**INSTITUTO POLITÉCNICO NACIONAL**

**ESCUELA SUPERIOR DE CÓMPUTO**

**DATA MINING (MINERÍA DE DATOS)**

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

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

**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** | Es 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. |
| **Huella digital** | Es el rastro de datos que un usuario deja al usar Internet (preferencias, intereses, deseos de compra, tendencias, frecuencia de consumo, horarios de interacción, redes sociales, ubicación de la conexión, dispositivos utilizados, etc.). |
| **Minería de datos** | Se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. |
| **Científico de datos** | Es aquel que lleva a cabo el análisis de grandes volúmenes de información, para así poder entenderlos e identificar relaciones ocultas, patrones y tendencias en ellos; con el objetivo de tomar decisiones más informadas a la vez que mejoran los productos y servicios de las organizaciones. |
| **Visualización de datos** | Es aquella que ayuda al usuario a examinar una gran cantidad de datos e identificar patrones o tendencias con la ayuda de gráficas o representaciones. Su objetivo es el de traducir de maneras fáciles, eficientes, precisas y decodificadas la información abstracta en representaciones visuales significativas. |
| **Internet de Contenido** | Es toda la información creada por los seres humanos para aumentar el conocimiento sobre temas particulares (páginas web tradicionales, artículos, enciclopedias como Wikipedia, YouTube, libros electrónicos, noticias, etc.). |
| **Internet de las personas** | Son todos los datos relacionados con la interacción social (correo electrónico, Facebook, Twitter, foros, LinkedIn, etc.). |
| **Internet de las cosas** | Son todos los objetos físicos conectados a la red. Son 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 (Google Maps, Google Earth, etc.). |

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

|  |  |
| --- | --- |
| **Plataforma** | **Cantidad-descripción** |
| Youtube | 500 horas de contenido |
| Email | 149,513 e-mails enviados |
| Facebook | 3.3 millones de posts |
| Gooogle | 3.8 millones de búsquedas |
| Instagram | 65,972 fotos subidas |
| Twitter | 448,800 tweets |
| Wordpress | 1,400 posts |
| WhatsApp | 29 millones de mensajes enviados |

1. **¿Por qué es importante valorar los datos, no subestimarlos? Ejemplifique.**

|  |
| --- |
| Se hace mención, que las grandes compañías como Facebook, Google y Twitter, a pesar de gastar grades cantidades de dinero para el mantenimiento de sus sistemas, sus usuarios finales no son quienes pagan este costo directamente, si no que los mismos son objeto de campañas publicitarias de otras compañías, las cuales pagan dichos costos para mostrar sus publicidades a cambio de obtener datos de los usuarios.  Se menciona el ejemplo de Twitter, en el cual se puede llevar a cabo el cálculo del valor monetario que tiene cada cuenta por si sola, basándose en: la cantidad de seguidores de dicha cuenta, los tweets que se han hecho en dicha cuenta, y la velocidad con la que dicha cuenta gana seguidores. Con esto se pudo se pudo determinar, que, por ejemplo, el valor que tiene la cuenta de Donald J. Trump es de 2 mil millones de dólares estadounidenses.  También se menciona el ejemplo de Netflix, el cuál desde su creación, les dio mucha importancia a los datos de sus usuarios, logrando así desarrollar un algoritmo que permitiese recomendar contenido de alto interés para cada uno de sus suscriptores. Dicho algoritmo logró recopilar tantos datos, que comenzó a ser capaz de predecir qué tan exitosos serían sus producciones originales.  Estos ejemplos nos muestran que, los datos de los usuarios son de gran importancia para estas compañías, pues con ellos pueden conocer las tendencias de cada uno de sus usuarios y elaborar estrategias que les permitan construir un modelo de negocios exitoso, con el uso de sus datos. |

1. **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** |
| * “Trabajadores del conocimiento”. * Entender los datos y sus relaciones. * Tomar decisiones bien informadas. * Mejorar productos y servicios. * Pensamiento creativo. | * El centro de producción de la riqueza es el saber y no el capital. | * Estadística * Matemáticas * Programación * Solución de problemas | * Colectar datos * Etiquetar datos * Limpiar datos * Organizar datos * Construir y modelar datos. * Modelar datos para patrones. * Refinar algoritmos. |

1. **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% | Modelado de datos para patrones |
| 9% | Refinar algoritmos |
| 8% | Otras actividades |

Los datos fueron obtenidos de una encuesta realizada por la compañía *Crowd Flower*, en el año 2017, en la que se entrevistaron a 179 científicos de datos de todo el mundo.

1. **Describa y ejemplifique en qué consisten las Áreas que maneja el científico de datos.**
   1. **Big Data**

|  |
| --- |
| Es 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. |

* 1. **Minería de datos**

|  |
| --- |
| Es la extracción de conocimientos de grandes cantidades de datos. Se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. |

* 1. **Visualización**

|  |
| --- |
| Es aquella que ayuda al usuario a examinar una gran cantidad de datos e identificar patrones o tendencias con la ayuda de gráficas o representaciones. Su objetivo es el de traducir de maneras fáciles, eficientes, precisas y decodificadas la información abstracta en representaciones visuales significativas.  La Visualización proviene de las ciencias exactas, las cuales han tenido la necesidad de recurrir al campo de la creatividad y del are, con el propósito de representar con fines estético algún aspecto de la realidad. Es una combinación entre CIENCIA y ARTE. |

1. **Características de las relaciones**
   1. **Relación 1. Big data y minería de datos**

|  |
| --- |
| Los algoritmos y modelos de entrenamiento y prueba desarrollados en el área de Minería de Datos pueden ser aplicados sobre el gran cúmulo de datos (Big Data). |

* 1. **Relación 2. Big Data y visualización de datos**

|  |
| --- |
| 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. |

* 1. **Relación 3. Minería de datos y visualización de datos**

|  |
| --- |
| Se puede implementar la Minería de Datos sobre cantidades de datos mesuradas, en proyectos donde se apliquen algoritmos y con ellos se obtenga un producto. |

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

|  |
| --- |
| Un científico de datos puede interactuar en cualquiera de las tres áreas y en sus respectivas intersecciones, siempre y cuando se posean las habilidades, conocimientos, técnicas y experiencia para dar respuesta a las necesidades de las organizaciones. |

1. **Conclusiones**

|  |
| --- |
| En la actualidad, existe una alta demanda de científicos de datos, ya que estos se encargan de interpretar las grandes cantidades de datos producidos por los usuarios que hacen uso de las diferentes plataformas de Internet, y así, las empresas para las que trabajan puedan tomar decisiones inteligentes para mantener una alta competitividad en el mercado. |

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

1. **Describa cinco ejemplos en donde se aplique la minería de datos. Agregue la fuente de donde obtiene la información.**

|  |
| --- |
| **Marketing:** Se utiliza para mejorar la segmentación del mercado por medio del análisis de las relaciones entre parámetros tales como la edad de los clientes, género, gustos, etc.; para así entender su comportamiento y dirigir campañas de marketing que se adapten a ellos. |
| **Banca:** Los bancos recurren a ella para entender mejor los riesgos de mercado. Es común ver aplicada la *Minería de Datos* en sistemas de antifraude para analizar transacciones, movimientos de tarjetas, patrones de compra y datos financieros de los clientes. |
| **Medicina:** Para diagnósticos más precisos. Analizando el historial médico del paciente, exámenes físicos y patrones de terapias se pueden prescribir tratamientos más efectivos. |
| **Comercio Minorista:** Muchos supermercados analizan los patrones de compra conjunta para identificar asociaciones en productos y así decidir cómo situarlos en los diferentes pasillos y estanterías para maximizar las compras. También se usa para detectar cuáles son las ofertas más valoradas por los clientes. |
| **Radio y Televisión:** Se aplica en tiempo real sobre los registros de audiencia de la TV y la radio, para analizar información referente a las visualizaciones, retransmisiones y programación en los canales. De esta manera se puede determinar que programas emitir en que horarios, entender la conducta y los intereses de su audiencia, y obtener datos valiosos para sus anunciantes para que estos puedan llegar a más clientes potenciales. |

**Fuente:** I. Corporativa, “Descubre Cómo el 'Data Mining' Predecirá Nuestro comportamiento,” Iberdrola. [Online]. Available: https://www.iberdrola.com/innovacion/data-mining-definicion-ejemplos-y-aplicaciones. [Accessed: 25-Feb-2023].

1. **Busque información actualizada de la cantidad de información que producen en 60 segundos las siguientes plataformas. Agregue las fuentes de donde toma la información.**

|  |  |
| --- | --- |
| **Plataforma** | **Cantidad-descripción** |
| Youtube | 500 horas de contenido |
| Email | 231 millones e-mails enviados |
| Facebook | 3.3 millones de posts |
| Gooogle | 5.9 millones de búsquedas |
| Instagram | 66,000 fotos subidas |
| Twitter | 575,000 tweets |
| Wordpress | 1,400 posts |
| WhatsApp | 16 millones de mensajes enviados |
| Otra (Netflix) | 452,000 horas de contenido en streaming visto |

**Fuente:** S. Heitman, “What happens in an Internet minute? [2022 statistics],” LocaliQ, 18-Oct-2022. [Online]. Available: https://localiq.com/blog/what-happens-in-an-internet-minute/. [Accessed: 25-Feb-2023].