**Lección 3: Características y tipos de datos**

Fuente: zetta.

Cuando hablamos de datos, debemos tener en cuenta que estos también pueden ser objetos o cualquier tipo de cosa que pueda ser utilizada como dato para hacer que el sistema aprenda. Dichos datos no pueden ser ingresados al sistema en su forma original, estos deben ser previamente procesados para que el sistema haga uso de la información concreta en la cual el usuario se encuentra interesado.

Cada dato posee diferentes características que lo describen y que representan sus propiedades, por ejemplo: si se habla de una persona, esta posee características como color de ojos, estatura, edad, estudios, etc.; debido a esto, al analizar un tema específico se debe determinar cuál o cuáles de estas características se adecuan al tema de análisis y a las necesidades de las partes interesadas. Esta actividad es responsabilidad del experto, el cual se encarga de brindarle la información al sistema.

Dado que esta selección de características es realizada por un individuo, puede darse un margen de error en la misma, para ello existen **técnicas estadísticas** que ayudan a determinarlas y en muchas ocasiones se descubren características que son representativas dentro del tema de estudio, aunque el experto considere lo contrario.

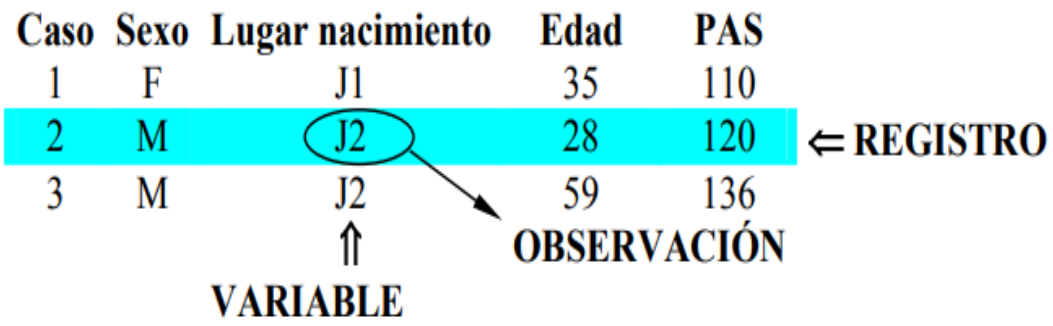
Los datos se pueden extraer de diversas fuentes, imágenes, videos, grabaciones, etc. Por ejemplo, se habla de analizar un video, será necesario tener un sistema que tenga la capacidad de extraer información de allí para posteriormente procesarla y aprender de la misma. La manera como se haga uso de los datos dependerá del tema que se esté procesando, debido a ello habrá datos que deberán ser clasificados o etiquetados para que la información que estos estén brindando se ajuste al tema de estudio, por ejemplo, si se tiene un conjunto muy grande de noticias, pero el usuario se encuentra interesado en aquellas que traten de temas económicos, el mismo deberá clasificarlas y así filtrar aquellas que son de su interés.

Es posible identificar tipos de datos dentro de una base de datos, determinar su tipo permite entender sus características para así lograr clasificarlas, por ello a continuación se presentan las características de los datos y sus tipos.

**Características de los datos**

1. **Unidad de análisis o de observación:** la unidad de análisis puede ser una región, un país, una institución, un objeto o persona. Todo aquello a lo que se le pueda extraer información y configurar una base de datos podrá ser objeto de análisis.
2. **Variable:** la variable corresponderá a cualquier característica de la unidad de análisis que se desee estudiar, donde dicha característica puede ser convertida en un número.
3. **Valor de una variable:** es la medición o número que logra describir la característica de interés de la unidad de análisis.
4. **Caso o registro:** corresponde al número de mediciones que se realizan sobre la unidad de análisis.

Para entender mejor estas características se presenta el siguiente ejemplo:




Fuente: UBA

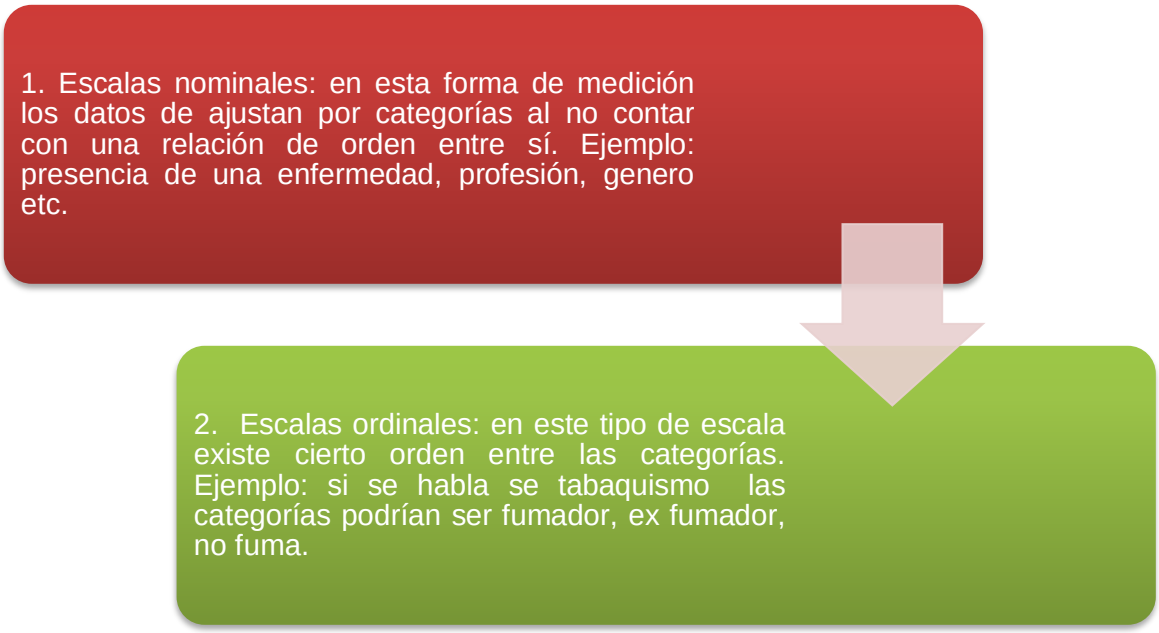
Cuando se maneja un conjunto de datos, es importante identificar la cantidad de variables que están registradas y de qué manera fueron registradas, esto va a permitir la definición de una estrategia de análisis. En el ejemplo que se acaba de mostrar se puede notar que algunas variables son números y otras son letras que muestran categorías. Estos diferentes tipos de datos se presentan a continuación.

**Tipos de datos**

1. **Datos cualitativos o categóricos**

Los datos cualitativos o categóricos hacen representación de un atributo o cualidad que lo que hace es clasificar a cada caso en una de varias categorías. Por ejemplo, el caso más sencillo es cuando se clasifica en uno de dos grupos como enfermo/sano, fumador/no fumador. Este tipo de datos son llamados binarios o dicotómicos. En ocasiones este tipo de clasificación no es suficiente cuando se presenta la necesidad de dar más opciones por ejemplo al hablar del color de ojos, grupo sanguíneo, etc.

Existen dos escalas que pueden ser utilizadas al medir este tipo de variables:




Es muy importante lograr hacer esta distinción entre datos cuantitativos discretos o continuos para tomar una decisión a la hora de seleccionar un método de análisis estadístico, dado que algunos solo están diseñados para datos continuos.

Aunque esta clasificación logra mostrar los tipos de variables que se pueden presentar a la hora de desarrollar un estudio, no se pueden desconocer otros tipos de variables que también son utilizadas como:

- Porcentajes: este tipo de datos son el resultado de tomar el cociente entre dos cantidades, por ejemplo, peso observado/ peso deseable. Este tipo de datos pueden ser considerados variables continuas, pero podrían causar problemas a la hora de desarrollar el análisis e especial al tomar valore superiores o inferiores que 100%.
- Escalas análogas visuales: este tipo de variables se presentan al tratar de conseguir categorías ordinales de una característica no medible, por ejemplo, pedirle al encuestado que indique un nivel de satisfacción, dolor o bienestar., para ello se utiliza la escala visual.

- Scores: estos son utilizados como indicadores de la condición en que se encuentra un individuo basándose en la observación de variables seleccionadas, las cuales generalmente son categóricas. Por ejemplo: la condición de un paciente se determinado la dar puntuación en base a síntomas y signos.
- Datos censurados: se dice que una observación es censurada cuando no pudo ser medida exactamente, pero que se conoce que está más allá de cierto límite, es decir, se conoce una cota inferior o superior para el dato. Por ejemplo: los estudios de seguimiento para conocer el tiempo de supervivencia.

Como se ha mencionado, es de vital importancia tener pleno conocimiento de que tipos de datos se están utilizando para poder definir la forma correcta como estos deben ser manipulados, de ello dependerán los resultados que se esperan obtener. Clasificar y definir las variables le permitirá al sistema de aprendizaje automático tener un proceso de aprendizaje óptimo, pues este lleva a cabo su trabajo con la información que le sea provista pero la calidad de dicha información dependerá del experto o plataforma que se la esté suministrando.

**Referencias:**


- Orellana, L. (marzo de 2001). Recuperado el 03 de octubre de 2018, de http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf
- Pita Fernández, P. D. (06 de marzo de 2001). Estadística descriptiva de los datos. Recuperado el 03 de octubre de 2018, de https://www.fisterra.com/gestor/upload/guias/10descriptiva2.pdf