Contexto de aplicación

La clasificación de especies de ballenas mediante imágenes forma parte de las aplicaciones de la inteligencia artificial en el ámbito de la **investigación biológica**, la **gestión ambiental** y la **protección de especies en peligro de extinción**. Tradicionalmente, la identificación de ballenas ha requerido la participación de expertos marinos realizando observaciones directas o analizando manualmente grandes volúmenes de fotografías. Este proceso es costoso, demorado y susceptible a errores humanos.

La automatización de la clasificación mediante modelos de visión por computadora permite analizar enormes cantidades de datos de manera eficiente y precisa, optimizando el tiempo y los recursos disponibles. Además, esta tecnología facilita el monitoreo continuo de las poblaciones de ballenas, contribuyendo a detectar cambios en sus patrones migratorios, áreas de alimentación y reproducción, así como también a identificar amenazas emergentes derivadas del cambio climático, la contaminación o la actividad humana.

Objetivo de Machine Learning

Predecir la especie de una ballena a partir de una imagen proporcionada. Se trata de un problema de clasificación supervisada de imágenes, donde a cada imagen de entrada le corresponde una etiqueta que representa su especie.

Dataset

Tipo de datos: Imágenes en formato JPEG con tamaño entre 170KB y 1KB .

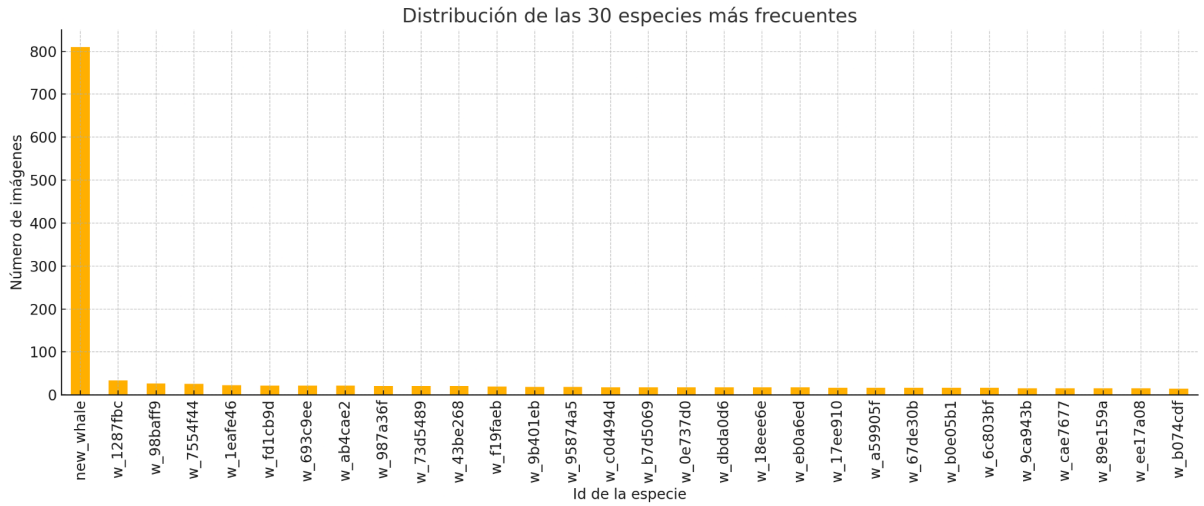
Tamaño del dataset: Cantidad de imágenes 9850, con un tamaño en disco de 289MB

Distribución de clases: La variable objetivo del conjunto de datos (`Id`) representa las diferentes especies de ballenas. Al analizar la distribución de clases, se observa que las imágenes no están distribuidas de manera equilibrada entre las distintas especies.

- Algunas especies cuentan con varios cientos de imágenes, mientras que otras apenas tienen unas pocas instancias.

- En el análisis inicial, las 30 especies más frecuentes concentran un número significativamente mayor de imágenes en comparación con el resto.

- Además, se identifican múltiples especies que poseen menos de 10 imágenes en el conjunto de entrenamiento.



Métricas de desempeño de machine learning

Exactitud (Accuracy): define la proporción de predicciones correctas respecto al total de muestras evaluadas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall (Sensibilidad o Exhaustividad): Indica qué proporción de los casos positivos reales fueron correctamente identificados.

$$Recall = \frac{TP}{TP+FN}$$

Precision (Precisión): Indica qué proporción de las predicciones positivas fueron correctas.

$$Precisión = \frac{TP}{TP+FP}$$

F1-Score (Media armónica de Precision y Recall): Mide el equilibrio entre Precision y Recall. Es útil en datasets desbalanceados.

$$F1 - Score = 2 \times \frac{Precisión \times Recall}{Precisión + Recall}$$

Balanced Accuracy (Exactitud Balanceada): Es decir, el promedio del Recall de la clase positiva y la clase negativa.

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Donde:

- **TP**: Verdaderos Positivos
- **TN**: Verdaderos Negativos
- **FP**: Falsos Positivos
- **FN**: Falsos Negativos

Métricas de desempeño del negocio

Tasa de Detección Correcta de Especies: Proporción de especies de ballenas correctamente identificadas frente al total de especies evaluadas.
Relacionada directamente con el **Recall macro** (promedio de recalls por clase).

$$Tasa\ de\ Deteccion = \frac{textEspeciescorrectamenteidentificadas}{textTotaldeespecies}$$

Tasa de Error de Clasificación: Porcentaje de especies mal clasificadas, lo cual puede derivar en **decisiones erróneas** en programas de conservación o monitoreo.

$$Error\ de\ Clasificacioon = 1 - Accuracy$$

Referencias y resultados previos

- **Fuente principal**:
  - Whale Categorization Playground - Kaggle: https://www.kaggle.com/competitions/whale-categorization-playground

- **Resultados previos**:
  - Modelos de clasificación de imágenes basados en **redes neuronales convolucionales (CNNs)** y **transfer learning** (como ResNet50 o EfficientNet) han mostrado desempeños destacados en problemas similares de clasificación de especies animales.

  - Algunos notebooks públicos en Kaggle aplican **augmentación de datos**, **normalización**, y **optimizadores como Adam** para mejorar la precisión.