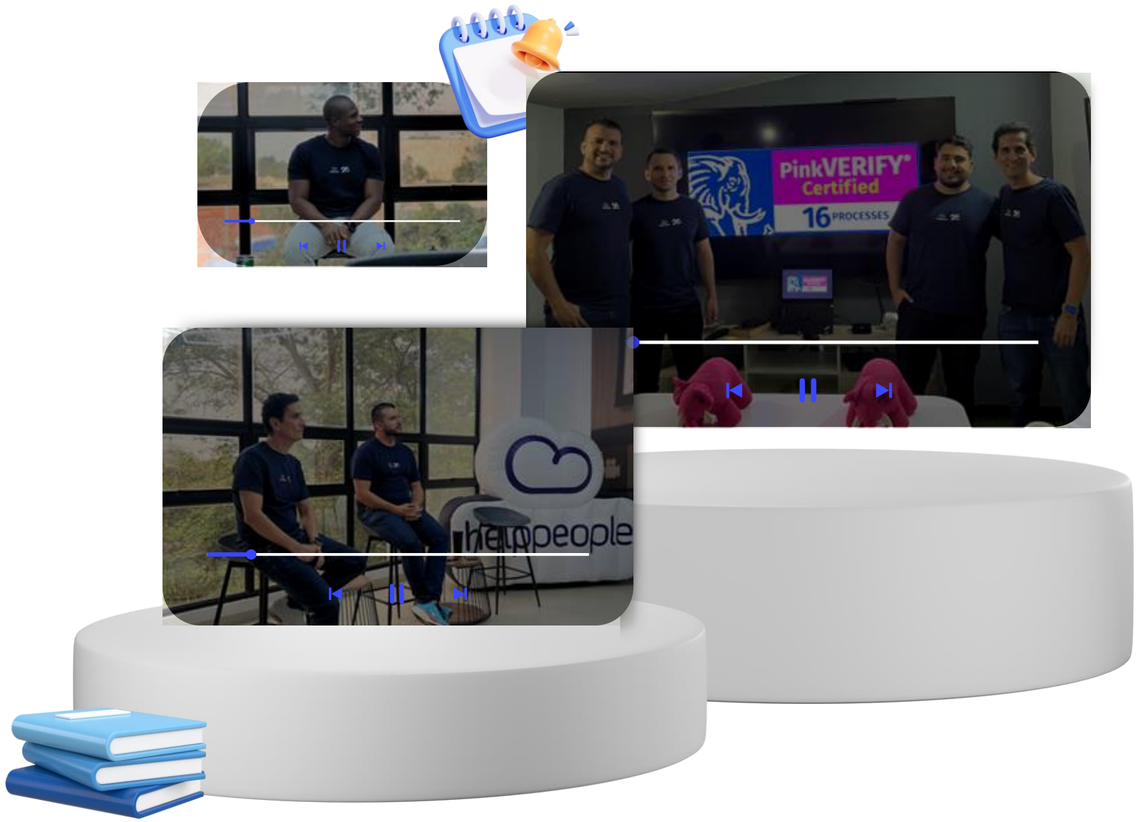
Transformando el caos en información valiosa

Limpieza de datos.

Bloque 4

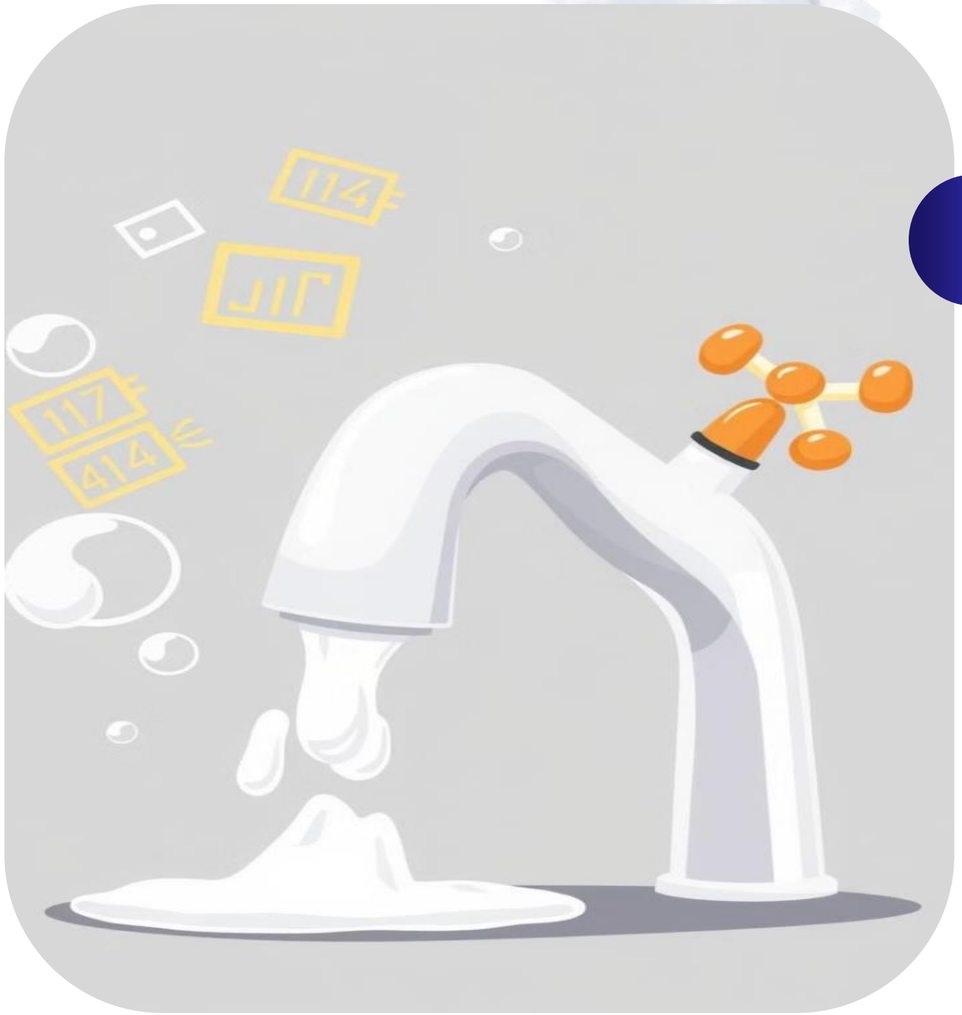


helppeople | Universidad de San Buenaventura

Bloque 4

¿Por qué limpiar datos ?

De la basura a la información útil



helppeople | Universidad de San Buenaventura

Bloque 4

En el mundo de la inteligencia artificial, la calidad de los datos es fundamental. Si alimentamos un modelo con datos desordenados, inconsistentes o incompletos, sus resultados serán inexactos y poco confiables. La limpieza de datos es el proceso que garantiza que la información utilizada sea precisa, estructurada y lista para el análisis.

helppeople | UNIVERSIDAD DE SAN BUENAVENTURA

La mala calidad de los datos afecta la IA más de lo que crees.

Un modelo de IA que entrenó a un asistente virtual para reclutamiento de personal discriminaba a mujeres porque los datos de entrenamiento contenían sesgos históricos.

Bloque 4

Problemas comunes en datos sucios

helppeople | Universidad de San Buenaventura

Bloque 4

Datos faltantes

Los datos faltantes son uno de los problemas más comunes en el procesamiento de datos. Ocurren cuando algunas observaciones están incompletas, lo que puede distorsionar los resultados. Para manejar esta situación, existen diversas estrategias: eliminarlos, estimar su valor mediante métodos estadísticos o marcarlos como desconocidos.

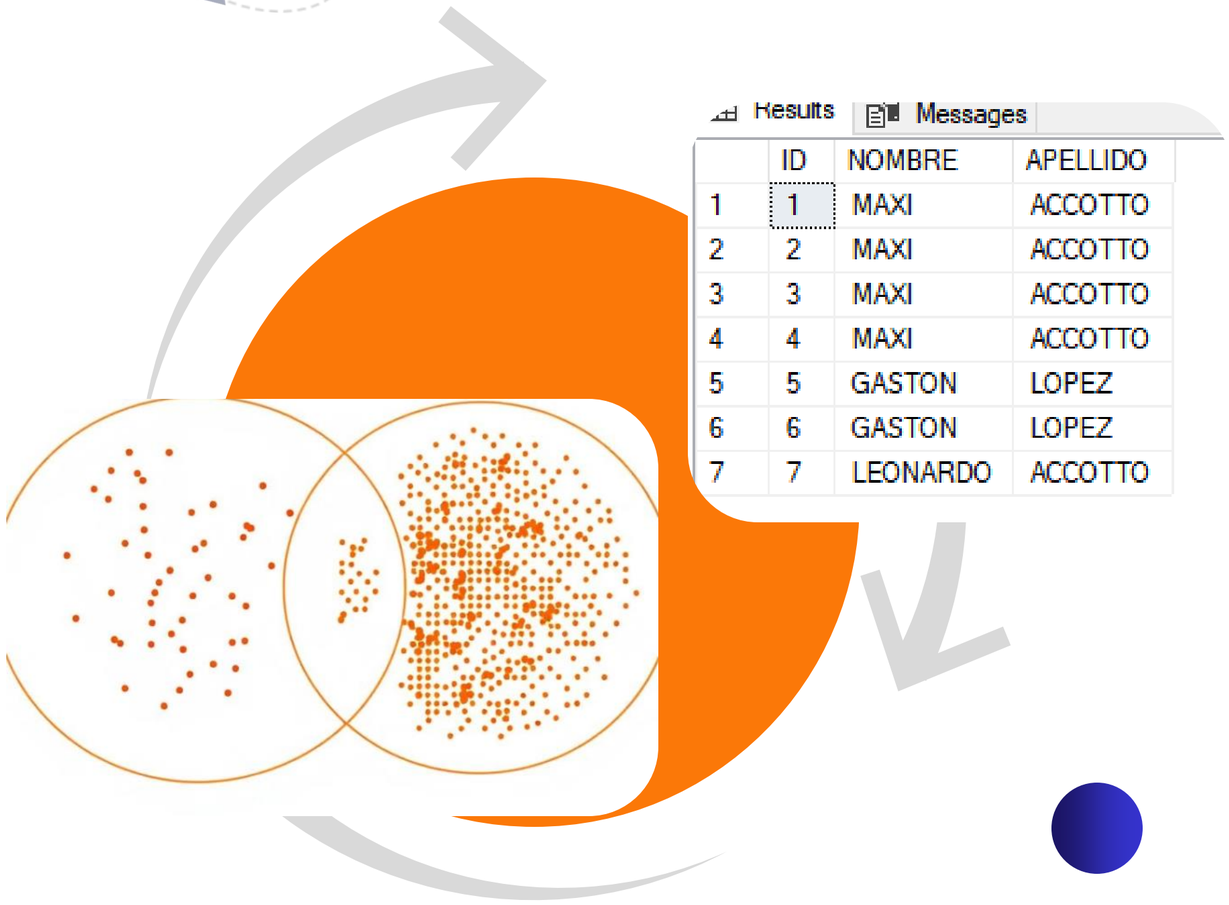
helppeople | Universidad de San Buenaventura

Según estudios, los científicos de datos dedican hasta el 85% de su tiempo a preparar y limpiar datos antes de analizarlos.

Bloque 4

Datos duplicados

Cuando los datos se repiten dentro de un conjunto, los resultados del análisis pueden verse distorsionados. La duplicación ocurre por errores de entrada, integraciones incorrectas o procesos automatizados deficientes. Empresas como Amazon han reportado que hasta el 15% de sus reseñas de productos pueden ser duplicadas, lo que afecta la calidad de sus recomendaciones



Results | Messages

	ID	NOMBRE	APELLIDO
1	1	MAXI	ACCOTTO
2	2	MAXI	ACCOTTO
3	3	MAXI	ACCOTTO
4	4	MAXI	ACCOTTO
5	5	GASTON	LOPEZ
6	6	GASTON	LOPEZ
7	7	LEONARDO	ACCOTTO

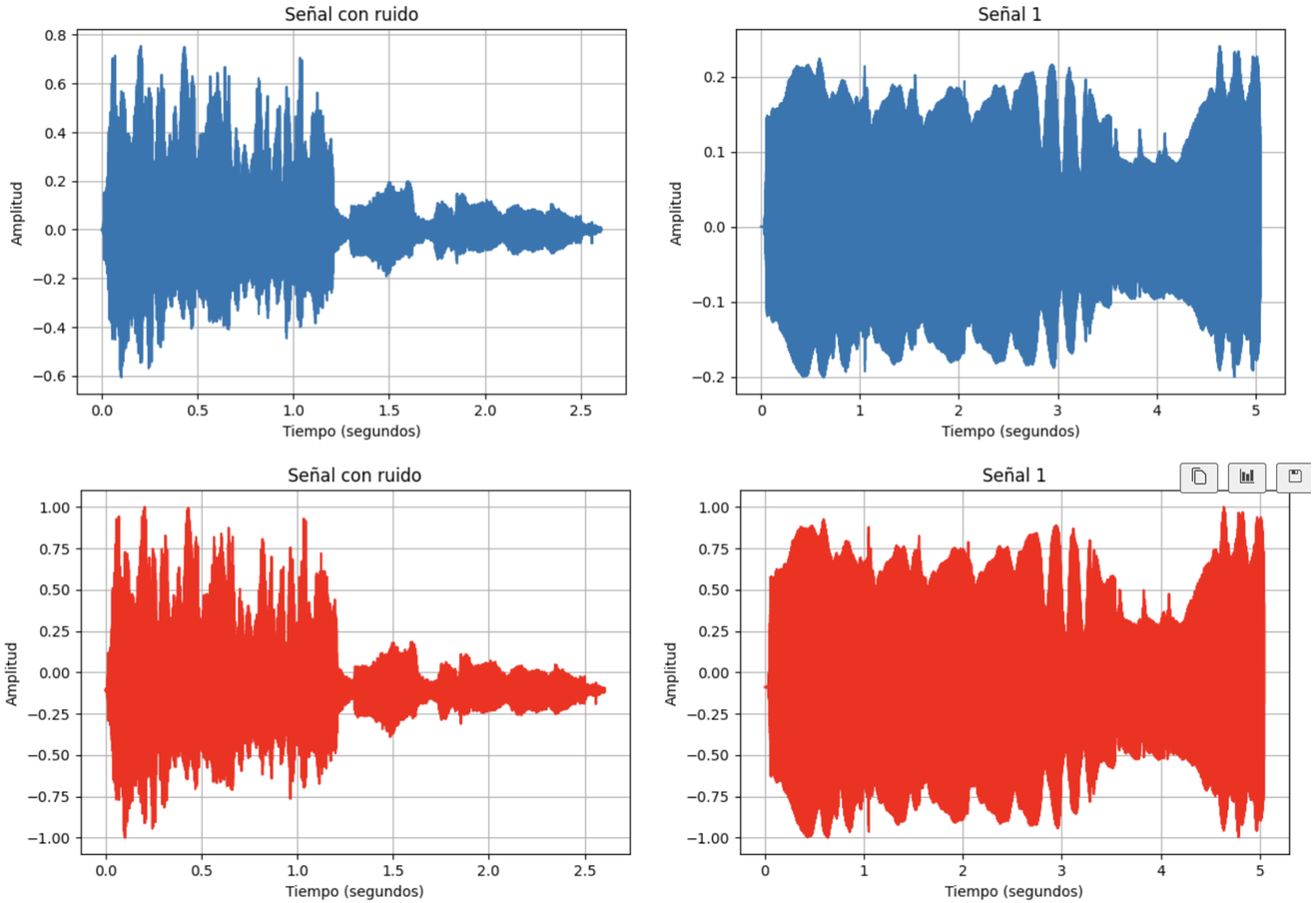
helppeople | UNIVERSIDAD DE SAN BUENAVENTURA

Se estima que en EE.UU. las empresas pierden más de 3 billones de dólares al año debido a errores en los datos, como registros duplicados o información inexacta.

Técnicas de limpieza

Normalización

Los datos provienen de diversas fuentes y, sin normalización, pueden presentarse en formatos inconsistentes. Un mismo país puede estar registrado como "EE.UU.", "USA" o "Estados Unidos", generando problemas en el análisis. La normalización unifica estos formatos y estandariza la información.



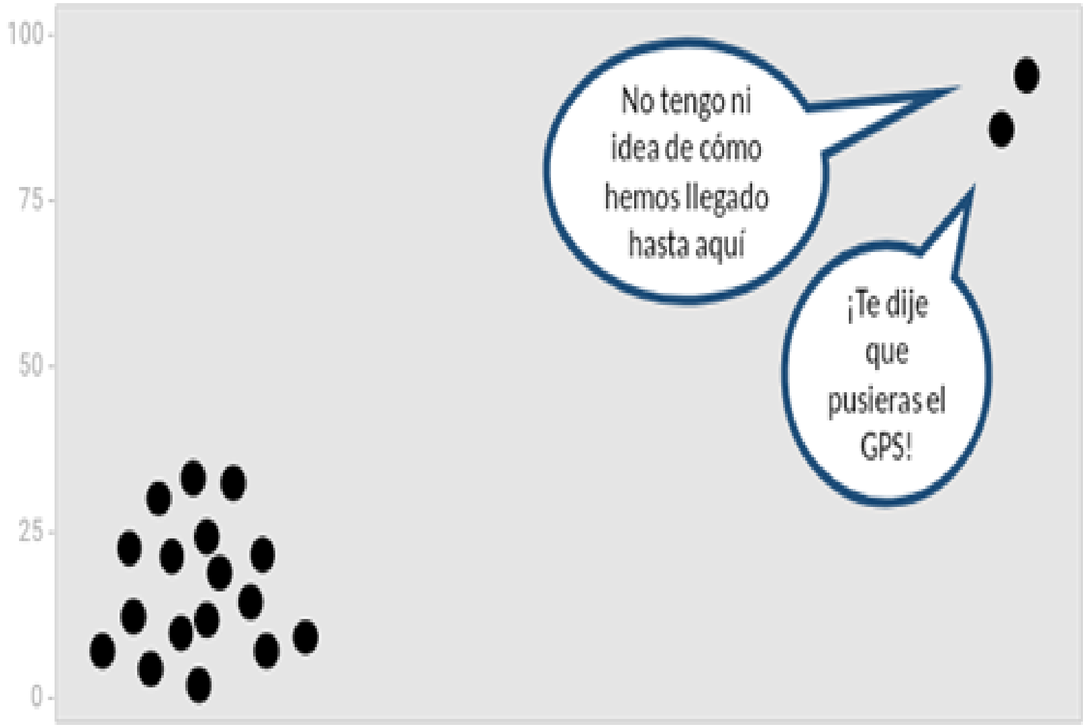
La normalización de datos puede cambiar resultados electorales

En 2000, miles de votos en Florida no fueron contados correctamente debido a problemas en el formato de datos, lo que influyó en la elección presidencial de EE.UU.

Outliers

Datos que desafían el patrón

Los outliers son valores que se alejan significativamente de la norma y pueden indicar errores o patrones ocultos en los datos. Por ejemplo, en un registro médico, una temperatura de 60°C es claramente un error, mientras que una de 41°C podría indicar una condición crítica. Es crucial identificar si estos valores deben eliminarse o si aportan información relevante.



Los valores atípicos pueden hacer historia

- En 2009, los datos de tarjetas de crédito revelaron un gasto inusual en cafeterías por parte de empleados de Goldman Sachs, lo que terminó ayudando a predecir la crisis financiera.

Herramientas automáticas

Limpiar datos manualmente puede ser un proceso tedioso y propenso a errores. Afortunadamente, existen herramientas como OpenRefine, Pandas y Trifacta que automatizan gran parte del trabajo.

Implementarlas en el flujo de trabajo garantiza mayor precisión y rapidez en la preparación de datos para su análisis.

Herramientas como OpenRefine, Pandas o Trifacta hacen el 80% del trabajo. Es como tener una lavadora de datos

Google Maps limpia más de un millón de direcciones al día

Para evitar problemas con nombres de calles duplicados o mal escritos, Google Maps estandariza y corrige diariamente enormes volúmenes de datos.

Bloque 4

Conclusiones

La limpieza de datos es un paso fundamental en cualquier proyecto de inteligencia artificial, ya que la calidad de los datos determina la precisión y confiabilidad de los modelos. Identificar y corregir problemas como datos faltantes, duplicados, formatos inconsistentes y valores atípicos permite extraer información valiosa y tomar decisiones basadas en evidencia. Además, el uso de herramientas automatizadas optimiza este proceso, reduciendo errores y mejorando la eficiencia. Ahora que conoces la importancia de los datos limpios.

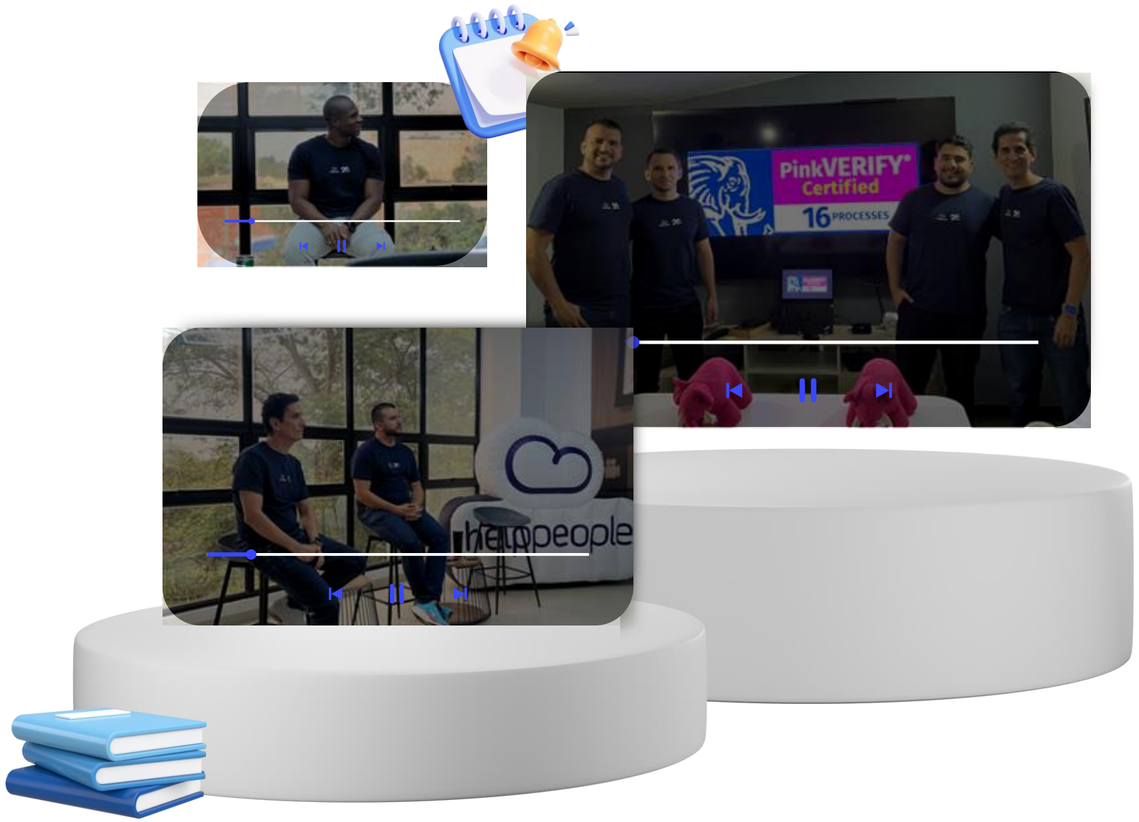
¿cómo crees que la mala calidad de los datos puede afectar decisiones críticas en sectores como la salud o las finanzas?

Explora herramientas como OpenRefine o Pandas y comienza a limpiar tus propios datasets para mejorar tus análisis

helppeople | Universidad de San Buenaventura

Hoy aprendimos a convertir datos sucios en información valiosa.

Recuerda realizar las actividades propuestas para el bloque, revisar el material complementario y completar la evaluación.



helppeople | UNIVERSIDAD DE SAN BUENAVENTURA

¿El siguiente paso?

Aprender a contar historias con ellos. ¡En el próximo bloque descubriremos el arte de la visualización de datos!"

¡Nos vemos en el último bloque!

helppeople | Universidad de San Buenaventura