Ejercicio 1: Descripción de los Datos y la Tarea

Responda las siguientes preguntas:

1. ¿De qué se trata el conjunto de datos?
2. ¿Cuál es la variable objetivo que hay que predecir? ¿Qué significado tiene?
3. ¿Qué información (atributos) hay disponibles para hacer la predicción?
4. ¿Qué atributos imagina ud. que serán los más determinantes para la predicción?
5. ¿Qué problemas observa a priori en el conjunto de datos? ¿Observa posibles sesgos, riesgos, dilemas éticos, etc? Piense que los datos pueden ser utilizados para hacer predicciones futuras.

**No hace falta escribir código para responder estas preguntas.**

1. Los datos (Boston house prices dataset-taken from the StatLib library which is maintained at Carnegie Mellon University) se obtuvieron en 1970. La base de datos en estudio reporta los precios de propiedades (casas) en Boston, e información que estarían vinculada con el valor de dichas casas a través de ciertos atributos/variables/características. La base de datos cuenta con 13 columnas (del tipo numéricos o categóricos), más la variable objetivo que es el valor de la casa. Reporta 506 propiedades (filas) Cada fila de la base de datos nombra un suburbio o una ciudad de Boston.
2. La variable objetivo a predecir es MEDV (Median value of owner-occupied homes in $1000's). la traducción literal seria Valor medio de las viviendas ocupadas por sus propietarios en $ 1000 y representa una una estimación del valor (precio) de casas ocupadas por sus dueños (en miles de dólares)

Para hacer la predicción se dispone de información de 13 features en total, los cuales presentan información sobre distintos características:

* Características económicas y demográficas de la ciudad, como la tasa del impuesto al valor total de la propiedad por $ 10,000 (TAX), la tasa de criminalidad per cápita (CRIM), tasa de estudiantes por docente de la ciudad (PTRATIO), porcentaje de la población de menor nivel socio-económico (LSTAT), y una variable ligada a la proporción de población negra en la ciudad (calculada a partir de la siguiente fórmula: B= 1000(Bk - 0.63)^2 ).
* Características de la zona como la proporción de terrenos residenciales divididos en lotes de más de 25,000 pies cuadrados (ZN), proporción de negocios comerciales no minoristas por ciudad (INDUS), el número promedio de habitaciones por vivienda (RM), la proporción de unidades ocupadas por sus dueños que fueron construidas antes de 1940 (AGE)
* Características de tipo ambiental y de localización geográfica: la cercanía al río Charles (CHAS), concentración de óxidos nítricos (NOX), la distancia ponderada a 5 centros comerciales de Boston (DIS), índice de accesibilidad a las carreteras radiales (circunvalación) (RAD).

1. A priori, se puede pensar que la ubicación es algo importantísimo en la determinación del valor de la propiedad. Por tanto variables como el índice de accesibilidad, cercanía al río charles (problemas de inundación?) y distancia a los centros comerciales podrían tener efecto significativo sobre la MEDV. También es posible que tenga relevancia en la variable target en que zona se encuentra la propiedad (la proporción de negocios no minoristas, la proporción de casas ocupadas por sus dueños. Obviamente el tamaño de la casa ( cantidad de habitaciones) debería tener efecto sobre MEDV
2. Sin conocer las características de la ciudad de Boston, estimamos que hay variables que podrían introducir sesgos y análisis poco éticos de la base dato La indicación del suburbio por ejemplo si bien es importante podría ser vista como una variable que clasificaría la propiedad en buena-mala. Pero más al grano hay variables más negativas desde el punto de vista ético como el porcentaje de población negra (tomando en cuenta que el data set es de 1970, se supone que hoy día no es políticamente correcto y obviamente no ético incluir(o incluso medir) esa variable/característica