QUE ES EL DATA SCIENCE: CONCEPTOS Y DIFERENCIAS

Julio Lopez-Nunez.
Ingeniero en Informática – Universidad Francisco de Aguirre.
Candidato a Doctor en Política Educativa – Universidad de Playa Ancha. Valparaíso-Chile.
Julio.Lopez@alumnos.upla.cl
ORCID ID: https://orcid.org/0000-0002-7920-1563

RESUMEN

La irrupción de conceptos como *Big Data*, *Business Intelligence y, Machine Learning, estaría provocando distorsiones respecto a la comprensión de los ámbitos en donde opera cada uno de ellos. Confusiones de estudiantes de carreras tecnológicas o, simplemente en anuncios de la prensa especializada, provocan una necesidad de aclarar cada uno de los conceptos que giran en torno a la gestión de los datos. El presente trabajo ilustra sobre las definiciones básicas en torno a los conceptos ya mencionados*. Así, se concluye con una propuesta de ejercicio para complementar la comprensión vía actividades que giran en torno a los conceptos que dan cuerpo a este escrito.

Palabras Claves: *Big Data, Business Intelligence, Machine Learning, Data Science.*

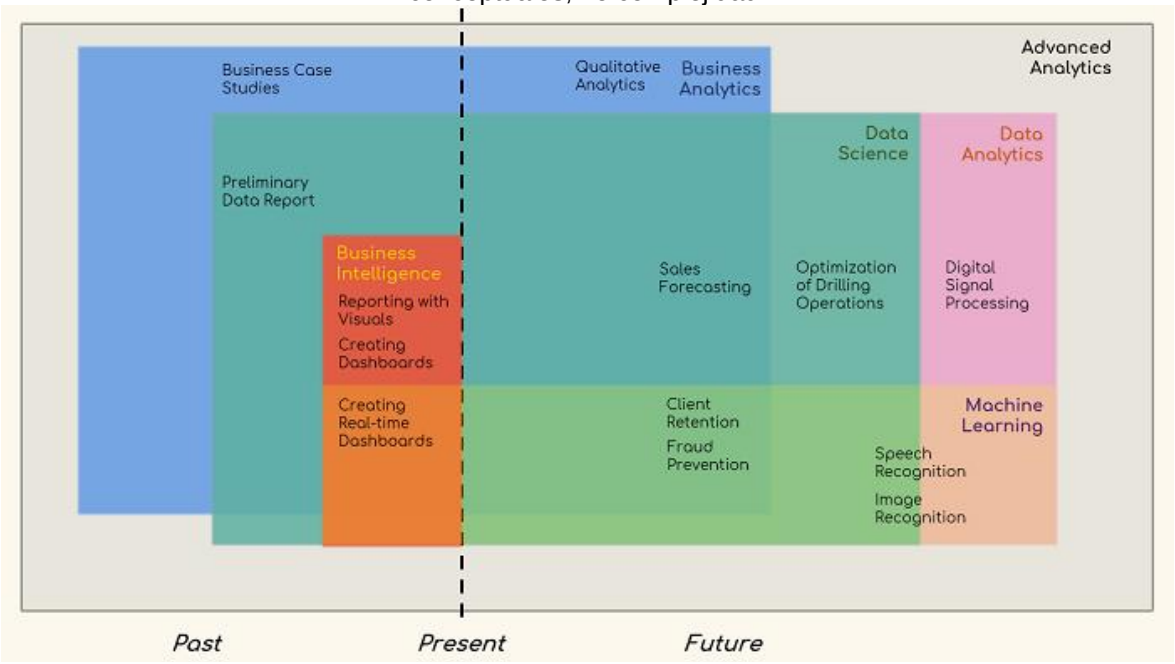
INTRODUCCIÓN

Desde mediados de los setenta que nos encontramos inmersos en una sociedad, que desde una perspectiva política, económica, filosofía y comunicacional, es denominada como la sociedad de la información (Nora y Minc, 1980; Lacroix y Tremblay, 1995; Negroponte, 1995; Miège, 1998; Castells, 2000). Sin duda, y desde una perspectiva netamente informatizada, es de dominio público que todo sistema computacional se ajusta a un comportamiento de “caja negra”, donde ingresan datos y sale información. Así entonces, en esta era predominada por el conocimiento y la información, el tratamiento de los datos resulta crucial.

Así, la teoría basada en el conocimiento (Grant, 1991; 1996, Kogut & Zander, 1992) visualiza a la empresa como comunidad que da cuenta de un acervo que gira en torno a los conocimientos. Con esto, se posiciona como la principal ventaja competitiva el conocimiento, llevando a que los datos se convirtieron en el principal activo de las grandes compañías. En esta era donde lo físico no es lo principal, el formato y tratamiento de los datos resulta ser una tarea crucial para toda empresa.

Junto a lo anterior, debemos incorporar a esta ecuación el volumen de los datos a procesar y, como enfrentamos la velocidad y variedad de datos que se generan por cada segundo. Así todo, desde la aparición del concepto *IoT (del inglés, Internet of Things[1])* propuesto hace más de 20 años por Kevin Ashton (Ramírez *et al.*, 2019), se hace patente lo señalado anteriormente: la preocupación se centra en el volumen, velocidad y variedad de los datos a procesar.

Imagen 1: La posición, tamaño y color de los rectángulos muestran similitudes y diferencias conceptuales, no complejidad.



Fuente: https://www.kdnuggets.com/2018/05/data-science-machine-learning-business-analytics.html

[1] Concepto que se refiere a una interconexión digital de objetos cotidianos con internet

Con todo, analizar los datos bajo metodologías que consideren las características que se enunciaron en el párrafo anterior, permitirá tener resultados veraces para ir en apoyo de la toma de decisiones (Mayer-Schönberger, 2013). Es en este punto donde nace el desafío. ¿Qué metodología, para el procesamiento de datos, debemos adoptar?

Entonces, como punto de partida se hace necesario aclarar la orientación que trasciende en las metodologías que se relacionan con el tratamiento de grandes almacenes de datos, estructurados y no estructurados. Con esto, hacemos un reconocimiento a todos los conceptos que comúnmente se tienden a ubicar bajo una raíz común reconocida como Ciencia de Datos (Ver Imagen 1). Así, de esta forma lo que se busca es resaltar el hecho que no existiría una claridad respecto a los límites que dividen cada uno de los conceptos mostrado en la Imagen 1. Junto a lo anterior, la propia Imagen 1 puede ser una guía para comprender como cada estos conceptos de apoyan, y así construir el derrotero hacia la correcta gestión de los datos.

DEFINICIONES

Ciencia de datos es un término con más de 20 años de existencia, ya desde 1997 fue acuñado en la exposición llamada *Statistics = Data Science?* (Jeff Wu, 1997). Así es como, dicho trabajo describe la función estadística como una componente que agrupa los procesos de recolección de datos, análisis y modelado de datos, junto a la toma de decisiones (Diggle, 2015). De esta forma, la ciencia de datos corresponde a un área que involucra el trabajo que se ajusta a métodos científicos y, que involucran procesos y sistemas para extraer conocimiento de los datos. Todo lo anterior, se debe reconocer como una extensión de la minería de datos, el aprendizaje automático, y la analítica predictiva.

En tanto, el *Big Data* no se resume exclusivamente a “muchos datos”, en realidad este concepto nos apoya para lograr un mejor entendimiento de cómo se deben tratar los datos. A esto, se debe incluir la forma de obtener todos los datos necesarios para abordar un correcto análisis. En esta era donde el volumen de

datos que se generan supera con creces las capacidades de procesamiento tradicional, el *Big Data* nos hace reflexionar sobre el cómo capturar, analizar y almacenar conjuntos de datos que se caracterizan por su gran volumen, su amplia variedad y, la velocidad en su proceso de generación (Diggle, 2015). Por último, es necesario señalar que el *Big Data* presta especial atención, adicionado a lo ya señalado, a la veracidad de los datos. El cuidado particular sobre esta cuarta característica del *Big Data*, hace referencia al grado de fiabilidad de los datos a ser procesados (Ramírez *et al.*, 2019).

En tanto, la Inteligencia de negocios *(Business Intelligence*[2]*)* corresponde a un concepto que debe ser entendido como un conjunto de productos y servicios que permite a los usuarios finales, analizar de forma expedita y sencilla grandes conjuntos de datos. Con esto, el *Business Intelligence* es posible caracterizarlo como la vía para transformar los datos en Información y la Información en conocimiento (Curto y Caralt, 2018). De esto último, es necesario resaltar el hecho que las tareas asociadas al *Business Intelligence* se descompone en un 20% de actividades relacionadas con los procesos de normalización y, un 80% de análisis de la información (Cano, 2007). Por último, resulta de toda necesidad señalar que el *Business Intelligence* analiza hechos pasados, con tal de comprender el presente y, tomar decisiones basadas en el conocimiento previo (Experiencia).

Con lo expuesto y, con el fin de completar una visión general sobre los conceptos asociados a la Ciencia de Datos, el recorrido finaliza con un concepto que podría concentrar la mayor atención de la comunidad: La inteligencia Artificial.

Tal vez, resulta del todo natural asociar la inteligencia artificial con la robótica, esto atribuido por la injerencia del cine de ciencia ficción. Sin embargo y, no despreciando lo señalado, los dispositivos dotados de inteligencia artificial podrán desarrollar procesos emulando el comportamiento humano (Russell y Norvig, 2009). Imagine el chat de atención al cliente de su banco, el cual no es operado

[2] También reconocido como inteligencia empresarial, inteligencia de negocios o BI (del inglés business intelligence). Se trata de un conjunto de estrategias, aplicaciones, datos y productos los cuales están enfocados a la administración y creación de conocimiento a través del análisis de los datos. (Dedić et al.; 2016)

por un ser humano, sino mas bien por un software que entrega respuestas estándar, que nada solucionan (hasta ahora solo orientan).

En esta línea, el *Machine Learning* corresponde a una rama de la Inteligencia Artificial, la cual se encarga de generar algoritmos dotados de una capacidad de aprender. Con esto, evitamos que los programadores inviertan demasiado tiempo en explorar los escenarios posibles que pueden enfrentar los procesos de cualquier industria. Así, los algoritmos Machine Learning, se clasifican en Modelos lineales, Modelos de árbol y Redes Neuronales (Sandoval, 2017). Y, respecto a su ámbito de acción, el *machine learning* usa el pasado para predecir el futuro.

Es en este punto donde se torna crucial establecer algunas de las diferencias que separan el *Machine Learning* y el *Business Intelligence*. Lo anterior resulta interesante cuando nos enfocamos en los ámbitos de acción de estos dos conceptos.

El *Machine Learning*, permite detectar patrones en grandes volúmenes de datos individuales. Con esto, el *Machine Learning* se convierte en aplicaciones predictivas, las cuales apoyan los procesos decisionales. Junto a lo anterior, esta técnica dota a los algoritmos con una capacidad de aprendizaje automático, lo cual los habilita para adaptarse a los posibles cambios que impacten a la organización (Sandoval, 2018).

Por el contrario, el *Business Intelligence* realiza tareas asociadas a la recolección de datos, los cuales provienen generalmente de base de datos transaccional, espacio destinado generalmente para registrar los resultados de las operaciones del negocio. En esta línea, el proceso de recolección de datos es una acción con un alto grado de participación humana. Aquí es donde florecen las actividades de extracción, transformación y carga de datos (ETL)

Así es como, con tal de almacenar el resultado de las actividades ETL, ingresa a este cuadro el concepto de *Data Warehouse*[3]. Con lo señalado, en algún instante

[3] Colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza

el proceso ETL podría cruzar la frontera y, ser denominado bajo el rotulo de Minería de Datos (*Data Mining*). Por esto, es necesario comprender que las tareas propias de la minería de datos son aquellas que permite develar la información útil que se encontraría en repositorios de datos de gran tamaño. Así es como, el concepto de Minería de datos se encontraría en el límite que lo separa de otros campos como lo son la Estadística, Bases de Datos, *Machine Learning*, Inteligencia Artificial, Visualización de Datos y las Matemáticas (Dávila y Sánchez, 2012).

Con esto, la minería de datos trabaja con datos, los cuales podrían ir desde una matriz con pocas observaciones numéricas, hasta una matriz compleja de millones de observaciones con miles de variables. En sí, la minería de datos utiliza métodos computacionales especializados, con tal de hacer visible estructuras significativas y útiles en los datos (McCue, 2007).

Continuando, los analistas disponen de variadas técnicas de visualización de datos, todos ellos basados en los registros almacenados en un *Data Warehouse*. Así es como, estas herramientas permiten crear paneles visuales (D*ashboards*[4]) para hacer accesible la información, encontrando en este punto el enlace con el *Business Intelligence.*

Dicho lo anterior, es necesario destinar algunas líneas para un último concepto, los *DataMart*. Su relevancia encuentra espacio cuando evocamos las aportaciones de Curto y Caralt (2018) quienes identifican un *DataMart* como un subconjunto de datos pertenecientes a un *Data Warehouse*. Mientras, la finalidad de estos subconjuntos radica en dar respuesta a un determinado análisis, función o necesidad, con una población de usuarios específica.

En esta línea, los *DataMart* pueden ser clasificados, según el tipo de función que desarrollan, en dos dimensiones. A saber:

Cubos OLAP (*On-Line Analytical Processing*), los cuales se estructuran agregando requisitos aportados por un área, por diferentes dimensiones o, por diversas

[4] Interfaz gráfica que permite realizar análisis de datos, generalmente provenientes de diferentes fuentes.

estructuras multidimensionales. Para todo lo anterior, el insumo principal de datos proviene de variadas fuentes, entre ella las bases de datos, informes de negocios, ventas y mercadotecnia, entre otros. Su objetivo principal radica en las funciones capaces de analizar los datos almacenados.

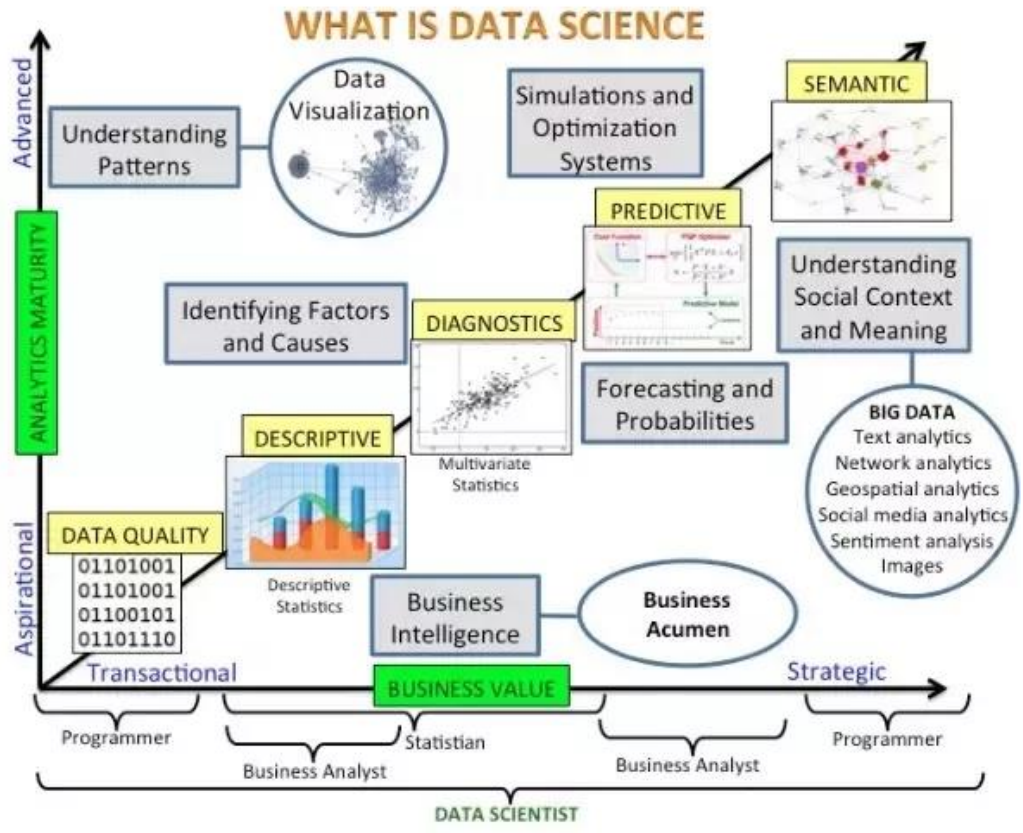
En tanto, los Cubos OLTP (On Line Transaction Processing), corresponden a *DataMart* construidos en base a la ingesta automatizada de datos, los cuales entregan la posibilidad de integrar y procesar datos. En concreto estos *DataMart* describen su función principal como aquella capaz de interactuar sobre los registros de datos, posibilitando su modificación y/o agregación.

Dicho esto, el análisis de los datos, independiente de su origen, almacenamiento o cantidad, nos permiten concluir supuestos que viajan desde la descripción hasta la interpretación del comportamiento humano. Con esto, la ciencia de datos viajara desde un nivel básico hacia el avanzado en tanto cual sea la gestión que realiza sobre los datos (según su perfil técnico dentro la organización), iniciando en la captura de datos para realizar una representación gráfica de ellos, llegando a realizar análisis explicativos y predictivos (ver Imagen 2). Pese a esto, es necesario considerar que algunos autores señalan que la calidad de los datos es más importante que su volumen (Mircea y Stoica, 2016).

Dado el contexto actual, el reconocimiento de los datos como el principal activo de la empresa, la proliferación de herramientas disponibles para el tratamiento de datos y, la penetración de los conceptos descritos es necesario señalar hasta donde es posible llegar con el análisis de los datos. En epítome, la gestión de los datos permite realizar pronósticos del comportamiento humano.

Así es como, es posible señalar que los análisis descriptivos, los cuales se basan en análisis de datos históricos y actuales, permiten determinar las relaciones y tendencias para identificar acciones posibles para hacer frente a un evento decisional (Mircea y Stoica, 2016). De esta forma, este tipo de análisis corresponde a un aprendizaje supervisado que sólo puede clasificar lo sucedido con anterioridad.

Imagen 2: Fronteras entre *Business Intelligence* y *Data Science.*



Fuente: https://www.linkedin.com/pulse/business-intelligence-data-science-fuzzy-borders-rubens-zimbres/

En tanto, un segundo tipo de análisis corresponde al tipo predictivo, el cual tiene su ámbito de acción sobre los datos pasados para determinar una probabilidad de ocurrencia. Este tipo de análisis aporta una guía a los procesos decisionales, pero siempre basado en una probabilidad. Con esto, este tipo de análisis corresponde a un sistema de aprendizaje no supervisado, el cual referencia el punto en donde se encuentra y, adonde se va, según el tipo de decisión que se tome (Garaigordobil *et al.*, 2003).

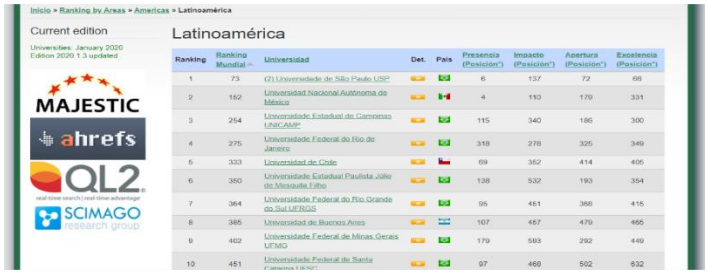
Continuando, el análisis prescriptivo se encarga de analizar todos los elementos de una decisión, combinando los resultados que entregarían los análisis descriptivos y predictivos. Se trataría de un análisis basado en un aprendizaje no supervisado y que permite recomendar acciones basadas en los primeros dos análisis ya descritos (Rosemann *et al.*, 2012).

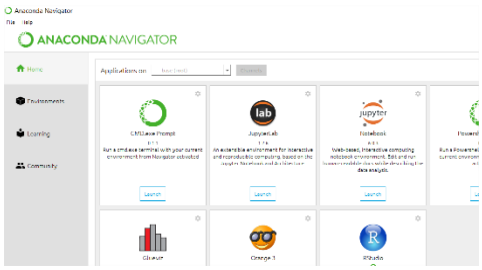
Finalmente, el análisis causal desentraña los motivos por los cuales se producen los resultados de la empresa, una medida que busca encontrar las causas que originan los problemas (Finnie y Barker, 2005)

PROPUESTA PRACTICA

Considerando las definiciones aportadas para los conceptos de *Data Warehouse*, ETL y *Data Mart*, utilice Pentaho Data Integration, con tal de ejecutar tareas de Webscraping[5] sobre la información publicada en https://www.webometrics.info/es

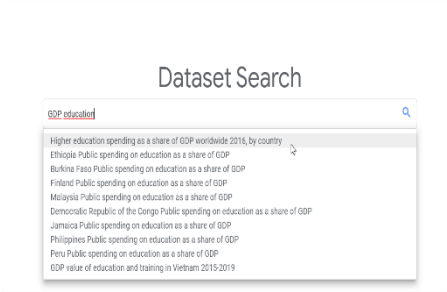
En particular, se solicita que almacene los datos publicados en el último ranking de universidades latinoamericanas en *Web Metrics*.





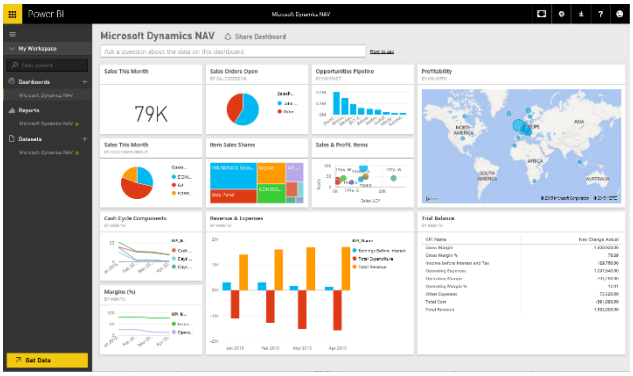
Los datos recuperados desde WebMetrics deberán ser tratados con Anaconda. La finalidad será mostrar las medidas de centralidad (media y mediana), de variabilidad (desviación estándar y coeficiente de variabilidad) y de forma (asimetría y curtosis).

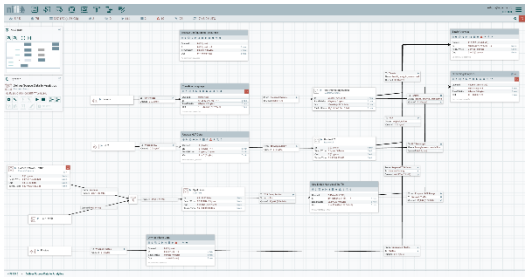
[5] Técnica utilizada mediante programas de software para extraer información de sitios web.



Posteriormente, se solicita que genere una vista de los datos correspondientes al Gasto en educación como porcentaje del Producto Interno Bruto. Utilice el servicio de *Google DataSetSearch* para obtener los datos solicitados https://datasetsearch.research.google.com/

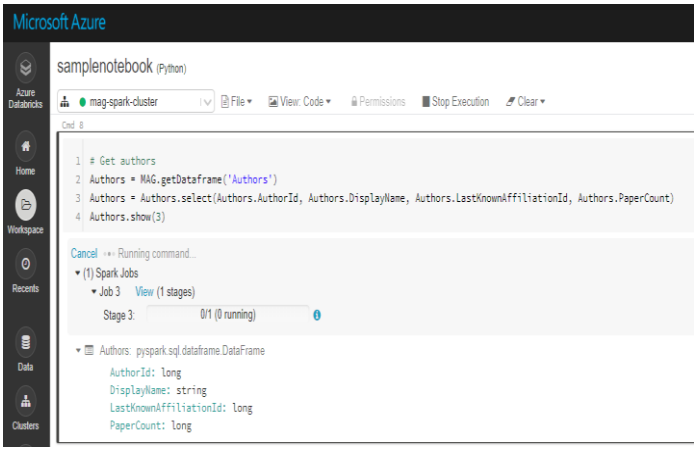
Para lo anterior, construya un tablero (*Dashboard)* utilizando la herramienta *PowerBI de Microsoft* (https://powerbi.microsoft.com/es/).
Este tablero deberá incluir los datos de Web Metrics y DataSetSearch.





Una vez construido el dashboard, realice un análisis sobre la presencia del tópico “educación”, en las publicaciones en español que muestra Twitter.

Una vez conocido el gasto en educación, la importancia de la educación para los ciudadanos latinoamericanos y, la posición que ocupan las universidades de cada país en un ranking mundial. Construya un modelo de regresión lineal que permita explicar el resultado de las universidades en el ranking publicado por WebMetrics. Para esto debe utilizar el servicio de Databricks.



REFERENCIAS

Cano, J. L.(2007). BUSINESS INTELLIGENCE: Competir Con Información. Barcelona.

Castells, Manuel (2000), La era de la información. La sociedad red,vol. I, Siglo XXI, México

Curto, J. y Caralt, J.(2018). Introducción al Business Intelligence. Barcelona: Editorial UOC.

Dávila, F., y Sánchez, Y. (2012). Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. Revista Cubana de Informática Médica, 4(2), 174-183.

Dedić, N., & Stanier, C. (2016, December). Measuring the success of changes to existing business intelligence solutions to improve business intelligence reporting. In International Conference on Research and Practical Issues of Enterprise Information Systems (pp. 225-236). Springer, Cham.

Diggle, P. (2015). Statistics: a data science for the 21st century. J. R. Statist. Soc. A (2015) 178, Part 4, pp. 793–813.

Finnie, G. y Barker, J. (2005). Real-Time Business Intelligence in Multi-Agent Adaptive Supply Network. ePublications, pp. 1-5, 2005.

Garaigordobil, M., Cruz, S. y Pérez, J. (2003) A correlational and predictive analysis of self-concept with other behavioural, cognitive and emotional factors of personality during adolescence, Studies in Psychology, 24:1, 113-134, DOI: 10.1174/021093903321329102

Grant, R.M. (1997). “Dirección Estratégica. Conceptos, Técnicas y Aplicaciones”. Civitas, Madrid. (Título original: “Contemporary Strategy Analysis: Concepts, Techniques, Applications”. 2nd. Edition. BlackwellmPublishers. Cambridge). Jeff Wu,C.F. (1997) Statistics=Data Science?. University of Michigan.

Kogut, B. y Zander, U. (1992). “Knowledge of the Firm, Combinate Capabilities, and the Replication of Technology”. Organization Science, Vol. 3 (3), pp. 383-397.

Lacroix, Jean-Guy y Gaëtan Tremblay (1995), Les autoroutes de l’information. Un produit de la convergence, Presses de l’Université du Québec, Canadá.

Mayer-Schönberger, V., Cukier, K. y Iriarte, A. (2013). Big data. Madrid: Turner.

McCue, C. (2007). Data Mining and Predictive Analysis.

Mircea, M. y Stoica, M. (2016). Combining Business Intelligence with Cloud Computing to Delivery Agility in Actual Economy. CiteSeerX, pp. 1-16.

Miège, Bernard (2000), Les industries du contenu face à l’ordre informationnel, Presses Universitarires de Grenoble, Francia.

Negroponte, Nicholas, Ser digital, México, Océano, 1995.

Nora, Simón y Alan Minc (1980), Informatización de la sociedad,Fondo de Cultura Económica, México.

Ramírez, M., Vázquez, S., Manrique, E., Ramírez, H.(2019). Business Intelligence and Big Data. En, 14th Iberian Conference on Information Systems and Technologies (CISTI).19 – 22 June 2019, Portugal.

Rosemann, M., Eggert, M., Voigt, M., y Beverungen, D. (2012). Leveraging social network data for analytical CRM strategies-The introduction of social BI. AIS Electronic Library, pp. 95-101.

Russell, J., y Norvig, P. (2009). Artificial intelligence: a modern approach (3.ª edición). Upper Saddle River, N.J.: Prentice Hall. ISBN 0-13-604259-7.

Sandoval, J. (2017). Machine learning algorythms for analysis and data prediction, 2017 IEEE 37th Central America and Panama Convention (CONCAPAN XXXVII), Managua, 2017, pp. 1-5, doi: 10.1109/CONCAPAN.2017.8278511.

Sandoval, L. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. Revista Tecnológica; no. 11. Disponible: http://redicces.org.sv/jspui/handle/10972/3626