A1.5 Solución de problemas

Para esta actividad trabajaremos con la base de datos de calificaciones que utilizamos en la lectura interactiva L1.1 Aprendizaje estadístico-automático. Nos interesa tratar de predecir la calificación final de estudiantes de un curso, a partir de su información demográfica y sus calificaciones de los primeros dos periodos.

Utilizaremos el archivo de nombre "A1.5 Calificaciones.csv", donde podrás encontrar información para 395 estudiantes, con 10 variables en total. Los datos se descargaron del Student Performance Data Set en el UCI Machine Learning Repository, y podrás encontrar más información sobre los mismos en el siguiente enlace.

La base de datos cuenta con la siguiente información:

- "**Escuela**". Indica si el estudiante en cuestión asistía a la escuela Gabriel Pereira (GP) o a la escuela Mousinho da Silveira (MS).
- "**Sexo**". F para mujeres y H para hombres.
- "**Edad**". Edad del estudiante, en años.
- "**HorasDeEstudio**". Cantidad de horas de estudio: 1 indica menos de dos horas, 2 indica de dos a cinco horas, 3 indica de cinco a diez horas, 4 indica más de diez horas.
- "**Reprobadas**". Indica la cantidad de materias reprobadas previamente.
- "**Internet**". Si el estudiante tenía acceso (yes) o no (no) a internet en su casa.
- "**Faltas**". Cantidad de veces que faltó a clases.
- "**G1**". Calificación del primer periodo, escala del 0 al 20.
- "**G2**". Calificación del segundo periodo, escala del 0 al 20.
- "**G3**". Calificación final, escala del 0 al 20.

Desarrolla los siguientes puntos en una *Jupyter Notebook*, tratando, dentro de lo posible, que cada punto se trabaje en una celda distinta. Los comentarios en el código siempre son bienvenidos, de preferencia, aprovecha el *markdown* para generar cuadros de descripción que ayuden al lector a comprender el trabajo realizado.

1. Importa los datos del archivo "Calificaciones.csv" a tu ambiente de trabajo. Este archivo lo encontrarás en la misma página donde descargaste esta plantilla. Imprime en consola el tipo de dato de cada variable del *data frame*.

2. Transforma todas las variables categóricas, de forma que los nuevos datos sean útiles para generar un modelo de regresión lineal múltiple. Presta especial atención a variables que, aunque parecen cuantitativas (contienen números), realmente son cualitativas (los números representan una clase). Imprime las primeras 5 observaciones de la base de datos modificada, demostrando que las variables cualitativas desaparecieron y fueron reemplazadas por variables adecuadas.

3. Identifica valores atípicos para la variable "Faltas", utilizando el método de Tukey con k=3. Imprime en consola todas las observaciones que se consideren atípicas, y tras revisar las características de dichas observaciones, agrega una línea de texto que describa qué planeas hacer con dichos valores y por qué. Realiza la acción descrita en caso de ser necesario.

4. Genera una matriz de correlaciones para encontrar potenciales problemas de colinealidad. Genera un *heatmap* para visualizar de forma más sencilla los resultados. Determina si es necesario eliminar una o múltiples variables, explicando tu razonamiento en una línea de texto. Realiza la acción descrita en caso de ser necesario.

5. Incluye términos de interacción para al menos dos pares de variables, las que te llame más la atención analizar con esta metodología. Trata de evitar incluir interacciones para todos los pares de variables posibles. Imprime en consola las primeras 5 observaciones de la base de datos con los nuevos términos.

6. Entrena un modelo de regresión lineal múltiple en un subconjunto de datos que corresponda al 80% de los datos totales e imprime en consola un resumen de los resultados obtenidos. Posteriormente, usando dicho modelo, predice la calificación final del 20% de las observaciones que no se usaron para entrenar. Genera una gráfica de dispersión de las calificaciones finales reales contra las calificaciones finales estimadas por el modelo en los datos de prueba. Agrega una línea de texto donde des una conclusión sobre los resultados del modelo, con base en la información que se muestra en la gráfica.