Introducción a la minería de datos

Universidad Veracruzana

Facultad de Estadística e Informática

Maestría en Ingeniería de Software

Dr. Ángel Juan Sánchez García

angelsg89@hotmail.com

angesanchez@uv.mx

Xalapa, Ver. Octubre de 2024

¿Qué tenemos hasta ahora?

- Agentes reactivos
- Agentes que planifican
- Agentes que razonan
- ¿Qué sigue…?
- Agentes que generan conocimiento

¿Qué sabemos de bases de datos?

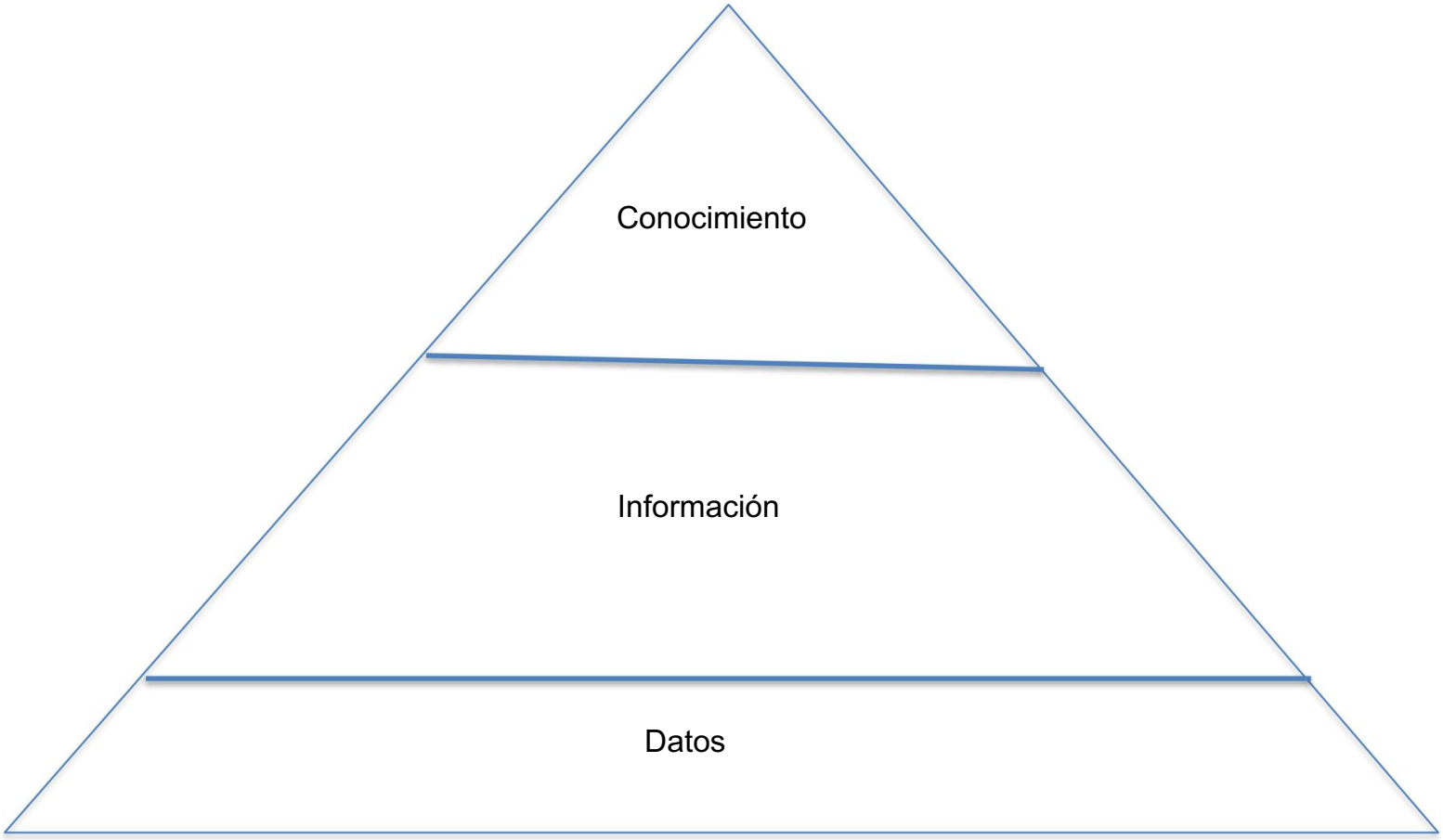
- ¿Diseñar?
- ¿Construir?
- ¿Implementar?
- ¿Acceder?
- ¿Analizar?

¿Datos = Información?

- Datos: Atributos, observaciones, características de algo, combinados sin ningún significado.

- Información: Datos procesados.

Minería de datos: Conocimiento



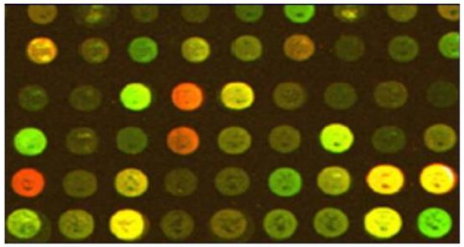
¿Por qué extraer datos? Perspectiva comercial

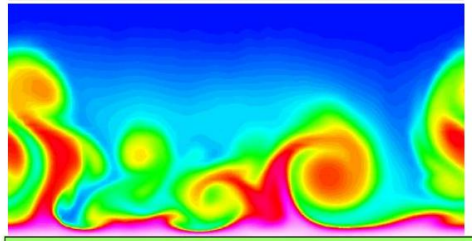
- Se están recopilando muchos datos y almacenado
  - Datos web, comercio electrónico
  - compras en el departamento/ tiendas de comestibles
  - Transacciones bancarias/con tarjeta de crédito

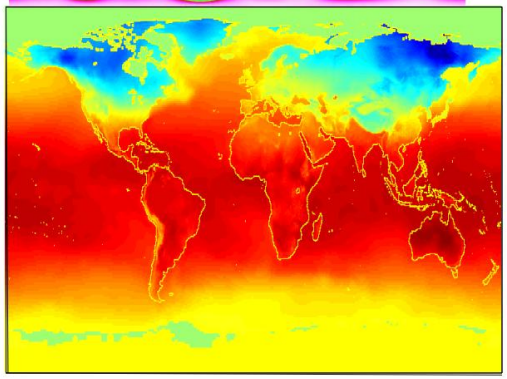
- Las computadoras se han vuelto más baratas y más potentes.
- La presión competitiva es fuerte

Maestría en Ingeniería de Software

¿Por qué extraer datos? Perspectiva científica





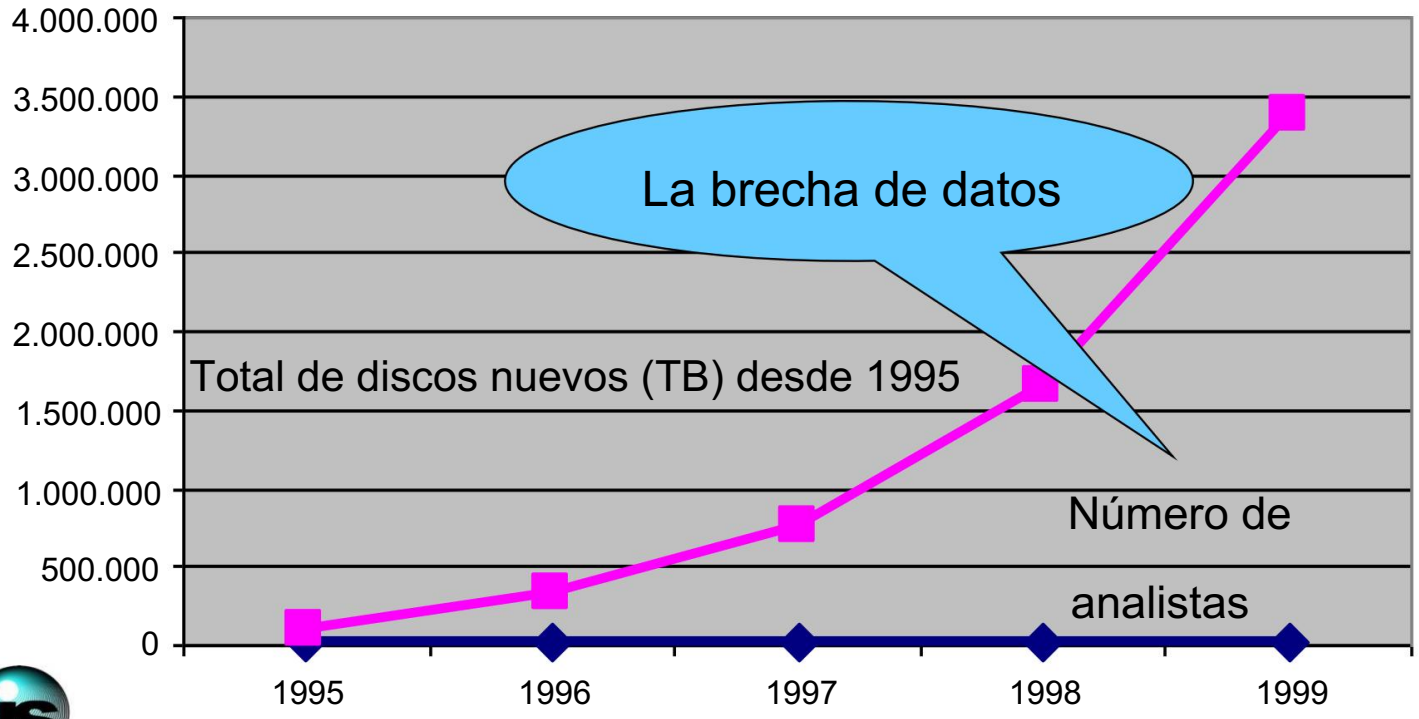


- Datos recopilados y almacenados a enormes velocidades (GB/hora)
  - sensores remotos en un satélite
  - telescopios escaneando los cielos
  - microarrays que generan datos de expresión genética
  - simulaciones científicas que generan terabytes de datos
- Las técnicas tradicionales no son viables para datos brutos
- La minería de datos puede ayudar a los científicos
  - en la clasificación y segmentación de datos
  - en la formación de hipótesis

Maestría en Ingeniería de Software

Minería de grandes conjuntos de datos: motivación

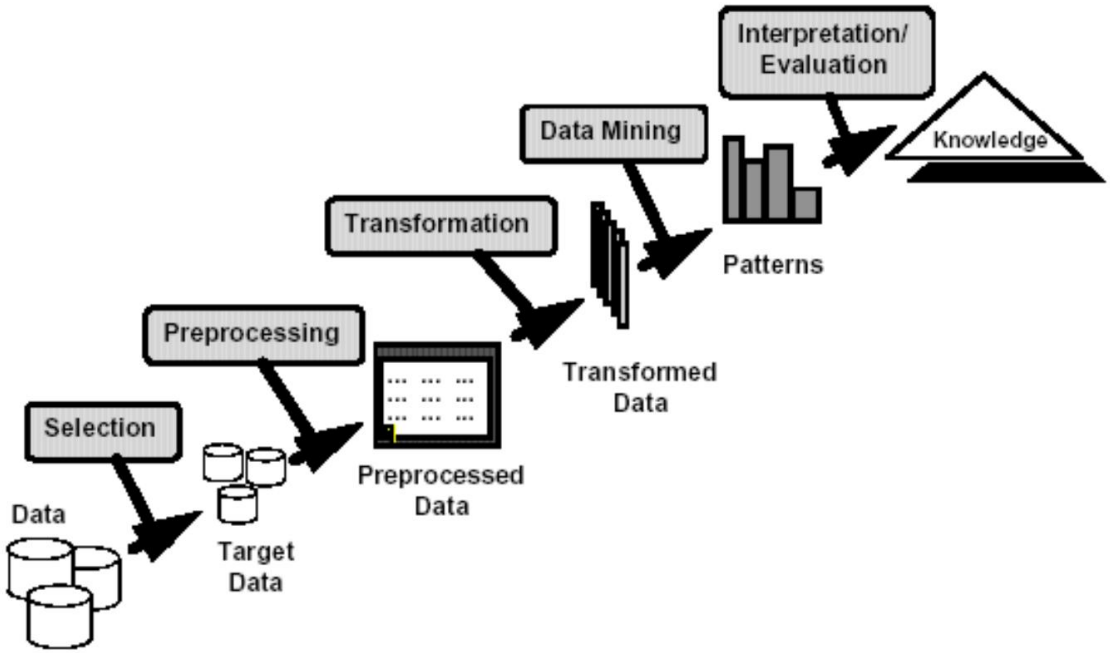
- A menudo hay información “oculta” en los datos que se no es fácilmente evidente
- Los analistas humanos pueden tardar semanas en descubrir información útil
- Gran parte de los datos nunca se analizan en absoluto



De: R. Grossman, C. Kamath, V. Kumar, “Minería de datos para aplicaciones científicas y de ingeniería”

¿Qué es la minería de datos?

- Muchas definiciones
  - Extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos
  - Exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos para descubrir patrones significativos



Maestría en Ingeniería de Software

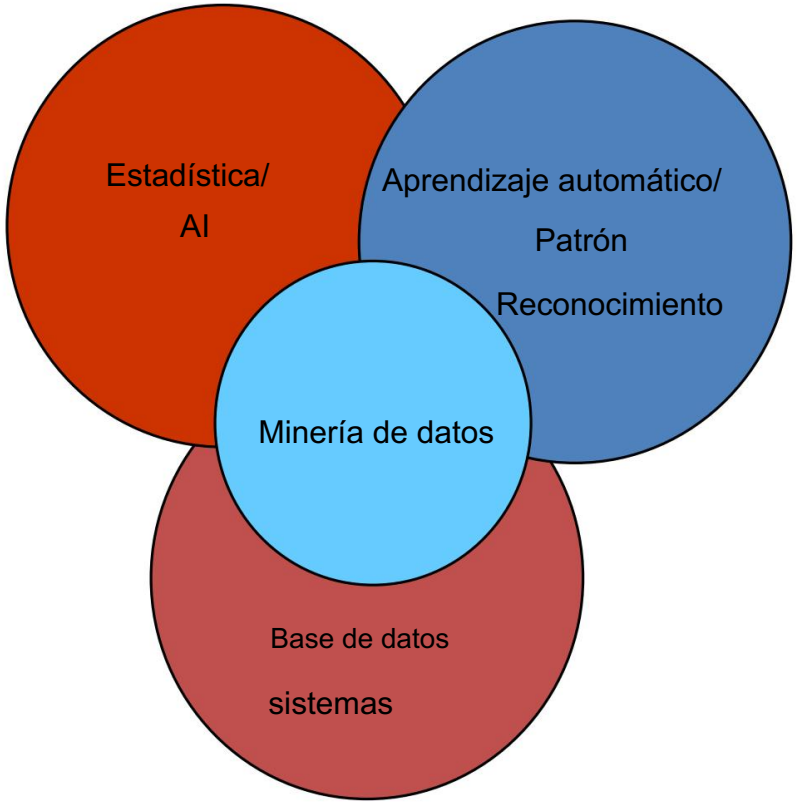
¿Qué es (no) la minería de datos?

- ¿Qué no es Minería de Datos?
  - Busque el número de teléfono en el directorio telefónico
  - Consultar un motor de búsqueda web para obtener información sobre "Amazonas"

- ¿Qué es la minería de datos?
  - Ciertos nombres son más frecuentes en ciertas ubicaciones de EE. UU.
  - Agrupar documentos similares devueltos por el motor de búsqueda según su contexto (por ejemplo, selva amazónica, Amazon.com).

Orígenes de la minería de datos

- Extrae ideas del aprendizaje automático/IA, reconocimiento de patrones, estadísticas y sistemas de bases de datos.



Maestría en Ingeniería de Software

Tareas de minería de datos

- Métodos de predicción
  - Utilizar algunas variables para predecir lo desconocido o lo futuro. valores de otras variables.

- Métodos de descripción
  - Encontrar patrones interpretables por humanos que describan la datos.

De [Fayyad, et.al.] Avances en el descubrimiento de conocimientos y la minería de datos, 1996

Tareas de minería de datos...

- Agrupamiento [Descriptivo]
- Descubrimiento de reglas de asociación

[Descriptivo] • Descubrimiento de patrones

secuenciales [Descriptivo] •

Clasificación [Predictivo] • Regresión [Predictivo]

Definición de agrupamiento

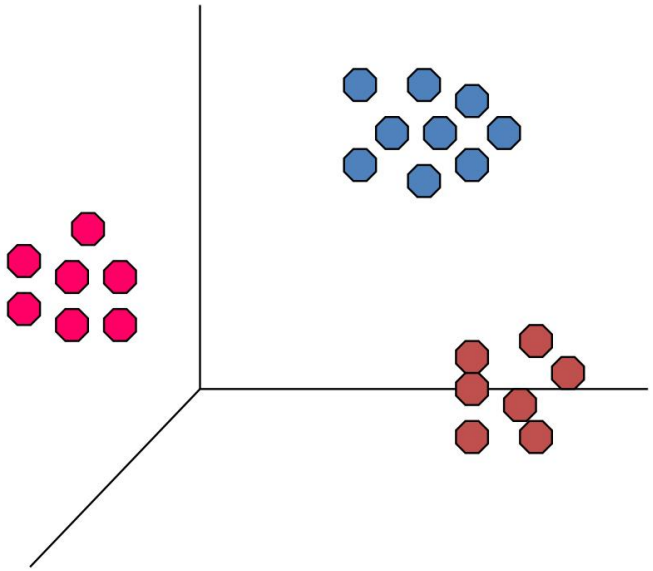
- Dado un conjunto de puntos de datos, cada uno con un conjunto de atributos y una medida de similitud entre ellos, encuentre grupos tales que
  - Los puntos de datos de un grupo son más similares entre sí. otro.
  - Los puntos de datos en grupos separados son menos similares entre sí otro.
- Medidas de similitud:
  - Distancia euclidiana si los atributos son continuos.
  - Otras medidas específicas del problema.

Ilustración de la agrupación en clústeres

- Agrupamiento basado en la distancia euclidiana en el espacio 3D.

Las distancias entre clústeres se minimizan

Las distancias entre clústeres se maximizan



Agrupamiento: Aplicación 1

- Segmentación del mercado:
  - Objetivo: subdividir un mercado en subconjuntos distintos de clientes, donde cualquier subconjunto pueda ser seleccionado como un objetivo de mercado que se alcanzará con una combinación de marketing distinta.
  - Enfoque: -

    Recopilar diferentes atributos de los clientes en función de su información geográfica y relacionada con su estilo de vida.

    - Encontrar grupos de clientes similares.
    - Medir la calidad del agrupamiento observando los patrones de compra de los clientes del mismo grupo en comparación con los de grupos diferentes.

Agrupamiento: Aplicación 2

- Agrupación de documentos:
  - Objetivo: encontrar grupos de documentos que sean similares entre sí en función de los términos importantes que aparecen en ellos.
  - Enfoque: Identificar términos frecuentes en cada documento. Generar una medida de similitud basada en la frecuencia de los diferentes términos. Utilízala para agrupar.
  - Ganancia: La recuperación de información puede utilizar los clústeres para relacionar un nuevo documento o término de búsqueda con documentos agrupados.

Ilustración de la agrupación de documentos

- Puntos de agrupación: 3204 artículos de Los Angeles Times.
- Medida de similitud: cuántas palabras son comunes en estos documentos (después de algún filtrado de palabras).

Categoría	Total	Artículos colocados correctamente
Financiero	555	364
Extranjero	341	260
Nacional	273	36
Metro	943	746
Deportes	738	573
Entretenimiento	354	278

Descubrimiento de reglas de asociación: definición

- Dado un conjunto de registros, cada uno de los cuales contiene cierto número de elementos de una colección determinada;
  - Producir reglas de dependencia que predecirán la ocurrencia de un elemento en función de las ocurrencias de otros elementos.

Elementos	TID
1	Pan, Coca-Cola, Leche
2	Cerveza, Pan
3	Cerveza, Coca-Cola, Pañal, Leche
4	Cerveza, Pan, Pañal, Leche
5	Coca-Cola, pañal, leche

Reglas descubiertas:

{Leche} --> {Coca-Cola}

{Pañal, Leche} --> {Cerveza}

Descubrimiento de reglas de asociación: Aplicación 1

- Marketing y promoción de ventas:
  - Sea la regla descubierta

    {Bagels, …} --> {Papas fritas}

  - Papas fritas como consecuente => Se puede utilizar para determinar ¿Qué se debe hacer para aumentar sus ventas?
  - Bagels en el antecedente => Se puede usar para ver cuál Los productos se verían afectados si la tienda deja de vender bagels.
  - Bagels en antecedente y Papas fritas en consecuente => ¡ Se puede usar para ver qué productos se deben vender con Bagels para promover la venta de Papas fritas!

Descubrimiento de reglas de asociación: Aplicación 2

- Gestión de estanterías de supermercados.
  - Objetivo: Identificar los artículos que se compran juntos Suficientes clientes.
  - Enfoque: Procesar los datos del punto de venta recopilados con lectores de códigos de barras para encontrar dependencias entre los artículos.
  - Una regla clásica --

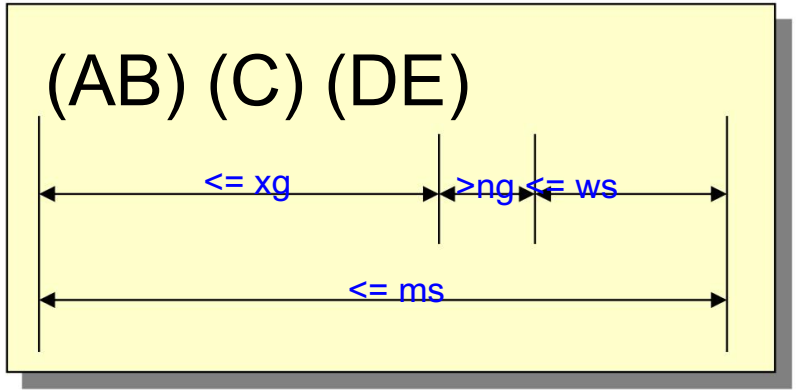
    - Si un cliente compra pañales y leche, es muy probable que compre cerveza.

    - Así que no te sorprendas si encuentras paquetes de seis apilados uno al lado del otro. ¡pañales!

Descubrimiento de patrones secuenciales: definición

• Dado un conjunto de objetos, cada uno asociado con su propia línea de tiempo de eventos, encuentre reglas que predigan fuertes dependencias secuenciales entre diferentes eventos.

• Las reglas se forman al descubrir primero patrones. Las ocurrencias de eventos en los patrones son gobernado por restricciones de tiempo.



Descubrimiento de patrones secuenciales: ejemplos

- En los registros de alarmas de telecomunicaciones,
  - (Problema del inversor: Corriente de línea excesiva)

    (Alarma_Rectificador) --> (Alarma_Incendio)
- En las secuencias de transacciones en el punto de venta,
  - Librería de informática:

    (Introducción a Visual C) (Introducción a C++) -->

    (Perl para principiantes, Tcl Tk)
  - Tienda de ropa deportiva:

    (Calzado) (Raqueta, Racketball) --> (Chaqueta deportiva)

Clasificación: Definición

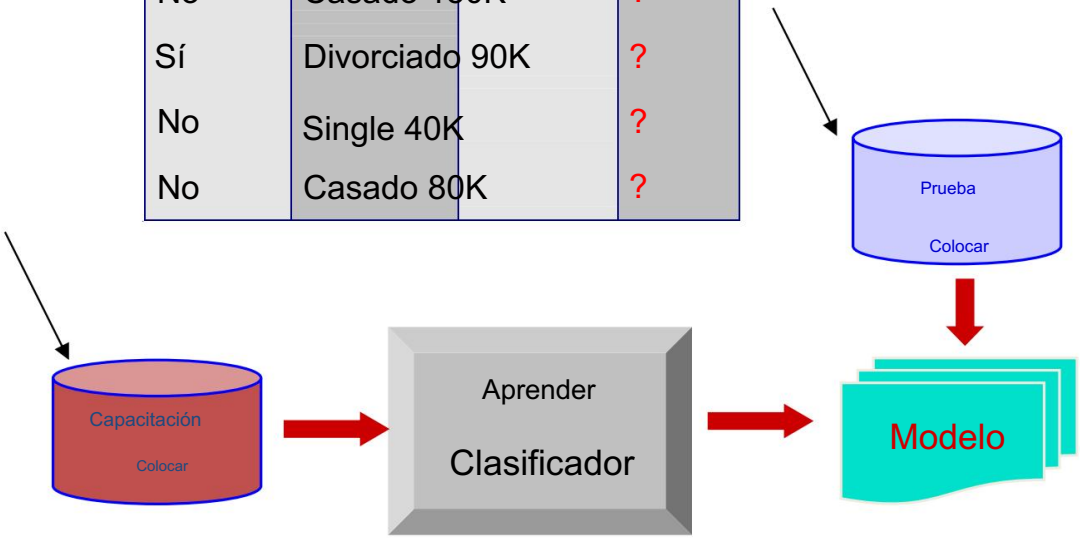
- Dada una colección de registros (conjunto de entrenamiento)
  - Cada registro contiene un conjunto de atributos, uno de los atributos es la clase.
- Encontrar un modelo para el atributo de clase como función de los valores de otros atributos.
- Objetivo: a los registros nunca antes vistos se les debe asignar una clase con la mayor precisión posible.
  - Se utiliza un conjunto de pruebas para determinar la precisión del modelo.

    Generalmente, el conjunto de datos dado se divide en conjuntos de entrenamiento y de prueba, siendo el conjunto de entrenamiento el utilizado para construir el modelo y el conjunto de prueba el utilizado para validarlo.

Ejemplo de clasificación

	categorical	categorical	continuous	class
	Reembolso	Estado civil	Ingresos imponibles	
1	Sí	Individual	125K	No
2	No	Casado	100K	No
3	No	Sencillo	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Sencillo	85K	Sí
9	No	Casado	75K	No
10	No	Sencillo	90K	Sí

Reembolso matrimonial	Estado	Imponible	Trampa de ingresos
No	Sencillo	75K	?
Sí	Casado	50K	?
No	Casado	150K	?
Sí	Divorciado	90K	?
No	Single	40K	?
No	Casado	80K	?



Maestría en Ingeniería de Software

Clasificación: Aplicación 1

- Marketing directo
  - Objetivo: Reducir el costo del envío de correo al dirigirse a un grupo de consumidores con probabilidades de comprar un nuevo producto de telefonía celular.
  - Enfoque:
    - Utilice los datos de un producto similar introducido anteriormente.
    - Sabemos qué clientes decidieron comprar y cuáles decidieron
      De lo contrario, esta decisión de {comprar, no comprar} constituye el atributo de clase.
    - Recopilar diversos datos demográficos, de estilo de vida y de interacción con la empresa.
      información relacionada sobre todos esos clientes.
      - Tipo de negocio, donde se alojan, cuánto ganan, etc.
    - Utilice esta información como atributos de entrada para aprender un modelo de clasificador.

De [Berry & Linoff] Técnicas de minería de datos, 1997

Clasificación: Aplicación 2

- Detección de fraude
  - Objetivo: Predecir casos fraudulentos en transacciones con tarjetas de crédito.
  - Enfoque:
    - Utilizar las transacciones de tarjetas de crédito y la información del titular de su cuenta como atributos.
      - ¿ Cuándo compra un cliente, qué compra, con qué frecuencia paga? tiempo, etc.
    - Etiquetar las transacciones pasadas como fraudulentas o justas. Esto forma el atributo de clase.
    - Aprender un modelo para la clase de las transacciones.
    - Utilice este modelo para detectar fraudes mediante la observación de transacciones con tarjetas de crédito. en una cuenta.

Clasificación: Aplicación 3

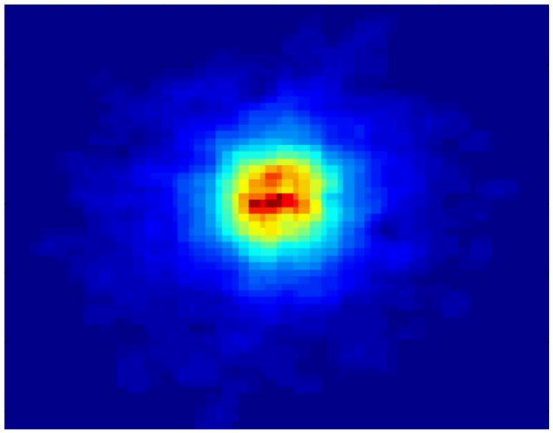
- Catalogación de estudios del cielo
  - Objetivo: Predecir la clase (estrella o galaxia) de los objetos del cielo, especialmente los visualmente débiles, basándose en las imágenes del estudio telescópico (del Observatorio Palomar).
    - 3000 imágenes con 23.040 x 23.040 píxeles por imagen.
  - Enfoque:
    - Segmentar la imagen.
    - Medir atributos de imagen (características): 40 de ellos por objeto.
    - Modelar la clase basándose en estas características.
    - Historia de éxito: Se pudieron encontrar 16 nuevos cuásares de alto corrimiento al rojo, algunos de los cuales ¡Objetos más lejanos que son difíciles de encontrar!

De [Fayyad, et.al.] Avances en el descubrimiento de conocimientos y la minería de datos, 1996

Clasificación de galaxias

Cortesía: http://aps.umn.edu

Temprano



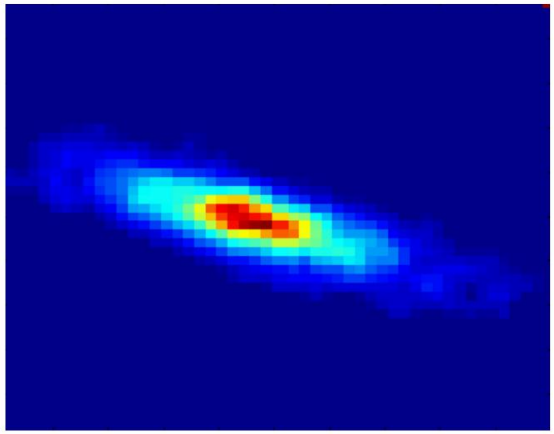
Clase:

- Etapas de Formación

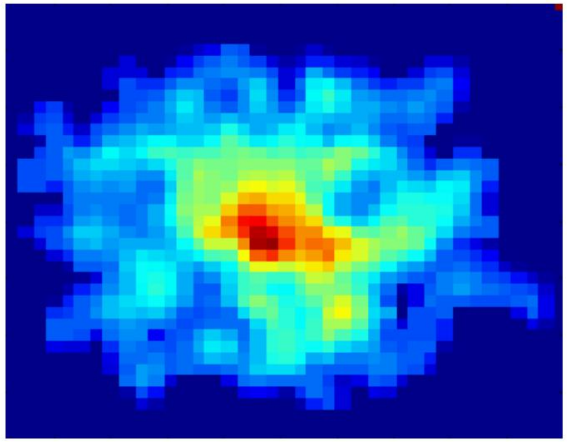
Atributos:

- Características de la imagen,
- Características de las ondas de luz recibidas, etc.

Intermedio



Tarde



Tamaño de los datos:

- 72 millones de estrellas, 20 millones de galaxias
- Catálogo de objetos: 9 GB
- Base de datos de imágenes: 150 GB

Maestría en Ingeniería de Software

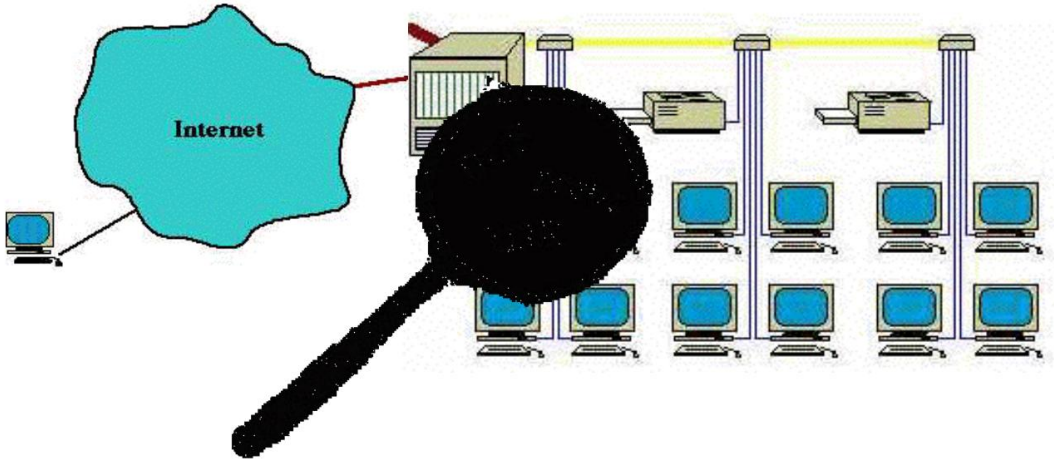
Regresión

- Predecir un valor de una variable continua dada basándose en los valores de otras variables, asumiendo un modelo de dependencia lineal o no lineal.
- Ampliamente estudiado en estadística y campos de redes neuronales.
- Ejemplos:
  - Predecir montos de ventas de nuevos productos en función del gasto publicitario.
  - Predecir la velocidad del viento en función de la temperatura, la humedad, la presión del aire, etc.
  - Predicción de series temporales de índices bursátiles.

Detección de desviaciones/anomalías

- Detectar desviaciones significativas del comportamiento normal • Aplicaciones:
  - Detección de fraude con tarjetas de crédito

  - Intrusión en la red Detección



El tráfico de red típico a nivel universitario puede alcanzar más de 100 millones de conexiones por día.

Desafíos de la minería de datos

- Escalabilidad
- Dimensionalidad •

Datos complejos y heterogéneos • Calidad

de los datos •

Propiedad y distribución de los datos •

Preservación de la

privacidad • Transmisión de datos