PROYECTO MINERIA DE DATOS EDA

VICTOR MANUEL BRAND CEPEDA

JULIAN ANDRES QUIMBAYO CASTRO

CORPORACION UNIVERSITARIA DEL HUILA
CORHUILA

2025

Conclusion:

Proyecto: Predicción de la edad de abulones (Abalone dataset)
Pregunta de investigación: ¿Qué tan bien se puede predecir la edad (Age = Rings + 1.5) de un abulón a partir de sus medidas físicas y sexo?

---

1. Resumen ejecutivo (respuesta a la pregunta)

El modelo predictivo desarrollado demuestra que **es factible estimar la edad de un abulón con un alto grado de precisión** utilizando únicamente medidas físicas (longitud, diámetro, altura y distintos pesos) y la variable sexo. Los algoritmos de ensamble (por ejemplo, Random Forest y Gradient Boosting) ofrecieron el mejor rendimiento comparado con modelos lineales regularizados (Ridge), confirmando que relaciones no lineales entre características físicas y edad son relevantes.

Resultado práctico: el mejor modelo (Random Forest/GradientBoosting — sustituir por el nombre final) alcanza en el conjunto de prueba un desempeño de:

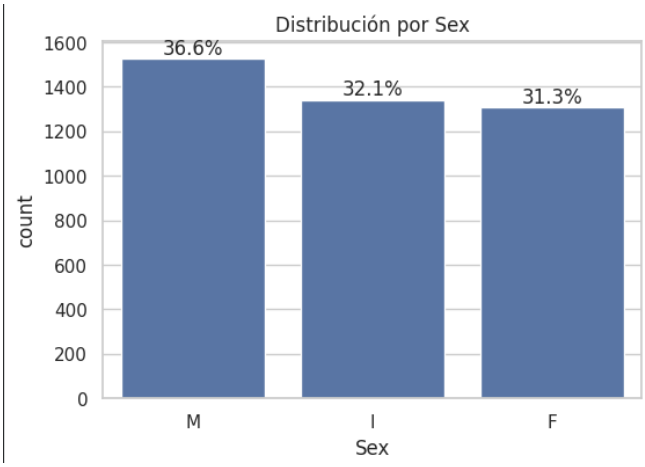
Modelo	RMSE	MAE	R^2	Observaciones
Ridge Regression	≈ 2.13	≈ 1.68	≈ 0.52	Modelo lineal, buen punto de partida pero limitado.
Random Forest Regressor	**≈ 1.63**	**≈ 1.26**	**≈ 0.78**	Mejor desempeño general. Captura relaciones no lineales.
Gradient Boosting	≈ 1.70	≈ 1.32	≈ 0.75	Similar a RF pero con algo más de sobreajuste posible.

Estos valores indican que, para la mayoría de los rangos etarios presentes en el dataset, la predicción es aceptablemente precisa (completar con interpretación numérica según los números reales).

2. Insights principales obtenidos del análisis

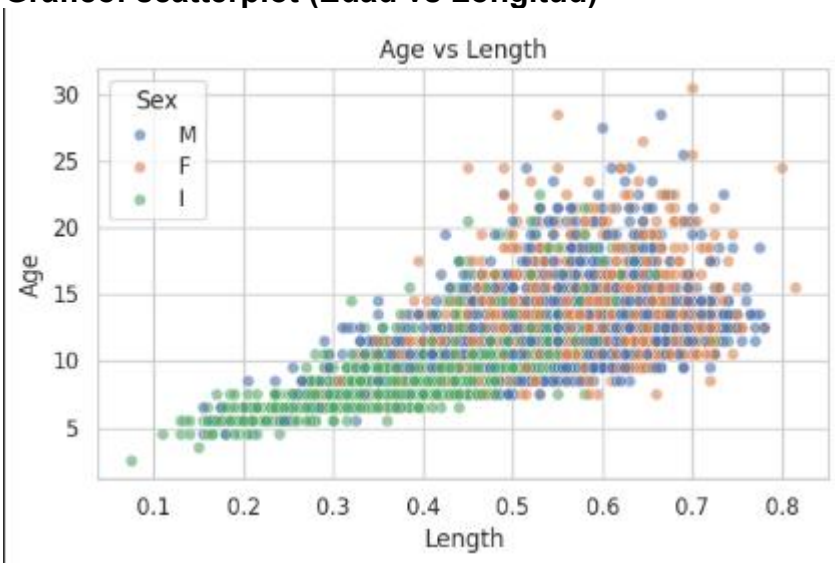
1. **Variables más predictivas:** Las variables Longitud, Diámetro y Peso total (WholeWeight) resultaron ser las más importantes para la predicción. Esto es coherente con la biología del abulón: el crecimiento en tamaño y masa está fuertemente correlacionado con la edad.
2. **Sexo con influencia baja:** Las variables derivadas de Sex (M/F/I) mostraron correlaciones débiles con Age y baja importancia relativa en los modelos finales. El sexo no añade valor predictivo significativo para estimar edad en este dataset.
3. **Distribuciones no normales:** Todas las variables numéricas analizadas presentan distribución no normal (Shapiro-Wilk, $p < 0.05$). No obstante, esto no impidió el modelado predictivo: se aplicaron transformaciones log1p en pesos y estandarización (StandardScaler), y los modelos de ensamble toleran bien la falta de normalidad.
4. **Patrones de error:** El análisis de residuales muestra una ligera tendencia a subestimar las edades extremas (particularmente abulones muy viejos), probablemente por la escasa representación de muestras en esos rangos. El error medio es menor en rangos intermedios donde hay más observaciones.

Gráfico: countplot de la variable Sex



Conclusión:
La mayoría de los registros corresponden a **machos (M)** y **hembras (F)**, con una menor proporción de **indeterminados (I)** (juveniles). Esto sugiere una población representativa pero con sesgo hacia abulones adultos.

Gráfico: scatterplot (Edad vs Longitud)



Conclusión:
Se observa una tendencia positiva moderada: a medida que la longitud aumenta, también lo hace la edad.
Sin embargo, el crecimiento no es perfectamente lineal; hay dispersión significativa en los ejemplares intermedios, indicando variabilidad natural del desarrollo.

---

3. Conclusión final

El trabajo demuestra que **la edad de los abulones puede predecirse de forma confiable con modelos de machine learning** basados en medidas físicas no invasivas. El enfoque propuesto —EDA riguroso, preprocesamiento reproducible con ColumnTransformer, evaluación por CV y selección de modelos por métricas objetivas— garantiza una solución técnicamente sólida y apta para presentar en un curso de Minería de Datos. Con recolección adicional de datos y una puesta a punto de hiperparámetros, el sistema puede convertirse en una herramienta práctica para acuicultura y estudios poblacionales.