



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS
DIPLOMADO EN MINERÍA DE DATOS

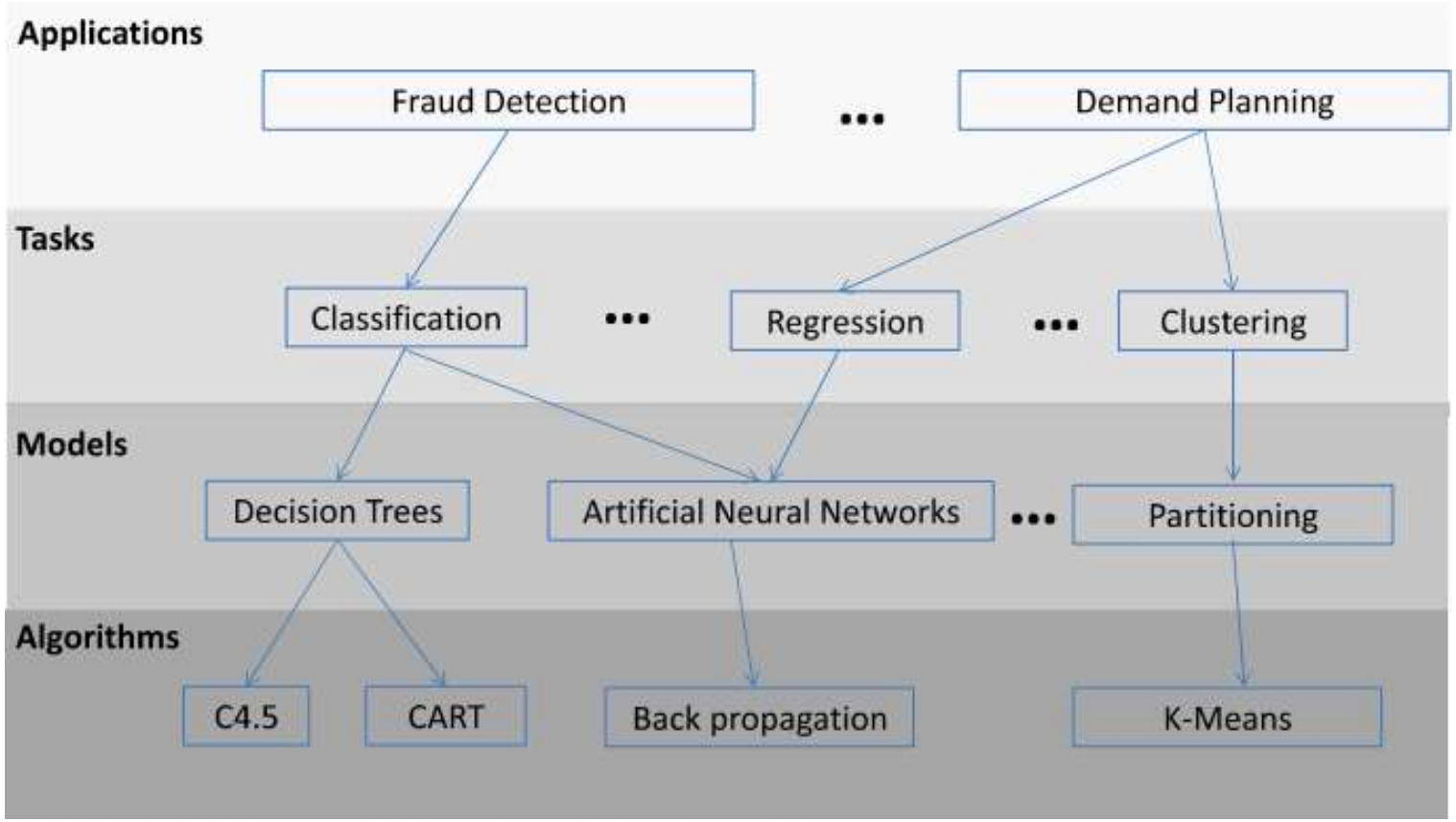
**Módulo 5**. *Minería de Datos*

# Minería de Datos: Clasificación

**Gerardo Avilés Rosas**
gar@ciencias.unam.mx

# Introducción

- La minería de datos agrupa seis actividades: ***Clasificación, Estimación, Predicción, Asociación, Agrupación, Descripción y Visualización***.
- Las tres primeras tareas son ejemplos de la ***minería de datos dirigida*** o ***aprendizaje supervisado***.



# …Introducción

- Como se ha visto, las BD contienen una buena cantidad de información escondida que puede ser usada para tomar **decisiones inteligentes**.
- **Clasificación** y **predicción** son dos formas de análisis de datos que se utilizan para **extraer modelos** que describan importantes clases de datos o predigan tendencias futuras en los mismos.
- Los **modelos de clasificación *predicen*** etiquetas categóricas (*discretas y sin ordenar*):

***Préstamo bancario***

# …Introducción

- Los **modelos de predicción** trabajan con funciones continuas:

- La mayoría de estos algoritmos han sido propuestos en los campos de **máquinas de aprendizaje**, **reconocimiento de patrones** y **estadística**; muchos de ellos **residen en memoria** (*asumen un tamaño pequeño de los datos*).

# ¿Qué es la clasificación?

- Consiste en **predecir** un **resultado determinado** con base en una **entrada dada**.
- **Asigna** a un objeto una **cierta clase** en función de la **similitud** con ejemplos previos de otros objetos.


# …¿Qué es la clasificación?

- En cualquiera de estos ejemplos, se busca construir modelo **(*clasificador*)** que permita predecir etiquetas categóricas:

  ***seguro, riesgo, si, no, tratamiento A, tratamiento B, tratamiento C, etc***.

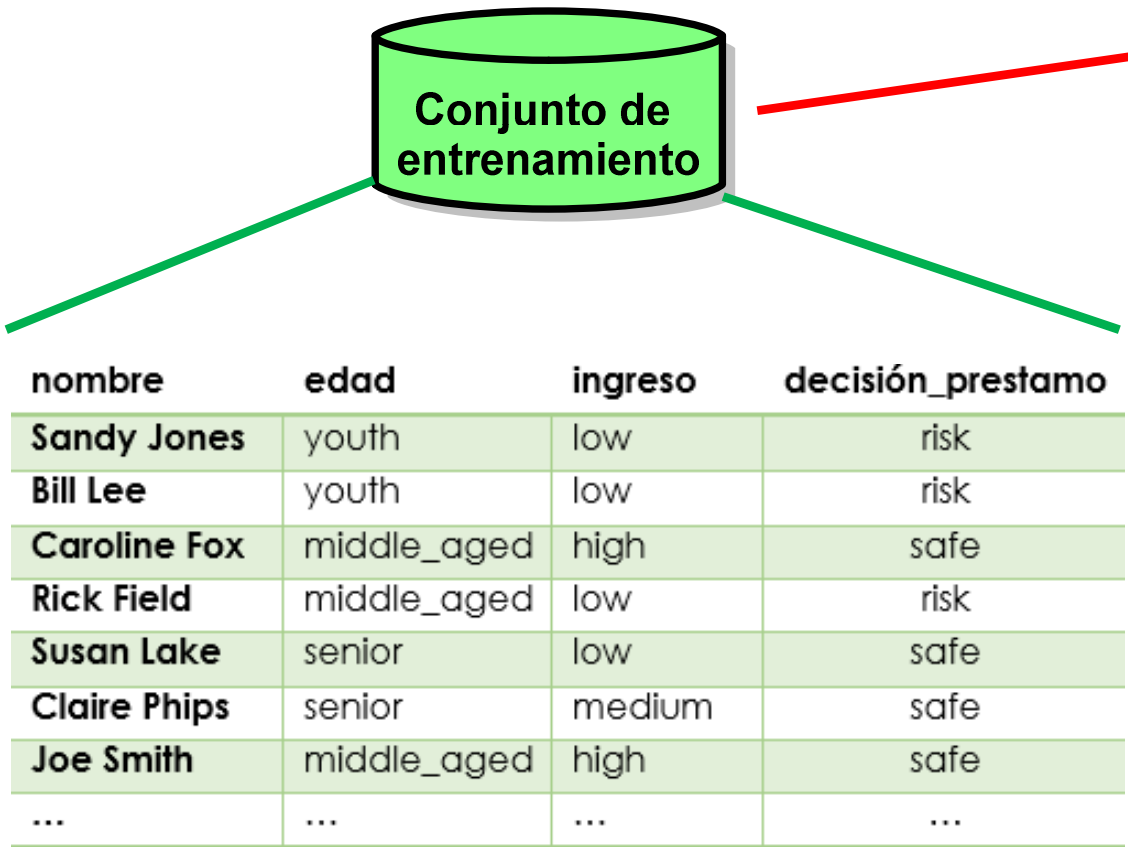
- Estas categorías representar por medio de **valores discretos**, donde el ordenamiento entre los mismos no tiene ningún significado.

# Clasificación: primera fase

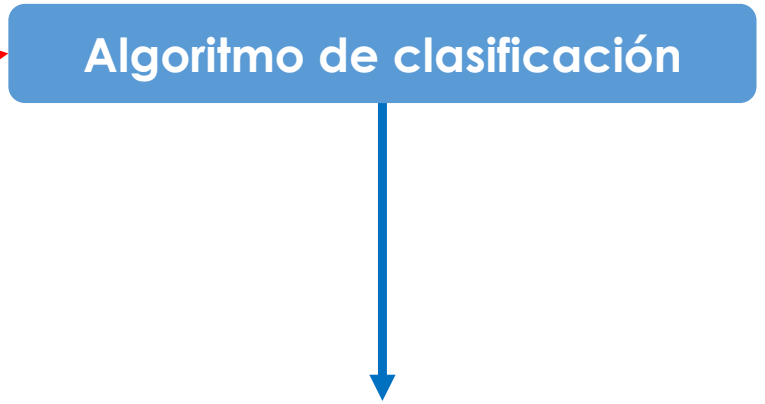
Se trata de un proceso de dos fases, en **primer lugar:**

- Se construye un **modelo** que permita predecir una etiqueta de ***clase***.
- Se utiliza un ***algoritmo de clasificación,*** que construye un ***clasificador*** a través del aprendizaje sobre un **conjunto de entrenamiento** (conjunto de tuplas y sus **etiquetas de clase asociadas** → también se conoce como ***fase de entrenamiento***.
- Una ***tupla X*** representa un ***vector de atributos n-dimensional*** que representa ***n mediciones*** hechas sobre una tupla de **n atributos**.
- Cada ***tupla X*** **pertenece** a una **clase predefinida** (***etiqueta de clase***): *atributo categórico cuyo valor sirve como una categoría*.
- Las tuplas que componen el **conjunto de entrenamiento** se conocen como ***tuplas de entrenamiento*** (*se seleccionan de la BD a través de un análisis*).

# …Clasificación: primera fase




| nombre | edad | ingreso | decisión_prestamo |
|---|---|---|---|
| Sandy Jones | youth | low | risk |
| Bill Lee | youth | low | risk |
| Caroline Fox | middle_aged | high | safe |
| Rick Field | middle_aged | low | risk |
| Susan Lake | senior | low | safe |
| Claire Phips | senior | medium | safe |
| Joe Smith | middle_aged | high | safe |
| … | … | … | … |





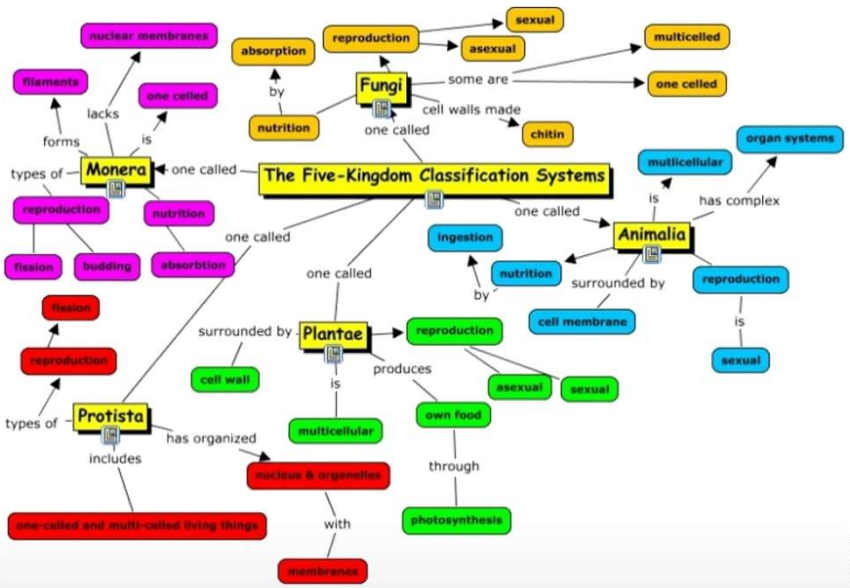
**IF** edad = *youth* **THEN** decisión_prestamo = *risk*
**IF** ingreso = *high* **THEN** decisión_prestamo = *safe*
**IF** edad = *middle_aged* **AND** ingreso = *low*
**THEN** decisión_prestamo = *risk*
…

# …Clasificación: primera fase

- Dado que las etiquetas de clase son proporcionadas para la fase de entrenamiento, este paso es también conocido como ***aprendizaje supervisado***.
- Este primer paso puede verse como un ***mapeo o función c = f(X)***, que puede predecir la etiqueta de clase asociada ***c*** a una ***tupla X*** dada:

  El objetivo del mapeo debe permitir **separar las clases de datos**

- Típicamente este mapeo se representa a través de:
  - ***Reglas de clasificación***
  - ***Árboles de decisión***
  - ***Fórmulas matemáticas***

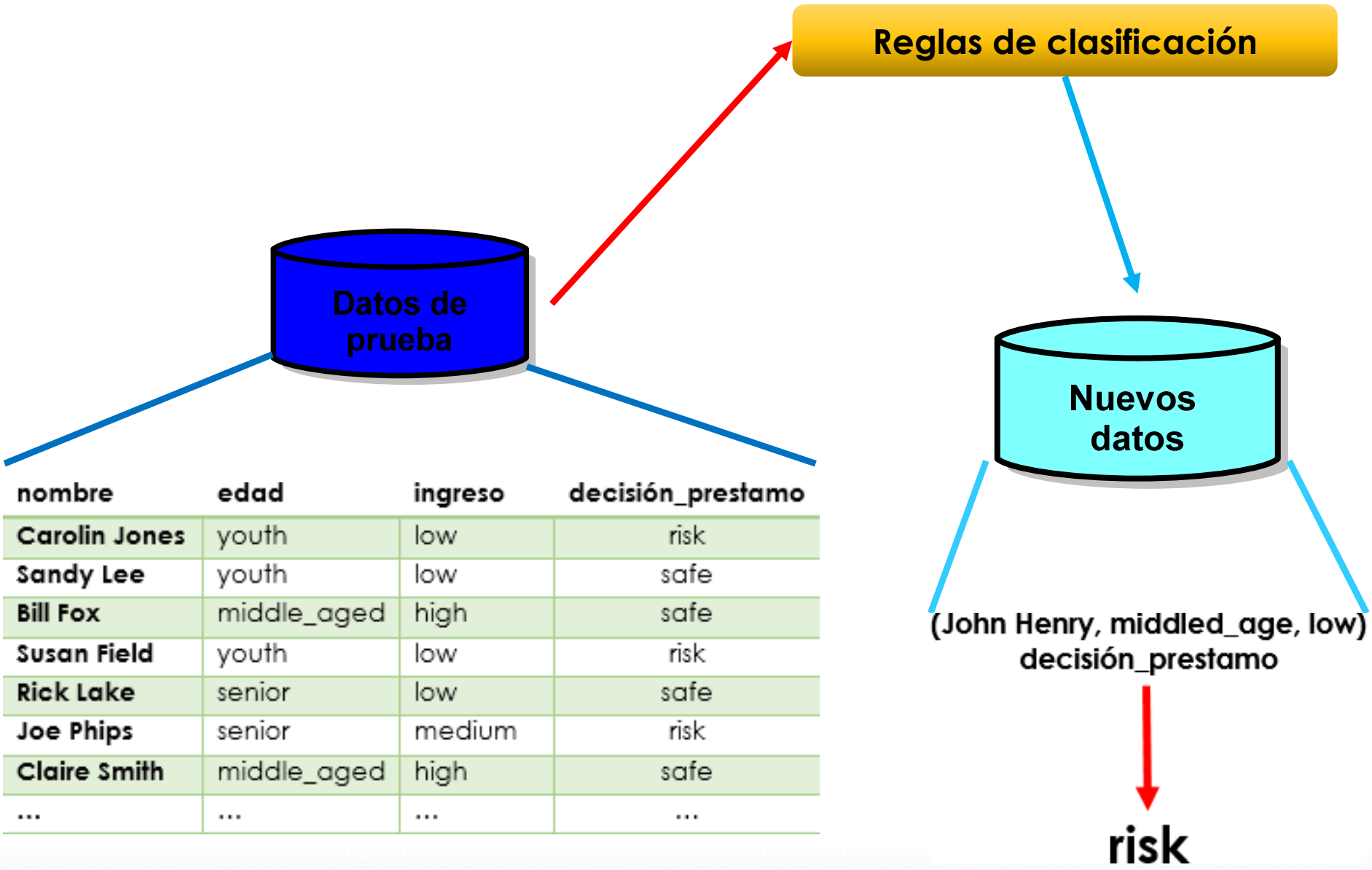


# Clasificación: segunda fase

En la **segunda fase**:

- El **modelo creado** se utiliza para clasificar y así poder ***estimar la exactitud predictiva*** *del clasificador*.
- Se puede utilizar el **conjunto de entrenamiento** para medir la exactitud, pero en este caso, se obtiene una ***estimación bastante optimista***, ya que el clasificador tiende a ***sobreajustar*** los datos.
- Por esta razón se utiliza un ***conjunto de prueba*** (*tuplas de prueba y sus etiquetas de clase asociadas*). Las tuplas se seleccionan de manera aleatoria y son **independientes** del conjunto de tuplas de entrenamiento.
- La ***exactitud del clasificador*** en un conjunto de prueba dado es el porcentaje de tuplas que son correctamente clasificadas por el **clasificador**. Las etiquetas de clase asociadas de cada tupla son comparadas con las clases que predijo el clasificador en la fase de aprendizaje.
- ***Si la exactitud es aceptable, se puede utilizar para clasificar futuras tuplas***.

# …Clasificación: segunda fase




| nombre | edad | ingreso | decisión_prestamo |
|---|---|---|---|
| Carolin Jones | youth | low | risk |
| Sandy Lee | youth | low | safe |
| Bill Fox | middle_aged | high | safe |
| Susan Field | youth | low | risk |
| Rick Lake | senior | low | safe |
| Joe Phips | senior | medium | risk |
| Claire Smith | middle_aged | high | safe |
| … | … | … | … |

# Preprocesamiento de datos

- **Limpieza de datos:**

  Es necesario ***preprocesar*** los datos a fin **de remover o reducir el ruido** y tratar los **valores perdidos** (*missing values*), pues aunque la mayoría de los métodos de clasificación disponen de algunos mecanismos para para manejar este tipo de datos, esto puede ayudar a **reducir la confusión** durante el aprendizaje.

- **Análisis de relevancia:**

  **Muchos de los atributos** en los datos pueden ser **redundantes** o bien irrelevantes, de manera que es importante detectar a aquellos que no contribuyen con la tarea de clasificación. Este tipo de análisis nos puede ayudar a **mejorar la eficiencia y escalabilidad**.

- **Transformación y reducción de datos:**

  Los datos se pueden **normalizar** (*para datos que involucran mediciones de distancia*) o bien **generalizar** (*principalmente utilizado para atributos que poseen valores continuos*).

# Comparación y evaluación

- **Exactitud:**

  Habilidad de predecir las etiquetas de clase de datos nuevos (previamente invisibles). Se estimada usando uno o más conjuntos de prueba que son independientes del conjunto de entrenamiento.

- **Velocidad:**

  Costo computacional involucrado en la generación y uso del clasificador.

- **Robustez:**

  Habilidad del clasificador de realizar predicciones correctas dados datos con ruido o con valores perdidos.

- **Escalabilidad:**

  Habilidad de construir un clasificador que pueda trabajar con grandes cantidades de datos.

- **Interpretabilidad:**

  Nivel de entendimiento y de visión que es proporcionado por el clasificador. Se trata de un aspecto subjetivo y por lo tanto es más difícil de asegurar.


¡Gracias!