Repaso 7

Las características de las 560 muestras en el conjunto de datos de Housing se encuentran resumidas a continuación, y han sido tomadas de la fuente original compartida previamente en <https://archive.ics.uci.edu/ml/datasets/Housing>:

* CRIM: Tasa de criminalidad per cápita por ciudad.
* ZN: Proporción de suelo residencial ocupado para terrenos de más de 2300 m2
* INDUS: Proporción de hectáreas de negocios no minoristas por ciudad.
* CHAS: Variable ficticia de Charles River (=1 si el tramo limita el rio; si no=0).
* NOX: Concentración de óxido nítrico (partes por 10 millones).
* RM: Número medio de habitaciones por vivienda.
* AGE: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
* DIS: Distancias ponderadas hasta cinco centros de empleo de Boston.
* RAD: Índice de accesibilidad a autopistas radiales
* TAX: Tasa de impuesto a la propiedad de valor total de 10000
* PTRATIO: Ratio de profesor por alumno por ciudad.
* B: 1000(Bk- 0.63) ^2, donde Bk es la proporción de personas [de ascendencia afroamericana] por ciudad
* LSTAT: Porcentaje de estatus inferior de la población
* MEDV: Valor medio de casas ocupadas por sus propietarios por 100

Para el resto de esta sección, tendremos en cuenta los precios de las casas (MEDV) como nuestra variable TARGET.

Repaso 8

Caso 1

\*\*Acerca de la información\*\*

El objetivo es construir un modelo para predecir si en cierta zona habrá o no habrá reclamo de clientes (indicentes). Para ello cada registro contiene información de:

\*\*RSRP:\*\* Nivel de Señal medido en dBm. Generalmente una buena señal es de -70 dBm, y luego se va degradando hacia valores muy malos por debajo de -115 dBm. Mientras más negativo es el valor peor es la señal. Con respecto a este valor, es importante señalar que los \*\*valores perdidos\*\* fueron codificados como \*\*0\*\*.

\*\*Redireccion:\*\* Si una red 4G hay redirecciones o transiciones hacia 3G es un indicativo de que hay problemas en la señal. Esta variable indica la cantidad de conexiones que se van a 3G.

\*\*CQI:\*\* Indicador de la calidad del enlace de radiofrecuencia. Los valores van de 1 a 15. 6 o 7 son valores malos, a partir de 12 hay una buena calidad. A menor calidad hay propensión de mayor reclamos.

\*\*Incidentes:\*\* Variable objetivo, que indica la cantidad de reclamos que existen en la zona. Podemos recodificar 1: Reclamo, 0: No Reclamo. Hay valores 0, nulos, vacíos que habría que limpiar. Adicionalmente hay pocos indicentes (Reclamos) con relación a la cantidad de registros, por lo que habría que como tarea previa (preprocesamiento) realizar un \*\*balanceo de datos\*\*.

Caso 2

**Edad**

**Job**

**Educación**

**Meses**

**Duración**

**Y** 0: no se realiza el préstamo //// 1:se realiza el préstamo

Repaso 9

Las características de las 560 muestras en el conjunto de datos de Housing se encuentran resumidas a continuación, y han sido tomadas de la fuente original compartida previamente en <https://archive.ics.uci.edu/ml/datasets/Housing>:

* CRIM: Tasa de criminalidad per cápita por ciudad.
* ZN: Proporción de suelo residencial ocupado para terrenos de más de 2300 m2
* INDUS: Proporción de hectáreas de negocios no minoristas por ciudad.
* CHAS: Variable ficticia de Charles River (=1 si el tramo limita el rio; si no=0).
* NOX: Concentración de óxido nítrico (partes por 10 millones).
* RM: Número medio de habitaciones por vivienda.
* AGE: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
* DIS: Distancias ponderadas hasta cinco centros de empleo de Boston.
* RAD: Índice de accesibilidad a autopistas radiales
* TAX: Tasa de impuesto a la propiedad de valor total de 10000
* PTRATIO: Ratio de profesor por alumno por ciudad.
* B: 1000(Bk- 0.63) ^2, donde Bk es la proporción de personas [de ascendencia afroamericana] por ciudad
* LSTAT: Porcentaje de estatus inferior de la población
* MEDV: Valor medio de casas ocupadas por sus propietarios por 100

Para el resto de esta sección, tendremos en cuenta los precios de las casas (MEDV) como nuestra variable TARGET.

Repaso 11

**Ejercicio KNN**

Para nuestro ejercicio tomaremos 257 registros con Opiniones de usuarios sobre una app (Reviews). Utilizaremos 2 columnas de datos como fuente de alimento del algoritmo. Recuerden que sólo tomaré 2 features para poder graficar en 2 dimensiones, PERO para un problema «en la vida real» conviene tomar más características de lo que sea que queramos resolver. Esto es únicamente con fines de enseñanza. Las columnas que utilizaremos serán:

-Wordcount con la cantidad de palabras utilizadas -SentimentValue con un valor entre -4 y 4 que indica si el comentario fue valorado como positivo o negativo.

Nuestras etiquetas, serán las estrellas que dieron los usuarios a la app, que son valores discretos del 1 al 5. Podemos pensar que si el usuario puntúa con más estrellas, tendrá un sentimiento positivo, pero no necesariamente siempre es así.