UNIVERSIDAD NACIONAL DEL NORDESTE

FACULTAD DE CIENCIAS EXÁCTAS, NATURALES Y AGRIMENSURA

LICENCIATURA EN SISTEMAS DE INFORMACIÓN

BASE DE DATOS II

Grupo Nº 6

Integrantes:	**LU:**
● Avellanal, Facundo Nahuel	41549
● Bacigaluppe, Ricardo Maximiliano	46833
● Espíndola, David Gabriel	46350
● Fernández, Denis Abraham	34881
● Jordán Villalva, María de los Angeles	50482
● Merlo, Veronica Soledad	50497

Tema: Fase de selección en Minería de Datos

Año 2022

Contenido

Índice de figuras

Índice de tablas

INTRODUCCIÓN

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos.

No obstante, la minería de datos es ya un concepto muy evolucionado que necesita ser aproximado conceptualmente por etapas.

El proceso de extracción del conocimiento comienza con la recopilación e integración de la información a partir de unos datos iniciales de que se dispone (fase de selección de datos). Las primeras fases del proceso son muy importantes porque determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. La información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas, generalmente ordenada en almacenes de datos. El análisis posterior será mucho más sencillo si la fuente es unificada, accesible y desconectada del trabajo transaccional. Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa. La disponibilidad de grandes volúmenes de información en esta fase nos lleva a la necesidad de usar técnicas de muestreo para la selección de datos.

PROCESO DE EXTRACCIÓN DEL CONOCIMIENTO

La minería de datos es sólo una etapa del proceso de extracción de conocimiento a partir de datos (KDD Knowdlege Discovery in Databases), el cual constituye el primer modelo que define el descubrimiento de conocimiento en bases de datos como un “proceso”, compuesto por distintas etapas y fases que van desde la preparación de los datos hasta la interpretación y difusión de los resultados.

Los patrones deberían ser válidos para nuevos datos, novedosos en el sentido que deberían aportar un nuevo conocimiento al dominio de aplicación y potencialmente útiles para el usuario final o tomador de decisiones. KDD es un proceso iterativo e interactivo. Iterativo ya que la salida de alguna de las fases puede retroceder a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o más generalmente un experto en el dominio del problema, debe ayudar a la preparación de los datos y validación del conocimiento extraído.

El modelo de proceso KDD se resume en las siguientes cinco fases:

1. Selección de los datos sobre los que se trabajará.
2. Pre-procesamiento de los datos, donde se realiza un tratamiento de los datos incorrectos y ausentes.
3. Transformación de los datos y reducción de la dimensionalidad.
4. Minería de datos, donde se obtienen los patrones de interés según la tarea de minería que llevemos a cabo (descriptiva o predictiva).
5. Interpretación y evaluación del nuevo conocimiento en el dominio de aplicación.

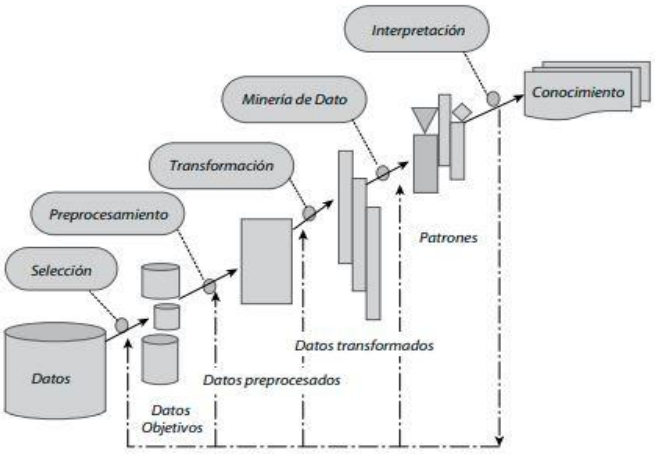

Figura 1 - Etapas del proceso KDD

Fase de Selección

En la fase de selección se integran y recopilan los datos, se determinan las fuentes de información que pueden ser útiles y donde conseguirlas, se identifican y seleccionan las variables relevantes en los datos y se aplican las técnicas de muestreo adecuadas.

Las primeras fases de la minería de datos determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:

- En bases de datos y otras fuentes muy diversas, tanto internas como externas.
- Muchas de estas fuentes son las que se utilizan para el trabajo transaccional.
- Se requiere un historial suficiente (1, 5 o 10 años dependiendo del ámbito).
- El nivel de detalle (granularidad) para la minería de datos ha de ser alto.
- Volúmenes de datos muy grandes. El análisis posterior será mucho más sencillo si la fuente es unificada, accesible y desconectada del trabajo transaccional.

El análisis posterior será mucho más sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional.

El proceso subsiguiente de Minería de Datos depende mucho de la fuente, como ser:

- OLAP (Base de Datos orientadas al procesamiento analítico).
- OLTP (base de Datos orientadas al procesamiento de transacciones).
- Datawarehouse (Almacén de datos).
- ROLAP o MOLAP (Análisis de datos a través de un modelo de datos multidimensional).

Depende también del tipo de usuario, en donde reconocemos dos tipos:

- Picapedreros (O granjeros), los cuales se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anormales, etc.
- Exploradores, los cuales se encargan de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

La información interna de la organización se puede obtener en diferentes formatos:

- Bases de datos operacionales
- Hojas de cálculo
- Informes internos: estratégicos
- Reglas de negocio

Aparte de información interna, los almacenes de datos pueden recoger información externa:

- Demografías (censo), páginas amarillas, psicografías (perfiles por zonas), uso de Internet, información de otras organizaciones.
- Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
- Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas deportivas, catástrofes, etc.
- Bases de datos externas compradas a otras compañías.

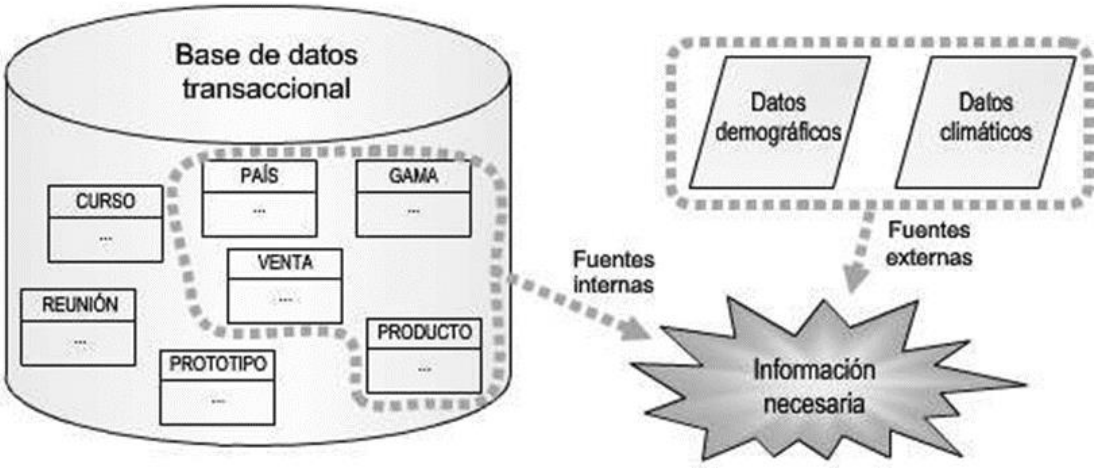

Figura 2 - Modelo de extracción de la información

FUENTES DE DATOS

Data Warehouse

Generalmente la información que se quiere investigar se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas. Muchas de estas fuentes son las que se utilizan para el trabajo diario (base de datos operacionales o transaccionales). Sobre estas mismas bases de datos de trabajo ya se puede extraer conocimiento porque se utilizan para mantener el trabajo transaccional diario de los sistemas de información originales (conocido como OLTP, On-Line Transactional Processing) y para análisis de los datos en tiempo real sobre la misma base de datos (conocido como OLAP, On-Line Analytical Processing). Pero este análisis perturba el trabajo transaccional diario de los sistemas de información originales ya que este tipo de base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos, tarea esta última que debe hacerse por la noche o fines de semana.

Los almacenes de datos o Data Warehouse permiten disponer de sistemas de información de apoyo a la toma de decisiones (DSS o Decision Support Systems) y de bases de datos que permitan extraer conocimiento de la información histórica almacenada en la organización. Se trata de bases de datos diseñadas con un objetivo de explotación (orientadas al análisis) distinto al de las bases de datos de los sistemas operacionales (orientadas al proceso). Un almacén de datos es una colección de datos diseñada para dar apoyo a la toma de decisiones orientada hacia la información relevante de la organización (se diseña para consultar eficientemente información relativa a las actividades básicas de la organización como ventas, compras y producción, y no para soportar los procesos que se realizan en ella como gestión de pedidos, facturación, etc.), integrada (integra datos recogidos de diferentes sistemas operacionales de la organización y/o fuentes externas), variable en el tiempo (los datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente) y no volátil (los datos almacenados no son actualizados, solo son incrementados).

Los almacenes de datos presentan como ventajas claras para las organizaciones la rentabilidad de las inversiones realizadas para su creación, el aumento de la competitividad en el mercado y aumento de la productividad de los técnicos de dirección, siendo los principales problemas la infravaloración del esfuerzo necesario para su diseño y creación, la infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos, el incremento continuo de los requisitos de los usuarios y la privacidad de los datos.

Las componentes típicas de un almacén de datos son un Sistema ETL (Extraction, Transformation, Load), un Repositorio Propio de Datos con información relevante o metadatos, Interfaces y Gestores de Consulta que permitan acceder a los datos conectándose

sobre ellos herramientas más sofisticadas (OLAP, EIS, minería de datos) y Sistemas de Integridad y Seguridad que se encargan de un mantenimiento global, copias de seguridad, etc. El Sistema ETL realiza las funciones de extracción de las fuentes de datos (transaccionales o externas), transformación (limpieza, consolidación) y la carga de almacén de datos.

Las herramientas de explotación de los almacenes de datos han adoptado un modelo multidimensional de datos. Son típicas las herramientas de OLAP, que presentan al usuario una visión multidimensional de los datos (esquema multidimensional) para cada actividad que es objeto de análisis. El usuario realiza consultas a la herramienta OLAP seleccionando atributos de este esquema multidimensional sin conocer la estructura interna (esquema físico) del almacén de datos. La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema (p.ej. mediante una sentencia SELECT). De esta forma se favorece la fase de selección en el proceso de extracción del conocimiento.

Las herramientas de OLAP se caracterizan por ofrecer una visión multidimensional de los datos (matricial), no imponer restricciones sobre el número de dimensiones, ofrecer simetría para las dimensiones, permitir definir de forma flexible (sin limitaciones) sobre las dimensiones: restricciones, agregaciones y jerarquías entre ellas, ofrecer operadores intuitivos de manipulación y ser transparentes al tipo de tecnología que soporta el almacén de datos (ROLAP o MOLAP). Los sistemas ROLAP se implementan sobre la tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento (índices de mapas de bits, índices de JOIN, técnicas de particionamiento de datos, optimizadores de consultas, extensiones de SQL, etc.).

Sistemas OLAP

OLAP, conocido como procesamiento analítico en línea (On-Line Analytical Processing por sus siglas en inglés), es una solución utilizada en el campo de la Inteligencia de Negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ellos utiliza estructuras de datos diversas, normalmente multidimensionales que contienen datos resumidos de grandes Bases de Datos o Sistemas Transaccionales (OLTP).

La principal razón para utilizar OLAP para las consultas es su rapidez de respuesta teniendo en cuenta que las consultas pueden realizarse a multitablas. Cabe destacar que OLAP se destaca a la hora de ejecutar sentencias SQL del tipo SELECT.

En cuanto a su funcionalidad, debemos tomar al sistema OLAP como un concepto de cubo multidimensional el cual está compuesto por hechos numéricos o medidas que se van clasificando en dimensiones. El cubo es típicamente creado a partir de un esquema de estrella y las medidas se obtienen de los registros de una tabla de hechos y las dimensiones se derivan en las dimensiones de los cuadros. El modelo de datos multidimensional simplifica a los usuarios formular consultas complejas, arreglar datos en un reporte, cambiar de datos resumidos a datos detallados y filtrar o rebanar los datos en subconjuntos significativos.

El análisis OLAP te permite “navegar” fácilmente por la información, solicitando con el detalle preciso y con los filtros adecuados. Se puede hacer de manera dinámica, fácil, sobre la marcha, sin necesitar asistencia, rápido, y utilizando el lenguaje de negocio.

Existen tres modelos de almacenamiento:

- MOLAP (OLAP Multidimensional): en estos sistemas los datos se encuentran almacenados en una estructura multidimensional. Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado. Estos valores precalculados o agregaciones son la base de las ganancias de desempeño de este sistema. Algunos sistemas utilizan técnicas de comprensión de datos para disminuir el espacio de almacenamiento en disco debido a valores precalculados.

Tabla 1 - Ventajas y Desventajas de MOLAP

Ventajas	**Desventajas**
Mayor performance en el procesamiento de consultas.	Tamaño limitado para la arquitectura del cubo.
Poco tiempo de cálculo realizado en el momento	No puede acceder a los datos que no estén en el cubo.
Puede escribir sobre la base de datos.	No puede explotar el paralelismo en las bases de datos.
Posibilita hacer cálculos más complicados.	-----

- ROLAP (OLAP Relacional): son sistemas en los cuales los datos se encuentran almacenados en una base de datos relacional. Típicamente, los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas. Los esquemas

más comunes sobre los que se trabaja son estrella o copo de nieve, aunque es posible trabajar sobre cualquier base de datos relacional.

Tabla 2 - Ventajas y Desventajas de ROLAP

Ventajas	**Desventajas**
Uso total de la seguridad e integridad de la base de datos.	Consultas más lentas.
Escalable para grandes volúmenes.	Construcción cara.
Los datos pueden ser compartidos por aplicaciones SQL.	Los cálculos están limitados a las funciones de la base de datos.
Datos y estructuras dinámicas.	-----

- HOLAP (OLAP Híbrido): estos sistemas mantienen los registros detallados en la base de datos relacional, mientras que los datos resumidos o agregados se almacenan en una base de datos multidimensional separada.

Sistemas OLTP

Son los sistemas conocidos como On-Line Transactional Processing. Estos procesan las transacciones en tiempo real de un negocio. Contienen estructuras de datos optimizadas para la introducción y edición de los datos. Su principal desventaja es que proporciona capacidades muy limitadas para la toma de decisiones. Son una herramienta tecnológica capaz de soportar el procesamiento, administración y mantenimiento diario de transacciones generadas por los negocios de una compañía a nivel corporativo para ofrecer altos niveles de disponibilidad, seguridad y confiabilidad.

OLTP es una gran herramienta, pero hay cuestiones que se deben tener en cuenta ya que pueden suponer un problema, como ser la seguridad, los costos económicos o los de tiempo.

- Seguridad: OLTP tiene como una de sus ventajas la seguridad, pero su alta disponibilidad hace que sea más susceptible a intrusos y hackers.
- Simplicidad: Otra de las ventajas es que hace que las cosas sean más simples para la empresa. Entre sus habilidades se destacan la reducción de documentación, previsiones de ingresos, previsiones de gastos de forma rápida y precisa, entre otros.

- Eficiencia: En cuanto a eficiencia las OLTP se destacan por dos características. La primera es que amplía la base de consumidores para una organización y la segunda es que los procesos individuales se ejecutan mucho más rápido.

Se podría decir que estos sistemas definen el comportamiento operacional de un entorno operacional de gestión:

- Altas, bajas, modificaciones, consultas.
- Consultas rápidas y encuestas.
- Poco volumen de información.
- Transacciones rápidas.
- Gran volumen de concurrencia.

Diferencias entre OLTP y Data Warehouse

Tabla 3 - Diferencias entre OLTP y Data Warehouse

Sistema Operacional (OLTP)	**Almacén de Datos (DW)**
Almacena datos actuales	Almacena datos históricos
Almacena datos de detalle	Almacena datos de detalle y datos agregados a distintos niveles
Bases de datos medianas (100MB – 1GB)	Bases de datos grandes (100GB – 1TB)
Los datos son dinámicos (actualizables)	Los datos son estáticos
Los procesos (transaccionales) son repetitivos	Los procesos no son previsibles
El número de transacciones es elevado	El número de transacciones es bajo o medio
Tiempo de respuesta pequeño (segundos)	Tiempo de respuesta variable (segundos –horas)
Dedicado al procesamiento de transacciones	Dedicado al análisis de datos
Orientado a los procesos de la organización	Orientado a la información relevante
Soporta decisiones diarias	Soporta decisiones estratégicas
Sirve a muchos usuarios (administrativos)	Sirve a técnicos de dirección

Diferencias entre OLTP y OLAP

OLAP es un término que utilizan los diseñadores de base de datos para describir un enfoque dimensional al proceso de información.

Una base de datos dimensional está optimizada para la recuperación y el análisis de datos. Cualquier dato nuevo que se cargue en la base de datos se suele actualizar por lotes, de diversas fuentes.

En cambio, en los sistemas OLPT tienden a organizarse datos alrededor de procesos específicos. Definen el comportamiento habitual de un entorno operacional de gestión y ejecutan las operaciones del día.

Tabla 4 - Diferencias entre OLTP y OLAP

Características	**OLTP**	**OLAP**
Tamaño BD	Gigabytes	Gigabytes a Terabytes
Origen de datos	Interno	Interno y Externo
Actualización	Actual	Histórico
Consultas	Predecible	Ad hoc
Actividad	Operacional	Analítico

SELECCIÓN DE DATOS MEDIANTE MUESTREO

Al hablar de métodos de muestreo nos referimos al conjunto de técnicas estadísticas que estudian la forma de seleccionar una muestra lo suficientemente representativa de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un error medible y acotable. A partir de la muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas estimadores, que se convierten en variables aleatorias al considerar la variabilidad de las muestras. Los errores se cuantifican para medir la precisión de éstos. La metodología que permite inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas previamente elegidas por métodos de muestreo formales, se denomina inferencia estadística.

Es muy importante tener en cuenta que para medir el grado de representatividad de la muestra es necesario utilizar muestreo probabilístico, es decir cuando puede establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar, esto es, cuando la selección de muestras constituya un fenómeno aleatorio probabilizable. Dicha selección se verificará en condiciones de azar, siendo susceptible de medida la incertidumbre derivada de la misma. Esto permitirá medir los errores cometidos en el proceso de muestreo.

Existen varios tipos de muestreo, dependiendo de que la población estadística sea finita o infinita. Considerando solamente el muestreo en poblaciones finitas, a la población inicial que se desea investigar se denomina población objetivo, pero el muestreo de toda la población objetivo no siempre es posible debido a diferentes problemas que no permiten obtener información de algunos de sus elementos, con lo que la población que realmente es objeto de estudio o población investigada no coincide con la población objetivo.

Por otro lado, para seleccionar la muestra, necesitaremos un listado de unidades de muestreo denominado marco que teóricamente debiera coincidir con la población objetivo. Un marco será más adecuado cuanto mejor cubra la población objetivo, es decir, cuanto menor sea el error de cobertura. Pero en los marcos son inevitables las desactualizaciones, las omisiones de algunas unidades, las duplicaciones de otras y la presencia de unidades extrañas y otras impurezas que obligan a su depuración.

Asimismo, también sería una meta que al eliminar del marco las unidades de las que no se puede obtener información se obtuviera la población investigada.

Tipos de Muestreo

- **Muestreo Aleatorio Simple:** Cualquier instancia tiene la misma probabilidad de ser extraída en la muestra. Dos versiones, con reemplazo y sin reemplazo.
- **Muestreo Aleatorio Estratificado:** El objetivo de este muestreo es obtener una muestra balanceada con suficientes elementos de todos los estratos, o grupos. Una versión simple es realizar un muestreo aleatorio simple sin reemplazamiento de cada estrato hasta obtener los n elementos de ese estrato. Si no hay suficientes elementos en un estrato podemos utilizar en estos casos muestreo aleatorio simple con reemplazamiento (sobremuestreo).
- **Muestreo de Grupos:** El muestreo de grupos consiste en elegir sólo elementos de unos grupos. El objetivo de este muestreo es generalmente descartar ciertos grupos que, por diversas razones, pueden impedir la obtención de buenos modelos.
- **Muestreo Exhaustivo:** Para los atributos numéricos (normalizados) se genera al azar un valor en el intervalo posible; para los atributos nominales se genera al azar un valor entre los posibles. Con esto obtenemos una instancia ficticia y buscamos la instancia real más similar a la ficticia. Se repite este proceso hasta tener n instancias. el objetivo de este método es cubrir completamente el espacio de instancias.

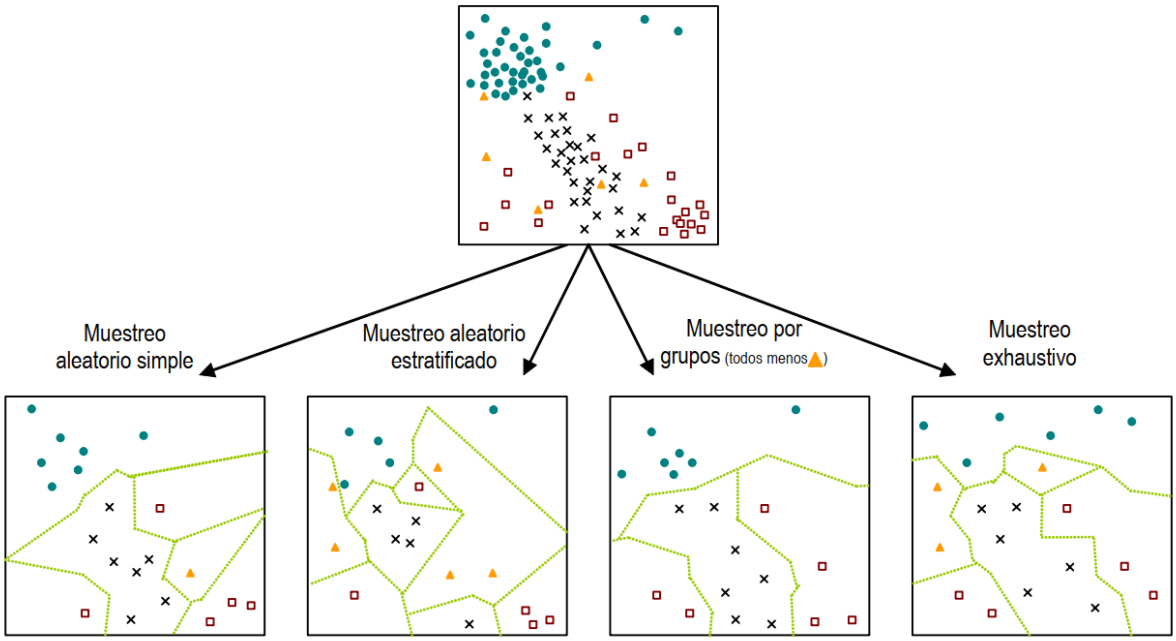

Figura 3 - Tipos de Muestreo

CONCLUSIÓN

Hay muchos factores que determinan la utilidad de los datos como la exactitud, la integridad, la consistencia, la actualidad. Los datos tienen que ser de calidad para satisfacer el propósito previsto. Por ello, el preprocesamiento inicial es crucial en el proceso de minería de datos, y de ahí la importancia de entender la fase de selección de datos.

En la fase de selección se integran y recopilan los datos, se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas, se identifican y seleccionan las variables relevantes en los datos y se aplican las técnicas de muestro adecuadas. Todo ello se facilita disponiendo de un almacén de datos con la información en formato común y sin inconsistencias.

En esta fase, se pretende realizar un análisis para determinar qué se debe hacer con los datos de poca relevancia e innecesarios o datos que no se ajustan al comportamiento normal de la mayoría, datos faltantes o perdidos, a fin de eliminar todo aquello considerado ruido para el dominio en estudio y asegurar la calidad del conocimiento que se vaya a generar.

BIBLIOGRAFÍA

[1] Daniel Santín Gonzáles, César Pérez. Minería de Datos, Técnicas y Herramientas. España: Editorial Paraninfo, 1 ene. 2007.

[2] J. Hernández Orallo y C. Ferri Ramírez, “Minería de Datos y Extracción de Conocimiento De Base de Datos”, [En línea]. Disponible: http://users.dsic.upv.es/~jorallo/docent/doctorat/t2a.pdf.