

Juan Fernando Jaramillo

Santiago del Campo

Juan David Carvajal

Germán Carvajal

**Fase 2:** Recopilación de información

Con el objetivo de tener claridad en algunos conceptos o términos, a continuación, se plantean una serie de definiciones relacionadas con el problema a tratar.

**Datos:**

Como se dice en Tecnologías Información, los datos son cualquier hecho, número o texto que puede ser procesado por una computadora.

Hoy en día, las organizaciones están acumulando grandes cantidades, y cada vez mayores, de datos en diferentes formatos y diferentes bases de datos. Esto incluye:

* **Datos operativos o transaccionales**, Tales como, ventas, costos, inventario, nómina y contabilidad
* **Datos no operacionales**, Tales como ventas de la industria, datos del pronóstico del tiempo y datos macroeconómicos
* **Metadatos** - Datos sobre los datos en sí, como el diseño de base de datos lógica o los datos de las definiciones del diccionario

**Información:**

Los patrones, asociaciones, o relaciones entre todos estos datos pueden proporcionar información. Por ejemplo, el análisis del punto de venta de datos de transacciones de venta puede dar información sobre qué productos se venden y cuándo. (Información, 2018)

**Conocimiento:**

La información puede ser convertida en conocimiento acerca de los patrones históricos y las tendencias futuras. Por ejemplo, la información resumida sobre las ventas de supermercados minoristas puede ser analizada a la luz de los esfuerzos de promoción para facilitar el conocimiento del comportamiento de compra del consumidor. Por lo tanto, un fabricante o distribuidor puede determinar qué elementos son los más susceptibles a los esfuerzos de promoción. (Información, 2018)

**Almacenes de datos:**

Los dramáticos avances en la captura de datos, potencia de procesamiento, transmisión de datos y capacidades de almacenamiento están permitiendo a las organizaciones integrar sus diversas bases de datos en almacenes de datos. El almacenamiento de datos se define como un proceso de gestión de datos centralizado y su recuperación. El almacenamiento de datos, como la minería de datos, es un término relativamente nuevo, aunque el concepto en sí ha existido por años.

El almacenamiento de datos representa una visión ideal de mantener un repositorio central de todos los datos de la organización. La centralización de los datos es necesaria para maximizar el acceso de los usuarios y el análisis. Los dramáticos avances tecnológicos hacen que esta visión sea una realidad para muchas empresas. Y, los avances igualmente espectaculares en el software de análisis de datos, permite a los usuarios acceder a estos datos libremente. El software de análisis de datos es lo que apoya la minería de datos. (Información, 2018)

**Análisis de datos:**



El Análisis de Datos (*Data Analysis, o DA*) es la ciencia que examina datos en bruto con el propósito de sacar conclusiones sobre la información. El análisis de datos es usado en varias industrias para permitir que las compañías y las organizaciones tomen mejores decisiones empresariales y también es usado en las ciencias para verificar o reprobar modelos o teorías existentes. El análisis de datos se distingue de la extracción de datos por su alcance, su propósito y su enfoque sobre el análisis. Los extractores de datos clasifican inmensos conjuntos de datos usando software sofisticado para identificar patrones no descubiertos y establecer relaciones escondidas. El análisis de datos se centra en la inferencia, el proceso de derivar una conclusión basándose solamente en lo que conoce el investigador.

**Minería de datos:**

La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos. (Owen Duncan, 2018)

Estos patrones y tendencias se pueden recopilar y definir como un *modelo de minería de datos*. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

* Pronóstico: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
* Riesgo y probabilidad: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
* Recomendaciones: determinación de los productos que se pueden vender juntos y generación de recomendaciones.
* Búsqueda de secuencias: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
* Agrupación: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.
* La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. Definir el problema
2. Preparar los datos
3. [Explorar los datos](https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017#ExploringData)
4. Generar modelos
5. Explorar y validar los modelos
6. Implementar y actualizar los modelos

(Owen Duncan, 2018)

**Técnicas de Minería de datos:**

En el ámbito de la investigación las técnicas de data mining pueden ayudar a los científicos a clasificar y segmentar datos y a formar hipótesis. El data mining permite encontrar información escondida en los datos que no siempre resulta aparente, ya que, dado el gigantesco volumen de datos existentes, gran parte de ese volumen nunca será analizado. (Cloud, 2014)

Las técnicas de data mining pueden ser de dos tipos:

* **Métodos descriptivos**– Buscan patrones interpretables para describir datos. Son los siguientes: clustering, descubrimiento de reglas de asociación y descubrimiento de patrones secuenciales. (Cloud, 2014)

Los métodos descriptivos se han utilizado, por ejemplo, para ver que productos suelen adquirirse conjuntamente en el supermercado.

* **Métodos predictivos-** Usan algunas variables para predecir valores futuros o desconocidos de otras variables. Son los siguientes: clasificación, regresión y detección de la desviación. (Cloud, 2014)

Dentro de los métodos de análisis de asociación se encuentran unos que son los mas utilizados y que posiblemente sean usados en el desarrollo del programa.

**Fuerza Bruta:** Este algoritmo o método básicamente genera y prueba todas las posibilidades de grupos de elementos que puedan ser formados con las entradas actuales. En cuanto a precisión y exactitud, es el mejor método, así como es fácil de implementar. Su desventaja es que genera y prueba demasiadas combinaciones de elementos, por lo que es temporal y espacialmente muy complejo, no permitiendo el análisis de grupos de datos de mas de unas pocas decenas.

**Principio A Priori:** Este principio nos dice que cualquier elemento que sea infrecuente, es decir, que no cumple la ley de soporte, generara o será subconjunto de otros grupos, que tampoco serán frecuentes, por lo que descarta todos estos grupos.

Es mucho mas eficiente que el algoritmo de fuerza bruta, sin perder mucha precisión.

**Algoritmo de Crecimiento FP:** Este método es radicalmente diferente a los anteriores, pues genera una estructura auxiliar a partir de los datos, un árbol que contiene los elementos, asi como una representación de las veces que aparece un elemento o un grupo de elementos, esto mediante sobreposición de objetos en el árbol, y al reducir significativamente el tamaño de datos, además de la misma naturaleza de la estructura, permite extraer las asociaciones directamente.

(P.N Tan, 2005)

**Cadenas de Márkov:** Es un proceso estadístico de predicción de estados futuros a partir de los datos inmediatamente anteriores, que puede ser aplicado al numero de veces y forma de aparición de los grupos de elementos del problema actual, y al ser estrictamente probabilístico, es extremadamente eficiente.

Así mismo, hay diferentes formas de generar los grupos de objetos más frecuentes utilizados por los métodos de fuerza bruta, y especialmente por el principio a priori, que no son mutuamente excluyentes, como:

**Generación fuerza bruta:** Generar todas las posibles combinaciones y después descartarlas con el principio a priori.

**F(k-1) X F(1):** Utilizar el nivel anterior y el actual para generar todas las combinaciones posibles de elementos.

**F(k-1) X F(k-1):** Utilizar combinaciones del nivel anterior para generar todas las posibles combinaciones del nivel actual y después probar las que cumplan las reglas.

**Aplicación máxima:** Generar la siguiente generación de grupos de datos a partir de los grupos de datos que tienen mas frecuencia.

**General a especifico:** Pasar de un grupo amplio de datos, como los grupos de las combinaciones de tamaño anterior, a un grupo de mayor tamaño y menor cantidad de ocurrencias.

**Especifico a general:** Pasar de grupos de elementos de menor cantidad total, como los de mayor ocurrencia, a un grupo mas general, combinando los elementos de los grupos específicos.

**Clases equivalentes:** Separar los datos por características que los hacen iguales, como el tamaño del grupo de datos, y después generar conexiones que los unan.

**Primero profundidad:** Recorrer los datos para generar las conexiones de forma descendente, generando el siguiente nivel con los datos del anterior sin antes terminar de analizar todo el nivel anterior.

**Primero anchura:** Recorrer los datos para generar las conexiones de forma lateral, generando el siguiente nivel con los datos del anterior primero terminando de analizar todo el nivel anterior.

(P.N Tan, 2005)

**Estado del Arte:**

Existen ya varias aplicaciones que realizan análisis y minería de datos. Entre ellas se encuentran Sisense, Oracle Data Minig, RapidMiner, Micrososft SharePoint, IBM Cognos, entre otros. (FinancesOnline, 2018)

Algunos de ellos, generan análisis gráficos, como el Sisense, pero la mayoría de estos generar análisis mediante métodos predictivos, que, dependiendo de varias entradas, pueden predecir los mejores clientes, anomalías, oportunidades de venta, etc. (FinancesOnline, 2018)

# Referencias

Cloud, C. (8 de Septiembre de 2014). *Clinic Cloud*. Obtenido de Que es el Data Mining?: https://clinic-cloud.com/blog/data-mining-que-es-definicion-mineria-de-datos/

FinancesOnline. (3 de Septiembre de 2018). *FinancesOnline*. Obtenido de Top 15 Data Mining Software Systems: https://financesonline.com/top-15-data-mining-software-systems/

Información, T. (10 de Septiembre de 2018). *Tecnologías Información*. Obtenido de Minería de datos: https://www.tecnologias-informacion.com/mineria-de-datos.html#

Owen Duncan, S. C. (30 de Abril de 2018). *Microsoft Docs*. Obtenido de Conceptos de Minería de Datos: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017

P.N Tan, M. S. (2005). Association analysis: Basic concepts and algorithms. *Introduction to Data Mining*, 327-414.