Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Agosto, 2022

Introducción (1)

- ¿Para qué sirve la Ciencia de Datos?
  - Para obtener información valiosa de los datos
- ¿Científico de datos?
  - Ayuda a convertir datos sin procesar en información.
  - Habilidades en: analítica, aprendizaje automático, minería de datos y estadística, experiencia en algoritmos y programación…
  - Habilidad más importante que debe poseer un científico de datos es:
    - Poseer la capacidad de explicar el significado de los datos de una manera que pueda ser fácilmente entendida por otros y a su vez que los datos cuenten historias y solucionen problemas basados en ellos.

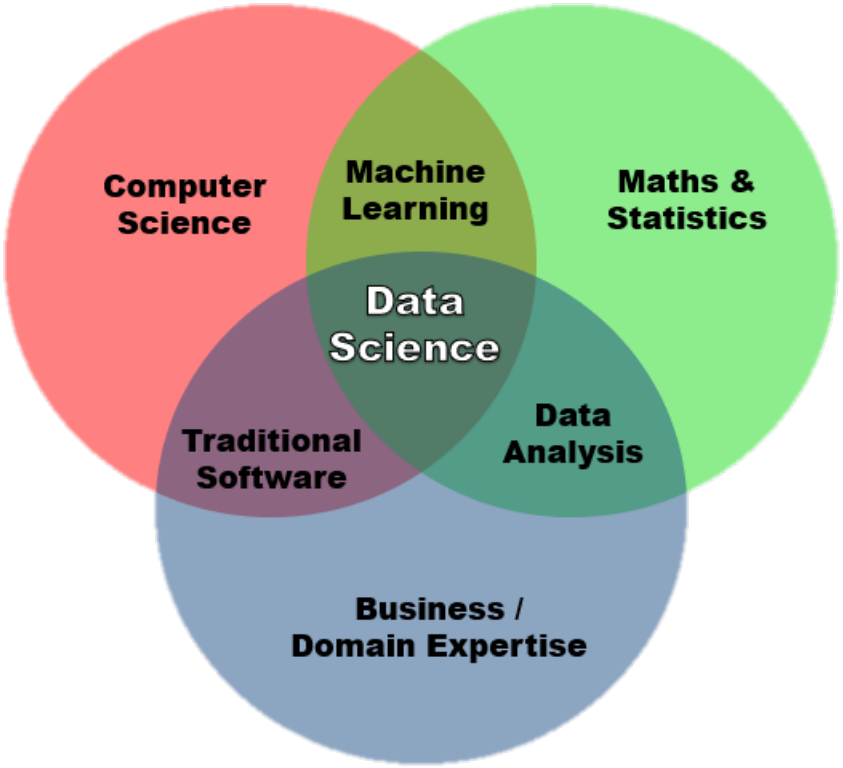
¿Qué es la Ciencia de Datos? (1)

- Es un conjunto de principios, definición de problemas, algoritmos y procesos para extraer patrones no obvios y útiles en grandes volúmenes de datos. Muchos de estos principios se han desarrollado en campos relacionadoscon Machine Learning (ML) y Minería de Datos (MD). Estos términos a menudo se usan indistintamente…
  - La CD se centra en mejorar la toma de decisiones a través del análisis de datos, un aspecto muy destacable y central de esta disciplina. Sin embargo aunque la CD se basa en otros campos, ésta tiene un alcance más amplio.
  - Por ejemplo, ML se centra en el diseño y evaluación de algoritmos para extraer patrones de datos y clasificarlos.
  - La MD en general se ocupa del análisis de datos que a menudo implica un énfasis en aplicaciones comerciales.
  - La CD toma estas consideraciones en cuenta, pero también asume otros desafíos como la captura, limpieza y transformación de datos no estructurados. Por ejemplo, desde las redes sociales, datos web, así como el uso de tecnologías de big data para almacenar grandes conjuntos de datos no estructurados.

¿Qué es la Ciencia de Datos? (2)

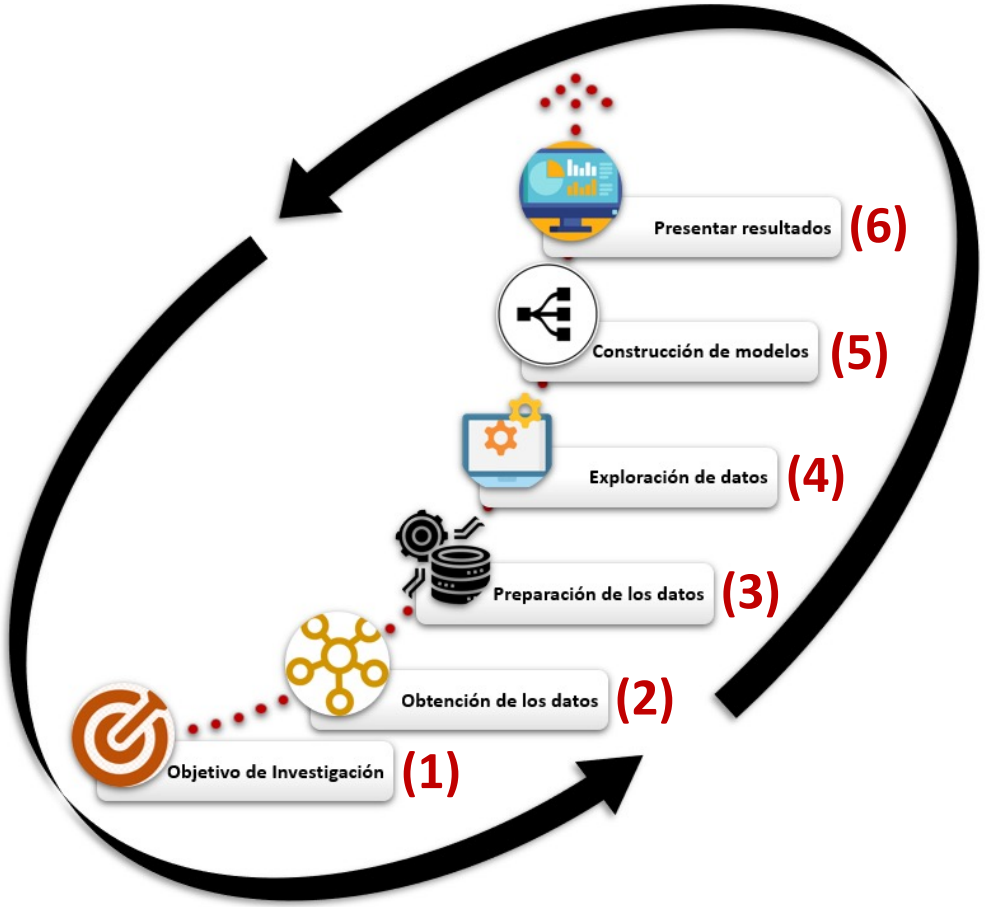
- Usando la CD, podemos extraer diferentes tipos de patrones. Por ejemplo, aquellos que nos ayuden a identificar grupos de clientes siguiendo un comportamiento y gustos similares (segmentación de clientes en marketing) y en terminología de CD a esto se le denomina (agrupación o clusters).
- Alternativamente se puede extraer un patrón que identifique productos que son comprados juntos con frecuencia, y esto es un proceso de MD, conocido como regla de asociación, o tal vez requerimos extraer patrones que identifiquen eventos extraños como un reclamo de seguros, reclamos fraudulentos, y esto es un proceso conocido como detección de anomalías y valores atípicos.
- Finalmente es posible que necesitemos identificar patrones que nos ayuden a clasificar cosas… Entonces en general, la CD es una disciplina que nos permite convertir datos sin procesar en entendimiento, comprensión y conocimiento.

¿Qué es la Ciencia de Datos? (3)



- Podemos ver el diagrama de Venn, como la unión de varias disciplinas de varias ciencias del conocimiento, tales como:
  - Ciencia de la computación, matemáticas y estadística que entre ellas estas habilidades combinadas producen el ML.
  - Las habilidades del negocio con la ciencia de la computación generan la programación o desarrollo de software, utilizado en negocios tradicionalmente.
  - La combinación de estadística y matemáticas, con conocimientos del negocio, entonces formaliza el análisis de datos.
  - Por tanto, la CD es una combinación de habilidades en cada una de estas especialidades y ponerlas a disposición de la empresa para responder interrogantes del negocio.

Proceso de la Ciencia de Datos (1)



- Todo proyecto de CD nace de una necesidad específica:
  - (1). Definición de hipótesis y descripción de qué queremos obtener como resultados.
  - (2). Cuáles son los posibles datos que puede arrojar esa respuesta o que se pueden procesar para obtener esas respuestas a esas interrogantes o comprobar o descartar esa hipótesis.
    - Datos de interés
      - Distintas fuentes de datos…
      - Datos de diversas disciplinas…

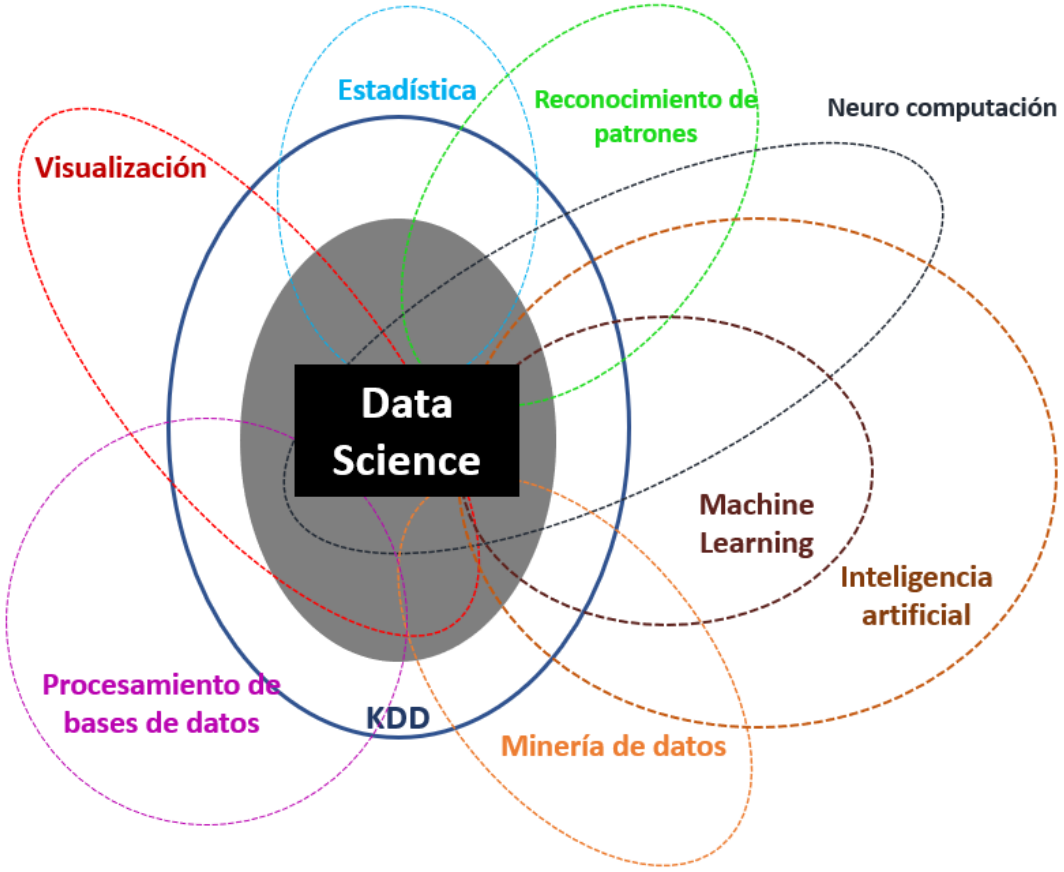
Proceso de la Ciencia de Datos (2)

- (3). Hacer una transformación, una estandarización, un modelado de datos, en donde se necesite. Por ejemplo: estandarizar formatos de fechas, formatos de variables, combinar o calcular nuevas variables a partir de los datos, etc. Por tanto, es un proceso importante el estandarizar y luego explorar sobre ellos para saber ¿qué es lo que existe? ¿cómo están? ¿cuáles son sus patrones o posibles patrones que van a estar o qué vamos a identificar dentro de los datos?
  - Algo muy típico dentro de los datos, es que tenemos muchos valores que están perdidos o valores faltantes, errores de formato, de almacenamiento, entre otros. Estos detalles deben ser limpiados o transformados para contar con "datos estandarizados", para posteriormente realizar una exploración de datos…

Proceso de la Ciencia de Datos (3)

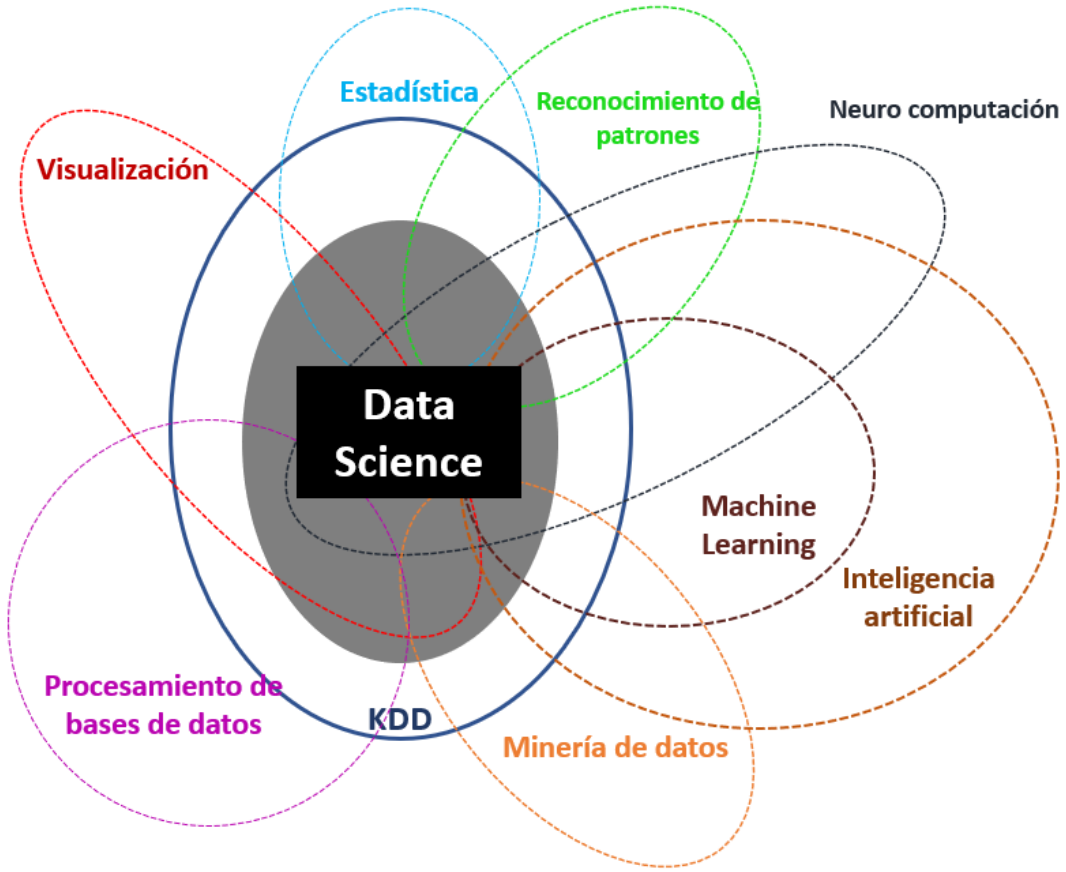
- (4). Esta tarea se da en distintas etapas como:
  - La exploración visual, realizando pruebas estadísticas dentro de esos datos, conocer la distribución, mínimos, máximos, promedios, conocer la densidad, entre otros.
- (5). Luego de ya explorados, se construye el modelo de datos que va a utilizarse para responder a diversas preguntas o definiciones planteadas en el objetivo, descartando aquellos datos que no se utilizarán para aplicar los algoritmos para desarrollar el modelo que va a procesar esos datos y los convertirá en información que se va a presentar…
- (6). Esta tarea consiste en resumir el producto final, conocer que técnicas de visualización pueden utilizarse, cómo presentar esta información con base en la audiencia a quién se dirige; ya sea gráficamente o textualmente.
  - En este sentido que sea cognitivamente entendible para las personas; es decir, elegir el modelo más adecuado que explique la información, darle valor al proceso que hubo detrás, mediante gráficas, dashboards que definan las conclusiones del objetivo de la investigación
    - Ejemplo: Our World in Data, Information is beautiful...

Especialidades en la Ciencia de Datos (1)



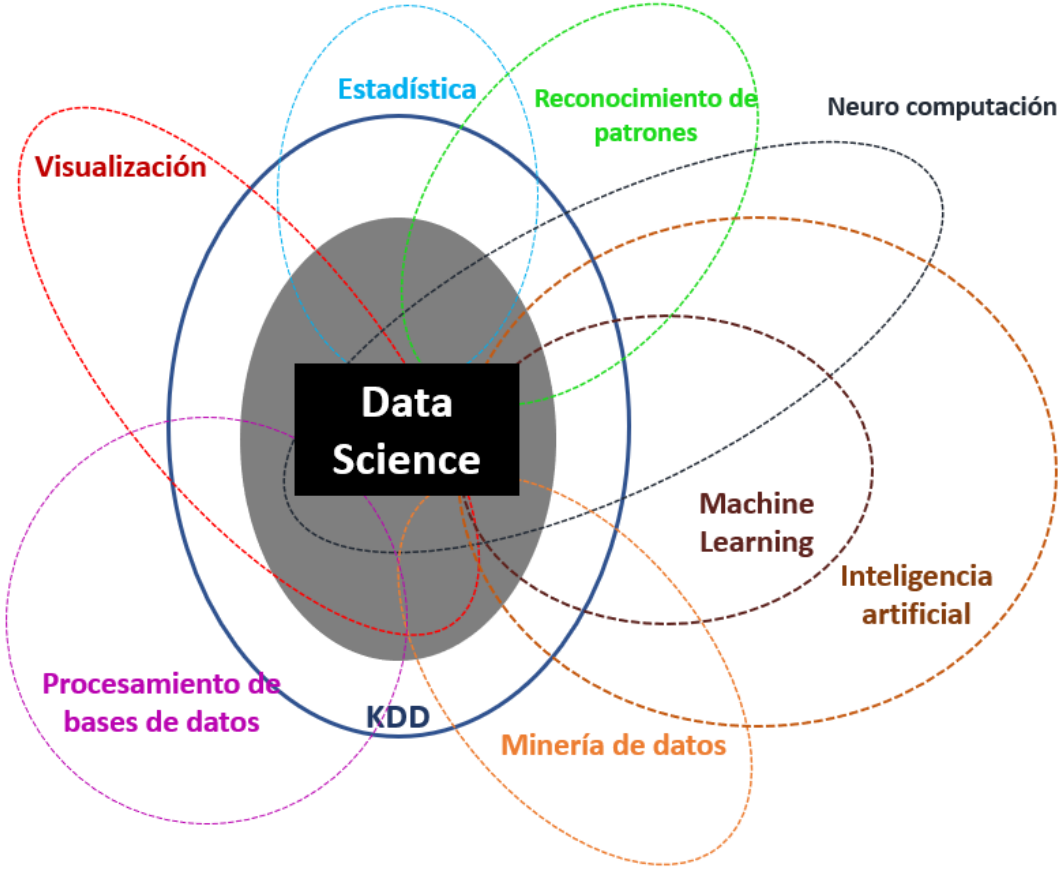
- La CD engloba e integra muchas disciplinas para aportar en Knowledge Discovery Database:
  - Estadística para obtener un orden y análisis de un conjunto de datos y para obtener explicaciones y predicciones sobre fenómenos observados. Los métodos estadísticos permiten recolectar información para luego analizarla y extraer de ellos conclusiones relevantes. La estadística es la mejor representación de la CD y su principal objetivo es mejorar la comprensión de los hechos a partir de la información disponible, tiene como característica fundamental su transversalidad, por lo cual la hace aplicable a estudios de diversas disciplinas.

Especialidades en la Ciencia de Datos (2)



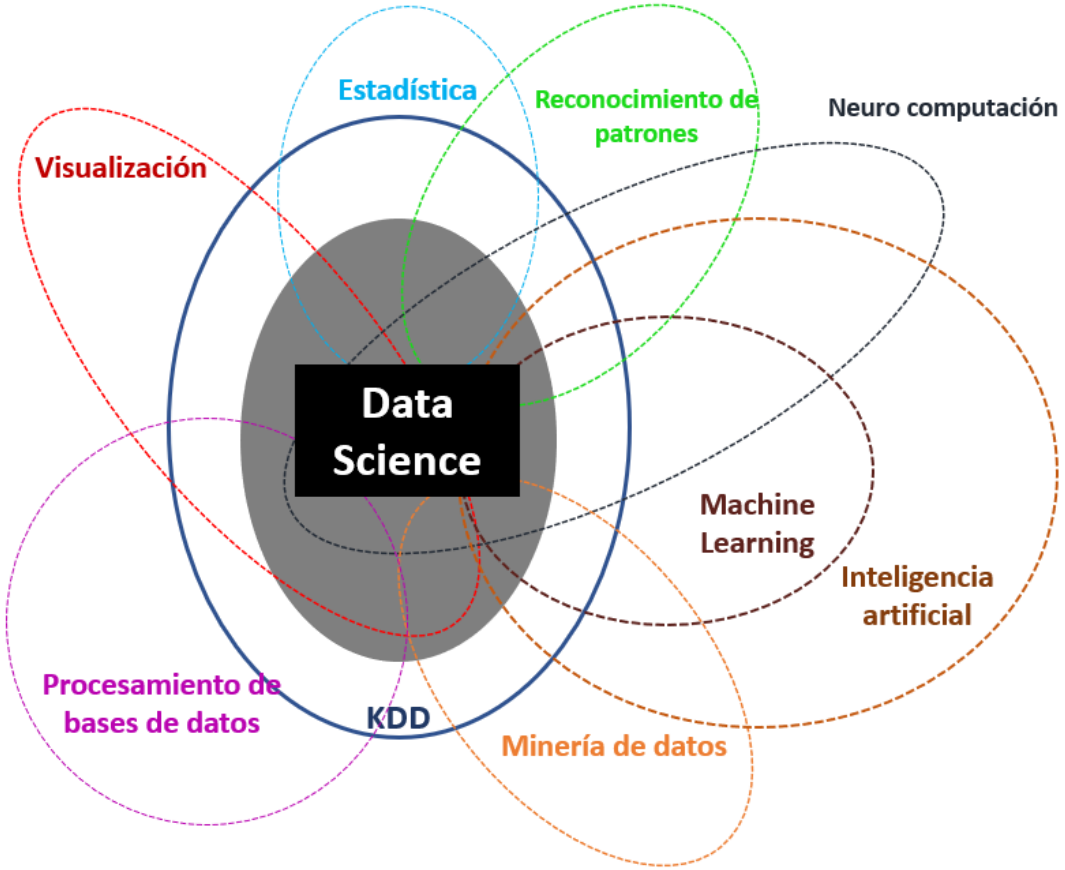
- **Reconocimiento de patrones.** Es el procesamiento de información para solucionar diversos problemas, algunos de éstos resueltos por los humanos y otros son con máquinas, a través de métodos y algoritmos. La tarea fundamental del RP es la de clasificar objetos en un número específico de categorías o clases, dependiendo del dominio de aplicación. Estos objetos se denotan con el término genérico de patrones.
- **Neurocomputación.** Rama científica interdisciplinaria que enlaza diversos campos de la biofísica, neurociencia, ciencia cognitiva, ingeniería eléctrica ciencias de la computación y las matemáticas, y que en general persiguen recrear en forma visual las redes neuronales y sus interacciones en nuestro cerebro.
- **Machine Learning.** Es una disciplina dentro del ámbito de la IA, enfocada a crear sistemas que aprenden automáticamente. ***"Aprender"*** en este contexto significa identificar patrones complejos en millones de datos. La máquina que realmente **"aprende"** es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente este contexto implica que estos sistemas se mejoran en forma autónoma con el tiempo, sin la intervención humana.

Especialidades en la Ciencia de Datos (3)



- **Inteligencia Artificial.** Es una disciplina que trata de crear sistemas capaces de aprender y razonar con un ser humano. Lo que se busca es aprender de la experiencia, investigar cómo resolver problemas ante unas condiciones dadas, contrastar información y realizar tareas lógicas.
- **Minería de Datos.** Es un enfoque que intenta descubrir patrones en grandes volúmenes de datos, utiliza los métodos de la IA, ML, estadística y SBD para extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Dentro de la MD existe una rama en crecimiento la "minería de textos" (*text mining*), la cual refiere al proceso de analizar y descubrir información nueva a partir de textos, por medio de la identificación de patrones o correlación entre los términos, para encontrar información que no está explícita dentro del texto. Además, brinda la posibilidad de procesar textos, tomando como fuentes de datos páginas web, libros digitales, emails, reseñas, lista de artículos y productos, redes sociales, entre otros.

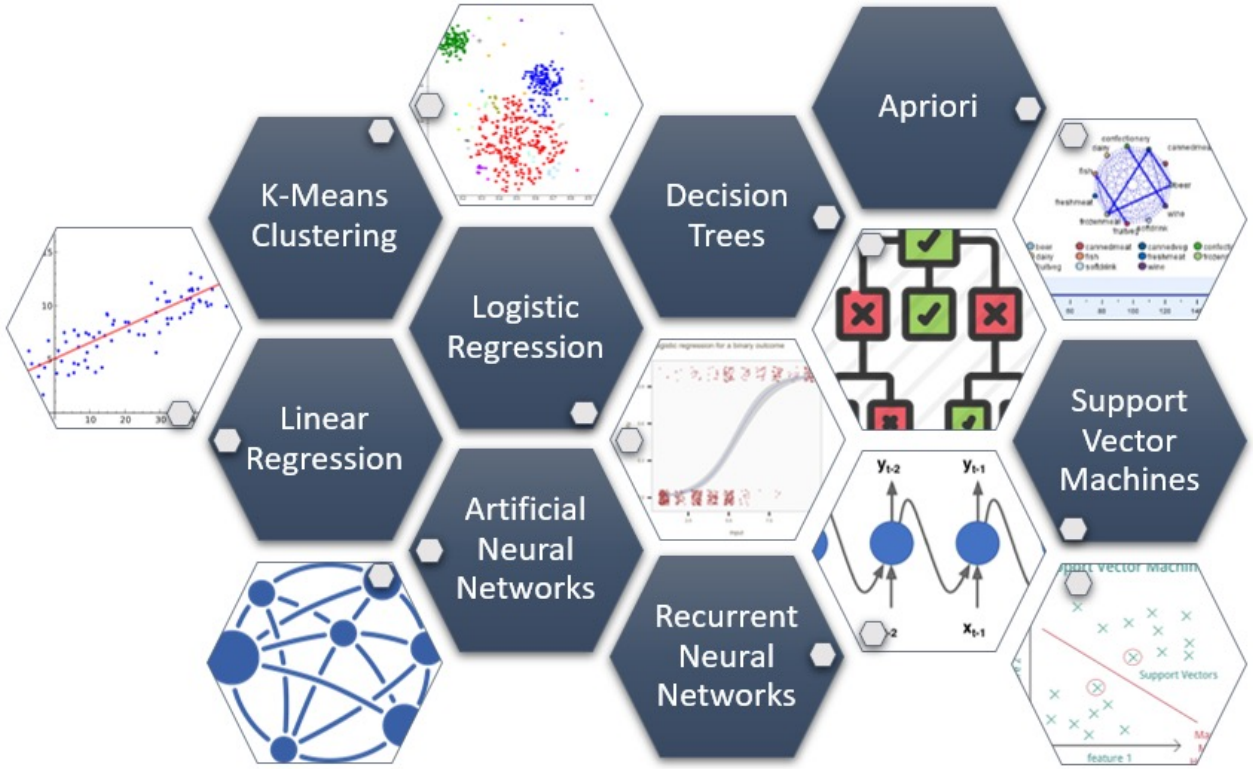
Especialidades en la Ciencia de Datos (4)



- **Procesamiento de BD.** Trata sobre el almacenamiento, manipulación de elementos de datos para producir información significativa. Todos ellos apuntan sobre el tratamiento de datos y la CD de alguna forma maneja las estructuras de datos que se almacenan en una BD.
- **Visualización.** Un simple gráfico brinda más información a la mente del analista de datos que cualquier otro dispositivo. Por tanto, está comprobado científicamente que la información presentada en forma gráfica es más comprensible para el cerebro humano y que también un gráfico es la mejor representación de mucha información contenida en un gran conjunto de datos. La visualización es la representación gráfica de información y datos al usar elementos visuales tales como: cuadros, gráficos, mapas, etc. La visualización de datos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. En el mundo del Big Data las herramientas y tecnologías de visualización de datos son esenciales para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

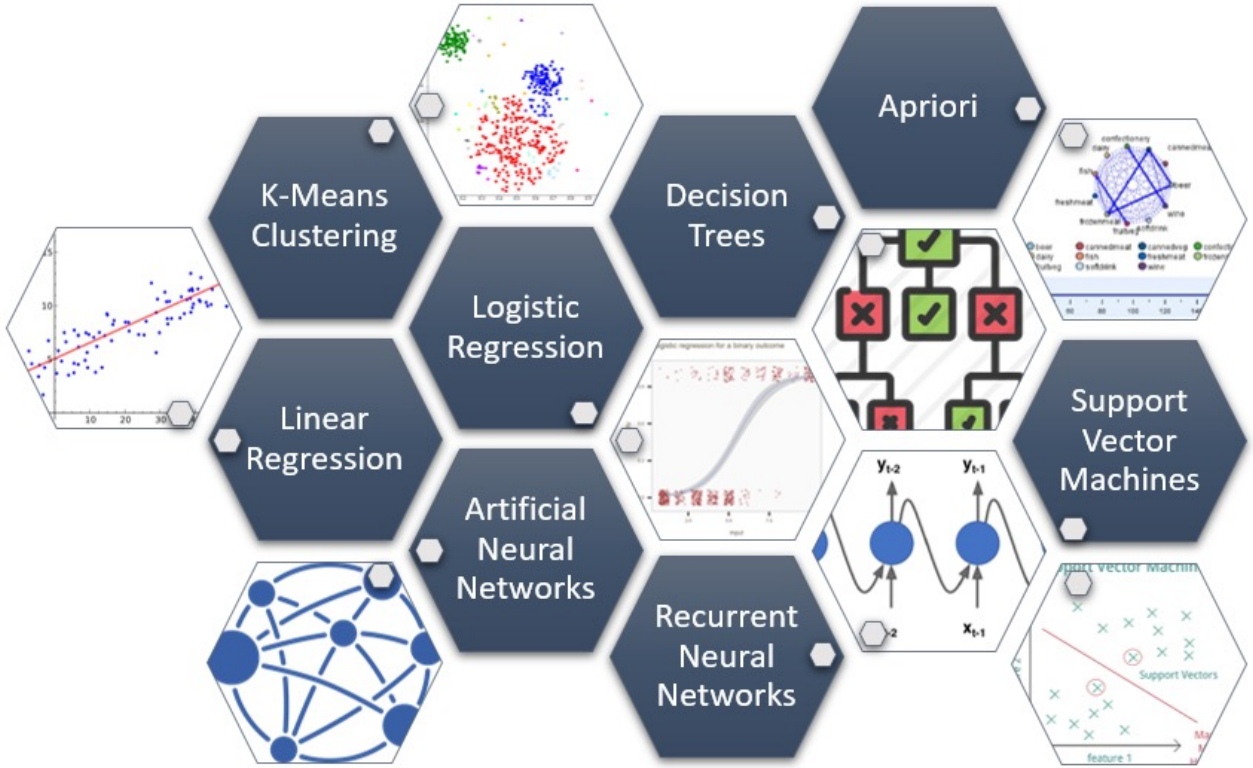
Es importante resaltar que la integración y la relación entre todas estas disciplinas dentro de la CD, buscan la generación de conocimiento, y esa generación es para ayudar y facilitar la toma de decisiones en todos los ámbitos en los cuales se puede desarrollar cada una de estas disciplinas, en general la Ciencia de Datos.

Principales Algoritmos en la Ciencia de Datos (1)



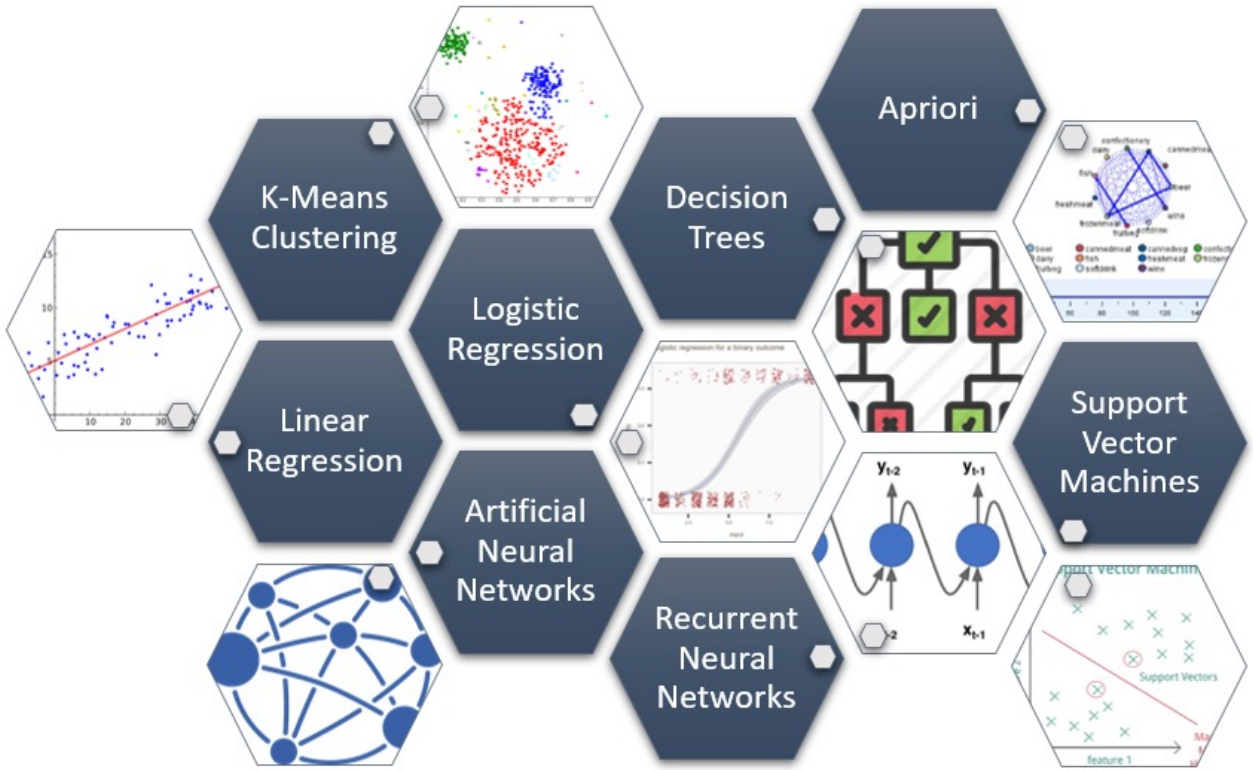
- **Modelos de Regresión.** Son modelos matemáticos que buscan determinar la relación entre una variable dependiente (*y*), con respecto a otras variables explicativas o independientes (*x*). El modelo de regresión se suele usar en ciencias sociales con el fin de determinar si existe o no relación causal entre una variable dependiente (*y*) y un conjunto de otras variables explicativas (*x*). Asimismo, el modelo busca determinar cual será el impacto sobre la variable dependiente ante un cambio de las variables explicativas o independientes. Se busca tratar de predecir por ejemplo, en variables continuas o discretas, se puede predecir por ejemplo: el retiro no muy común de un cliente, el riesgo alto o bajo de morosidad, o tal vez obtener una cantidad de seguidores o la cantidad de clientes a futuro, haciendo un modelo de regresión lineal o regresiones logísticas.

Principales Algoritmos en la Ciencia de Datos (2)



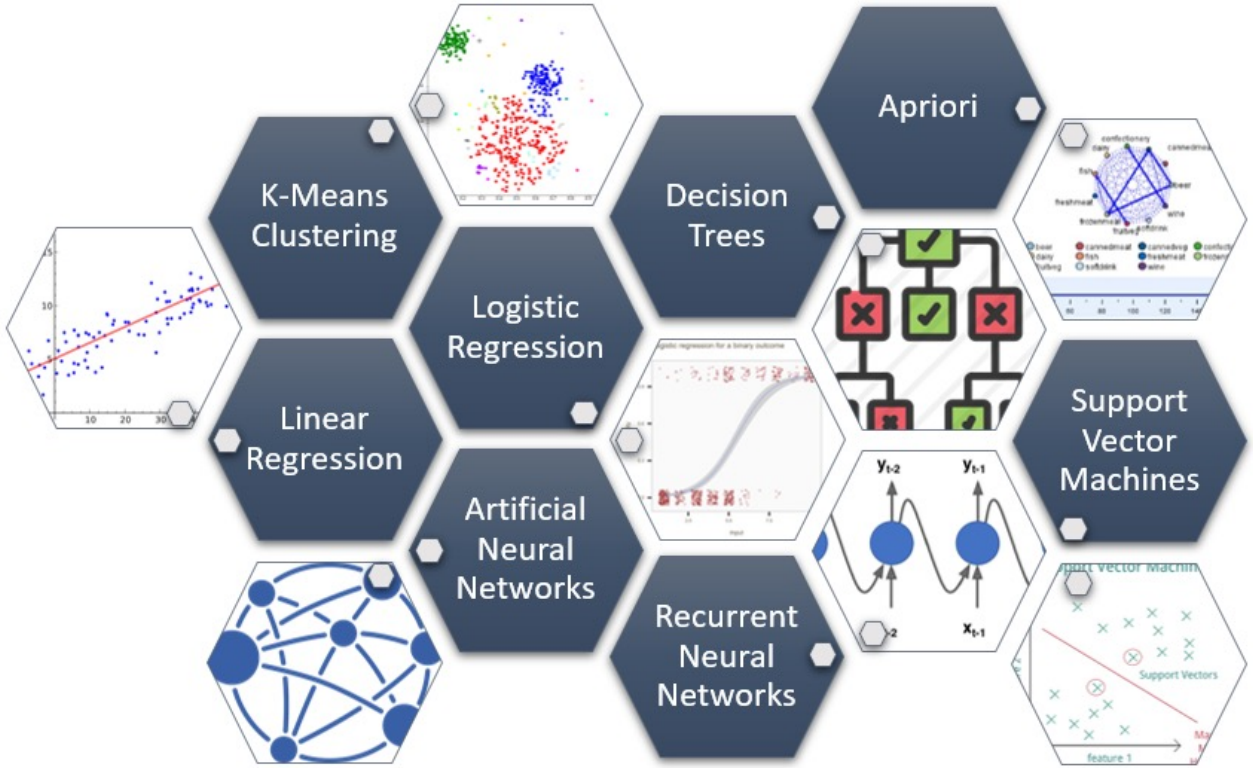
- **Redes Neuronales Artificiales.** Es un modelo simplificado que emula el modo en que el cerebro humano procesa la información. Una RNA funciona simulando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de las neuronas humanas. Entonces, estas unidades de procesamiento se organizan en capas, principalmente una capa neuronal de entrada, una o varias capas ocultas y una capa de salida con la unidad o unidades que representan el campo o campos de destino. Estas unidades se conectan con fuerza de conexión variables o ponderaciones y los datos de entrada se presentan en la primera capa y los valores se propagan desde la neurona, hasta cada neurona de la siguiente capa. Al final se envía un resultado de la capa de salida, entonces una red aprende examinando los registros de manera individual, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Entonces este proceso se repite varias veces y la red sigue mejorando su procesamiento hasta que haya alcanzado uno o varios criterios de parada. Su precisión y capacidad de procesamiento hacen que sean ampliamente usadas en problemas del mundo real, del mundo de negocios, y en ciencia de datos.

Principales Algoritmos en la Ciencia de Datos (3)



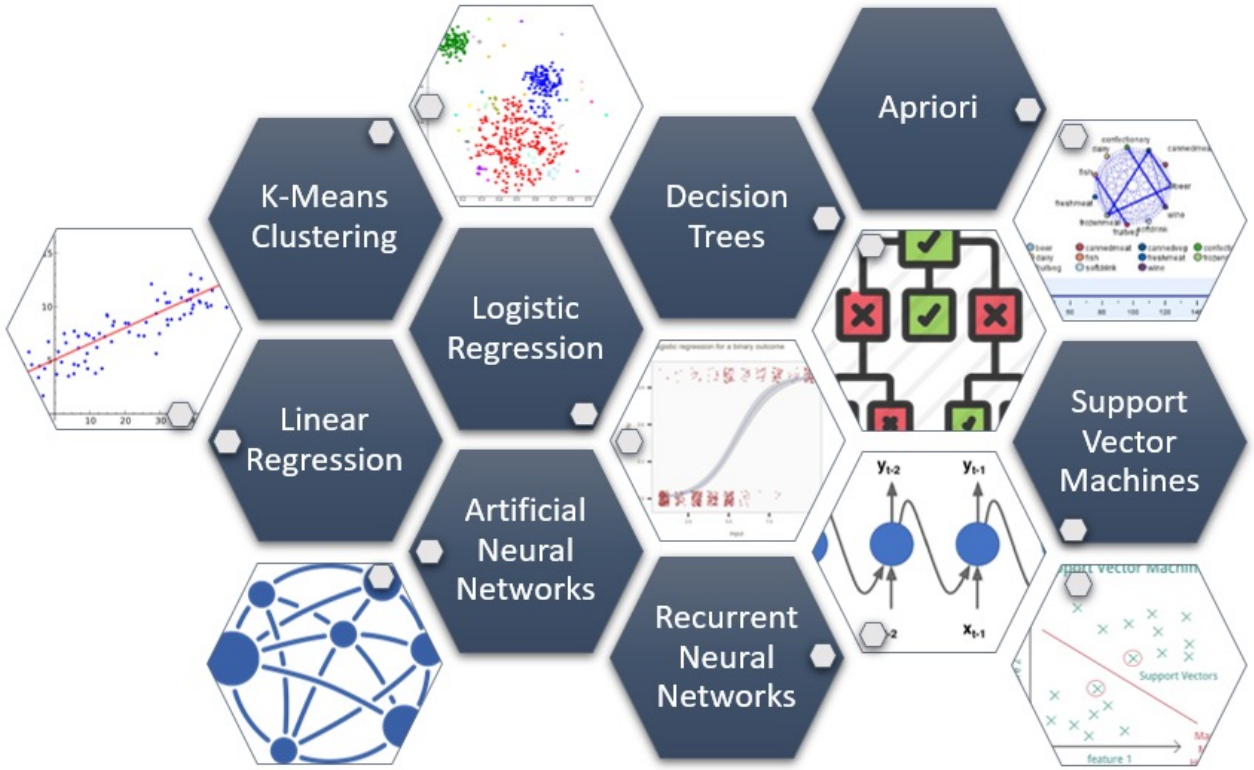
- **Árboles de Decisión.** Son métodos analíticos que a través de una representación esquemática da las alternativas disponibles para facilitar y mejorar la toma de decisiones, especialmente cuando existen riesgos, costo beneficio y múltiples opciones. En este sentido, un AD va brindando los caminos a tomar según el escenario que se va presentando. Entonces, los AD son muy vistosos porque muestran un esquema de cómo y cuál es la mejor decisión a tomar cuando se presenta un escenario u otro.
- **Clustering.** El algoritmo de *clustering* más utilizado es *K-means* y la tarea principal que tiene un algoritmo de cluster es buscar agrupar un conjunto de objetos no etiquetados para lograr construir subconjuntos de datos conocidos como clusters. Entonces lo que se busca es por ejemplo, segmentar clientes con la finalidad de que se puedan ofrecer servicios personalizados o más adaptados según a la categoría de cliente que tiene una empresa, encontrar similitudes entre diversos perfiles de usuarios y en redes sociales y clasificarlos, entre otros.

Principales Algoritmos en la Ciencia de Datos (4)



- **Métodos apriori.** Son métodos que buscan una relación entre un algoritmo de minería de datos para buscar relaciones entre elementos y encontrar reglas de asociación de esos elementos y asocian la combinación de éstos en diferentes dominios de aplicación. Por tanto, se generan una gran cantidad de reglas que permiten ver la relación que tienen por ejemplo: las ventas en un retail, ventas de un productos, entre otros. Pueden responder a preguntas como: ¿qué probabilidad tiene que se de la venta de un producto? ¿cuántas veces se ha dado en el tiempo? Estos algoritmos son muy poderosos para sistemas de recomendación, que integran diversas aplicaciones orientadas a perfiles de usuario (i.e. Netflix, Amazon, Music, etc.).

Principales Algoritmos en la Ciencia de Datos (5)



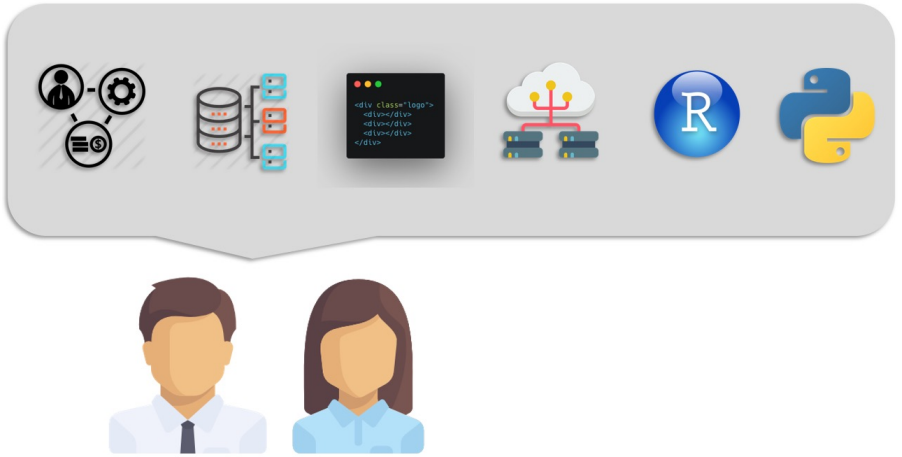
- **Máquinas de Soporte Vectorial.** Son métodos propiamente relacionados con problemas de clasificación y regresión, en donde se toma un conjunto de datos de entrada y se predice para cada una de las entradas, las dos clases de salida a las cuales pertenece, por lo que es un clasificador no probabilístico lineal binario, ya que solo escoge entre dos opciones. Entonces busca tener gran precisión y responder problemas más complejos que tal vez a través de modelos de regresión lineal o ligística no pudieran responderse. Por tanto, estos son de los principales algoritmos que se pueden usar en la ciencia de datos, en donde el mismo algoritmo puede tener distintas aplicaciones dentro de un dominio de aplicación y ayuda no solamente a predecir, sino a clasificar entre otras muchas tareas.

Principales Algoritmos en la Ciencia de Datos (6)

Análisis de Regresión

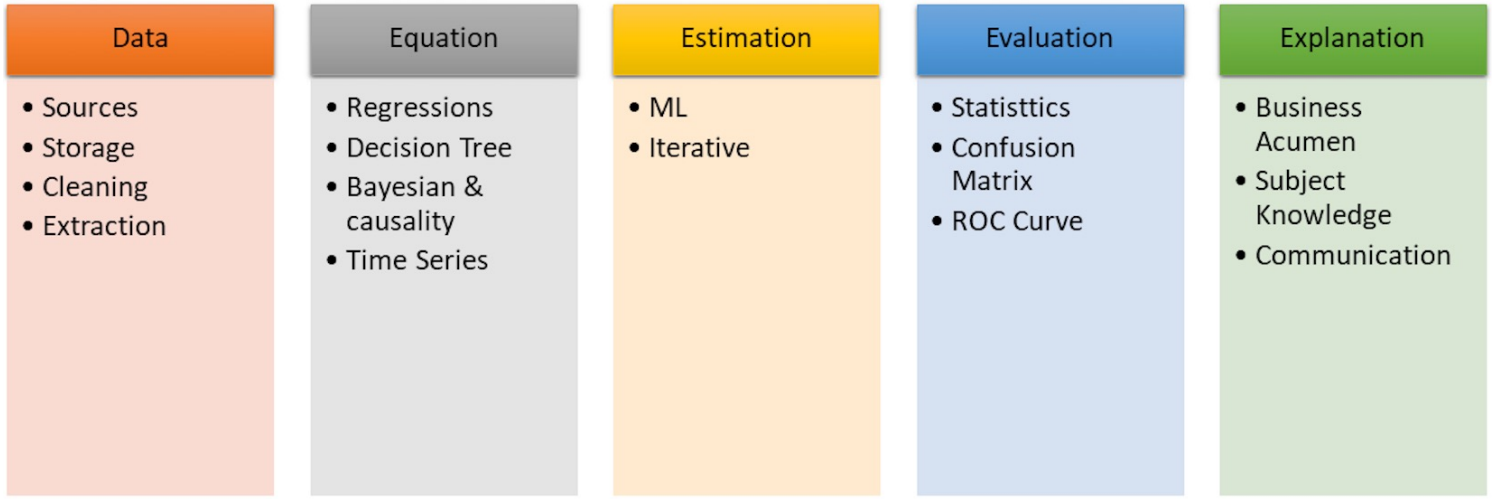
Científico de Datos (1)

- Un **científico de datos** es un especialista en estadística que pone en práctica estos conocimientos a través de la programación de software para extraer el máximo valor de los datos, ya sea desde una fuente a múltiples fuentes de datos; en donde estos datos pueden ser:
  - Estructurados,
  - semi-estructurados o
  - no estructurados.



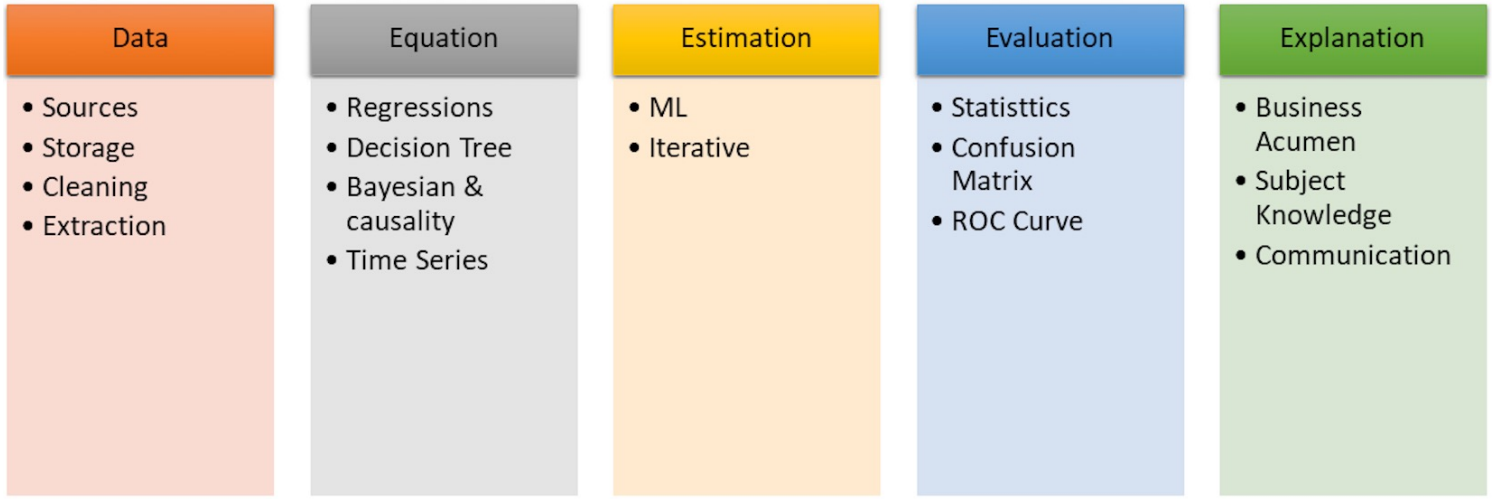
- Tareas principales de un científico de datos:
  - Analizar los datos del dominio de aplicación, identificar importancias y capturar aquellos que generan valor.
  - Limpieza de los datos, para darle una estructura analizable.
  - Evaluación de los modelos estadísticos para determinar la validez de los análisis.
  - Utilizar ML para construir mejores algoritmos predictivos.
  - Pruebas y mejora continua de la precisión de los modelos de ML.
  - Construir visualizaciones de datos para resumir la conclusión de un análisis avanzado.
  - Responder a requerimientos basados en las reglas del negocio con los datos.
  - Ayudar a mejorar la toma de decisiones.

Habilidades de un Científico de Datos (1)



- Habilidades en el manejo de:
  - Fuentes de datos y conocer la diversidad. Poder manejarse en distintas fuentes de datos estructurados, semi y no estructurados, así como en el conocimiento de los mismos datos y la limpieza de éstos.
  - Transformación de los datos para contar con una estandarización de los datos, una limpieza, sustitución de éstos y extracción de los mismos, ya sea de fuentes de datos internas o externas al dominio de aplicación, realizando estas tareas mediante APIs, conexiones web, conexiones con otros datos, etc.
  - Ecuaciones en los modelos de funciones específicas tales como regresión, árboles de decisión, probabilidad bayesiana, análisis de causalidad, series de tiempo, entre otros.
  - Entendimiento de los procesos de estimación basados en ML y procesos iterativos.

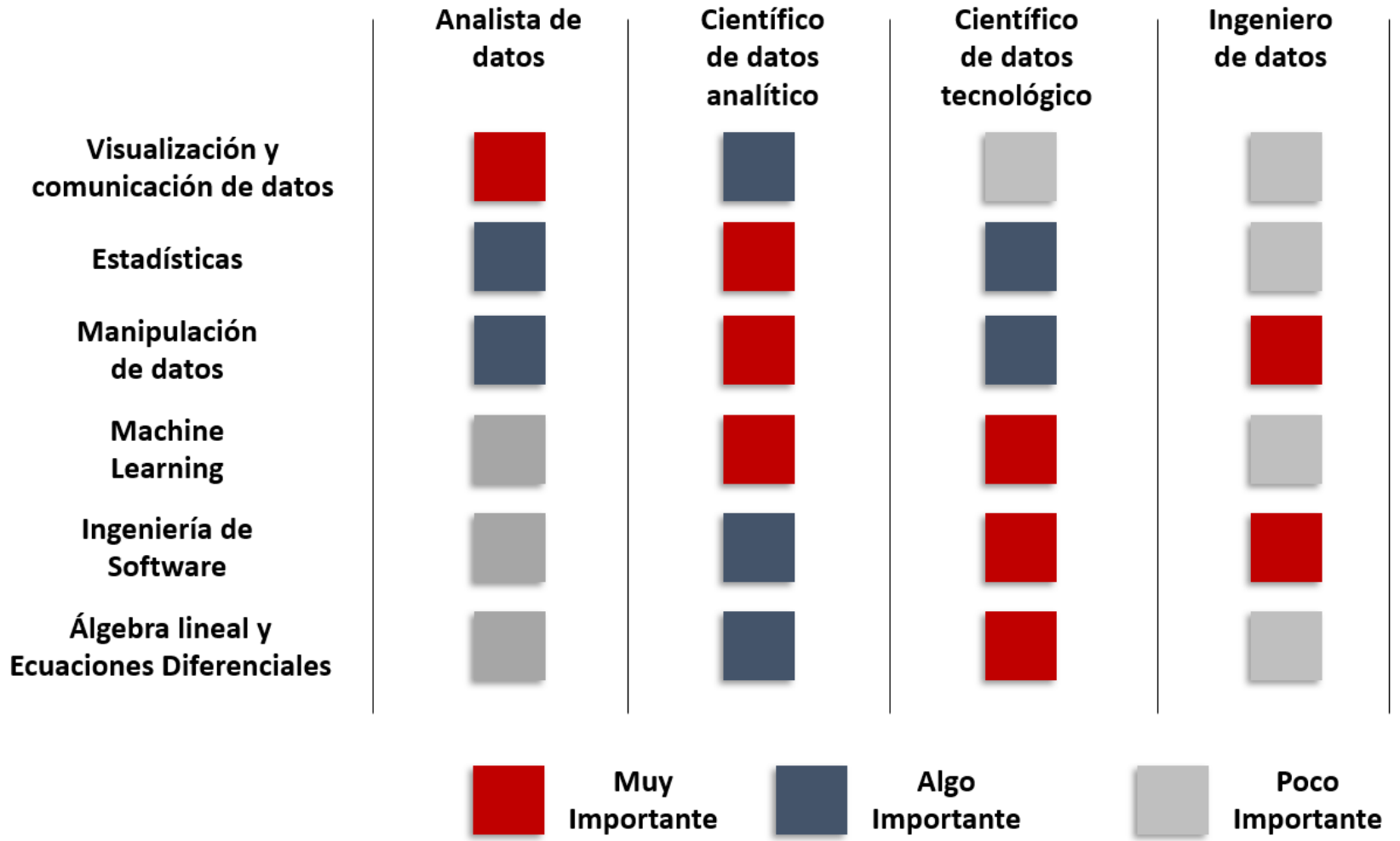
Habilidades de un Científico de Datos (2)



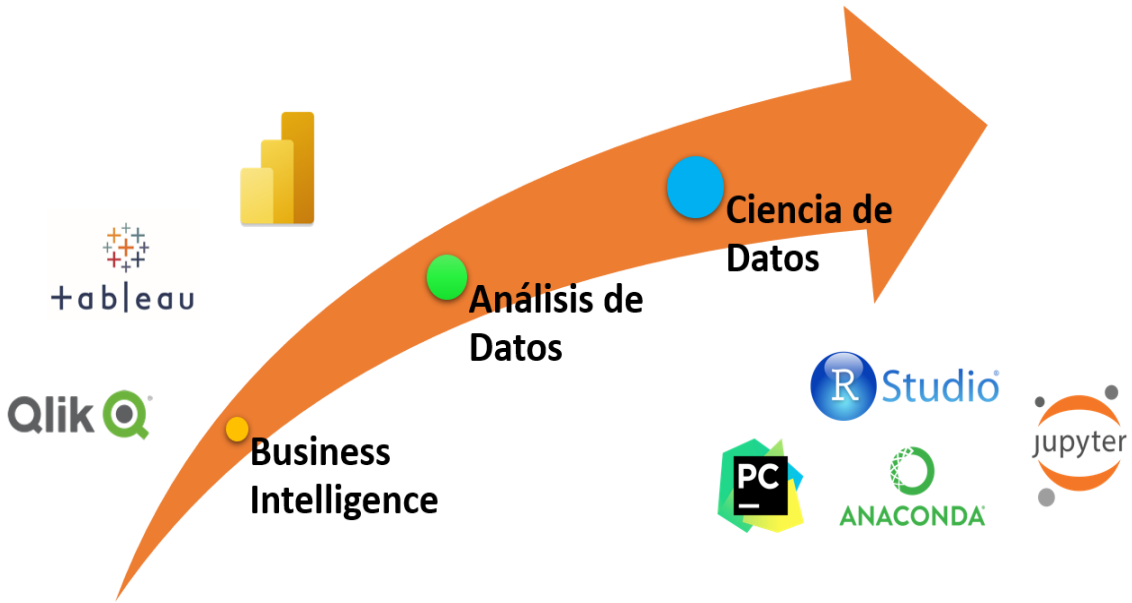
- Habilidades en el manejo de:
  - La evaluación, procesamiento de resultados para saber si es correcto o no la metodología aplicada. Por tanto, se requiere de una base estadística para usar herramientas como la matriz de confusión, la curva ROC para entender la precisión de los modelos que se obtienen a través del procesamiento de los datos.
  - La explicación de resultados ¿por qué es importante conocer el dominio, las reglas del negocio? Para conocer de lo que se habla, conocimiento técnico con relación a los cálculos que se realizan, saber comunicar, es decir, conocimiento del dominio para explicar claramente una interpretación de los resultados y cómo fueron éstos obtenidos.

En resumen, un científico de datos debe tener la habilidad y conocimiento de técnicas de modelado estadístico, conocimientos en matemáticas; particularmente álgebra lineal, desarrollo de software principalmente relacionado con lenguajes como R y Python, así como competencias en visualización de datos.

Perfiles en un Científico de Datos (1)

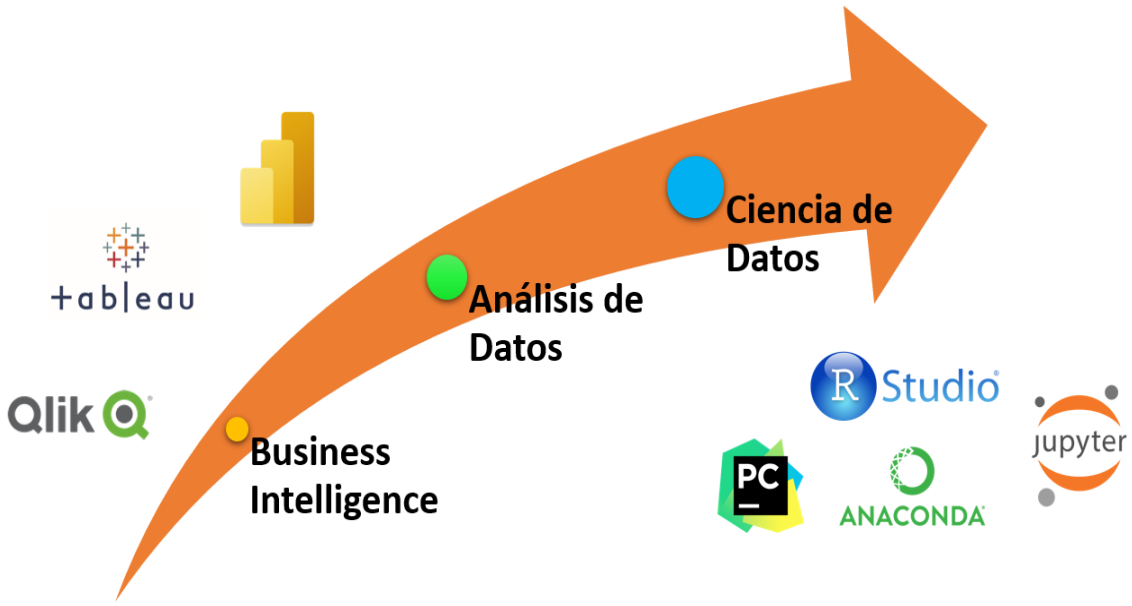


Evolución del Análisis de Datos (1)



- La evolución del análisis de datos ha sufrido varios cambios con respecto al panorama de la analítica:
  - Muchas personas iniciaron en el manejo de datos a través de Business Intelligence, utilizando aplicaciones como: Qlik, ClickSense,Tableau, Power BI, iniciando con la interacción de los datos, creación de información gráfica sobre las principales áreas de un dominio de aplicación, procesos o tareas de análisis.
  - Se inició el trabajo con indicadores como KPIs, brindando información sobre el funcionamiento de algunas áreas o tareas específicas
  - ¿Cómo se ha hecho en el pasado para llegar al resultado actual? Entonces, cuando ya se tiene una madurez dentro de la Inteligencia del Negocio y se ha generado una cultura de "alfabetización" de los datos, donde ya existen los mecanismos, búsquedas y usos de elementos para tomar decisiones, generación de gráficas, dashboards presentados en diversas plataformas, entre otros. Entonces pueden existir ya modelos de datos generados a partir de BI, conectados a una fuente de datos original que es un sistema transaccional del sitio, así como a fuentes de datos externas como redes sociales, sitios web, entre otros.

Evolución del Análisis de Datos (2)



- La evolución del análisis de datos ha sufrido varios cambios con respecto al panorama de la analítica:
  - Ya cuando se tiene ese grado de madurez de BI, estamos enfocados en una comprensión del pasado o de una manera descriptiva de cómo se ha llegado a la situación actual. Al contar con este marco de trabajo se puede avanzar a la generación de pronósticos, la realización de segmentaciones, entre otros.
  - De esta manera se avanza formalmente al análisis de datos, incorporando algunos modelos de análisis, pruebas estadísticas que permitan evaluar la calidad de esos modelos, sus predicciones e incorporando dentro de las plataformas algoritmos que ayuden a realizar estos análisis.
  - Entre las principales plataformas se tienen R y Python, que generan nuevos resultados y ofrecen una nueva experiencia a usuarios finales, a través de dashboards, informes utilizando herramientas de estadística empresarial, etc.
  - Cuando se tiene implementado el marco de trabajo del análisis de datos, es necesario utilizar otras herramientas para crear ciencia de datos, con la intención que estos modelos de análisis sean reproducibles, más científicos en la explicación y tengan un sustento de pruebas estadísticas para satisfacer una hipótesis.
  - Entre las herramientas y las plataformas más comunes se tienen:
    - R Studio que es la interfaz para R por excelencia
    - Interfaces para Python como Spyder, Anaconda, Jupyter Notebook.
      - Jupyter Notebook es un entorno de desarrollo interactivo basado en la web para cuardernos, código y datos. Con una interfaz flexible que permite al usuario configurar y organizar flujos de trabajo de CD, computación científica, periodismo computacional y ML. Un diseño modular permite agregar extensiones para ampliar y enriquecer la funcionalidad, con lo cual se busca generar también estadística predictiva y empezar a realizar inferencias.

R como lenguaje para Ciencia de Datos (1)

- R posee unas características especiales que lo hacen versátil para el manejo de elementos estadísticos, en concreo para operaciones con matrices y vectores, lo que facilita la manipulación de bases de datos. Por tanto, R permite manipular (seleccionar, recodificar y recuperar) datos rápidamente. De hecho existen algunos paquetes diseñados para ello como plyr, lo que hacen que este lenguaje de programación sea más hábil y eficiente en la preparación de los datos para su posterior análisis.
  - Fue un lenguaje diseñado específicamente para hacer análisis estadístico, es muy preciso y exacto para el análisis de datos.
  - Dispone de una gran cantidad de paquetes para la creación de gráficos, lo que aporta capacidades avanzadas en la visualización de los datos y los resultados del análisis. Incluye un paquete básico para funciones gráficas y se pueden agregar otros como lattice o ggplot.
  - Para Machine Learning, R tiene implementados una gran cantidad de algoritmos, como consecuencia de las diferentes líneas de investigación de grupos que dieron pie a su creación, debido precisamente al hecho de que R nació en el ámbito académico.
  - R tiene un enfoque orientado al análisis estadístico, lo que lo hace muy útil para la minería de datos. Es un multiparadigma orientado a objetos, vectorial y multiplataforma, tiene una gran cantidad de desarrolladores que lo mejoran y enriquecen.
  - R tiene una curva de aprendizaje más lenta a comparación de Phyton.

R como lenguaje para Ciencia de Datos (2)

- Espacios colaborativos para el científico de datos
  - Ejemplo de manipulación de datos COVID en lenguaje R.
    - Se hacen las tareas de extracción de datos, de conexión de datos, transformación de esos datos y visualización de los datos a través de una gráfica. En este ejemplo no se aplicó ninguna prueba estadística para validar esos datos, ni se aplicó ninguna distribución sobre éstos. Simplemente los datos fueron conectados, transformados, limpiados y se creó un modelo simple, el cual contiene solo las variables de interés para posteriormente crear la visualización de esa información.
  - Ejemplo con datos de población mundial.

GitHub

RPubs

Recursos

- Recursos de Python para #DataScience y #MachineLearning
  - Programación básica en Python:
    - Listas, Tuplas, Diccionarios, Condicionales, Loops, etc.
    - Estructuras de Datos y Algoritmos
    - NumPy Arrays
    - Regular expressions (Regex)
  - Manipulación de Datos
    - Pandas
    - SQLAlchemy
  - Visualización de Datos
    - Matplotlib
    - Seaborn
    - Plotlib
    - Python Graph Library
  - Machine Learning / Deep Learning
    - Scikit-learn Tutorial
    - Deep Learning Tutorial
    - Kaggle Kernels
    - Otros cursos