INSTITUTO POLITÉCNICO NACIONAL

**ESCUELA SUPERIOR DE CÓMPUTO**

**DATA MINING (MINERÍA DE DATOS)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Semestre escolar: | 2023-2 (Febrero-Junio 2023) |  | Grupo: | 3CV15 |

|  |  |
| --- | --- |
| Nombre: | Mauro Sampayo Hernández |

**Artículo:**

Sahu, Hemlata; Shrma, Shalini & Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*. Volume 1, Issue 3.

**Actividades a realizar con el artículo:**

1. **Definiciones:**

|  |  |
| --- | --- |
| **Palabra** | **Definición** |
| **Minería de datos** | Es un proceso para la extracción de información implícita y conocimiento que es potencialmente útil y que las personas no conocen de antemano, y dicha extracción se realiza sobre datos incompletos, ruidosos, borrosos y aleatorios.  Es el uso de técnicas automatizadas de análisis de datos para descubrir relaciones previamente no detectadas entre los elementos de los datos. La Minería de Datos involucra a menudo el análisis de datos almacenados en *data warehouses*. |
| **KDD (Knowledge Discovery in Databases)** | Es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos almacenados en Bases de Datos. |

|  |
| --- |
| Diferencia entre minería de datos y el análisis de datos tradicional |
| La diferencia esencial entre la *Minería de Datos* y *el análisis de datos tradicional* es que, la *Minería de Datos* extrae información y descubre conocimientos con la premisa de no tener una suposición clara. |

1. **Describa cada una de las etapas del KDD.**

|  |  |
| --- | --- |
| **Etapa** | **Descripción** |
| **Limpieza de datos** | Se remueven o eliminan los datos con ruido y los datos irrelevantes que se encuentran en la colección de datos. |
| **Integración de datos** | Diversas fuentes de información, usualmente heterogéneos, son combinadas dentro de una fuente de datos en común. |
| **Selección de datos** | Se decide cuales son los datos relevantes para el análisis y se recuperan. |
| **Transformación de datos** | También se le conoce como “Consolidación de Datos”. Los datos seleccionados son transformados en formas apropiadas para el proceso de minería.  En esta etapa, la información que fue seleccionada es transformada de una forma apropiada para el proceso de minería. |
| **Minería de datos** | Es el paso crucial, en el cual se aplican técnicas inteligentes para extraer patrones potencialmente útiles. |
| **Evaluación de patrones** | Se identifican los patrones estrictamente interesantes que representan el conocimiento basado en las medidas proporcionadas. |
| **Representación del conocimiento** | Es la fase final, en la que el conocimiento descubierto es presentado al usuario de manera visual, por medio de técnicas de visualización que ayuden a los usuarios a entender e interpretar los resultados del proceso de la Minería de Datos. |

1. **Tareas de la minería de datos:**

|  |  |
| --- | --- |
| **Técnica** | **Descripción** |
| **Agrupamiento (Clustering)** | Consiste en descubrir grupos y estructuras en un conjunto de datos los cuales sean de alguna u otra forma “similares”, sin hacer uso de estructuras conocidas en los datos.  Es una técnica utilizada para colocar elementos de datos en grupos relacionados sin conocimiento previo de las definiciones de cada grupo. |
| **Clasificación** | Consiste en la generalización de una estructura conocida para aplicarla a nuevos datos. Es una forma de analizar datos para generar modelos que describan clases de datos. |
| **Regresión** | Consiste en tratar de encontrar una función que modele un conjunto de datos con el menor error posible.  La *Regresión* toma un conjunto de datos numéricos y desarrolla una fórmula matemática que se ajuste a los datos dados. Una vez hecho esto, se pueden usar los resultados obtenidos para predecir comportamientos futuros, insertando nuevos datos de entrada sobre la fórmula desarrollada para obtener una predicción    La principal limitación de esta técnica es que solo funciona apropiadamente con datos cuantitativos continuos (peso, velocidad, edad, etc.), por lo que no funcionará aplicándola con datos categóricos con un orden no significativo (color, nombre, género, etc.) |
| **Reglas de asociación** | Consiste en la búsqueda de relaciones entre variables. Por ejemplo, un supermercado podría recopilar datos sobre los hábitos de compra de sus clientes y determinar que productos fueron comprados juntos frecuentemente, y usar dicha información con propósitos de marketing. |

1. **Explicación del proceso general de la minería de datos**

|  |  |
| --- | --- |
|  | 1. **Data preparation:** Consiste en dos procesos:  * **Data collection:** Es la recolección de datos que pueden provenir de transacciones existentes, sistemas de procesamiento o de data warehouses. Estos datos pueden tener cierto grado de “polución”, esto quiere decir que puede tener datos inconsistentes o faltantes. * **Data collation:** Se eliminan los datos con ruido y los datos inconsistentes de los datos recolectados.  1. **Data Mining:** Hace uso de herramientas de minado y técnicas pata encontrar patrones y tendencias sobre la colección de datos. 2. **Information expression:** Hace uso de la visualización y de tecnologías de expresión de comunicación y conocimiento, para mostrar los resultados del proceso de la Minería de Datos; y de esta manera facilitar la precisión y eficacia de la toma de decisiones. 3. **Analysis and decisión-making:** Se analizan los resultados del proceso de Minería de Datos y se ajustan a las estrategias de la toma de decisiones, combinándolas con la situación actual. |

1. **Explicación de cada una de las capas de la arquitectura**

|  |  |
| --- | --- |
|  | **Data Layer (Capa de Datos):**  Puede ser una Base de Datos o un sistema de data warehouse Es una interfaz para todas las fuentes de datos. Los resultados de la Minería de Datos son almacenados en esta capa, para que estos puedan ser presentados a los usuarios finales en la forma de reportes o en otros tipos de visualización de datos. |
| **Data mining application layer:**  Se utiliza para recuperar datos de una Base de Datos. En esta se pueden también llevar a cabo rutinas de transformación de datos a algún formato deseado. |
| **Front-end (Interfaz de usuario):**  Provee al usuario final una interfaz de usuario intuitiva y amigable con la que pueda interactuar con el sistema de minado de datos, y pueda también visualizar los resultados de este proceso. |

1. **Técnicas de minería de datos:**

|  |  |
| --- | --- |
| **Técnica** | **Descripción** |
| **Árboles de decisión** | Hace uso de un árbol de decisión como modelo predictivo que relaciona las observaciones hechas sobre un elemento para llegar a una conclusión acerca del valor objetivo del atributo.  Cada *nodo interno (nodo no hoja)* denota una prueba en un atributo, cada *rama* representa un resultado de la prueba y cada *nodo terminal (nodo hoja)* tiene una etiqueta de clase. |
| **Sistema de soporte de decisiones** | Es un sistema de información basado en cómputo que soporta actividades de toma de decisiones de negocios u organizaciones. Sirven a la gestión, operaciones y niveles de organización y ayudan a la toma de decisiones las cuales pueden cambiar rápidamente y no estar especificadas de antemano. |
| **Redes neuronales** | Se utiliza para la clasificación, agrupamiento (clustering), minería de características, predicción y reconocimiento de patrones. Este método imita la estructura neuronal de los animales, basado en el modelo M-P y la regla de aprendizaje de Hebbien, por lo que en esencia es una matriz distribuida estructurada. |
| **Agrupamiento k-medias** | Es un algoritmo de Minería de datos y Machine Lerning usado para agrupar observaciones dentro de grupos de observaciones relacionadas, sin ningún conocimiento previo de las mismas. Es comúnmente utilizado en medicina, imágenes, biometría y campos relacionados. |

1. **Explique seis aplicaciones de la minería de datos**

|  |  |
| --- | --- |
| 1) | **Marketing/Retail.** La Minería de Datos ayuda a las compañías de marketing a construir modelos basados en datos históricos, para predecir cuál será la respuesta del público sobre nuevas campañas de marketing. A través de esta predicción, los especialistas de marketing pueden tener un enfoque más informados para vender productos rentables a clientes potenciales. |
| 2) | **Marketing/Retail.** Por medio del Análisis de Canasta de Mercado, las tiendas y supermercados pueden conocer que productos son comprados con frecuencia juntos por los clientes, y así poder tener un arreglo de productos más apropiados.  distintos patrones que podría ser efectivo en la vigilancia. |
| 3) | **Finanzas/Bancos.** La Minería de Datos puede brindar información a las instituciones financieras, relacionadas a prestamos e informes de crédito. Por medio de la construcción de modelos, usando como base la información financiera de clientes anteriores con características comunes, se puede estimar que tan riesgoso es el otorgarle un crédito a un cliente que cuenta con cierta situación financiera. |
| 4) | **Finanzas/Bancos.** La Minería de Datos puede ayudar a los bancos a detectar transacciones fraudulentas con tarjetas de crédito, y así evitar que los propietarios de dichas tarjetas pierdan su crédito. |
| 5) | **Manufacturación.** Aplicando la Minería de Datos sobre los datos de la Ingeniería Operativa, los fabricantes pueden detectar equipos defectuosos y determinar parámetros de control óptimos. |
| 6) | **Gobierno.** La Minería de Datos puede ayudar a las agencias gubernamentales con el análisis de los registros de transacciones financieras, para poder construir patrones que puedan detectar actividad criminal o lavado de dinero. |

1. **Desventajas y desventajas de la minería de datos**

|  |  |
| --- | --- |
| **Ventajas** | **Desventajas** |
| En el marketing, se pude usar para construir modelos que permitan entender el comportamiento de los clientes de una empresa, y así elaborar campañas de marketing exitosas o aumentar las ventas de ciertos productos. | **Problemas de privacidad.** Muchas empresas recopilan información de sus clientes de muchas maneras diferentes, para poder entender sus tendencias y comportamientos. Sin embargo, la información que estas poseen podría ser filtrada o vendida a terceros, los cuál podría resultar en que estos datos sean utilizados de manera poco ética. |
| En las finanzas, se puede usar para construir modelos que puedan determinar si un cliente es apto para recibir un crédito o no, basado en los historiales bancarios de clientes anteriores; así como también detectar posibles transacciones fraudulentas con tarjetas de crédito, y evitar así que los propietarios de estas no pierdan su crédito. | Problemas de Seguridad. Las empresas poseen información sobre sus empleados y clientes, incluida información sensible tales como números de seguridad social y datos de tarjetas de crédito. Está información es muy atractiva para los hackers informáticos, quienes pueden cometer delitos tales como el robo de tarjetas de crédito o de identidad con el uso de está información, pudiendo así afectar de gran manera a los propietarios de dicha información si es que su información es hackeada y robada. |
| En el área de facturación, se puede usar para detectar equipos defectuosos y estableces parámetros de control óptimo para manufacturar productos de alta calidad. | Mal uso de la información/información inexacta. La información recabada por medio de la Minería de Datos podría llegar a ser explotada por personas o compañías con propósitos no éticos o discriminatorios, llegando afectar así la toma de decisiones y causando consecuencias graves. |
| En las agencias gubernamentales, se puede usar para analizar registros de transacciones financieras, y crear modelos que puedan detectar el lavado de dinero o actividades criminales. |  |

1. **Retos de la minería de datos**

|  |
| --- |
| * Escalabilidad * Dimensionalidad * Manejo de datos Complejos y Heterogéneos * Calidad de los datos * Propiedad y Distribución de los Datos * Preservación de la Privacidad * Streaming de datos |

1. **Futuro de la minería de datos**

|  |
| --- |
| En los últimos años, la Minería de Datos se ha ido consolidando como una de las principales diciplinas de la Ciencia Computacional, teniendo un amplio impacto industrial. Sin lugar a duda, la investigación sobre la Minería de Datos va a continuar e incluso aumentar en las próximas décadas, apuntando a la transformación de datos históricos en conocimiento.    A pesar de que la Minería de Datos aún se encuentra en sus inicios, una amplia gama de industrias, como la venta al por menor, la atención a la salud, y la aeroespacial; están utilizando herramientas y técnicas de Minería de Datos para poder aprovechar los datos históricos, y así ayudar a los analistas a detectar patrones, tendencias, hechos significativos, excepciones y anomalías, que de otro modo podrían pasar desapercibidas. |

1. **Conclusiones**

|  |
| --- |
| La Minería de Datos, es posiblemente una de las herramientas más útiles e incluso indispensables de la Ciencia de la Computación, ya que está es la base para la generación del conocimiento a partir del análisis de datos masivos que en un principio podría parecer que no tienen mucho sentido, pero que pueden aportar información muy útil, que puede derivar en el éxito de modelos para marketing, detección de fraudes, toma de decisiones, etc. |