ANÁLISIS DE DATOS AGROPECUARIOS

IVÁN RAMÍREZ-MORALES / BERTHA MAZON-OLIVO

Editorial UTMACH

REDES 2017

COLECCIÓN EDITORIAL

Análisis de Datos Agropecuarios

Iván Ramírez-Morales
Bertha Mazon-Olivo

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

302 pag; 22X19cm - (Colección REDES 2017)

Título: Análisis de Datos Agropecuarios. / Iván Ramírez-Morales / Bertha Mazon-Olivo (Coordinadores)

ISBN: 978-9942-24-120-7

Publicación digital

Título del libro: Análisis de Datos Agropecuarios.

ISBN: 978-9942-24-120-7

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

Introducción

El análisis de datos es un proceso complejo que trata de encontrar patrones útiles y relaciones entre los datos a fin de obtener información sobre un problema específico y de esta manera tomar decisiones acertadas para su solución.

Las técnicas de análisis de datos que son exploradas en el presente libro son actualmente utilizadas en diversos sectores de la economía. En un inicio, fueron empleadas por las grandes empresas a fin de incrementar sus rendimientos financieros.

El libro se basa en la aplicación de la especialización inteligente, de este modo, gracias al trabajo colaborativo, se combina al sector agropecuario con las tecnologías, matemáticas, estadística y las ciencias computacionales, para la optimización de los procesos productivos.

La idea de descubrir la información oculta en las relaciones entre los datos, incentiva a encontrar aplicaciones para el sector agropecuario, por ejemplo los obtenidos de una producción avícola, o los datos que se generan durante los procesos de fermentación, los parámetros físicos y químicos del suelo, del agua y de las plantas, los datos de sensores, de espectrometría, entre otros.

En la actualidad, este sector se ha mantenido con su producción habitual sin un destacado repunte ni diferenciación, a pesar de existir herramientas científicas que han permitido desarrollar dispositivos tecnológicos y sus aplicaciones.

Este libro ha sido el resultado de la sistematización de las experiencias individuales de un equipo humano con objetivos comunes y una historia académica multidisciplinar, cuyos hallazgos de investigación han sido publicados en revistas científicas y conferencias de alto impacto. El área temática sobre la que se centra este texto es en técnicas de extracción, procesamiento y análisis de datos del ámbito agropecuario, se combinan para entregar al lector una obra de calidad y alto valor científico.

Así, el presente libro está concebido desde diferentes puntos de vista de profesionales agrónomos, informáticos, electrónicos, matemáticos, estadísticos y empresarios. Todos buscan un objetivo en común: "descubrir el conocimiento oculto en los datos que proporcione una ventaja competitiva". Se aborda el ciclo completo del proceso de obtención de conocimiento a partir de datos crudos del sector agropecuario, con la finalidad de apoyar la toma de decisiones. Este ciclo involucra procesos de: selección de los datos (extracción, comunicación, almacenamiento), pre-procesamiento, transformación, aplicación de modelos y/o técnicas de análisis, presentación e interpretación de resultados. El enfoque temático del libro es el siguiente:

Capítulo 1: Ciencia de Datos en el sector Agropecuario.- En este capítulo se aborda una revisión desde los inicios del análisis de datos en el sector agropecuario hasta el progreso actual que se ha dado en esta área del conocimiento que se considera como la nueva revolución en la agricultura y la ganadería de precisión.

Capítulo 2: Obtención de datos en sistemas agropecuarios.- El enfoque del capítulo es la generación de datos crudos en los sistemas agropecuarios, aplicando métodos y técnicas básicas donde se registran información de: número de unidades producidas, cantidad de nutrientes, variables climáticas, muestreo y monitoreo de organismos vivos, entre otros.

Capítulo 3: Internet de las cosas (IoT).- Este capítulo aborda los sistemas de telemetría para obtención de datos y control de dispositivos, aplicando tecnologías como: redes de sensores inalámbricos (dispositivos electrónicos, sensores, actuadores y puertas de enlace), protocolos de comunicación, centros de procesamiento de datos (cloud computing) y aplicaciones IoT para el sector agropecuario.

Capítulo 4: Matemáticas aplicadas al sector agropecuario.- Este capítulo explica los procedimientos para la creación de modelos matemáticos determinísticos que representen procesos asociados al sector agropecuario, como una alternativa de solución en la ingeniería.

Capítulo 5: Estadística básica con datos agropecuarios.- El capítulo se enfoca en los atributos, escalas de medición de las variables, su influencia en la elección del procedimiento estadístico a desarrollar, así como, el papel de las medidas de resumen, estimación puntual y prueba de hipótesis en la investigación científica.

Capítulo 6: Estadística predictiva con datos agropecuarios.- El capítulo considera las principales técnicas de la estadística avanzada aplicada al sector agropecuario, con el propósito de establecer predicciones que permita tomar mejores decisiones.

Capítulo 7: Inteligencia de negocios en el sector agropecuario.- El capítulo comprende la obtención de conocimiento a partir de datos crudos con la finalidad de apoyar la toma de decisiones en empresas del sector agropecuario. Involucra procesos de extracción, transformación y almacenamiento de datos en nuevos almacenes (Data warehouse - Big Data), distribución y análisis de la información con técnicas: multidimensional OLAP y tableros de control (dashboards).

Capítulo 8: Inteligencia Artificial aplicada a datos agropecuarios.- El capítulo trata sobre las principales técnicas de machine learning aplicadas a los datos agropecuarios, entre éstas se destacan: las redes de neuronas artificiales, máquinas de soporte de vectores, vecinos más cercanos, análisis de componentes principales, entre otros.

01 Capítulo Ciencia de datos en el sector agropecuario

Iván Ramírez-Morales; Bertha Mazon-Olivo; Alberto Pan

En el año 2015 de acuerdo a la revista Fortune, 151 startups de tecnología fueron financiados con $ 976 millones para enfocarse en el análisis de datos agropecuarios para la producción de alimentos. Transnacionales como Monsanto, Dupont y Archer Daniels Midland están invirtiendo importantes sumas de dinero en programas de ciencia de datos agrícolas.

Iván Ramírez-Morales: Doctor en Medicina Veterinaria y Zootecnia por la Universidad Agraria de la Habana, Máster en Desarrollo Comunitario por la Universidad Nacional de Loja y Doctor en Tecnologías de la Información y de las Comunicaciones por la Universidade A Coruña, ha realizado varios cursos en Brasil, Japón, Perú y Argentina. Fue Oficial de Territorio del Programa Marco ART/PNUD de la ONU, y Director de Planificación del Gobierno Provincial de El Oro. Actualmente es Profesor Titular en la Universidad Técnica de Machala, su área de investigación se centra en el uso de tecnologías para el mejoramiento de la productividad agropecuaria, Cuenta a la fecha más de 15 publicaciones indexadas, varias de ellas en revistas de alto impacto en los índices de JCR y SJR.

Bertha Mazon-Olivo: Ingeniera en Sistemas y Magíster en Informática Aplicada por la Escuela Superior Politécnica de Chimborazo. Profesora Titular en la Universidad Técnica de Machala. Es estudiante del programa doctoral en Tecnologías de la Información y las Comunicaciones en Universidade da Coruña, España. Sus líneas de investigación son: Internet de las Cosas, Ciencia de Datos y Desarrollo de Aplicaciones Informáticas. Cuenta con varias publicaciones indexadas.

Alberto Pan: Director Técnico de Denodo y Profesor Asociado de la Universidad de A Coruña. Recibió una Licenciatura en Ciencias de la Computación en la Universidad de A Coruña en 1996 y un Ph.D. en Informática por la misma universidad en 2002. Sus intereses de investigación están relacionados con la extracción e integración de datos y la automatización de la web. Alberto ha dirigido varios proyectos a nivel nacional y regional en el campo de la integración de datos y acceso a la Web oculta. También es autor o coautor de numerosas publicaciones en revistas científicas y actas de congresos.

Este movimiento de inversiones tiene lógica, ya que en los últimos años han emergido un número creciente de maquinarias, equipos y sistemas de monitoreo y control para el sector agropecuario que generan datos, y estos datos requieren ser interpretados. Entre los dispositivos generadores de datos se encuentran los tractores, arados, sembradoras, cosechadoras, las redes de sensores, estaciones meteorológicas, los espectrómetros portátiles, drones, cámaras, imágenes de satélite, invernaderos y galpones inteligentes, entre otros.

Por este motivo el Departamento de Agricultura de los Estados Unidos ha invertido en su iniciativa OpenAg, que ha hecho públicos sus datos. Esta situación implica un necesario esfuerzo multidisciplinario entre profesionales de las ciencias de datos y del sector agropecuario, para dar respuesta a la demanda mundial generada y promover una producción de alimentos más eficiente.

Los datos que se pueden recolectar o generar, con el uso de las Tecnologías de la Información y Comunicación (TIC) en la agricultura, pueden ser (Bendre, Thool, & Thool, 2015; Wolfert, Ge, Verdouw, & Bogaardt, 2017): datos en campo (características físico-químicas del suelo, topografía, datos de productividad), datos del clima, datos derivados de la interpretación de imágenes de cámaras o satelitales (variabilidad espacial y/o temporal del cultivo), datos de mapas de rendimiento, datos de mapas con prescripciones de aplicación de insumos, datos históricos, otros datos internos o externos. La integración y análisis de estos datos, pueden servir para generar decisiones automáticas (que se ejecutan en dispositivos, robots o maquinaria) o apoyar las decisiones humanas.

La agricultura de precisión permite incrementar la eficiencia y productividad aplicando un enfoque científico durante todas las fases del ciclo de producción de plantas y animales. Se plantea que la próxima gran revolución en la agricultura de precisión será a través de la ciencia de datos.

La optimización de los procesos productivos cada vez van evolucionando al fusionar los métodos, técnicas y tecnologías de la Agricultura de Precisión con la Ciencia de Datos, permi-

tiendo disminuir el uso de recursos como: energía, agua, fertilizantes, plaguicidas, etc.; lo que conlleva a producir más alimentos con menos esfuerzo, costos e impactos ambientales.

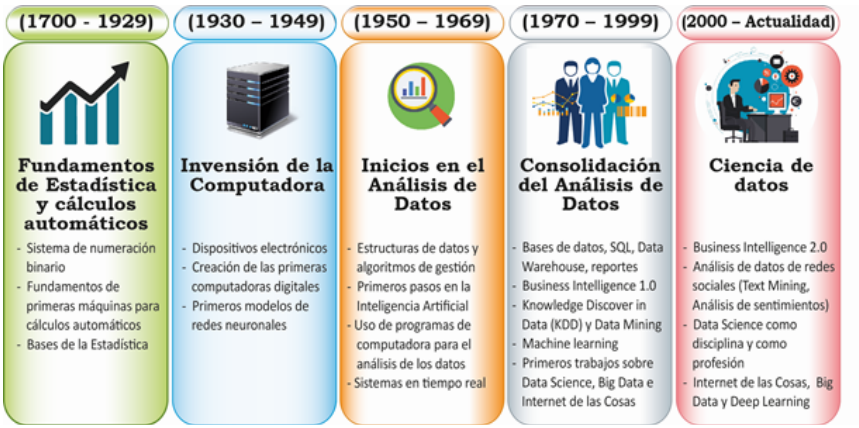
En la actualidad emergen nuevas técnicas y algoritmos para el procesamiento de los datos. La aplicación de estas técnicas involucra una diversidad de disciplinas.

Este capítulo se enfoca en la sistematización y comprensión de la evolución de los procesos de análisis de datos, los fundamentos teóricos de la Ciencia de Datos, los tipos de análisis de datos, las técnicas enfocadas al análisis de los datos agropecuarios aplicando las TIC, las herramientas y demás terminología asociada.

Evolución de la ciencia de datos

La Ciencia de Datos es una disciplina relativamente joven, en comparación con otras disciplinas con las que guarda relación como la Estadística y las Ciencias de la Computación. Varios aportes relacionados con la evolución de Data Science (Capgemini, 2015; Jifa & Lingling, 2014; Press, 2013) fueron revisados para elaborar el resumen de hitos más relevantes que se describen a continuación (ver Imagen 1.1):

Imagen 1.1: Evolución de la Ciencia de Datos



Fuente: Elaboración propia

Hitos importantes en la evolución de la ciencia de datos

Fundamentos de estadística y cálculos automáticos (1700 – 1929)

A continuación se describen los hitos que establecieron las bases teóricas de la Estadística y de los cálculos automáticos, como los primeros aportes para la construcción de un computador.

- 1703: Leibniz G. propuso la utilización del sistema de numeración binario para cálculos sencillos. En la actualidad es la base de los cálculos automáticos que realiza un computador.
- (1791-1871): Charles Babbage es conocido como el padre de la computación. Diseñó e implementó parcialmente una máquina de diferencias mecánicas para calcular tablas de números. Además, diseñó una máquina analítica para ejecutar programas de tabulación o computación.
- (1815-1852): Ada Lovelace es conocida como la primera programadora de la historia que contribuyó con Babbage.
- 1847: George Boole siguió los pasos de Leibniz y propuso el Álgebra de Boole, que consiste en aritmética aplicada al sistema de numeración binario.
- 1749: Gottfried Achenwall introdujo el término alemán Statistik, que designaba originalmente el análisis de datos del Estado, es decir, la "ciencia del Estado". Sin embargo, no fue hasta el siglo XIX cuando el término Estadística adquirió el significado de recolectar y clasificar datos, concepto introducido por el militar británico Sir John Sinclair (1754-1835).
- 1805. Legendre propone los métodos de mínimos cuadrados para análisis de regresión que consiste en un proceso estadístico para estimar las relaciones entre variables.

- 1890: Herman Hollerith usó las tarjetas perforadas en el censo de Estados Unidos, en los siguientes años, fueron el medio para el ingreso y almacenamiento de datos.

Invención de la computadora (1930 – 1949)

En estas dos décadas surgió definitivamente la invención de los dispositivos electrónicos y las primeras computadoras digitales.

- 1930's a 1940's: Se crearon unas pocas computadoras con fines militares, académicos e investigativos.
- 1936: Alan Turing propuso la teoría de "Máquina Universal de Turing" donde se estableció los principios del proceso de cómputo de cualquier computador digital.
- 1937: Claude Shannon a través de su Teoría de la Información, hizo posible la aplicación del álgebra de Boole en los dispositivos electrónicos, que fueron usados en la construcción de las primeras computadoras.
- 1943: Warren McCulloch y Walter Pitts, proponen los primeros modelos de redes neuronales, basadas en modelos matemáticos e informáticos.
- 1949, Von Neuman propone "La Arquitectura de Von Neuman" que buscó mejorar la arquitectura del computador ENIAC para dar origen al EDVAC y al primer computador comercial UNIVAC I, y sentó las bases del resto de computadoras digitales inventadas hasta la actualidad.

Inicios en el análisis de datos (1950 – 1969)

Creación de los primeros modelo de datos, algoritmos de tratamiento y predicción e inicios de la inteligencia artificial.

- 1950's: En esta década surgen algoritmos de ordenación y búsqueda en estructuras de datos, el procesamiento por lotes, el almacenamiento temporal de datos.
- 1950: Primer modelo de predicción meteorológica propuesta por un equipo meteorólogos estadounidenses, empleando la computadora ENIAC.

- 1950: Turing dio los primeros pasos en la inteligencia artificial con su artículo "Computing Machinery and Intelligence", donde hizo la siguiente pregunta ¿puede pensar una máquina? El enfoque de Turing consistió en ver a la inteligencia artificial como una imitación del comportamiento humano.
- 1957 - Frank Rosenblatt diseñó la primera red neuronal para computadoras (el perceptrón), que simula los procesos mentales del cerebro humano.
- 1958: Hans Peter Luhn (de IBM), en el artículo "A Business Intelligence System", define la Inteligencia de Negocios, como: "la habilidad de aprender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada".
- 1959: Edsger Dijkstra propone el algoritmo Dijkstra de cálculo de la ruta más corta o mínima basada en la teoría de grafos para resolver problemas de transporte y logística.
- 1962: John Tukey escribió sobre "The Future of Data Analysis", donde argumenta la importancia del uso de programas de computadora en el análisis de los datos.
- 1960's: Surgimiento de sistemas en tiempo real, dando lugar al procesamiento de datos en tiempo real.
- 1965: Ingo Rechenberg, introdujo una técnica que llamó estrategia evolutiva y es el punto de partida de los algoritmos genéticos o computación evolutiva.
- 1969: Edgar Codd definió la teoría de las Base de datos relacionales (DBMS).

Consolidación del análisis de datos (1970 – 1999)

La estadística tradicional es tratada con la tecnología informática, ejecutándose tareas como: almacenamiento, procesamiento y análisis de datos.

- 1970's: Auge de las primeras bases de datos y de las aplicaciones empresariales.

- 1974: Peter Naur, publica el libro "Concise Survey of Computer Methods" con los resultados de una encuesta de métodos informáticos para el procesamiento de datos utilizados en varias aplicaciones. Además, hace algunas definiciones importantes como los conceptos de dato y ciencia de datos. En el mismo año, Donald Chamberlin define el lenguaje estructurado de consultas en bases de datos (SQL).
- 1977: The International Association for Statistical Computing (IASC), vincula la estadística tradicional con la tecnología informática moderna y el conocimiento de expertos del dominio para convertir los datos en información y conocimiento.
- 1980s: Ralph Kimball y Bill Inmon, proponen el concepto de Data Warehouse y se crean los primeros sistemas de reportes.
- 1989: Gregory Piatetsky-Shapiro, organizó el primer workshop, relacionado con el descubrimiento del conocimiento en datos, titulado "Knowledge Discover in Data (KDD)", convirtiéndose en los siguientes años en un evento anual de ACM SIGKDD "Conference on Knowledge Discovery and Data Mining"[1] hasta la actualidad.
- 1989: Howard Dresner, populariza el término Business Intelligence (BI).
- 1990s: Proliferación de múltiples aplicaciones BI de escritorio dando lugar a Business Intelligence 1.0. También es la década del aprendizaje automático "Machine Learning"[2] que pasa de un enfoque basado en el conocimiento a un enfoque basado en datos. Los científicos comienzan a crear programas para que las computadoras analicen grandes cantidades de datos y obtengan conclusiones, o "aprendan", de los resultados.

[1] http://www.kdd.org/

[2] http://todobi.blogspot.com/2016/02/una-breve-historia-del-machine-learning.html

- 1996: en una reunión organizada en Japón por los miembros de International Federation of Classification Societies (IFCS), se volvió a mencionar la ciencia de datos en la conferencia titulada "Data science, classification, and related methods".
- 1996: Fayyad, Piatetsky-Shapiro y Smyth publicaron un trabajo titulado "From Data Mining to Knowledge Discovery in Databases" donde diferencian los conceptos de KDD y Data Mining. KDD se refiere al proceso general de descubrir el conocimiento útil a partir de datos, y la minería de datos se refiere a un paso particular en este proceso. La minería de datos es la aplicación de algoritmos específicos para extraer patrones de datos.
- 1997: El profesor C. F. Jeff Wu dio la conferencia inaugural titulada "Statistics = Data Science?" para su nombramiento en la Universidad de Michigan, donde hace una petición de renombrar a la estadística como ciencia de datos y a los estadistas como científicos de datos.
- 1999: J. Zahavi en su trabajo "Mining Data for Nuggets of Knowledge" critica a los métodos estadísticos convencionales de que sólo funcionan bien con pequeños conjuntos de datos; pero, las bases de datos actuales pueden involucrar millones de filas y decenas de columnas de datos introduciendo el término "Big Data", aduciendo que la escalabilidad es un gran problema en la minería de datos. En este mismo año, Kevin Ashton, en una conferencia en Procter & Gamble, habló por primera vez de Internet de las Cosas o "Internet of Things".

Ciencia de datos (2000 – hasta la actualidad)

Uso generalizado de la analítica de datos, Inteligencia de Negocios, Big Data, Internet de las Cosas, Machine Learning y surgimiento de Deep Learning.

- 2000's – 2010's: Consolidación de las aplicaciones Business Intelligence 2.0 mediante plataformas BI (Oracle, SAP, IBM, Microsoft, Tableau, Qlik, etc.) que gestionan y analizan información de Data Warehouse y datos no estructurados. Crecimiento de la Web 2.0 o era de las

redes sociales como: Linkedin, Facebook, Twitter, Instagram; y es el comienzo de la explosión de datos.

- 2001: William Cleveland, presenta el trabajo "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" como una disciplina independiente que se extiende de la estadística e incorpora avances en computación de datos.
- 2002: Lanzamiento de las revistas "Data Science Journal"[3] y en el 2003 "Journal of Data Science"[4] .
- 2004 - 2005: Primeros desarrollo de plataformas de Big Data e Internet de las Cosas (MapReduce, Hadoop, etc.).
- 2006 - Geoffrey Hinton acuña el término Deep Learning "aprendizaje profundo" para explicar nuevos algoritmos que permiten a las computadoras ver y distinguir objetos y texto en imágenes y videos.
- 2009: Mike Loukides escribió "What is Data Science?"[5] .
- 2012: Tom Davenport y Patil en su publicación "Data Scientist: The Sexiest Job of the 21st Century" en Harvard Business Review[6], menciona que Data Science se ha convertido en una opción muy atractiva de estudio de grados de Máster y/o PhD debido a que los científicos de datos son muy demandados y mejor pagados en muchas empresas que tienen que tratar con un gran volumen y diversidad de datos.

Cabe resaltar que desde finales de los noventa, los aportes sobre ciencia de datos, como: artículos científicos, conferencias y libros se han incrementado considerablemente, proponiéndose una variedad de métodos, técnicas y algoritmos para los diferentes tipos de análisis de datos. A la par, se han desarrollado varias herramientas informáticos y lenguajes de programación que sirven para hacer los cálculos más eficientes.

[3] https://datascience.codata.org/

[4] http://www.jds-online.com/journal/

[5] https://www.oreilly.com/ideas/what-is-data-science

[6] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Fundamentos de la ciencia de datos

Ciencia de datos (data science)

La ciencia de datos consiste en la aplicación de métodos científicos para construir algoritmos y sistemas que permiten detectar patrones y descubrir conocimiento útil para la toma de decisiones. Involucra procesos de integración y análisis de datos de distintas fuentes y en una variedad de formatos, a fin de construir modelos que ayudan a identificar y comprender fenómenos complejos.

Varios autores han contribuido en la definición de Ciencia de Datos. A continuación, se citan algunos:

- "La ciencia que trata con los datos, una vez que se han establecido, mientras que la relación de los datos con lo que representan se delega a otros campos y ciencias" (Naur, 1974).
- "La extracción de conocimiento útil de los datos para resolver problemas empresariales mediante un proceso sistemático con etapas bien definidas" (Provost & Fawcett, 2013).
- "Disciplina que crea sistemas y algoritmos para descubrir conocimiento, detectar patrones, generar información útil y/o realizar predicciones a partir de datos a gran escala" (Molina-Solana, Ros, Dolores Ruiz, Gomez-Romero, & Martin-Bautista, 2017).
- "Extracción de conocimiento accionable directamente de los datos a través de un proceso de descubrimiento o formulación y prueba de hipótesis" (NIST, 2015).

Según Gartner (2014) en su reporte gráfico "Hype Cycle for Emerging Technologies Maps the Journey to Digital Business"[7] , Data Science se muestra como una tecnología o disciplina emergente, de gran expectativa por la comunidad de científicos, profesionales, empresarios y personas que les interesa obtener un valor agregado de sus datos.

[7] https://www.gartner.com/newsroom/id/2819918

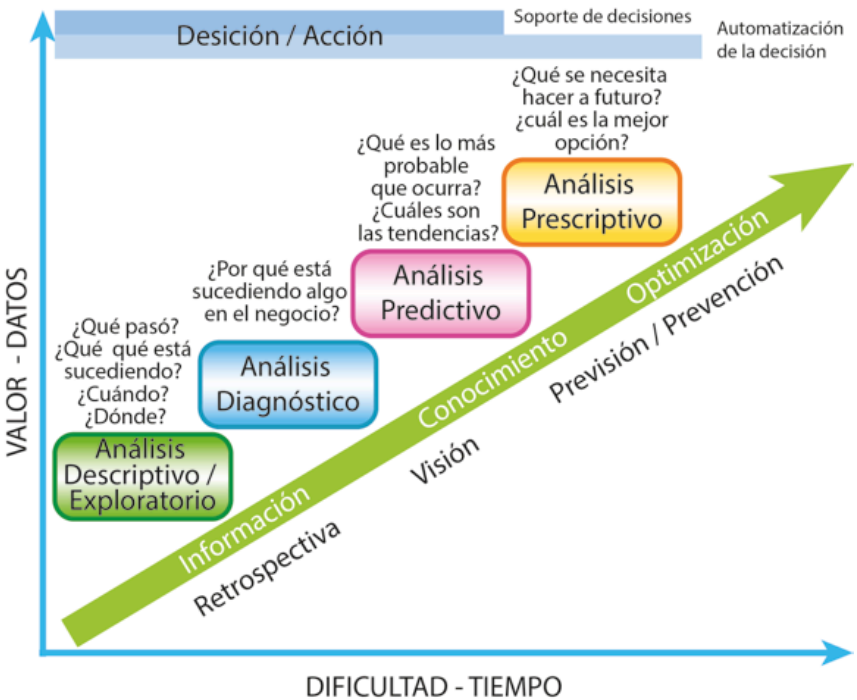
Análisis de datos (data analytics)

El proceso analítico de datos es la síntesis del conocimiento a partir de la información (NIST, 2015). El análisis de datos se traduce como el proceso de obtención, transformación y modelado de datos, con el fin de determinar patrones de comportamiento que ayuden en la toma de decisiones.

Tipos de análisis de datos

Según (Gartner, 2012; Loury, 2014; National Academi of Science, 2017; Sivarajah, 2017), los métodos analíticos de datos se clasifican en: Análisis descriptivo y exploratorio, Análisis Diagnóstico, Análisis Predictivo y Análisis Prescriptivo (ver Imagen 1.2).

Imagen 1.2: Tipos de análisis de datos



Fuente: Elaboración propia

A continuación se describen los tipos de análisis de datos:

- El análisis descriptivo y exploratorio. Implican el resumen y la descripción de patrones de conocimiento utilizando métodos de visualización de datos estáticos y dinámicos. Las fuentes de datos sin procesar (raw data) pueden ser conjuntos de datos (data sets) en distintos formatos: hoja de cálculo (xls, xlsx, ods, ect.), archivos de texto (CSV, TXT, XML, JSON, HTML, etc.), bases de datos relacionales, informes de sistemas transaccionales, etc. A menudo, este tipo de análisis sirve para crear informes de gestión que se ocupan de modelar comportamientos pasados o presentes; responden a las preguntas: ¿qué pasó en el negocio?, ¿qué ocurrió y qué está sucediendo?, ¿cuándo?, ¿dónde?

  Los análisis descriptivos son la base de información de una organización; es decir, son las principales aplicaciones de la inteligencia empresarial o de negocios (Mazon-Olivo et al., 2017). Los métodos más comunes de análisis descriptivo son: informes y consultas estáticas o personalizadas (reports & queries statics or Ad hoc), métodos estadísticos básicos (media, moda, desviación estándar, varianza, medición de frecuencia de eventos específicos, etc.), análisis multidimensional OLAP (On-Line Analytical Processing - procesamiento analítico en línea), tableros de control y cuadros de mandos (dashboards, scorecards) y otras técnicas de visualización de datos.

- Análisis diagnóstico. Consiste en sondear datos para certificar o rechazar proposiciones comerciales o hipótesis. Este tipo de análisis se basa en información descriptiva para comprender: ¿Por qué está sucediendo algo en el negocio?

- Análisis predictivo. Este tipo de análisis busca descubrir patrones y capturar relaciones entre los datos; pronosticar el resultado probable de una situación dada y generar un modelo estadístico de los datos actuales e históricos para determinar las posibilidades futuras, en base a técnicas de aprendizaje supervisado, no supervisado y

semi-supervisado; responden a las preguntas ¿Qué es lo más probable que ocurra a futuro? ¿Cuáles son las tendencias?

- Análisis prescriptivo. Este tipo de análisis se realiza para encontrar cuál es la solución entre una gama de variantes. Su tarea es optimizar recursos y aumentar la eficiencia operativa. Las soluciones prescriptivas ayudan a los analistas de la empresa, en la toma de decisiones mediante la determinación de acciones y la evaluación del impacto con respecto a los objetivos, los requisitos y las limitaciones del negocio. El análisis prescriptivo se basa en la aplicación de reglas de negocio, aprendizaje automático y procedimientos de modelado computacional, para intentar responder a la pregunta: ¿qué se puede hacer para que algo suceda?, ¿cuál es la mejor opción de decisión?, ¿qué acciones tomar?

Un tipo de análisis prescriptivo es el Análisis preventivo que consiste en tener la capacidad de tomar medidas preventivas sobre eventos que pueden influir indeseablemente en el desempeño de la organización, por ejemplo, identificar los posibles riesgos y recomendar estrategias de mitigación en el futuro.

Campos y técnicas de análisis de datos

Las Técnicas de análisis de datos más prioritarias se resumen el Cuadro 1.1, están clasificadas por categorías, tipo de análisis, preguntas y campos específicos de acción. Los aportes de varios autores (National Academi of Science, 2017; Provost & Fawcett, 2013; Sivarajah, 2017) y algunos sitios web[8] fueron revisados para elaborar este cuadro.

[3] https://www.nap.edu/read/23670/chapter/6

https://www.sv-europe.com/blog/10-reasons-organisation-ready-prescriptive-analytics/

http://www.healthcareimc.com/main/making-sense-of-analytics/

https://twitter.com/doug_laney/status/611172882882916352

Categorías del análisis de datos:

Retrospectiva (mirada al presente o pasado). Consiste en descubrir patrones de información o conocimiento en bases de datos (transaccionales estructuradas y no estructuradas o pre-procesadas como data warehouse) que registran hechos o transacciones de la organización. Se busca apoyar en la decisión a los mandos medios y estratégicos, con información procesada y organizada.

Visión (mirada al presente o pasado para predecir el futuro). Conlleva procesos de diagnóstico, predicción y de apoyo a la decisión en una organización; además del descubrimiento de patrones de información, la búsqueda de causas y efectos a posibles problemas encontrados, la comprobación de hipótesis, clasificación de datos en grupos de interés, identificación de tendencias, y/o se predicción de información faltante pasada, presente o futura.

Cuadro 1.1. Campos y técnicas de análisis de datos

*Cat.	**Ta	Preguntas	Campo Específico	Técnicas
Retrospectiva	Análisis descriptivo y exploratorio	¿Qué pasó en el negocio?	Descriptivo	Reportes y consultas estáticas mediante lenguaje estructurado de consultas (SQL)
		¿Qué ocurrió y qué está sucediendo? ¿Cuándo? ¿Dónde?	Descriptivo de apoyo a la decisión Inteligencia de Negocios	Consultas y reportes personalizados (Ad Hoc) Análisis multidimensional (OLAP) Indicadores clave de desempeño (KPI's) Tableros de control y cuadros de mando (dashboards y scorecards) Métodos estadísticos básicos

*Cat.	**Ta	Preguntas	Campo Específico	Técnicas
Visión	Análisis Diagnóstico	¿Por qué está sucediendo algo en el negocio?	Descriptivo, exploratorio y apoyo a la decisión Inteligencia de Negocios Minería de datos descriptiva Análisis situacional Causa y efecto	Análisis OLAP con distintos niveles de detalle (slice and dice, drill and down, across) Tableros de control y cuadros de mando (dashboards y scorecards) Monitoreo automático y alertas Clustering o segmentación, reglas de asociación, patrones secuenciales Pruebas de hipótesis Minería de textos, análisis de sentimientos u opiniones Otras técnicas
	Análisis Predictivo	¿Qué información que se desconoce? ¿Qué es lo más probable que ocurra a futuro? ¿Cuáles son las tendencias?	Predictivo y apoyo a la decisión Minería de datos predictiva Modelos de análisis de tendencias Evaluación de probabilidad y de riesgos Análisis de big data (datos a gran escala) Análisis de datos no estructurados: imágenes, audios, videos, archivos .pdf, etc.	Aprendizaje de máquina (machine learning), aprendizaje profundo (deep learning) Regresión lineal y no lineal Árboles de decisión Métodos bayesianos Redes neuronales Series temporales Máquina de soporte vectorial (Support-vector machines) Otras técnicas

*Cat.	**Ta	Preguntas	Campo Específico	Técnicas
Previsión / prevención	Análisis Prescriptivo	¿Qué se necesita hacer a futuro? ¿Cuál es la mejor opción de decisión? ¿Qué acciones tomar? ¿Cómo actuar?	Previsión, recomendación Automatización de la decisión Formulación de estrategias Planificación en base a escenarios o la mejor opción Sistemas de recomendación	Métodos de simulación: simulación de eventos discretos, simulación de Monte Carlo, modelos estocásticos con incertidumbre Modelos de optimización Motor de reglas (rules engine) y Procesador complejo de eventos (complex event processor) Programación Lineal y no lineal Otras técnicas

* CAT=Categoría, ** TA=Tipo de Análisis

Fuente: Elaboración propia

Previsión / prevención (de la visión a la actuación). Además de apoyar a la decisión, se pretende tomar acciones o recomendar estrategias para resolver problemas puntuales; dependiendo del caso, se busca automatizar la decisión; por ejemplo, en un sistema de telemetría aplicado a la agricultura que utiliza tecnologías de Internet de las Cosas, puede activar el riego en el momento oportuno, basándose en datos de sensores de humedad del suelo. Otros ejemplos son los sistemas de recomendación de fertilizantes para nutrición de plantas, según los requerimientos nutricionales de un cultivo, determinar tipo y dosis de fertilizantes a aplicar y el presupuesto. Otros casos, son los sistemas de predicción/prevención de riesgos en la producción (alertas de plagas, inundaciones, incendios, etc.), sistemas que ayudan a generar estrategias de optimización en el uso de recursos, etc.

Disciplinas que se relacionan con la ciencia de datos

La ciencia de datos es una disciplina relativamente joven e interdisciplinaria; guarda relación con otras disciplinas (ver Imagen 1.3) como la Estadística, Matemática y las Ciencias de la Computación. En el contexto de las Ciencias de la computación, los campos específicos interrelacionados son: Inteligencia de Negocios, Minería de datos, Aprendizaje Automático, Inteligencia Artificial, Redes Neuronales, Bases de datos, Datos a Gran Escala (Big data), entre otros (Costa & Santos, 2017; Kamilaris, Kartakoullis, & Prenafeta-boldú, 2017; Leading Edge, 2015).

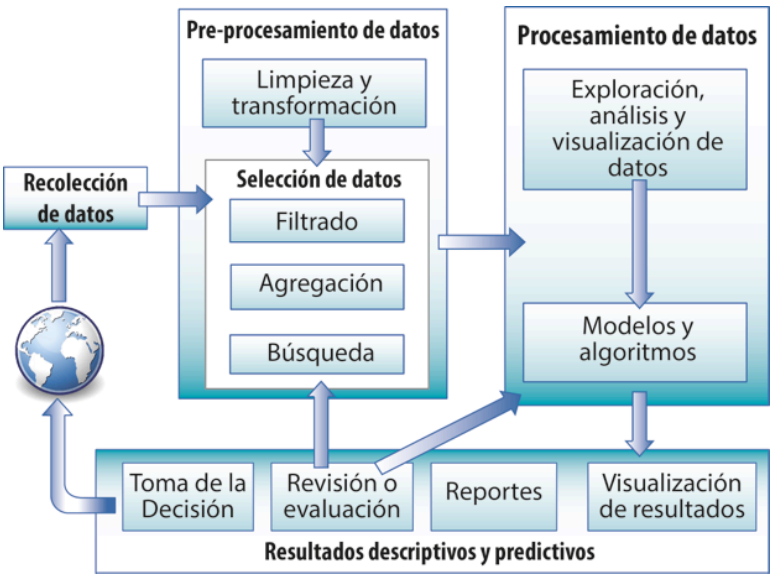
Imagen 1.3: Disciplinas que se relacionan con Ciencia de Datos

Fuente: Elaboración propia

El ciclo de vida de la ciencia de datos

El ciclo de vida de la ciencia de los datos comprende un conjunto de procesos que transforman los datos crudos en conocimiento accionable (NIST, 2015).

Imagen 1.4: El ciclo de vida de la Ciencia de Datos



Fuente: (Larson & Chang, 2016; Provost & Fawcett, 2013),

En la Imagen 1.2, se observan las etapas del ciclo de vida de Data Science:

- Recolección de datos crudos del mundo real en diversidad de formatos.
- Pre-procesamiento de datos. Comprende tareas de edición, limpieza y transformación (data munging), ajuste de datos para detectar omisiones, verificación de legibilidad y consistencia para su codificación y almacenamiento, con el propósito de garantizar la calidad de datos (data quality). Seguidamente, ciertos datos son seleccionados aplicando filtros, agregaciones y búsqueda parcial con el fin de realizar tareas de procesamiento.
- Procesamiento de datos. Las tareas que se realizan son: exploración, análisis y visualización de los datos, creación de modelos y algoritmos confiables para obtener resultados.
- Obtención de Resultados. Éstos pueden ser descriptivos o predictivos y deben ser debidamente evaluados

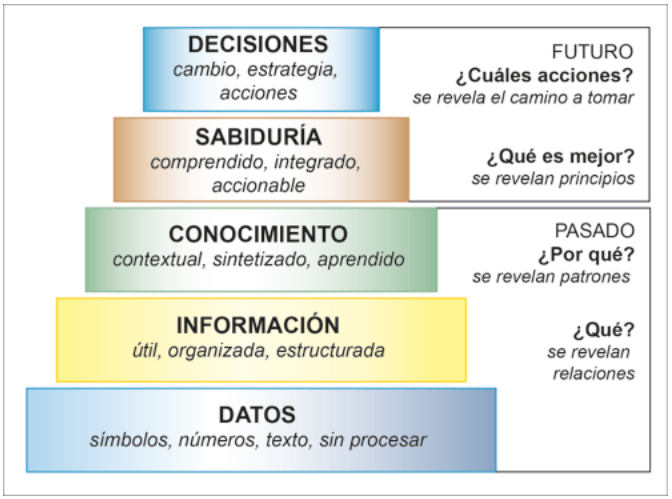
y/o revisados para generar conclusiones, nuevas teorías o tomar decisiones que implican cambios con afectación al mundo real.

Jerarquía del conocimiento en ciencia de datos

Hay varios enfoques sobre el proceso de transformación de los datos en información y conocimiento. García-Marco (2011) explica en su trabajo, la pirámide de la información como un constructo clave de la ciencia de la información; hace una crítica a otros modelos similares y presenta un modelo ampliado desde sus hallazgos y perspectiva. La jerarquía DICS, explicada en (García-Marco, 2011; Jifa & Lingling, 2014), consiste en la cadena de conversión: Dato –> Información –> Conocimiento –> Sabiduría. El cómo los datos se convierten en decisiones, tiene que ver con los procesos explicados en la sección del Ciclo de Vida de la Ciencia de Datos. Los datos que representan hechos reales no relacionados, son procesados para encontrar relaciones y generar información; el análisis de la información permite descubrir patrones muy útiles (conocimiento), que al ser presentados de manera oportuna a la persona adecuada y con la experiencia necesaria, pueden aprovechar para actuar con sabiduría y tomar las decisiones más acertada en beneficio de su organización. En la Imagen 1.5 se aprecia la pirámide del conocimiento, que va desde los datos hasta la toma de decisiones[9] .

[9] http://information4dummies.blogspot.com/2014/04/modelo-yo-conceptos-de-representacion.html

Imagen 1.5: La pirámide del conocimiento



Fuente: (Larson & Chang, 2016; Provost & Fawcett, 2013)

A continuación se explica la diferencia entre datos, información y conocimiento:

Datos

Los datos son el recurso más abundante del planeta; se calcula que a nivel mundial, se generan a diario 2.5 Exabytes de datos (BBC News, 2014; Khoso, 2016). En todo el mundo, las empresas recopilan datos de sus transacciones diarias; los gobiernos recopilan regularmente datos de censos e informes de incidencias en los departamentos de policía; a diario en las redes sociales, millones de personas suben fotos, videos y envían mensajes de texto. Este diluvio de datos crece rápidamente y cada vez más, con las nuevas tecnologías como el Internet de las Cosas (IoT), las redes de sensores inalámbricas (WSN) y los objetos inteligentes (Smart Object). En el 2013, la cantidad total de datos en el mundo, fue de 4,4 Zetta bytes (Zb), y se estima unos 44 Zb para el 2020 y 163 Zb para el 2025 (BBC News, 2014; Khoso, 2016).

Los datos que se obtienen de diversas fuentes internas o externas de una organización y aquellos que aún no han sido procesados se los denomina datos crudos (raw data) o

datos en bruto. También, son considerados como la materia prima que se procesa para obtener información y conocimiento que es útil para la toma de daciones. Los datos representan características o atributos de un objeto o hecho de la vida real. A continuación se citan algunos aportes de autores respecto a la definición de dato:

- Naur P. (1974). "Los datos son una representación formal de hechos o ideas que pueden ser comunicados o manipulados por algún proceso".
- Según Flores (1981). "Los datos se describen además como una representación del mundo real".
- "El valor de los datos reside en su uso" National Science Foundation (NSF): Bits of power (1997).
- Collis (2009), "los datos se definen como una representación de hechos que pueden ser recogidos, registrados y utilizados como base para la toma de decisiones".
- Hernández (2014), define al dato como la "unidad mínima semántica que por sí sola no representa información".

Clases de datos. Los datos según su formato se clasifican en:

- Estructurados. Tiene un estructura bien definida, por ejemplo una base de datos relacional o almacén de datos (data warehouse).
- Semi-estructurados. Conjuntos de datos con un formato diferente a las bases de datos relacionales. Por ejemplo archivos de texto plano (CSV, JSON, XML, HTML, etc.), archivos de hojas de cálculo (XLS, XLSX, ODS, otras), reportes de sistemas informáticos, etc.
- No estructurados. Datos sin una estructura estándar, por ejemplo: imágenes, audios, videos, documentos de textos, etc.

Datos según el origen:

- Internos. Datos generados en la propia organización; por ejemplo datos de compras, producción, ventas, etc.
- Externos. Datos que se obtienen de sistemas externos.

Por ejemplo de sistemas gubernamentales (seguro social, servicios de recaudación de impuestos), redes sociales, datos en tiempo real provenientes de dispositivos inteligentes (Smart objects) utilizados en entornos de Internet de las Cosas (IoT), entre otros.

Datos según su valor:

- Cuantitativos (numéricos). Son de dos tipos: 1) discreto, corresponde a valores enteros, por ejemplo: número de plantas por hectárea en un cultivo de cacao, número de cajas de banano cosechadas por hectárea, etc.) y, 2) continuo, corresponde a valores numéricos con decimales. Por ejemplo: temperatura, humedad del suelo, etc.
- Cualitativos. O también conocidos como alfanuméricos, representan un texto con varios caracteres o un sólo carácter. Pueden combinarse letras, símbolos incluso números. Se clasifican en 1) ordinales, por ejemplo: fases fenológicas de un cultivo de banano y, 2) nominales por ejemplo: nombre del cultivo, nombre de un fertilizante, etc.
- Lógicos, representan un valor lógico que puede ser "verdadero" o "falso", "SI" o "NO". Por ejemplo ¿la solicitud ha sido aprobada? La respuesta puede ser Si o No

Información

Este término se refiere a un conjunto de hechos que tienen un significado, un propósito y un formato adecuado para la toma de decisiones. La información[10] es el resultado del procesamiento y análisis de los datos. Los datos depurados, transformados, organizados, relacionados o clasificados se convierten en información. El concepto de información responde a preguntas como: ¿qué?, ¿quién?, ¿dónde? y ¿cuándo? La información es capaz de cambiar opiniones, pensamientos y criterios de acuerdo a la forma de ser percibida por el receptor (Guillén, López Ayuso, Paniagua, & Cadenas, 2015).

[9] http://infomedicsa.blogspot.com/2016/04/dikw-datos-informacion-conocimiento.html
http://mizrablogs.blogspot.com/2017/02/capitulo-ii-datos-en-medicina.html

Conocimiento

Es la realidad de un objeto, captada y entendida por un sujeto. La información se transforma en conocimiento al ser interpretada de forma reflexiva y con base en la experiencia. El conocimiento responde a las preguntas ¿cómo?, ¿por qué?

Según Guillén et al. (2015), el conocimiento:

> *"...surge de ideas verificadas y validadas por convención"; es decir, "...es el resultado de procesar información y hallar ciertos patrones invariantes que generan un cuerpo coherente de juicios acerca del mundo".*

Tipos de conocimiento

El conocimiento según el origen puede ser:

- Tácito. Cuando el conocimiento se basa en la experiencia personal, éste resulta sencillo de aplicarlo pero difícil explicarlo debido a que permanece en nuestro inconsciente de forma desarticulada.
- Explícito. Cuando el conocimiento es fácilmente explicado de manera escrita u oral; se encuentra estructurado y es fácil explicar y compartir con los demás. Ejemplo: libros, artículos científicos, etc.

Roiger (2017) presenta la clasificación del conocimiento según la forma de extracción:

- Evidente: Es aquella información que se puede obtener de forma sencilla. Por ejemplo a través de una consulta en una base de datos a través de un sistema informático, se puede verificar las facturas emitidas por la venta de un producto en una fecha determinada.
- Multi-dimensional: Es la información capaz de ser representada mediante varias perspectivas o vistas, brindando cierta estructura a los datos analizados. La técnica utilizada es el análisis mediante cubos OLAP. Por ejemplo: *Total de ventas de cajas de banano por, año, semestre y mes. Dónde: total ventas es la medida de la información, y las vistas son: producto=banano, año, semestre y mes*

- Oculto: Es aquella información que no se encuentra visible de manera superficial, es desconocida, por lo tanto, no evidente, pero útil si se aplican técnicas correctas para su descubrimiento. Entre las principales técnicas se encuentra las de Minería de Datos (Data Mining) y Estadística. Por ejemplo: En base a la información de ventas de camarón durante los últimos cinco años, determinar la proyección de ventas para el próximo año.
- Profundo: Es la información que se encuentra inmersa en los datos, pero con un grado de seguridad potencialmente alto, donde se debe poseer una especie de clave para entenderla e interpretarla. Se aplican técnicas como Aprendizaje de Máquina (Machine learning), Aprendizaje profundo (Deep Learning) y otras técnicas de inteligencia artificial.

Sabiduría (Wisdom).

Consiste en la capacidad de intuir, comprender e interpretar el conocimiento para planificar o tomar la mejor decisión posible a corto, mediano o largo plazo. Es la capacidad de emplear el juicio basado en principios para saber ¿cuál es la mejor decisión a tomar?

El perfil del científico de datos (data scientist)

En la actualidad, el universo de los datos es abundante, crece cada vez y a pasos agigantados, trayendo consigo nuevas profesiones como el de científico de datos (data scientist). Al examinar la literatura científica, algunos autores han contribuido en la comprensión del perfil profesional del científico de datos. Davenport & Patil (2012) presentan el trabajo de científico de datos como el más atractivo del siglo XXI; además, describen el perfil y cómo las organizaciones pueden explotar su potencial. Dhar (2013) también describe el campo y las habilidades que debe tener un científico de datos. Provost y Fawcett (2013) propusieron relacionar el campo de la Ciencia de Datos con otros temas como Big Data y la toma de decisiones basada en datos e identificaron los principios fundamentales presentes en la Ciencia

de Datos. El Instituto Nacional de Estándares y Tecnología (NIST, 2015), define a un científico de datos como un profesional con suficiente conocimiento de las necesidades del negocio (empresa o institución), conocimiento del dominio del problema o del contexto de los datos, habilidades analíticas (estadística, matemática), manejo de herramientas de software e ingeniería de sistemas para administrar los procesos de datos según el ciclo de vida de la ciencia de los datos. Según Costa & Santos (2017), los conocimientos que debe saber el científico de datos son:

- Gestión de los datos. Características, esquemas y estructura de los datos, modelado de datos, administración de sistemas gestores de datos, manejo de grandes volúmenes de datos.
- Seguridad, privacidad y ética en el manejo de los datos.
- Teorías, métodos y herramientas computacionales. Programación de algoritmos, gestores de almacenamiento de datos, almacenes de datos (data warehouse), inteligencia artificial, aprendizaje de máquina (machine learning), aprendizaje profundo (Deep Learning).
- Capacidades personales y sociales. Perspicacia para los negocios, comunicación, emprendimiento, creatividad, trabajo interdisciplinario.
- Diseño de sistemas computacionales. Sistemas distribuidos, redes de datos, procesamiento paralelo/distribuido, manejo de servidores y computación en la nube, escalabilidad en sistemas distribuidos.
- Procesamiento de flujos de datos. Captura y extracción de datos, limpieza y transformación de datos, procesamiento de datos, minería de datos, visualización de datos y comunicación de resultados, análisis de datos (análisis descriptivo / exploratorio de datos, análisis de métodos predictivos /prescriptivos, análisis automatizado, etc.)
- Investigación de tópicos relacionados al campo de estudio. Estadística, matemática, aplicación del método científico, ciencias computacionales, dominio del problema relacionado con los datos, etc.

En el Cuadro 1.2, se aprecian algunas de las competencias del científico de datos, algunas de ellas definidas en los trabajos de: Davenport & Patil (2012), Dhar (2013), NIST (2015) y Costa & Santos (2017) .

Cuadro 1.2: Competencias del científico de datos

Competencias Generales	Competencias específicas
Asegura flujos eficientes de datos	Realiza captura eficiente de los datos
	Limpia y transforma datos
	Asegura la calidad de los datos
	Integra datos de diversidad de fuentes y formatos
	Resuelve problemas basándose en el análisis de datos
	Extrae valor de los datos
	Respeta la privacidad, seguridad y ética de los datos
Identifica patrones y tendencias en los datos	Explora y visualiza datos o grandes volúmenes de datos
	Analiza datos aplicando técnicas y herramientas adecuadas
	Define y prueba hipótesis
	Realiza experimentos con los datos y responde a preguntas específicas
	Obtiene y valida conclusiones
	Encuentra patrones y realiza predicciones basado en conocimiento oculto en los datos
Diseña, construye, implementa y optimiza artefactos de datos	Crea y programa algoritmos para el tratamiento de datos
	Diseña e implementa plataformas hardware & software para gestión y análisis de datos
	Implementa herramientas de inteligencia de negocios (análisis, visualización de datos, minería de datos, etc.)
	Crea modelos estadísticos / matemáticos de análisis de datos
	Crea soluciones efectivas, escalables y robustas basadas en datos
Trabaja con diferentes datos	Maneja grandes volúmenes de datos
	Maneja datos de distintas estructuras y formatos

Competencias Generales	Competencias específicas
Comunica y disemina aportes a la ciencia de datos	Comunica las mejores prácticas de gestión y análisis de datos
	Realiza contribuciones con otros grupos relacionados con proyectos de datos
	Comunica, por escrito o verbalmente los resultados y hallazgos de sus de investigaciones relacionadas con datos
Contribuye en la gestión y mejora del rendimiento de un negocio	Comprende los requerimientos del usuario
	Identifica amenazas de la competencia basándose en los datos
	Define reglas y medidas para monitorear productos
	Sugiere nuevas estrategias y acciones a tomar para cambiar la dirección del negocio
	Contribuye a la mejora del rendimiento del negocio
	Informa y brinda soporte al personal de mandos medios y estratégicos del negocio
	Identifica tendencias y oportunidades
	Encuentra respuestas a preguntas importantes del negocio
	Presenta estrategias y recomienda acciones de mejora del rendimiento de la producción
	Advierte de posibles riesgos del negocio a la alta gerencia y aconseja las acciones a tomar
	Asiste en el replanteamiento de nuevos retos del negocio.

Herramientas para el científico de datos

Existen una variedad de herramientas tanto para la gestión de datos como para la visualización y analítica de datos. En el Cuadro 1.3 se muestra un listado de las herramientas que más presencia tienen en el mercado de la ciencia de datos.

Cuadro 1.3: Herramientas para el científicos de datos

Tipo de Herramienta	Herramienta
Gestión de datos[11]	- Gestores de bases de datos relacionales: PostgreSQL, Oracle DB, Mysql, Microsoft SqlServer, Teradata, Netezza, etc. - Gestores de bases de datos NoSQL: MongoDB, Couchbase, Amazon DynamoDB, Casandra, etc. - Big data: Hadoop, Mahout, Hive, Spark, Kafka, Apache Cloudera[12], Apache Hortonwork[13] , etc. - Motores de búsqueda: Elasticsearch, Microsoft Azure Search, Splunk, Google Search Appliance, etc.
Análisis y visualización de datos	- Extracción Transformación y Carga(ETL) e integración de datos: Pentaho Data Integration[14], Talend[15], Informatica[16], Denodo[17], etc. - Inteligencia de Negocios: Microsoft PowerBI[18], Pentaho[19], QlikView[20], Tableau[21], Microstrategy[22] , Oracle BI[23], IBM Cognos[24], Eclipse BIRT, JasperReport, etc. - Estadística y Minería de datos: R, Python, SPSS, Minitab, Orange, RapidMiner, Weka, Knime, etc. - Aprendizaje de Máquina (machine learning)[25]: MatLab, Mathematica, SPSS, TensorFlow, Microsoft Azure Machine Learning, Apache Mahout, OpenCV, KNIME, R, Python, etc.

11 https://db-engines.com/en/ranking/relational+dbms
12 https://www.cloudera.com/
13 https://es.hortonworks.com/
14 https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html
15 https://es.talend.com/
16 https://www.informatica.com/products/data-integration.html
17 https://www.denodo.com/en
18 https://powerbi.microsoft.com/es/
19 https://www.hitachivantara.com/go/pentaho.html
20 https://www.qlik.com/us/
21 https://www.tableau.com/
22 https://www.microstrategy.com/es
23 https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html
24 https://www.ibm.com/products/cognos-analytics
25 https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Campos de aplicación de la ciencia de datos

La ciencia de datos es un campo interdisciplinario y transversal, que puede aplicarse en todo ámbito, donde se requiere analizar datos como estrategia para descubrir el conocimiento oculto y aprovechable para tomar decisiones. La ciencia de datos está cambiando la manera de cómo nos ocupamos de: negocios, producción, marketing, publicidad, mejora de la calidad de los servicios, investigación científica, etc.

Aplicación de la ciencia de datos en el sector agropecuario

En el caso del sector agropecuario, aunque un trabajador experimentado es capaz de supervisar personalmente los procesos productivos de animales y cultivos agrícolas, hay tareas que requieren de mucho esfuerzo si se realizaran manualmente; por ejemplo: detectar la presencia de enfermedades, estimar la producción en base a datos históricos, optimizar el uso de recursos, mejorar la calidad de la producción, pronosticar posibles riesgos de desastres naturales como sequías, inundaciones, heladas, etc. En la actualidad gracias al uso de tecnologías y la aplicación de la ciencia de datos, es posible el monitoreo de una unidad de producción agropecuaria (FAO, 2016), las 24 horas, 7 días a la semana, 365 días al año. La Agricultura de Precisión, el Internet de las Cosas y la Ciencia de Datos aplicadas eficientemente trae beneficios sustanciales, como el ahorro de tiempo y dinero a los agricultores, y la mejora de la producción.

En este capítulo se presentó un panorama general de lo que se puede hacer con la Ciencia de Datos. Primero se explica la evolución y los fundamentos teóricos, luego se explican los tipos de análisis de datos y las técnicas de análisis según el campo específico, seguido se determinan las disciplinas que guardan relación con la ciencia de datos, el ciclo de vida de data science, la jerarquía del conocimiento, el perfil del científico de datos y sus competencias y finalmente las herramientas que pueden utilizarse así como los campos de aplicación de data science en varios sectores como el agropecuario.

Referencia bibliográfica

BBC News. (2014). Big Data: Are you ready for blast-off? Retrieved November 30, 2017, from http://www.bbc.com/news/business-26383058

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015). Big Data in Precision Agriculture : Weather Forecasting for Future Farming. *In 2015 1st International Conference on Next Generation Computing Technologies* (pp. 4–5). Dehradun, India. http://doi.org/10.1109/NGCT.2015.7375220

Capgemini. (2015). A brief history of Data Science. Retrieved October 13, 2017, from https://whatsthebigdata.com/2015/02/17/history-of-data-science-infographic/

Collis, J., & Hussey, R. (2009). Business research: *A practical guide for undergraduate and postgraduate students*. (3rd ed.). Palgrave Macmillan: Hampshire. Retrieved from https://www.researchgate.net/publication/38177413_Business_research_A_practical_guide_for_undergraduate_and_postgraduate_students

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management*, (xxxx). http://doi.org/10.1016/j.ijinfomgt.2017.07.010

Davenport, T., & Patil, D. J. (2012). Data Scientist The Sexiest Job of the 21st Century Meet the people who can coax treasure out of messy, unstructured data. Retrieved May 17, 2018, from http://billsynnotandassociates.com.au/images/stories/documents/data_scientist.pdf

Dhar, B. V. (2013). Data Science and Prediction. *Communications of the ACM, 56(12), 64–73.* http://doi.org/10.1145/2500499

FAO. (2016). *Programa mundial del censo agropecuario 2020. Volumen 1. Programa, definiciones y conceptos.* Retrieved from http://www.fao.org/3/a-i4913s.pdf

Flores, I. (1981). *Data base architecture*. New York.: Van Nostrand Reihold Company.

García-Marco, F.-J. (2011). La Pirámide de la Información Revisitada: Enriqueciendo el Modelo Desde la Ciencia Cognitiva. *El Profesional de La Informacion*, 20(1), 11–24. http://doi.org/10.3145/epi.2011.ene.02

Gartner. (2012). Analytic Ascendancy Model. Retrieved December 12, 2017, from http://www.growwithfarm.com/evolving-analytics-from-descriptive-to-prescriptive/

Guillén, M. A., López Ayuso, B., Paniagua, E., & Cadenas, J. M. (2015). Una revisión de la Cadena Datos-Información-Conocimiento desde el Pragmatismo de Peirce. *Documentación de Las Ciencias de La Información*, *38*(Dic), 153–177. http://doi.org/10.5209/rev_DCIN.2015.v38.50814

Hernández Mendo, A., Castellano, J., Camerino, O., Jonsson, G., Villaseñor, Á., Lopes, A., & Anguera, M. T. (2014). Programas informáticos de registro , control de calidad del dato , y análisis de datos. *Psicología Del Deporte*, *23*(1), 111–121.

Jifa, G., & Lingling, Z. (2014). Data, DIKW, Big Data and Data Science. *Procedia Computer Science*, *31*, 814–821. http://doi.org/10.1016/j.procs.2014.05.332

Kamilaris, A., Kartakoullis, A., & Prenafeta-boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*(January), 23–37. http://doi.org/10.1016/j.compag.2017.09.037

Khoso, M. (2016). How Much Data is Produced Every Day? Retrieved November 30, 2017, from http://www.northeastern.edu/level-blog/2016/05/13/how-much-data-produced-every-day/

Larson, D., & Chang, V. (2016). International Journal of Information Management A review and future direction of agile , business intelligence , analytics and data science. *International Journal of Infor-*

mation Management, 36(5), 700–710. http://doi.org/10.1016/j.ijinfomgt.2016.04.013

Leading Edge. (2015). Data science: the new monetization model for analytics industry. Retrieved December 4, 2017, from http://www.leadingedgeprovider.com/2016/07/data-science-the-new-monetization-model-for-analytics-industry/

Loury, J. (2014). Evolving Analytics: From Descriptive to Prescriptive. Retrieved December 11, 2017, from http://www.growwithfarm.com/evolving-analytics-from-descriptive-to-prescriptive/

Mazon-Olivo, B., Rivas, W., Pinta, M., Mosquera, A., Astudillo, L., & Gallegos, H. (2017). Dashboard para el soporte de decisiones en una empresa del sector minero. *Conference Proceedings - Universidad Técnica de Machala*, *1*, 1218–1229. Retrieved from http://investigacion.utmachala.edu.ec/proceedings/index.php/utmach/article/view/219/191

Molina-Solana, M., Ros, M., Dolores Ruiz, M., Gomez-Romero, J., & Martin-Bautista, M. J. (2017). Data science for building energy management: A review. *Renewable & Sustainable Energy Reviews*, *70*(December 2016), 598–609. http://doi.org/10.1016/j.rser.2016.11.132

National Academi of Science. (2017). Overview of Data Science Methods. In *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions* (pp. 53–79). The National Academies Press. http://doi.org/10.17226/23670

Naur, P. (1974). *Concise Survey of Computer Methods*. Lund: Studentlitteratur.

NIST. (2015). NIST Special Publication 1500-1 NIST. Big Data Interoperability Framework : Volume 1 , Definitions. *National Institute of Standards and Technology*, *1*, 32. http://doi.org/10.6028/NIST.SP.1500-1

Press, G. (2013). A Very Short History Of Data Science. Retrieved October 10, 2017, from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

Provost, F., & Fawcett, T. (2013). *Data Science for Business. What you need to know about Data Mining and Data-Analytic thinking*. O'Reilly Media.

Roiger, R. (2017). *Data Mining: A Tutorial-Based Primer (Segunda Ed)*. United States of America: CRC Press.

Sivarajah, U. et al. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. http://doi.org/10.1016/J.JBUSRES.2016.08.001

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, *153*, 69–80. http://doi.org/10.1016/j.agsy.2017.01.023

Análisis de Datos Agropecuarios
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: 978-9942-24-120-7