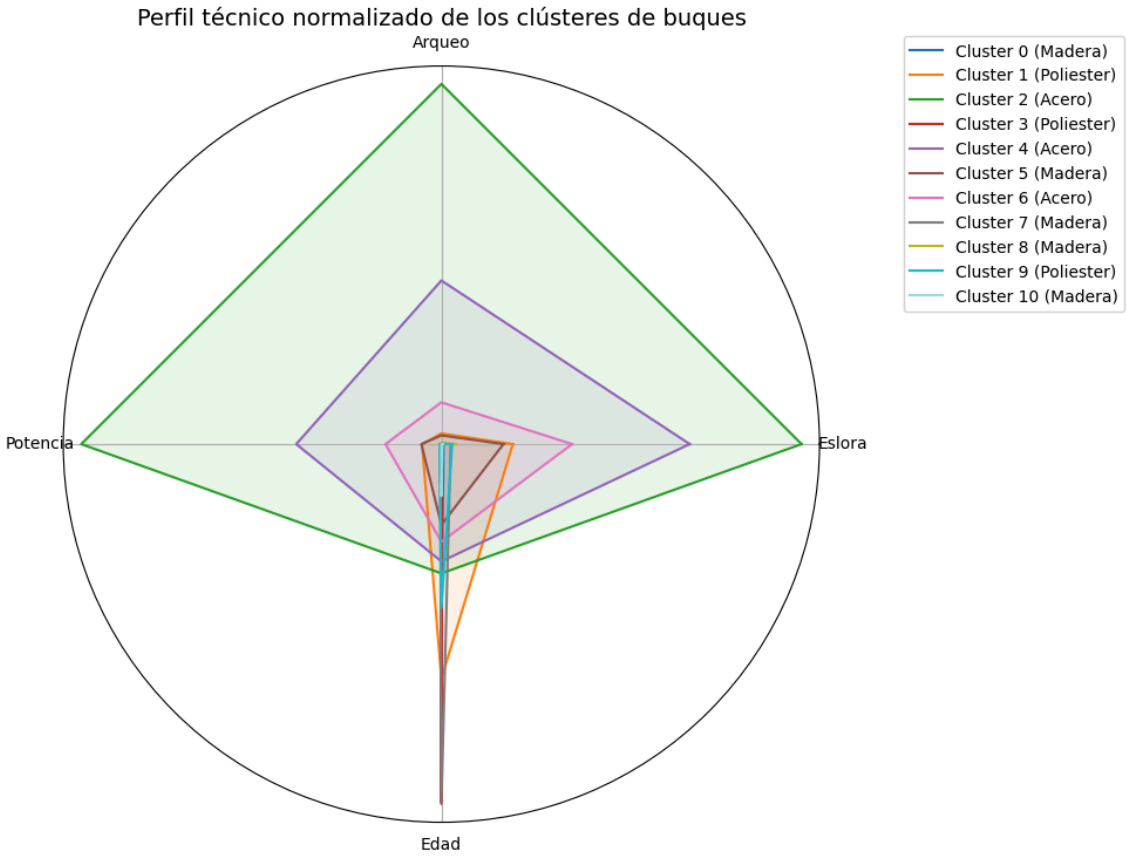


Tras realizar el entrenamiento, observamos que aunque alrededor de los 18 grupos las métricas son un poco más altas, decidimos tomar 11 grupos como el número óptimo, ya que presenta un *silhouette score* de 0.40 y no segmenta tanto la flota como 18 grupos diferentes.

Si representamos en 2 dimensiones mediante un PCA los diferentes grupos, podemos observar cómo se produce cierto solapamiento en los valores más pequeños de los componentes principales, mientras que cuando aumentan estos valores, los grupos quedan mejor diferenciados.



Esto se refleja también en el siguiente gráfico de radar, en el que se pueden ver las diferencias entre las características técnicas de los diferentes grupos:



Finalmente, si comparamos las medias de las características técnicas (moda en el caso del material del casco) podemos ver que las diferentes agrupaciones tienen características bien diferenciadas:

cluster	eslora_total	arqueo_gt	potencia_kw	Edad_buque	Material del casco (moda)
0	6.64	2.52	21.28	18.75	Madera
1	21.49	77.21	209.27	23.16	Poliester
2	83.28	2534.71	3555.61	14.69	Acero
3	6.18	1.89	21.97	33.18	Poliester
4	59.34	1152.64	1444.77	13.72	Acero
5	19.46	63.31	216.40	10.79	Madera
6	34.16	295.97	568.47	12.15	Acero
7	8.17	4.65	31.75	33.27	Madera
8	9.27	8.79	47.64	4.29	Madera
9	8.48	4.40	35.37	17.47	Poliester
10	6.40	2.29	19.51	8.46	Madera

- **Cluster 8**: Buques muy pequeños y muy recientes, con eslora y potencias bajas
- y edad < 4 años.
- **Clusters 0, 3, 7, 9**: Buques pequeños con baja potencia. Eslora < 10 m, arqueo < 5 GTs y potencia < 50 KWs, donde predomina la madera y poliéster como material de casco.
- **Cluster 3 y 7:** También formado por buques pequeños con eslora < 10 m, estos están más envejecidos, con una edad > 30 años
- **Clusters 1 y 5:** Buques medianos de madera o poliester, con una eslora de aproximadamente 20 m, arqueo < 80 GTs y potencia de 220 KWs aproximadamente.
- **Cluster 6:** Buques medianos de acero y mayor potencia, sobre los 35 m de eslora y mayor capacidad (300 GTs de arqueo), con una gran potencia, alrededor de los 570 KWs.
- **Clusters 2 y 4:** Buques grandes, entre los 60 y 80 m de eslora media y con potencias entre 1500 y 3500 KWs con gran capacidad de arqueo, donde predomina el acero como material del casco.

7. Resolución del problema

El objetivo principal era clasificar el tipo de arte de pesca de cada buque a partir de sus características técnicas y segmentar la flota en tipologías homogéneas, además del análisis de sus características técnicas.

- **Clasificación supervisada**: se implementó un flujo de trabajo que incluyó limpieza, imputación, transformaciones y estandarización de variables numéricas, codificación de categóricas y partición estratificada de datos.

    - Se entrenaron varios algoritmos basados en bagging y boosting, seleccionando LightGBM como base.

- Mediante optimización de hiperparámetros y sobremuestreo SMOTE, se obtuvo un modelo final con una buena relación entre precisión y sensibilidad (precision = 0.67, recall = 0.71).

- **Segmentación no supervisada:** usando KNN, se definieron 11 clusters que agrupan buques con perfiles técnicos similares.
    - Cada cluster corresponde a combinaciones de eslora, arqueo, potencia y material de casco, desde buques muy nuevos y pequeños hasta grandes embarcaciones de alta potencia.
    - Esta herramienta ofrece una guía para el diseño de políticas de modernización y asignación de recursos según las necesidades específicas de cada grupo.

En conjunto, la resolución del problema se materializa en:

1. Un modelo predictivo entrenado y validado, listo para integrarse en sistemas de gestión, capaz de anticipar el tipo de arte de pesca de un buque.
2. Un conjunto de tipologías de la flota que facilita análisis de segmentación, planificación estratégica y seguimiento de modernización.
3. Documentación y análisis exploratorio (código Python, modelos .pkl, gráficos y datos procesados) disponibles para replicación y extensión del estudio.

Conclusiones finales

- **Calidad y preparación de los datos**
    - Se verificó la ausencia de duplicados y se depuraron columnas con alto porcentaje de valores nulos o inconsistentes (p. ej. ‘IMO’, ‘IRCS’, ‘tipo_auxiliar’), garantizando un dataset limpio y homogéneo.
    - La imputación de ceros en ‘eslora_total’, ‘arqueo_gt’ y ‘potencia_kw’ mediante KNN con k = 15 preservó las propiedades estadísticas originales, mantuvo la coherencia de las relaciones multidimensionales y evitó distorsiones debidas a outliers.
- **Análisis exploratorio**
    - La flota española ha envejecido: una parte significativa de los buques supera los 30 años, aunque existe una renovación notable en segmentos de 0–10 años.

- Se aprecia una tendencia hacia embarcaciones de mayor eslora y arqueo en los buques más recientes, junto con picos de potencia que reflejan modernización tecnológica.

- El material del casco ha evolucionado: disminuye el poliéster y crece la fibra de vidrio, mientras que el uso de madera se mantiene para buques pequeños de las últimas dos décadas.

- **Relaciones entre variables**

  - Fuerte correlación positiva entre eslora, arqueo y potencia (0.80–0.94), coherente con la física naval y que confirma la validez de los datos tras la limpieza.

  - La prueba chi-cuadrado mostró dependencia significativa entre comunidad autónoma y tipo de arte de pesca ($p < 0.05$), lo que sugiere diferencias regionales en la modalidad pesquera.

  - El test de Kruskal-Wallis confirmó que la eslora varía según el tipo de arte ($p < 10^{-16}$), destacando que artes menores agrupan buques muy pequeños, mientras que arrastre, cerco y desconocido presentan las mayores esloras.

- **Modelado supervisado**

  - Se compararon Decision Tree, Random Forest, LightGBM y CatBoost con métrica de balanced_accuracy; LightGBM destacó con 0.6657.

  - Tras optimización por Grid Search y sobremuestreo SMOTE, el modelo final (LightGBM + SMOTE) alcanzó precision = 0.67, recall = 0.71 y accuracy = 0.87, equilibrando la identificación de la clase mayoritaria con la detección de las minoritarias.

  - El modelo clasifica mejor arrastre, artes menores y “desconocido”; presenta dificultades en cañeros, palangre y artes fijas, debido a similitudes técnicas entre estos buques y la falta de datos de comportamiento (GPS).

- **Modelado no supervisado**

  - El análisis de clustering (KNN con 11 clusters) segmentó la flota en grupos técnicamente diferenciados, desde buques muy pequeños y recientes (cluster 8) hasta grandes buques de acero con alta potencia (clusters 2 y 4).

  - Estas tipologías facilitan la toma de decisiones en políticas de modernización, asignación de ayudas y detección de embarcaciones atípicas.

- **Recomendaciones**
    - Incorporar datos de geolocalización (trayectorias GPS) mejoraría la diferenciación de artes pesqueras y aumentaría la precisión de los modelos supervisados.
    - Ampliar el análisis a variables medioambientales o de impacto económico podría enriquecer la comprensión de la sostenibilidad y eficiencia de la flota.
    - Mantener actualizado el registro y fomentar la digitalización de las características técnicas impulsaría futuros estudios de evolución de la flota.

8. Código

El código del proyecto, realizado en *Python*, se encuentra en la carpeta *./source*, diferenciado en limpieza y modelado de los datos. Las diferentes salidas .csv como las estadísticas descriptivas o los resultados del dataset clusterizado se encuentran en la carpeta *./output*, así como los modelos entrenados y listos para su uso en formato .pkl.

Los gráficos e imágenes generados durante el análisis se encuentran en la carpeta *./figures* y finalmente los datos utilizados en la carpeta *./data*.

Todo ello se encuentra disponible en el siguiente enlace al repositorio de *Github* https://github.com/LucasZV/tipologia_PR2

9. Video

El video explicativo del proyecto se encuentra en https://drive.google.com/file/d/1LhnIlgajqVJOJiwh-lgPPLZytvJRy4zg/view?usp=sharing

10. Contribuciones

Contribuciones	**Firma**
Investigación previa	Julio UQ, Lucas ZV
Redacción de las respuestas	Julio UQ, Lucas ZV

Desarrollo del código	Julio UQ, Lucas ZV
Participación en el vídeo	Julio UQ, Lucas ZV