**TAREA 46**

**estadística-descriptiva | análisis-exploratorio | dato-estructurado | dato-semiestructurado | dato-noestructurado**

**Estadística**: La estadística es una parte de las matemáticas, nos permite el análisis de los datos, de manera que sea más fácil entenderlos y sacar conclusiones. Para poder sacar conclusiones necesitamos: recolectar datos, organizarlos, describirlos, presentarlos y por último, concluir o generalizar. A través de la estadística conocemos mejor en mundo que nos rodea y gracias a ella podemos tomar decisiones efectivas o hacer propuestas de mejora.

**Estadística descriptiva**: Es una rama de la estadística que se encarga de informar sobre lo que pasa en un momento dado con una muestra concreta, se basa en la recolección, organización, descripción y presentación de los datos.

**Estadística inferencial**: Es la rama de la estadística que se encarga de inferir o predecir lo que va a suceder en una población de la que hemos extraído una muestra, es en esta muestra donde hemos recopilado los datos, a través de test matemáticos y de probabilidad somos capaces de hacer una predicción y también de calcular el porcentaje de error que podemos cometer.

**Análisis exploratorio**: en el análisis de datos existen varias etapas fundamentales

* **Objetivo**: el primer paso para un buen análisis es establecer desde el principio cuál es el objetivo del mismo, para qué lo vamos a llevar a cabo, que es lo que quiero saber.
* **Recolección de datos**: todo el análisis posterior va a depender de esta etapa, es **necesario tener claro qué tipo de datos necesito y como los voy a conseguir**
* **La limpieza de los datos**: también es necesario definir en qué formato voy a guardar los datos, qué observaciones me interesa guardar y cuáles no, crear si es necesario nuevas variables combinando las ya existentes.
* **La exploración:** en esta etapa sacaremos el máximo partido de nuestros datos, los presentaremos en tablas o gráficas y los describiremos mediante la estadística descriptiva, para descubrir que es lo que nos tienen que decir.
* mediante técnicas de **estadística inferencial** como el contraste de hipótesis y los cálculos de las probabilidades de error, podemos trasladar lo que hemos observado en la muestra a la población de la que la hemos extraído.
* **Las conclusiones**: tras el análisis, se enumeran las conclusiones a las que hemos llegado en el estudio, lo que ayudará a la toma decisiones.

**Dato estructurado**: Son los datos con una estructura conocida, siguen un patrón, normalmente se organizan en tablas con columnas y filas y estas tablas se pueden relacionar. Son los datos propios de las bases de datos relacionales que se consultan a través del lenguaje SQL. Son los datos más fáciles de analizar. Suelen ser archivos de texto plano, con filas y columnas.

**Dato semi-estructurado**: En este caso los datos no siguen una estructura predeterminada, pero sí los podemos organizar a través de etiquetas, que permite agruparlos por jerarquías. También se los conoce como datos no relacionales o no SQL. Algunos ejemplos son los archivos HTML, JSON, correos electrónicos, bases de datos no relacionales.

**Datos no estructurados**: todos los datos que no pertenecen a ninguna de las categorías anteriores, es decir, la mayoría de datos. Pueden tener o no una estructura interna, pero no atienden a ningún patrón.

Pueden ser datos de texto o no textuales, pueden haber sido generados por una máquina o por una persona, se pueden almacenar en una base de datos NoSQL, o directamente en un “datalake”.

Algunos ejemplos son: ficheros de texto tipo Word, archivos pdf, imágenes, videos, audios, datos de redes sociales, datos Web, datos redes sociales.