Laboratorio: Resolver un problema de regresión

**Objetivos de la actividad**

En esta actividad vas a profundizar en las distintas técnicas que se pueden aplicar para abordar un problema de regresión. Además, profundizarás en tus conocimientos sobre las librerías *statsmodels* y *scikit-learn* de Python, al crear un conjunto de datos mediante simulación.

**Descripción de la actividad y pautas de elaboración**

El primer paso consiste en crear un conjunto de datos ficticio. Generaremos nuestro conjunto con un número semilla compuesto por ocho cifras a través de nuestro número de identificación. Si el número de identidad tiene más de esas cifras, nos quedaremos con las primeras ocho empezando desde la derecha. Si tiene menos de ocho cifras replicaremos las primeras hasta obtener exactamente ocho. Además, para evitar los dígitos cero y uno, si alguna de las cifras es menor que dos la sustituiremos por ese número. Aplicando estos cambios tendremos el número de semilla a través del documento de identidad preparado para la resolución del laboratorio. Veamos algunos ejemplos:

* Ejemplo 1:

1234567890 34567890 34567892

* Ejemplo 2:

12345678 22345678

* Ejemplo 3:

304156 30415630 32425632

Para lo anterior se puede crear una función que lo realice automáticamente o podemos hacerlo manualmente.

Una vez tenemos el número del documento de la semilla vamos a crear el conjunto de datos para el problema de regresión con la función *sklearn.datasets.make\_regression* de la librería *scikit-learn*.

Accede a la librería de *scikit-learn* en el siguiente enlace: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_regression.html>

* Crea el conjunto de datos empleando los siguientes argumentos:
  + n\_samples = 200 + 10 · primer dígito semilla
  + n\_features = 10 + segundo dígito semilla + tercer dígito semilla
  + n\_informative = 10 + segundo dígito semilla
  + bias = 2
  + noise = 10 \* cuarto dígito semilla
  + random\_state = semilla
  + shuffle = False

Entonces, se pide de forma justificada:

* ¿Qué significa cada uno de los parámetros de entrada en la función make\_regression usados para crear nuestro *dataset*?
* Describe tu conjunto de datos, aplica los métodos info, describe, head, tail y los que creas adecuados y obtén un análisis gráfico de todas las variables (predictoras y respuesta).
* Divide el conjunto de datos en 200 observaciones para el entrenamiento y el resto para realizar la validación de los distintos métodos de regresión aplicados.
* Obtén un modelo de regresión lineal múltiple. ¿Son todas las variables predictoras significativas?
* Realiza una selección de variables mediante un algoritmo de tipo *stepwise*, donde en cada paso se elimine la variable predictora menos significativa atendiendo a su p.valor hasta que todas las variables del modelo sean significativas.
* Realiza una regresión de tipo Ridge. Prueba distintos valores de α.
* Realiza una regresión de tipo Lasso. Prueba distintos valores de α.
* Realiza una regresión mediante la red elástica. Prueba distintos valores de r y obtén el valor óptimo de r y de α mediante validación cruzada.
* Comprueba con la muestra de validación con cuál de los modelos se obtiene un menor error cuadrático medio.
* No olvides añadir las referencias utilizadas para la elaboración del trabajo al final del documento.

**Extensión**: entrega de dos archivos:

* Un archivo ipynb con las operaciones realizadas. Formato en el aula.

apellido1\_apellido2\_laboratorio.ipynb

Ejemplo: gonzalez\_sanchez\_laboratorio.ipynb

* Un archivo pdf con el análisis de los resultados obtenidos. Máximo 10 páginas.

apellido1\_apellido2\_laboratorio.pdf

Ejemplo: gonzalez\_sanchez\_laboratorio.pdf

**Rúbrica:**

|  |  |  |  |
| --- | --- | --- | --- |
| Resolver un problema de regresión | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | El conjunto de datos y su descriptivo son correctos. | 2 | 20% |
| Criterio 2 | El modelo de regresión lineal múltiple y la selección de variables se realizan de manera adecuada. | 2 | 20% |
| Criterio 3 | Se aplican correctamente los modelo de regresión de Ridge, Lasso y red elástica. | 3 | 30% |
| Criterio 4 | La validación se ha realizado correctamente. | 1 | 10% |
| Criterio 5 | La entrega, la presentación, el formato y la bibliografía son correctos. | 2 | 20% |
|  |  | **10** | **100 %** |