TEMA 1

MÓDULO:
TÉCNICAS AVANZADAS DE PREDICCIÓN

PREPARACIÓN DEL TABLÓN DE MODELIZACIÓN

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Entender la necesidad de modelización de un determinado fenómeno.
- Identificar los pasos previos para realizar una correcta modelización.
- Entender el fenómeno que queremos estudiar y las variables con las que contamos y que tendría sentido introducir en nuestro modelo de predicción.

1. PREPARACIÓN DEL TABLÓN DE MODELIZACIÓN

Supongamos que trabajamos en el departamento de finanzas de un banco y necesitamos saber si un nuevo cliente es propenso a impagar, o en el departamento de marketing y necesitamos identificar oportunidades de multiequipamiento, o en el departamento de logística de una empresa de trasportes y necesitamos patrones de buena conducción , o en un equipo farmacéutico de ensayos clínicos, o que una empresa aeronáutica nos pide analizar los resultados procedentes de sensores instalados en sus aviones, o que trabajamos en un equipo de investigación para analizar la incidencia de un componente en el cáncer. Para todos estos casos, además de un correcto conocimiento de la materia particular de cada disciplina, es necesario conocer técnicas de predicción que nos ayuden a entender problemas y **tomar decisiones.**

¿Para qué necesitamos realizar una modelización?

Usando términos matemáticos, podemos plasmar la realidad de forma simplificada en una modelización. Es decir, la modelización es una herramienta para poder desmigar la realidad, a veces demasiado intrincada, y analizar sus partes.

¿Qué objetivos perseguimos cuando modelizamos?

- Queremos entender el pasado: ¿por qué ha ocurrido un suceso y, repetidamente, en una zona?
- Explicar la relación entre variables: ¿de qué manera incrementan las ventas los anuncios publicitarios?
- O bien queremos discriminar o clusterizar unos perfiles: ¿qué clientes son más sensibles a un tipo de publicidad?

Para todo ello, utilizaremos técnicas matemáticas diversas.

¿Qué fases encontramos en el proceso de modelización?

1. Primero tenemos que reconocer el problema.
2. Segundo, veremos si tenemos información relevante para resolverlo.
3. Formalizaremos matemáticamente. Construiremos ecuaciones y trabajaremos las variables.

4. Resolveremos el problema matemático.

5. Interpretaremos el resultado y validaremos el modelo.

Este proceso resultará familiar a quien haya trabajado antes con el método científico porque es muy similar: trabajando los datos disponibles, contestamos una pregunta y extraemos conclusiones.

IMPORTANTE

Vamos a construir modelos probabilísticos o estocásticos:

$$\text{Modelo determinista: } Y = f(X)$$

$$\text{Modelo Estocástico: } Y = g(X, u)$$

donde u es una variable aleatoria no observable y g es una función.

Antes de comenzar el trabajo de modelizado y una vez definido el problema que queremos resolver, necesitamos:

1. **Datos:** en esta sección abordaremos cómo crear un buen tablón de modelización.

2. **Herramientas:** software para tratar y manipular información. (Durante este tema, el código que veremos será en R.).

3. **Conocimiento:** sobre técnicas y algoritmos para poder sacar conclusiones veraces y robustas.

Fuente www.datasciencecentral.com

¿Cuál es nuestro punto de partida?

Partimos de unos datos, bien sean internos o externos. Dichos datos proporcionan un resultado empírico de lo que ya ha ocurrido o de algo que ya existe. Ahora, tenemos que prepararnos para la modelización de los mismos.

Tenemos:

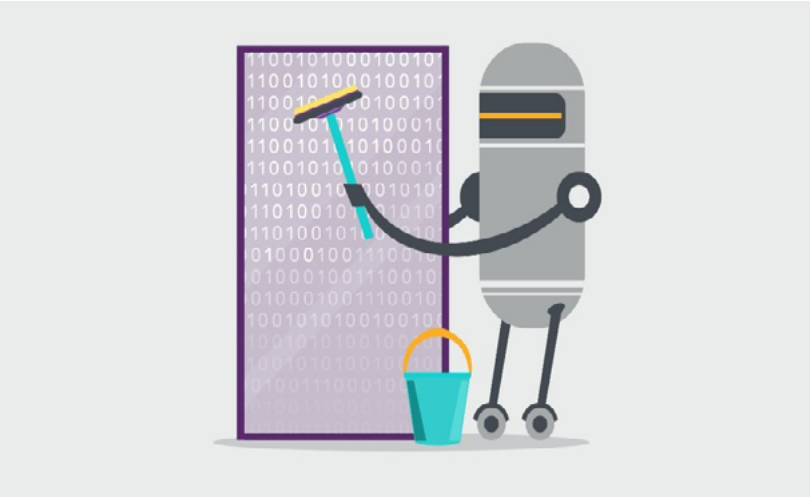
- **Variables explicativas/exógenas:** son las variables que tenemos disponibles para explicar. Las utilizaremos como información para ver su relación con la variable explicada.

- **Variable explicada/endógena:** es la que queremos predecir. Nuestra variable respuesta.

1.1 PREPROCESADO DE DATOS

La cruda realidad: los datos nunca vienen limpios. **¡¡NUNCA!!:**

- **Inconsistencias:** son diferentes métricas en una misma variable (metros / kilómetros / tiempo dentro de una misma variable).
- **Errores:** son los valores vacíos o sin sentido.
- **Sin Fundamento:** son valores que necesitan de un suavizado antes de poder utilizarlos:



Fuente https://bit.ly/2GEahRv

RECUERDA

Una vez tengáis el problema en la cabeza para resolverlo, dedicad horas a meteros en los datos, bucead en ellos en busca de errores, de detalles que puedan viciar el resultado, posteriormente. El tiempo dedicado en este apartado puede resultar tedioso, pero todo depende de esto. Si tenemos datos malos, el resultado va a ser malo por muy buena que sea la técnica utilizada para modelizar.

ACTIVIDAD

SUPER ONLINE

Contamos con los datos de una empresa de compras online (supermercado online). Esta empresa cuenta con poca información sobre sus clientes, puesto que no quiere avasallarlos con cuestionarios ni llamadas. En los próximos apartados, veremos si esta información de la que disponemos es interesante. Contamos con:

- **Compras:** compras medias durante el último año de la persona en la empresa.
- **Edad:** edad actual del cliente.

- **Antigüedad:** nº de años que lleva en la compañía el cliente.
- **Género:** indica si es varón o mujer.
- **Ingresos:** variable externa comprada a un servicio de business analytics que proporciona ingresos estimados de cada persona al mes.
- **Long:** longitud del domicilio del cliente.
- **Lat:** latitud del domicilio del cliente.

Fuente www.periodicopalacio.com

Objetivo: el objetivo que tiene la empresa es el de modelizar la variable compras con las diferentes variables explicativas que contamos. Con esta modelización, podrá definir campañas de marketing en el futuro y ofrecer, selectivamente, descuentos y promociones a aquellos clientes más fidelizados por importe:

- ¿Consumen todos los perfiles más o menos por igual?
- ¿Qué perfiles harán mayores compras en el futuro?
- ¿A qué perfiles tenemos que dirigirnos para ofrecer los productos?
- ¿A quién deberían ir dirigidas las campañas de marketing?

```
tabla<-read.csv("./Data/table_5.01.csv",sep=",")[,-1]
pander(head(tabla), split.cell = 70, split.table = Inf,digits=2)
```

COMPRAS	EDAD	ANTIGUEDAD	GENERO	INGRESOS	LONG	LAT
628	29	1	0	2455	-3.8	41
548	59	2	0	806	-3.7	40
366	31	4	1	2390	-3.7	40
519	45	3	0	1178	-3.8	40
574	59	3	1	1911	-3.6	40
588	49	5	0	2586	-3.7	41

Vamos a utilizar el siguiente árbol:

1. Verificar si hay registros duplicados. Dejar registros únicos.
2. Verificar consistencia de datos:
   - Si hay demasiadas inconsistencias en una variable, eliminamos variable.
   - Si podemos resolverla a mano con alguna regla, modificamos Variable.
   - Si no podemos resolverlo a mano, contemplamos como NA.

3. Tratamiento de los Missing Values:

   - Si hay demasiados NAs en una variable, eliminamos variable.
   - Si los NAs son consistentes en registros determinados, eliminamos registros.
   - Si hay pocos, establecemos regla para rellenarlos.
   - Rellenamos con un aleatorio dentro de la distribución de la variable. Sencillo.
   - Rellenamos con un valor medio. Sencillo.
   - Rellenamos con un modelo de regresión: con los registros que tienen información correcta, intentamos estimar aquellos que no tienen información.

4. ¿Hay información correlacionada? Entonces, eliminamos variables redundantes. (*vid. infra* punto 5.1.3).

5. ¿Necesitamos transformar alguna variable? Entonces, ¡buscamos el sentido estadístico! ¿Nos interesan todos los datos? (*vid. infra* punto 5.1.3).

Los pasos 1, 2 y 3 se han visto a lo largo de otros temas previos al máster. Por lo tanto, nos centramos en los pasos 4 y 5. Asumimos una base de datos trabajada y corregida, anteriormente.

1.2 COMPLETAMOS LA BASE DE DATOS

La estrategia de datos dentro de una empresa es fundamental para definir la dirección que queremos que tomen los proyectos. Tanto la calidad como la cantidad de datos toman un papel clave en el posicionamiento de la empresa. Tener información de calidad nos proporciona una capacidad de aprendizaje y de análisis inigualable. A su vez, nos posiciona fuertes frente a los competidores, en un mercado cada vez más ágil y desarrollado.

Supongamos una empresa aseguradora de coches, cuya misión "analítica"" es la de discriminar a aquellos clientes que van a tener siniestros de los que no van a tener. Cuanta más información tenga sobre los clientes, mejor podrá predecir quienes de ellos van a ser rentables y quiénes no. Igualmente, la competencia hará lo mismo para intentar predecir aquellos perfiles más interesantes, económicamente. Por lo que se produce una guerra por el dato de calidad que posicione a la empresa en un nivel de competitividad elevado. A los perfiles que interesen a la empresa, se les ofrecerá mejores precios, y, a los perfiles que interesen menos, malos precios elevados.

Vamos a distinguir, ahora, entre **información externa e interna.** La **información externa** será siempre mucho más amplia que la información interna que tengamos, pero esto puede ser un arma de doble filo: hay que seleccionar bien la información que queremos trabajar, no siempre es mejor tener grandes cantidades de información si esta no va a ser relevante a nuestras intenciones, o si, finalmente, obtenemos tanta información que resulta abrumadora para trabajar bien los datos; sin embargo, la información externa existe y está ahí, no debemos olvidarla como recurso. Para obtenerla, podemos recurrir a distintas técnicas: compra de datos, APIs o Scrapping.

No obstante, nos vamos a detener un momento en una pieza de **información interna** que se encuentra a disposición de todas las empresas hoy en día, y que ha sido tradicionalmente infravalorada, pero que ofrece grandes posibilidades en cuanto a análisis. Nos estamos refiriendo a la dirección o el código postal del cliente y que merece la pena saber explotarlo para sacarle el máximo beneficio posible.

Conociendo la localización del cliente:

- Sabemos la distancia entre puntos de venta y cliente.
- Posición económica del cliente y sus alrededores.
- Hábitos y comportamiento.
- Estilo de vida.
- Qué le rodea a tu cliente.

Uno de los objetivos de la **modelización** es explicar la relación entre variables:

- ¿De qué manera incrementan las ventas los anuncios publicitarios?
- ¿Incrementa la probabilidad de asistir a urgencias el hecho de vivir cerca de un hospital?
- ¿Afecta al precio de la vivienda vivir cerca de una oficina bancaria?
- ¿Es mayor el precio de la gasolina en zonas donde hay pocas gasolineras?

Resulta sencillo utilizar como datos la dirección o código postal del cliente y los servicios que se encuentran a su alrededor para crear modelos que expliquen estas relaciones.

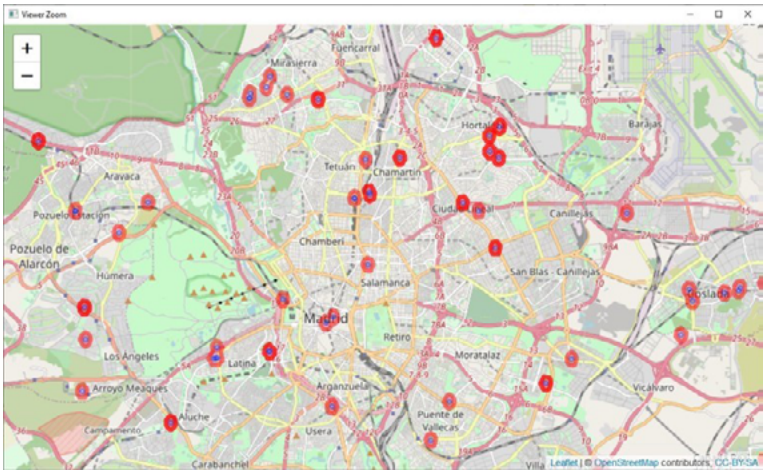
Veamos, pues, una **herramienta** que puede ayudarnos en este cometido. Os presento **Open Street Maps.** Es una herramienta que nos va a ayudar a incorporar mucha información espacial en nuestra base de datos.

Simplemente vamos a descargar los centros comerciales de Madrid para agregar sus referencias geográficas a nuestra base de datos con la que estamos trabajando.

IMPORTANTE

No bajéis información en exceso. Primero pensad la información que tiene sentido estadístico meter en los modelos que, a continuación, veremos (es mejor tener menor cantidad de datos, pero relevantes, que mucha información que nos resulte abrumadora).

```
#Nos vamos a descargar la localización de los cc de Madrid.
datos<-Descarga_OSM(ciudad="Madrid, Spain",key='shop',value = "mall")
#Leaflet(datos[[1]]) %>% addTiles() %>% addPolygons(data = datos[[2]], co
l = "red",label =datos[[3]] ) %>% addCircles()
```



Esta información la podemos adjuntar a nuestra base de datos de múltiples formas.

- Podemos plantear la distancia mínima de cada persona a un centro comercial (Dist_Min).
- Podemos plantear nº de centros comerciales en un radio de 4 km. Es decir, la densidad de centros comerciales en dicho radio. (Dens).

El paquete geosphere te da un cálculo de distancias entre puntos euclídeo corregido por la curvatura de la tierra. Hay muchos otros paquetes o APIs que te devuelven la distancia tanto en kilómetros como en tiempo entre dos puntos. Aquellas que consisten en llamar a una API son mucho más exigentes en tiempos computacionales, así que hay que valorar a priori la que elegimos. Yo os dejo aquí una forma de hacerlo con la función "distm":

```
#Adjuntamos la informacion
coordenadas<-as.data.frame(gCentroid(datos[[2]], byid=TRUE)@coords)
Distancias<-distm(cbind(tabla$LONG,tabla$LAT),cbind(coordenadas$x,coorden
adas$y),fun = distCosine )/1000
tabla$Dist_Min<-round(apply(Distancias,1,min),1)
tabla$Dens<-apply((Distancias<4)*1,1,sum)

pander(head(tabla), split.cell = 60, split.table = Inf,digits=2)
```

COMPRAS	EDAD	ANTIGUEDAD	GENERO	INGRESOS	LONG	LAT	Dist_Min	Dens
628	29	1	0	2455	-3.8	41	6.8	0
548	59	2	0	806	-3.7	40	1.9	5
366	31	4	1	2390	-3.7	40	1.2	4
519	45	3	0	1178	-3.8	40	0.8	4
574	59	3	1	1911	-3.6	40	3.5	1
588	49	5	0	2586	-3.7	41	4.8	0

Para nutrir la base de datos, también vamos a bajar información de hospitales de Madrid:

```
#Nos vamos a descargar la localización de los cc de Madrid.
datos_hospitales< Descarga_OSM(ciudad="Madrid, Spain",key='building',valu
e = "hospital")

coordenadas<-as.data.frame(gCentroid(datos_hospitales[[2]], byid=TRUE)@co
ords)
Distancias<-distm(cbind(tabla$LONG,tabla$LAT),cbind(coordenadas$x,coorden
adas$y),fun = distCosine )/1000
tabla$Dist_Min_h<-round(apply(Distancias,1,min),1)
tabla$Dens_h<-apply((Distancias<4)*1,1,sum)
```

1.3 ÚLTIMOS PASOS ANTES DE MODELIZAR

Recordemos que nuestro objetivo con la base de datos consiste en predecir las compras, por lo tanto, nuestros esfuerzos por dejar la base de datos preparada irán con foco en las compras.

Verificamos la correlación (Pearson) de las variables:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

donde **x** e **y** son el conjunto de variables que estamos analizando.

Más adelante, veremos la razón por la que es tan importante verificar correlaciones antes de modelizar. En el caso de encontrar variables totalmente correlacionadas, las buenas prácticas aconsejan deshacernos de una de ellas antes de empezar a modelizar, puesto que estaríamos incluyendo información redundante en el modelo y podríamos incumplir las hipótesis de partida en las que se basa la regresión, además de que las conclusiones del modelo podrían llevarnos a equívoco.

Ejemplos de correlaciones que podríamos encontrar en un dataset:

- Edad y Antigüedad del carnet de conducir.
- Precio del petróleo y precio de la gasolina.
- Nº de ventas y Nº de clientes.
- Kilómetros recorridos y gasto en combustible.

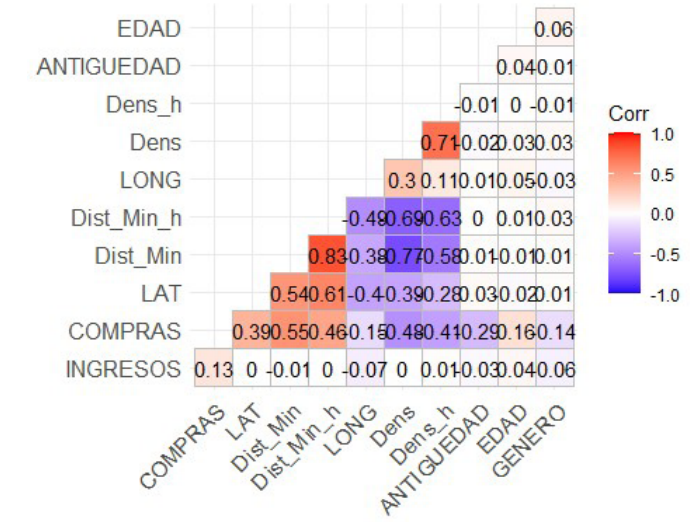
SABÍAS QUE...

No es necesario tener la información de ambas variables explicativas para predecir un determinado suceso. Con saber una de ellas, nuestro modelo tendría información suficiente para sacar el mejor partido a la información.

Además, el análisis de correlaciones nos va a dar una intuición de aquellas variables más influyentes en nuestra variable respuesta/explicada (compras).

Encontramos grandes correlaciones, sobre todo, entre las variables descargadas **Distancia Mínima y Densidad de centros comerciales.** Con elegir una de las dos variables para la modelización, bastará para sacar el potencial deseado:

```
cr <- cor(tabla, use="complete.obs")
ggcorrplot(cr, hc.order = TRUE,type = "lower",lab = TRUE)
```



Nuestro siguiente paso consiste en realizar un análisis univariante de las variables en nuestra base de datos. Visualizaremos las variables en dos sentidos:

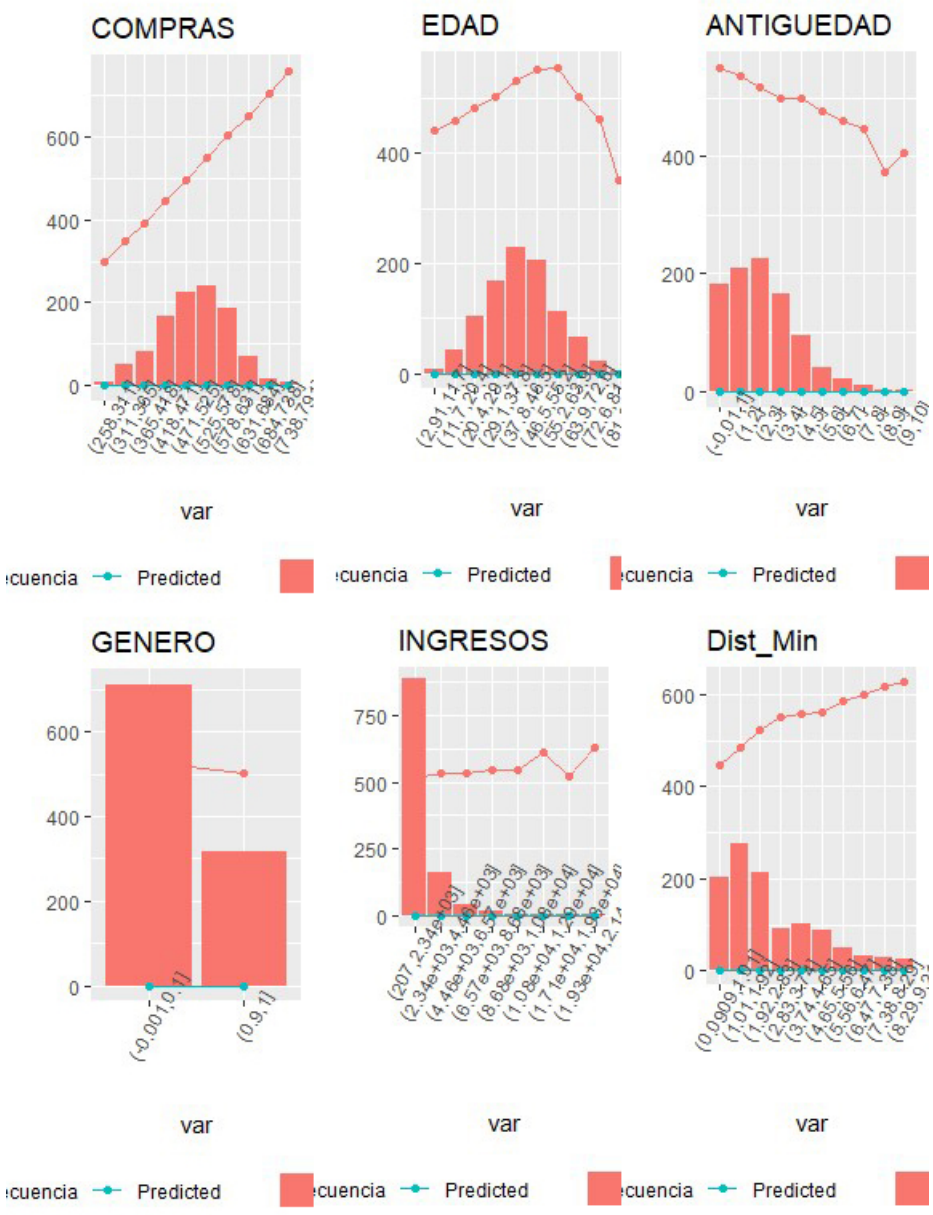
- Análisis de distribución: histograma de la variable.
- Relación de cada uno de los nodos de dicha distribución con respecto a la variable que queremos analizar: compras. Es decir, por cada una de las barras del histograma, vamos a ver cómo se comporta la variable compras, de tal forma que podamos tener una idea, a priori, de la relación entre la variable explicada y las variables explicativas.

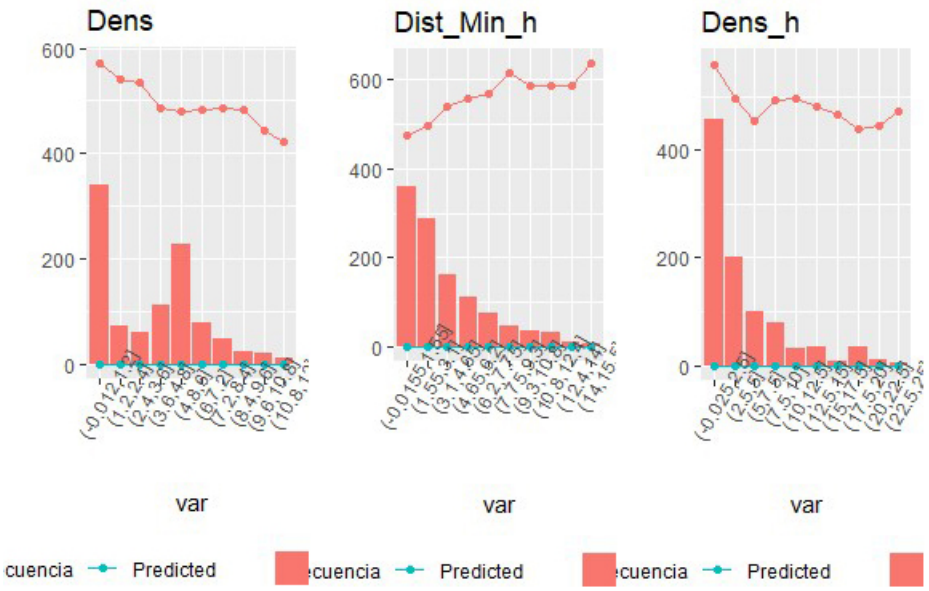
Para ello, os he preparado una función Hist que requiere de 4 inputs:

- El primero es la tabla con la que estamos trabajando.
- El segundo es la variable respuesta que queremos inferir.
- El tercero (aún no hemos llegado), si tuviésemos algún tipo de predicción de la variable respuesta.
- El cuarto es la variable que queremos estudiar en un entorno univariante.

Visualizamos las variables que estamos contemplando. Solo vamos a analizar las que nos interesan inicialmente:

```
tabla1<-dplyr::select(tabla,-LONG,-LAT)
for (i in 1:ncol(tabla1)){
pr<-Hist(tabla1,response = tabla1[,1],predicted = 0,var = tabla1[,i],n=i,
breaks = 10)
plot(pr)
}
```





En los gráficos resultantes, podemos ver, más o menos, las mismas conclusiones que nos daba el estudio de correlaciones. Sin embargo, nos llaman la atención dos comportamientos:

- La variable **edad** tiene una correlación positiva, sin embargo, en edades avanzadas, la gente tiende a gastar menos en sus compras online.
- La variable **ingresos** tiene una distribución con cola larga. Quiere decir que hay muchos datos entre 200 y dos mil euros, pero luego tenemos también datos hasta cantidades de dinero desorbitadas. Ante esto, tenemos que proponer una solución. No es lo mismo pasar de ganar 200 euros al mes a 400 euros al mes, que pasar de ganar 30.000 a 30.200. En teoría, el efecto (comprar más de lo que ya compraba) en el consumidor va a ser, prácticamente, despreciable en el segundo caso, mientras que en el primer caso, donde está multiplicando por dos sus ingresos, puede que tenga un efecto importante.

Entonces, ¿Cómo podemos corregir la variable?

Tenemos tantas opciones como reescalados de la variable se nos ocurran.

Pero, ¿Qué es reescalar una variable?

Pues transformar la variable con cualquier tipo de función que se nos ocurra que tenga sentido matemático y económico.

¿Por ejemplo?

$$variable.reescalada = ln(variable.original)$$

Aunque no hayamos comenzado aún a adentrarnos en el terreno de la regresión, creo que es importante remarcar un aspecto que más tarde necesitaremos.

SABÍAS QUE...

Cuando reescalamos una variable para entender mejor un proceso o sacar mejor partido a los datos, estamos modificando el proceso entero de modelización y la manera de interpretar el resultado cambia.

¿Cómo podemos entonces entender nuestro modelo ante transformaciones de la variable?

IMPORTANTE

Si nuestro modelo es:

$$Y = X\beta$$

- Si no hacemos ningún reescalado, entonces:

$$\Delta Y = \beta \Delta X$$

- Si reescalamos la X:

$$\Delta Y = (\beta/100)\Delta X\%$$

- Si reescalamos la Y:

$$\Delta Y\% = 100\beta\Delta X$$

- Si reescalamos la X y la Y:

$$\Delta Y\% = \beta\Delta X\%$$

¿Puedes poner otro ejemplo de reescalado que no sean logaritmos?

$$variable.reescalada.normalizada = \frac{variable.original - \mu}{\sigma}$$

Siendo μ la media y σ la desviación típica.

Como más de uno estará suponiendo a estas alturas, el reescalado puede dar como resultado datos depurados que nos aportarán información valiosa, o datos que lleven a engaño o imprecisiones. Queda a criterio del analista aplicar reescalados que tengan sentido económico o de negocio.

RECUERDA

Utilizad el buen criterio y conocimiento previo que tengáis sobre los datos para aplicar modificaciones sobre los mismos.

PIENSA UN MINUTO

¿Qué transformación podríamos aplicar a la variable ingresos? Razona la lógica de dicha transformación.

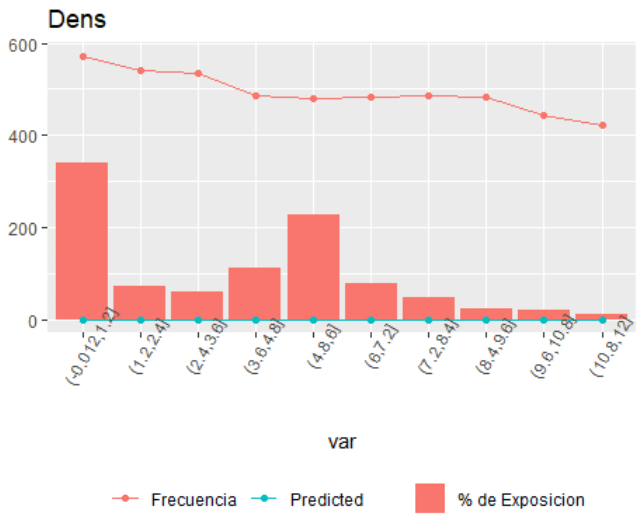
En nuestro caso vamos a aplicar el logaritmo neperiano a la variable **Ingresos.** No es lo mismo sobre 400 euros mensuales añadir 100, que sobre 2000 añadir 100, por lo tanto, un cambio de escala es interesan-

te, de momento, a nivel teórico.

Por ahora, la vamos a dejar calculada, para cuando llegue el momento de **modelizar** y así podremos ver si la variable bruta tiene más o menos sentido que la variable reescalada con su logaritmo natural.

Vamos a visualizar las variables que estamos contemplando. Solo analizamos las que nos interesan inicialmente:

```
tabla$Log_Ing<-log(tabla$INGRESOS)
pr<-Hist(tabla,response = tabla[,1],predicted = 0,var = tabla[,9],n=9,bre
aks = 10)
plot(pr)
```



1.4 MODELOS SUPERVISADOS VS. MODELOS NO SUPERVISADOS.

La mayoría de problemas estadísticos los podemos clasificar como *Supervisados* o *No Supervisados.*
La característica fundamental de los modelos Supervisados es que para cada uno de los registros tenemos una variable respuesta asociada. Nuestro objetivo será crear un modelo que vincule la respuesta con los factores explicativos.

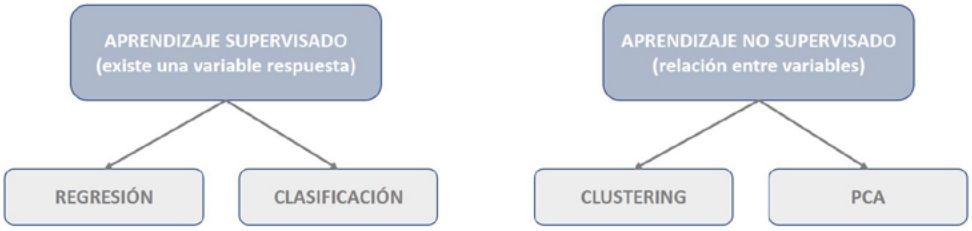
Problemas supervisados:

- Predecir si determinados pacientes tienen una enfermedad, basándonos en su historial clínico.
- Predecir los crímenes de cada ciudad europea, basándonos en el gasto social.
- Predecir el nº de siniestros de coche, basándonos en las características del coche.

Los modelos **No Supervisados,** por el contrario, carecen de variable respuesta. En estos modelos, buscamos entender la relación entre variables o entre observaciones de nuestros datos.

Problemas no supervisados:

- Segmentación de los clientes de una compañía, basándonos en su perfil.
- Simplificación de un conjunto de datos con atributos similares.
- Detección de "outliers" que no encajan en ningún perfil determinado.



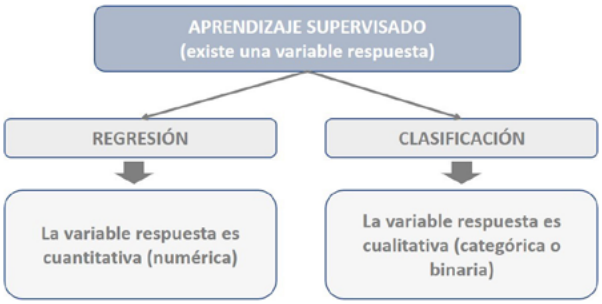
1.5 PROBLEMA DE REGRESIÓN VS PROBLEMA DE CLASIFICACIÓN

En este tema, nos vamos a centrar en aprender técnicas supervisadas, en concreto de regresión y clasificación (a través del modelo lineal).

Normalmente nos referimos a **técnicas de regresión** cuando la variable respuesta, es decir, la variable que queremos explicar o predecir a través de otras variables es cuantitativa.

Una **variable cuantitativa** es aquella que contiene valores numéricos (precio de la vivienda, nº de casos de Covid, tasa de paro).

Cuando nos referimos a **técnicas de clasificación,** es porque la variable respuesta es **cualitativa,** que quiere decir categórica o que contiene clases: tipos de fruta, género, colores). Cuando la variable que queremos predecir es binaria, también consideramos dicho problema como problema de clasificación. (Si,No / 1,0).



IDEAS CLAVE

- La modelización estadística es una representación de la realidad de manera simplificada donde buscamos entender un determinado fenómeno a través de toda la información disponible.

- La información con la que contamos en nuestras bases de datos es la suma de la información directamente asociada al fenómeno a estudiar y al esfuerzo por recopilación de información externa que hagamos. La labor del analista consistirá decidir con qué variables se puede trabajar y merece la pena almacenar.

- Los problemas supervisados los podemos dividir en problemas de regresión y de clasificación.