Desafío - Enfermedad en la sangre

En este desafío validaremos nuestros conocimientos de métodos de regularización. Para lograrlo, necesitarás aplicar regulación Ridge, Lasso y Elastic Net**.**

Lee todo el documento antes de comenzar el desarrollo **individual**, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos.

Tiempo asociado: 4 horas cronológicas.

Descripción

En este desafío utilizaremos una base de datos recolectada en un laboratorio de donación de sangre, en la que contamos con información de diferentes mediciones incluido el sexo del donador y un indicador de donador sospechoso. Las variables regresoras son:

1. **X (Patient ID/No.):** Número o identificación única asociada a cada paciente en la base de datos.
2. **Category (diagnosis):** Categoría que indica el diagnóstico relacionado con la hepatitis C para cada paciente.
3. **Age (in years):** Edad del paciente en años al momento de la detección o registro en la base de datos.
4. **Sex (f,m):** Género del paciente, indicado por "f" para femenino y "m" para masculino.
5. **ALB:** Albúmina, una proteína producida por el hígado. Los niveles de albúmina pueden indicar la función hepática.
6. **ALP:** Fosfatasa alcalina, una enzima que puede estar relacionada con la función hepática y otros procesos.
7. **ALT:** Alanina aminotransferasa, una enzima que puede indicar daño hepático.
8. **AST:** Aspartato aminotransferasa, una enzima que también puede indicar daño hepático.
9. **BIL:** Bilirrubina, un pigmento amarillo que puede aumentar en casos de problemas hepáticos.
10. **CHE:** Colinesterasa, una enzima que puede estar asociada con la función hepática y otros procesos.
11. **CHOL:** Colesterol, un lípido que puede estar relacionado con la salud del hígado.
12. **CREA:** Creatinina, un producto de desecho que se filtra a través de los riñones.
13. **GGT:** Gamma-glutamil transferasa, una enzima que puede estar relacionada con la función hepática.
14. **PROT:** Proteínas totales, que pueden incluir varias proteínas, como albúmina y globulinas.

La variable objetivo es **Category,** y almacena las siguientes categorías: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'

Aplicando los conceptos y herramientas aprendidas hasta ahora, deberás realizar las siguientes tareas:

1. Importa las librerías necesarias para entrenar modelos de regresión logística con validación cruzada y Extreme Gradient Boosting, carga los datos y prepáralos. Para ello:
    a. elimina la columna 'Unnamed: 0'.
    b. codifica la variable objetivo **Category** en dos categorías: una para Category='0=Blood Donor', a la que debes asociar valor 0, y un 1 para los demás valores. Asigna estos valores en una nueva columna llamada **target**.
    c. codifica en otra columna, con nombre **suspect,** con valor 1 cuando Category='0s=suspect Blood Donor', y aigna 0 en caso contrario. Elimina finalmente la columna Category.
    d. Revisa si la base de datos contiene valores ausentes. En caso que sea así, aplica los siguientes criterios:
        i. si el porcentaje de valores ausentes para alguna variable es menor a 1%, entonces elimina las filas que contengan estos valores ausentes
        ii. si el porcentaje de valores ausentes es mayor, entonces reemplaza estos por el promedio según la clase a la que pertenezca el valor ausente.
    e. Realiza un análisis descriptivo para cada variable regresora, usando gráficos para representar distribuciones y boxplot para revisar posibles outliers. Construye un heatmap que muestre las correlaciones entre las variables regresoras y describe.
    f. Transforma las variables regresoras por medio de la estandarización. Muestra el antes y después de esta usando boxplots, y realiza una segmentación de la muestra en 33% para test y el resto para entrenamien

2. Desarrolla un modelo regresión logística con validación cruzada usando 5-fold, con regularización Elastic Net. Busca para los siguientes hiper parámetros:

    - Cs: valores entre 0.01 y 5.0, con 200 valores lineales.
    - l1_ratio: valores entre 0 y 1, con 200 valores lineales.

    Muestra los valores óptimos encontrados, y las métricas precisión, recall, f1-score y accuracy. Describe cuáles son las tres características con mayor incidencia en la predicción de sujeto o muestra con sangre con posible enfermedad.

3. Implementa un modelo Extreme Gradient Boosting con búsqueda de hiper parámetros de grilla, con regularización para la combinación L1 y L2. Considera para esto:

   a. reg_lambda: valores entre 0.0 y 2.0, con 10 valores lineales.
   b. reg_alpha: valores entre 0 y 1, con 10 valores lineales.
   c. learning_rate: valores entre 0.1 a 10, con 20 valores lineales.

4. Muestra el nivel de importancia de los atributos, los valores óptimos de los hiper parámetros y las métricas precisión, recall, f1-score y accuracy. Compara los resultados de este modelo con el de regresión logística.

Requerimientos

1. Importa las librerías, carga los datos y los prepara para lo solicitado **(1 punto)**
2. Implementa y evalúa un modelo ElasticNet, buscando sus parámetros más adecuados. **(4 puntos)**
3. Implementa y evalúa un modelo Extreme Gradient Boosting, buscando sus parámetros más adecuados. **(4 puntos)**
4. Compara los rendimientos de los modelos anteriores, a partir de sus métricas respectivas. **(1 punto)**

 ¡Mucho éxito!

Consideraciones y recomendaciones
Debes entregar tu trabajo en un archivo Jupyter, en el que realices las tareas pedidas e incluyas las explicaciones necesarias en cada paso.