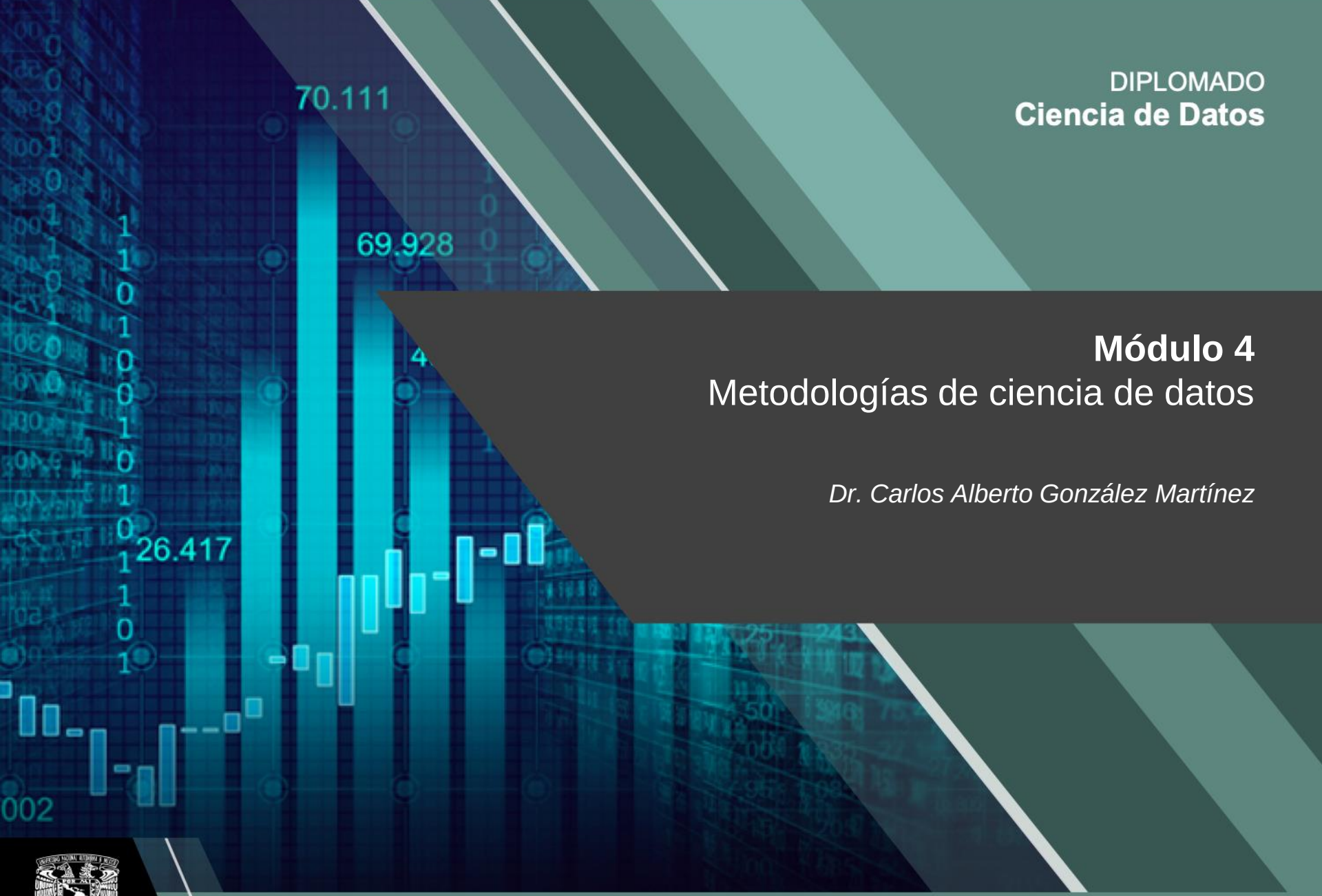


DIPLOMADO
Ciencia de Datos

Módulo 4
Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

KDD

Objetivo

El participante identificará y reconocerá el proceso de extracción de conocimiento, a partir de bases de datos para su análisis y aportación a la ciencia.

KDD

Contenido

1. Fases del proceso de KDD
2. Etapas de preprocesamiento
3. La fase de minería de los datos

KDD

A finales de los ochenta apareció un nuevo campo de investigación llamado **KDD** (Knowledge Discovery in Databases)

KDD es el proceso no trivial de identificar patrones, a partir de los datos con las siguientes características:

- Válidos
- Novedosos
- Potencialmente útiles
- Comprensibles

La revolución digital ha permitido que la captura de datos sea fácil, y su almacenamiento tenga un costo casi nulo.

KDD

El Descubrimiento de conocimiento en bases de datos (**KDD**, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático, en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice.

Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (*data mining*) y presentar resultados.

Autores como Fayyad, Piatetsky-Shapiro y Smith (1996, p. 89), lo definen como “El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos”.

KDD

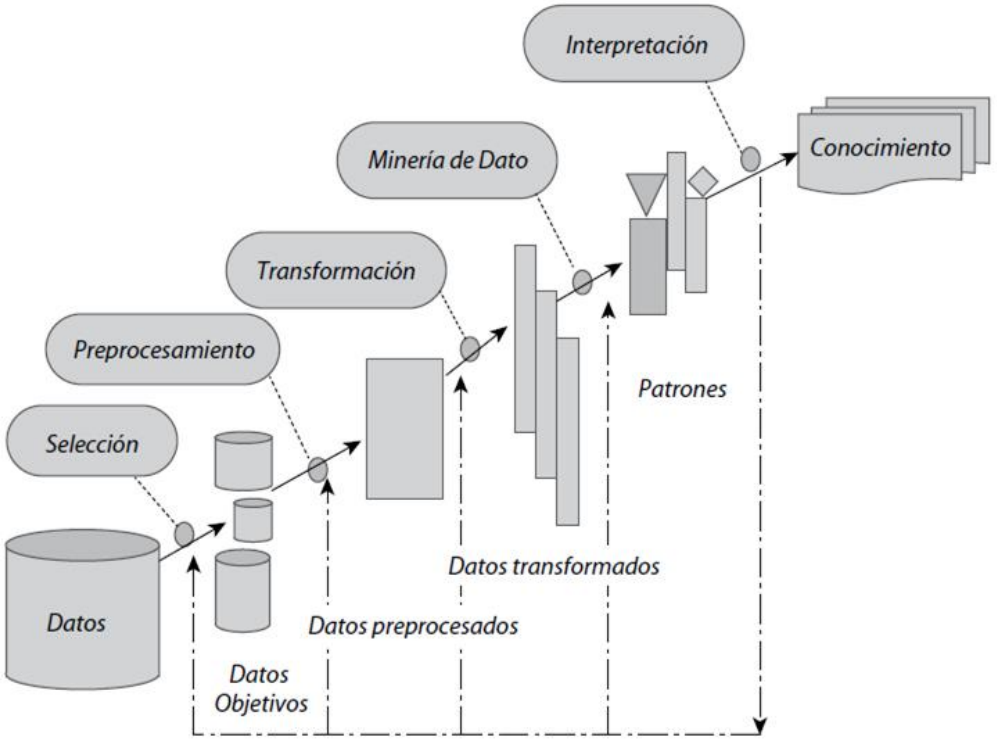
Contenido

1. **Fases del proceso de KDD**
2. Etapas de preprocesamiento
3. La fase de minería de los datos

KDD

1. Fases del proceso de KDD

Atendiendo a la visión que fija la minería de datos como parte del KDD, este proceso constaría de las fases que se observan en el siguiente esquema:



KDD

1. Fases del proceso de KDD

Datos (Localización y extracción)

Es necesario localizar y conocer los datos con los que se tratará el problema. Es necesario saber dónde están los datos, el significado de los mismos, cómo extraerlos y el volumen aproximado.

Los orígenes pueden ser muy variados: provenir internamente de la organización, de fuentes de datos públicas o incluso de proveedores de información externa.

Cómo extraerlos también va a influir en una productivización posterior. No es lo mismo que tengamos que acceder a ellos vía API a que se almacenen *on-premise* en nuestra empresa.

KDD

1. Fases del proceso de KDD

Etapa de selección

En esta etapa se requiere previamente conocer las fuentes de información más importantes y quiénes tienen control y acceso sobre ellas. También es relevante considerar las diferentes fuentes de datos, volumetría y formatos utilizados.

En la etapa de selección, una vez identificado el conocimiento relevante y prioritario, y definidas las metas del proceso kdd, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de éste, sobre el cual se realiza el proceso de descubrimiento.

La selección de los datos varía de acuerdo con los objetivos del negocio.

KDD

1. Fases del proceso de KDD

Etapa de preprocesamiento/limpieza

En la etapa de preprocesamiento/limpieza (*data cleaning*) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos y se seleccionan estrategias para el manejo de datos desconocidos (*missing* y *empty*), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.

En esta etapa, es de suma importancia la interacción con el usuario o analista.

En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos.

KDD

1. Fases del proceso de KDD

Etapa de transformación/reducción

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos, dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación, para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos (Fayyad *et al*., 1996).

La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos por la discretización de valores continuos (por ejemplo, edad por un rango de edades).

KDD

1. Fases del proceso de KDD

Etapa de transformación/reducción

La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento). Se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía y muestreo, entre otras (Han y Kamber, 2001).

Para la eliminación vertical se puede apoyar en la correlación entre variables.

KDD

1. Fases del proceso de KDD

Etapa de minería de datos

El objetivo de la etapa minería de datos es la búsqueda y el descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación (Quinlan, 1986), (Wang, Iyer y Scott, 1998), clustering (Ng y Han, 1994), (Zhang, Ramakrishnan, Livny, 1996), patrones secuenciales (Agrawal y Srikant, 1995) y asociaciones (Agrawal y Srikant, 1994), (Srikant y Agrawal, 1996), entre otras.

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos.

KDD

1. Fases del proceso de KDD

Etapa de minería de datos

Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas.

Entre las tareas predictivas están la clasificación y la regresión.

Los modelos descriptivos identifican patrones que explican o resumen los datos. Sirven para explorar las propiedades de los datos examinados.

Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los *clustering* y las correlaciones.

KDD

1. Fases del proceso de KDD

Etapa de interpretación/evaluación de datos

En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones.

Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles, en términos que sean entendibles para el usuario.

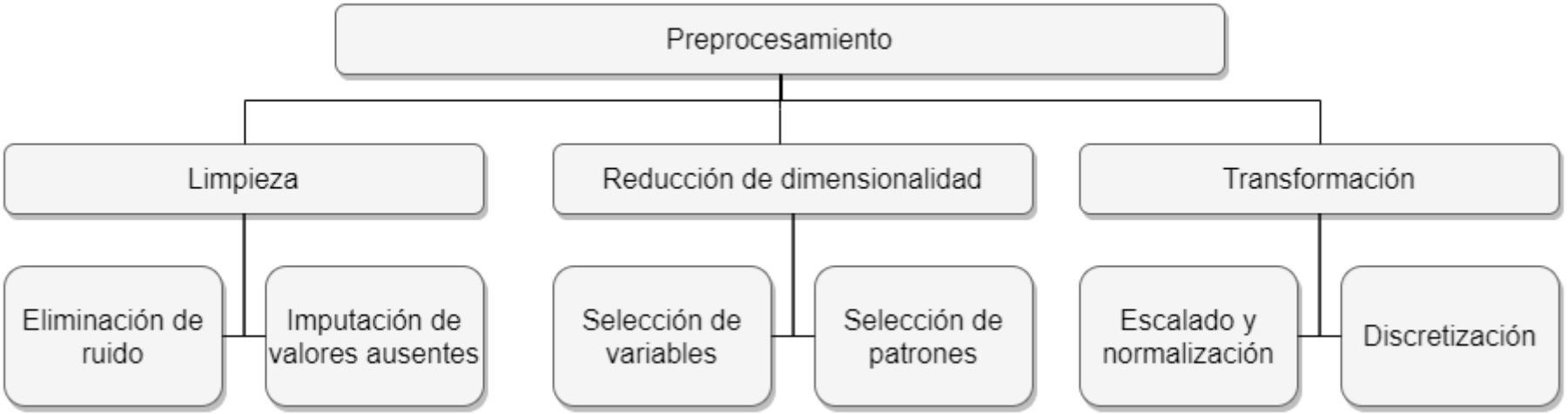
KDD

Contenido

1. Fases del proceso de KDD
2. **Etapas de preprocesamiento**
3. La fase de minería de los datos

KDD

2. Etapas de preprocesamiento



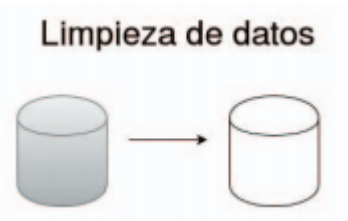
KDD

Etapa de preprocesamiento/limpieza

Esta etapa considera:

Limpieza

- Limpieza y cribado
- Acciones ante datos anómalos (outliers)
- Acciones ante datos faltantes (missing values)

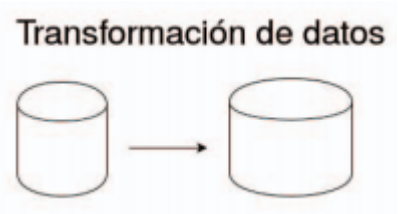


Reducción de dimensionalidad

- Selección de variables
- Selección de patrones

Transformación

- Escalado
- Discretización



KDD

Etapa de preprocesamiento/limpieza

- **Limpieza y cribado**

Limpieza (data cleansing) y ***criba*** (selección) de datos:

Se debe eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba). Se utilizan métodos estadísticos casi exclusivamente.

- Histogramas (detección de datos anómalos).
- Selección de datos (muestreo, ya sea verticalmente, eliminando atributos u horizontalmente, eliminando tuplas).
- Redefinición de atributos (agrupación o separación).

KDD

Etapa de preprocesamiento/limpieza

- **Acciones ante datos anómalos (outliers)**

***Ignorar*:** Algunos algoritmos son robustos a datos anómalos (por ejemplo: árboles)

***Filtrar* (eliminar o reemplazar) la columna:** Solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna es reemplazarla por una columna discreta, diciendo si el valor era normal o outlier (por encima o por debajo).

***Filtrar la fila*:** Claramente sesga los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.

KDD

Etapa de preprocesamiento/limpieza

- **Acciones ante datos anómalos (outlíers) (Cont):**

***Reemplazar el valor*:** Por el valor 'nulo' si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier, o por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.

***Discretizar*:** Transformar un valor continuo en uno discreto (por ejemplo: muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

KDD

Etapa de preprocesamiento/limpieza

- **Acciones ante datos faltantes (missing values)**

***Ignorar*:** Algunos algoritmos son robustos a datos faltantes (por ejemplo: árboles).

***Filtrar* (eliminar o reemplazar) la columna:** Solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad. Preferible a eliminar la columna, es reemplazarla por una columna booleana diciendo si el valor existía o no.

Reemplazar el valor por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.

Modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

KDD

Etapa de preprocesamiento/Reducción de dimensionalidad

- **Selección de variables y patrones**

No todos los datos obtenidos de las fuentes originales son necesariamente relevantes para el KDD, por lo que la selección de aquellos que realmente son útiles es otro paso más en el preprocesamiento.

Dos de las tareas más usuales en esta fase son la **selección de variables y la selección de patrones.** Esencialmente, consisten en eliminar aquellos datos que, por estar repetidos o pueden estimarse a partir de otros, no aportan mejora en la extracción de conocimiento. Estas dos técnicas forman parte de los métodos de **reducción de dimensionalidad.**

KDD

Etapa de preprocesamiento/Reducción de dimensionalidad

- **Selección de variables y patrones**

La selección de variables la determina el o los objetivos del proyecto.

Una vez que se tiene en claro cuáles son las variables independientes a utilizar, se confirma que los datos correspondientes ya se encuentren limpios y se validan por medio de la correlación de Pearson, que dichas variables sean independientes.

Las variables dependientes son las que son afectadas por las independientes y tienen un valor superior a 0.6 de correlación con las variables independientes.

KDD

Etapa de preprocesamiento/Transformación

- **Escalado y discretización**

Transformación de los datos: una vez que los datos están limpios y no contienen redundancias, aspectos de los que se ocupan las operaciones previas, podría pensarse que ya pueden usarse para el aprendizaje de un modelo. No obstante, hay acciones que podrían mejorar los datos de cara a hacer más efectivo ese aprendizaje.

Entre ellos están la **normalización, el escalado y la discretización.** Se trata de operaciones que transforman los datos originales, casi siempre de manera reversible, produciendo una nueva versión más conveniente para el análisis KDD.

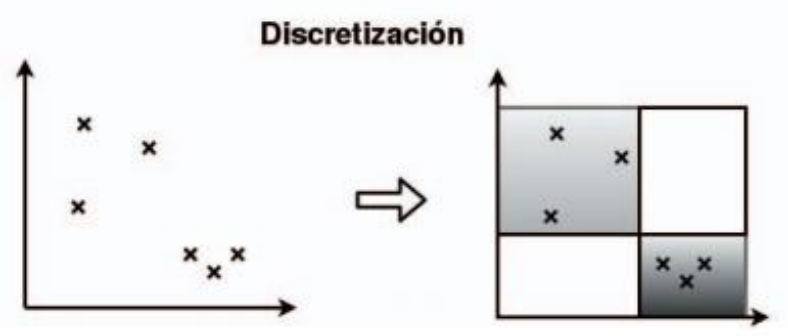
KDD

Etapa de preprocesamiento/Transformación

- **Discretización**

La discretización de datos es una técnica utilizada para pasar de un dato continuo a uno discreto. Esto favorece la ejecución de algunos algoritmos, dependiendo del lenguaje en que se utilice.

La discretización de datos es una técnica utilizada para la aplicación de muchos algoritmos de aprendizaje automático. En las bases de datos, cuando se tienen atributos continuos, puede complicar el aprendizaje. Al convertir una variable continua en una discreta, los algoritmos pueden trabajar con frecuencias (Apolinar, I., 2020).



KDD

Etapa de preprocesamiento/Transformación

ChiMerge

Este método de discretización utiliza el enfoque de juntar intervalos. Este método tiene dos características importantes: primero, que las frecuencias relativas de clase deben ser bastante parecidas dentro de un intervalo (de lo contrario, se debe dividir el intervalo).

Dos intervalos adyacentes no deben tener similares frecuencias relativas de clase (de lo contrario, se debe juntar). Este método utiliza la prueba de independencia $\chi 2$ para saber si dos tributos Fi son independientes. Para dos intervalos adyacentes, si la prueba $\chi 2$ concluye que la clase es independiente de los intervalos, los intervalos se deben juntar. Si la prueba $\chi 2$ concluye que no son independientes y además la diferencia en frecuencia relativa de la clase es estadísticamente significativa, los dos intervalos deben continuar separados.

KDD

Etapa de preprocesamiento/Transformación

- **Escalado**

El escalado entre variables se refiere al proceso de ajuste de valores, que permita la ejecución de algoritmos de modelos entre ellos, manteniendo la proporcionalidad y la distribución de los datos.

En ocasiones, el problema con el que nos encontramos es que **el rango de alguna de las variables es mayor que el del resto**.

Para facilitar esto y hacer que todos los datos preserven una distribución similar, se utiliza el escalado estándar, que transforma los datos originales en otros cuya distribución estadística tiene una media 0 y una desviación estándar igual a 1 (Bolanos, L., 2020).

KDD

Contenido

1. Fases del proceso de KDD
2. Etapas de preprocesamiento
3. **La fase de minería de los datos**

KDD

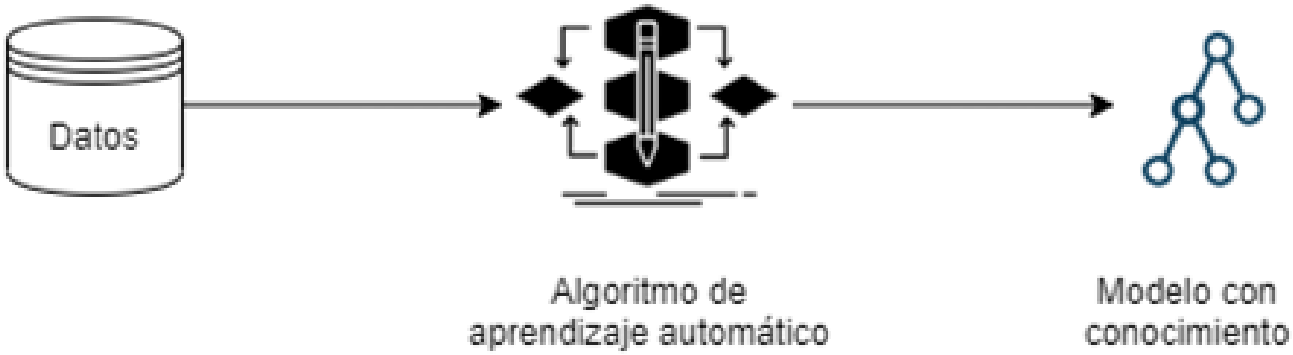
3. La fase de minería de los datos

Tras el preprocesamiento, los datos están preparados para la siguiente fase. En ésta se usa un **algoritmo de minería de datos**, a fin de extraer de éstos el conocimiento que no resulta obvio ni es trivial, aunque esté implícito en ellos.

KDD

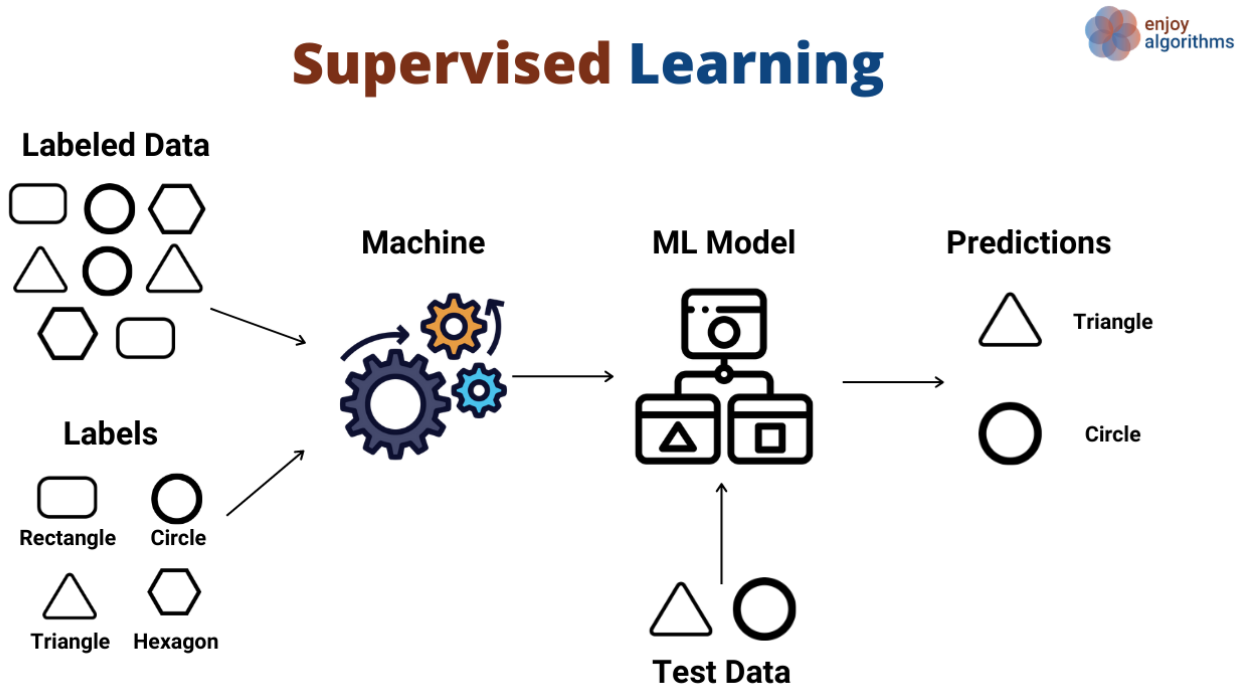
Los algoritmos de minería de datos se clasifican en supervisados y no supervisados o automáticos.

Un **algoritmo de aprendizaje automático** procesa los datos y lo que genera como salida no son nuevos datos, como es habitual en la mayoría de algoritmos de ordenador, sino un modelo que representa el conocimiento extraído.



KDD

En los **algoritmos de aprendizaje supervisado** se genera un modelo predictivo, basado en datos de entrada y salida. El aprendizaje supervisado es una técnica para deducir una función, a partir de datos de entrenamiento.

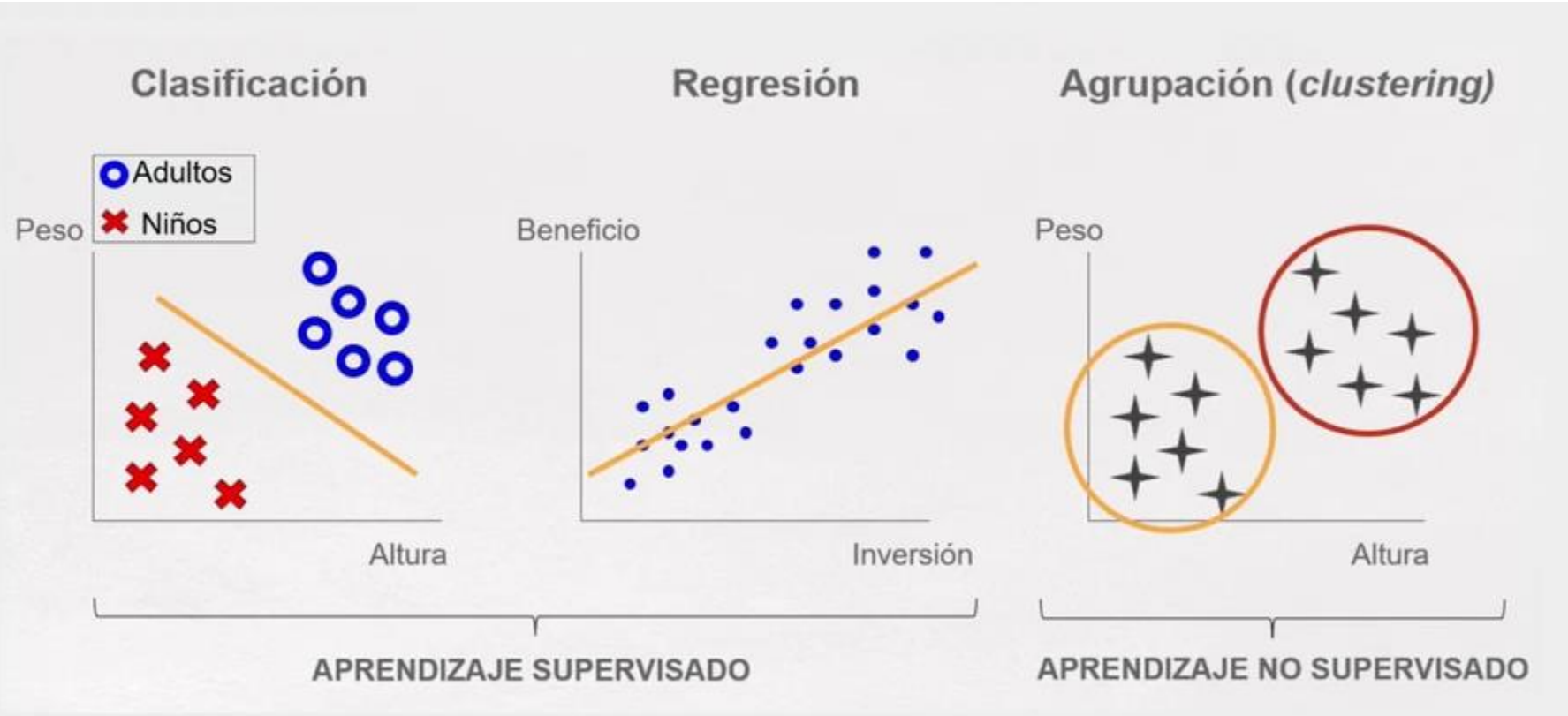


KDD

En la tabla siguiente se muestran algunas de las técnicas de minería de datos en ambas categorías:

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Detección de Desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (Clustering)
Series temporales	Reglas de Asociación
	Patrones Secuenciales

KDD


Clasificación
Regresión
Agrupación (clustering)
Adultos
Niños
Peso
Altura
Beneficio
Inversión
Peso
Altura
APRENDIZAJE SUPERVISADO
APRENDIZAJE NO SUPERVISADO

KDD

La siguiente tabla muestra algunas de las técnicas más comunes de Minería de Datos

Métodos estadísticos	ANOVA
	Prueba Ji cuadrado
	Análisis de componentes principales
	Análisis de clusters
	Análisis discriminante
	Regresión lineal
	Regresión logística
Arboles de decisión	CHAID
	CART
Reglas do asociación	
Redes de neuronas artificiales	
Algoritmos genéticos	
Otros	Lógica difusa
	Series temporales

KDD

Métodos Estadísticos:

- **ANOVA:** Análisis de la Varianza, contrasta si existen diferencias significativas entre las medidas de una o más variables continuas, en grupos de población distintos.

- **Ji cuadrado:** Contrasta la hipótesis de independencia entre variables.

- **Componentes principales:** Permite reducir el número de variables observadas a un menor número de variables artificiales, conservando la mayor parte de la información sobre la varianza de las variables.

- **Análisis de clústeres:** Permite clasificar una población en un número determinado de grupos, sobre la base de semejanzas y diferencias de perfiles existentes entre los diferentes componentes de dicha población.

KDD

Métodos Estadísticos (Cont):

- **Análisis discriminante:** Método de clasificación de individuos en grupos que previamente se han establecido, que permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definan la pertenencia al grupo.

- **Regresión Lineal:** Técnica básica del Data Mining. Un modelo de regresión lineal se implementa identificando una variable dependiente (y) y todas las variables independientes (X1, X2,...). Se asume que la relación entre estas y aquella es lineal. Todas las variables han de ser continuas. El resultado es la ecuación de la recta que mejor se ajusta al juego de datos y esta ecuación se interpreta o se usa para predicción.

- **Regresión Logística:** Puede trabajar con variables discretas. También requiere que todas las variables sean lineales.

KDD

Árboles de decisión:

Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones, mediante la ruptura y la subdivisión sistemática de la información contenida en el conjunto de datos. El árbol de decisión se construye partiendo el conjunto de datos en dos (CART) o más (CHAID) subconjuntos de observaciones, a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado, utilizando el mismo algoritmo.

El método CHAID (Chi Squared Automatic Interaction Detector) es útil en aquellas situaciones en las que el objetivo es dividir una población en distintos segmentos, basándose en algún criterio de decisión.

KDD

Reglas de asociación:

Derivan de un tipo de análisis que extrae información por coincidencias. Este análisis, a veces llamado "cesta de la compra“, permite descubrir correlaciones o coocurrencias en los sucesos de la base de datos a analizar y se formaliza en la obtención de reglas de tipo: SI ... ENTONCES...

Redes Neuronales *(“Neural Networks") :*

Las Redes Neuronales constituyen una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro.

KDD

Redes Neuronales *(“Neural Networks")* (Cont.):

Las redes neuronales son una nueva forma de analizar la información, con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender patrones y características dentro de los datos.

Una vez adiestradas las redes neuronales, pueden hacer previsiones, clasificaciones y segmentación.

Las redes neuronales se construyen estructurando en una serie de niveles o capas compuesta por nodos o "neuronas". Poseen dos formas de aprendizaje derivadas del tipo de paradigma, que usan: el supervisado y el no supervisado.

KDD

Algoritmos Genéticos *(“Genetic Algorithms”) :*

Los Algoritmos Genéticos son otra técnica que debe su inspiración, de nuevo, a la Biología, como las Redes Neuronales.

Estos algoritmos representan la modelización matemática de cómo los cromosomas, en un marco evolucionista, alcanzan la estructura y la composición más óptima en aras de la supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies, que se plasma en mutaciones y cambios en los genes o cromosomas.

Los Algoritmos Genéticos hacen uso de las técnicas biológicas de reproducción (mutación y cruce), para ser utilizadas en todo tipo de problemas de búsqueda y optimización.

KDD

Lógica Difusa *(“fuzzy logic"):*

La Lógica Difusa surge de la necesidad de modelizar la realidad de una forma más exacta, evitando precisamente el determinismo o la exactitud.
La Lógica permite el tratamiento probabilístico de la categorización de un colectivo.

La Lógica Difusa es aquella técnica que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos, en los que categorizamos un colectivo, o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

KDD

Series Temporales:

Consisten en el estudio de una variable a través del tiempo para, a partir de ese conocimiento, y bajo el supuesto de que no van a producirse cambios estructurales, poder realizar predicciones.

Suelen basarse en un estudio de la serie en ciclos, tendencias y estacionalidades, que se diferencian por el ámbito de tiempo abarcado, para, por composición, obtener la serie original.

Se pueden aplicar enfoques híbridos con los métodos anteriores, en los que la serie se puede explicar no sólo en función del tiempo, sino también como combinación de otras variables de entorno más estables y, por lo tanto, más fácilmente predecibles.

Referencias bibliográficas

Agrawal, R. y Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *vldb Conference*, Santiago de Chile.

Agrawal, R. y Srikant, R. (1995). Mining Sequential Patterns. *The 11th International Conference on Data Engineering icde*, Taipei, República de China.

Beltran, B., (2019). Minería de datos. Recuperado de http://bbeltran.cs.buap.mx/NotasMD.pdf

Charte, F., (2020). *Cómo es el proceso de extraer conocimiento a partir de bases de datos*. Recuperado de https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx

Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning Journa*l, *1*(1), 81-106.

Referencias bibliográficas

Ng, R. y Han, J. (1994). Efficient and Effective Clustering Method for Spatial Data Mining. *vldb Conference*. Santiago de Chile, Chile.

Srikant R. y Agrawal, R. (1996). *Mining quantitative association rules in large relational tables, acm sigmod, Montreal*. Recuperado de http://rakesh.agrawal-family.com/papers/sigmod-96qassoc.pdf

Wang, M., Iyer, B. y Scott, J. (1998). Scalable Mining for Classification Rules in Relational Databases. I*nternational Database Engineering and Application Symposium - Ideas*. Cardiff, Wales.