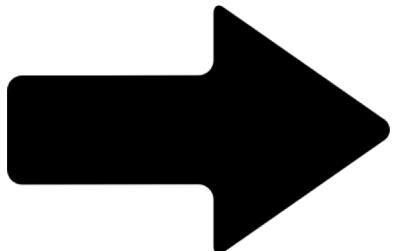


MODELO DE MINERÍA DE DATOS

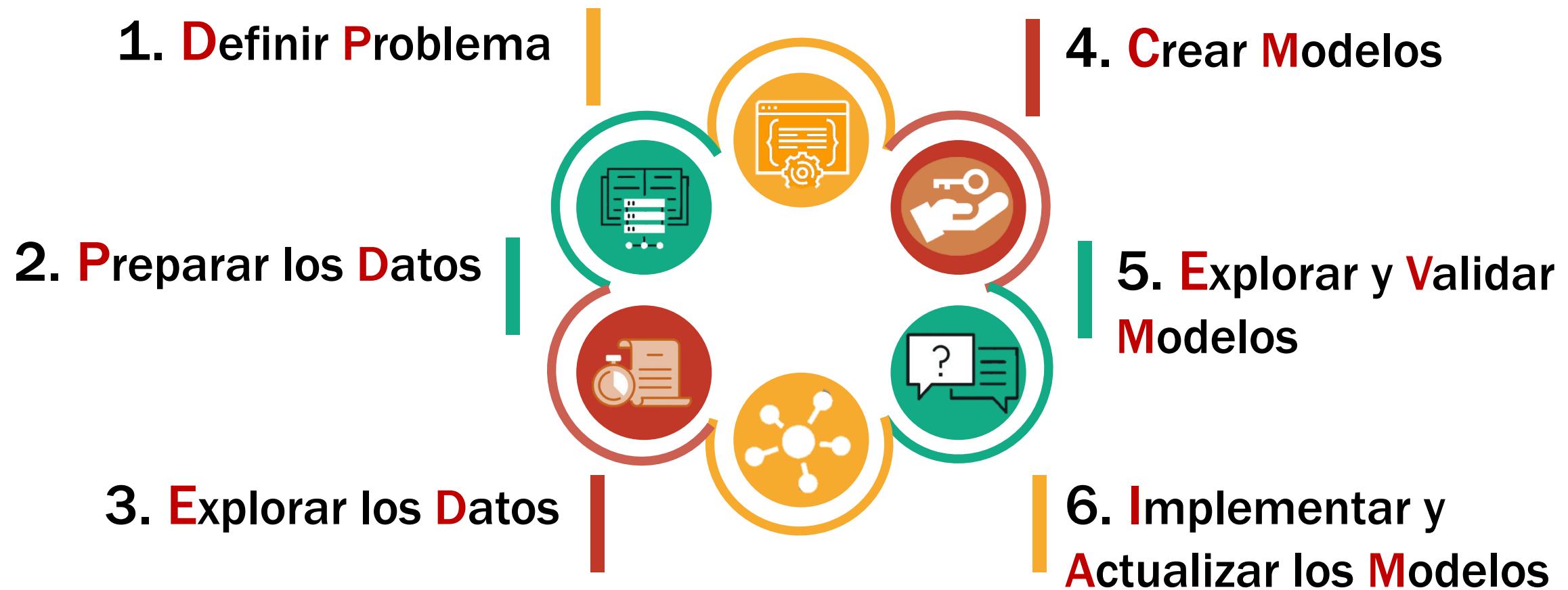
3

MODELO DE MINERÍA DE DATOS

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo.



Este proceso se puede definir mediante los seis pasos básicos siguientes:



- **Definir Problema**

Consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo.

Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos.



Estas tareas se traducen en preguntas como las siguientes:

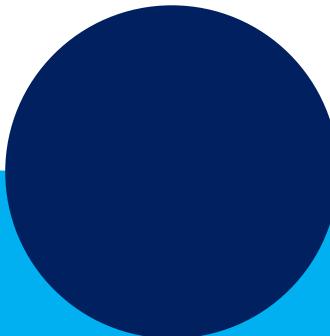


- 1 ¿Qué está buscando?
- 2 ¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?
- 3 ¿Qué tipos de relaciones intenta buscar?
- 4 ¿Qué resultado o atributo desea predecir?
- 5 ¿Desea realizar predicciones a partir del modelo o solo buscar asociaciones y patrones interesantes?



- 6 ¿Qué tipo de datos tiene y qué tipo de información hay en cada columna?
- 7 En caso de que haya varias tablas, ¿cómo se relacionan?
- 8 ¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?
- 9 ¿Cómo se distribuyen los datos? ¿Los datos son estacionales?
- 10 ¿Los datos representan con precisión los procesos de la empresa?

EJERCICIO PRÁCTICO



- **Ejemplo:**

En la página de Kaggle se seleccionó una base de datos:

Breast Cancer Wisconsin (Diagnostic) Data Set (<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>)

Imaginándonos nosotros representar al Hospital de Wisconsin, se plantea un objetivo, problema y la posible solución como se muestra en el ejemplo de abajo.



Breast Cancer Wisconsin (Diagnostic) Data Set

Objetivo: Reducir el tiempo que se tarda un experto en identificar malignidad dentro de una anomalía de cáncer de mama.

Problema planteado: Se requieren muchas revisiones para asegurar la malignidad de una anomalía por parte de varios expertos, lo cual ocupa tiempo que podría utilizarse para empezar a tratar al paciente.

Solución: Desarrollar una herramienta de aprendizaje de máquina que permita identificar malignidad de anomalías dependiendo de sus características.

Enlaces a bases de datos:

- a) Google Play Store (<https://www.kaggle.com/lava18/google-play-store-apps>)
- b) Coronavirus (<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>)
- c) Criticas de vinos (<https://www.kaggle.com/zynicide/wine-reviews>)
- d) Clasificación de plantas (<https://www.kaggle.com/uciml/iris>)
- e) Shows de Netflix (<https://www.kaggle.com/shivamb/netflix-shows>)