Análisis Avanzado de Datos

EVALUACIÓN

29 AL 31 DE AGOSTO

CRITERIOS DE EVALUACIÓN:

Comprensión y preparación de datos

- Calidad del análisis exploratorio y cómo se han preparado los datos para el modelado.

Implementación de modelos de regresión y su validación por medio de métricas de calidad

- Correcta selección y entrenamiento de modelos de regresión.
- Uso apropiado de métricas para evaluar la calidad del modelo.

Métodos de regularización

- Implementación y justificación de métodos de regularización como Ridge y Lasso.

Métodos de validación cruzada y Bootstrap

- Uso efectivo de técnicas de validación cruzada para ajuste de hiperparámetros y evaluación del modelo.
- Implementación de métodos de bootstrap para estimación de intervalos de confianza y robustez del modelo.

Los criterios para autoevaluación, coevaluación y heteroevaluación, se define a partir de la rúbrica anexa en la plataforma Moodle del curso

PARTE 1 (50 pts)

El dataset ***CrabAgePrediction_Subset2.csv***. es conjunto de datos que contiene varias mediciones físicas de cangrejos junto con su edad. Estas mediciones incluyen la longitud, el diámetro, la altura, el peso total, el peso sin concha, el peso de las vísceras y el peso de la concha. También se proporciona información sobre el sexo de cada cangrejo, pero por ahora esta característica deja por fuera.

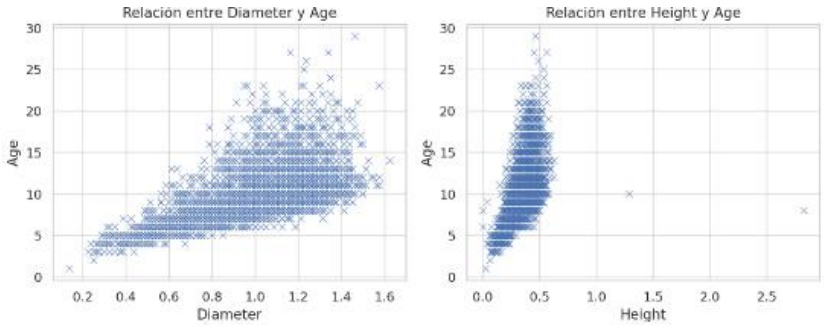
Objetivo: Desarrollar un modelo que pueda predecir con precisión la edad de los cangrejo, **machos y Hembras,** en función de las características que están en el dataset y evaluar su rendimiento Finalmente, el modelo se entregará para su evaluación independiente con un conjunto de datos no revelado que posee el profesor.

Análisis Avanzado de Datos

1.A Análisis exploratorio de los datos (10 puntos):

Visualizar la relación entre las diferentes características y la variable objetivo ("Age"). Y verifica la correlación entre las características y la variable objetivo, defina que característica no va tomar como predictores y cuales si, recuerda no tomar predictores categóricos, pero si tener cuidado con su análisis

Un ejemplo de lo que se espera que visualicen y entren es



Y que analizan sean frase de este estilo *"… Weight y Shell Weight: Estas características tienen coeficientes positivos significativos, lo que indica que un aumento en estas medidas generalmente resulta en un cangrejo más viejo.* Las categorías 'M' (macho) y 'F' (hembra) tienen coeficientes positivos, mientras que 'I' (indeterminado) tiene un coeficiente negativo …*"*

***Recuerde**: son libre de seleccionar el mejor método para la validación (conjunto de validación, validación cruzada)*

1.B. Construye un modelo de regresión (lineal múltiple o polinomial) para predecir la edad del cangrejo y evalúa los modelos utilizando métricas como RMSE y MAE (25 pts)

Se espera que script de júpiter se evidencia todos los modelos entrenados y al final indicar cuál es el modelo seleccionado entregando el valor de métricas de calidad y los coeficientes del polinomio o de línea multivariable

1.C. Validación y entrenamiento final, (15 pts)

Con el modelo seleccionado vuelva a entrenar con todo el conjunto de datos y entregar los nuevos coeficientes y la métrica de calidad finales esperadas

Entregables:

- **Archivo de Notebook**: Envíe un archivo de cuaderno de Jupyter (**.ipynb**) o Google Colab, titulado **parte1.ipynb**. Este archivo debe dividirse en secciones que aborden los ítems 1.a, 1.b y 1.c. En cada sección, asegúrese de explicar detalladamente tanto la selección de datos como los métodos de validación empleados para cada etapa del análisis.

Evaluación

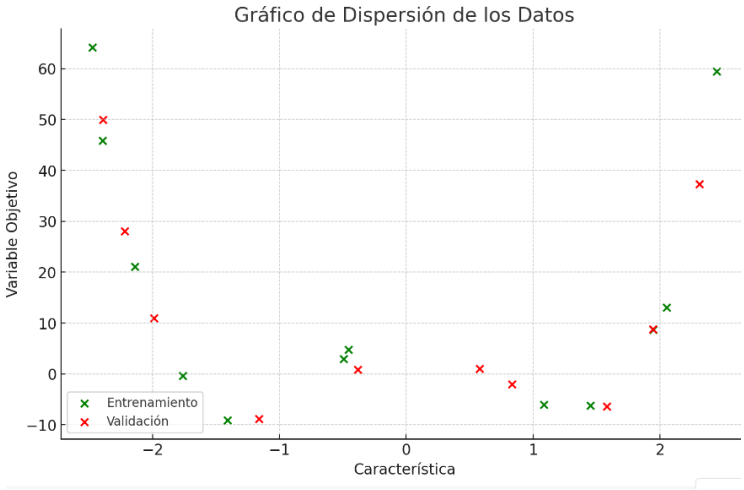
Análisis Avanzado de Datos

PARTE 2 (30 pts)

En la siguiente grafica tenemos los datos aleatorios, Desarrollar un modelo de regresión polinomial para predecir la variable objetivo. Validar el rendimiento del modelo utilizando técnicas de validación cruzada y bootstrap.

Los datos los descargar del Moodle y tiene el nombre de **part2xVal, part2yVal, part2xtrain, part2xtrain** y están distribuido de la siguiente manera:

- Conjunto de entrenamiento de características (x_train): 12 observaciones
- Conjunto de validación de características (x_val): 10 observaciones
- Conjunto de entrenamiento de etiquetas (y_train): 12 observaciones
- Conjunto de validación de etiquetas (y_val): 10 observaciones



2.A Regresión Polinomial y Validación Cruzada (10pts)

- Implementar modelos de regresión polinomial de varios grados.
- Utilizar validación cruzada para evaluar el rendimiento del modelo en el conjunto de entrenamiento.
- Reportar métricas MSE y el MAE.

2.B Mejora del Modelo con Regularización (10pts)

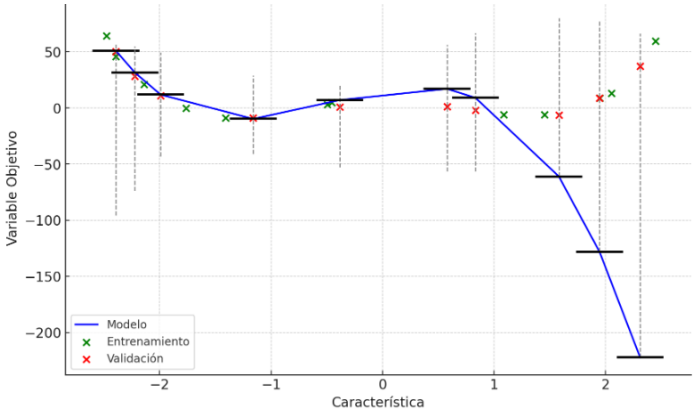
1. Experimentar con métodos de regularización Ridge o Lasso con los modelos polinomiales.
2. Utilizar validación cruzada para encontrar los mejores parámetros de regularización.
3. Reporta e hiperparametro lamda que mejor resultado da junto con el polinomio

2.C Evaluación con Bootstrap

- Implementar el procedimiento de bootstrap para estimar la distribución la distribución MSE del modelo seleccionado en los dos pasos anteriores.

Evaluación

- Grafica los resultados del bootstrap para calcular un intervalo de confianza para esta métrica, ver la siguiente imagen como ejemplo:



Entregables:

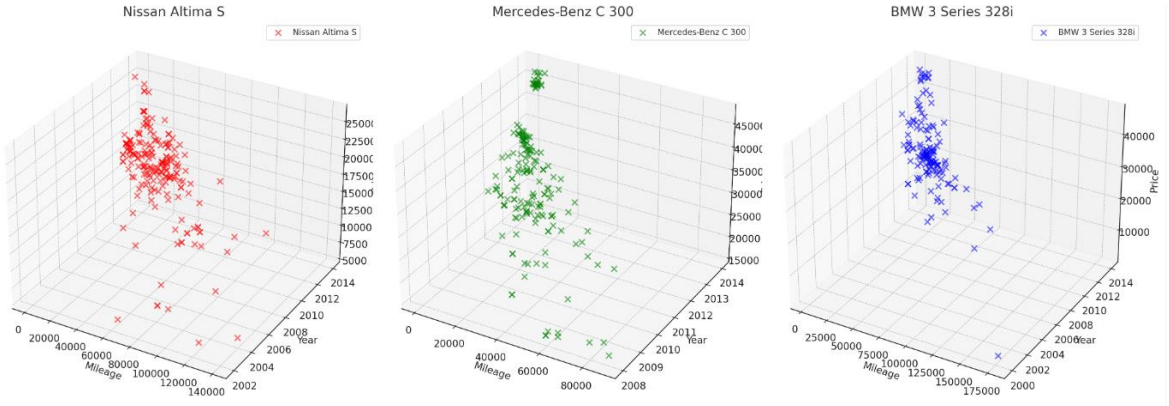
- **Archivo de Notebook**: Proporcione un archivo de cuaderno de Jupyter (**.ipynb**) o Google Colab, titulado **parte2.ipynb**. Este archivo debe estar estructurado en secciones que correspondan a los ítems 2.a, 2.b y 2.c. Cada sección debe incluir detalles sobre la selección del modelo empleado.

PARTE 3 (20 pts)

Desarrollar el modelo de regresión para predecir el precio de venta de vehículos de una de las marcas específicas en el conjunto de datos LondonCars2014. Los predictores claves a considerar son el kilometraje del vehículo y el año de fabricación. El rendimiento del modelo se evaluará en función del valor del Error Cuadrático Medio de la Raíz (RMSE). Cuanto menor sea el RMSE, mejor será el modelo

3.A Selección de la Marca: Elija una de las marcas y un modelo de automóviles disponibles en el conjunto de datos para centrar su análisis.

Marca	Model
Nissan	Altima S
Mercedes-Benz	C 300
BMW	3 Series 328i
Lexus	RX 350
Infiniti	G 37
Mercedes-Benz	E 350
Lexus	ES 350
Mercedes-Benz	ML 350
Honda	Civic LX



3.B Selección de Modelo: Está en libertad de emplear cualquier método de regresión que considere apropiado para lograr el objetivo, incluyendo regularización.

3.C Validación del Modelo: Utilice técnicas de validación como la validación cruzada para asegurarse de que el modelo es robusto y generalizable a nuevos datos.

3.D Presentación de Resultados: Documente claramente todas las etapas del análisis, desde la exploración de datos hasta la evaluación del modelo, y presente sus resultados de una manera coherente y fácil de seguir.

Entregables:

- **Archivo de Notebook**: Proporcione un archivo de cuaderno Jupyter (**.ipynb**) o Google Colab, nombrado como **parte3.ipynb**. Este archivo debe contener las secciones 3.a, b y c y la Documentación de cada sección debe venir acompañada con una explique clara del enfoque adoptado, los resultados obtenidos y cualquier conclusión relevante