KDD: ¿Qué es el Knowledge Discovery in Databases o KDD?

Cuando hablamos de grandes cantidades de datos, el Descubrimiento de Conocimiento en Bases de Datos o KDD se refiere al proceso de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles.

Es probable que no hayas escuchado anteriormente el concepto de KDD. También es probable que te suene más conocido el término Minería de Datos, ¿cierto? Sin embargo, la Minería de Datos es solamente uno de los pasos en ese camino más amplio hacia el descubrimiento del conocimiento latente en tus datos.

KDD vs Minería de Datos

Knowledge Discovery implica la evaluación e interpretación de patrones y modelos para tomar decisiones con respecto a lo que constituye conocimiento y lo que no lo es. Por lo tanto, el KDD requiere de un amplio y profundo conocimiento sobre tu área de estudio.

Por otra parte, la Minería de Datos, exploración de datos o Data Mining, no requiere tanto conocimiento sobre el área de estudio, sino más conocimiento técnico.

Como mencionamos anteriormente, la Minería de Datos es un paso que forma parte del KDD e implica el análisis de grandes cantidades de datos observacionales, para encontrar relaciones insospechadas.

*El Data Mining se ocupa de reunir los datos de manera novedosa, entendible y útil para el propietario o usuario final.*

Los Pasos del KDD

Para entender mejor este proceso, vamos a descomponerlo en sus 6 pasos fundamentales.

Para cada paso, te mostraremos sus componentes o etapas y algunas sugerencias o recomendaciones.

Paso 1

**COMPRENSIÓN DEL DOMINIO DEL ESTUDIO Y ESTABLECIMIENTO DE OBJETIVOS**

* *Desarrollo de un entendimiento sobre el dominio*
* *Descubrimiento de conocimiento previo que sea relevante*
* *Definición del objetivo del KDD*

Como en cualquier tipo de investigación, es fundamental tener muy claros los límites y objetivos de lo que pretendemos. Es muy fácil perder el rumbo en el océano infinito de datos a nuestra disposición.

En este paso es cuando reconocemos las fuentes de información más importantes y quienes tienen control sobre ellas. También es relevante incluir toda la metadata relacionada, dimensionar la cantidad de datos, y formatos.

Te recomendamos que toda la información más importante que se encuentre solamente en medios físicos sea digitalizada, previo a iniciar las actividades de KDD.

Paso 2

**CREACIÓN DE UN SET DE DATOS (DATASET) OBJETIVO**

* *Selección e integración de los datos objetivo provenientes de fuentes múltiples y heterogéneas.*

Los datos que necesitamos están ahí afuera. Datos relevantes al dominio y objetivos de nuestro estudio pueden existir, por ejemplo, en bases de datos relacionales, colecciones de documentos, correos electrónicos, fotografías, clips de vídeo, bases de datos de procesos, bases de datos de transacciones de clientes, registros web, o web logs, etc. Es importante homogeneizar los formatos para que los datos sean más fáciles de procesar y analizar.

Paso 3

**LIMPIEZA Y PROCESAMIENTO DE DATOS**

* *Eliminación de ruido y datos aislados o outliers.*
* *Uso del conocimiento previo para eliminar las inconsistencias y los duplicados.*
* *Escogencia y uso de estrategias para manejar la información faltante en los datasets.*

Los datasets disponibles en la actualidad usualmente están incompletos (valores de atributos faltantes), tienen ruido (errores y datos aislados o outliers), o presentan inconsistencias (discrepancias en los datos recolectados).

Estos “datos sucios” pueden confundir el proceso de minería y conducir a resultados inválidos o poco confiables.

El preprocesamiento y la limpieza tienen el objetivo de mejorar la calidad de los datos y los resultados de la minería. Recuerda que la implementación de análisis complejos y el minado de grandes cantidades de datos puede tomar mucho tiempo, así que lo que podamos hacer para acortar ese tiempo será siempre de provecho.

Paso 4

**MINERÍA DE DATOS**

¡Y por fin llegamos a la Minería de Datos!

Cuando “minamos” en realidad lo que hacemos es convertirnos en exploradores. Nos adentramos en la inmensidad de los datos y descubrimos poco a poco los patrones o modelos presentes en ellos; las relaciones.

Y en esta exploración, una de nuestras herramientas más útiles son los algoritmos.

*¿Qué es un algoritmo? Básicamente, un algoritmo es una serie de instrucciones o reglas establecidas en un programa informático que nos permiten llegar a un resultado o solución.*

En el caso de la minería de datos, un algoritmo nos permite procesar un set de datos para obtener nueva información sobre ese mismo dataset.

En general, la minería de datos comprende tres pasos: la selección de la tarea, la selección del algoritmo (o algoritmos) y su uso.

PASO 1: SELECCIONANDO LA TAREA

Para saber qué algoritmos serán más útiles, primero es necesario saber qué tarea queremos realizar.

Puede ser que estemos buscando resultados estadísticos (como mediana o media por ejemplo), o bien, que nuestro objetivo sea realizar una predicción, una asociación, o identificar secuencias de datos.

Cada uno de estos objetivos nos guiará durante el siguiente paso.

PASO 2: SELECCIONANDO EL ALGORITMO O ALGORITMOS

Para trabajar con algoritmos necesitamos contar con conocimiento en varios campos, entre ellos las ciencias de la computación, estadística, machine learning, optimización, y otros.

En este paso nos preguntamos:

¿Cuál es el mejor algoritmo para buscar modelos y patrones en mis datos? Por ejemplo, elegimos entre un algoritmo de K-means o K-medoid.

¿Cuáles son mis parámetros y criterios de evaluación? Por ejemplo, el número de clusters.

¿Coincide el algoritmo que elegí con el objetivo general de mi proceso de KDD?

PASO 3: Y FINALMENTE ¡USAMOS LOS ALGORITMOS!

La implementación de los algoritmos es un proceso relativamente automatizado, que se desarrolla en el dataset objetivo.

*El algoritmo busca patrones y modelos que nos interesen, siguiendo sus reglas preestablecidas, que pueden incluir árboles de clasificación, modelos de regresión, clusters, modelos mixtos, entre otros.*

Paso 5

**INTERPRETACIÓN DE LOS PATRONES MINADOS**

Es importante que comprendamos la diferencia entre dos términos clave:

* *Patrones: son estructuras locales que hacen declaraciones sólo sobre un espacio restringido por variables. Esto tiene aplicaciones importantes en detección de anomalías como la detección de faltas en procesos industriales o de fraudes en el sistema bancario.*
* *Modelos: son estructuras globales que hacen declaraciones sobre cualquier punto en el espacio de medición. Por ejemplo, los modelos pueden predecir el valor de alguna otra variable.*

En la etapa de interpretación, hallamos los patrones y modelos en los datos analizados.

Los resultados deben presentarse en un formato entendible. Por esta razón las técnicas de visualización son importantes para que los resultados sean útiles, dado que los modelos matemáticos o descripciones en formato de texto pueden ser difíciles de interpretar para los usuarios finales.

Desde este punto del proceso es posible regresar a cualquiera de los pasos anteriores.

Paso 6

**UTILIZACIÓN DEL CONOCIMIENTO DESCUBIERTO**

¡Bingo! Descubrimos el conocimiento oculto en nuestros datos. Ahora es el momento de usar ese conocimiento para tomar mejores decisiones.

1. **Entendimiento del problema:** se trata de hablar con el cliente, conocer sus necesidades, conocer su negocio o actividad, conocer qué datos relevantes tiene disponibles y cuáles serían necesarios pero no están disponibles, etc.

2. **Entendimiento de los datos:** hay que saber qué significan los datos, si son continuos o discretos, qué tipo de valores toman,  qué utilidad futura pueden tener y saber si están bien capturados o no.

3. **Preparación de datos:** se trata de reflexionar sobre cómo guardar los datos. Típicamente hablaremos de tablas con filas y columnas, pero hay que ver cómo se organizan las tablas, cómo se interrelacionan entre ellas, etc. En definitiva organizar los datos para poder sacarles partido.

4. **Modelamiento:** una vez se tienen los datos organizados hay que definir los algoritmos que se van a utilizar para tratar los datos. Una vez tratados, los datos nos devolverán información útil.

5. **Evaluación:** los resultados obtenidos deben de ser sometidos a comprobación, verificar que están libres de errores, ratificar que son útiles para los objetivos perseguidos, etc.

6. **Despliegue funcional-comercial:** una vez se tiene automatizada la captura y tratamiento de datos para obtener unos resultados, se desarrollan herramientas, normalmente en forma de aplicaciones informáticas que permiten generar alertas, informes, estadísticas, etc. que tienen una utilidad directa para la toma de decisiones y sistema de información del cliente.