



# La Ciencia de Datos y sus aplicaciones

DRA. MARÍA DEL PILAR ANGELES
CENTRO VIRTUAL DE COMPUTACIÓN, UNAM
PILARANG@UNAM.MX

# Ciencia de datos

Campo (inter/trans)disciplinario que involucra

- métodos científicos
- Procesos
- Sistemas

para extraer conocimiento o

mejor entendimiento de datos

- estructurados o no estructurados

Utiliza diversas disciplinas para el análisis descriptivo, predictivo, prescriptivo de datos como:

- Estadística computacional
- minería de datos
- aprendizaje automático

También se define La ciencia de datos como "Un concepto para unificar estadísticas, análisis de datos, aprendizaje automático y sus métodos relacionados para comprender y analizar los fenómenos reales"

# Arquitecto de datos

Verifica recursos de infraestructura, personal, tipo de integración, tipo de análisis, cantidad de datos, tiempo de respuesta, tipo de información y legislaciones que se deben cumplir según el sector.

Determina qué sistemas de gestión de datos son apropiados según la estrategia de negocio.

Propone software, hardware, middleware que permitan la implementación de la solución.

# Ingeniero de datos

Los ingenieros de datos a menudo luchan con problemas asociados con la integración de bases de datos y conjuntos de datos no estructurados y desordenados. Su objetivo final es proporcionar datos limpios y utilizables a quien lo requiera.

El ingeniero debe tener una comprensión clara de cómo es el ciclo de vida de los datos y adecuarlos para reducir el componente de error humano.

Los ingenieros de datos limpian, preparan y optimizan los datos para el consumo.

Una vez que los datos se vuelven útiles se entregan a los científicos de datos.

# Científico de datos

Es una persona formada en ciencias matemáticas y computacionales con experiencia en cierta área de negocio o conocimiento que puede identificar que algoritmos y parámetros de análisis son los adecuados según la información con la que se cuenta para lograr ciertos objetivos.

Además debe ser el enlace entre la estrategia de negocio, los métodos científicos, su interpretación y aplicación para lograr dichos objetivos.

# Científico de datos

Deben identificar que tarea es la mas conveniente y por cada tarea que técnica es la mejor, si no existe una, entonces diseñar y programar algoritmos que se ajusten a los datos y proporcionen el modelo matemático requerido.

Pueden realizar una variedad de análisis y técnicas de visualización para comprender verdaderamente los datos y, eventualmente, contar una historia, realizar predicciones o descubrir conocimiento a partir de los datos.

# Trabajo en equipo

La comunicación entre un ingeniero de datos y un científico de datos es vital.

El científico le indica el formato deseado de los datos dependiendo de la técnica de minería de datos a utilizar.

Típicamente, los datos no sólo se almacenan en una base de datos que aguarda el consumo.

Los datos tienen que ser optimizados para el caso de uso del científico de datos.

# Trabajo en equipo

Los ingenieros de datos pueden trabajar en estrecha colaboración con arquitectos de datos (para determinar qué sistemas de gestión de datos son apropiados) y científicos de datos (para determinar qué datos son necesarios para el análisis).

El NY Times escribió que entre el 50 y el 80 por ciento del trabajo de un científico de datos es la limpieza de datos.

# Competencias del Científico de Datos

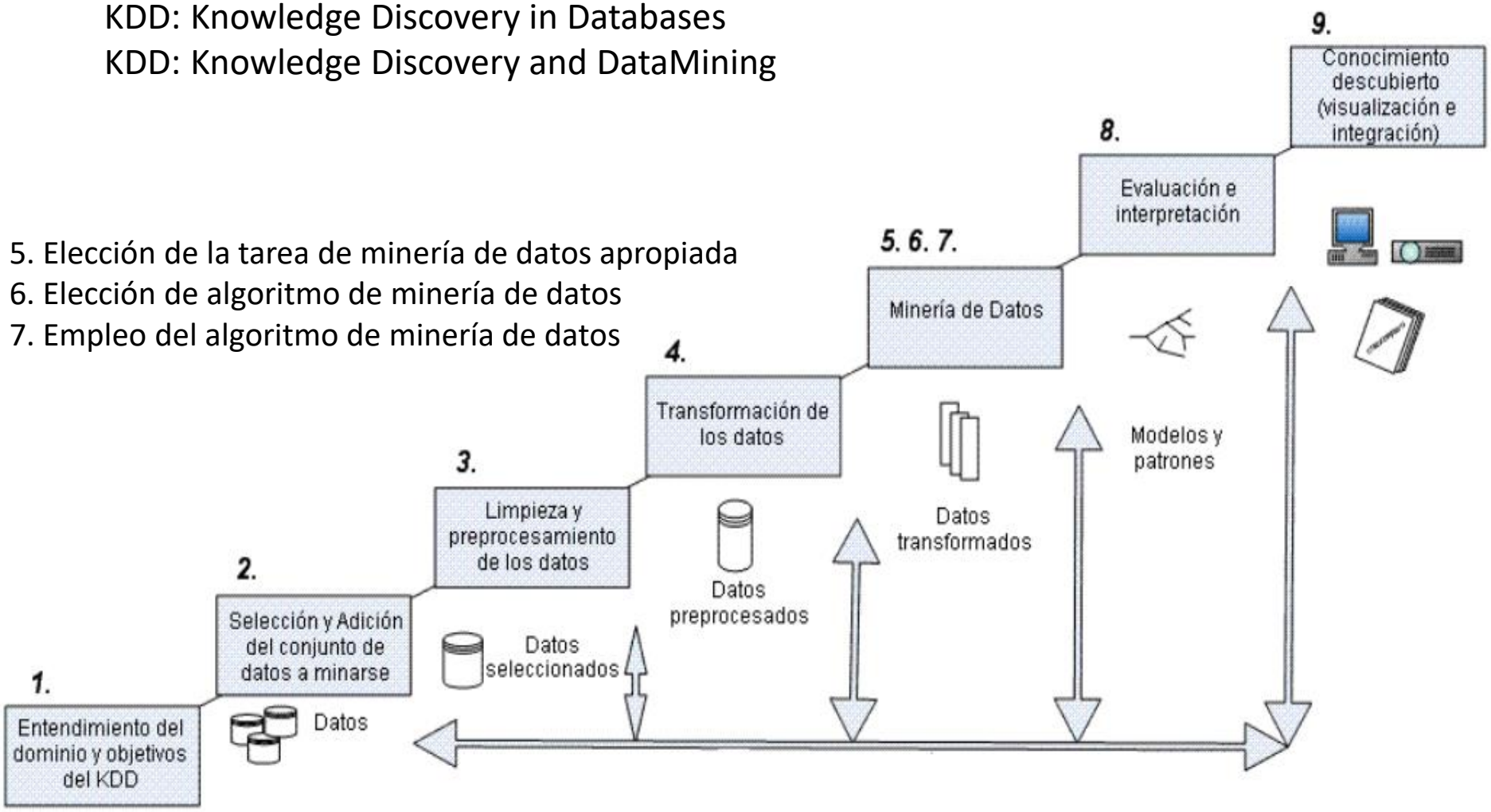
Generar alternativas de solución a las problemáticas en el manejo de datos e información de los sectores productivos, públicos y privados, utilizando modelos analíticos, herramientas computacionales y matemáticas.

- Seleccionar, adaptar y utilizar herramientas computacionales y matemáticas pertinentes para la recolección, extracción, almacenamiento, integración y manejo de distintos tipos de datos masivos e información conducentes a la resolución del problema.
- Identificar las características de los datos para determinar y establecer criterios y métricas para la correcta evaluación de la calidad de los datos en el contexto del problema.
- Modelar la información y los datos con base en estrategias de pre procesamiento y representación de datos.
- Traducir el modelo a toma de decisiones/acciones
- Mejora continua de todo el proceso

# Antecedente a la ciencia de datos
# Proceso de descubrimiento de conocimiento

KDD: Knowledge Discovery in Databases
KDD: Knowledge Discovery and DataMining

5. Elección de la tarea de minería de datos apropiada
6. Elección de algoritmo de minería de datos
7. Empleo del algoritmo de minería de datos

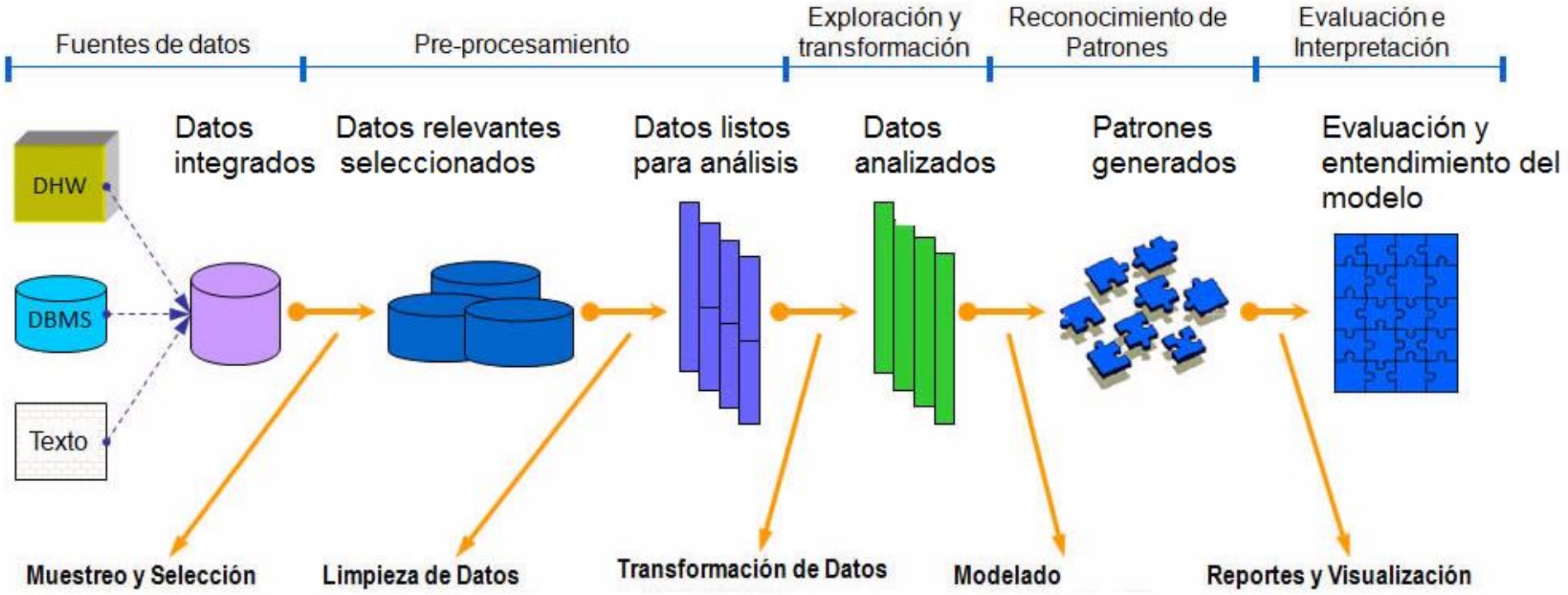


# En términos de los datos…

Una vez realizado el análisis del negocio y definido una estrategia, se plantean los objetivos técnicos que preguntas/conocimiento se deben responder/descubrir a partir de qué datos.

Procesamiento de datos

- Preparación y curado de los datos que permitan el modelado y tratamiento de éstos



# Paso 1. Desarrollar un entendimiento del dominio de la aplicación.

El encargado del proyecto KDD necesita entender y definir los objetivos del usuario final y el ambiente en el cual tomará lugar el proceso de descubrimiento de conocimiento (incluyendo conocimiento previo relevante).

A medida que avanza el proceso KDD, se puede hacer una revisión de este paso.

Teniendo entendido los objetivos del KDD, se inicia el pre-procesamiento de los datos, definidos en los tres próximos pasos.


# Paso 2 Selección de datos

Seleccionar y crear un conjunto de datos en el cual se ejecutará el descubrimiento.

Teniendo definidos los objetivos, se deben determinar los datos que serán usados para el descubrimiento del conocimiento.

Averiguar qué datos están disponibles.

La integración de todos los datos para el descubrimiento de conocimientos en un conjunto de datos.

Este proceso es muy importante porque la minería de datos aprende y descubre conocimiento de los datos disponibles. Si faltan algunos datos importantes, el estudio entero puede fallar.

Por otra parte, recoger, organizar y operar repositorios complejos de datos es costoso y hay una compensación con la posibilidad de comprender mejor los fenómenos.

# Paso 3. Limpieza y pre-procesamiento de los datos.

En esta etapa, la confiabilidad en los datos se eleva. Incluye la claridad en los datos, tal como el manejo de valores faltantes y la remoción de ruido o datos anómalos.

Puede convertirse en la mayor parte (en términos de tiempo invertido) de un proyecto de KDD.

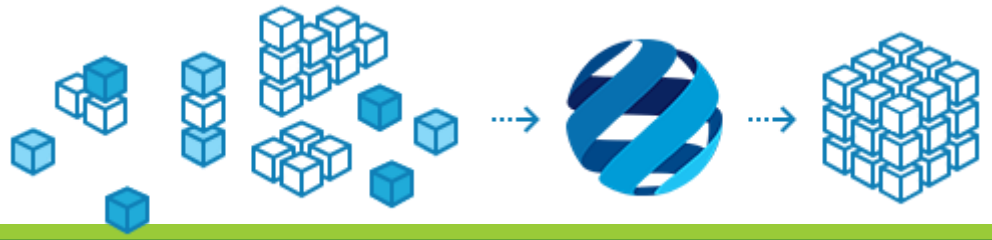
Puede involucrar métodos estadísticos complejos o el uso de algoritmos de Data Mining (DM) en este contexto. Por ejemplo, si uno sospecha que cierto atributo de confiabilidad insuficiente o tiene muchos datos faltantes, entonces este atributo puede convertirse en el objetivo de un algoritmo de minería de datos supervisado. Se desarrollará un modelo de predicción para este atributo y a continuación los datos faltantes se pueden predecir.

# Paso 4. Transformación de los datos

En esta etapa, se preparan mejores datos para la minería.

Selección de características, extracción y registro de muestras), y transformación de atributos (p.e. discretización de atributos numéricos y transformación funcional). Este paso puede ser crucial para el éxito del proyecto KDD entero, y suele ser específico para cada proyecto.

Teniendo completos los cuatros pasos mencionados, los siguientes cuatro pasos están relacionados con la parte de minería de datos, en donde el enfoque está en los aspectos algorítmicos para cada proyecto.



# ¿Qué es la minería de datos?

**Minería de datos:**

proceso que intenta descubrir patrones a partir de conjuntos de datos y que utiliza por ejemplo:

**Computación**

Manejo de Datos

inteligencia artificial

Aprendizaje automático

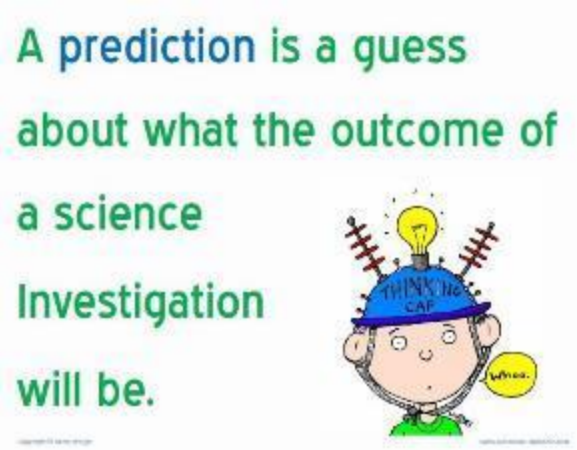
**Matemáticas**

Estadística, distancias, etc.

**Tareas:**

Descripción, Clasificación, Estimación, Predicción, Agrupación, Asociación

# Paso 5. Elección de la tarea de minería de datos apropiada

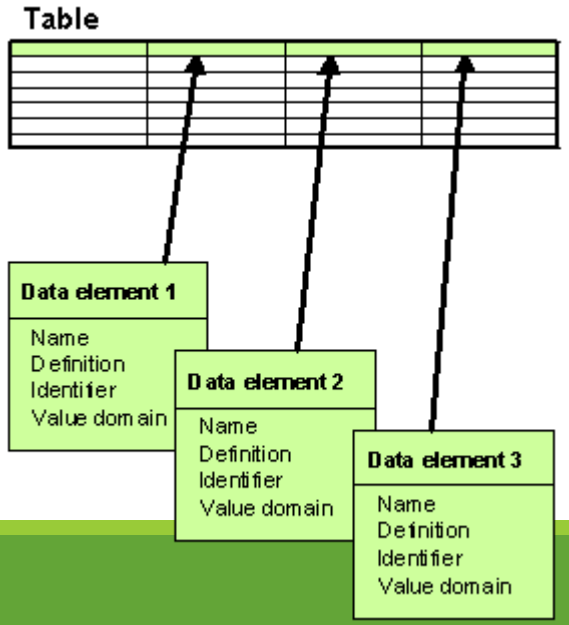



Consiste en decidir cuál tipo de tarea de minería usar, por ejemplo, clasificación, regresión o agrupamiento. Esto depende sobre todo de los objetivos de KDD, y también en los pasos anteriores.

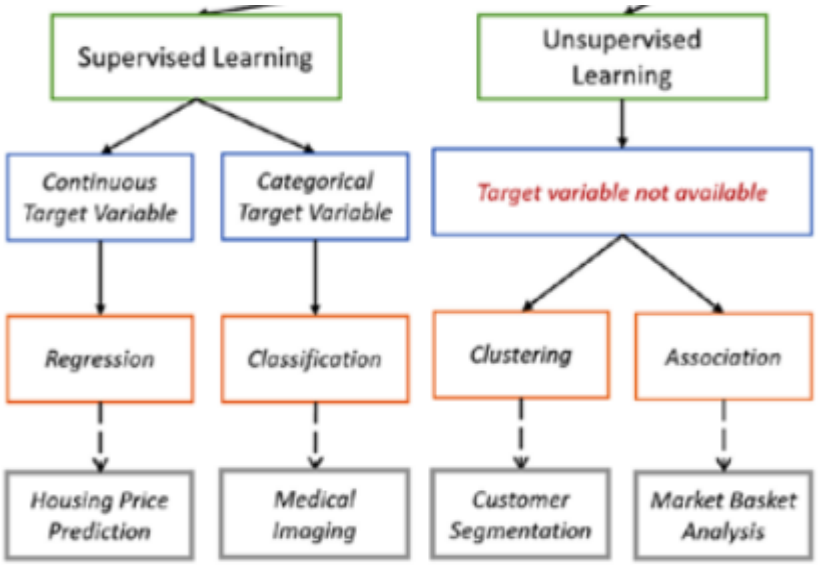
Hay dos objetivos principales en Minería de Datos: predicción y descripción.

La predicción se refiere a menudo a minería de datos supervisada.

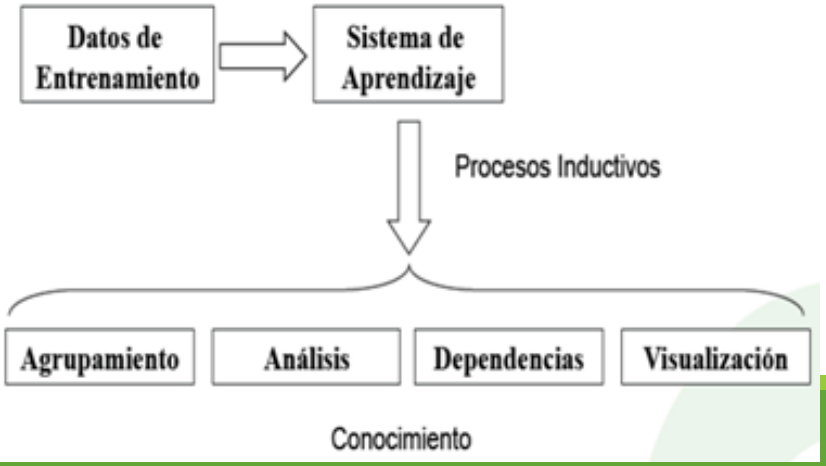
La descripción incluye aspectos no supervisados y visualización de la minería de datos.



# Paso 5. Elección de la tarea de minería de datos apropiada




**Aprendizaje no supervisado**




Método supervisado : se parte de un conocimiento previo de los datos.

Método no supervisado : se buscan automáticamente grupos de valores para que después el usuario intente encontrar las correspondencias entre esos grupos seleccionados automáticamente y las categorías que le puedan ser de interés.

La mayoría de técnicas de MD están basadas en aprendizaje inductivo, en donde un modelo se construye explícita o implícitamente por la generalización desde un número suficiente de ejemplos de entrenamiento. El supuesto subyacente del enfoque inductivo es que el modelo entrenado es aplicable a casos futuros.

# Paso 6. Elección de algoritmos de minería de datos

Técnicas de minería de datos

| Supervisados | No supervisados |
|---|---|
| Árboles de decisión | Detección de desviaciones |
| Inducción neuronal | Segmentación |
| Regresión | Agrupamiento |
| Series temporales | Reglas de asociación |
| | Patrones secuenciales |

Fuente: (Moreno *et al.*, 2001).

Teniendo la estrategia, ahora se decide la táctica. Esta etapa incluye la selección del método específico que se usará para buscar patrones. Por ejemplo:

Considerando la precisión es más conveniente el uso de redes neuronales.

Considerando entendimiento, es mejor con árboles de decisión.

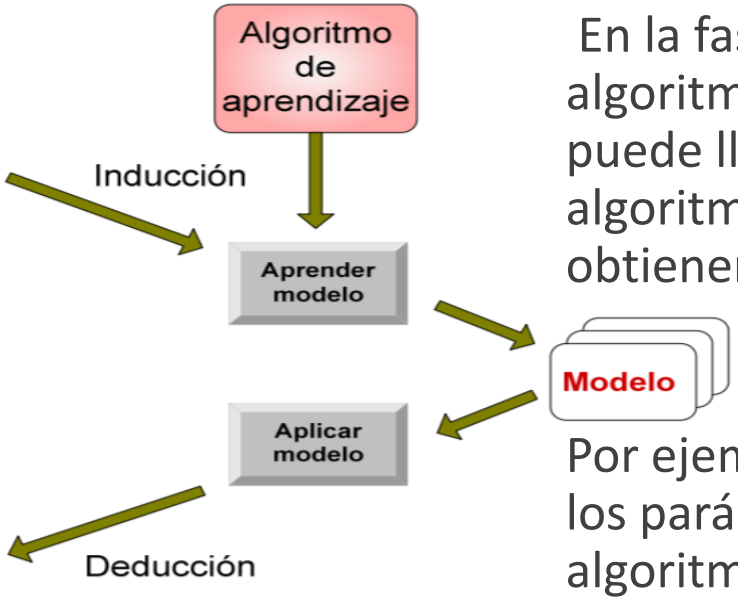
# Paso 7. Empleo del algoritmo de minería de datos

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Conjunto de entrenamiento

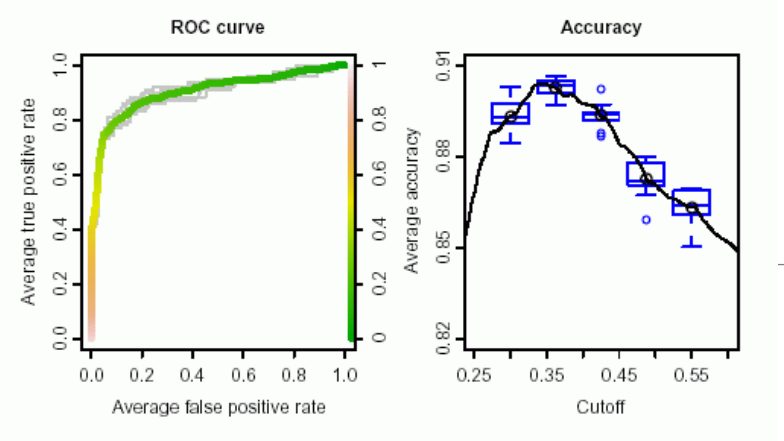
| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Conjunto de prueba




En la fase de implementación del algoritmo de minería de datos se puede llegar a necesitar emplear el algoritmo varias veces hasta que se obtienen resultados satisfactorios.

Por ejemplo, por la puesta a punto de los parámetros de control del algoritmo, tal como el número mínimo de instancias en una sola hoja de un árbol de decisiones.




# Paso 8. Evaluación

En esta etapa se evalúan e interpretan los patrones minados (reglas, confiabilidad, etc.), con respecto a los objetivos definidos en el primer paso.

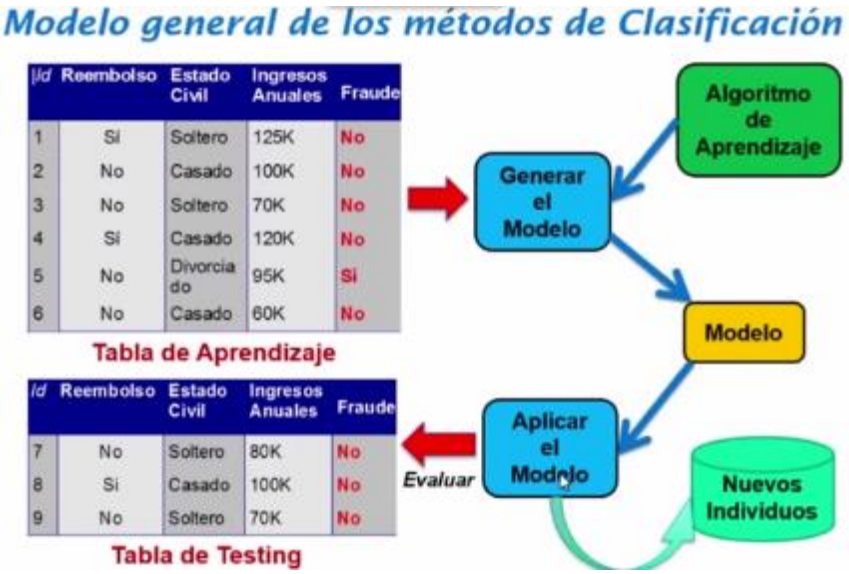
Aquí se consideran los pasos de pre-procesamiento con respecto a sus efectos en los resultados del algoritmo de minería (por ejemplo, adición de características en el paso de transformación, y repetir desde allí). Este paso se centra en la comprensibilidad y utilidad del modelo inducido. Durante la evaluación, el conocimiento descubierto es también documentado para su uso más adelante.

El último paso es el uso y retroalimentación general sobre los patrones y resultados de descubrimiento, obtenidos por la Minería de Datos.

# Paso 8. Evaluación

Medir, interpretar y en su caso graficar exactitud, precisión, recall, especifidad, f-measure

- Modificar condiciones de umbrales, corpus, etc, para generar mediciones que permitan mejor rendimiento




| ID | Type | Predicted |
|---|---|---|
| 1 | spam | spam |
| 2 | spam | spam |
| 3 | normal | normal |
| 4 | normal | spam |
| 5 | spam | spam |
| 6 | spam | spam |
| 7 | normal | spam |
| 8 | spam | normal |
| 9 | normal | normal |
| 10 | normal | normal |
| 11 | spam | spam |
| 12 | spam | spam |
| 13 | spam | normal |
| 14 | spam | normal |
| 15 | normal | normal |

| ID | Type | Predicted |
|---|---|---|
| 1 | spam | spam |
| 2 | spam | spam |
| 3 | normal | normal |
| 4 | normal | spam |
| 5 | spam | spam |
| 6 | spam | spam |
| 7 | normal | spam |
| 8 | spam | normal |
| 9 | normal | normal |
| 10 | normal | normal |
| 11 | spam | spam |
| 12 | spam | spam |
| 13 | spam | normal |
| 14 | spam | normal |
| 15 | normal | normal |

| pred.\true. | normal | spam |
|---|---|---|
| normal | TP | FP |
| spam | FN | TN |

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

# Paso 9. Uso del conocimiento descubierto

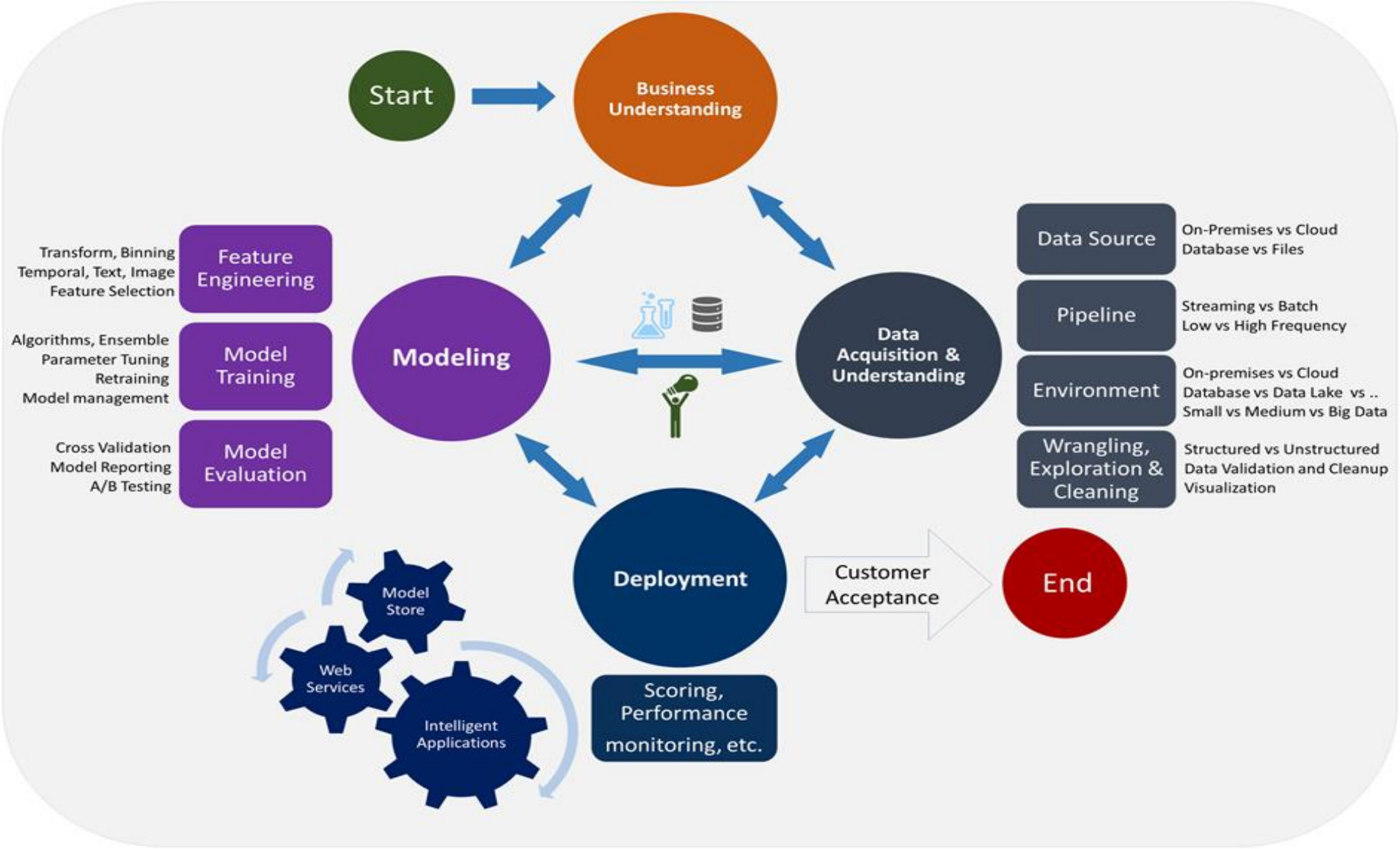

| Id | Sueldo | Casado | Vehículo | Hijos | Alq/Prop | Sindicato | Bajas/Año | Antigüedad | Sexo |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | Si | No | 0 | Alquiler | No | 7 | 15 | H |
| 2 | 2000 | No | Si | 1 | Alquiler | Si | 3 | 3 | M |
| 3 | 1500 | Si | Si | 2 | Propia | Si | 5 | 10 | H |
| 4 | 3000 | Si | Si | 1 | Alquiler | No | 15 | 7 | M |
| 5 | 4000 | Si | Si | 0 | Propia | Si | 1 | 6 | H |
| 6 | 2500 | No | No | 0 | Alquiler | Si | 3 | 16 | M |
| 7 | 2000 | No | Si | 0 | Alquiler | Si | 0 | 8 | H |
| 8 | 800 | No | Si | 0 | Propia | Si | 2 | 6 | M |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Se enfoca en la incorporación del conocimiento en otros sistemas para posteriores acciones.

El conocimiento llega a ser activo en el sentido que se pueden hacer cambios al sistema y medir los efectos. En realidad el éxito de este paso determina la efectividad del proceso entero de KDD.

Hay muchos retos en este paso, tales como pérdida de las condiciones de laboratorio bajo las cuales han operado. Por ejemplo, el conocimiento fue descubierto desde una cierta fotografía estática (usualmente una muestra) de los datos, pero ahora los datos son dinámicos. La estructura de los datos pueden cambiar (ciertos atributos pueden no estar disponibles), y se puede modificar el dominio de los datos (un atributo puede tener un valor que no fue asumido antes).

# Resumen: Ciclo de vida de Ciencia de Datos



# Disciplinas que usa la minería de datos para servir a la ciencia de datos

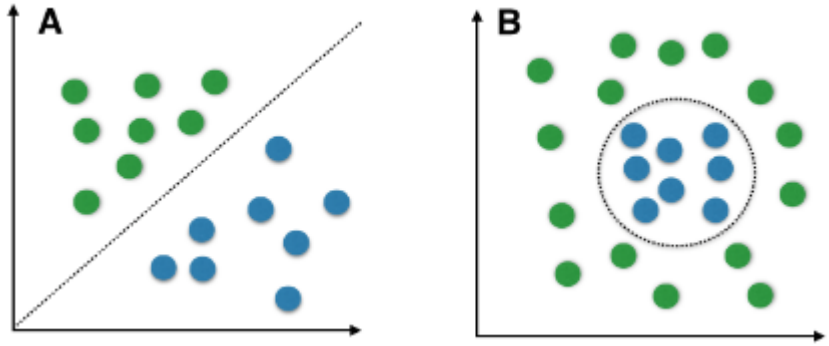
ESTADÍSTICA
DATOS Y ALGO MÁS

**Naïve Bayes Classifier (I)**

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$ MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(d \mid c)P(c)}{P(d)}$$ Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$ Dropping the denominator




Problemas lineales No-lineales


Machine Learning Algorithms

| | Unsupervised | Supervised |
|---|---|---|
| Continuous | • Clustering & Dimensionality Reduction<br>○ SVD<br>○ PCA<br>○ K-means | • Regression<br>○ Linear<br>○ Polynomial<br>• Decision Trees<br>• Random Forests |
| Categorical | • Association Analysis<br>○ Apriori<br>○ FP-Growth<br>• Hidden Markov Model | • Classification<br>○ KNN<br>○ Trees<br>○ Logistic Regression<br>○ Naive-Bayes<br>○ SVM |

**Estadística**
**Manejo de datos**
**Recuperación de información (textos, documentos,etc)**
**Big Data**
**Aprendizaje de máquina, etc.,etc.**

# Tareas y Técnicas de minería de datos

| | |
|---|---|
| Descripción | Técnicas estadísticas (media, moda, mediana, desviación estándar, mínimo , máximo, rango, correlaciones) y gráficas, algoritmos genéticos |
| Clasificación | Redes neuronales (back propagation), árboles de decisión (ID3, C4.5, C5.0, CART), k-nn (k vecinos más cercanos), naive bayes, técnicas estadísticas |
| Estimación | Técnicas estadísticas (regresión lineal simple, correlación, regresión múltiple), árboles de decisión, k-nn, redes neuronales |
| Predicción | Técnicas estadísticas, redes neuronales, árboles de decisión, k-nn, algoritmos genéticos |
| Agrupación (Clustering) | Jerárquico, K-nn, K-means, Red Kohonen, Fuzzy C-means |
| Asociación | Apriori (all, some, dynamic some), GRI, FP Grow |

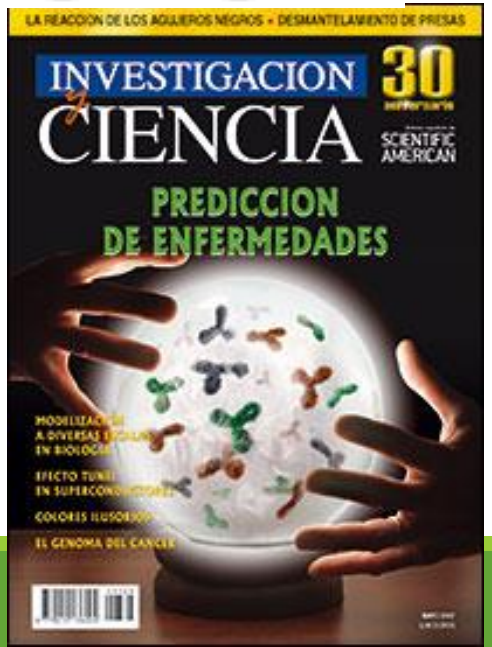
# Tareas y Técnicas de minería de datos

| Nombre | PREDICTIVO | | DESCRIPTIVO | | |
|---|---|---|---|---|---|
| | Clasificación | Regresión | Agrupamiento | Reglas de asociación | Correlaciones/ Factorizaciones |
| Redes neuronales | ✓ | ✓ | ✓ | | |
| Árboles de decisión ID3, C5.0 | ✓ | | | | |
| Árboles de decisiones CART | ✓ | ✓ | | | |
| Otros árboles de decisión | ✓ | ✓ | ✓ | ✓ | |
| Redes de Kohonen | | | ✓ | | |
| Regresión lineal y logarítmica | | ✓ | | | |
| Regresión logística | ✓ | | | ✓ | |
| Kmeans | | | ✓ | | |
| Apriori | | | | ✓ | |
| Naive Bayes | ✓ | | | | |
| Vecinos más próximos | ✓ | ✓ | ✓ | | |
| Análisis factorial y de componentes principales | | | | | ✓ |
| Twostep, Cobweb | | | ✓ | | |
| Algoritmos genéticos y evolutivos | ✓ | ✓ | ✓ | ✓ | ✓ |
| Máquinas de vectores de soporte | ✓ | ✓ | ✓ | | |
| CN2 rules (cobertura) | ✓ | | | ✓ | |
| Análisis discriminante multivariante | ✓ | | | | |

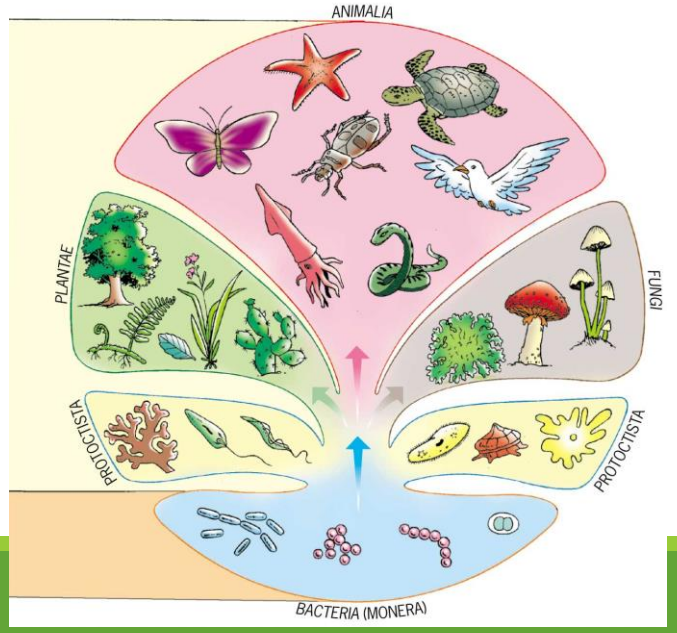
# Campos de aplicación











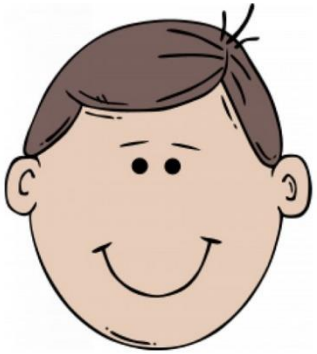
# Ofrecer productos/servicios lo más personalizados posible:

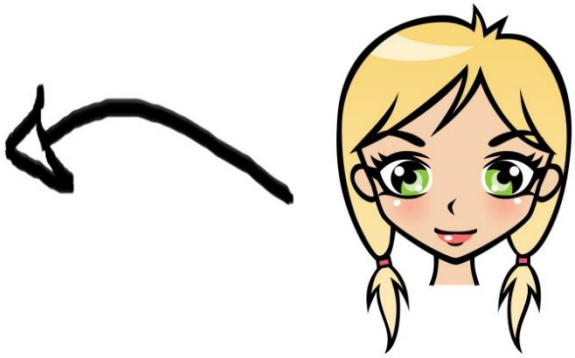
A partir del análisis del comportamiento de los usuarios, las empresas tendrán información necesaria para elaborar productos que respondan a necesidades específicas.

# Predecir comportamientos y establecer recomendaciones

La tecnología permite identificar los patrones de consumo de los usuarios con en base a la información que estos vuelcan a la red.

Es así como las compañías pueden ofrecer los productos indicados, en el momento oportuno.



# Ejemplos de aplicaciones

Descubrir patrones de comportamiento/consumo de los clientes:

- Fuga de clientes
- Autorizar créditos o transacciones bancarias (Análisis de riesgos)
- Detectar fraudes
- Campañas de mercadotecnia más efectivas de productos y/o servicios
- Recomendaciones

# Tan sencillo como cálculo de distancia y similitudes

Existen diversos cálculos de distancia:

Distancia Euclideana:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Distancia de Manhattan:

$$\sum_{i=1}^{n}|x_i - y_i|$$

Distancia de Chebychev:

$$max_{i=1..n}|x_i - y_i|$$

Valores Continuos (conveniente normalizar entre 0-1 antes)

Similitud coseno:

cada ejemplo es un vector y la distancia es el coseno del ángulo que forman

Valores Continuos. No es necesario normalizar

Si el ángulo entre los vectores es 0, el coseno es 1

Cualquier ángulo entre los vectores daría valor menor a 1

SI los vectores son ortogonales, el coseno se anula

Si tienen sentido contrario nos daría -1 entonces su rango cerrado es [-1,1]

Coseno de A producto punto B

||A||B||

# Recomendación por distancias

Dada la siguiente tabla de calificaciones de bandas musicales (spotify, amazon music, google music,etc.)

Se desea hacer recomendaciones a Héctor sobre bandas musicales que sea posible que le gusten y que no ha escuchado.

Se calcula la distancia de Héctor con clientes similares (suponga Verónica y Jorge solamente) y se toman las bandas de aquellos mas cercanos para recomendarlos a Héctor

| Conjunto Musical | Angelica | Benito | Claudia | Daniel | Hector | Jorge | Samuel | Verónica |
|---|---|---|---|---|---|---|---|---|
| Queen | 3.5 | 2 | 5 | 3 | | | 5 | 3 |
| The Beatles | 2 | 3.5 | 1 | 4 | 4 | 4.5 | 2 | |
| Guns´n´roses | | 4 | 1 | 4.5 | 1 | 4 | | |
| AC/DC | 4.5 | | 3 | | 4 | 5 | 3 | 5 |
| Nirvana | 5 | 2 | 5 | 3 | | 5 | 5 | 4 |
| Led Zeppelin | 1.5 | 3.5 | 1 | 4.5 | | 4.5 | 4 | 2.5 |
| Pink Floyd | 2.5 | | | 4 | 4 | 4 | 5 | 3 |
| The Rolling Stones | 2 | 3 | | 2 | 1 | 4 | | |

Ej.

Distancia euclidiana entre **Héctor y Verónica:**

$$= \sqrt{(4-5)^2 + (4-3)^2} = \sqrt{1+1} = \sqrt{2} = 1.414$$

Distancia euclidiana entre **Héctor y Jorge**

$$= \sqrt{(4-4.5)^2 + (1-4)^2 + (4-5)^2 + (4-4)^2 + (1-4)^2}$$

$$= \sqrt{(-0.5)^2 + (-3)^2 + (-1)^2 + (0)^2 + (-3)^2}$$

$$= \sqrt{.25+9+1+0+9} = \sqrt{19.25} = 4.387$$

**¿Qué bandas se le recomendarían a Héctor?**

**Las de Verónica que no haya escuchado o calificado Héctor**

**“Héctor te recomiendo escuchar Queen o Nirvana”**

# Análisis de canasta de mercado:

Si se obtiene cuales productos se compran juntos, quienes los compran y durante que temporada.

Distribución de los productos en anaqueles, tiendas, estados

- Control de inventario y oferta con base a la demanda
- Fidelización de los clientes

# Reglas de asociación

| | Leche | Cuernitos | Mayonesa | Café | Pasta | Puré | Tostadas |
|---|---|---|---|---|---|---|---|
| $T_1$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $T_2$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_3$ | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| $T_4$ | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| $T_5$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| $T_6$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $T_7$ | 1 | 1 | 0 | 1 | 0 | 0 | 0 |

Una regla de asociación es de la forma $\alpha \rightarrow \beta$, donde $\alpha$ y $\beta$ son dos conjuntos disjuntos de artículos, también puede expresarse: SI $\alpha$ ENTONCES $\beta$ **(Leche, cuernitos) → (café)**

**MEDIDAS DE LA CALIDAD DE LA REGLA:**

Sea el Soporte el número de instancias X que la regla predice correctamente con respecto al total D.

Sea la Confianza (precisión, confidence): El número de veces que la regla se cumple cuando puede aplicarse.

# Ejemplo de análisis de canasta de mercado

Sea la siguiente información sobre los tickets de compra de un supermercado. Se desea identificar la mejor distribución de los productos en la tienda con base al consumo. Consideremos solo leche, café y cuernitos.

| | Leche | Cuernitos | Mayonesa | Café | Pasta | Puré | Tostadas |
|---|---|---|---|---|---|---|---|
| $T_1$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $T_2$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_3$ | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| $T_4$ | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| $T_5$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| $T_6$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $T_7$ | 1 | 1 | 0 | 1 | 0 | 0 | 0 |

Sea A leche, cuernitos y B=café. Si consideramos la regla ($A \Rightarrow B$)

(Leche, cuernitos) →(café)

Obtener el soporte y la confianza

La regla se cumple 3 veces de un total de 7 transacciones. soporte ($A \Rightarrow B$)=3/7

La regla se cumple en un 43% de las transacciones.

El 43% de todas las transacciones de compra si se compra leche y cuernitos se compra café

# Ejemplo de análisis de canasta de mercado

Considerando la regla (Leche, cuernitos) →(café) ; obtener su confianza

| | Leche | Cuernitos | Mayonesa | Café | Pasta | Puré | Tostadas |
|---|---|---|---|---|---|---|---|
| $T_1$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| $T_2$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| $T_3$ | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| $T_4$ | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| $T_5$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| $T_6$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| $T_7$ | 1 | 1 | 0 | 1 | 0 | 0 | 0 |

De 4 transacciones en las que se compró leche y cuernitos (se cumple el antecedente), en 3 se compró café (se cumple la regla). Es decir, la regla se cumple en un 75% de las transacciones en las que podía aplicarse. O bien:

Confianza : soporte (A ∪ B ) / soporte(A) ; ¾

La regla (Leche, cuernitos) -> (café) Suporte 0.43 y una confianza de 0.75

Interpretación: Cuando el cliente compra leche y cuernitos el 75% de las veces compra café

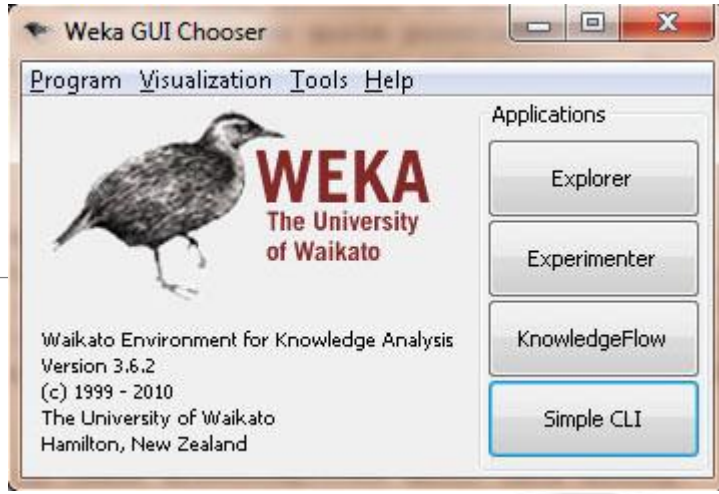
Decisión: El departamento de lácteos y panadería deben ubicarse cerca y en el camino a ellos, en los pasillos ofertar café

# P.e. Análisis de textos

Algoritmos que tokenicen corpus, modelos de n-gramas

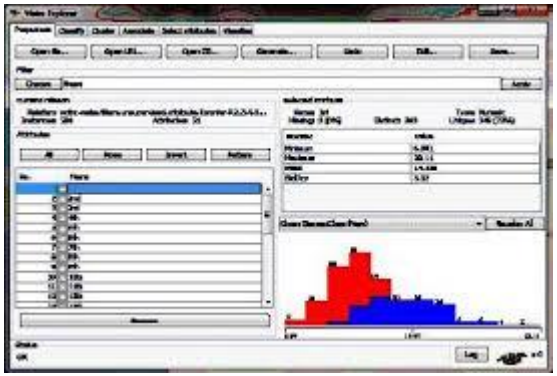
Evaluación de similitud de documentos

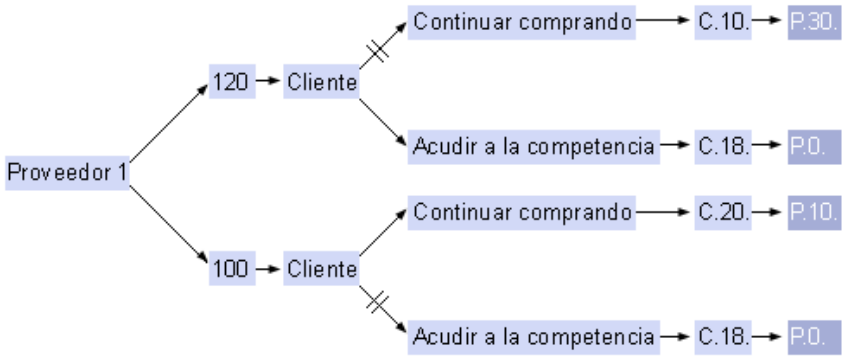
Modelos de predicción usando diferentes clasificadores como N-Bayes, Maquinas de soporte vectorial, arboles de decisión




Natural Language Tool Kit (NLTK)
Basic Text Analytics


| | Positive | Negative |
|---|---|---|
| Predicted Positive | 69 (98.6%) | 10 (14.3%) |
| Predicted Negative | 1 (1.43%) | 60 (85.7%) |






# Minería de textos

Considere un sistema de recuperación de información que contiene los siguientes documentos de dichos populares:

D1: Con vino añejo y pan de hoy se pasa el invierno
D2: Pan a hartura vino a mesura
D3: Pan de ayer y vino añejo, mantienen hombre sano
D4: Al pan pan y al vino vino

Si nosotros queremos saber en que documentos puedo encontrar si el hombre debe consumir pan y vino para estar sano.

**PASOS**

**1) Reducir las palabras a un formato de raíz común.**
**2) Eliminar preposiciones, verbos y determinantes**
**3) Elaborar la matriz correspondiente dentro de un modelo de Espacio Vectorial.**
**4) Reducir la consulta a un vector de términos**

La consulta en términos de la matriz: Q = (1,0,1,0,0,0,0,0,1,1)

| | vino | añejo | pan | hoy | invierno | hartura | Mesura | ayer | Hombre | sano |
|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| D2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| D3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| D4 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# Minería de textos

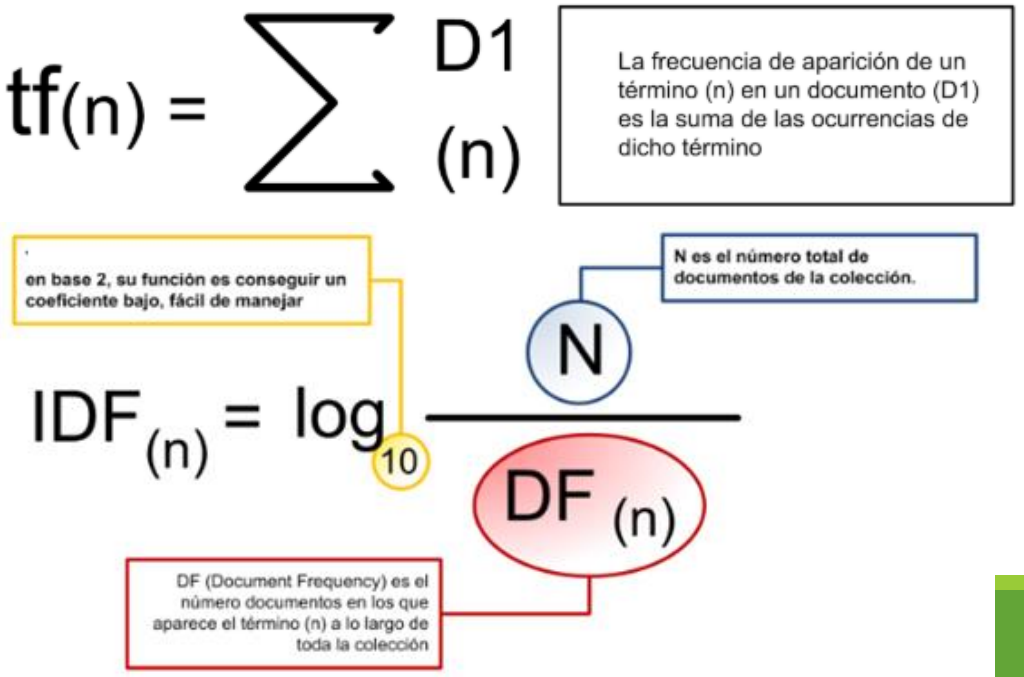
**5) Obtener la Matriz tf-idf de términos y documentos en el Espacio Vectorial con los pesos calculados y considerando la pregunta Q.**

La sig. matriz contiene el tf(n) y el DF(n)

| | vino | añejo | pan | hoy | invierno | hartura | Mesura | ayer | Hombre | sano |
|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| D2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| D3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| D4 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DF | 4 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Cálculo de frecuencias inversas

Idf (vino) = Log (4/4) = log (1) = 0
Idf (añejo) = Log (4/2) = log 2 = 0.301
Idf (pan) = Log (4/4) = log 1 = 0
Idf (hoy) = Log (4/1) = log 4 = 0.602
Idf (invierno) = Log (4/1) = log 4 = 0.602
Idf (hartura) = Log (4/1) = log 4 = 0.602
Idf (mesura) = Log (4/1) = log 4 = 0.602
Idf (ayer) = Log (4/1) = log 4 = 0.602
Idf (hombre) = Log (4/1) = log 4 = 0.602
Idf (sano) = Log (4/1) = log 4 = 0.602



Matriz TF*IDF; Recuerde que Q= (1,0,1,0,0,0,0,0,1,1)

| | vino | añejo | pan | hoy | invierno | hartura | Mesura | ayer | Hombre | sano |
|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 1*0 | 1*0.301 | 0 | 0.602 | 0.602 | 0 | 0 | 0 | 0 | 0 |
| D2 | 0 | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0 | 0 | 0 |
| D3 | 0 | 0.301 | 0 | 0 | 0 | 0 | 0 | 0.602 | 0.602 | 0.602 |
| D4 | 2*0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q | 1*0 | 0 | 1*0 | 0 | 0 | 0 | 0 | 0 | 1*0.602 | 1*0.602 |

**6) Calcular similitud por producto escalar entre el vector pregunta Q y los vectores de los documentos.** Hay que multiplicar componente a componente de los vectores y sumar los resultados. El modo más sencillo de obtener la similitud es por medio del producto escalar de los vectores (es decir, multiplicando los componentes de cada vector y sumando los resultados).

| | vino | añejo | pan | hoy | invierno | hartura | Mesura | ayer | Hombre | sano | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1,Q | 0*0 | 0.301*0 | 0*0 | 0.602*0 | 0.602*0 | 0*0 | 0*0 | 0*0 | 0*0.602 | 0*0.602 | 0 |
| D2,Q | 0 | 0 | 0 | 0 | 0 | 0.602*0 | 0.602*0 | 0 | 0 | 0 | 0 |
| D3,Q | 0 | 0.301*0 | 0 | 0 | 0 | 0 | 0 | 0.602 | 0.602*0.602 | 0.602*.602 | =0.724 |
| D4,Q | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**7) Ordene la respuesta en función de los resultados de similitud obtenidos.**

RESPUESTA CORRECTA D3

Con estos valores de similitud, se obtiene que la pregunta “¿Cómo el hombre debe consumir pan y vino para estar sano? Se responde con el documento D3 que dice Pan de ayer y vino añejo, mantienen hombre sano

# Uso en sector salud

En la actualidad, la investigación farmacéutica y los diversos dispositivos utilizados en hospitales, tales como marcapasos y equipos sofisticados de diagnóstico de salud, incrementan los volúmenes de información a un ritmo vertiginoso.

Dichos datos pueden ser aprovechados para avanzar en el conocimiento, detección temprana y curación de enfermedades.

# Uso en sector salud

Predecir diagnósticos

Prevenir y pronosticar afecciones.

Mejorar la eficacia de los medicamentos

Mejora en la administración de medicamentos intrahospitalarios

Anticiparse a los problemas de salud  en los centros hospitalarios

# Predecir diagnósticos
# Prevenir y pronosticar afecciones

A través de la implementación de soluciones de inteligencia de negocios (BI) y big data, el enorme caudal de datos generado día a día por los pacientes es capitalizado por organizaciones de salud, hospitales y centros de investigación a nivel mundial.

La información generada –sea estructurada (como resultados de análisis) o no estructurada (como imágenes médicas)– puede ser digitalizada y almacenada para su posterior estudio.

Compartir y analizar registros de historias clínicas, por medio de la aplicación de tecnologías disruptivas, permite predecir diagnósticos, así como prevenir y pronosticar afecciones.

# Naive Bayes

La clasificación bayesiana es un método basado en estadísticos. Su funcionamiento usa el cálculo de probabilidades a partir del teorema de Bayes.

P(Ai) es la probabilidad a priori de la hipótesis Ai.

P(B) es la probabilidad de observar el conjunto de entrenamiento B, cuando es usado para clasificar. Como es la total, P(B) = P(A ^ B) ; P(A)P(B)

P(B|Ai) es la probabilidad de observar el conjunto de entrenamiento B en un universo donde se verifica la hipótesis Ai.

P(Ai|B) es la probabilidad a posteriori de Ai, cuando se ha observado el conjunto de entrenamiento B.

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Donde:

$P(A_i)$ = Probabilidad a priori

$P(B/A_i)$ = Probabilidad condicional

$P(B)$ = Probabilidad Total

$P(A_i/B)$ = Probabilidad a posteriori

# Determinar si un paciente tiene cáncer por Naive Bayes

Sabemos que sólo el 0.8% de las personas en los Estados Unidos tienen esta forma de cáncer.

La prueba presente un resultado binario, es decir da POS o NEG.

Cuando la enfermedad está presente la prueba devuelve un resultado correcto POS en el 98% de las veces, o devuelve un resultado correcto NEG el 97% del tiempo en los casos cuando la enfermedad no está presente.

Hipótesis: El paciente tiene cáncer o bien, El paciente no tiene ningún tipo de cáncer.

# Calculo de probabilidades

- Sabemos que sólo el 0,8% de la población en los EE.UU. tienen esta forma de cáncer. $P(cancer) = 0.008$
- El 99,2% de las personas no tienen este cáncer. $P(\neg cancer) = 0.992$
- Cuando la enfermedad está presente la prueba devuelve un resultado correcto POS el 98% del tiempo. $P(POS|cancer) = 0.98$
- Cuando la enfermedad está presente la prueba devuelve un resultado incorrecto NEG un 2% del tiempo. $P(NEG|cancer) = 0.02$
- La prueba devuelve un resultado correcto NEG el 97% del tiempo en los casos cuando la enfermedad no está presente. $P(NEG|\neg cancer) = 0.97$
- Devuelve un resultado incorrecto POS el 3% del tiempo en los casos cuando la enfermedad no está presente $P(POS|\neg cancer) = 0.03$

Suponga que Rita fue al doctor, le hicieron una prueba de sangre para saber si tenía cáncer. Y ésta arrojo un resultado Positivo (POS).

Dado que la prueba es 98% efectiva. Use el Teorema de Bayes para determinar que tan probable es que Rita tenga cáncer.

*Recuerde que:*

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

$P(cancer) = 0.008$

$P(\neg cancer) = 0.992$

$P(POS|cancer) = 0.98$

$P(POS|\neg cancer) = 0.03$

$P(NEG|cancer) = 0.02$

$P(NEG|\neg cancer) = 0.97$

$$P(\text{cáncer} \mid POS) = \frac{P(cancer)\, P(POS \mid cancer)}{P(cancer)\, P(POS \mid cancer) + P(\neg\, cancer)\, P(POS \mid \neg\, cancer)}$$

# Buscando la máxima probabilidad a posteriori:

P(*cancer*) *P*(*POS* | *cancer*) = (.008) .98= .0078

P(*POS* | ¬ *cancer*) P(¬ *cancer*) = .03(.992) = .0298

Si queremos saber la probabilidad exacta, podemos normalizar estos valores haciéndolos que sumen 1:

$$P(\text{cáncer} \mid POS) = \frac{P(cancer)\,P(POS \mid cancer)}{P(cancer)\,P(POS \mid cancer) + P(\neg cancer)\,P(POS \mid \neg cancer)}$$

$$P(cancer \mid POS) = \frac{0.0078}{0.0078 + 0.0298} ; = 0.21$$

Rita tiene un 21% de probabilidad de tener cáncer.

# Anticiparse a los problemas de salud en los centros hospitalarios

Por ejemplo, análisis de imágenes médicas

- Naive Bayes
- Redes neuronales

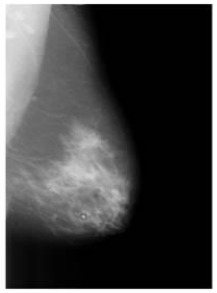
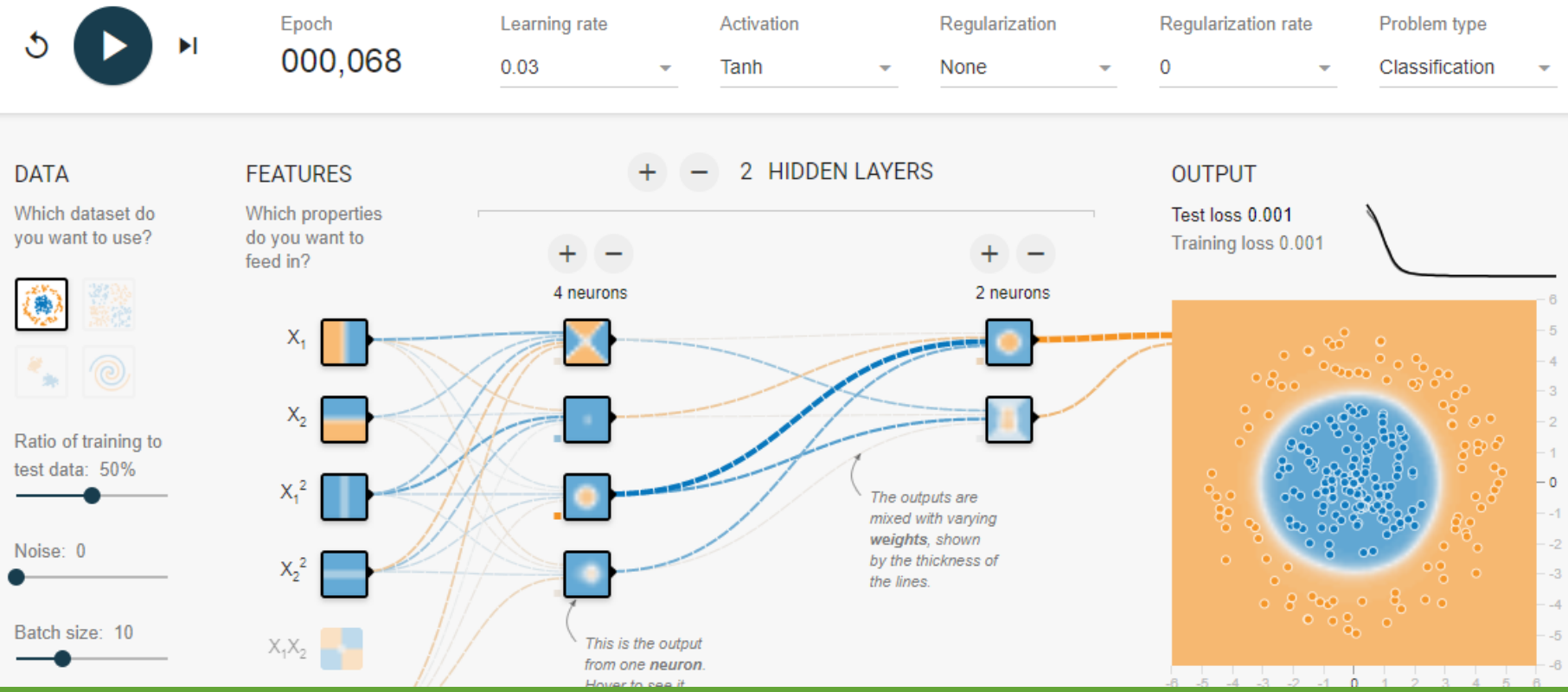


**Figura 3.1.** Tumor benigno, calcificación.

**Figura 3.2.** Zoom de 3.1 en la región comprometida. Puede observarse la resolución de DDSM.



# Uso en industria de mercadotecnia

Mejorar la relación con los clientes es una de las principales razones que motivan la implementación la Ciencia de datos, ya sea para realizar acciones de fidelización o para identificar nuevos potenciales. Por eso, estas soluciones se han convertido en una herramienta fundamental para ayudar a los profesionales de marketing a lograr los objetivos de negocio.

# Uso en mercadotecnia


Las campañas basadas en el comportamiento de clientes y reconocimiento de emociones en videos permiten:

- Predecir comportamientos
- Segmentación de mercado
- Ofrecer productos/servicios lo más personalizados posible


# Segmentación de mercado

Permite elaborar campañas dirigidas a un público objetivo mucho más definido que en el pasado. A partir del análisis, los profesionales de marketing pueden saber quién compra qué producto, cuándo y dónde, y así orientar sus esfuerzos en base a eso.

- Campañas de mercadotecnia dirigidas al segmento correcto
- Calculo de precio de acuerdo al segmento

# Estrategia de mercadotecnia directa por segmentación

Paso 1: Cargar y preparar datos de campañas de mercadotecnia anteriores, incluyendo atributos personales (o propios del objeto de estudio, por ejemplo, edad, sexo, área) y atributos de comportamiento (uso de productos y servicios, sitio web, etc.).

Paso 2: Determinar qué factores influyen en la respuesta a las campañas de mercadotecnia, establecer, normalizar y ordenar pesos

Paso 3: Entrenar y validar el modelo de respuesta del cliente. Por ejemplo, por validación cruzada en paralelo, que realice diez intentos, que genere muestreo estratificado. Al entrenar el modelo que sea con Naive Bayes y corrección de Laplace. En pruebas aplicar el modelo. Determinar rendimiento por clasificación binomial, medir precision, etc.

# Estrategia de mercadotecnia directa por segmentación

Paso 4: Cargar datos que contengan destinatarios potenciales para nuevas campañas. Aplicar el modelo de respuesta del cliente para identificar y orientar a los destinatarios que son los más propensos a responder a la campaña de mercadotecnia de la manera deseada.

Paso 5: Experiencia: Normalmente, omitir destinatarios que hubieran respondido, incurre en un costo mayor que el envío de una campaña a alguien que no responde. Contabilizar esos costos, calcular y aplicar el umbral de confianza óptimo.

Salidas: factores de influencia, clientes calificados con probabilidad de responder.

Interpretación: Explicar cuantos y cuales clientes son mas factibles de responder ante nuevas campañas.

# Plan de estudios de la Licenciatura en Ciencia de Datos

## Modalidad presencial

## Título que se otorga:
## Licenciado(a) en Ciencia de Datos

# Plan de estudios

## Objetivo general

La Licenciatura en Ciencia de Datos formará profesionales capaces de seleccionar, extraer, preparar, analizar, evaluar y comunicar cantidades masivas de datos de cualquier tipo de manera ética y responsable para la toma de decisiones inteligentes y la resolución de problemas complejos en los sectores científicos, tecnológicos, empresariales y sociales.

# Objetivos particulares

1. Interpretar datos estructurados y no estructurados para resolver problemas complejos.
2. Construir técnicas que permitan visualizar de forma eficaz la información obtenida de los análisis de los datos.
3. Elaborar modelos matemáticos que permitan entender mejor los problemas que se presenten en diversos fenómenos, como, por ejemplo: físicos, biológicos o sociales.
4. Utilizar sus conocimientos científicos y tecnológicos con ética y responsabilidad para contribuir al uso eficiente y responsable de los recursos naturales, humanos y financieros.

# Perfil de ingreso

**Conocimientos:** Matemáticas básicas, Computación, Inglés a nivel de comprensión oral y escrita (A2), Redacción

**Habilidades:** Capacidad de observación, abstracción, búsqueda, análisis y síntesis de la información.

**Actitudes:** Disposición a la mejora continua en la elaboración de sus trabajos., Curiosidad, Flexibilidad y adaptabilidad

**Intereses:** Análisis matemático, estadístico, económico administrativo y computacional. Interés para realizar actividades experimentales y de investigación.

**Valores:** Responsabilidad, Proactividad, Tolerancia hacia otras opiniones

# Perfil de egreso

**Conocimientos**: Construir modelos matemáticos que permitan entender mejor problemas de diversas áreas, físicos, biológicos o sociales. Diseñar soluciones de infraestructura de tecnologías para la información.

**Habilidades**: Pensar de manera crítica y matemática, resolver problemas abstractos, visualizar y comunicar hallazgos analíticos, programar en lenguajes computacionales, manipular grandes conjuntos de datos, comunicarse con personas de distintas disciplinas.

**Actitudes**: Uso eficiente y responsable de los recursos naturales, humanos y financieros. Proponer soluciones con responsabilidad, honestidad y alto sentido ético.

**Valores**: Actuar con responsabilidad, honestidad, justicia y sentido ético en su ejercicio profesional

# Áreas de trabajo

**Inteligencia de negocios y gobierno de datos**: inteligencia de negocio para la toma de decisiones

**Diseño de sistemas de visualización de datos:** análisis, diseño y desarrollo de algoritmos de visualización para la correcta interpretación y manejo de información.

**Diseño de software**: estadístico, de negocio, exploratorio de datos, de inteligencia artificial, aprendizaje de máquina.

**Investigación y desarrollo tecnológico:** proyectos de investigación para la generación de nuevo conocimiento y nuevas tecnologías.

**Manejo de grandes bases de datos:** administración e integración de diversas fuentes de datos.

**Construcción de modelos estadísticos predictivos:** acorde a los requerimientos de las empresas o instituciones.

# Carrera de ingreso indirecto

## Origen

- Actuaría, Física, Ciencias de la Computación, Ingeniería en Computación, Matemáticas, Matemáticas Aplicadas, Matemáticas Aplicadas y Computación
- Facultad de Ciencias
- Facultad de Ingeniería
- FES Aragón
- FES Acatlán

## Conocimientos comunes en los primeros 4 semestres

- Álgebra y cálculo diferencial e integral
- Algoritmos, programación básica
- Probabilidad y estadística

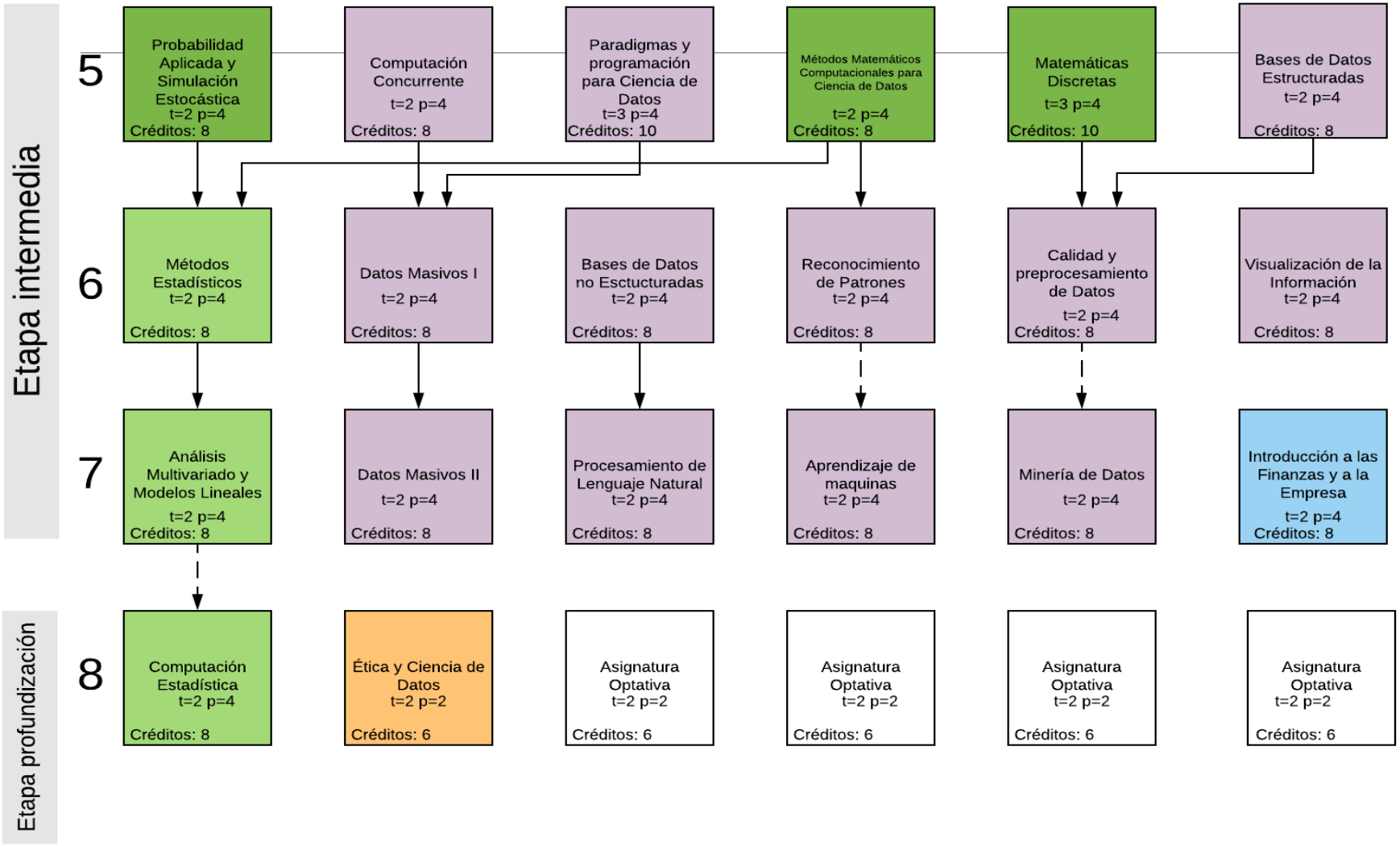
## Carácter interdisciplinario

- Perfil de los aspirantes provenientes de diversas carreras
- El carácter propio de la Ciencia de Datos
- Plan de estudios
- Entidades responsable y participantes
- Áreas de aplicación

# Tiempos y créditos

Créditos de carrera de origen en un rango de **170 a 232**.
Del quinto al octavo semestre un total de **24 asignaturas, 20 obligatorias y 4 optativas, constituyendo un total de créditos de 186 créditos.**
Total de créditos en un rango de **356** a **418** créditos, dependiendo de carrera de origen:

| | |
|---|---|
| Actuaría-Facultad de Ciencias: | **412** |
| Actuaría-FES Acatlán: | **418** |
| Ciencias de la Computación-Facultad de Ciencias: | **374** |
| Física-Facultad de Ciencias: | **374** |
| Ingeniería en Computación-Facultad de Ingeniería: | **364** |
| Ingeniería en Computación-FES Aragón: | **356** |
| Matemáticas- Facultad de Ciencias: | **378** |
| Matemáticas Aplicadas-Facultad de Ciencias: | **410** |
| Matemáticas Aplicadas y Computación-FES Acatlán: | **406** |


Etapa intermedia
Etapa profundización
5
6
7
8
Probabilidad Aplicada y Simulación Estocástica t=2 p=4 Créditos: 8
Computación Concurrente t=2 p=4 Créditos: 8
Paradigmas y programación para Ciencia de Datos t=3 p=4 Créditos: 10
Métodos Matemáticos Computacionales para Ciencia de Datos t=2 p=4 Créditos: 8
Matemáticas Discretas t=3 p=4 Créditos: 10
Bases de Datos Estructuradas t=2 p=4 Créditos: 8
Métodos Estadísticos t=2 p=4 Créditos: 8
Datos Masivos I t=2 p=4 Créditos: 8
Bases de Datos no Esctucturadas t=2 p=4 Créditos: 8
Reconocimiento de Patrones t=2 p=4 Créditos: 8
Calidad y preprocesamiento de Datos t=2 p=4 Créditos: 8
Visualización de la Información t=2 p=4 Créditos: 8
Análisis Multivariado y Modelos Lineales t=2 p=4 Créditos: 8
Datos Masivos II t=2 p=4 Créditos: 8
Procesamiento de Lenguaje Natural t=2 p=4 Créditos: 8
Aprendizaje de maquinas t=2 p=4 Créditos: 8
Minería de Datos t=2 p=4 Créditos: 8
Introducción a las Finanzas y a la Empresa t=2 p=4 Créditos: 8
Computación Estadística t=2 p=4 Créditos: 8
Ética y Ciencia de Datos t=2 p=2 Créditos: 6
Asignatura Optativa t=2 p=2 Créditos: 6
Asignatura Optativa t=2 p=2 Créditos: 6
Asignatura Optativa t=2 p=2 Créditos: 6
Asignatura Optativa t=2 p=2 Créditos: 6

Asignaturas de la etapa de profundización:

| Campo de Profundización | Asignaturas |
|---|---|
| Algoritmos computacionales y Sistemas de Información | Aprendizaje de Máquina y Minería de Datos Avanzados<br>Temas selectos de Sistemas de Información<br>Seguridad de la Información |
| Estadística | Series de tiempo<br>Temas Selectos de Estadística |
| Investigación Científica | Introducción a la Investigación Científica<br>Temas selectos de Visualización computacional<br>Temas selectos de Ciencia de Datos |
| Procesamiento de Lenguaje Natural | Minería de textos<br>Temas Selectos de Procesamiento de Lenguaje Natural<br>Temas Selectos de Visión Computacional |
| Tópicos especiales | Temas selectos de tópicos especiales |

| Campo de aplicación | Asignaturas |
|---|---|
| Biología | Bioinformática<br>Ciencia de Datos en Biología |
| Ciencia Social | Ciencia Social Computacional |
| Finanzas Corporativas | Estrategias de portafolios de inversión utilizando Ciencia de Datos<br><br>Temas selectos de Finanzas Corporativas |
| Mercadotecnia | Temas selectos de Ciencia de Datos en Mercadotecnia |
| Tópicos especiales | Temas selectos de Ciencia de Datos en Área Diversa |

El potencial de la ciencia de datos es muy grande, diversas disciplinas siguen en constante inovación de sus algoritmos  al servicio de la minería de datos y diversas tecnologías se siguen proponiendo para el manejo de grandes cantidades de datos  semi-estructurados y no estructurados.

Por otro lado, el internet de las cosas permite la generación de grandes cantidades de datos, imágenes o textos.

La inteligencia de negocios continua buscando nuevas estrategias para el modelado y mejora del negocio.

Los requerimientos de tipo y velocidad de análisis, la cantidad y tipo de datos, asi como el nivel de experiencia de negocio hacen de la Ciencia de datos un reto constante.

Gracias!

pilarang@unam.mx