



Profundizar en el análisis exploratorio de los datos mediante un ejercicio práctico.

Analizar datos del Covid-19

Últimamente se escucha mucho que los **datos** son el petróleo del s. XXI y no les falta razón ya que a partir de esta materia prima se debería poder extraer conocimiento (si los datos tienen la calidad suficiente):

- 1.- Imagina que a partir de los sensores de un tren somos capaces de simular el desgaste futuro de los rodamientos y hacer un mantenimiento predictivo que evite accidentes. Todo esto gracias al análisis de los datos.
- 2.- Imagina que...

Son increíbles las posibilidades que puede ofrecer un correcto análisis de los datos y la IA está muy ligada a este análisis.

Sí, los datos son la materia prima sobre la cual se **construyen las soluciones de IA** pero antes es importante hacer un primer análisis exploratorio de los datos para conocer bien el entorno y después ya se pasará de construir modelos "inteligentes". Ese primer análisis estadístico-descriptivo es importante para adquirir el conocimiento de campo necesario para posteriormente modelar y desplegar soluciones de IA. Si no somos capaces de entender el dato objeto de estudio difícilmente podremos construir modelos fiables de IA.

Existe otro aspecto que también es importante, **la calidad de los datos**: Disponer de datos que reflejen aquella realidad que se quiere acometer es primordial en la mayoría de soluciones de IA (no siempre). Si no se dispone de datos históricos fiables habrá que generarlos.

Además, en muchas ocasiones, las fuentes de datos pueden ser muy diversas (*máquinas industriales, redes sociales, sistemas de gestión empresarial, entrevistas sociológicas, etcétera*) y suele ser necesario dar sentido a todas esas fuentes y organizarlas bien para construir más adelante modelos probabilísticos que nos sirvan para "predecir el futuro", por poner un ejemplo.

Hoy en día, existe un interesante movimiento entorno a los datos que busca poner a disposición de la gente datos de diferentes fuentes: es el movimiento opendata y actualmente prácticamente todas las instituciones públicas y muchas privadas están publicando en abierto muchos datos que se pueden explotar:

- 1.- NASA
- 2.- OMS
- 3.- New York
- 4.- ...

Para la siguiente tarea disponemos de datos abiertos sobre los efectos del Covid-19 en Araba, Bizkaia y

Gipuzkoa. Se pide que analicéis estos datos y que expliquéis que a ocurrido con el Covid-19 en estas provincias. Debéis subir vuestro trabajo en un PDF a vuestro repositorio. Recuerda que tal y cómo nos indican en el [siguiente enlace](#) las etapas que se deben seguir para un correcto análisis de los datos son los siguientes:

El Problema. Todo nace de un problema, de una necesidad real. En esta etapa definirás el foco del estudio.

La Recolección. Es el diseño de un método para obtener DATOS Más técnicamente es un proceso de experimentación. Puede ser una encuesta, pruebas en laboratorios, con pacientes, nutrirse de datos de marketing en redes sociales etc?

La Limpieza. Homogeneizar los datos en cuanto a formato, deshacer observaciones que no te interesan y almacenar las más útiles.

La Exploración. Utilizando la estadística descriptiva podrás traducir los DATOS a gráficos y características que puedas interpretar.

El Análisis. Con la ayuda de la estadística inferencia encontrarás conclusiones de un conjunto grande (población) con la información de una pequeña parte de este conjunto (muestra). El contraste de hipótesis es la herramienta más famosa de esta etapa. Pero existen otras técnicas como: la predicción, la clasificación o los métodos de causa-efecto. Entre otros.

La Conclusión. Interpretarás los resultados del análisis y listarás las conclusiones. Estarás mucho más cerca del objetivo que te habrás planteado al inicio.

En vuestro trabajo debéis:

- 1.- Ser objetivos y neutrales en el análisis.
- 2.- Debéis ofrecer una visibilidad completa del fenómeno Covid-19 en Araba, Bizkaia y Gipuzkoa.
- 3.- Debéis procurar describir tendencias y relaciones entre las variables objeto de estudio.
- 4.- El documento debe ayudar al lector a la fácil comprensión del tema y a interpretar los resultados de modelos estadísticos más complejos mediante modelos simples.
- 5.- El documento no debe extenderse más de 20 páginas y debe contener un índice.

Para realizar una presentación de calidad os podéis ayudar de las herramientas que consideréis oportunas. Hoy en día, por ejemplo, se dispone de diferentes plataformas para la visualización descriptiva de los datos: [PowerBI](#), [tableau](#), [Qlik](#), etcétera. Incluso con el propio [excel](#) o su análogo gratuito el [Apache Open Office Calc](#) se pueden hacer cosas muy interesantes.

El propio python también tiene muchas librerías para el análisis descriptivo de datos como por ejemplo: [pandas](#) (para la manipulación de tablas), [matplotlib](#) (para la generación de gráficos), ...

¿Nos podéis ayudar a entender mejor qué a ocurrido con el Covid-19 en Araba, Bizkaia y Gipuzkoa?

NOTA:

- 1.- Para realizar esta tarea se recomienda tener acaba la tarea "Estadística descriptiva para novatos".
- 2.- Los ficheros que se añaden son una referencia. Recomendamos trabajar con datos más actualizados descargados desde este enlace: [datos abiertos](#)

#HASHTAGS *(etiquetas de ayuda para búsqueda de información relevante)*

#estadística-descriptiva #media #moda #mediana #campa-de-gauss #plataformas-de-BI #análisis-exploratorio #calidad-del-dato #opendata

LINKS DE INTERÉS

<https://www.youtube.com/watch?v=XzeBtodosKQ>
https://www.youtube.com/watch?v=W1_eCwuYkAI
<https://www.youtube.com/watch?v=CmSZwmoGnJY>
<https://www.youtube.com/watch?v=cucK3A0BXEk>

DICCIONARIO

herramientas-de-Business-Intelligence | calidad-del-dato | análisis-exploratorio | estadística-descriptiva |
histograma | gráfico-de-barras | gráfico-circular | gráfico-de-dispersión | gráfico-de-caja-y-bigotes |
cartograma

PUNTUACIÓN

Programación: 3

Redes: 1

Seguridad: 1

Algoritmia: 6