MINERIA DE DATOS.

Inteligencia de Negocios:

Según el Data Warehouse Institute, la Inteligencia de Negocios, o BI por sus siglas en inglés, se define como: "La combinación de tecnología, herramientas y procesos, que me permiten transformar los datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial."

Es decir, el conjunto de procesos tecnológicos que nos permite analizar los datos en busca de información relevante que, acumulada y asimilada a la estructura organizacional, nos permita tomar decisiones estratégicas relacionadas con el negocio.

Dentro de estas tecnologías, dos piezas claves son:

- La correcta extracción, transformación y carga de los datos.
- El correcto almacenamiento de estos datos para la toma de decisiones.
- La minería de datos sobre estos datos.
- La presentación de los resultados obtenidos de la minería de datos.

Los dos primeros elementos ya los hemos analizado como parte de los capítulos anteriores, y el tema de la presentación de los resultados producidos por los procesos de la minería de datos los veremos más adelante, pero… ¿Qué es la minería de datos?

Minería de Datos:

Según Frawley, Piatetsky-Shapiro y Matheus, Data Mining es "la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos".

Según Herb Edestein Data Mining es "el descubrimiento de patrones y relaciones ocultas en sus datos."

En esencia, son todas aquellas técnicas que nos permiten encontrar patrones ocultos en nuestros datos.

Métodos y algoritmos de Minería de Datos:

En la actualidad, los principales métodos de Minería de Datos son:

- Asociación: Afinidad de artículos en mis datos, también conocido como "Análisis de Canasta"
- Clusterización: Definir o encontrar grupos en mis datos (**no** confundir con clasificar).
- Predictivos:
  - Clasificación: Predicción de "tipo" (categorías) en los valores de mis datos.
  - Predicción: Predicción numérica en mis datos.

Cada uno de estos métodos, está asociado con algún tipo de modelo estadístico o "algoritmo", siendo en ocasiones homónimos. Entre los más comunes tenemos:

Asociación:

- Análisis de Correlación.

Clusterización:

Particionamiento: Agrupación en "n" grupos con características específicas.

- Modelos de Densidad: Clusterización de proximidad, donde se ubican núcleos de polarización y se establecen ciertos umbrales para considerar su pertenencia.
- Modelos Jerárquicos: Agrupación basado en características comunes partiendo de lo general a lo particular.

Clasificación:

Decidir, de manera discreta, es decir, en base a un conjunto de valores distintos específicos, a cuales de éstos grupos o categorías pertenece. Por ejemplo: "¿Debo darle un crédito a tal o cual persona?", "¿Debo o no jugar en ciertas condiciones climáticas?", "¿En qué categoría debo asignar a un cliente: Normal, Preferente o Senior…?"

Las técnicas más usadas son:

- Arboles de decisión. Se usan árboles de decisión para predecir en qué "clase" definida encajará cierto comportamiento.
- Redes neuronales: Estos algoritmos son particularmente útiles para el reconocimiento y discretización de valores, se basan en una serie de elementos de entrada (perceptrones de entrada) y otra serie de elementos de salida (perceptrones de salida), entre ellas hay una serie de una o más capas de procesamiento, compuesta de varios elementos que guardan ciertos valores numéricos en base a ciertas reglas de producción o "reglas delta", la red neuronal puede ser entrenada para afinar el reconocimiento de patrones, no obstante en este tipo de modelos queda más que patente que el precio que se tiene que pagar por aprender es olvidar…

- Modelos Bayesiano. Se utiliza el teorema de Bayes para establecer la probabilidad de que el elemento propuesto encaje en cierta "clase", aunque su principal uso es la comprobación de hipótesis.
- Algoritmos genéticos. Basados en el concepto de la "Selección Natural", una población es inicializada de manera aleatoria a través de ciertas reglas, y se va entrenando con cierto grupo de datos. En cada iteración se introducen ciertas "mutaciones" (ligeros cambios en las reglas) y "extinciones" (eliminación o ajustes de ciertos resultados)

La efectividad de la predicción se realiza a través de la llamada "Matriz de Confusión" donde, después de haber divido un conjunto de datos estadísticos, con su respectivo resultado final, en dos muestras: modelado (50%-70%) y evaluación (50%-30%), se genera el modelo a partir de la primera muestra de los datos, y se comprueba la efectividad de dicho modelo con la otra porción de la muestra en función del siguiente diagrama y fórmula:

		REAL: POSITIVO	REAL: NEGATIVO
PREDICCION	POSITIVO	VERDADERO POSITIVO (TP)	FALSO POSITIVO (FP)
PREDICCION	NEGATIVO	FALSO NEGATIVO (FN)	VERDADERO NEGATIVO (TN)

$$PRECISION = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Normalmente una precisión de más de 70% puede usarse en casos de negocio, no obstante, en el caso de las redes neuronales, se recomienda su uso cuándo éstas puedan predecir mejor el resultado que un ser humano.

Predicción:

Este tipo de estimación futura, o a-priori, se distingue de la clasificación, en que la predicción implica valores numéricos, y normalmente se representan como una función que representa una aproximación estadística de los datos representados.

- Regresión Estadística. Este tipo de modelos estadísticos es de lo más comunes y ampliamente usados, consiste en identificar tendencias que nos permiten definir funciones para poder predecir, con cierto grado de precisión, valores de la variable dependiente para valores no capturados de la variable, o variables independientes. Si el valor a predecir está dentro del rango de valores capturados, se dice que se está intrapolando el valor, si la predicción corresponde a rangos fuera del intervalo capturado se habla de una extrapolación.

Cuando los valores de la variable, o variables independientes, están dentro del rango de los valores usados para el modelado, se dice que estamos frente a una interpolación, cuán los valores evaluados están por arriba, o por abajo, de dicho rango de valores, se dice que se trata de una extrapolación.

Existen diferentes alternativas para implementar estos modelos, desde herramientas gráficas que permiten diseñar, probar e implementar dichos modelos como SPSS, hasta herramientas OpenSource como el Lenguaje R.

Conceptos vistos:

Business Intelligence.
Data Mining
- Asociación.
- Clusterización.
- Clasificación.
- Predicción.

Algoritmo estadístico.
- Particionamiento.
- Modelos Jerárquicos.
- Modelos de densidad.
- Arboles de decisión.
- Modelos Bayesianos.
- Algoritmos genéticos.
- Redes neuronales.
- Regresión.
  - Intrapolación.
  - Extrapolación.