Unidad IV

Adquisición del conocimiento.

4.1 Introducción a la minería de datos

La **minería de datos** (es la etapa de análisis de "Knowledge Discovery in Databases" o KDD), es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y gestión de datos, procesamiento de datos, el modelo y las consideraciones de inferencia, métricas de Intereses, consideraciones de la Teoría de la complejidad computacional, post-procesamiento de las estructuras descubiertas, la visualización y actualización en línea.

El término es una palabra de moda, y es frecuentemente mal utilizado para referirse a cualquier forma de datos a gran escala o procesamiento de la información (recolección, extracción, almacenamiento, análisis y estadísticas), pero también se ha generalizado a cualquier tipo de sistema de apoyo informático decisión, incluyendo la inteligencia artificial, aprendizaje automático y la inteligencia empresarial. En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo". Incluso el popular libro "La minería de datos: sistema de prácticas herramientas de aprendizaje y técnicas con Java" (que cubre todo el material de aprendizaje automático) originalmente iba a ser llamado simplemente "la máquina de aprendizaje práctico", y el término "minería de datos" se añadió por razones de marketing. A menudo, los términos más generales "(gran escala) el análisis de datos", o "análisis" -. o cuando se refiere a los métodos actuales, la inteligencia artificial y aprendizaje automático, son más apropiados.

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis cluster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en el análisis adicional o, por ejemplo, en la máquina de aprendizaje y análisis predictivo. Por ejemplo, el paso de minería de datos podría identificar varios grupos en los datos, que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son

parte de la etapa de minería de datos, pero que pertenecen a todo el proceso KDD como pasos adicionales.

Los términos relacionados con la obtención de datos, la pesca de datos y espionaje de los datos se refieren a la utilización de métodos de minería de datos a las partes de la muestra de un conjunto de datos de población más grandes establecidas que son (o pueden ser) demasiado pequeñas para las inferencias estadísticas fiables que se hizo acerca de la validez de cualquier patrón descubierto. Estos métodos pueden, sin embargo, ser utilizados en la creación de nuevas hipótesis que se prueban contra poblaciones de datos más grandes.

4.2 Técnicas para el pre-procesamiento de datos: limpiado, reducción y normalización.

La reducción de riesgo se logra a través de la implementación de Medidas de protección, que basen en los resultados del análisis y de la clasificación de riesgo.

Las medidas de protección están divididos en medidas **físicas y técnicas**,**personales** y **organizativas**.

En referencia al Análisis de riesgo, el propósito de las medidas de protección, en el ámbito de la Seguridad Informática, solo tienen un efecto sobre los componentes de la Probabilidad de Amenaza, es decir aumentan nuestra capacidad física, técnica, personal y organizativa, reduciendo así nuestras

vulnerabilidades que están expuestas a las amenazas que enfrentamos. Las medidas normalmente no tienen ningún efecto sobre la Magnitud de Daño, que depende de los Elementos de Información y del contexto, entorno donde nos ubicamos. Es decir, no se trata y muy difícilmente se puede cambiar el valor o la importancia que tienen los datos e informaciones para nosotros, tampoco vamos a cambiar el contexto, ni el entorno de nuestra misión.

La **normalización** o **estandarización** es la redacción y solo aprobación de normas que se establecen para garantizar el acoplamiento de elementos construidos independientemente, así como garantizar el repuesto en caso de ser necesario, garantizar la calidad de los elementos fabricados, la seguridad de funcionamiento y trabajar con responsabilidad social.

La **normalización** es el proceso de elaborar, aplicar y mejorar las normas que se aplican a distintas actividades científicas, industriales o económicas con el fin de ordenarlas y mejorarlas. La asociación estadounidense para pruebas de materiales (ASTM) define la normalización como el proceso de formular y aplicar reglas para una aproximación ordenada a una actividad específica para el beneficio y con la cooperación de todos los involucrados.

Según la ISO (International Organization for Standarization) la normalización es la actividad que tiene por objeto establecer, ante problemas reales o potenciales, disposiciones destinadas a usos comunes y repetidos, con el fin de obtener un nivel de ordenamiento óptimo en un contexto dado, que puede ser tecnológico, político o económico.

La normalización persigue fundamentalmente tres objetivos:

- Simplificación: se trata de reducir los modelos para quedarse únicamente con los más necesarios.
- Unificación: para permitir el intercambio a nivel internacional.
- Especificación: se persigue evitar errores de identificación creando un lenguaje claro y preciso.

Las elevadas sumas de dinero que los países desarrollados invierten en los organismos normalizadores, tanto nacionales como internacionales, es una prueba de la importancia que se da a la normalización.

4.4 Herramienta para análisis del conocimiento, selección de datos, extracción de regla

El aumento del volumen y variedad de información
que se encuentra informatizada en base
s de datos digitales ha crecido espectacularmente en la última década.

• Gran parte de esta información es histórica
, es decir, representa transacciones o situaciones que se han producido.
• Aparte de su función de “memoria de la organización”, la información histórica es útil para predecir la información futura.