TP 1: Regresión lineal

Fecha y hora límite de primera entrega: 14 de octubre de 2022 a las 23:59 p.m.

Fecha y hora límite de entrega con penalización: 28 de octubre de 2022 a las 23:59 p.m.

Fecha y hora de devolución general: 5 de noviembre de 2022 a las 11 a.m.

INSTRUCCIONES

Deberán realizar el trabajo en un **RNotebook** y entregarlo en formato R Notebook o HTML

El **RNotebook** debe:

- Tener el siguiente nombre: eea2022_tp1_**apellido_nombre**
- Mostrar todo el código que escriban. NO USAR `echo=FALSE`
- Ser ordenado de acuerdo a las consignas propuestas

Una vez terminado el notebook deberán enviarlo por mail a eea.uba@gmail.com

CRITERIOS DE EVALUACION

- Explicar los procedimientos y decisiones en el texto
- Comentar el código
- Llegar a los resultados esperados
- Recomendamos fuertemente usar las funciones de `tidyverse`

En caso que los resultados no sean los esperados y no logremos identificar las fuentes de error podemos pedirles que nos compartan el archivo .Rmd y ciertas bases de datos que vayan generando.

DATOS

Los datos con los que se trabajará en este TP provienen de la 3° Encuesta Mundial de Salud Escolar (EMSE) provistos por el Ministerio de Salud (**link**) de la República Argentina. Esta encuesta trata sobre temas de salud y hábitos de las personas en la escuela secundaria que pueden impactar en su salud.

Los datasets que se comparten corresponden a un recorte (muestra) del dataset original, luego del tratamiento de valores atípicos e ingeniería de atributos.

Las variables incluidas son:

- **record**: ID de la observación
- **edad**: edad en años
- **genero**: género de la persona
- **nivel_educativo**: nivel educativo en que se encuentra la persona
- **altura**: altura en centímetros
- **peso**: peso en kilogramos
- **frecuencia_hambre_mensual**: variable categórica que indica la frecuencia con la que la persona considera que pasó hambre en el último mes porque no había suficiente comida en su hogar
- **dias_consumo_comida_rapida**: cuántos días comió en un restaurante de comida rápida en la última semana
- **edad_consumo_alcohol**: edad en qué la persona comenzó a consumir alcohol

- **consumo_diario_alcohol**: cantidad de tragos que la persona habitualmente toma por día
- **dias_actividad_fisica_semanal**: cantidad de días que la persona realizó una actividad física por un total de al menos 60 minutos en la última semana

- **consumo_semanal_frutas**: cantidad de veces que la persona consumió frutas en la última semana
- **consumo_semanal_verdura**: cantidad de veces que la persona consumió verduras en la última semana

- **consumo_semanal_gaseosas**: cantidad de veces que la persona consumió gaseosas (al menos un vaso) en la última semana

- **consumo_semanal_snacks**: cantidad de veces que la persona consumió snacks/comida salada en la última semana
- **consumo_semanal_comida_grasa**: cantidad de veces que la persona consumió comidas altas en grasas en la última semana

CONSIGNAS

El objetivo general del trabajo es poder crear una serie de modelos lineales para explicar y predecir el **peso** de los estudiantes según la información que proporciona la EMSE.

1) **Análisis exploratorios**

Leer el archivo "encuesta_salud_train.csv". ¿Qué puede mencionar sobre su estructura y variables?

¿Cómo es la correlación entre las variables numéricas? Utilice y analice en detalle algún gráfico que sirva para sacar conclusiones sobre la asociación de variables realizando apertura por género. En particular, ¿cómo es la correlación entre la variable a explicar (**peso**) y el resto de las variables numéricas?

Para las categorías de la variable frecuencia de hambre mensual, analice gráficamente la distribución en términos de frecuencia relativa de:

a) El consumo semanal de verdura.
b) El consumo semanal de comida grasa.

¿Cuáles son las principales características que observa en estos gráficos?

2) **Modelo inicial**

Se plantea que una primera alternativa para modelar el peso es:

$$E(peso) = \beta_0 + \beta_1 altura + \beta_2 edad + \beta_3 genero + \beta_4 diasActividadFisicaSemanal + \beta_5 consumoDiarioAlcohol$$

¿Cuál es la interpretación de cada uno de los coeficientes estimados? ¿Son significativos? ¿El modelo resulta significativo para explicar el peso? ¿Qué porcentaje de la variabilidad explica el modelo?

3) **Modelo categóricas**

Se sugiere probar un modelo que incopore el consumo semanal de snacks y una interacción entre el género y la edad, en lugar de actividad física y consumo de alcohol:

$$E(peso) = \beta_0 + \beta_1 altura + \beta_2 edad + \beta_3 genero + \beta_4 consumoSemanalSnacks + \beta_5 genero \cdot edad$$

Además se pide explícitamente que la categoría *"No comí comida salada o snacks en los últimos 7 días"* de la variable **consumoSemanalSnacks** se encuentre como nivel/categoría basal.

¿Cuál es la interpretación de los coeficientes estimados para las categorías de **consumoSemanalSnacks** y **genero . edad**? ¿Son significativas? ¿Qué porcentaje de la variabilidad explica el modelo?

En caso de detectar que existen categorías no significativas de la variable **consumoSemanalSnacks** evaluar si la variable es significativa en su conjunto y, en caso afirmativo, proponer una redefinición de las mismas

que permita obtener una mayor proporción de categorías significativas individualmente. Luego, analizar si existen cambios en la variabilidad explicada por el modelo.

4) **Modelos propios y evaluación**

Realizar 2 modelos lineales múltiples adicionales y explicar brevemente la lógica detrás de los mismos (se valorará la creación y/o inclusión de variables nuevas).

Evaluar la performance del **modelo inicial**, el **modelo categóricas** con las categorías redefinidas de la variable *consumoSemanalSnacks* y los modelos desarrollados en este punto en el dataset de entrenamiento y evaluación (usar dataset "encuesta_salud_test.csv"). La evaluación de performance consiste en comparar la performance en términos del R cuadrado ajustado, RMSE y MAE sobre el set de entrenamiento y en términos de RMSE y MAE sobre el set de evaluación.

¿Cuál es el mejor modelo para nuestro objetivo de predecir el peso? ¿Por qué?

5) **Diagnóstico del modelo**

Analizar en profundidad el cumplimiento de los supuestos del modelo lineal para el **modelo inicial**.

6) **Modelo Robusto**

Leer el archivo "encuesta_salud_modelo6.csv". Este último consiste en el dataset original de train con la incorporación de algunas observaciones adicionales que pueden incluir valores atípicos. En particular, observar la relación entre peso y altura ¿Qué ocurre con estos nuevos datos?

Entrenar el **modelo inicial** con estos nuevos datos y comentar qué se observa en los coeficientes estimados y las métricas de evaluación (R cuadrado ajustado, RMSE y MAE) respecto al modelo entrenado con el set de entrenamiento original.

Entrenar un **modelo robusto** con la misma especificación que el modelo inicial sobre los nuevos datos. Comparar los coeficientes y su performance (RMSE y MAE) respecto al modelo inicial no robusto entrenado en este punto. ¿Qué puede concluir al respecto?

Nota: los registros que se suman en este punto son observaciones ficticias que se generaron a partir de observaciones reales del set de datos original.