EJERCICIO DE CLASE INTRODUCCIÓN (ARTÍCULOS)

Grupo: 3CV19 **Fecha:** Lunes 23 de agosto de 2021

Equipo No. 1

Nombres de los participantes:

RIVERA PAREDES FERNANDO DANIEL

CASTRO CRUCES JORGE EDUARDO

LOPEZ GARCIA FELIPE DE JESUS

REYES PÉREZ EDWARD DANIEL

BECERRIL HERNANDEZ ALDO

Actividades a realizar con el artículo:

Moreno Salinas, José Gerardo. (2017). Científico de datos: codificando el valor oculto e intangible de los datos. Revista Digital Universitaria. Vol. 18, Núm. 7, septiembre-octubre 2017.

1) **Definiciones:**

Palabra	Definición
Big data	El gran cúmulo de datos compuesto por diferentes tipos, estructuras y relaciones de datos, que a su vez tienen veloces tasas de generación y dispersión, y el procesarlos con tecnologías convencionales para su posterior análisis
Huella digital	Información o datos públicos de los usuarios en Internet
Minería de datos	Extracción y procesamiento de conocimientos útiles alojados en grandes cantidades de datos
Científico de datos	Aquella persona que posee las capacidades, habilidades, el pensamiento creativo y la tecnología para procesar, analizar y visualizar los grandes volúmenes de datos
Visualización de datos	Un medio para explorar y dar una representación a los datos cuantitativos recurriendo al campo de la creatividad y del arte
Internet de Contenido	Es toda la información creada por los seres humanos para aumentar el conocimiento sobre temas particulares
Internet de las personas	Son todos los datos relacionados con la interacción social
Internet de las cosas	Son todos los objetos físicos conectados a la red. Todas las cosas que tienen una identificación única y una presencia en una estructura similar a Internet
Internet de la ubicación	Refiere a todos los datos que tienen una dimensión espacial

2) **Cantidad de información que producen en 60 segundos las siguientes plataformas, según el reporte del año 2017.**

Plataforma	Cantidad-descripción
Youtube	500 horas de videos subidos
Email	149, 513 emails enviados
Facebook	3.3 millones de posts
Google	3.8 millones de búsquedas
Instagram	65, 972 fotos subidas
Twitter	448, 800 tweets
Wordpress	1, 440 posts
WhatsApp	29 millones de mensajes enviados

3) ¿Por qué es importante valorar los datos, no subestimarlos? Ejemplifique.

Porque los datos pueden utilizarse de diferentes maneras para generar, como lo hace twitter, que es una red orientada a la comunicación de noticias, eso permite introducir publicidad por parte de twitter, consumida por los usuarios y así poder pagar los gastos de mantenimiento de la plataforma. O como el caso de Netflix que su éxito se debe en mayor parte a su algoritmo de recomendación, el cual tiene como base el historial de reproducción de los usuarios.

4) Identificar las características de cada uno de los siguientes aspectos:

Científico de datos	Economía del conocimiento	Áreas de conocimiento del científico de datos	Actividades del científico de datos
- Colectar, etiquetar, limpiar y organizar los datos. - Construir y modelar los datos. - El modelado de datos para patrones. - Refinar algoritmos	- Usa información y conocimiento para generar valor. - Poseen las capacidades, las habilidades, el pensamiento creativo y la tecnología para procesar, analizar y visualizar las grandes bases de datos. - Entienden en su máxima expresión los datos y sus relaciones.	- Big Data para procesar datos. - Minería de datos para analizar e identificar relaciones ocultas, patrones y tendencias. - Visualización de datos para explicar y socializar mejor la información obtenida.	- Mejorar los productos y servicios de las organizaciones. - Analizar y visualizar grandes bases de datos.

5) Estimación de los tiempos en el desarrollo de actividades del científico de datos. Comente de dónde obtuvieron estos datos.

Porcentaje	Actividad
51%	Colectar, etiquetar, limpiar y organizar datos.
19%	Construir y modelar los datos.
10%	El modelado de datos para patrones.
9%	Refinar algoritmos.
8%	Otras actividades.

Nota:Los datos anteriores se obtuvieron de una encuesta que realizó la compañía Crowd Flower (2017) a 179 científicos de datos de todo el mundo, con esto se logró identificar la distribución de actividades que toman mayor tiempo para poder ser realizadas.

6) Describa y ejemplifique en qué consisten las Áreas que maneja el científico de datos.

a. Big Data

Puede describirse como un conjunto de datos masivos y complejos con una tasa de generación de datos veloz, provenientes de distintas fuentes. Cuya importancia radica en brindar respuestas completas y con mayor fiabilidad de los datos gracias a la cantidad de información que ofrece, lo que puede ayudar en un sentido empresarial a tomar una mejor decisión

b. Minería de datos

Es un proceso que consiste en el análisis de grandes volúmenes de datos que permiten extraer características, relaciones e interacciones entre los datos, concretar una conclusión y tener la capacidad de predecir resultados.

c. Visualización

La visualización de datos es la presentación de datos en formato ilustrado o gráfico. Permite ver la analítica presentada de forma visual, de modo que puedan captar conceptos difíciles o identificar patrones.

7) Características de las relaciones

a. Relación 1. Big data y minería de datos

Existe una estrecha relación entre ambas áreas, ya que los algoritmos y modelos de entrenamiento y prueba desarrollados por el área de minería de datos deberán ser implementados en los grandes cúmulos de datos (Big Data), sobre todo cuando se tiene una amplia serie de datos históricos.

b. Relación 2. Big Data y visualización de datos

La relación entre el área de Big Data y la visualización de datos es la que busca definir la mejor interpretación y visualización de grandes cúmulos de datos y sus relaciones, de forma que al usuario le resulte más fácil entenderlos.

c. Relación 3. Minería de datos y visualización de datos

La minería y la visualización de datos también pueden trabajar en una dimensión donde no necesariamente se procesen un gran cúmulo de datos (Big Data), se pueden implementar proyectos para una cantidad mesurada de datos, donde se apliquen algoritmos y con éstos obtengamos un producto.

d. Relación 1, 2 y 3. Big Data, minería y visualización de datos

Por ejemplo, ¿podemos encontrar algún valor en la enorme cantidad de información registrada por las máquinas dispensadoras de refrescos enlatados? Considerando que en los Estados Unidos existen 15 000 máquinas dispensadoras y que cada una puede despachar 150 latas únicas. En cada transacción se registra una cadena de datos: hora, ubicación y preferencias del usuario (Big Data). Con éstos datos se visualizaron patrones de consumo total y se encontraron altos consumos de bebidas los fines de semana y lo contrario, durante los días entre semana (Visualización de datos). Con toda la explotación de los datos pudieron entender mejor lo que sucede: el analizar los dispensadores individualmente mostró características inesperadas (Minería de datos).

8) Conclusiones

Especializarse como científico de datos, te llevará a explorar al menos una de sus tres áreas principales (Big Data, Minería de datos o Visualización de los datos), que con la llegada de la Web 2.0, los usuarios finales han marcado en diferentes plataformas de internet una Huella Digital, las cuales, empresas interesadas han sabido utilizar a lo largo de los años. El perfil de un científico de datos es muy importante para el tratamiento, análisis, visualización y determinación del valor de los datos, ya que permite que a un usuario final se le pueda brindar un producto o servicio adecuado a sus necesidades, considerando que si los datos no se utilizan de una manera inteligente, podría llevar a una pérdida de capital enorme para las empresas o inclusive la bancarrota.

==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0

Actividades a realizar con el 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.

1) Definiciones:

Palabra	Definición
Minería de datos	Es un proceso de extraer información implícita y conocimiento el cual es potencialmente útil, dicha extracción es de datos incompletos, ruidosos, aleatorios, etc.
KDD	Se refiere a la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos en las bases de datos.
Análisis de datos tradicional	Mejor descrita como la solicitud de una consulta convencional, un informe, o bien, una aplicación de análisis en línea.

2) Describa cada una de las etapas del KDD.

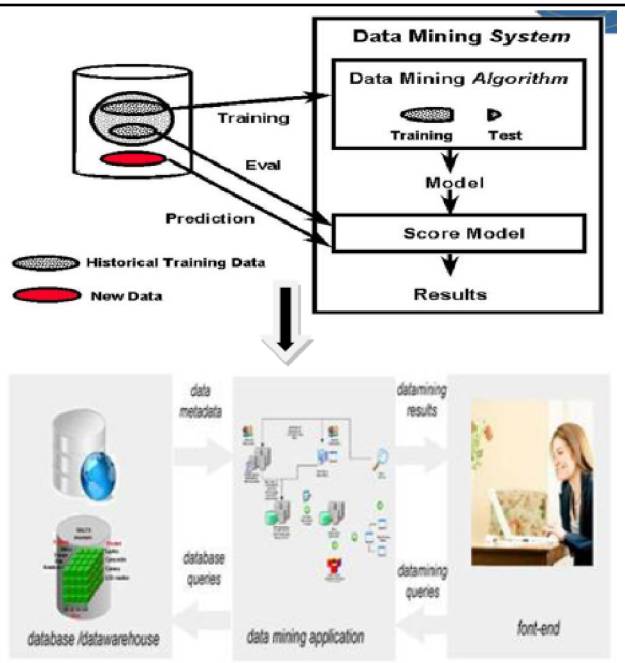
Etapa	Descripción
Limpieza de datos	Es una fase en la que se eliminan los datos de ruido y datos irrelevantes de la colección.
Integración de datos	En esta fase, se pueden combinar múltiples fuentes de datos, a menudo heterogéneas, en una fuente común.
Selección de datos	En este paso, se deciden los datos relevantes para el análisis y se recuperan de la recopilación de datos.
Transformación de datos	Es una fase en la que los datos seleccionados se transforman en formas adecuadas para el procedimiento de minería.
Minería de datos	Es el paso crucial en el que se aplican técnicas inteligentes para extraer patrones potencialmente útiles.
Evaluación de patrones	En este paso, se identifican los patrones estrictamente interesantes que representan el conocimiento sobre la base de las medidas dadas.
Representación del conocimiento	Es la fase final en la que el conocimiento descubierto se representa visualmente para el usuario. Este paso esencial utiliza técnicas de visualización para ayudar a los usuarios a entender e interpretar los resultados de la minería de datos.

3) Tareas de la minería de datos:

Técnica	Descripción
Agrupamiento	El clustering o agrupamiento es el proceso de particionar un conjunto de datos (u objetos) en un conjunto de subclases significativas llamadas grupos (clusters). Un grupo es una colección de objetos de datos que son similares a otros y así pueden ser tratados colectivamente como un grupo.
Clasificación	Se dedica a generalizar la estructura conocida y así aplicarla a los nuevos datos. Es una manera de predecir nuevas instancias a partir de atributos ya existentes de esa instancia.
Regresión	En esta técnica se utilizan modelos matemáticos como fórmulas para poder predecir el futuro de datos cuantitativos continuos así como el peso, velocidad y edad. Sin embargo esto también es una desventaja puesto en tipos de datos categóricos donde el orden no importa, no es posible poder aplicar esta técnica.
Reglas de asociación	Se tiene como objetivo principal el determinar relaciones no evidentes entre atributos categóricos, esta es una tarea de tipo descriptiva. La idea es recopilar cierto tipo de información para al final utilizarla para un fin en específico.

4) Explicación de cada una de las capas de la arquitectura

	Capa de datos: Descrita como un medio donde se encuentran almacenados los datos (Base de Datos, Sistemas de Almacenamientos de Datos) recopilados a través del proceso de la Minería de Datos, los cuales, son presentados al usuario final mediante la Visualización de Datos

	
	Capa de aplicación de la minería de datos: Capa que se encarga de recuperar parte de los datos almacenados en la Capa de Datos, los cuales, se procesan y generan una transformación en los datos mediante algunos algoritmos de Minería de Datos.
	Interfaz de usuario (*front-end*) Capa encargada de proporcionar al usuario final una interfaz intuitiva y amigable, donde generalmente se presentan resultados del proceso de Minería de Datos

5) Técnicas de minería de datos:

Técnica	Descripción
Árboles de decisión	Utilizado en estadísticas, minería de datos y aprendizaje automático, utiliza un árbol de decisiones como modelo predictivo que asigna las observaciones sobre un elemento a conclusiones sobre el valor objetivo del elemento. Los nombres más descriptivos para tales modelos de árboles son árboles de clasificación o árboles de regresión. En estas

	estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan conjunciones de características que conducen a esas etiquetas de clase.
Sistema de soporte de decisiones	Es un sistema de información basado en computadora que respalda las actividades de toma de decisiones empresariales u organizativas. Los DSS sirven a los niveles de gestión, operaciones y planificación de una organización y ayudan a tomar decisiones, que pueden cambiar rápidamente y no fácilmente especificadas de antemano.
Redes neuronales	El método de red neuronal se utiliza para clasificación, agrupamiento, extracción de características, predicción y reconocimiento de patrones. Imita la estructura de las neuronas de los animales, se basa en el modelo M-P y la regla de aprendizaje de Hebbien, por lo que en esencia es una estructura matricial distribuida. A través de la minería de datos de entrenamiento, el método de la red neuronal calcula gradualmente (incluyendo iteraciones repetidas o cálculos acumulativos) los pesos de la red neuronal conectada.
Agrupamiento k-medias	La agrupación en clústeres de K-means es un algoritmo de minería de datos / aprendizaje automático que se utiliza para agrupar observaciones en grupos de observaciones relacionadas sin ningún conocimiento previo de esas relaciones. El algoritmo k-means es una de las técnicas de agrupamiento más simples y se usa comúnmente en imágenes médicas, biometría y campos relacionados.

6) Desventajas y desventajas de la minería de datos

Ventajas	Desventajas
Marketing/Retail: Ayuda a las empresas a hacer modelos predictivos con datos de sus usuarios y mejorar el mercado meta.	Privacy Issues: Debido al crecimiento de internet la información sobre los usuarios está más expuesta siendo contraproducente para la privacidad.
Finance / Banking: Ayuda a los bancos a crear modelos para saber que tan buenos son sus clientes dependiendo de su historial crediticio.	Security issues: Al tener nuestra información en internet se corre el riesgo de una práctica desleal por parte de 3ros, como robo de identidad, acceso a cuentas bancarias, etc.
Manufacturing: Ayuda a hacer revisión de control de calidad dependiendo del historial de productos elaborados.	Misuse of information/inaccurate information: Un mal uso de información puede ser el atacar a un grupo o persona. Además de no usar las métricas y datos correctos puede producir predicciones incorrectas.
Governments: Permite prevenir lavado de dinero o actividades criminales basado en los datos que proveen los bancos de los usuarios al gobierno.	

7) Retos de la minería de datos

- Escalabilidad
- Dimensionalidad
- Datos complejos y heterogéneos
- Calidad de los datos
- Propiedad y distribución de los datos
- Preservación de la privacidad
- Flujo de datos

8) Futuro de la minería de datos

La minería de datos se mantendrá en gran uso por las próximas décadas, debido a que es una herramienta que ha beneficiado mucho sobre todo a los departamentos de marketing ya que los ayuda a conocer mejor a sus usuarios y a así logrando trazar sus estrategias futuras para poder conseguir mayores ganancias.
Ya se tienen varias técnicas para poder aplicar la minería de datos pero es probable que se lleguen a encontrar algunas más debido al gran uso que tiene la minería de datos.

9) Conclusiones

La importancia de la minería de datos en el campo de la investigación es muy significativa para los distintos campos o especializaciones, tales como la medicina, economía, política, computación, entre otros. Minería de datos no solo se trata de realizar una búsqueda o consulta , requiere de un análisis a profundidad que permita integrar de una manera útil cada una de las características que se presentan en los datos de diversos formatos, que nos permitan ver relaciones entre los datos o bien ser prevenidos para futuros eventos.

==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0==0

Describa cinco ejemplos en donde se aplique la minería de datos.

Agricultura
Vigilancia
Agencias de seguridad nacional
Relación cuantitativa estructura-actividad
Análisis de cliente