


Metodología para Ciencia de Datos



Rubén Pizarro Gurrola

rpizarro@itdurango.edu.mx

Enero 2025

Problema a resolver o Áreas de oportunidad?



Las situaciones, problemas, deficiencias, áreas de oportunidad se pueden abordar y/o resolver con datos?



¿Qué hay que resolver, o sobre que área de oportunidad trabajar?



¿Qué análisis realizar?

Análisis exploratorio y descriptivo,
Análisis Predictivo,
Análisis preescriptivo

Contexto de los datos



Contexto de los datos

- ¿en dónde están los datos?
- ¿en que formatos y en que lugar, herramienta?
- ¿qué datos existen?
- cómo están los datos, estructurados, no estructurados?
- ¿quién los tiene? O ¿a quien le pertenecen?
- ¿qué significado tienen los datos?, todo en su conjunto y sus características?



Cargar datos

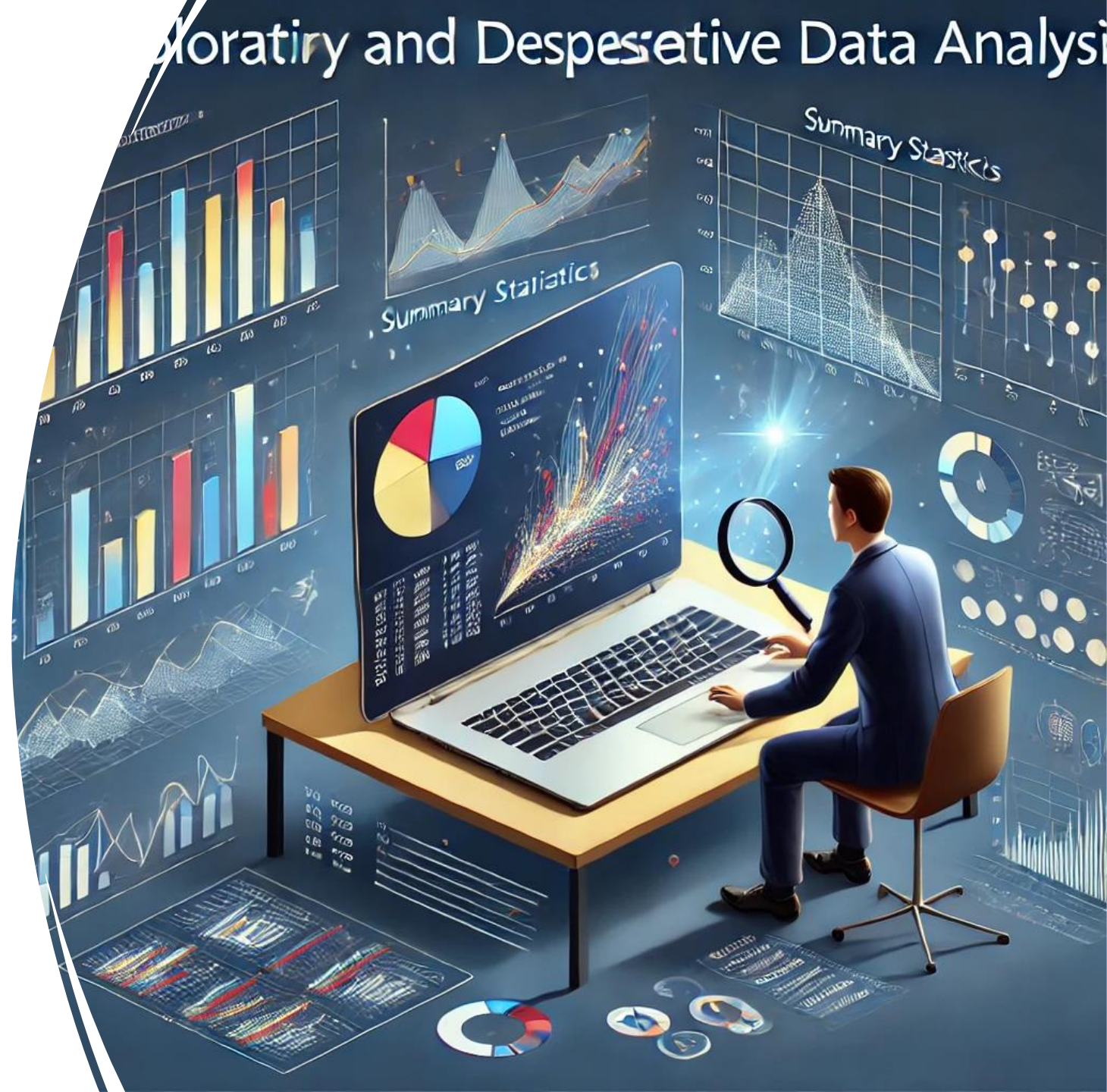
- El proceso de cargar datos es una etapa crucial en el análisis de datos donde se obtienen los datos necesarios desde sus fuentes originales y se transfieren a una herramienta o sistema para su análisis. Es una parte clave de la fase de preparación y preprocesamiento en cualquier flujo de trabajo de análisis de datos.



Cargar datos

- Identificar la fuente de datos: Bases de datos (SQL, NoSQL).
 - Archivos locales (CSV, Excel, JSON, XML, etc.).
 - APIs (interfaces de programación que permiten acceder a datos en línea).
 - Sensores o dispositivos IoT. Web scraping (recolección de datos desde páginas web).
- Conexión a la fuente de datos:
 - Establecer conexión con la base de datos o API. Configurar credenciales de acceso si es necesario.
- Extracción de datos: Descargar o consultar los datos en bruto desde la fuente.

Análisis inicial exploratorio descriptivo



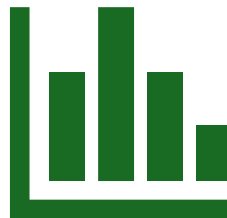
Análisis inicial exploratorio descriptivo



Descripción de datos

Textual

Visualización



Identificar calidad de los datos

¿hay datos malos, sucios, inconsistentes, no homogéneos, faltantes,

errores, deficientes, atípicos?

Hay que escalar datos?, cual fórmula para escalar y porqué?

Hay que estandarizar datos?, convertir a valores z los datos numéricos continuos?

Análisis descriptivo

- El **análisis descriptivo** es un tipo de análisis de datos que se enfoca en resumir y describir las características principales de un conjunto de datos. No intenta hacer predicciones ni establecer relaciones causales, sino proporcionar una visión clara y estructurada de los datos disponibles.

Análisis descriptivo

Medidas de tendencia central:Media:

- El promedio de los datos.
- Mediana: El valor central en un conjunto ordenado de datos.
- Moda: El valor que más se repite.

Medidas de dispersión:

- Rango: Diferencia entre el valor máximo y el mínimo.
- Varianza: Indica cómo se distribuyen los datos alrededor de la media.
- Desviación estándar: Mide cuánto se alejan los datos de la media.

Medidas de forma de distribución:

- Sesgo (skewness): Indica si los datos están distribuidos de manera simétrica o asimétrica.
- Curtosis: Mide la "altura" de la distribución (forma de las colas).

Preparar datos



Preparar datos

Transformar

Factorizar

Eliminar o actualizar

Limpiar

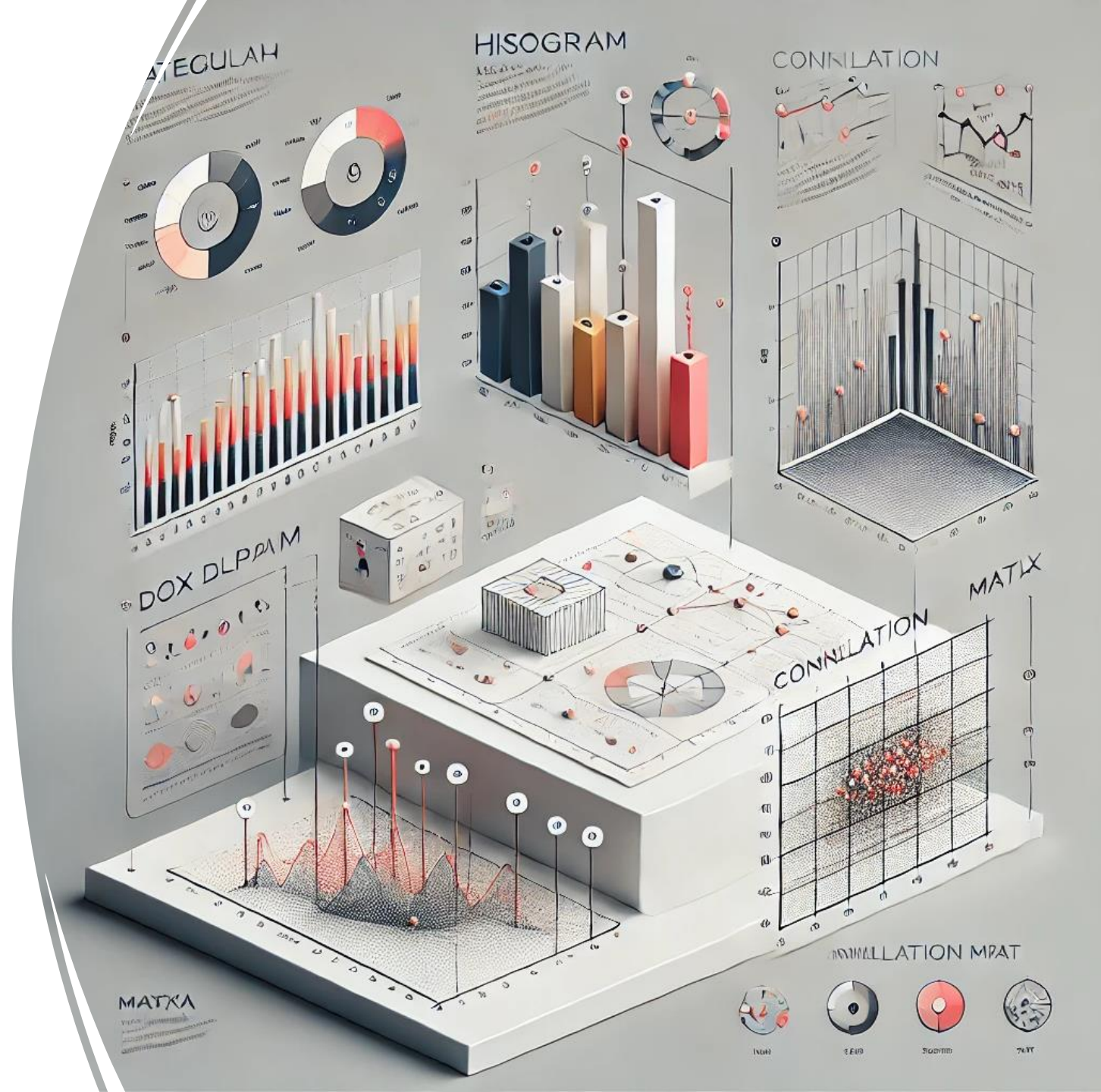
Renombrar

Escalar

Estandarizar

Otros

4. Visualización de datos



4.

Visualización de datos

- La **visualización de datos** es el proceso de representar información y datos de forma gráfica o visual. Permite a los analistas y usuarios explorar y comprender rápidamente patrones, tendencias y relaciones dentro de los datos.
- En el análisis de datos, la visualización:
- Ayuda a **identificar patrones, anomalías y correlaciones**.
- Facilita la **comunicación de hallazgos** a audiencias no técnicas.
- Permite tomar decisiones informadas de manera eficiente.
- Utiliza herramientas como gráficos, tablas y diagramas para **resumir y presentar datos complejos** de manera clara y accesible.

Visualización de datos

1. Comparación de datos

Para comparar valores entre categorías o grupos:

Gráficos de barras: Comparar valores categóricos como ventas por producto.

Gráficos apilados de barras: Mostrar la composición dentro de una categoría.

2. Distribución de datos

Para mostrar cómo los valores se distribuyen:

Histogramas: Representar la frecuencia de rangos de datos.

Diagramas de densidad: Indicar concentraciones de valores.

Boxplot (diagrama de caja): Visualizar la dispersión y detectar valores atípicos.

3. Relaciones entre variables

Para analizar cómo interactúan dos o más variables:

Gráficos de dispersión (scatter plot): Mostrar correlaciones o tendencias.

Diagramas de burbujas: Comparar tres dimensiones de datos.

Visualización de datos

4. Composición de datos

- Para analizar proporciones o partes de un todo:
- **Gráficos de pastel:** Mostrar porcentajes entre categorías.
- **Gráficos de área apilados:** Evaluar cambios acumulativos en el tiempo.

5. Análisis temporal

- Para observar tendencias a lo largo del tiempo:
- **Gráficos de líneas:** Visualizar cambios continuos (ventas mensuales, por ejemplo).
- **Gráficos de velas (candlestick):** Usado en análisis financiero para ver precios de apertura, cierre, máximo y mínimo.

6. Correlación o matriz de relaciones

- Para examinar múltiples relaciones al mismo tiempo:
- **Matrices de correlación:** Evaluar el nivel de relación entre varias variables.
- **Gráficos de calor (heatmaps):** Representar intensidades en una matriz.

Análisis predictivo



Análisis predictivo

- El **análisis predictivo** es una rama de la analítica de datos que utiliza técnicas estadísticas y algoritmos de aprendizaje automático para identificar patrones y realizar predicciones sobre eventos futuros. Se basa en datos históricos y actuales para generar modelos que anticipen resultados y tendencias.



Visualización de datos para análisis predictivo

- Gráfico de regresión:
 - Para visualizar relaciones entre variables y observar cómo un cambio afecta a otro.
- Gráficos de series temporales:
 - Para datos que evolucionan con el tiempo, como ventas mensuales.
- Curvas ROC y AUC:
 - Para evaluar la precisión de modelos clasificatorios.
- Gráficos de dispersión con tendencias:
 - Para visualizar patrones predictivos.
- Gráficos de importancia de variables:
 - Para identificar qué variables influyen más en el modelo.

Interpretación

- La **interpretación de datos** es el proceso de asignar significado a los resultados obtenidos del análisis de datos, relacionándolos con el contexto del problema o la pregunta de investigación. Es una etapa clave en el análisis de datos, ya que transforma números y estadísticas en información útil para la toma de decisiones.



Interpretación

- Dar significado a los resultados: Entender qué implican los valores, patrones y tendencias encontrados.
- Conectar los datos con el contexto: Relacionar los hallazgos con el problema o el fenómeno que se está estudiando.
- Generar conclusiones y recomendaciones: Basadas en los datos, ayudar a formular estrategias o decisiones.
- Comunicar hallazgos: Presentar los resultados de manera comprensible para audiencias técnicas y no técnicas.

Comunicación

- La comunicación de datos es el proceso de transmitir de manera efectiva los hallazgos, conclusiones y recomendaciones obtenidas del análisis de datos. Su objetivo principal es transformar los resultados técnicos en información clara y comprensible para diferentes audiencias, desde expertos en el área hasta públicos generales o tomadores de decisiones.



Comunicación

- Facilitar la comprensión: Hacer que los resultados sean accesibles y relevantes, incluso para personas no técnicas.
- Soportar la toma de decisiones: Proveer la información necesaria para definir estrategias, resolver problemas o aprovechar oportunidades.
- Resumir información clave: Destacar los datos más importantes y los hallazgos significativos.
- Persuadir y motivar acciones: Utilizar datos y evidencias para respaldar propuestas y recomendaciones.



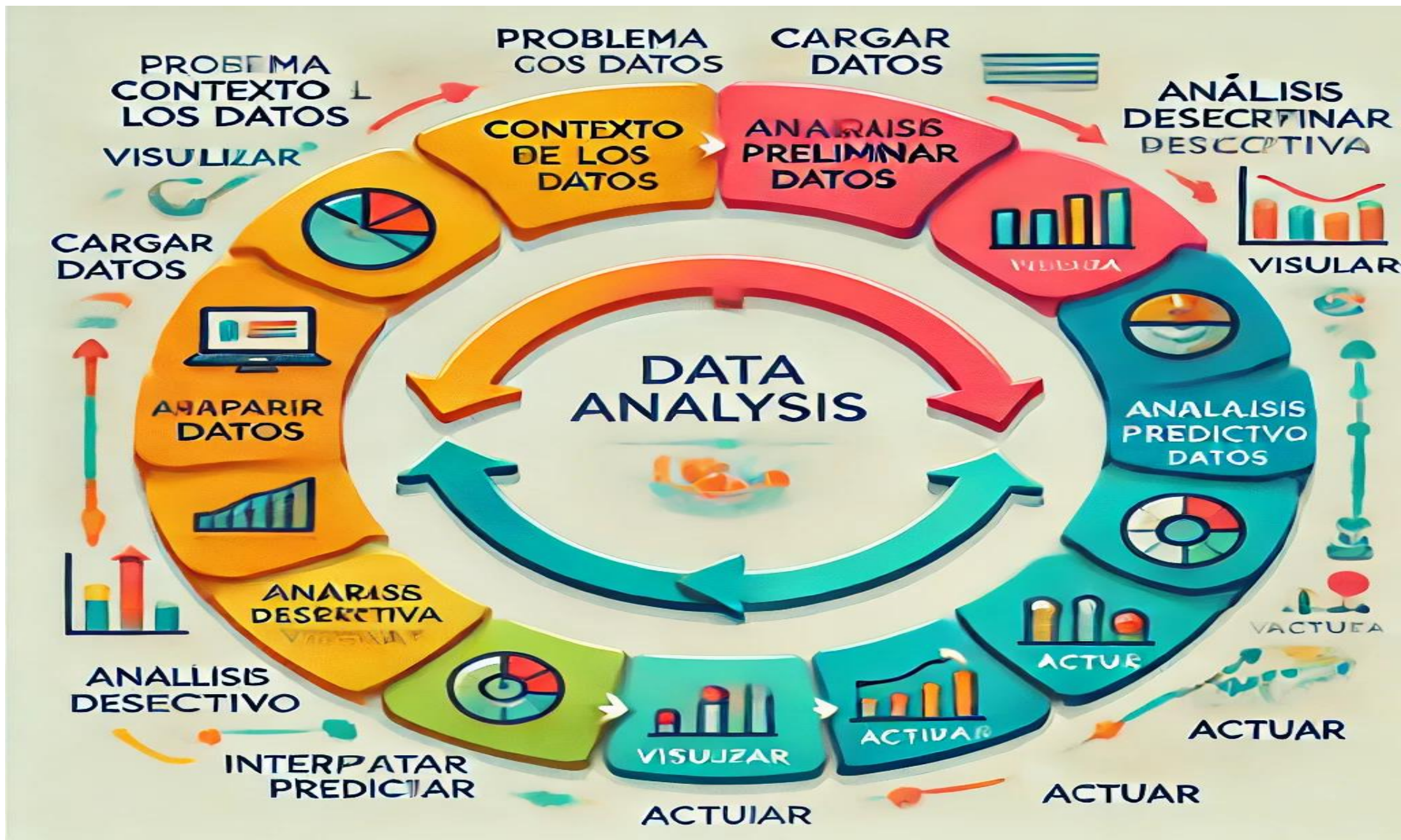
Actuación

- **convertir los hallazgos y conclusiones obtenidos del análisis en decisiones concretas o acciones prácticas.** Es el paso final en el proceso de análisis de datos, donde la información recopilada y comprendida se utiliza para generar un impacto tangible en la organización, proyecto o problema en cuestión.



Actuación

- Interpretar los resultados en contexto: Entender cómo los hallazgos del análisis se relacionan con los objetivos o problemas planteados.
- Definir objetivos claros: Basarse en los resultados para establecer metas alcanzables y específicas.
- Elaborar planes que detallen qué cambios o iniciativas se implementarán.
- Determinar qué acciones tendrán el mayor beneficio y son factibles de implementar. Ejecutar las estrategias diseñadas.
- Automatizar procesos después de identificar ineficiencias manuales. Monitorear y ajustar:
- Evaluar el impacto de las acciones basadas en métricas clave y ajustar según sea necesario.





Gracias