TALLER REGRESIÓN

Objetivos

1. Aplicar las técnicas de regresión lineal vistas en clase, y comunicar los resultados de manera efectiva.

Caso

Los ejecutivos de la compañía Mashable en aumentar la cantidad de artículos que se comparten en la red. El archivo OnlineNewsPopularity.csv contiene observaciones que se hicieron alrededor de 2 años.

Los datos iniciales se obtuvieron de:
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

Información de atributos:

0. url:	URL of the article
1. timedelta:	Days between the article publication and the dataset acquisition
2. n_tokens_title:	Number of words in the title
3. n_tokens_content:	Number of words in the content
4. n_unique_tokens:	Rate of unique words in the content
5. n_non_stop_words:	Rate of non-stop words in the content
6. n_non_stop_unique_tokens:	Rate of unique non-stop words in the content
7. num_hrefs:	Number of links
8. num_self_hrefs:	Number of links to other articles published by Mashable
9. num_imgs:	Number of images
10. num_videos:	Number of videos
11. average_token_length:	Average length of the words in the content
12. num_keywords:	Number of keywords in the metadata
13. data_channel_is_lifestyle:	Is data channel 'Lifestyle'?
14. data_channel_is_entertainment:	Is data channel 'Entertainment'?
15. data_channel_is_bus:	Is data channel 'Business'?
16. data_channel_is_socmed:	Is data channel 'Social Media'?
17. data_channel_is_tech:	Is data channel 'Tech'?
18. data_channel_is_world:	Is data channel 'World'?
19. weekday_is_monday:	Was the article published on a Monday?
20. weekday_is_tuesday:	Was the article published on a Tuesday?
21. weekday_is_wednesday:	Was the article published on a Wednesday?
22. weekday_is_thursday:	Was the article published on a Thursday?

23. weekday_is_friday:	Was the article published on a Friday?
24. weekday_is_saturday:	Was the article published on a Saturday?
25. weekday_is_sunday:	Was the article published on a Sunday?
26. is_weekend:	Was the article published on the weekend?
27. shares:	Number of shares (target)

ACTIVIDADES

Plantear las hipótesis de negocio que permitan profundizar en el objetivo principal de la compañía Mashable

1. Realice la depuración de datos que haga falta y exponga de manera detallada la manera que realizo la corrección

2. Genere 4 hipótesis de negocios que se puedan responder usando regresión lineal, donde se especifica cuales son las variables independientes y cuales las variables dependientes

3. Calcule los estadísticos descriptivos de todas las variables

4. Calcule los coeficientes de correlación de Pearson y Spearman entre todas las variables que relacionadas con las hipótesis de negocio. ¿Hay diferencias?

5. Realice la estimación por mínimo cuadrados de la regresión lineal entre la variables endógenas y exógenas expuestas en el punto 2.

6. Interprete los siguientes resultados de los modelos:
   a. Coeficientes estimados ¿Son todos significativos?
   b. R-cuadrado ¿Tiene un buen ajuste el modelo?
   c. ¿Cómo se comportan los residuos, tienen distribución normal?

7. Realice la depuración del modelo de tal forma que le permita especificar los factores que determinan la cantidad de artículos que comporten en la red dado el dataset de la compañía Mashable.

8. Con base en los resultados, ¿Qué propuesta puede realizar como científico de datos al área creativa para aumentar la cantidad de artículos que se comparten en la red ?