TP-2

Laboratorio de datos 2023 (comisión: G. Solovey)

Los datos para hacer este TP están en el archivo `tp2.RData` que pueden cargar en R con el comando `load("tp2.RData")`.

Parte 1: Regresión

Elaborar un modelo de regresión para predecir el número de usos diarios del sistema de Ecobici. El dataset `clima_ecobici` contiene información registrada por la estación meteorológica del Aeroparque durante todos los días del 2022 y el número total de usos diarios de bicicletas del sistema Ecobici en la CABA. Las variables del dataset son:

- `date`: fecha
- `tavg`: temperatura promedio (en grados Celcius)
- `tmin`: temperatura mínima (en grados Celcius)
- `tmax`: temperatura máxima (en grados Celcius)
- `prcp`: precipitación (en mm)
- `wdir`: dirección del viento
- `wspd`: velocidad del viento (en km/h)
- `pres`: presión atmosférica (en hPa)
- `n`: número de bicicletas utilizadas

Nota: Este dataset se construyó con todos los datos de usos del sistema Ecobici, no a partir del dataset reducido que se usó en el TP-1.

El modelo que tienen que proponer debe tener contener 2 de las siguientes variables predictoras:

- temperatura del día (promedio, mínima o máxima).
- si llueve o no llueve.
- presión atmosférica.
- velocidad del viento.
- si es día laborable o no laborable.

Para construir el modelo sigan estos pasos:

a. Realizar un análisis exploratorio que justifique la elección de las variables predictoras.
b. Ajustar el modelo.
c. Reportar adecuadamente el modelo resultante. Interpretar los coeficientes encontrados.
d. Hacer visualizaciones apropiadas.

Parte 2: Clasificación

El objetivo de este ejercicio es desarrollar un clasificador de noticias en “reales” o “fake-news”. El dataset contiene información sobre 150 noticias, algunas reales y otras falsas. Tienen una variedad de predictores pero vamos a trabajar sólo con tres:

- `title_has_excl`: variable binaria que indica si el título de la noticia tiene o no signos de exclamación.
- `negative`: porcentaje estimado de palabras en el título que tienen connotaciones negativas.
- `title_words`: número de palabras en el título.

La variable respuesta está en la variable `type`.

1

Realizar visualizaciones apropiadas para convencerse de que las variables predictoras podrían servir para clasificar las noticias en “reales” vs. “fake-news”.

2

Separar el dataset en un conjunto de entrenamiento y uno de prueba. Construir dos clasificadores (uno con k-NN y otro con un árbol de decisión) usando las tres variables mencionadas en 1. Justificar la elección de los parámetros del modelo (por ejemplo, el valor de k en k-NN). Reportar las matrices de confusión y accuracy de los modelos.

3

Supongamos que se publica un nuevo artículo que tiene un título de 15 palabras sin signos de exclamación y el 6% de sus palabras tienen connotaciones negativas. Calcular la probabilidad de que el artículo sea “fake-news”.

Sobre el informe

- Entreguen un html self-contained.
- Utilicen los paquetes que vimos en clase.
- Usen gráficos acordes al tipo de datos que analizan.
- Los gráficos deben ser claros (por ejemplo tamaño de las letras, leyendas de los ejes y colores/formas, uso de colores/formas cuando sea necesario).
- Revisen la redacción, faltas de ortografía y signos de puntuación.