Caso Práctico Final Evaluable

Code

Tomaremos el dataset Salaries.csv

El conjunto de datos consiste en los salarios de nueve meses recogidos de 397 profesores universitarios en los EE.UU. durante 2008 y 2009. Además de los salarios, también se recogió el rango del profesor, el sexo, la disciplina, los años desde el doctorado y los años de servicio. Así, hay un total de 6 variables, que se describen a continuación.

```
  1. rank: Categórica - de profesor asistente, profesor asociado o catedrático
  2. discipline: Categórica - Tipo de departamento en el que trabaja el profesor,
ya sea aplicado (B) o teórico (A)
  3. yrs.since.phd: Continuo - Número de años desde que el profesor obtuvo su doct
orado
  4. yrs.service: Continuo - Número de años que el profesor ha prestado servicio a
l departamento y/o a la universidad
  5. sex: Categórico - Sexo del profesor, hombre o mujer
  6. salary: Continuo - Sueldo de nueve meses del profesor (USD)
```

El objetivo de esta práctica consiste en realizar un estudio íntegro del dataset para terminar implementando un modelo lineal regularizado que realice predicciones sobre el salario a percibir de un profesor. Asimismo, se pedirá aprovechar la explicabilidad de estos modelos y los estudios estadísticos realizados para arrojar intuiciones y dependencias en los datos.

Para ello, se pide al estudiante que realice los siguientes pasos:

1. Carga los datos. Realiza una inspección por variables de la distribución de salarios en función de cada atributo visualmente. Realiza las observaciones pertinentes. ¿Qué variables son mejores para separar los datos?
2. ¿Podemos emplear un test paramétrico para determinar si las medias de salarios entre hombres y mujeres son las mismas o difieren? Ten en cuenta que, en tanto que se pide usar un test paramétrico, se deberá determinar si las muestras cumplen con las hipótesis necesarias.
3. Divide el dataset tomando las primeras 317 instancias como train y las últimas 80 como test. Entrena un modelo de regresión lineal con regularización Ridge y Lasso en train seleccionando el que mejor **MSE** tenga. Da las métricas en test. Valora el uso del One Hot Encoder, en caso de emplearlo arguméntalo.
4. Estudia la normalidad de los residuos del modelo resultante, ¿detectas algún sesgo?
5. ¿Qué conclusiones extraes de este estudio y del modelo implementado? ¿Consideras correcto el rendimiento del mismo?

¡Mucho ánimo y espero que disfrutéis de esta última práctica!