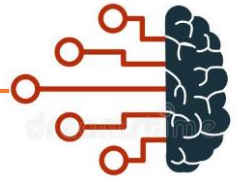
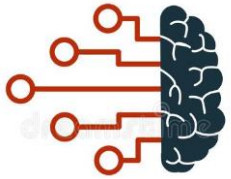


Introducción
Machine Learning

Lic. Patricia Rodríguez Bilbao

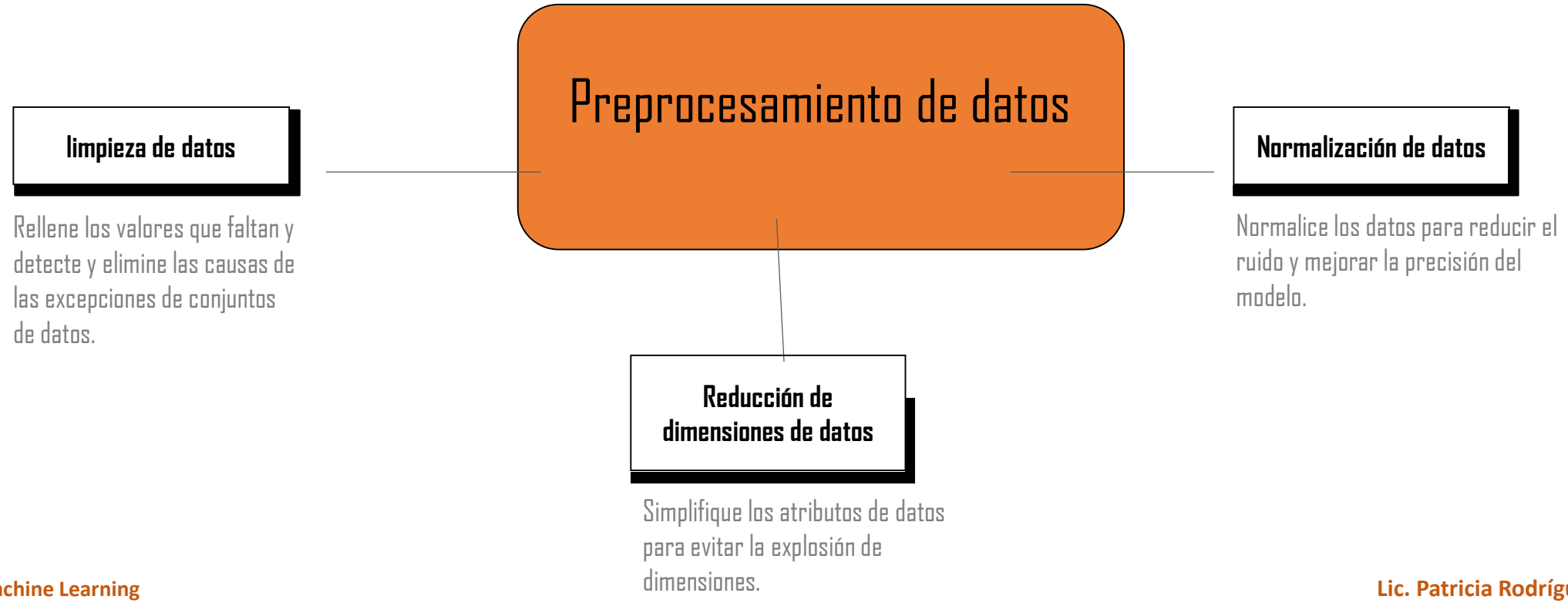
Contenido



1. Introducción
2. Conceptos
3. Definiciones
4. Clasificación
5. Taller

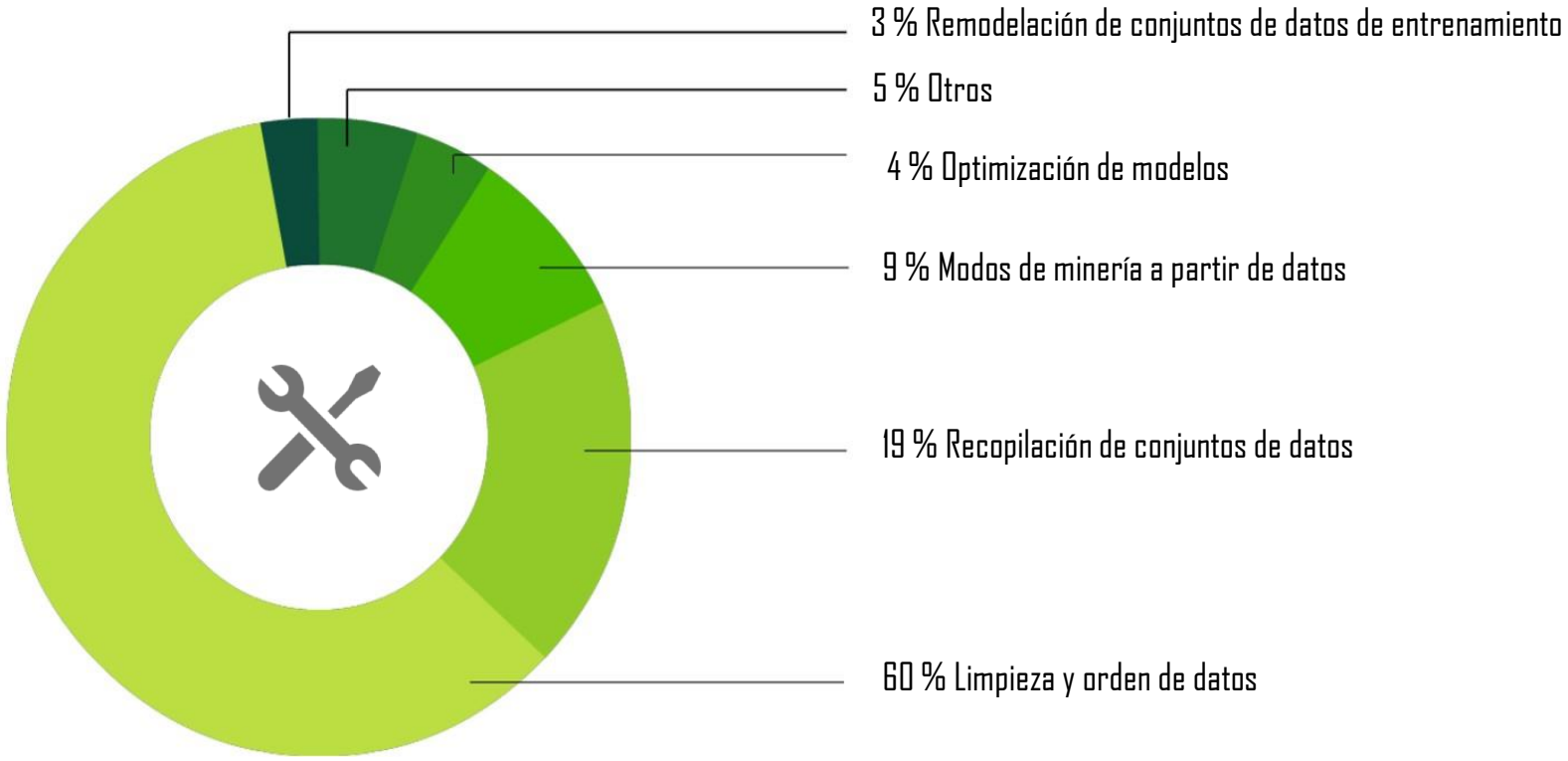
Importancia del procesamiento de datos

Los datos son cruciales para los modelos. Es el techo de las capacidades del modelo. Sin buenos datos, no hay un buen modelo.



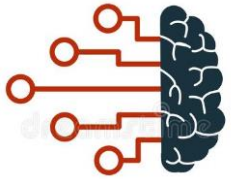
Tareas de limpieza de datos

Estadísticas sobre el trabajo de los científicos en M.L



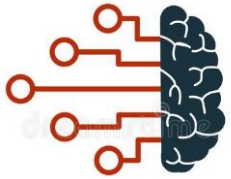
Informe CrowdFlower Data Science 2016

Limpieza de datos



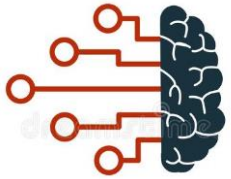
- La mayoría de los modelos de ML procesan características, que suelen ser representaciones numéricas de variables de entrada que se pueden utilizar en el modelo.
- En la mayoría de los casos, los datos recogidos pueden ser utilizados por algoritmos sólo después de ser preprocesados. Las operaciones de preprocesamiento incluyen las siguientes:
  - Filtrado de datos
  - Procesamiento de datos perdidos
  - Procesamiento de posibles excepciones, errores o valores anormales
  - Combinación de datos de múltiples fuentes de datos
  - Consolidación de datos

Datos sucios



- Generalmente, los datos reales pueden tener algunos problemas de calidad.
  - Incompletitud: contiene valores que faltan o los datos que carecen de atributos
  - Ruido: contiene registros incorrectos o excepciones.
  - Inconsistencia: contiene registros inconsistentes.

Datos sucios



#	Id	Name	Birthday	Gender	Is Teacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Valor faltante

Valor inválido

Valor que debería estar en otra columna

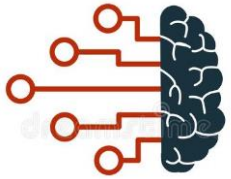
Error de ortografía

Item duplicado-Inválido

Formato incorrecto

Dependencia de atributos

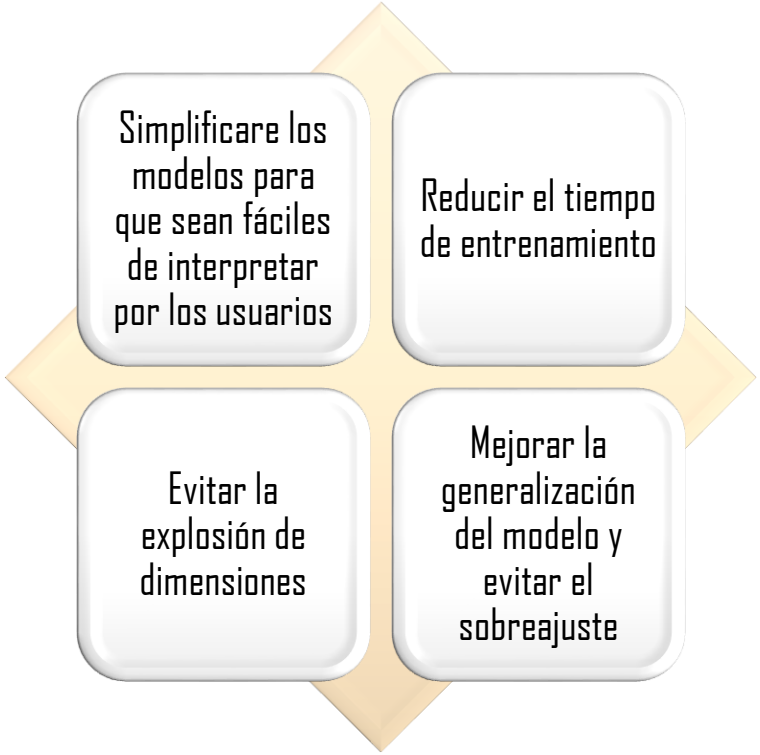
Conversión de datos



- Después de ser preprocesados, los datos deben ser convertidos en un formulario de representación adecuado para el modelo de Machine Learning. Los formularios comunes de conversión de datos incluyen los siguientes:
  - Con respecto a la clasificación, los datos de las categorías se codifican en la representación numérica correspondiente.
  - Los datos de valor se convierten en datos de categoría para reducir el valor de las variables (para la segmentación de edad).
  - Otros datos
    - En texto, la palabra se convierte en un vector de palabras a través de incrustaciones de palabras (generalmente usando el modelo word2vec, el modelo BERT, etc.)
    - Procesar datos de imagen (espacio de color, escala de grises, cambio geométrico, función Haar y mejora de imagen)
  - Ingeniería de características
    - Normalizar características para garantizar los mismos rangos de valores para las variables de entrada del mismo modelo.
    - Expansión de características : Combine o convierta variables existentes para generar nuevas caracteristicas, como el promedio.

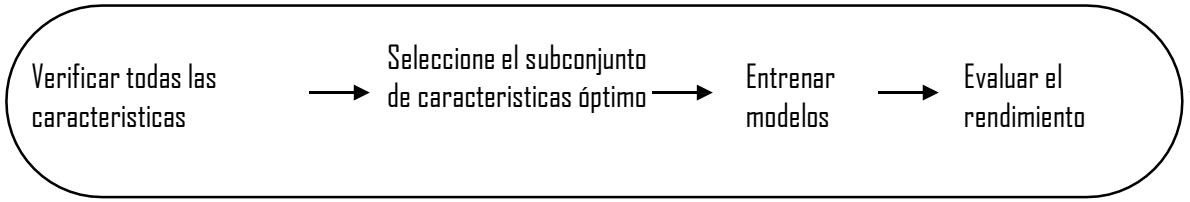
Necesidad de la selección de datos

- Generalmente, un conjunto de datos tiene muchas características, algunas de las cuales pueden ser redundantes o irrelevantes para el valor a predecir.
- La selección de características es necesaria en los siguientes aspectos:



Métodos de selección de características - filtro

- Los métodos de filtrado son independientes del modelo durante la selección de características



Procedimiento de un método de filtro

Al evaluar la correlación entre cada característica y el atributo de destino, estos métodos utilizan una medida estadística para asignar un valor a cada característica. Las características se clasifican por puntuación, lo que ayuda a preservar o eliminar características específicas.

Métodos comunes

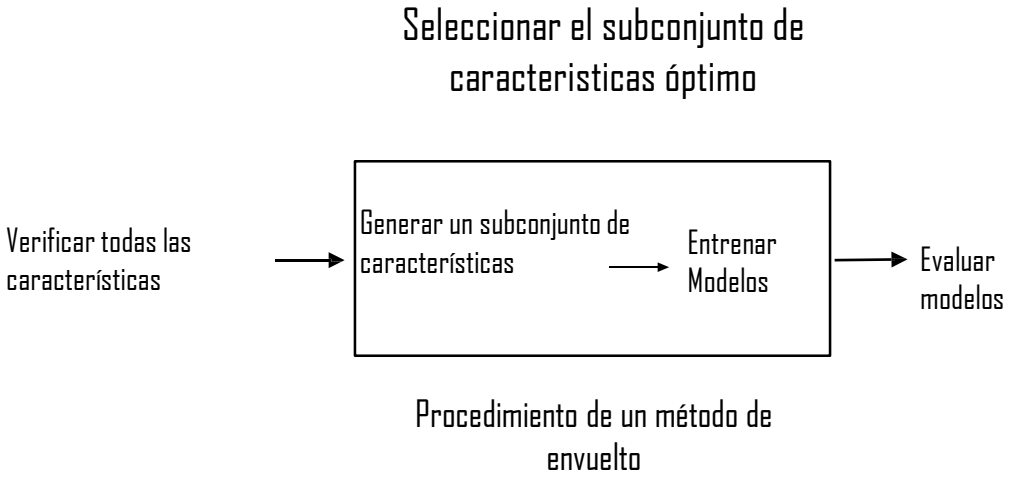
- Coeficiente de correlación de Pearson
- Coeficiente de chi-cuadrado
- Información mutua

Limitaciones

- El método de filtrado tiende a seleccionar variables redundantes ya que no se tiene en cuenta la relación entre las características.

Métodos de selección – Wrapper (envuelto)

- Los métodos de envuelto utilizan un modelo de predicción para puntuar subconjuntos de características



Procedimiento de un método de envuelto

Los métodos de envuelto consideran la selección de características como un problema de búsqueda para el cual se evalúan y comparan diferentes combinaciones. Un modelo predictivo se utiliza para evaluar una combinación de características y asignar una puntuación basada en la precisión del modelo.

Métodos comunes

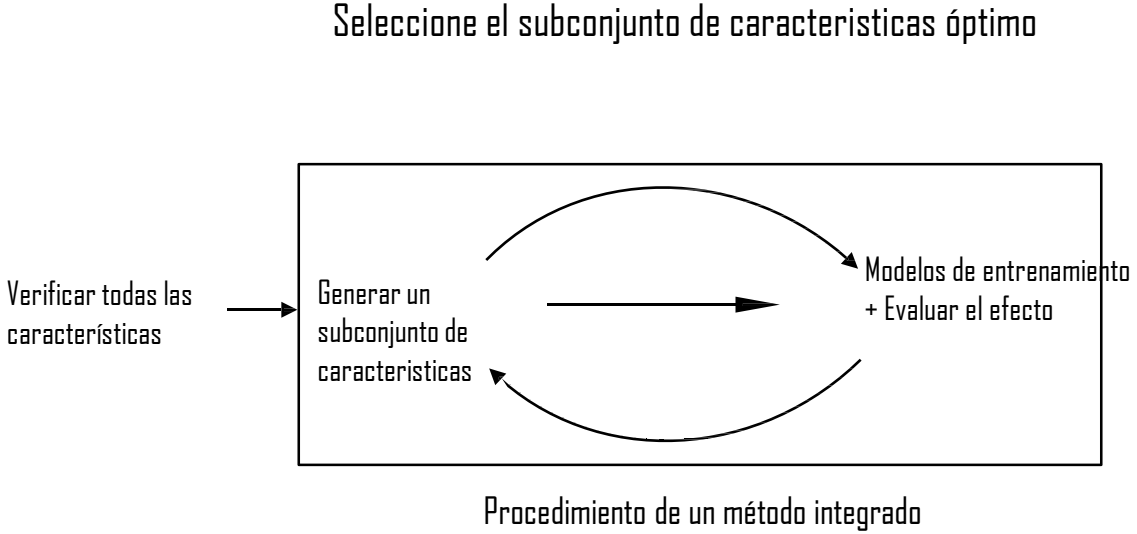
- Eliminación de funciones recursivas (RFE)

Limitaciones

- Los métodos de envuelto entrenan un nuevo modelo para cada subconjunto, lo que da lugar a un **gran número de cálculos.**
- Normalmente se proporciona un conjunto de características con el mejor rendimiento para un tipo específico de modelo.

Métodos de selección – Embedded (integrados)

Los métodos integrados consideran la selección de características como parte de la construcción de modelos



Procedimiento de un método integrado

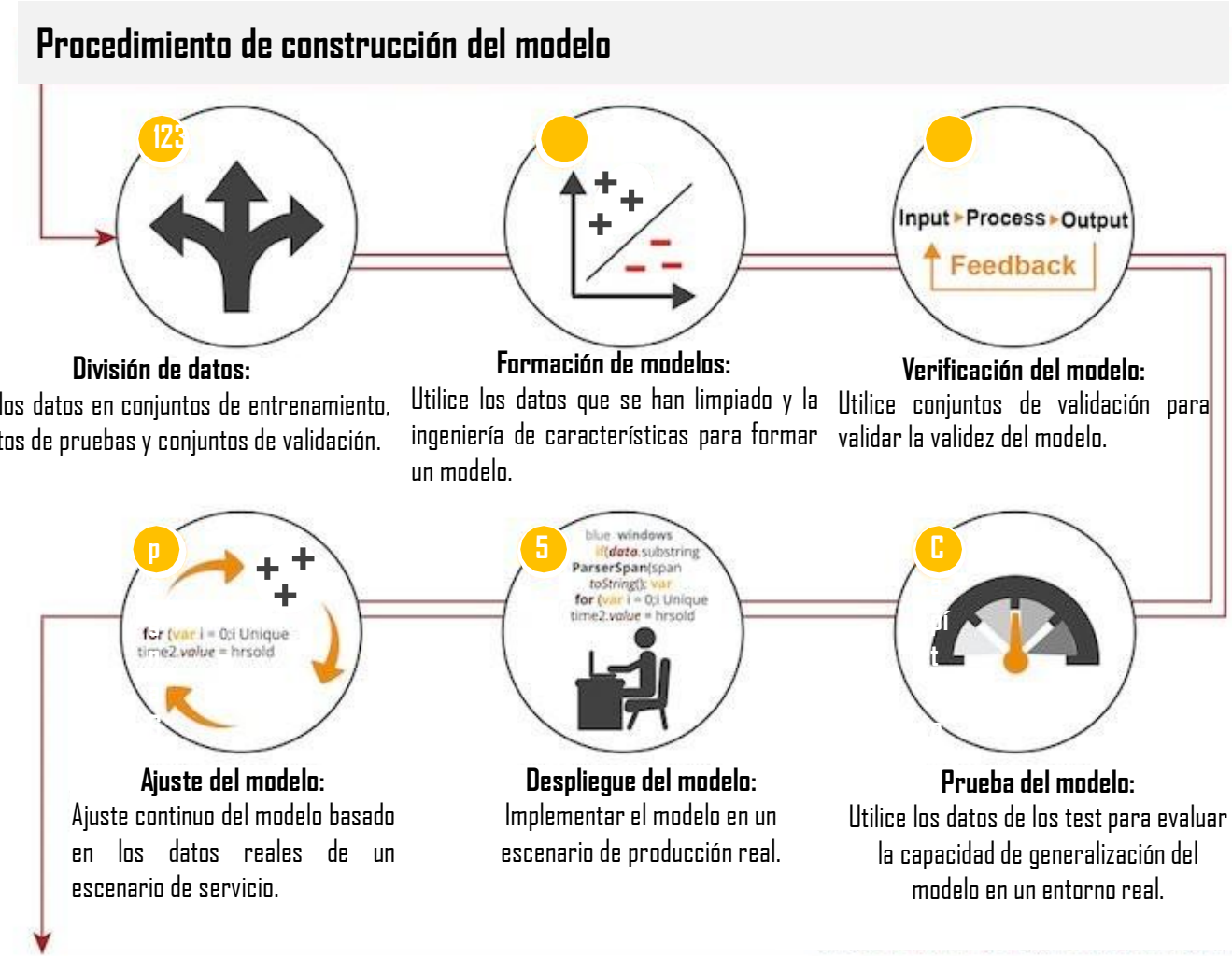
El método de selección de características incorporadas más común es el **método de regularización**.

Los métodos de regularización también se llaman métodos de penalización que introducen restricciones adicionales en la optimización de un algoritmo predictive, que presiona al modelo hacia una menor complejidad y reducen el número de características.

Métodos comunes

- Regresión Lasso
- Regresión Ridge

Procedimiento construcción de un modelo



División de datos:
Divida los datos en conjuntos de entrenamiento, conjuntos de pruebas y conjuntos de validación.

Formación de modelos:
Utilice los datos que se han limpiado y la ingeniería de características para formar un modelo.

Verificación del modelo:
Utilice conjuntos de validación para validar la validez del modelo.

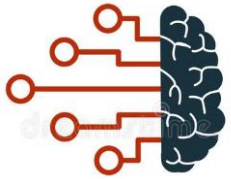
Ajuste del modelo:
Ajuste continuo del modelo basado en los datos reales de un escenario de servicio.

Despliegue del modelo:
Implementar el modelo en un escenario de producción real.

Prueba del modelo:
Utilice los datos de los test para evaluar la capacidad de generalización del modelo en un entorno real.

¿Qué es un buen modelo?

- **Capacidad de generalización**
  ¿Puede predecir con precisión los datos reales del servicio?
- **Interpretabilidad**
  ¿Es fácil interpretar el resultado de la predicción?
- **Velocidad de predicción**
  ¿Cuánto tiempo se tarda en predecir cada pieza de datos?
- **Practicidad**
  ¿La velocidad de predicción sigue siendo aceptable cuando el volumen de servicio aumenta con un gran volumen de datos?



Gracias……