Ciencia de datos: una revisión del estado del arte

Ing. Naivy Pujol Méndez [1] y Msc. Joelsy Porven Rubier [2]

[1, 2] Universidad de las Ciencias Informáticas (UCI). Habana, Cuba

[1] npujol@uci.cu

[2] jporven@uci.cu

Recibido: 25 feb. 2018 Aceptado: 11 jun. 2018

RESUMEN

Con un crecimiento explosivo en datos no estructurados y estructurados, las organizaciones buscan formas de innovar a través del análisis y de la ciencia de datos; la disponibilidad de Big Data permite a las organizaciones de todas las industrias aprovechar el análisis de datos. Por tanto, el objetivo de este artículo es realizar una revisión del estado del arte referente a la ciencia de datos. Se realizó un estudio inicial para determinar los temas y términos más representativos en el campo de la ciencia de datos y se utilizaron los métodos de investigación analítico-sintético e histórico-lógico para examinar los elementos fundamentales y característicos de la ciencia de datos y los científicos de datos; y para determinar los diferentes procesos, soluciones, herramientas y la evolución de estas en el transcurso del tiempo. Las principales conclusiones arribadas se encuentran: la amplia aplicación de la ciencia de datos, trae como consigo que existan muchas soluciones diferentes, estrechamente relacionados con el área de aplicación y las características del problema; propiciado por Big Data en la mayoría de las ocasiones se utiliza el aprendizaje automático para resolver los problemas; las técnicas más utilizados son los siguientes: regresión lineal, k-Nearest Neighbors (k-NN), k-means, regresión logística, redes bayesianas, máquina de soporte vectorial y redes neuronales.

Palabras claves: Ciencia de datos; Científico de datos; Aprendizaje automático

ABSTRACT

Data science: a review of the state of the art. With an explosive growth in unstructured and structured data, organizations look for ways to innovate through analysis and data science; The availability of Big Data allows organizations of all industries to take advantage of data analysis. Therefore, the objective of this article is to perform a review of the state of the art regarding data science. An initial study was carried out to determine the most representative topics and terms in the field of data science and analytical-synthetic and historical-logical research methods were used to examine the fundamental and characteristic elements of data science and scientists. of data; and to determine the different processes, solutions, tools and the evolution of these in the course of time. The main conclusions reached are: the wide application of data science, brings with it the existence of many different solutions, closely related to the area of application and the characteristics of the problem; propitiated by Big Data in most of the occasions the automatic learning is used to solve the problems; The most used techniques are the following: linear regression, k-Nearest Neighbors (k-NN), k-means, logistic regression, Bayesian networks, vector support machine and neural networks.

Keywords: Data science; Data scientist; Machine learning

INTRODUCCIÓN

El avance de las Tecnologías de la Información y las Comunicaciones (TIC), ha elevado la cantidad de información almacenada a partir de las diferentes áreas de la vida común. En muchos casos, la ciencia se está quedando atrás del mundo comercial en cuanto a la capacidad de inferir el significado de los datos y tomar medidas basadas en ese

significado (Hey, 2012). Durante años, le empresa Gartner[1] ha informado las tendencias de investigación de las tecnologías de la información. En el octubre del 2017, se celebró el *Gartner Symposium/ITxpo 2017: innovation and disruption*, en esta ocasión se dieron a conocer las 10 principales tendencias tecnológicas estratégicas para 2018; la inteligencia artificial, las aplicaciones y el análisis inteligente estaban entre ellas. Con respecto a estos temas David W. Cearley, vicepresidente de la compañía, expresó: "En los próximos años, cada aplicación, aplicación y servicio incorporará inteligencia artificial en algún nivel… se ha convertido en el próximo gran campo de batalla en una amplia gama de mercados de software y servicios...para agregar de datos valor comercial en nuevas versiones en forma de análisis avanzados, procesos inteligentes y experiencias de usuario avanzadas". Con un crecimiento explosivo de los datos no estructurados y estructurados, las organizaciones buscan formas de innovar a través del análisis y de la ciencia de datos (Blei and Smyth, 2017**)**.

La ciencia de los datos a comenzado una nueva era; (Hey, 2012) en el libro *The Fourth Paradigm – Data-Intensive Scientific Discovery* lo define como el cuarto paradigma debido a la alta aplicabilidad en la solución de problemas. Existe un creciente interés en las organizaciones por extraer información y producir conocimiento a partir de la cantidad masiva de datos creados diariamente (Fayyad et al., 2017). La disponibilidad de *Big Data* permite a las organizaciones de todas las industrias aprovechar el análisis de datos, con el fin de extraer conocimiento procesable que pueden utilizarse para la toma de decisiones y predicciones comerciales sólidas (Molina-Solana et al., 2017). Al utilizar Big Data, el análisis empresarial abre el potencial predictivo del análisis de datos para mejorar la gestión estratégica, la eficiencia operativa y el rendimiento financiero (Newman et al., 2016). Pero no es solo la masividad lo que hace que todos estos datos nuevos sean interesantes o plantee desafíos (Van der Aalst, 2016), son datos en sí, y su comportamiento en tiempo real (Rupp et al., 2017), los convierten en componentes básicos en la búsqueda de conocimiento. Una característica importante de los datos es la alta diversidad con la que cuentan. Estos pueden ser desde los tradicionales: numérico, categórico o binario hasta más complejos como los son: texto (correos electrónicos, *tweets*, artículos científicos, comentarios), registros (datos a nivel de usuario, datos de eventos con marcas de tiempo, *logs*), datos de ubicación geográfica, red, sensores o imágenes.

Por tanto, los principales desafíos científicos que dan paso al surgimiento la ciencia de los datos, están dados por la necesidad de analizar datos diversos, incompletos y desordenados; conjuntos de datos muy grandes, que cambian en el tiempo (Kormos et al., 2017); y la necesidad de encontrar hallazgos que impulsen decisiones sobre operaciones y productos en las organizaciones. Hoy, la ciencia de datos está entrando en una nueva era, donde la tecnología de la información ahora es capaz de soportar negocios basados en datos, en tiempo real (Norbert et al., 2017) con el fin de facilitar decisiones informadas basadas en evidencia científica confiable para proporcionar herramientas a los responsables de las políticas y las decisiones (Dalkir and Beaulieu, 2017).

En Internet, los sistemas de recomendación de productos en Amazon, amigos en Facebook, de películas en Netflix, música en Spotify y más; los análisis de comportamientos en el medio ambiente (Rupp et al., 2017), migraciones (Knudson et al., 2016), bioinformáticos (Baxevanis and Ouellette, 2004) y muchas otras aplicaciones en las diversas áreas son ejemplos de la aplicación de la ciencia de datos. Influyendo en todos los sectores de la industria y la academia, desde las empresas hasta la educación y la atención médica, desde el sector científico hasta el gobierno, y está revolucionando estas industrias a medida que continúa creciendo en importancia.

La amplia aplicación de la ciencia de datos, trae como consigo que existan muchas soluciones diferentes, estrechamente relacionados con el área de aplicación y las características del problema. Por tanto, el objetivo de este artículo es realizar una revisión del estado del arte referente a la ciencia de datos. La estructura del artículo será la siguiente: se describe los fundamentos, características y el proceso de la ciencia de datos; ¿Qué es un científico de datos?, ¿Cuáles son sus competencias? y cuáles son las principales soluciones y herramientas relacionadas a este tema. Por último, se realizará una discusión para llegar a conclusiones sobre los elementos estudiados relacionados con la ciencia de datos.

---

1 Gartner Inc. es una empresa consultora y de investigación de las tecnologías de la información con sede en Stamford, Connecticut, Estados Unidos.

METODOLOGÍA COMPUTACIONAL

Se realizó un análisis inicial para determinar los temas y términos más representativos en el campo de la ciencia de datos. Los métodos de investigación empleados son el analítico-sintético e histórico-lógico; el método analítico-sintético se empleó para examinar los elementos fundamentales y característicos de la ciencia de datos y los científicos de datos; y el método histórico-lógico se utilizó para determinar los diferentes procesos, soluciones, herramientas y la evolución de estas en el tiempo.

Se revisaron un total de 225 trabajos sobre la ciencia de datos, provenientes de revistas como: *Data Science Journal, Expert Systems with Applications, Future Generation Computer Systems, Journal of Machine Learning Research, Journal of Systems and Software, Neurocomputing, Revista Cubana de Ciencias Informáticas, etc...* También se revisaron 41 eventos como: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1st Europe Summer School: Data Science, Proceedings of the 2014 Conference on Designing Interactive Systems, Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, etc...* También de revisaron 17 tesis de posgrado y 81 libros o secciones de libros. Entre los libros sobresalientes de este tema según *google scholar* se encuentran: *Doing Data Science: Straight Talk from the Frontline* con 223 citas*, Process Mining: Data Science in Action* con 210 citas*, Data Science for Business: What you need to know about data mining and data-analytic thinking* con 343 citas. En la revisión se tuvo en cuenta la fecha de publicación de los mismos y su tipo. En la figura 1 se observa la relación tipo de documento-fecha de publicación que se tuvo en cuenta en el análisis.

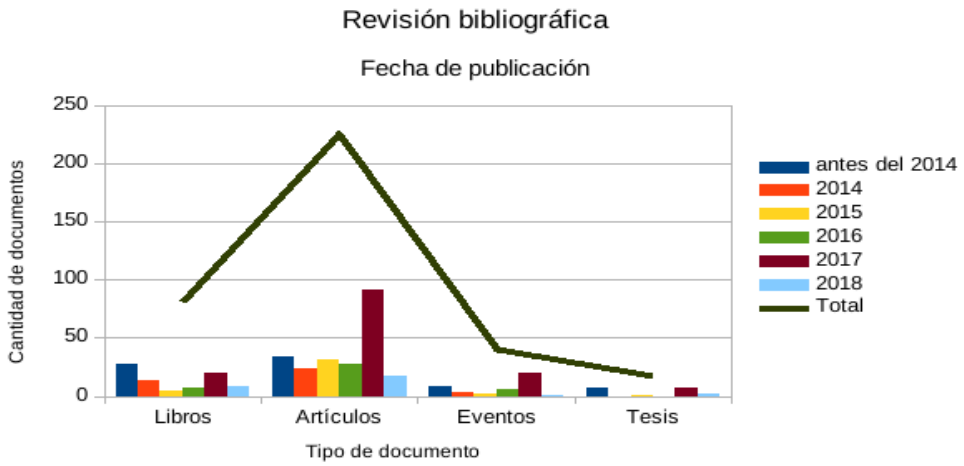

Figura 1. Revisión bibliográfica. Tipo de documento-fecha de publicación.

Ciencia de datos. Definiciones.

El término **ciencia de datos** se acuñó a principios del siglo XXI; se le atribuye a (Cleveland, 2001), donde lo define como una continuación de algunos campos de análisis de datos como la estadística, pone la nueva disciplina propuesta en el contexto de la informática y el trabajo contemporáneo en la minería de datos (Cleveland, 2001). También en ese mismo año (Breiman and others, 2001), planteó *"Si nuestro objetivo como campo es utilizar datos para resolver problemas, entonces debemos alejarnos de la dependencia exclusiva de los modelos de datos y adoptar un conjunto de herramientas más diverso".*

Según (Hazen et al., 2014) se define como campo emergente de la ciencia de datos combina, ciencias de la computación, estadística, matemática, y la experiencia ciencia del comportamiento para provocar puntos de vista de datos de la empresa. Otro elemento agregado posteriormente es su relación con la minería de datos (Provost and Fawcett, 2013), la minería de procesos (Van der Aalst, 2016) y el aprendizaje automático (Chojnacki et al., 2017). En la figura 2 se muestra la evolución de los diferentes elementos que conforman la ciencia de datos.

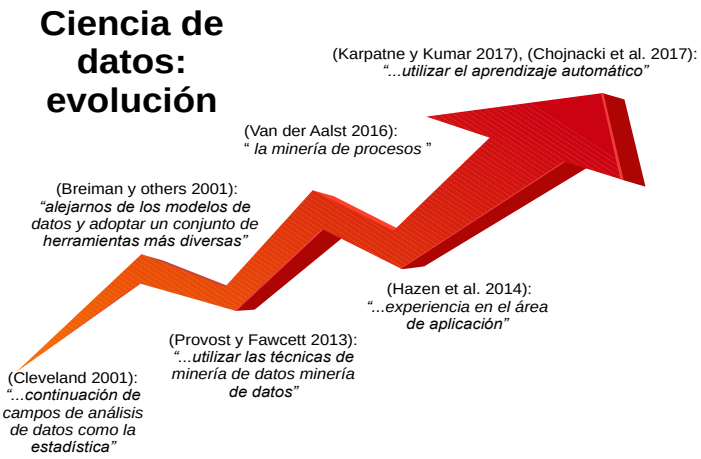

Figura 2. Evolución de la Ciencia de datos.

Varias son las técnicas y teorías dentro de las áreas de matemáticas, estadística e informática empleadas en la ciencia de datos (Schutt and O'Neil, 2013), además, agrega conocimiento de dominio no especificado propio del área de aplicación de la solución. En realidad, esto implica una amplia gama de disciplinas, una por campo de dominio (Ayankoya et al., 2014). En la figura 3 se muestra el diagrama de Venn, que ilustra la relación de las diferentes disciplinas con la ciencia de datos; informática/ciencias de la computación, matemáticas y estadística, y dominio específico del negocio.

De cada una de estas disciplinas son necesarias varias teorías y técnicas para la construcción de las soluciones (Blei

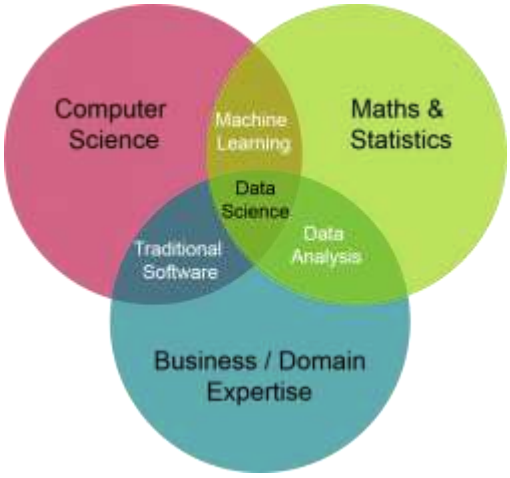

Figura 3. Disciplinas que componen la Ciencia de datos. (Tomado de *(Ayankoya et al., 2014)*)

and Smyth, 2017):

- Estadística: modelado estadístico, diseño experimental, muestreo, agrupamiento, reducción de datos, intervalos de confianza, pruebas, modelado, modelado predictivo.

- Matemáticas: investigación de operaciones, negocio analítico (gestión de inventario y previsión, optimización de precios, cadena de suministro, control de calidad, optimización de rendimiento).

- Análisis de datos: dominio de tecnologías y herramientas emergentes, optimización y arquitectura de bases de datos/memoria/ archivos, API, optimización de flujos de datos, computación paralela.

- Informática: algoritmos, computación, complejidad, optimización.

- Negocios, optimización de infraestructuras, ciencias de decisión (diseño de tableros, selección y definiciones de métricas)

Otro elemento importante abordado en la figura 3 es el dominio de aplicación, a diferencia de las otras disciplinas estas pueden ser muy variadas de un problema a otro (Kempler and Mathews, 2017). Sin embargo, según (Schoenherr and Speier-Pero, 2015) las habilidades comunicativas juegan un papel definitivo con el fin de mejorar esta área. Después de analizar las diferentes definiciones estudiadas y las disciplinas que involucra, se propone la siguiente definición: la ciencia de datos es el campo interdisciplinario con bases en informática, estadística y matemáticas, que se ocupa de la teoría, la práctica y la comunicación de los resultados con el fin de extraer conocimiento relevante de los datos (elaboración propia). Este nuevo campo implica nuevos especialistas, con habilidades muy variadas, las personas que se dedican a la ciencia de datos se les conoce como científico de datos.

Científico de datos

Los científicos de datos son los encargados de hacer frente a los grandes proyectos de la ciencia de datos en todos los niveles (Hey, 2012). Deben poseer habilidades en las disciplinas relacionas con la ciencia de datos (Schutt and O'Neil, 2013). Tiene que ser capaz de estudiar las diversas fuentes de información disponibles en una organización; extraer datos a partir de diversos formatos (Blei and Smyth, 2017); depurarlos, analizarlos, idear y desarrollar algoritmos; realizar inferencias, preparar y comunicar los resultados de dichos análisis y transmitir conclusiones que ayude a tomar mejores decisiones (Cao, 2017).

Las actividades que realiza dependen del ámbito en el cual se desempeñe: académico o industrial (Schuff, 2018). Un científico de datos académicos es un científico entrenado en cualquier área del conocimiento, desde las ciencias sociales hasta la biología; trabaja con grandes cantidades de datos (Schutt and O'Neil, 2013) y debe lidiar con problemas computacionales complejos planteados por la estructura, tamaño, desorden, la complejidad y naturaleza de los datos (Van der Aalst, 2016), mientras que simultáneamente resuelven un problema del mundo real. Un científico jefe de datos, en el ámbito industrial, debe establecer la estrategia de datos de la empresa (Schutt and O'Neil, 2013); administración del equipo de ingenieros, científicos y analistas; debe comunicar a los líderes los resultados del producto (Dhar, 2013); patentar soluciones innovadoras y establecer objetivos de investigación (Ayankoya et al., 2014); así como, configuración de la ingeniería y la infraestructura para recopilar datos, su uso en la toma de decisiones y su integración en el producto (Giama and Papadopoulos, 2018).

Proceso de la ciencia de datos

Para organizar las soluciones que utilizan ciencia de datos (Schutt and O'Neil, 2013) propone los elementos fundamentales de cada una de las etapas que deben seguir los proyectos de ciencia de datos, se describen a continuación (Dalkir and Beaulieu, 2017):

1. Identificar el problema: Tipo de problemas y métrica utilizada para medir el éxito. Identificar personas clave dentro de su organización y fuera de ella. Obtener especificaciones, requisitos, prioridades, presupuestos ¿Qué tan precisa debe ser la solución? ¿Necesitamos todos los datos?

2. Identificar las fuentes de datos disponibles: Extraer y verificar la muestra, realizar análisis exploratorios. Evaluar la calidad y el valor disponible en los datos, identificar problemas técnicos y encontrar soluciones alternativas.

3. Identificar si se necesitan fuentes de datos adicionales: ¿Cuáles? ¿Cuánto cuesta? ¿Tiempo real? ¿Necesita diseño experimental?

4. Preparación y análisis de datos: Limpieza de datos, explorar metodologías, seleccionar variables y modelos, detectar / eliminar valores atípicos, validar la metodología elegida, medir la precisión, proporcionar intervalos de confianza y proporcionar visualización.

5. Implementación y desarrollo: FSSRR: rápido, simple, escalable, robusto, reutilizable, depuración, ¿Necesitas crear una API para comunicarte con otras aplicaciones?

6. Comunicación de los resultados: Integración y visualización, discusión de posibles mejoras (con estimaciones de costos), proporcionar entrenamiento, documentación de código y metodología.

7. Mantenimiento: Pruebe el modelo o la implementación; pruebas de estrés, actualizaciones regulares.

En la fase de "implementación y desarrollo" son varias las herramientas utilizadas. Entre las tareas más importantes llevadas a cabo mediante los modelos estadísticos se encuentran: interpretar parámetros; los estadísticos proporcionan intervalos de confianza y distribuciones posteriores para parámetros y estimadores (Schutt and O'Neil, 2013); y están interesados en capturar la variabilidad o incertidumbre de los parámetros (Amato et al., 2018). Los modelos estadísticos hacen suposiciones explícitas sobre los procesos y distribuciones que generan datos, y usted usa los datos para estimar los parámetros (Gould et al., 2018). Propiciado por *Big Data* es necesario en la mayoría de las ocasiones utilizar el aprendizaje automático (Dhar, 2013), (Karpatne and Kumar, 2017), (Lee et al., 2018).

Aprendizaje automático en la ciencia de datos.

El aprendizaje automático, (la intersección de las ciencias de la computación/informática y las estadísticas/matemáticas en la figura 1) aporta una nueva perspectiva que conduce a nuevos conocimientos y ningún conocimiento de dominio previo puede ser potencialmente ventajoso (Dalkir and Beaulieu, 2017). Esto permite superar el sesgo de dominio de aplicación, facilitando el uso de la ciencia de los datos en contextos muy diferentes (Kempler and Mathews, 2017); de manera de que se pueda aprender sobre la empresa y el dominio junto mediante preguntas a los expertos del dominio.

Las técnicas de aprendizaje automático se utilizan en gran medida para predecir, clasificar o agrupar (Mueller and Massaron, 2016). Las técnicas más utilizadas son los siguientes (Amato et al., 2018): regresión lineal, k-Nearest Neighbors (k-NN), k-means, regresión logística, redes bayesianas, máquina de soporte vectorial y redes neuronales; cada uno de estas técnicas serán analizados a continuación.

Regresión lineal: se usa para expresar la relación matemática entre dos variables o atributos, su salida es una variable continua. Debido a las pocas variables que puede manejar esta técnica no es tan utilizado, aunque hay investigaciones en que lo usan; por ejemplo, para predecir el número de bicicletas registradas en un mostrador de bicicletas, o para predecir el momento de la muerte del pez cebra y el ratón (Hunter et al., 2017). Para utilizar esta técnica es necesario hace las siguientes suposiciones: linealidad; términos de error distribuidos normalmente con media cero, independientes entre sí y con varianza constante entre los valores de x; por último, se asume que los predictores usando son los correctos (Duzhin and Gustafsson, 2018). Se utiliza fundamentalmente para predecir una variable conociendo a los demás o si queremos explicar o entender la relación entre dos o más cosas (Amato et al., 2018).

k-Nearest Neighbors (k-NN) o K-vecinos más cercanos: es una técnica de clasificación que necesita un grupo de objetos que han sido clasificados o etiquetados y otros objetos similares que aún no se han clasificado o etiquetado, y desea una manera de etiquetarlos automáticamente (Adeniyi et al., 2016). La técnica k-NN es un ejemplo de un enfoque no paramétrico; no necesita suposiciones de modelado sobre las distribuciones subyacentes generadoras de datos, y no intenta estimar ningún parámetro (Kim and Na, 2018). Para su uso se asumen que los datos se encuentran en algún espacio de características donde la noción de "distancia" tiene sentido (Schutt and O'Neil, 2013); los datos de entrenamiento han sido etiquetados o clasificados en dos o más clases (Amato et al., 2018); el valor de k elige el número de vecinos para usar. Esta técnica es utilizada en varias soluciones, como ejemplo de ellas se tienen: clasificación de un pequeño subconjunto de características de *microarrays* de ADN (C. Yu et al., 2017); para predecir el género basado en una colección de publicaciones de la vida real en páginas de *blog* reales (Chen et al., 2017), diseño de un clasificador utilizando *Spark* para análisis con *Big Dat*a (Maillo et al., 2017). Esta técnica tiene una amplia aplicación dentro del aprendizaje supervisado porque permite mediante un conjunto de datos previamente clasificados predecir la clase de otros.

K-means: es una técnica de aprendizaje no supervisada, donde el objetivo de la técnica es determinar la definición de la respuesta correcta mediante la búsqueda de agrupaciones o clusterización de los datos (L. Yu et al., 2017). Este es un ejemplo de aprendizaje no supervisado porque las etiquetas no son conocidas y en su lugar son descubiertas por el algoritmo ((Huang and Luo, 2016). esta técnica posee varios problemas conocidos: problemas de convergencia: la solución puede no existir, si el algoritmo cae en un bucle y la interpretabilidad; la respuesta puede no ser del todo útil, de hecho, ese es a menudo el mayor problema que presenta (Schutt and O'Neil, 2013). A pesar de estos problemas, es bastante rápido (en comparación con otras técnicas de agrupamiento) y existen amplias aplicaciones (Iyer, Zhou y Paul 2017). Entre las aplicaciones tenemos: para mejorar las características biológicas de las redes de co-expresión de genes WGCNA (Botía et al., 2017); para la agrupación difusas para redes inalámbricas de sensores (WSN) donde cada nodo está equipado con sensores (Kamper et al., 2017); para el análisis de conglomerados de la provincia de Indonesia con base en el combustible de cocina principal del hogar (Huda, 2017).

Regresión logística: es una técnica de aprendizaje no supervisada para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras (Cao, 2017). La regresión logística es intrínsecamente simple, tiene baja varianza y, por lo tanto, es menos propensa al ajuste excesivo (Wang and Priestley, 2017); y funciona mejor si hay un solo límite de decisión (Urbano et al., 2017). Entre los ejemplos de aplicación se tiene: clasificación binaria de cuentas de servicio vencidas; análisis de herramientas de datos científicos para la evaluación basada en sensores de la calidad de vida en la atención médica (Urbano et al., 2017); predecir revisiones de pacientes en el Departamento de Emergencia del Sistema de Salud de la Universidad de Virginia (Fowler et al., 2017).

Redes bayesianas: Las redes bayesianas modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas (Toyinbo et al., 2017). Dado este modelo, se puede hacer inferencia bayesiana (McNally et al., 2017); es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Estos modelos pueden tener diversas aplicaciones, para clasificación, predicción, diagnóstico, etc. Entre sus aplicaciones se tiene: análisis del politraumatismo y las lesiones relacionadas con explosiones (Toyinbo et al., 2017); sistema PrivBayes: revelar información privada a través de redes bayesianas (Zhang et al., 2017); identificar a TRIB1 como un nuevo regulador de la progresión y supervivencia del ciclo celular en células cancerosas (Gendelman et al., 2017).

Máquina de soporte vectorial: son un conjunto de técnicas de aprendizaje supervisados relacionados utilizados para la clasificación y la regresión (Anzola, 2016). Son sistemas de aprendizaje que utilizan como espacio de hipótesis, funciones lineales en espacios característicos de dimensión muy alta, ensayando algoritmos de aprendizaje de la teoría de la optimización que implementan un aprendizaje sesgado derivado a partir de la teoría del aprendizaje estadístico (Zhou et al., 2017). Entre las aplicaciones tenemos: predecir los riesgos de seguridad en pozos de fundación profundos en proyectos de infraestructura de metro (Zhou et al., 2017); clasificación de la transición de estado mediante el uso de un sensor Doppler de microondas para la detección errante (Shiba et al., 2017).

Redes neuronales: Las redes neuronales artificiales (ANN, por sus siglas en inglés) o las conexiones son sistemas informáticos inspirados en las redes neuronales biológicas (Esteva et al., 2017). Dichos sistemas aprenden (mejoran progresivamente el rendimiento en) tareas al considerar ejemplos, generalmente sin programación específica de la tarea (Esteva et al., 2017). Desarrollan su propio conjunto de características relevantes a partir del material de aprendizaje que procesan. Entre los ejemplos de aplicación se encuentran: clasificar las características específicas de un diente en base a datos de escáner 3D (Raith et al., 2017); clasificación textual generativa y discriminativa con redes neuronales recurrentes (Yogatama et al., 2017); pronóstico adaptativo de baterías de iones de litio (Sbarufatti et al., 2017).

Cada una de estas técnicas solucionan problemas diferentes en dependencia de los datos disponibles. Solo con los datos, sin testar previamente clasificado o etiquetado, es necesarios técnicas de aprendizaje no supervisado. Si, por el contrario, se tiene un conjunto clasificado entonces se está en presencia de aprendizaje supervisado.

CONCLUSIONES

La ciencia de datos es un campo interdisciplinario que permite encontrar hallazgos interesares en los conjuntos de los datos; influyendo significativamente en todos los sectores de la industria y la academia, a medida que continúa creciendo en importancia. La gran cantidad de herramientas que conforman la ciencia de los datos propicia que existan muchas soluciones diferentes, estrechamente relacionados con el área de aplicación y las características del problema. Propiciado por Big Data en la mayoría de las ocasiones se utiliza el aprendizaje automático para resolver los problemas; las técnicas más utilizadas son los siguientes: regresión lineal, k-Nearest Neighbors (k-NN), k-means, regresión logística, redes bayesianas, máquina de soporte vectorial y redes neuronales.

REFERENCIAS BIBLIOGRÁFICAS

1. Adeniyi, D.A., Wei, Z., Yongquan, Y., 2016. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Appl. Comput. Inform. 12, 90–108. https://doi.org/10.1016/j.aci.2014.10.001

2. Amato, G., Candela, L., Castelli, D., Esuli, A., Falchi, F., Gennaro, C., Giannotti, F., Monreale, A., Nanni, M., Pagano, P., Pappalardo, L., Pedreschi, D., Pratesi, F., Rabitti, F., Rinzivillo, S., Rossetti, G., Ruggieri, S., Sebastiani, F., Tesconi, M., 2018. How Data Mining and Machine Learning Evolved from Relational Data Base to Data Science, in: A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Springer, Cham, pp. 287–306. https://doi.org/10.1007/978-3-319-61893-7_17

3. Anzola, N.S., 2016. Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. ODEON 0, 113–172.

4. Ayankoya, K., Calitz, A., Greyling, J., 2014. Intrinsic Relations Between Data Science, Big Data, Business Analytics and Datafication, in: Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology, SAICSIT '14. ACM, New York, NY, USA, p. 192:192–192:198. https://doi.org/10.1145/2664591.2664619

5. Baxevanis, A.D., Ouellette, B.F.F., 2004. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons.

6. Blei, D.M., Smyth, P., 2017. Science and data science. Proc. Natl. Acad. Sci. 114, 8689–8692. https://doi.org/10.1073/pnas.1702076114

7. Botía, J.A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., Hardy, J., Lewis, C.M., Ryten, M., Weale, M.E., 2017. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. BMC Syst. Biol. 11, 47. https://doi.org/10.1186/s12918-017-0420-6

8. Breiman, L., others, 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231.

9. Cao, L., 2017. Data Science: A Comprehensive Overview. ACM Comput Surv 50, 43:1–43:42. https://doi.org/10.1145/3076253

10. Chen, J., Xiao, T., Sheng, J., Teredesai, A., 2017. Gender prediction on a real life blog data set using LSI and KNN, in: 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). pp. 1–6. https://doi.org/10.1109/CCWC.2017.7868410

11. Chojnacki, A., Dai, C., Farahi, A., Shi, G., Webb, J., Zhang, D.T., Abernethy, J., Schwartz, E., 2017. A Data Science Approach to Understanding Residential Water Contamination in Flint, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17. ACM, New York, NY, USA, pp. 1407–1416. https://doi.org/10.1145/3097983.3098078

12. Cleveland, W.S., 2001. Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. Int. Stat. Rev. 69, 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

13. Dalkir, K., Beaulieu, M., 2017. Knowledge Management in Theory and Practice. MIT Press.

14. Dhar, V., 2013. Data Science and Prediction. Commun ACM 56, 64–73. https://doi.org/10.1145/2500499

15. Duzhin, F., Gustafsson, A., 2018. Machine Learning-Based App for Self-Evaluation of Teacher-Specific Instructional Style and Tools. Educ. Sci. 8. https://doi.org/10.3390/educsci8010007

16. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.

17. Fayyad, U.M., Simoudis, E., Srivastava, A., 2017. Foreword to the Applied Data Science: Invited Talks Track at KDD-2017, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17. ACM, New York, NY, USA, pp. 7–8. https://doi.org/10.1145/3097983.3121426

18. Fowler, B., Rajendiran, M., Schroeder, T., Bergh, N., Flower, A., Kang, H., 2017. Predicting patient revisits at the University of Virginia Health System Emergency Department, in: 2017 Systems and Information Engineering Design Symposium (SIEDS). pp. 253–258. https://doi.org/10.1109/SIEDS.2017.7937726

19. Gendelman, R., Xing, H., Mirzoeva, O.K., Sarde, P., Curtis, C., Feiler, H.S., McDonagh, P., Gray, J.W., Khalil, I., Korn, W.M., 2017. Bayesian Network Inference Modeling Identifies TRIB1 as a Novel Regulator of Cell-Cycle Progression and Survival in Cancer Cells. Cancer Res. 77, 1575–1585. https://doi.org/10.1158/0008-5472.CAN-16-0512

20. Giama, E., Papadopoulos, A.M., 2018. Carbon footprint analysis as a tool for energy and environmental management in small and medium-sized enterprises. Int. J. Sustain. Energy 37, 21–29. https://doi.org/10.1080/14786451.2016.1263198

21. Gould, R., Wild, C.J., Baglin, J., McNamara, A., Ridgway, J., McConway, K., 2018. Revolutions in Teaching and Learning Statistics: A Collection of Reflections, in: International Handbook of Research in Statistics Education. Springer, Cham, pp. 457–472. https://doi.org/10.1007/978-3-319-66195-7_15

22. Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. 154, 72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

23. Hey, T., 2012. The Fourth Paradigm – Data-Intensive Scientific Discovery, in: E-Science and Information Management. Springer, Berlin, Heidelberg, pp. 1–1. https://doi.org/10.1007/978-3-642-33299-9_1

24. Huang, D., Luo, L., 2016. Consumer Preference Elicitation of Complex Products Using Fuzzy Support Vector Machine Active Learning. Mark. Sci. 35, 445–464. https://doi.org/10.1287/mksc.2015.0946

25. Huda, S.N., 2017. Cluster Analysis of Indonesian Province Based on Household Primary Cooking Fuel Using K-Means. IOP Conf. Ser. Mater. Sci. Eng. 185, 012016. https://doi.org/10.1088/1757-899X/185/1/012016

26. Hunter, M.C., Pozhitkov, A.E., Noble, P.A., 2017. Accurate predictions of postmortem interval using linear regression analyses of gene meter expression data. Forensic Sci. Int. 275, 90–101. https://doi.org/10.1016/j.forsciint.2017.02.027

27. Kamper, H., Livescu, K., Goldwater, S., 2017. An embedded segmental K-means model for unsupervised segmentation and clustering of speech. ArXiv170308135 Cs.

28. Karpatne, A., Kumar, V., 2017. Big Data in Climate: Opportunities and Challenges for Machine Learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17. ACM, New York, NY, USA, pp. 21–22. https://doi.org/10.1145/3097983.3105810

29. Kempler, S., Mathews, T., 2017. Earth Science Data Analytics: Definitions, Techniques and Skills. Data Sci. J. 16. https://doi.org/10.5334/dsj-2017-006

30. Kim, Y.-K., Na, K.-S., 2018. Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective. Prog. Neuropsychopharmacol. Biol. Psychiatry 80, 71–80. https://doi.org/10.1016/j.pnpbp.2017.06.024

31. Knudson, S., Sarkar, S., Ray, A., 2016. Connecting Data Science and Qualitative Interview Insights through Sentiment Analysis to Assess Migrants' Emotion States Post-Settlement. ArXiv160908776 Cs.

32. Kormos, M., Collura, M., Takács, G., Calabrese, P., 2017. Real-time confinement following a quantum quench to a non-integrable model. Nat. Phys. 13, 246. https://doi.org/10.1038/nphys3934

33. Lee, S.-I., Celik, S., Logsdon, B.A., Lundberg, S.M., Martins, T.J., Oehler, V.G., Estey, E.H., Miller, C.P., Chien, S., Dai, J., Saxena, A., Blau, C.A., Becker, P.S., 2018. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat. Commun. 9, 42. https://doi.org/10.1038/s41467-017-02465-5

34. Maillo, J., Ramírez, S., Triguero, I., Herrera, F., 2017. kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. Vol. Var. Veloc. Data Sci. 117, 3–15. https://doi.org/10.1016/j.knosys.2016.06.012

35. McNally, R.J., Mair, P., Mugno, B.L., Riemann, B.C., 2017. Co-morbid obsessive–compulsive disorder and depression: a Bayesian network approach. Psychol. Med. 47, 1204–1214. https://doi.org/10.1017/S0033291716003287

36. Molina-Solana, M., Ros, M., Ruiz, M.D., Gómez-Romero, J., Martin-Bautista, M.J., 2017. Data science for building energy management: A review. Renew. Sustain. Energy Rev. 70, 598–609. https://doi.org/10.1016/j.rser.2016.11.132

37. Mueller, J.P., Massaron, L., 2016. Machine learning for dummies, For dummies. John Wiley & Sons, Inc, Hoboken, NJ.

38. Newman, R., Chang, V., Walters, R.J., Wills, G.B., 2016. Model and experimental development for Business Data Science. Int. J. Inf. Manag. 36, 607–617. https://doi.org/10.1016/j.ijinfomgt.2016.04.004

39. Norbert, D., Andreas, G., Armin, K., Manuel, M., Andrea, H., 2017. Solutions for Cyber-Physical Systems Ubiquity. IGI Global.

40. Provost, F., Fawcett, T., 2013. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

41. Raith, S., Vogel, E.P., Anees, N., Keul, C., Güth, J.-F., Edelhoff, D., Fischer, H., 2017. Artificial Neural Networks as a powerful numerical tool to classify specific features of a tooth based on 3D scan data. Comput. Biol. Med. 80, 65–76. https://doi.org/10.1016/j.compbiomed.2016.11.013

42. Rupp, G.M., Opitz, A.K., Nenning, A., Limbeck, A., Fleig, J., 2017. Real-time impedance monitoring of oxygen reduction during surface modification of thin film cathodes. Nat. Mater. 16, 640. https://doi.org/10.1038/nmat4879

43. Sbarufatti, C., Corbetta, M., Giglio, M., Cadini, F., 2017. Adaptive prognosis of lithium-ion batteries based on the combination of particle filters and radial basis function neural networks. J. Power Sources 344, 128–140. https://doi.org/10.1016/j.jpowsour.2017.01.105

44. Schoenherr, T., Speier-Pero, C., 2015. Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential. J. Bus. Logist. 36, 120–132. https://doi.org/10.1111/jbl.12082

45. Schuff, D., 2018. Data Science for All: A University-Wide Course in Data Literacy, in: Analytics and Data Science. Springer, Cham, pp. 281–297. https://doi.org/10.1007/978-3-319-58097-5_20

46. Schutt, R., O'Neil, C., 2013. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc.

47. Shiba, K., Kaburagi, T., Kurihara, Y., 2017. Classification of State Transition by Using a Microwave Doppler Sensor for Wandering Detection 11, 3245.

48. Toyinbo, P.A., Vanderploeg, R.D., Belanger, H.G., Spehar, A.M., Lapcevic, W.A., Scott, S.G., 2017. A Systems Science Approach to Understanding Polytrauma and Blast-Related Injury: Bayesian Network Model of Data From a Survey of the Florida National Guard. Am. J. Epidemiol. 185, 135–146. https://doi.org/10.1093/aje/kww074

49. Urbano, J., Nogueira, P., Rocha, A.P., Cardoso, H.L., 2017. Analysis of Data Science Tools for Sensor-Based Assessment of Quality of Life in Health Care, in: Recent Advances in Information Systems and Technologies. Springer, Cham, pp. 446–455. https://doi.org/10.1007/978-3-319-56535-4_45

50. Van der Aalst, W.M., 2016. Process mining: data science in action. Springer.

51. Wang, Y., Priestley, J., 2017. Binary Classification on Past Due of Service Accounts using Logistic Regression and Decision Tree. Grey Lit. PhD Candidates.

52. Yogatama, D., Dyer, C., Ling, W., Blunsom, P., 2017. Generative and Discriminative Text Classification with Recurrent Neural Networks. ArXiv170301898 Cs Stat.

53. Yu, C., Wang, N., Yang, L.T., Yao, D., Hsu, C.-H., Jin, H., 2017. A semi-supervised social relationships inferred model based on mobile phone data. Future Gener. Comput. Syst. 76, 458–467. https://doi.org/10.1016/j.future.2016.11.027

54. Yu, L., Zhang, Y., Jian, G., Gutman, I., 2017. Classification for Microarray Data Based on K-Means Clustering Combined with Modified Single-to-Noise-Ratio Based on Graph Energy. J. Comput. Theor. Nanosci. 14, 598–606. https://doi.org/10.1166/jctn.2017.6248

55. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X., 2017. PrivBayes: Private Data Release via Bayesian Networks. ACM Trans Database Syst 42, 25:1–25:41. https://doi.org/10.1145/3134428

56. Zhou, Y., Su, W., Ding, L., Luo, H., Love, P.E.D., 2017. Predicting Safety Risks in Deep Foundation Pits in Subway Infrastructure Projects: Support Vector Machine Approach. J. Comput. Civ. Eng. 31, 04017052. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000700