# TRABAJO PRÁCTICO NRO 1

MINERIA DE DATOS

FECHA DE ENTREGA: 31 DE OCTUBRE DE 2024

AUTOR: CADER FERNANDA
PROFESOR: REA OSCAR

MINERÍA DE DATOS

Clase 6 – Práctico Evaluativo

Fecha de inicio: 24 de Octubre de 2024

Fecha de finalización: 31 de Octubre de 2024

Consignas:

De acuerdo a las clases vistas:

1. ¿A que necesidades debe responder la minería de datos?

2. ¿Cómo respondió la tecnología de bases de datos para mejorar el manejos

de grandes volúmenes de datos? Desarrolle.

3. ¿Que se entiende cuando hablamos de KDD? ¿Cómo se estructura?

4. ¿Qué se entiende por procesamiento transaccional y analítico?

5. Explique que se entinde por integración de datos. Mencione ejemplos

Materia: Desarrollo de programas de procesamiento de Datos – InSET 2

6. ¿Por qué es importante la discretización y la numerización?

7. ¿Qué se entiende por vista minable?

8. ¿Para que usamos técnicas de muestreo?

Desarrollo

## 1. ¿A qué necesidades debe responder la minería de datos?

La minería de datos debe responder a varias necesidades críticas en el contexto empresarial y de investigación. Entre ellas se incluyen:

- **Descubrimiento de patrones**: La minería de datos busca identificar patrones significativos y relaciones ocultas en grandes volúmenes de datos. Esto permite a las organizaciones entender mejor su comportamiento y tendencias.
- **Predicción y clasificación**: La minería de datos ayuda a predecir resultados futuros basándose en datos históricos. Por ejemplo, puede clasificar clientes en segmentos según sus comportamientos de compra, lo que permite una mejor personalización de las estrategias de marketing.
- **Toma de decisiones informada**: Las técnicas de minería de datos facilitan la toma de decisiones al proporcionar análisis basados en datos concretos, en lugar de suposiciones. Esto es especialmente importante en el contexto de los almacenes de datos, donde la información histórica se utiliza para la toma de decisiones estratégicas.
- **Identificación de anomalías**: La minería de datos permite detectar valores atípicos que pueden indicar fraudes, errores o comportamientos inusuales. Esta capacidad es fundamental para la prevención de riesgos y el aseguramiento de la calidad.
- **Optimización de procesos**: Al analizar datos sobre el rendimiento de procesos internos, las organizaciones pueden identificar áreas de mejora y optimizar sus operaciones para aumentar la eficiencia y reducir costos.

**2. ¿Cómo respondió la tecnología de bases de datos para mejorar el manejo de grandes volúmenes de datos? Desarrolle.**

La tecnología de bases de datos ha evolucionado significativamente para manejar grandes volúmenes de datos, respondiendo a la necesidad de eficiencia y efectividad en el almacenamiento y procesamiento de información. Las respuestas incluyen:

- **Desarrollo de bases de datos no relacionales**: Las bases de datos NoSQL (como MongoDB y Cassandra) han surgido para gestionar datos no estructurados o semi-estructurados, permitiendo el almacenamiento de información en formatos más flexibles. Esto es crucial para manejar datos masivos que no encajan bien en un modelo relacional tradicional.
- **Escalabilidad horizontal**: Las tecnologías modernas permiten la escalabilidad horizontal, donde se pueden añadir más servidores para distribuir la carga y manejar más datos, a diferencia de la escalabilidad vertical, que se limita a mejorar un solo servidor.
- **Almacenes de datos**: La creación de almacenes de datos se presenta como una solución centralizada para la integración de datos provenientes de diversas fuentes. Esto permite a las organizaciones consolidar grandes volúmenes de datos históricos y optimizar el análisis mediante la estructuración multidimensional.
- **OLAP (Procesamiento Analítico en Línea)**: Las tecnologías OLAP permiten realizar análisis complejos de datos rápidamente. A través de esquemas multidimensionales, los usuarios pueden explorar grandes conjuntos de datos de manera interactiva, facilitando informes y análisis en tiempo real.
- **Técnicas de optimización de consultas**: Se han desarrollado múltiples estrategias para mejorar la eficiencia de las consultas en bases de datos,

incluyendo índices, materialización de vistas y optimización de algoritmos de búsqueda, lo que permite un acceso más rápido a grandes volúmenes de datos.

**3. ¿Qué se entiende cuando hablamos de KDD? ¿Cómo se estructura?**

KDD, que significa "Knowledge Discovery in Databases" (Descubrimiento de Conocimiento en Bases de Datos), es el proceso de identificar patrones y extraer conocimiento útil a partir de grandes volúmenes de datos. Se estructura en varias etapas:

1. **Selección de datos**: Implica identificar y recopilar datos relevantes de diversas fuentes. Esto puede incluir bases de datos, archivos de registros, o datos de redes sociales.
2. **Preprocesamiento de datos**: Esta fase abarca la limpieza y transformación de datos, eliminando ruido y datos irrelevantes, y manejando valores perdidos o inconsistencias para asegurar la calidad de los datos.
3. **Transformación de datos**: Los datos son transformados para ser más adecuados para el análisis. Esto puede incluir la normalización, discretización y agregación de datos.
4. **Minado de datos**: Es la fase en la que se aplican técnicas de minería de datos para extraer patrones y modelos de los datos. Se utilizan algoritmos de aprendizaje automático, análisis estadístico y técnicas de inteligencia artificial.
5. **Evaluación e interpretación**: Los patrones descubiertos son evaluados para determinar su relevancia y utilidad. Esto implica interpretar los resultados en el contexto de la pregunta de investigación o el problema de negocio.

6. **Presentación del conocimiento**: Los hallazgos se presentan de manera comprensible a las partes interesadas, utilizando visualizaciones y reportes que facilitan la toma de decisiones.

**4. ¿Qué se entiende por procesamiento transaccional y analítico?**

El procesamiento transaccional y analítico son dos tipos distintos de procesamiento de datos:

- **Procesamiento transaccional (OLTP)**: Se refiere al manejo de transacciones en tiempo real que involucran operaciones como inserciones, actualizaciones y eliminaciones en bases de datos. OLTP se enfoca en la eficiencia y la rapidez de las transacciones, garantizando la integridad y consistencia de los datos. Por ejemplo, en un sistema de gestión de ventas, cada venta se registra como una transacción.
- **Procesamiento analítico (OLAP)**: Este tipo de procesamiento se centra en la consulta y análisis de grandes volúmenes de datos, permitiendo a los usuarios realizar análisis complejos y obtener información sobre tendencias y patrones. OLAP se utiliza para generar informes y análisis históricos que apoyan la toma de decisiones estratégicas. Por ejemplo, se pueden realizar análisis de ventas por región o por producto durante períodos específicos.

Ambos tipos de procesamiento son complementarios: el procesamiento transaccional se utiliza para gestionar la operación diaria de la organización, mientras que el procesamiento analítico permite la exploración y análisis de los datos acumulados para obtener conocimiento estratégico.

**5. Explique qué se entiende por integración de datos. Mencione ejemplos.**

La integración de datos se refiere al proceso de combinar datos de diferentes fuentes para ofrecer una vista unificada y coherente de la información. Este proceso es crucial para los almacenes de datos, donde la información proviene de múltiples sistemas y debe consolidarse para análisis.

Ejemplos de integración de datos incluyen:

- **Integración de bases de datos heterogéneas**: Un sistema puede recoger datos de varias bases de datos relacionales, archivos planos, y fuentes en la nube. Por ejemplo, una empresa puede integrar datos de su sistema de ventas, su ERP y su sistema de gestión de relaciones con clientes (CRM).
- **ETL (Extract, Transform, Load)**: Este proceso implica extraer datos de varias fuentes, transformarlos para cumplir con un formato y estructura específicos, y cargarlos en un almacén de datos. Por ejemplo, se pueden extraer datos de ventas y de inventario, transformarlos para un formato común y luego cargarlos en un almacén de datos para análisis.
- **Uso de APIs**: Las interfaces de programación de aplicaciones (APIs) permiten la integración de datos en tiempo real desde diferentes plataformas. Por ejemplo, una aplicación de marketing puede integrar datos de usuarios desde redes sociales mediante APIs para analizar el comportamiento del cliente.

**6. ¿Por qué es importante la discretización y la numerización?**

La discretización y la numerización son procesos cruciales en la minería de datos que permiten transformar datos continuos en formatos más manejables y analizables:

- **Discretización**: Consiste en convertir datos continuos en intervalos o categorías discretas. Esto es importante porque muchos algoritmos de minería de datos funcionan mejor con datos categóricos. Por ejemplo, en lugar de tener una variable continua de ingresos, se puede discretizar en categorías como "bajo", "medio" y "alto". Esto facilita la identificación de patrones y relaciones en los datos.
- **Numerización**: Se refiere a convertir datos categóricos en valores numéricos. Esto permite que los algoritmos de aprendizaje automático procesen los datos más eficientemente. Por ejemplo, en lugar de representar géneros como "masculino" y "femenino", se pueden codificar como 0 y 1, respectivamente.

Ambos procesos son esenciales para mejorar la calidad de los datos y facilitar su análisis, ya que permiten una mejor interpretación y modelado de los mismos.

**7. ¿Qué se entiende por vista minable?**

Una vista minable se refiere a una representación de datos que ha sido transformada y estructurada de tal manera que permite la aplicación efectiva de técnicas de minería de datos. Esta vista optimiza los datos para el análisis, facilitando el descubrimiento de patrones y la generación de conocimiento.

- **Características de una vista minable**:
    - **Estructura adecuada**: Los datos se organizan en un formato que se ajusta a los requisitos de los algoritmos de minería, como tablas de hechos y dimensiones en un almacén de datos.
    - **Consolidación de información**: Incluye información relevante y preprocesada que elimina ruidos y datos innecesarios, lo que mejora la precisión de los análisis.

- **Accesibilidad**: Facilita el acceso a datos históricos y actuales de manera eficiente, permitiendo a los analistas realizar consultas y análisis sin complicaciones.

Las vistas minables son esenciales para realizar análisis efectivos y obtener conocimientos valiosos a partir de los datos.

**8. ¿Para qué usamos técnicas de muestreo?**

Las técnicas de muestreo se utilizan para seleccionar una muestra representativa de un conjunto de datos más grande, lo que permite realizar análisis y obtener conclusiones sin tener que procesar todo el conjunto de datos. Esto es especialmente útil en situaciones donde el volumen de datos es tan grande que procesarlo en su totalidad sería impráctico o costoso.

- **Ventajas del muestreo**:
  - **Eficiencia**: Reduce el tiempo y los recursos necesarios para realizar análisis, ya que se trabaja con un subconjunto de datos.
  - **Costos**: Disminuye los costos asociados al almacenamiento y procesamiento de grandes volúmenes de datos.
  - **Facilitación de análisis exploratorios**: Permite realizar pruebas preliminares y análisis exploratorios sin el compromiso de procesar toda la base de datos.
- **Ejemplos de técnicas de muestreo**:
  - **Muestreo aleatorio**: Se seleccionan elementos al azar del conjunto de datos, asegurando que cada elemento tenga la misma probabilidad de ser elegido.

- **Muestreo estratificado**: Se divide el conjunto de datos en grupos (estratos) y se selecciona una muestra de cada grupo, asegurando que se representen adecuadamente todas las subpoblaciones.
- **Muestreo sistemático**: Se seleccionan elementos a intervalos regulares a lo largo del conjunto de datos, como cada décimo elemento.

El muestreo es una herramienta poderosa que permite realizar análisis significativos y tomar decisiones informadas sin la necesidad de procesar conjuntos de datos completos.