Análisis de Datos y Ciencia de Datos

Ing. Remigio Hurtado Ortiz, PhD.

2025

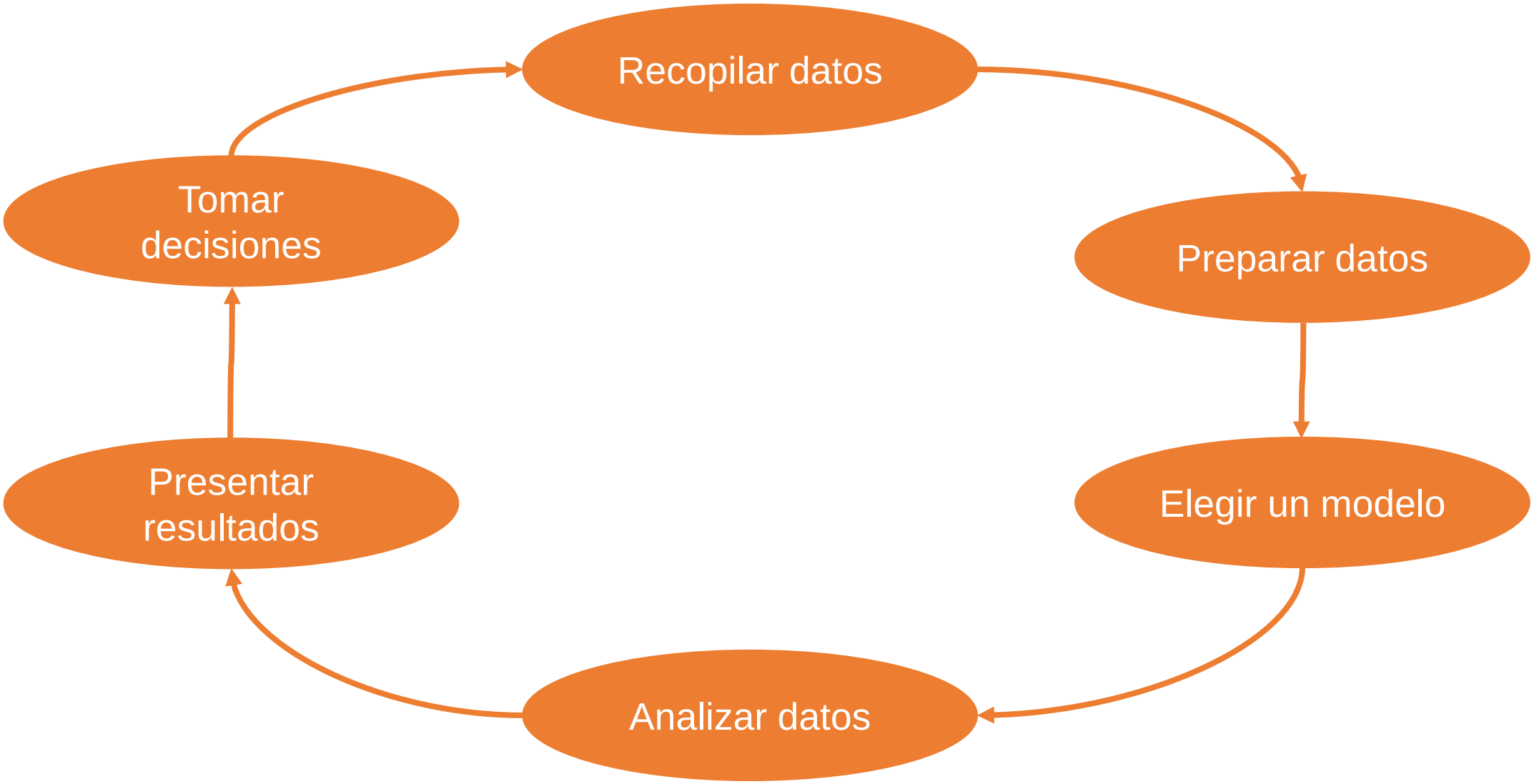
Análisis de datos y Ciencia de Datos

- La ciencia de datos es un campo interdisciplinario que se enfoca en el estudio y análisis de datos para obtener información valiosa y conocimiento útil. Combina elementos de estadísticas, matemáticas, programación y conocimiento de dominio para explorar conjuntos de datos, identificar patrones, realizar predicciones y tomar decisiones basadas en datos.
- El proceso típico de ciencia de datos incluye la recopilación, limpieza y organización de datos, seguido de su análisis exploratorio para comprender las características y relaciones dentro de los datos. Luego, se aplican modelos y algoritmos de aprendizaje automático para desarrollar modelos predictivos o descriptivos que pueden utilizarse para tomar decisiones informadas o generar recomendaciones.

Análisis de datos y Ciencia de Datos

1. Metodología
2. Tipos de variables
3. Preprocesamiento de datos: Transformación de variables, transformaciones numéricas, datos atípicos
4. Limpieza de datos
5. Tipos de análisis

Metodología de Análisis de Datos CRISP-DM



Ciclo de vida de proyectos de ciencia de datos

Entendiendo el negocio	Entendiendo los datos	Preparación de datos	Modelado	Evaluación	Despliegue
Determinar los objetivos del negocio Revisar situación: requerimientos, supuestos, restricciones, riesgos y contingencias, costos y beneficios Determinar metas de análisis de datos Producir plan del proyecto	Recolectar datos iniciales: técnicas de extracción de datos de distintas fuentes (bases de datos, imágenes, videos, web, redes sociales, IoT, etc.) Describir datos: análisis de distribución de datos, funciones de distribución, medidas de estadística descriptiva Explorar datos: resúmenes descriptivos y gráficos para descubrir relaciones entre los datos o problemas Verificar calidad de datos	Seleccionar datos Limpiar datos Construir datos: transformación de datos, nuevas variables (componentes o factores) Integrar datos Formatear datos Formar dataset y su descripción	Seleccionar técnicas de modelado (modelos estadísticos, machine learning) Generar diseño de pruebas Construir modelo: parámetros, modelos, descripciones de cada modelo Revisar modelo y sus parámetros	Evaluar resultados: revisar criterios de éxito – medidas de calidad Revisar el proceso Determinar siguientes pasos: posibles decisiones	Planificar despliegue Planificar monitoreo y mantenimiento Producir reporte final Revisar proyecto: documentación de experiencia

Tipos de análisis

Descriptivo

Predictivo

Prescriptivo

Tipo	Tareas	Preguntas
Descriptiva	Generación de informes estándar	¿Qué ocurrió?
	Informes ad hoc	¿Qué cantidad, con qué frecuencia, dónde?
	Consultas de datos	Exactamente, ¿cuál es el problema?
Predictivo	Simulación	¿Qué podría suceder?
	Pronósticos	¿Qué sucedería si continúan estas tendencias?
	Modelado predictivo	¿Qué sucederá a continuación?
Prescriptiva	Optimización	¿Cómo podemos obtener el mejor resultado?
	Optimizaciones bajo incertidumbre	¿Cómo podemos lograr el mejor resultado, dada la variabilidad?

Análisis Exploratorio

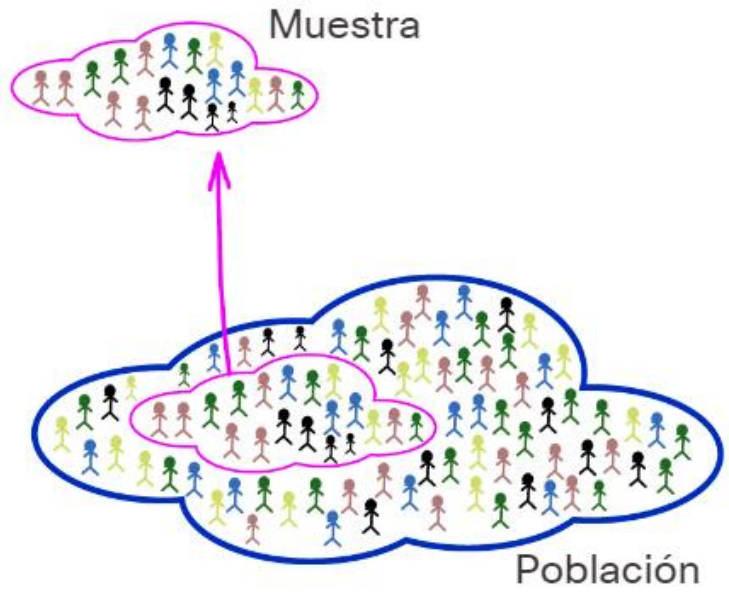
- **Carga de Datos:** Importar los datos desde diferentes fuentes, como archivos CSV, hojas de cálculo, páginas web, bases de datos, etc.
- **Exploración de Datos Iniciales:** Comprender la estructura de los datos, tamaño, cantidad de variables y observaciones, tipos de variables y la calidad de los datos (verificar si hay valores faltantes, duplicados o atípicos).
- **Resumen Estadístico:** Calcular estadísticas descriptivas básicas, como la media, mediana, desviación estándar, percentiles, para comprender la distribución y la variabilidad de los datos.
- **Visualización de Datos:** Crear gráficos y visualizaciones para representar los datos de manera efectiva. Esto puede incluir histogramas, diagramas de dispersión, gráficos de barras y diagramas de caja, entre otros.
- **Análisis de Variables Categóricas:** Si existen variables categóricas, realizar análisis de frecuencia para comprender la distribución de categorías.
- **Análisis de Correlación:** Evaluar las relaciones entre las variables mediante el cálculo de correlaciones.
- **Manejo de Datos Faltantes:** Decidir cómo manejar los valores faltantes, que puede incluir imputación (sustitución), eliminación o análisis de su impacto en los resultados.
- **Análisis de Outliers:** Identificar y comprender los valores atípicos que pueden afectar el análisis y la interpretación de los resultados.
- **Segmentación de Datos:** Si es relevante, dividir el conjunto de datos en segmentos o grupos con características similares para un análisis más detallado.
- **Generación de Hipótesis:** Basado en la exploración de datos, formular hipótesis iniciales sobre relaciones o patrones que se puedan investigar más adelante.
- **Presentación de Resultados:** Comunicar hallazgos iniciales a través de informes o presentaciones que resuman las observaciones y los posibles pasos siguientes en el análisis.

Tipos de variables

- Conjunto de datos (Dataset)
- Variables y observaciones

Ejemplo:

Dataset Statlog
Cantidad de observaciones (clientes): 1000
Cantidad de variables: 21

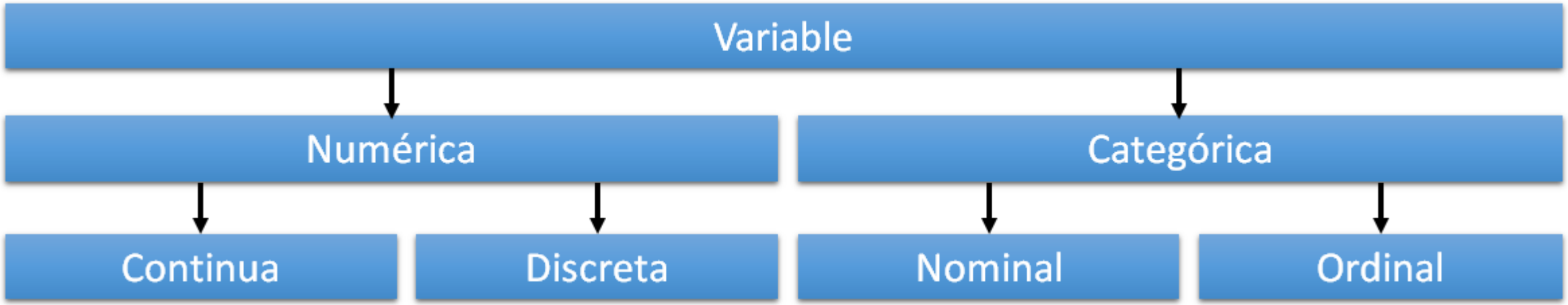


ESTADOCUENTACORRIENTE	PLAZOMESESCREDITO	HISTORIALCREDITO	PROPOSITOCREDITO	MONTOCREDITO	SALDOCUENTAAHORROS
A11	6	A34	A43	1169	A65
A12	48	A32	A43	5951	A61
A14	12	A34	A46	2096	A61
A11	42	A32	A42	7882	A61
A11	24	A33	A40	4870	A61

Tipos de variables

- Conjunto de datos (Dataset)
- Observaciones
- Variables y valores
- Tipos de variables

Algunos métodos estadísticos y visualizaciones de datos están diseñados para trabajar mejor con ciertos tipos de datos que con otros. Cómo se muestran mejor los resultados del análisis dependerá del tipo de variables utilizadas en los datos.



Tipos de variables

- **1. Categóricas:** Las variables categóricas indican el tipo o categoría de una observación. Por lo tanto, las variables categóricas son variables cualitativas y en la mayoría de conjuntos de datos están representadas por un valor no numérico.
- **1.1 Nominales:** la categoría indica la identidad del objetivo. No hay un grado o diferencia entre las categorías, lo que se enfatiza es el nombre. Algunos ejemplos de variables categóricas nominales son: sexo, país, color de ojos, etc. La transformación habitual para estas variables es una transformación "Coding" o conversión a un código binario. Este proceso consiste en pasar las categorías en formato one_hot, que consiste en poner tantos ceros como categorías, y para representar un valor se coloca un uno en la posición del valor, ejemplo:

  3 colores de ojos: azul, café, verde.

  azul = [1,0,0] café = [0,1,0] verde = [0,0,1]

  En Python para ejecutar esta transformación se utiliza: OneHotEncoder

Tipos de variables

- **1.2 Ordinales:** la categoría indica un orden. Algunos ejemplos de variables categóricas ordinales son: nivel de educación (primaria, secundaria, superior), las escalas de las encuestas de satisfacción (insatisfecho, neutro, satisfecho), calificación cualitativa (insuficiente, buena, muy buena, sobresaliente, muy sobresaliente), etc. La transformación para estas variables es más simple, las categorías en orden ascendente se transforman a un valor numérico empezando desde cero, ejemplo:

  Satisfacción sobre un producto: no me gusta, neutro, me gusta

  no me gusta: 0 neutro: 1 me gusta: 2

- En Python para ejecutar esta transformación se utiliza: OrdinalEncoder. Con esta clase y sus funciones se transforma automáticamente las categorías a un valor numérico. En el dataset utilizado en este artículo las categorías siguen una nominación que permite una transformación directa al valor numérico. Ejemplo: A30, A31, A32, A33 y A34, estas categorías se transforman a los valores 0, 1, 2, 3, 4, de manera respectiva. También es posible especificar otros valores numéricos, para ello, se debe indicar la lista de categorías en orden, en conjunto con los valores correspondientes. Ejemplo: X = [['no me gusta', 1], ['neutro', 3], ['me gusta', 5]].

Tipos de variables

- **2. Numéricas:** Las variables numéricas tienen valores que describen una cantidad medible. Las variables numéricas son variables cuantitativas.
- **2.1 Continuas:** estas son variables que son cuantitativas y pueden medirse a lo largo de una secuencia o un rango de valores. Existen dos tipos de variables continuas: las variables de intervalo y las variables de relaciones. Las variables de intervalo pueden tener cualquier valor dentro de un rango de valores. Algunos ejemplos son temperatura o tiempo. Las variables de relaciones son las variables de intervalo especiales donde un valor de cero (0) significa que no hay ninguna variable. Entre los ejemplos se incluyen ingresos o el volumen de ventas.
- **2.2 Discretas:** estos tipos de variables continuas son cuantitativos, pero tienen un valor específico de un conjunto de valores finito. Los ejemplos incluyen el número de sensores habilitados en una red, el número de automóviles en un estacionamiento, la cantidad de visitas a una página web, la cantidad de llamadas recibidas, etc.

Preparación/Preprocesamiento de datos

Coding de datos categóricos (código binario): Clases A (azul), V (verde), C (cafe) y N (negro). P=4 variables, que son P-1=3 variables binarias (x1,x2,x3). Este procedimiento siempre puede aplicarse pero puede lógicamente dar lugar a muchas variables. Conviene entonces ver si podemos agrupar las clases o categorías para evitar tener variables que casi siempre toman el mismo valor (cero si la categoría es poco frecuente o uno si lo es mucho).

CO	x_1	x_2	x_3
A	1	0	0
V	0	1	0
C	0	0	1
N	0	0	0

Naturalmente la variable CO podría también haberse codificado dando valores numéricos arbitrarios a las categorías, por ejemplo, A=1, V=2, C=3, N=4, pero esta codificación tiene el inconveniente de sugerir una graduación de valores que puede no existir. Sin embargo, cuando los atributos pueden interpretarse en función de los valores de una variable continúa tiene más sentido codificarla con números que indiquen el orden de las categorías. Por ejemplo, si tenemos empresas pequeñas, medianas y grandes, en función del número de trabajadores, tienen sentido codificarlas con los números 1, 2, y 3, aunque conviene siempre recordar que estos números sólo tienen un sentido de orden.

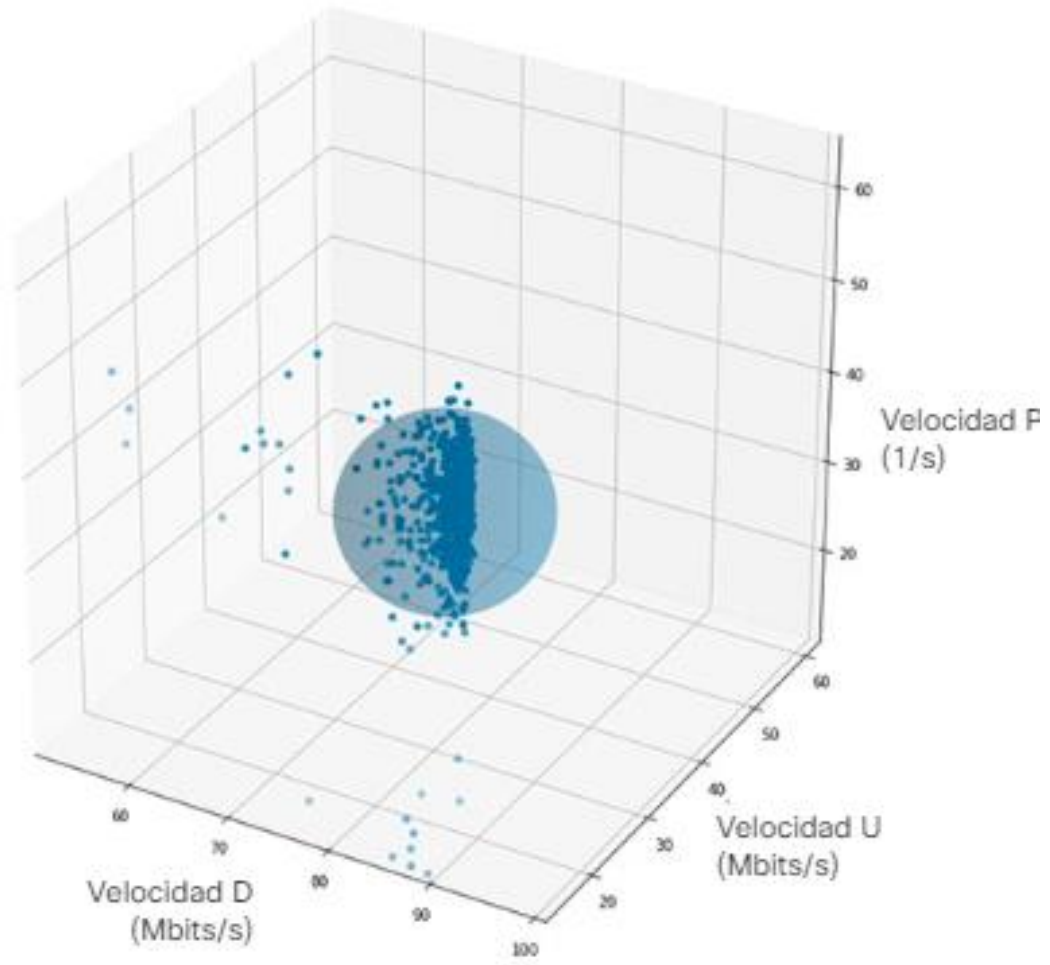
Normalización – transformación al rango entre 0 y 1
Scaling – transformación a umbrales mínimo y máximo (pueden ser diferentes de 0 y 1)
Estandarización - transformación de datos (mean=0, varianza=1)

Preparación/Preprocesamiento de datos
Limpieza de Datos - Detección de anomalías - Datos atípicos

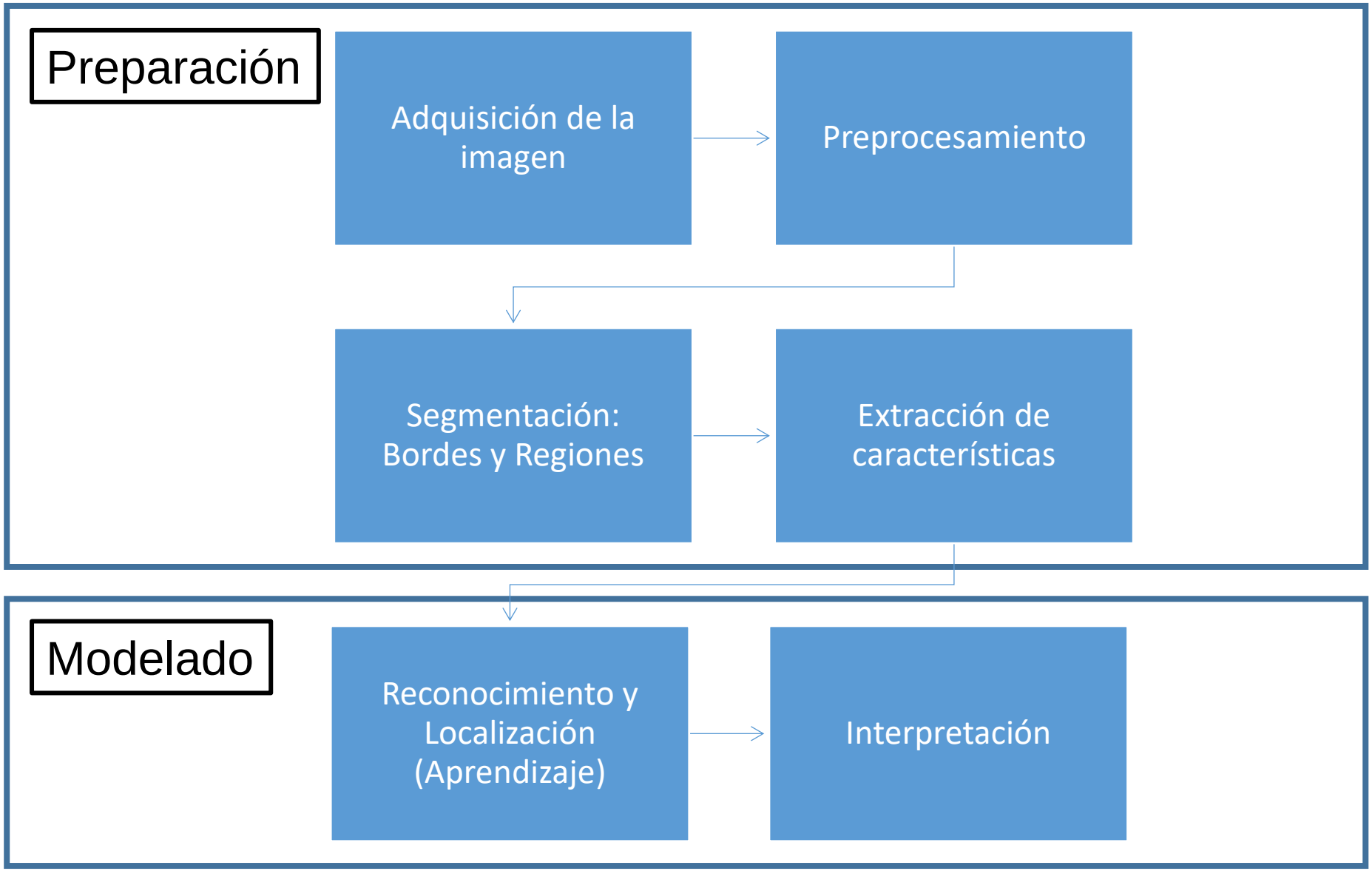
Las anomalías pueden representar datos que son anómalos, o valores que son anómalos. Los datos pueden dañarse o distorsionarse mediante muchos factores durante la medición, la transmisión o el almacenamiento. Estos valores se consideran valores atípicos. Se desvían tanto de los valores esperados que podrían distorsionar los resultados del análisis. Estas consideraciones se suelen eliminar del conjunto de datos después de un estudio detallado.

Para detectar anomalías, deberá identificar el límite de la decisión que define si un punto de datos es normal o es una anomalía. Para ello, primero se normalizan los datos de la distancia al establecer el trayecto más lejano a 1. Luego se determina un umbral entre 0 y 1 que defina el umbral para el límite de la decisión.

Normalización y Detección de anomalías (Visualización en 3D)



Proceso de Visión Artificial



Referencias

- Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).
- Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.
- Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.
- P. Joshi. (2017). Artificial intelligence with python. Packt Publishing Ltd.
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.