LAS CIUDADES INTELIGENTES

WILMER ILLESCAS ESPINOZA / SILVIA LANDÍN ÁLVAREZ / WASHINGTON FIERRO SALTOS

Editorial UTMACH

REDES 2017
COLECCIÓN EDITORIAL

Las ciudades inteligentes

Wilmer Illescas Espinoza
Silvia Landín Álvarez
Washington Fierro Saltos
Coordinadores

Primera edición en español, 2018

Este texto ha sido sometido a un proceso de evaluación por pares externos con base en la normativa editorial de la UTMACH

Ediciones UTMACH

Gestión de proyectos editoriales universitarios

148 pag; 22X19cm - (Colección REDES 2017)

Título: Las ciudades inteligentes. / Wilmer Illescas Espinoza / Silvia Landín Álvarez / Washington Fierro Saltos (Coordinadores)

ISBN: 978-9942-24-098-9

Publicación digital

Título del libro: Las ciudades inteligentes.

ISBN: 978-9942-24-098-9

Comentarios y sugerencias: editorial@utmachala.edu.ec

Diseño de portada: MZ Diseño Editorial

Diagramación: MZ Diseño Editorial

Diseño y comunicación digital: Jorge Maza Córdova, Ms.

Km. 5 1/2 Vía Machala Pasaje

www.utmachala.edu.ec

Machala - Ecuador

Índice

Capítulo V

Dedicatoria

El presente esfuerzo bibliográfico dedico a aquellas personas que trabajan día a día en el desarrollo de la Provincia de El Oro. Además, a quienes me han sabido guiar con sus acertadas opiniones en la ejecución de mi proyecto de vida.

Wilmer Illescas

A mi esposa y mis hijos Daniela, Sofía y Matías, este libro y toda mi vida.

Washington Fierro Saltos

A mis hijos Jaden Santistevan y Ayleen Arias que son parte de mi vida y motores de mi desarrollo profesional.

Silvia Landin

A Dios, a mi hermosa familia y a toda esa comunidad científica, académica y lectora, expongo este proyecto a todos los interesados en encontrar nuevos cambios en beneficio de la sociedad.

Jussen Facuy

Introducción

Actualmente, lograr que una determinada localidad, ciudad o provincia se considere una ciudad inteligente es el objetivo de una gobernanza eficiente. Esfuerzos en optimizar el uso adecuado de las energías, crecer en armonía con la naturaleza, y desarrollar su urbanismo de manera sostenible, forma parte de los objetivos de desempeño de gobiernos responsables. Para lograrlo, se necesita que las tendencias tecnológicas y la ciudadanía logren trabajar simbióticamente. El presente documento presenta a la ciudadanía los últimos avances de la tecnología informática y su integración en el desarrollo de las ciudades inteligentes.

Nuestro mundo gira y girará en torno a los datos, el progreso y la innovación de los datos están en el centro de una ciudad inteligente y de la nueva economía del conocimiento. Desde esta perspectiva el Capítulo 1 denominado la Ciencia de los Datos, analiza brevemente el concepto de ciudad inteligente o "Smart City", caracterizada por la aplicación de las Tecnologías de la Información y Comunicación (TIC) y el Internet de la cosas; en un segundo momento se define ampliamente la tecnología del Big Data, partiendo de las características y dimensiones relevantes; de las técnicas, algoritmos y herramientas de la Minería de Datos, para la extracción y descubrimiento de conocimiento útil a partir de grandes volúmenes de datos.

En el capítulo denominado “Sistema de Soporte a las Decisiones con la Web Semántica”, presenta los últimos avances para tomar decisiones en situaciones automatizables. La ventaja que brinda la Web Semántica es que la información generada tiene la capacidad de ser reutilizable porque se genera con código que es capaz de interoperar entre sistemas. Esto contribuye al ahorro energético y disminuir la obsolescencia del hardware, al no tener que volver a generar los datos. Adicionalmente, se contribuye a la afinación de la precisión de lo que recomiendan los sistemas informáticos, al no tener que depender, en su totalidad, de los ingresos de datos manuales.

El problema de los residuos e-garbage; los seres humanos consumimos recursos y desechamos aquello que no es útil, denominados residuos o comúnmente “basura”, desde perspectivas amplias y diversas, más eficiente, con el uso de sensores de medición, la distribución de pequeños agentes recolectores, el uso de grandes estaciones de desperdicios y recolectores de gran tamaño, estaríamos mejorando un sistema que por muchos años no ha sido eficiente.

El concepto de ciudad inteligente va asociado a mejorar la calidad de vida de los ciudadanos en diferentes áreas, tales como, participación ciudadana, movilidad, seguridad, polución ambiental, recolección de desechos, ecología, entre otros. Un área de especial interés en este capítulo es la de movilidad. Al respecto, se está hablando de problemáticas relacionadas al sistema de transporte público, el aumento de tráfico, el estacionamiento, la contaminación, entre otros. Dar solución a estos problemas es un desafío para los gobiernos locales, encargados de mantener la calidad de vida de sus ciudadanos concentrándose especialmente en la movilidad urbana.

El apartado ciudad inteligente presenta una conceptualización de una ciudad inteligente y la ciudadanía inteligente que permite el acceso a las TIC´s como herramientas de interacción de las personas con las diferentes áreas sociales, económicas, políticas, culturales, educativo entre otras que fomentan la responsabilidad, ética y seguridad en el uso de

la tecnología. Se aplicó un estudio en Ecuador para determinar la ciudadanía inteligente y su grado de compromiso con el uso de la tecnología. Se evidencia cómo ha evolucionado el uso de la tecnología en el ciudadano inteligente y actualmente cuáles son las oportunidades que la tecnología brinda, el reto, uso y modernización de las nuevas tecnologías. Se contribuye con información sobre la satisfacción y necesidades de las empresas sobre el uso de las TIC´s y la capacidad del personal en su uso. Por lo cual se concluye que el uso de la tecnología mejora la calidad de vida en el ciudadano.

Finalmente, se plantea la siguiente pregunta ¿cuándo parará la evolución tecnológica? ¿Cuál es el motor que impulsa su exponencial desarrollo? ¿Son sus prácticas sostenibles a largo plazo? A lo cual nos atrevemos a decir, que está en nuestras manos el tomar lo mejor que el campo científico nos provee, para innovar y desarrollarnos en perfecta armonía con la naturaleza, respetando los sueños y aspiraciones de las futuras generaciones, y hacer de éste un mundo mejor.

01 Capítulo

La Ciencia de los datos

Xavier Ochoa Chehab; Washington Fierro Saltos; Jonathan Cárdenas Benavides

Resumen

El Big Data se constituye en una tecnología emergente y juega un papel importante en el desarrollo de una ciudad inteligente (Smart City), al permitir gestionar y analizar eficientemente grandes volúmenes de datos generados en tiempo real, para explotar su valor, con el fin de ofrecer soluciones a las necesidades de los ciudadanos. El concepto de Smart City, propone un enfoque amplio e integrado de la inteligencia humana, la inteligencia colectiva y la inteligencia artificial para contribuir a mejorar la calidad de vida y bienestar social, pero el reto fundamental está en saber aprove-

Xavier Ochoa Chehab: Profesor principal en la Facultad de Ingeniería en Electricidad y Computación de la Escuela Superior Politécnica del Litoral (ESPOL) de Guayaquil, Ecuador. Grado de PhD en Ingeniería por la Universidad Católica de Lovaina (KULeuven). Actualmente, se desenvuelve como coordinador del Grupo de Investigación en Tecnologías para la Enseñanza y el Aprendizaje del Centro de Tecnologías de Información (CTI) de ESPOL.

Washington Fierro Saltos: Licenciado e Ingeniero en Sistemas e Informática, Magíster en Comunicación y Tecnologías Educativas en el ILCE de México. Docente titular de la Universidad Estatal de Bolívar.

Jonathan Cárdenas Benavides: Ingeniero en Informática y Ciencias de la Computación, Especialista en Redes de Comunicación de Datos, Magister en Informática Empresarial. Docente de la Universidad en Estatal de Bolívar.

char el gran volumen de datos e información que proporcionará una sociedad conectada a una variedad de dispositivos masivos de comunicación, con el fin de obtener patrones de comportamiento que permitan diseñar soluciones más eficientes a las ciudades.

Nuestro mundo gira y girará en torno a los datos, el progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento en los datos recopilados de manera oportuna y en una forma escalable. Los datos son tan valiosos si les podemos dar sentido, ellos revelan nuestros sentimientos, actitudes, conexiones sociales e intenciones, en definitiva los datos están en el centro de una cuidad inteligente y de la nueva economía del conocimiento.

Desde esta perspectiva el libro en el Capítulo 1, analiza brevemente el concepto de ciudad inteligente o "Smart City", caracterizada por la aplicación de las Tecnologías de la Información y Comunicación (TIC) y el Internet de la cosas; en un segundo momento se define ampliamente la tecnología del Big Data, partiendo de las características y dimensiones relevantes; de las técnicas, algoritmos y herramientas de la Minería de Datos, para la extracción y descubrimiento de conocimiento útil a partir de grandes volúmenes de datos.

Datos y Smart Cities

La visión de la Ciencia de los Datos y las Smart Cities

El siglo XXI estará caracterizado por el siglo de las ciudades inteligentes o Smart Cities, donde las urbes se constituirán en mega ciudades y en el centro de la actividad social, económica, cultural y artística. Según los últimos informes de la ONU, en el año 2050 estas ciudades concentrarán al 70% de la población mundial, esto significa que progresivamente el mundo y las ciudades dejarán de ser rurales para convertirse en urbanas, hasta tal punto que en los próximos 25 años pasaremos de los 7.300 millones a los 9.500 millones de personas habitando el planeta (GICI, 2015).

Ante esta realidad las principales ciudades del mundo buscan ser espacios más tecnológicos, verdes y transitables en el que el ciudadano es el eje del cambio y el principal beneficiado del nuevo paradigma urbano, para aquello la aplicación extensiva e intensiva de las Tecnologías de la Información y la Comunicación (TICs)[1] en los servicios públicos de: Gestión del suministro de consumo de energía o de agua, la mejora del transporte y la movilidad, la seguridad ciudadana y la protección civil, la creación de un entorno favorable para los negocios y la actividad económica de alto valor añadido, el gobierno de la ciudad, la transparencia y participación ciudadana, constituyen factores claves de transformaciones de la ciudad tradicional a una Smart City.

Smart City o ciudad inteligente es un concepto aún muy complejo de definir por sus características dinámicas y multidimensionales, como son la sostenibilidad e inclusión social y las nuevas tecnologías de Internet, pues la "inteligencia" de una ciudad puede emerger de sus ciudadanos, organizaciones o de la tecnología. Según (GICI, 2015) una Smart City "es aquella ciudad que mediante la incorporación de tecnologías, procesos y servicios innovadores, garantizan su sostenibilidad energética, medioambiental, económica y social, para mejorar la calidad de vida de las personas y favorecer la actividad empresarial y laboral". En esta misma línea (IBM, 2014), define a una "Ciudad inteligente" como la utilización inteligente de tecnología avanzada para detectar, examinar, procesar e integrar grandes volúmenes de información útil para dar respuestas a las necesidades diarias de los ciudadanos, incluyendo la seguridad, sistemas de transporte público y medio ambiente, salud pública y actividades industriales y comerciales entre otras.

Desde esta perspectiva las Tecnologías de la Información y la Comunicación se convierten en un eje transformador de la sociedad humana no sólo como un eje transversal sino como

[1] TIC. Las tecnologías de información y comunicación, son un conjunto de herramientas y programas informáticos, que sirven para facilitar la emisión, acceso, gestión y tratamiento de la información.

un eje directo, pues las ciudades inteligentes requerirán de nuevos dispositivos, de sensores, de redes de comunicaciones, de capacidad de almacenamiento y de procesamiento, de plataformas de gestión de servicios que permitan mejorar la prestación de los servicios de la ciudad, como la energía, el agua, el transporte, los residuos, el comercio, el turismo o el gobierno.

Como se puede visibilizar los servicios que puede prestar una Smart City se asocian a la integración de tecnologías en las que se pueden apoyar para hacer "ciudades inteligentes". Esta integración parte de las dimensiones de la inteligencia humana, a la inteligencia colectiva así como a la inteligencia artificial de los componentes físicos de la ciudad. La inteligencia de la ciudad se crea al interconectar redes digitales de telecomunicación (nervios), la inteligencia integrada en sistemas (cerebro), sensores y componentes físicos (órganos sensoriales), así como herramientas de software para la extracción del conocimiento y características cognitivas (William, 2007). Estas tecnologías abarcan desde las redes de comunicaciones por línea y por radio hasta los sistemas M2M (Machine to machine) de Internet de las cosas[2], que permiten la gestión de sensores y actuadores a lo largo y ancho de la ciudad. La gran capacidad de adquisición de datos a través de sensores desplegados por toda la ciudad, requerirán de una capacidad de almacenamiento y procesamiento, siendo para ello adecuado la aplicación de las tecnologías emergentes como el Big Data.

En una ciudad inteligente será preciso capturar, almacenar, procesar y analizar la gran cantidad de datos procedentes de fuentes muy diversas para poder transformarlos en conocimiento útil para la toma de decisiones y anticiparnos a lo que va a pasar, como por ejemplo definir las rutas óptimas en tiempo real para la recogida de basura, anticiparse a los atascos de un tráfico o realizar análisis de sentimientos para conocer el sentir de los ciudadanos en cada momento para

[2] Internet de las cosas. Por sus siglas en inglés (IoT), es un sistema de dispositivos computacionales interconectados con objetos y cosas cotidianos al Internet.

involucrarlos en las decisiones, son algunos casos de aplicación del Big Data. En esta misma línea Barbosa (2016), destaca algunas de las áreas en las cuales la magia del Big Data puede contribuir a mejorar los servicios en una ciudad inteligente:

- Seguridad ciudadana: Se podría mejorar la eficiencia y eficacia de las actuaciones de los cuerpos de seguridad a través de la correlación de toda la información procedente de los distintos sistemas instalados en la ciudad: desde cámaras de videovigilancia, geolocalización de coches de policía y bomberos, sensores de movilidad o de alertas, detectores de humo y fuego.
- Movilidad urbana: Mediante la captura y gestión de datos procedentes de cámaras repartidas por toda la ciudad, sensores instalados en autobuses, información meteorológica, datos originados en las redes sociales (como por ejemplo la organización de una manifestación a través de Twitter) se podría conseguir por ejemplo anticiparse a los atascos y tomar decisiones en tiempo real para redirigir la ruta de autobuses o interactuar con la red de semáforos e informar al ciudadano de la situación del tráfico.
- Gestión del agua: A través del análisis de los datos ofrecidos por una red de sensores de presión, PH y turbidez del agua ubicados en los sistemas de abastecimiento y saneamiento así como cámaras de vigilancia de plantas potabilizadoras sería posible detectar fugas y controlar la calidad del agua en todo momento.
- Energía y eficiencia energética: Gracias a los datos se puede lograr una eficiencia entre la capacidad de generación de energía y el consumo. Para aquello, tiene especial relevancia la integración de fuentes de energía renovable en la red eléctrica inteligente o Smart Grids.
- Residuos urbanos: El control de los contenedores (nivel de llenado) mediante sistemas de sensorización, el diseño de rutas eficientes de recogida de residuos en base a la información capturada sobre el estado de los contenedores, el control de las flotas de vehículos dedicados a esta recogida y sistemas de quejas en tiempo

real, se configuran como servicios esenciales para una gestión inteligente de los residuos.

- Análisis de sentimiento del ciudadano: Posibilidad de conocer la opinión de los ciudadanos y turistas sobre la ciudad a través del análisis en tiempo real de datos procedentes de distintas redes sociales, webs, call centers, etc., para conocer cuáles son los aspectos prioritarios que están demandando y poder responder a peticiones de forma inmediata.

En este entorno, el Big Data se convierte en una herramienta fundamental y en la piedra angular de innovación de las Smart Cities, pues asistiremos a una verdadera explosión de datos generados fundamentalmente por las interacciones de las personas en las redes sociales y de los miles de sensores y dispositivos conectados a Internet de las cosas , por lo tanto la exploración y el análisis de estas estructuras de datos a través de diferentes métodos y técnicas de minería de datos, permitirá mostrar nuevas dinámicas de comportamiento en la ciudad y también nuevas dinámicas humanas.

El Big Data

Nuestro mundo gira en torno a los datos, el progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable. Los datos son tan valiosos si les podemos dar sentido, los datos revelan nuestros sentimientos, actitudes, conexiones sociales, intenciones, pueden revelar lo que hicimos, lo que hacemos y lo que haremos, en definitiva los datos están en el centro de la sociedad y de la nueva economía del conocimiento. Los datos por sí sólo no bastan, requieren de arquitecturas y técnicas innovadoras para extraer conocimiento relevante, siendo el Big Data y la minería de datos con KDD[3]

[3] KDD. De las siglas Knowledge Discovery in Databases o descubrimiento de conocimiento en bases de datos, es un proceso metodológico que sirve para identificar un "modelo "o patrones válidos, útiles y entendibles para descubrir conocimiento a partir de los datos.

en una herramienta fundamental para encontrar, extraer, refinar, distribuir y monetizar esos datos. Es decir se trata de convertir los datos en información, conocimiento y decisiones.

De acuerdo a Manyika, y otros (2011) del "McKinsey Global Institute", los datos a gran escala o Big Data se definen como "las bases de datos cuyo tamaño está más allá de la capacidad que tiene el software tradicional para su manejo en términos de captura, almacenamiento, gestión y análisis". Además indican que el número de datos actualmente es inmanejable, esto ha implicado que las unidades de medida de información digital hayan crecido en zettabytes (ZB) una medida igual a un billón de gigabytes (GB). Por su parte, Gartner (2014), define el Big Data como "un gran volumen, velocidad o variedad de información que demanda formas costeables e innovadoras de procesamiento de información que permitan ideas extendidas, toma de decisiones y automatización del proceso".

El Big Data sigue creciendo diariamente, diversos estudios muestran que en zonas urbanas los seres humanos estamos generando más datos que los proyectados, es así que en el reporte de EdTech en el año 2013, menciona que desde el origen de la humanidad hasta el año 2003 se generaron 5 billones de exabytes de datos. Esa misma cantidad de información se generó en el año 2011 cada 2 días. O más aún, en el año 2013 esos 5 billones de exabytes de datos se están generando cada 10 minutos. Se estima que para el año 2020 cada individuo creará 1,7 megabytes de información nueva por segundo y además el universo de datos pasará de 4.4 zettabytes que existen actualmente a 44 zettabytes (44 billones de gigabytes) y también tendremos más de 6.100 millones de usuarios de smartphones en el mundo y al menos la tercera parte de todos los datos pasarán por la nube, una red de servidores conectada mediante el Internet. Las empresas capturan miles de millones de bytes de información sobre sus clientes, proveedores y sus operaciones. Millones de sensores conectados en red están presentes en dispositivos tales como teléfonos móviles, sistemas de detección o redes

sociales. Las personas, bien sea con teléfonos inteligentes (smartphones) o a través de redes sociales estimulan el crecimiento exponencial de la información.

En otro estudio realizado por IBM (2014), menciona que cada día se generan más de un quintillón de bytes (QB), que surgen de fuentes tan diferentes como los datos de clientes, proveedores, operaciones financieros en línea u obtenidos de dispositivos móviles, análisis de redes sociales, ubicación geográfica mediante GPS. En muchos países se gestionan gigantescas bases de datos, que contienen datos de impuestos, censo de población, registros médicos, etc. En tanto que la empresa Cisco (2014), en un estudio realizado entre el 2011 y el 2016, manifiesta que habría entre 19 mil millones de dispositivos conectados a la red, esto es más de 2 por habitante del planeta y el tráfico global de datos móviles alcanzará a 587 exabytes (mil millones de Gibabytes) anuales, esto significaría que los datos móviles crecerán anualmente un 78%; y, el número de dispositivos móviles que están conectados a Internet superarán la población de la tierra.

Toda esa cantidad de datos que se está generando a cada minuto y que sigue creciendo en forma exponencial, está transformando a nuestra sociedad y permitiendo a las organizaciones entender sus intereses, comportamientos y poder ofrecer un mejor servicio con propuestas de valor. Entonces contar con la información adecuada en el momento correcto es una condición indispensable hoy en día para realizar buenos análisis y tomar las mejores decisiones. Por tanto los datos se han convertido en el "nuevo petróleo" de la economía actual, que pueden ser empleados para diversos propósitos con alto beneficio.

La Comisión Europea (2014b), en su informe anual establece cómo a través del análisis con Big Data, cambiarán los estilos y servicios de una sociedad cada vez más digitalizada, es así que se:

- Transformarán las industrias de servicios, mediante la generación de una amplia gama de productos y servicios de información innovadores.

- Aumentarán la productividad de todos los sectores de la economía.
- Mejorarán la investigación y se acelerará la innovación.
- Lograrán reducciones de costos a través de servicios más personalizados.
- Aumentarán la eficiencia en el sector público.

Desde esta perspectiva el Big Data se está convirtiendo en un activo clave para el ámbito humano y la sostenibilidad de las ciudades inteligentes, así, se están desarrollando sistemas de información inteligentes que a partir de sensores electrónicos instalados a pie de calle permiten cambiar la duración de las luces de los semáforos en función de los datos que se recojan en tiempo real o identificar rápidamente picos de inflación de la economía de un país, sobre precios de productos comercializados en Internet (Esteban, 2014). Hoy en día es imposible hablar de Big Data sin mencionar las nuevas técnicas y tecnologías de MapReduce o Hadoop, diseñados para el manejo de información estructurada o semiestructurada con arquitecturas de procesamiento paralelo masivo.

Características del Big Data

Big Data es definido por IBM en términos de tres dimensiones claves que caracterizan a los datos: volumen, velocidad y veracidad, en tanto la empresa SAS identifica dos nuevas dimensiones adicionales a las “3V’s” que son la variabilidad y complejidad. Finalmente en los últimos años se empieza a hablar de las 7Vs donde se integra a las 5V, la visualización de los datos y el valor de los datos.

La dimensión de volumen se refiere a la cantidad masiva de datos que las organizaciones intentan aprovechar en pos de mejorar la toma de decisiones. Los volúmenes de datos continúan aumentado a un ritmo sin precedentes debido a las redes sociales, la movilidad que facilitan las redes inalámbricas / telefonía móvil, y los nuevos servicios de almacenamiento en la nube; se espera que en el año 2020 la cifra del volumen de datos supere 35 zettabytes (ZB). Ahora algunas

empresas están generan terabytes de datos cada hora de cada día del año y su desafío es aprovechar del poder de los datos para crear conocimiento e innovar.

La velocidad se refiere a los datos en movimiento por las constantes interconexiones que realizamos, es decir analiza no solo a la alta frecuencia con la que se crean, procesan y analizan los datos, sino a la necesidad de dar respuesta a la información en tiempo real, dado que los datos son creados cada segundo, estos se vuelven obsoletos rápidamente. Por esta razón, es importante hacer uso de estos datos lo más rápido posible, o sustituirlos por datos más actualizados que se generen en forma ágil.

La veracidad se puede entender como el grado de confianza y fiabilidad sobre los datos a utilizar. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de Big Data. Por lo tanto un alto volumen de información que crece a velocidad muy rápida y basada en datos estructurados o semi-estructurados y provenientes de una gran variedad fuentes, hacen inevitable dudar del grado de veracidad de los mismos.

La dimensión variedad, tiene que ver con la diversidad de los tipos de datos y de sus diferentes fuentes de obtención. Así, los tipos de datos pueden ser estructurados (Base de datos relacionales), semi-estructurados o no estructurados donde sus fuentes podrán provenir de archivos de texto, datos de la web, tweets, sensores de datos, audio, video streaming, archivos de logs, etc. Sin embargo la riqueza de datos aumenta el grado de complejidad tanto en su almacenamiento como en su procesamiento y análisis.

La complejidad o viabilidad se manifiesta en la naturaleza de los datos en sí. Es a la vez estructurado y no estructurado y proveniente de múltiples fuentes, lo que dificulta la vinculación de los datos, el emparejamiento, limpiar y transformar a través de los sistemas

La visualización se refiere al modo en el que los datos son presentados. Una vez que los datos son procesados mediante tablas u hojas de cálculo, necesitamos represen-

tarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar. Para que los datos sean comprendidos existen herramientas de visualización que ayudarán a comprender los datos gráficamente y en la perspectiva contextual.

El valor de los datos está en que sean accionables, es decir, que los responsables de las empresas puedan tomar la mejor decisión en base a datos, es decir el valor de los datos se obtiene cuando estos se transforman en información; este a su vez se convierte en conocimiento, y este en acción o en decisión.

En definitiva, el Big Data es una combinación de estas características que crea una oportunidad para que las ciudades y la sociedad, puedan obtener una ventaja competitiva con sus ecosistemas de datos en un mercado cada vez digitalizado. Probablemente como consecuencia de ello, Moro (2016) predice que en 2025 "todo el mundo estará haciendo Big Data con sus teléfonos móviles con acceso a datos abiertos y a herramientas de creación colaborativa", y también se ilusiona pensar que en un futuro cercano los gobiernos, empresas y agencias de las Naciones Unidas se verán en la necesidad de llegar a un acuerdo de compartición de datos para luchar contra los grandes problemas como: la pobreza, el acceso a los alimentos, las epidemias o el crimen organizado. En este contexto el Big Data es una tendencia ascendente y dominante para una sociedad innovadora e inteligente.

La Ciencia de los datos

El auge del Big Data ha evolucionado a un nuevo concepto de la Ciencia de los Datos y tiene una alta aplicabilidad en ciencias de la salud, marketing, negocios, mercados financieros, transporte, comunicaciones, redes sociales, etc.

El Data Science o la Ciencia de Datos incorpora diferentes elementos y se basa en las técnicas y teorías de muchos campos, incluyendo las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, com-

putación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y la informática de alto rendimiento con el objetivo de extraer el significado de datos y la creación de productos de datos.

Desde esta línea el Data Science está basado en algoritmos aplicados al problema de Big Data e incorpora un nuevo perfil profesional del data scientist en la sociedad. Esta evolución traspasa las fronteras en busca de utilizar todos los datos disponibles y relevantes generados por el ser humano para "extraer conocimiento" a través de encontrar correlaciones, patrones y predicciones aplicando algoritmos complejos, esto permitirá a más individuos imaginar cómo crear valor con la información disponible y dar vida a un diseño movido por la interacción humana con los datos.

La figura del científico de datos, es un término relativamente nuevo que emerge como respuesta al manejo y análisis de los grandes volúmenes de datos. Según Guerrero (2013) uno de los grandes científicos de datos del mundo, define a un Data Scientist, como "una persona con conocimientos en lenguajes de programación, matemáticas, estadística, métodos de optimización; y, que además tiene una experiencia práctica en el análisis de datos reales y en la elaboración de modelos predictivos".

Según un reciente estudio presentado por Delphi Analytics y citado en PowerData (2016), los datos almacenados se acumulan con una tasa de crecimiento anual del 28%, mientras que la fuerza de trabajo correspondiente a analistas de datos tan sólo aumenta en un 5,7% por año. Esto es corroborado por Gartner (2014) quien predice que los próximos años, 4,4 millones de empleos serán creados en torno a Big Data, en tanto que la formación u oferta de analistas de datos es bastante limitada frente a un crecimiento geométrico de demanda en las empresas que están adoptando cada vez más la Ciencia de Datos, sólo en Estados Unidos existe una escasez de 140.000 a 190.000 personas con experiencia analítica y 1,5 millones de gestores y analistas con los conocimientos necesarios para comprender y tomar decisiones basadas en el análisis de grandes datos. La revista Harvard

Business Review en el año 2015, calificó al "Data Scientist" como "la profesión más sexy del siglo XXI" que combinado con la Inteligencia Artificial y bases de datos en la nube, revolucionarán el mundo digital.

Minería de datos y descubrimiento de conocimiento en bases de datos.

La Minería de Datos (MD), se puede definir como el proceso de descubrir conocimiento útil y entendible desde grandes bases de datos por medio de modelos, técnicas y sistemas automatizados para la toma de decisiones.

De acuerdo con Peña (2014), la Minería de Datos es un proceso dedicado a escanear enormes repositorios de datos para generar información y descubrir conocimiento relevante para la toma de decisiones. La Minería de Datos está compuesta por un conjunto de técnicas y algoritmos que sirven para hacer análisis del conjunto de datos, descubrir patrones de datos, organizar la información, definir reglas y estructuras de asociación, estimar elementos desconocidos, clasificar objetos, y desvelar muchos tipos de resultados que no se producen fácilmente; de este modo, los resultados representan un valioso apoyo para la toma de decisiones.

Según Pérez, Caballero, Caro, Rodríguez, & Antequera (2014), la Minería de Datos es una parte importante de un proceso más amplio conocido como "Descubrimiento de conocimiento en bases de datos" (KDD en inglés) y su objetivo principal es la extracción de información oculta de un conjunto de datos. Esto puede ser alcanzado por técnicas de análisis de aprendizaje automático o semiautomático, lo que permite la extracción de patrones desconocidos. Estos pueden ser grupos de registros de datos (análisis clúster), inusuales registros (detección de anomalías) y las dependencias entre datos (asociación reglas). Por lo tanto, los patrones pueden ser vistos como un resumen de los datos de entrada y se pueden utilizar para su posterior análisis y toma de decisiones.

Desde este contexto la Minería de Datos busca encontrar relaciones, correlaciones, dependencias, asociaciones,

modelos, estructuras, tendencias, clases (clústeres), segmentos, los cuales se obtienen de grandes Bases de datos, para aquello utiliza modelos matemáticos, estadísticos y algorítmicos complejos para extraer información y/o patrones como parte del proceso KDD.

Tsai (2013) complementa los anteriores conceptos explicando que la Minería de Datos es un campo interdisciplinario que combina inteligencia artificial, gestión de bases de datos, visualización de datos, aprendizaje automático, algoritmos matemáticos y estadísticos. Esta tecnología ofrece diferentes metodologías para la toma de decisiones, resolución de problemas, el análisis, la planificación, el diagnóstico, la detección, la integración, la prevención, el aprendizaje y la innovación.

Aunque Data Science y Big Data son términos más actuales, desde 1989 se denomina a estas actividades similares como KDD (Knowledge Discovery from Databases) o descubrimiento de conocimiento en bases de datos. El KDD es el proceso completo de extracción de conocimiento a partir de patrones útiles en los datos, por lo tanto la Minería de Datos es sólo una etapa de ese proceso e informalmente se asocia la Minería de Datos con KDD.

Es importante destacar que la Minería de Datos, puede procesar distintos tipos de datos de diferentes fuentes como: Bases de datos relacionales, Bodegas de datos, Bases de datos transaccionales, Bases de datos orientadas a objetos y simbólicas, Bases de datos espaciales y/o temporales: telefonía móvil, Bases de datos de documentos (Text mining), Bases de datos multimedia (Imágenes, videos, sonidos), La World Wide Web (Web mining), Grandes volúmenes de datos no estructurados (Big Data, Social Big Data), entre otras.

Proceso de la Minería de datos

Dentro de un proceso de descubrimiento de conocimiento en bases de datos, uno de los aspectos fundamentales es considerar al usuario, ya que es él quien determina el dominio de la aplicación del problema y decide cómo y qué datos

se utilizaran en el proceso. Por lo tanto en un proceso global KDD descrito en la Figura 1.1, el usuario experto estará inmerso en cada una de las siguientes etapas (Natek & Zwilling, 2014):

a) Selección de datos: Consiste en buscar el objetivo del proceso de minería, identificando los datos que han de ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar la tarea. Es decir lo primero que se tiene q hacer es saber qué es lo que se desea obtener y determinar cuáles son los datos que facilitarán la información para lograr la meta.

b) Preparación de los datos: En este paso se limpia los datos anómalos, datos incompletos, el ruido y datos inconsistentes. Es decir, es este proceso se debe definir técnicas y estrategias para corregir errores en el conjunto de datos seleccionado, tratar la información faltante y unificar formatos.

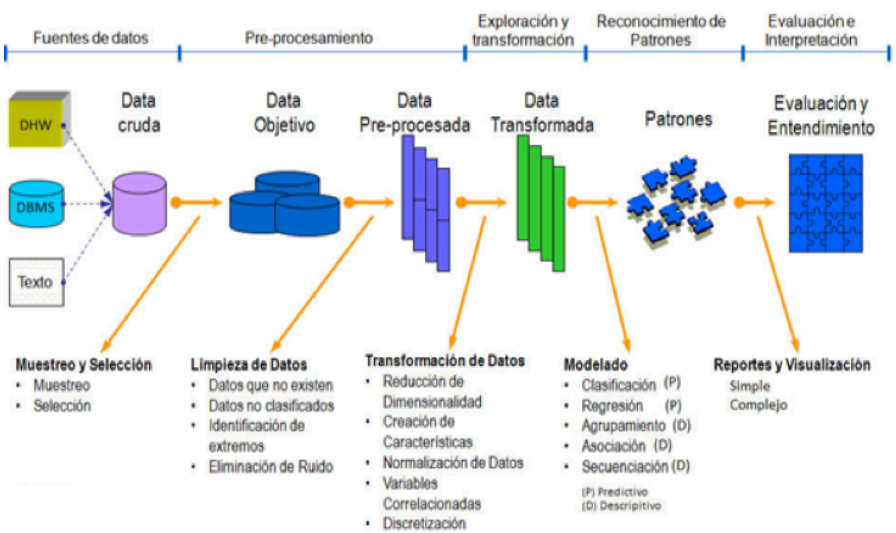
c) Transformación y almacenamiento de los datos: En este punto se transforman los datos a un formato más adecuado empleando reducción o agrupación de los datos en las características de interés para el posterior modelado. Existen diferentes métodos de transformación de variables continuas a discretas que se pueden agrupar por distintas aproximaciones: métodos locales, métodos globales, métodos supervisados y no supervisados.

d) Aplicación de algoritmos de Data Mining: Sobre los datos recogidos y preparados, se aplica una técnica adecuada de minería de datos según la hipótesis planteada y el análisis que se quiera hacer. Esto incluye la selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, regresión, etc. La selección de él o de los algoritmos a utilizar. Las técnicas seleccionadas permitirán generar modelos de minería de datos, y con ello descubrir patrones de información implícitos en los datos.

e) Interpretación y evaluación de patrones encontrados: Se identifican patrones interesantes que representan nuevos conocimientos y apoyándose en los expertos del negocio

para ver si se pueden tomar acciones con estos resultados. Para interpretarlos, es necesario visualizarlos de diversas formas, validando los patrones y modelos de datos, documentando los procedimientos y consideraciones de manera que se generen propuestas de valor para el negocio.

Figura 1 Proceso de minería de datos con KDD



Fuente: Molina Neyra & Murakami de la Cruz

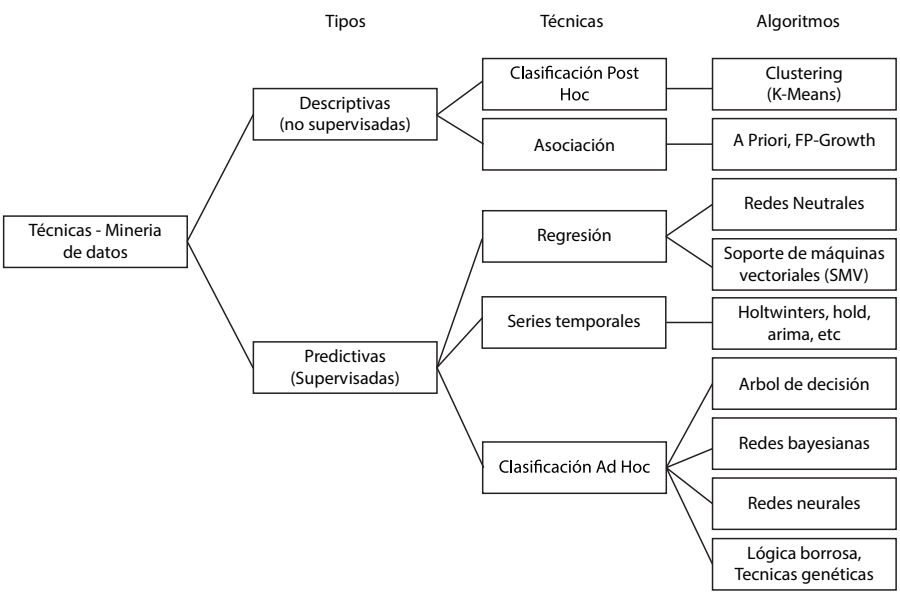
Técnicas de Minería de Datos

Una técnica constituye el enfoque conceptual para extraer la información de los datos mediante varios algoritmos. Cada algoritmo representa en la práctica, la manera de desarrollar una determinada técnica y modelo matemático y estadístico paso a paso, de forma que es preciso un entendimiento de alto nivel de los algoritmos para saber cuál es la técnica más apropiada para cada problema, así como es preciso también entender los parámetros y las características de los algoritmos para preparar los datos a analizar. En esta misma caracterización (Muhammad, Mohamudally, & Babajee, 2013), explican que el éxito de la investigación en el

campo de la Minería de Datos, está en el desarrollo y mejora de los algoritmos existentes, así como en la evaluación de los conocimientos descubiertos como un proceso de un solo paso o de múltiples pasos, desde diferentes enfoques como el álgebra relacional y la teoría de la información, entre otros.

Desde este breve contexto las técnicas de Minería de Datos se clasifican en dos grandes categorías: predictivas y descriptivas, como se muestra en la Figura 2:

Figura 2 Técnicas de la Minería de Datos



Fuente: Adaptado de Molina López & García Herrero

Técnicas descriptivas. Tratan de proporcionar información entre las relaciones de los datos y sus características encontrando patrones interpretables, entre ellas tenemos:

a) Agrupamiento o Clustering. López & Herrero (2012), manifiesta que un clustering permite la identificación de tipologías o grupos, donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de

valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. Es decir, agrupan a los datos bajo diferentes métodos y criterios. Los algoritmos más usadas son:

- Clustering Numérico (K-medias). Es un algoritmo que se caracteriza en especificar por adelantado cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan k elementos aleatoriamente, que representaran el centro o media de cada clúster. A continuación cada una de las instancias, ejemplos, es asignada al centro del clúster más cercano de acuerdo con la distancia Euclidea que le separa de él. Para cada uno de los clúster así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos clusters. Finalmente se repite el proceso completo con los nuevos centros de los clusters. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración.
- Clustering conceptual (COBWEB). Este algoritmo organiza incrementalmente las observaciones del conjunto de datos en un árbol de clasificación. Cada nodo en un árbol de clasificación representa una clase (concepto) y está etiquetado por un concepto probabilístico que resume las distribuciones de atributos y valores de los objetos clasificados bajo el nodo. Este árbol de clasificación se puede utilizar para predecir los atributos que faltan o la clase de un nuevo objeto. La descripción probabilística de este algoritmo se basa en los conceptos de predicibilidad y previsibilidad.
- Clustering Probabilístico (EM). Es un tipo de algoritmo de agrupamiento que permite estimar grupos de instancias y parámetros de distribuciones de probabilidad. Así pues, asigna a cada instancia una distribución de probabilidad de pertenencia a cada grupo. El algoritmo puede decidir

cuantos clusters crear basándose en una validación cruzada o se le puede especificar cuantos ha de generar.

b) Reglas de Asociación. Las asociaciones pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos como, por ejemplo, el comercial ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes y constituyen un pilar básico en la concepción de las ofertas y ventas cruzadas, así como del "merchandising". Por lo general esta forma de extracción de conocimiento se fundamenta en técnicas estadísticas, como los análisis de correlación y de variación. Uno de los algoritmos más utilizado es el algoritmo "A priori".

Técnicas predictivas. Cómo su nombre lo indica, intentan predecir o responder preguntas futuras con base en el comportamiento pasado. Este proceso intenta determinar los valores de una o varias variables a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. A continuación se muestran brevemente un conjunto de técnicas que específicamente sirven para la predicción:

a) Regresión. Persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística). Las regresiones se pueden utilizar por ejemplo para predecir comportamiento de la demanda futura, utilizando las ventas pasadas (Valcárcel, 2004). Entre los algoritmos más importantes se destacan:

- Redes Neuronales. Las redes neuronales son una nueva forma de examinar la información y tienen la capacidad de detectar y aprender patrones complejos y características dentro un conjunto de datos. Su comportamiento es parecido a nuestro cerebro aprendiendo de la experiencia y situaciones pasadas y aplicando dicho conocimiento a la resolución de problemas nuevos. Este aprendizaje se obtiene como resultado del adiestramiento ("training") y éste permite la sencillez y la potencia de

adaptación y evolución ante una realidad cambiante y muy dinámica. Una vez adiestradas las redes de neuronas pueden hacer previsiones, clasificaciones y segmentación. Presentan además, una eficiencia y fiabilidad similar a los métodos estadísticos y sistemas expertos

- Máquinas de soporte vectorial (SMV). Son un conjunto de algoritmos de aprendizaje supervisado. Dado un conjunto de datos de entrenamiento, podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

b) Series Temporales. Es el conocimiento de una variable a través del tiempo, para que a partir de ese conocimiento y con el supuesto de que no se producirán cambios poder realizar predicciones. También se llama Serie de Tiempo a un conjunto de mediciones de cierto fenómeno o experimento registradas secuencialmente en el tiempo (Ortíz, 2015). Entre los algoritmos más importantes se destaca:

- Holt-Winters. Utilizado habitualmente por muchas compañías para pronosticar la demanda a corto plazo cuando los datos de venta contienen tendencias y patrones estacionales de un modo subyacente. Esta técnica se basa en la atenuación de los valores de la serie de tiempo, obteniendo el promedio de estos de manera exponencial; es decir, los datos se ponderan dando un mayor peso a las observaciones más recientes y uno menor a las más antiguas (Coghlan, 2015).
- ARIMA. Significa modelos autorregresivos integrados de medias móviles, es un modelo estadístico para series temporales que se basa en la dependencia existente entre los datos, esto es, que cada observación en un momento dado es modelada en función de los valores anteriores. Es decir permite describir un valor como

una función lineal de datos anteriores y errores debidos al azar, además, puede incluir un componente cíclico o estacional. Box y Jenkins recomiendan como mínimo 50 observaciones en la serie temporal.

- ETS. Métodos de suavización exponencial, es un algoritmo que genera predicciones de la demanda de un producto en un periodo puntual. Holt-Winters y Box-Jenkins son dos de las técnicas más relevantes y útiles para realizar análisis y pronósticos de negocio, debido a su facilidad de uso y a sus resultados inmediatos.

c) Clasificación y predicción. Son dos tipos de análisis de datos, que se utilizan para clasificar datos y predecir tendencias. La clasificación de datos predice clases de etiquetas mientras la predicción de datos predice funciones de valores continuos. Aplicaciones típicas se incluyen análisis de riesgos de préstamos y predicciones de crecimiento. Algunas técnicas de clasificación incluyen:

- Arboles de decisión. Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos. Es decir un árbol de decisión tiene estructura jerárquica conformada por un conjunto de nodos, en donde cada nodo establece una condición o regla la misma que puede retornar verdadero o falso. Además permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.
- Red Bayesiana. Es una representación gráfica o grafo a cíclico dirigido y anotado que describe la distribución de probabilidad conjunta que gobierna un conjunto de variables aleatorias. Los nodos pueden representar cualquier tipo de variable, ya sea un parámetro medible (o medido), una variable latente o una hipótesis, esto permite establecer relaciones causales y efectuar predicciones. Una de las características principales de los métodos bayesianos es el uso de distribuciones de probabilidad

para cuantificar la incertidumbre de los datos que se desea modelar.

- Algoritmos Genéticos. Los algoritmos genéticos es otra técnica que combina la Biología y las Redes Neuronales. Estos algoritmos representan la modelización matemática de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más óptima en aras de la supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas. Los Algoritmos Genéticos hacen uso de las técnicas biológicas de reproducción (mutación y cruce) para ser utilizadas en todo tipo de problemas de búsqueda y optimización.
- Lógica Difusa ("fuzzy logic"). La lógica difusa surge de la necesidad de modelizar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud. La Lógica permite el tratamiento probabilístico de la categorización de un colectivo. Además es un algoritmo que permite y trata la existencia de barreras difusas o suaves entre los distintos grupos en los que categorizamos un colectivo o entre los distintos elementos, factores o proporciones que concurren en una situación o solución.

Se ha detallado brevemente algunas de las principales técnicas que se utilizan en la Minería de Datos, sin embargo existen muchas otras más como se pueden ver en la Tabla 1:

Tabla 1: Técnicas de minería de datos

Métodos Descriptivos	Métodos Predictivos
a. Visualización	a. Regresión estadística · Regresión Lineal · Regresión no lineal · Regresión · Regresión adaptativa lineal ponderada localmente
b. Aprendizaje no supervisado · Clustering Métodos no jerárquicos (Partición) Métodos jerárquicos (N-TREE) Métodos paramétricos (Algoritmo EM) Métodos no paramétricos (KNN, K-means Clustering, Centrioides, Redes Kohones, Algoritmo CobWeb, Autoclass)	b. Aprendizaje supervisado · Clasificación Arboles de Decisión, ID3, C4.5, CART · Inducción de reglas, · Redes Neurales(Simples y Multicapa) · Aprendizaje relacional y recursivo IFP(Inductive Functional Programming) IFLP (Inductive Functional Logic Programming), Aprendizaje de orden superior, Macro Average, Matrices de Coste y confusión, Analisis ROC(Receiver Operating Characteristics)
Asociación	
Asociación Secuencial	
c. Análisis estadístico · Estudio de Distribución de los datos · Detección de datos anómalos · Análisis de dispersión	
Correlaciones y estudios factoriales	
d. Heurísticas y Meta-heurísticas: Búsqueda Aleatoria, Escaladores de Colinas, Recocido Simulado, Búsqueda Tabú, Algoritmos Evolutivos, Algoritmos Meméticos (AG+EC), GRASP, Scatter Search, Path Relinking, Stigmergic Optimization (ACO, PSO), Búsqueda Gravitacional, Optimización de Colonia de Abejas Artificiales, Algoritmo de Fuegos Artificiales, Optimización de Bacterias Forrajeras, entre otros.	

Fuente: Adaptada de Justicia- Formas intermediarias de representación en minería de datos.

Es importante destacar que no existe un "mejor" modelo o algoritmo de minería de datos, depende del problema en estudio y de los datos disponibles y de las pruebas con diferentes clases de técnicas y algoritmos (Martínez, 2012).

Software para Minería de Datos

El proceso de extracción de patrones de comportamiento y de relación entre los datos es una tarea que puede tomar tiempo, pero se puede facilitar con el uso de software de Data Mining, empleados para analizar grandes cantidades de bases de datos. En el mundo informático existen una gran cantidad de herramientas de Software para el análisis de datos, entre ellas se destacan las de uso comercial y de código libre u Open Source.

Software de uso comercial

El software de uso comercial, son programas de minería de datos de pago, estas herramientas pueden llegar a costar cientos de miles de dólares dependiendo el nivel de usabilidad y complejidad del negocio. Entre las principales herramientas comerciales se destaca:

a) XLMiner. Es un complemento macro de extracción de datos para Excel, que permite análisis de datos de tipo transversal como secuencias temporales. Entre las principales características de XLMiner se encuentran:

- Manejo de bases de datos, con imputación de datos faltantes.
- Realización de predicciones, modelos ARIMA, Holt winters, Polinomiales, Arboles de decisión, análisis clúster, Redes neuronales, clasificador Bayes, K vecinos más cercanos, análisis discriminante, reglas de asociación, entre otras.

  XLMiner es de pago y accesible desde su web site: http://www.solver.com/xlminer

b) Matlab. MATLAB (abreviatura de MATrix LABoratory), es un programa propietario con un entorno integrado, utilizado

para llevar a cabo complejos cálculos matemáticos con visualización gráfica en 2D y 3D. Dispone de programas de apoyo especializado denominados Toolboxes y las de Simulink, que amplían el número de funciones para las áreas principales de la ingeniería y la simulación.

MATLAB dispone de un lenguaje de programación propio y puede realizar cualquier tipo de regresión o un proceso de validación cruzada. En relación a estos procesos de la Minería de Datos, se destaca las siguientes herramientas:

- Statistics Toolbox. Combina algoritmos estadísticos con interfaces gráficas interactivas en 2D o 3D.
- Nnet. Herramientas para el procesado de redes neuronales. Se subdivide principalmente en:

  - nnet\nnet - Neural Network Toolbox. Es un paquete de Matlab que contiene una serie de funciones para crear, trabajar visualización y simulación de redes neuronales artificiales.

  - nnet\nncontrol - Neural Network Toolbox Control System Functions. Provee un conjunto de funciones para medir y controlar el sistema de redes neuronales construido.

  - nnet\nndemos - Neural Network Demonstrations. Conjunto de muestras de redes neuronales.

  Para más información sobre MATLAB se puede acceder a la página web: http://www.mathworks.es/products/matlab/

c) IBM SPSS Modeler. Se trata de un producto comercial de la empresa IBM, que permite crear modelos predictivos para descubrir patrones y tendencias en datos estructurados o no estructurados. Posee algoritmos avanzados para obtener conocimientos en tiempo real, como la analítica de texto, analítica geoespacial, analítica de entidades, la gestión y optimización de decisiones. Además permite visualizar gráficamente el proceso llevado a cabo. En cuanto a técnicas de minería de datos, esta herramienta proporciona diferentes métodos como:

- Segmentación. K-medias, Kohonen, Anomalía, Bietápico.
- Asociación. A priori, CARMA, GRI y Análisis de Secuencia.
- Clasificación. Factorial discriminante, Red neuronal, C5.0, GLM, Redes bayesianas, Máquinas de Vectores de Soporte, Modelos de auto aprendizaje, Vecino más próximo, Árboles de decisión, Selección de características, etc.
- Predicción. Regresión lineal, Series de tiempo, Regresión de Cox y Logística.
- Automáticos. Auto numérico, Auto-clasificador, Auto-agrupación, Modelizador ARIMA automático.

Para obtener más información sobre IBM SPSS Modeler se puede consultar la web del fabricante disponible en: https://www.ibm.com/us-en/marketplace/spss-modeler

d) SAS Enterprise Miner. Ofrece uno conjuntos de algoritmos avanzados de modelado predictivo y descriptivo, incluyendo árboles de decisión, redes neuronales, splines de regresión, regresión lineal y logística, regresión por mínimos cuadrados parciales, y muchos más. Tiene una sencilla interfaz gráfica que integra el conjunto de herramientas necesario para la toma de decisiones. La solución Enterprise Miner se basa en la metodología SEMMA (Sample, Explore, Modify, Model, Assess) desarrollada por SAS Institute.

En un breve resumen, se trata de una de las herramientas potenciales del mercado para trabajar con grandes bases de datos; sin embargo, difiere por su alto precio de la licencia de uso. Para obtener más información de esta herramienta se puede acceder a través del siguiente enlace:

http://www.sas.com/technologies/analytics/datamining/miner/

e) Salford Systems Data Mining. Es una empresa especializada de software de minería de datos y consultoría que ofrece los siguientes productos:

- Software CART. Realiza una variedad de análisis de alta precisión de minería de datos, es la única herramienta

basada en árboles de decisión según la metodología desarrollada por la Universidad de Stanford y la Universidad de Berkeley en California.

- TreeNet. Basada en árboles de decisiones y es un sistema de aproximación de funciones y que también sirve como herramienta de exploración inicial de los datos.
- RandomForests. Ofrece modelos predictivos de alto rendimiento e incorpora nuevos análisis de clúster de métrica libre.
- SPM Salford Predictive Modeler. Cuenta con características adicionales orientadas a mejorar los modelos predictivos.

Para utilizar cada uno de estos programas tiene un costo de su licencia, para obtener más información acceda su web: http://www.salford-systems.com/

f) Oracle Data Mining (ODM). Es una herramienta desarrollada por la empresa Oracle y está basada en un esquema de flujo de trabajo, siendo la extensión SQLDeveloper, que analiza, explora los datos, construye y evalúa modelos, así como comparte estos modelos en aplicaciones en línea entregando resultados en tiempo real. La herramienta integra todas las etapas del proceso de la minería de datos y permite integrar los modelos en otras aplicaciones con objetivos similares.

ODM funciona dentro de la base de datos de Oracle y tiene un lenguaje de procedimiento integrado/ lenguaje de consulta estructurado (PL / SQL) e interfaces de Java de programación de aplicaciones (API). Además ODM ofrece dos versiones, una en la que a través de una interfaz gráfica los usuarios podrán aplicar las técnicas de minerías de datos y una versión en la que los desarrolladores podrán utilizar la API de SQP para crear aplicaciones a medida.

Por lo tanto se trata de la herramienta potente para trabajar con bases de datos de Oracle y hay que pagar por su licencia de uso. Para obtener más información se puede consultar dentro de la web de Oracle:

http://www.oracle.com/products/database/options/advanced-analytics/index.html

g) Tableu. Es un software interactivo de análisis y visualización de datos que explora bases de datos grandes y multidimensionales, y permite al usuario interactuar con facilidad con los datos para comparar, filtrar, conectar unas variables con otras, crear distintos tipos de gráficos de forma muy sencilla e incluso mapas en 3D, también permite realizar cálculos de todo tipo para resaltar datos importantes. La plataforma y los paneles que presenta la herramienta son muy visuales y personalizados para la compresión de los informes (dashboards) y toma de decisiones. Para obtener información se puede consultar dentro de la web en: https://www.tableau.com/es-es

Software de Código Abierto

El software de Código Abierto u Open Source, son aquellos programas de uso libre en cual sus características y dimensiones éticas y sociales lo hacen democrático, participativo y accesible. Así describimos algunas de las más relevantes:

a) R, R-Studio. Es uno de los programas más utilizados por la comunidad académica y científica. Es un lenguaje de programación vectorial pensado especialmente análisis estadístico y trabajar con grandes volúmenes de datos (minería de datos), fue desarrollado por John Chambers en los Laboratorios Bell de la Universidad de Stanford.

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, test estadísticos, análisis de seriales temporales, algoritmos de clasificación y agrupamiento, etc.). Para procesos de minería de datos R, posee gran cantidad de paquetes estadísticos como: Rattle, Caret, RDataMining, ggplot2. Tanto el programa como los paquetes estadísticos y su documentación asociada pueden descargarse a través de la web del proyecto R: http://www.r-project.org/ .

b) Rapid Miner. Es una herramienta de minería de datos desarrollado en Java, permite el desarrollo de procesos de

análisis avanzado de datos mediante el encadenamiento de 500 operadores a través de un entorno gráfico, contiene técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos, aprendizaje automático, también utiliza algoritmos incluidos en weka. Finalmente se trata de una herramienta multiplataforma que puede ser descargada junto con su documentación a través del enlace: http://www.knime.org/

c) WEKA (Waikato Environment for Knowledge Analysis). Es una herramienta para el aprendizaje automático y minería de datos diseñado en el lenguaje Java, soporta varias tareas típicas de minería de datos, especialmente pre procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o relación, donde cada punto marcado es etiquetado por un número fijo de atributos. WEKA proporciona acceso a bases de datos SQL utilizando conectividad de bases de datos Java y puede procesar el resultado devuelto como una consulta de base de datos. Su interfaz de usuario principal es el Explorer, pero la misma funcionalidad puede ser accedida desde la línea de comandos o a través de la interfaz de flujo de conocimientos basada en componentes.

d) Orange. Es un software amigable e intuitivo escrito en C++ y Python, diseñado para minería de datos, se basa en aprendizaje de máquina y de una programación visual y de script, es rápida y versátil para un análisis exploratorio de datos, preprocesamiento, filtros de información, modelación de datos, evaluación de modelos y técnicas de exploración. Además, proporciona componentes para:

- Entrada/salida de datos, soportando los formatos C4.5, assistant, retis y tab (nativo)
- Preprocesamiento de datos: selección, discretización, etc.

- Modelado predictivo: árboles de clasificación, regresión logística, clasificador de Bayes, reglas de asociación, etc.
- Métodos de descripción de los datos: mapas autoorganizados, clustering, k-means.
- Técnicas de validación cruzada del modelo.

Si se desea descargar y conocer más de la herramienta es recomendable visitar web site: http://orange.biolab.si/.

e) KNIME (Konstanz Information Miner). Es un software de integración de datos amigable, intuitiva y fácil de usar, está desarrollado sobre la plataforma Eclipse y programado en Java. Permite el procesamiento, análisis y exploración de datos, desde su plataforma. Permite crear visualmente flujos de datos, procesamiento de series de tiempo, ejecutar análisis selectivamente, estudiar los resultados, modelar y generar vistas interactivas, para facilitar la toma de decisiones a nivel gerencial.

KNIME también integra diversos componentes para el aprendizaje automático y minería de datos a través de:

a) Entrada de datos [IO > Read], Salida de datos [IO > Write]

b) Preprocesamiento [Data Manipulation], para filtrar, discretizar, normalizar, filtrar, seleccionar variables, etc.

c) Minería de datos [Mining], para construir modelos y reglas de asociación, clustering, clasificación, MDS, PCA.

d) Salida de resultados [Data Views], para mostrar resultados en pantalla ya sea de forma textual o gráfica.

Finalmente se trata de una herramienta multiplataforma que puede ser descargada junto con su documentación a través del enlace: http://www.knime.org/

A modo de conclusión

Una ciudad inteligente gira en torno al uso de las Tics para analizar los datos y tomar mejores decisiones, anticiparse a los problemas a fin de resolverlos de manera proactiva y coordinar los recursos para funcionar eficazmente. La visión

“Smart City” ofrece a las ciudades una serie de beneficios, no solo asociados al fomento de una mejor calidad de vida, sino también al desarrollo del talento de sus ciudadanos y al progreso empresarial de la ciudad.

El concepto de ciudad inteligente y su interrelación con el Big Data, ha dado un salto cualitativo y cuantitativo en la innovación y producción de conocimiento, centrado principalmente por el valor agregado de los datos generados por las interacciones de los ciudadanos en las redes sociales, de miles de sensores y dispositivos de comunicación conectados a Internet de las cosas. Es decir la gestión eficiente de toda información y el análisis que ofrece el Big Data, permitirá obtener predicciones e incluso recomendaciones que son de gran utilidad para los administradores y el día a día de los ciudadanos.

El aumento en volumen exponencial de los datos y su gran variedad nos plantea retos importantes que deberemos a futuro dar respuestas creativas como por ejemplo al asunto de: ¿Cómo ayudar a las ciudades y empresas a extraer el valor de los datos?, ¿Cómo democratizar el acceso a los datos?, ¿Cómo capturar todos los datos y presentarlos en tiempo real y con patrones relevantes?, ¿Cómo analizar la eficacia de los datos no estructurados?, ¿Cómo manejar las diferentes fuentes de información y el volumen de información?, ¿Qué técnicas, algoritmos, metodologías y plataformas para minar datos son las más apropiadas para cada problema o entorno?, etc.

Evidentemente para dar respuestas a estos desafíos se impone la necesidad de un nuevo profesional con un perfil de arquitecto y periodista de la información, el cual tendrá las habilidades y competencias digitales para “extraer conocimiento” a través de encontrar correlaciones, patrones y predicciones, aplicando algoritmos complejos para crear valor con la información disponible y dar vida a un diseño movido por la interacción humana con los datos.

Referencia bibliográfica

Barbosa, J. G. (2016). *Big data: piedra angular de las smart cities*. Obtenido de https://aunclicdelastic.blogthinkbig.com/wp-content/uploads/eBook-BIG-DATA-AunClicdelasTIC.pdf

Cisco. (2014). I*nternet será cuatro veces más grande en 2016*. Obtenido de http://www.cisco.com/web/ES/about/press/2012/2012-05-30-internet-net

Coghlan, A. (2015). *A Little Book of R for Times Series*. Cambrigde - Trust Sanger Institute: Cambridge, U.K.

Comisión Europea. (2014b). *Making Big Data work for Europe*. Obtenido de http://ec.europa.eu/digital-agenda/en/making-big-data-work-europe

Esteban, F. (2014). *Cinco ejemplos de cómo el 'Big Data' puede mejorar la sociedad*. Obtenido de http://www.daemonquest.com/es/blog/cinco-ejemplos-decomo-el-big-data-puede-mejorar-la-sociedad/

Gartner. (2014). *Big data*. Obtenido de http://www.gartner.com/it-glossary/big-data/

GICI. (2015). *Smart Cities documento de visión a 2030*. Obtenido de http://www.ptferroviaria.es/docs/Documentos/SMART%20CITIES_%20Documento%20de%20Visi%C3%B3n%202030_GICI.pdf

Guerrero, J. A. (2013). *La importancia de la ciencia de datos*. Obtenido de http://www.elconfidencial.com/tecnologia/2013-12-19/un-matematico-andaluz-desconocido-es-el-mejor-cientifico-de-datos-del-mundo_67675/

IBM. (2014). *¿Qué es Big Data?* Obtenido de http://www.ibm.com/developerworks/

López, J. M., & Herrero, J. G. (2012). *Técnicas de Analisis de Datos. Aplicaciones prácticas utilizando Micosoft Excel y Weka*. Obtenido

de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf

Manyika, J., Chui, M., Browm, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung-Byers, A. (2011). Big data: *The next frontier for innovation, competition*. McKinsey Global Institute.

Martínez, C. A. (2012). *Aplicación de Técnicas de Minería de Datos para mejorar el proceso de control de gestión en Entel*. Obtenido de http://repositorio.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1

Molina Neyra, C., & Murakami de la Cruz, S. (2008). *Implementación de una solución informática basado en M-Commerce aplicado a sistemas de distribución comercial*. Lima. Obtenido de Implementación de una solución informática basado en M-Commerce aplicado a sistemas de distribución comercial.

Moro, E. (2016). Big data. El poder de los datos. *Fundación Innovación Bankinter*, 66,68.

Muhammad, D., Mohamudally, N., & Babajee, D. (2013). *A Unified Theoretical Framework for Data Mining*. Procedia Computer Science(17), ISSN 104 – 113.

Natek, S., & Zwilling, M. (2014). S*tudent data mining solution–knowledge management system related to higher education institutions. Expert Systems with*. 6400–6407.

Ortíz, P. (2015). *Minería de datos con series de tiempo en el desarrollo e implementación del sistema inteligente que predice la producción de arroz en el ámbito de la gerencia regional de agricultura*. LAMBAYEQUE: Universidad Señor de Sipán.

Peña, A. A. (2014). *Educational data mining: A survey and a data mining-based analysis of recent works Expert Systems with Applications(41)*. ISSN: 1432–1462.

Pérez, T., Caballero, D., Caro, A., Rodríguez, P., & Antequera, T. (2014). *Applying data mining and Computer Vision Techniques to MRI*

to estimate quality traits in Iberian hams. Journal of Food Engineering(131), ISSN: 82–88.

PowerData. (2016). *Data Scientist: El papel de los científicos de datos en una organización*. Obtenido de http://cdn2.hubspot.net/hub/239039/file-1958963852-pdf/docs/[IC]_OFFER_-_EBOOK-_Data_Scientist_como_contratarlo.pdf

Tsai, H. (2013). *Knowledge management vs. data mining: Research trend, forecast and citation approach*. *Expert Systems with Applications(40)*. 160–3173.

Valcárcel, V. (2004). *Data Mining y el descubrimiento del conocimiento*. Industrial Data. ISSN 83-86.

William, M. (2007). *Intelligent citie, e-Journal on the Knowledge Society. ISSN 1885-1541*. Obtenido de http://www.uoc.edu/uocpapers/5/dt/eng/mitchell.pdf

Las ciudades inteligentes
Edición digital 2017-2018.
www.utmachala.edu.ec

Redes

Redes es la materialización del diálogo académico y propositivo entre investigadores de la UTMACH y de otras universidades iberoamericanas, que busca ofrecer respuestas glocalizadas a los requerimientos sociales y científicos. Los diversos textos de esta colección, tienen un espíritu crítico, constructivo y colaborativo. Ellos plasman alternativas novedosas para resignificar la pertinencia de nuestra investigación. Desde las ciencias experimentales hasta las artes y humanidades, Redes sintetiza policromías conceptuales que nos recuerdan, de forma empeñosa, la complejidad de los objetos construidos y la creatividad de sus autores para tratar temas de acalorada actualidad y de demanda creciente; por ello, cada interrogante y respuesta que se encierra en estas líneas, forman una trama que, sin lugar a dudas, inervará su sistema cognitivo, convirtiéndolo en un nodo de esta urdimbre de saberes.

UNIVERSIDAD TÉCNICA DE MACHALA
Editorial UTMACH
Km. 5 1/2 Vía Machala Pasaje
www.investigacion.utmachala.edu.ec / www.utmachala.edu.ec

ISBN: 978-9942-24-098-9