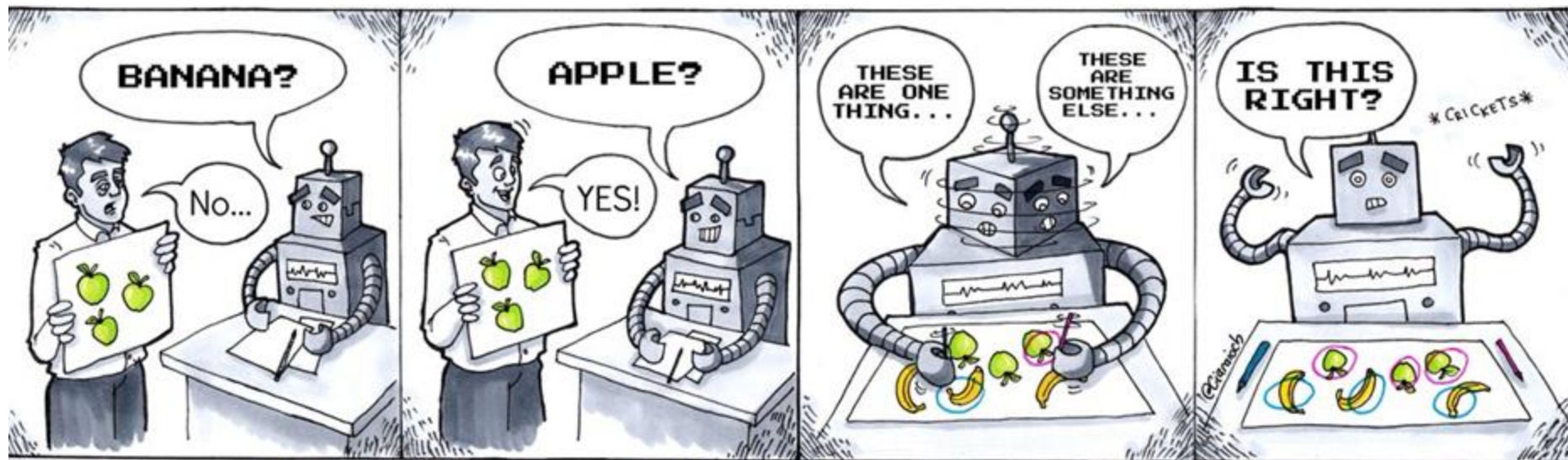
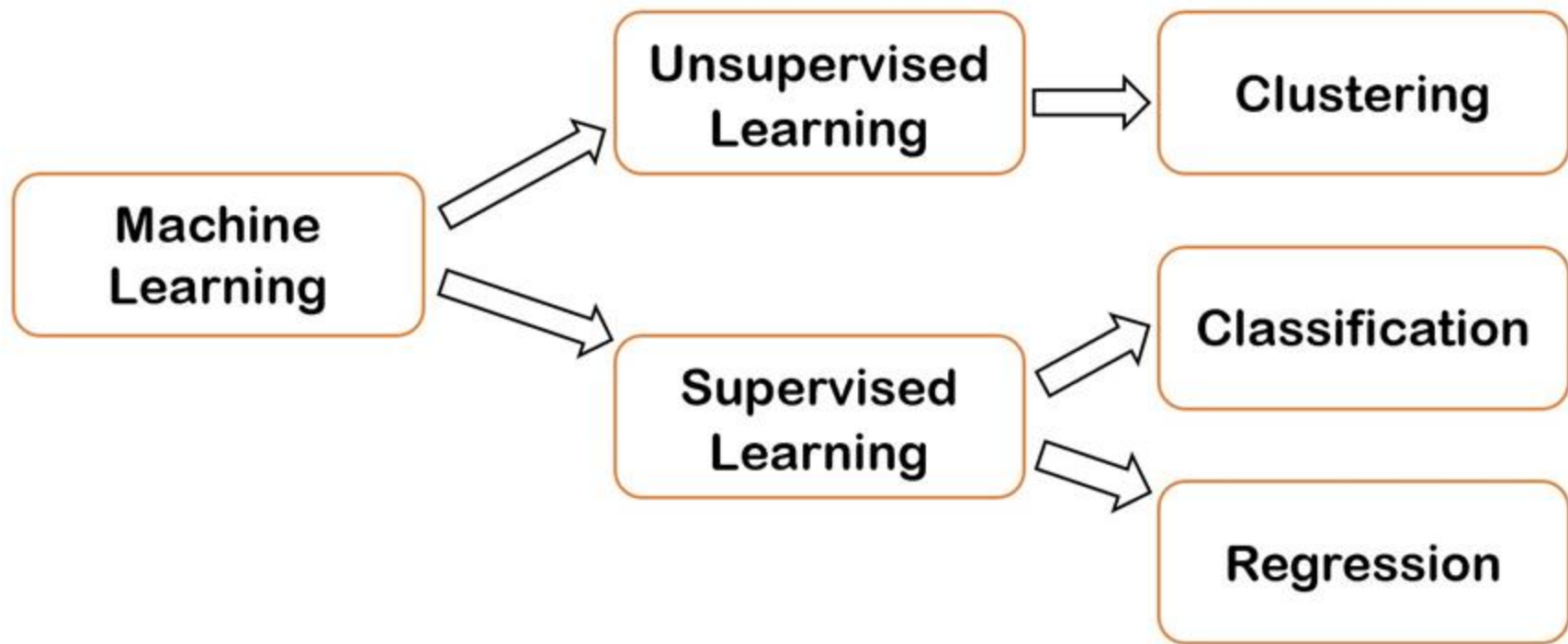


Introducción a la Ciencia de Datos



Supervised Learning

Unsupervised Learning



Preprocesamiento de Datos

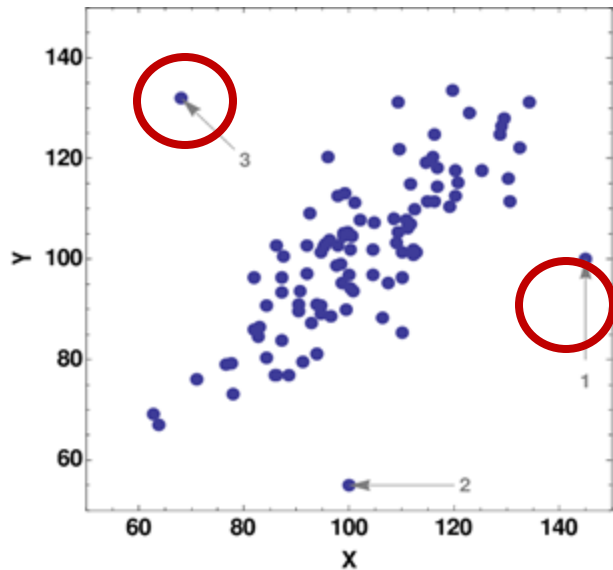
¿CÓMO SE GENERAN LOS VACÍOS?



- ✓ Omisiones durante encuestas
- ✓ Error al guardar datos en sistema
- ✓ Respuesta en alguna función programada. Ejemplos aquí

Es necesario eliminar los vacíos para evitar **errores de ejecución en el código** porque se requiere contar con la información completa

¿CÓMO SE GENERAN LOS OUTLIERS?



✓ **Outliers:** Datos que no se comportan igual que el resto

✓ Generados por error de digitación

✓ **Excepción** a la regla por

anomalía (factores externos)

Es necesario eliminar los outliers para evitar errores de precisión en el **modelo** porque admite como normal valores atípicos

FEATURE ENGINEERING

- ✓ Evaluar si existen **variables correlacionadas** para generar variables derivadas
- ✓ Evaluar mantener **variables relevantes** (*feature importance*)
- ✓ Agregar data externa si no cuenta con data suficiente
- ✓ Ejecutar **encoding o binning**
Esta etapa es considerada la **más importante** del preprocesamiento

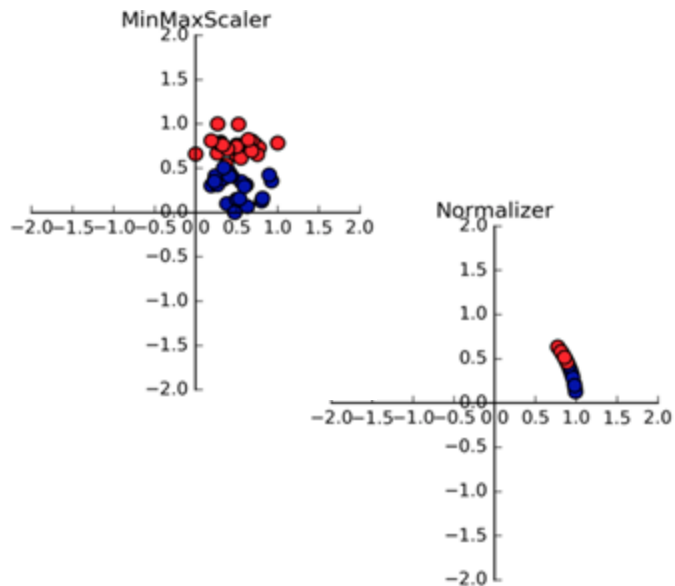
¿QUÉ SIGNIFICA ENCODEAR VARIABLES?



- ✓ **Encodear:** Reemplazar unos símbolos por algún valor que represente mejor su significado.
- ✓ Nosotros usaremos este concepto para traducir las palabras a números

Aprovecha para trasladar al modelo tu interpretación de las **variables categóricas**. **Verifica si el modelo lo requiere** para evitar errores al ejecutar.

¿NORMALIZAR O ESTANDARIZAR?



- ✓ **Normalizar:** Traslada los límites de los datos entre un mínimo y máximo específico

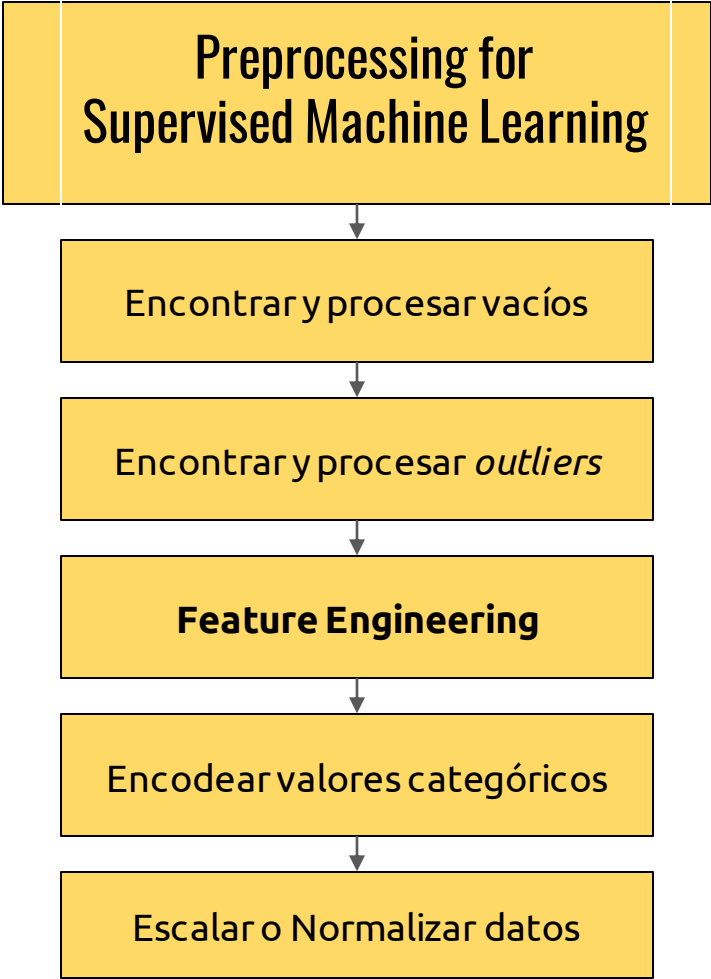
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- ✓ **Estandarizar:** Basado en la media y desviación estándar

$$X' = \frac{X - \mu}{\sigma}$$

Es obligatorio si el modelo predice en función de distancia. Más información aquí.

Preprocessing for Supervised Machine Learning



```
graph TD; A[Preprocessing for Supervised Machine Learning] --> B[Encontrar y procesar vacíos]; B --> C[Encontrar y procesar outliers]; C --> D[Feature Engineering]; D --> E[Encodear valores categóricos]; E --> F[Escalar o Normalizar datos];
```

The diagram is a vertical flowchart with six yellow rectangular boxes connected by downward-pointing arrows. The first box is wider than the others and contains the main title. The subsequent boxes contain specific preprocessing steps in Spanish, with 'Feature Engineering' being in bold. The final box is also wider than the intermediate steps.

Encontrar y procesar vacíos

Encontrar y procesar *outliers*

Feature Engineering

Encodear valores categóricos

Escalar o Normalizar datos