

Trabajo del Tema 4

PARTE I

Tienes un compañero de MD004 que afirma que puede predecir la nota final de la asignatura en función del tiempo que pasa jugando a videojuegos. Decides ponerlo a prueba y le preguntas por la ecuación de regresión que usaría para hacer la predicción. Te responde inmediatamente que un modelo de regresión lineal simple. Ahora, ¿cómo comprobarías si realmente comprende el modelo poniendo a prueba su conocimiento sobre regresión lineal simple? (1pt)

Finalmente desafías a tu compañero a mejorar su modelo y le pides que considere agregar otra variable predictora: "la cantidad de horas dedicadas al estudio". ¿qué argumentos en contra o a favor te podría dar? ¿cómo podrías evaluar si la nueva variable realmente mejora la capacidad predictiva del modelo? (1pt)

PARTE II

Se dispone de un dataset que contiene en formato csv datos de la población de Estados Unidos con el objetivo de generar un modelo de regresión lineal simple para predecir las tasas de mortalidad por cáncer (target_deathrate) en los condados de los Estados Unidos

Data: 202311_04_cancer_reg.csv

Se pide:

1. Exploración de los datos: análisis descriptivos de los datos y graficación comentando los resultados obtenidos(3pts)

- análisis descriptivo de la variable objetivo (métricas+gráficos) comentando los resultados
- análisis descriptivo de las variables explicativas (métricas+gráficos) comentando los resultados
- análisis bivalente entre la variable objetivo vs variable explicativa, escoged las 2 variables que a priori se observe que mejor ajustarían en un modelo de regresión lineal y justificad el porqué

2. Desarrollo del mejor modelo (3pts)

- generad un modelo de regresión lineal simple para cada una de las variables escogidas
- ¿qué modelo ajusta mejor?
- validad los diferentes supuestos que deben cumplir comentados en clase (¿hay alguna variable que no cumpla el criterio de homocedasticidad? ¿has detectado outliers?)
- escoge un modelo para su validación y justifica por qué?

3. Validación mediante el conjunto de Test (2pts)

- generad un gráfico de relación dato observado vs predicción
- comentad los valores R^2 , RMSE, MAE
- te quedarías con este modelo?

Entrega:

fichero Jupyter Notebook con MD004NombreApellidosAC04.ipynb

fichero .pdf con MD004NombreApellidosAC04.pdf

Plazo: 13/12/2023 19:00