Introducción a Ciencia de Datos y Machine Learning

Contents

BIENVENIDA

Objetivo

Brindar al participante los elementos teóricos y prácticos básicos alrededor de la programación para el análisis de datos. Aprenderá a distinguir las diferentes soluciones a problemas que pueden resolverse con algoritmos de machine learning y aprenderá a usar el conjunto de librerías en **R** más novedosas, estructuradas y ampliamente usadas para la manipulación, transformación y visualización de datos: *"TIDYVERSE"*.

Instructores

ACT. ARTURO BRINGAS

LinkedIn: arturo-bringas **Email:** act.arturo.b@ciencias.unam.mx

Actuario egresado de la Facultad de Ciencias con maestría en Ciencia de Datos por el ITAM. Se especializa en modelos predictivos y de clasificación de *machine learning* aplicado a seguros, banca, marketing, deportes, e-commerce y movilidad internacional. Ha sido consultor *Senior Data Scientist* para empresas y organizaciones como GNP, El Universal, UNAM, la Organización de las Naciones Unidas Contra la Droga y el Delito (UNODC), entre otros. Actualmente es profesor de *Ciencia de datos y Machine Learning* en AMAT y *Data Scientist Expert* en BBVA, en donde implementa soluciones de analítica avanzada con impacto global.

ACT. KARINA LIZETTE GAMBOA

LinkedIn: KaLizzyGam **Email:** lizzygamboa@ciencias.unam.mx

Actuaria egresada de la Facultad de Ciencias y candidata a Maestra en Ciencia de Datos por el ITAM.

Experiencia en áreas de analítica predictiva e inteligencia del negocio. Lead y Senior Data Scientist en consultoría en diferentes sectores como tecnología, asegurador, financiero y bancario. Es experta en entendimiento de negocio para la correcta implementación de algoritmos de inteligencia y explotación de datos. Actualmente se desarrolla como Arquitecta de Soluciones Analíticas en Merama, startup mexicana clasificada como uno de los nuevos unicornios de Latinoamérica. Senior Data Science en CLOSTER y como profesora del diplomado de Metodología de la Investigación Social por la UNAM así como instructora de cursos de Ciencia de Datos en AMAT.

Empresas anteriores: GNP, Actinver Banco y Casa de Bolsa, PlayCity Casinos, RakenDataGroup Consulting, entre otros.

Alcances del curso

Al finalizar este curso el participante será capaz de consumir, manipular y visualizar información proveniente de diversas fuentes de información para resolver problemas de propósito general asociados a los datos.

Requisitos:

- Computadora con al menos 8Gb Ram
- Instalar la versión más reciente de R
- Instalar la versión más reciente de RStudio

Temario:

1. Introducción a Ciencia de Datos

- Machine Learning, Bigdata, BI, AI y CD
- Objetivo de ciencia de datos

- Requisitos y aplicaciones
- Tipos de algoritmos
- Ciclo de vida de un proyecto

2. Manipulación de datos con Tidyverse

- Importación de tablas (readr)
- Consultas (dplyr)
- Transformación de estructuras (tidyr)

3. Concepto de Machine Learning

- Machine learning
- Análisis supervisado
- Análisis no supervisado
- Sesgo y varianza
- Partición de datos
- Preprocesamiento e ingeniería de datos

4. Algoritmos de Machine Learning

- Clustering: Kmeans, kmedoids, agnes
- Regresión Lineal
- Métricas de error
- Regresión logística
- Métricas de error
- KNN
- Árbol de decisión
- Random Forest
- Comparación de modelos

Duración y evaluación del curso

- El programa tiene una duración de 40 hrs.
- Las clases serán impartidas los días sábado, de 9:00 am a 1:00 pm
- Serán asignados ejercicios que el participante deberá resolver entre una semana y otra.
- Al final del curso se solicitará un proyecto final, el cual deberá ser entregado para ser acreedor a la constancia de participación.

Recursos y dinámica de clase

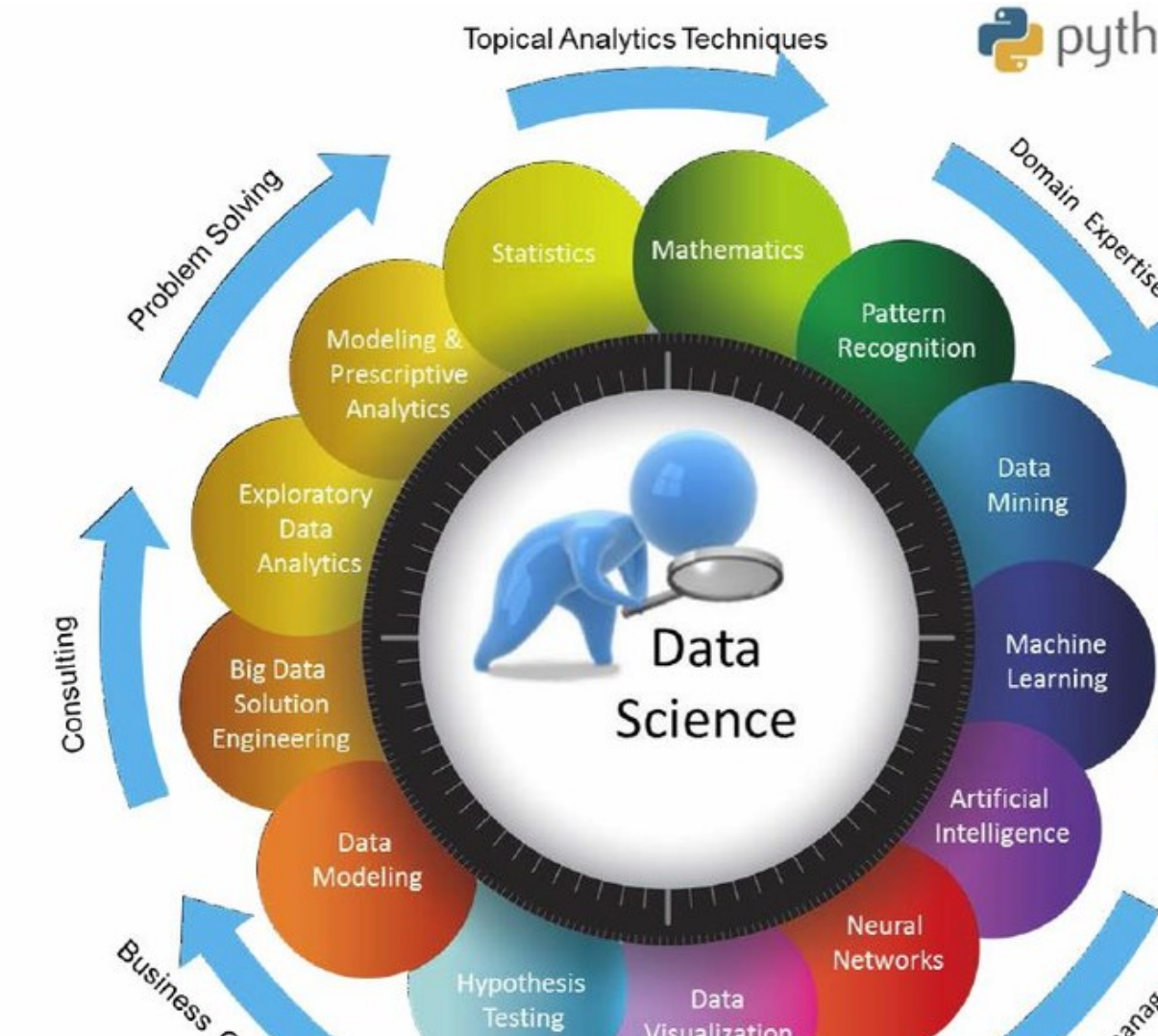
En esta clase estaremos usando:

- R da click aquí si aún no lo descargas
- RStudio da click aquí también
- Miro úsame
- Zoom Clases
  - Pulgar arriba: Voy bien, estoy entendiendo!
  - Pulgar abajo: Eso no quedó muy claro
  - Mano arriba: Quiero participar/preguntar ó Ya estoy listo para iniciar
- Grupo de WhatsApp El chismecito está aquí
- One Drive
- Notas de clase Revisame si quieres aprender

Chapter 1

Conceptos de Ciencia de Datos

1.1 ¿Qué es Ciencia de Datos?



Definiendo conceptos:

Estadística Disciplina que recolecta, organiza, analiza e interpreta datos. Lo hace a través de una población muestral generando estadística descriptiva y estadística inferencial.

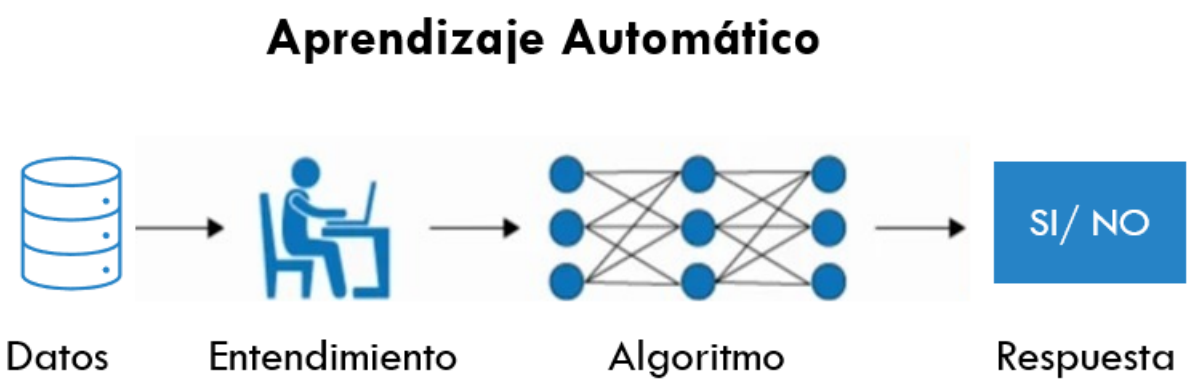
- La estadística descriptiva, como su nombre lo indica, se encarga de describir datos y obtener conclusiones. Se utilizan números (media, mediana, moda, mínimo, máximo, etc) para analizar datos y llegar a conclusiones de acuerdo a ellos.
- La estadística inferencial argumenta o infiere sus resultados a partir de las muestras de una población. Se intenta conseguir información al utilizar un procedimiento ordenado en el manejo de los datos de la muestra.
- La estadística predictiva busca estimar valores y escenarios futuros más probables de ocurrir a partir de referencias históricas previas. Se suelen ocupar como apoyo características y factores áltamente asociados al fenómeno que se desea predecir.

Business Intelligence: BI aprovecha el software y los servicios para transformar los datos en conocimientos prácticos que informan las decisiones empresariales estratégicas y tácticas de una organización. Las herramientas de BI acceden y analizan conjuntos de datos y presentan hallazgos analíticos en informes, resúmenes, tableros, gráficos, cuadros, -indicadores- o KPI's y mapas para proporcionar a los usuarios **inteligencia detallada sobre el estado del negocio**. BI esta enfocado en analizar la historia pasada para tomar decisiones hacia el futuro.

¿Qué características tiene un KPI?

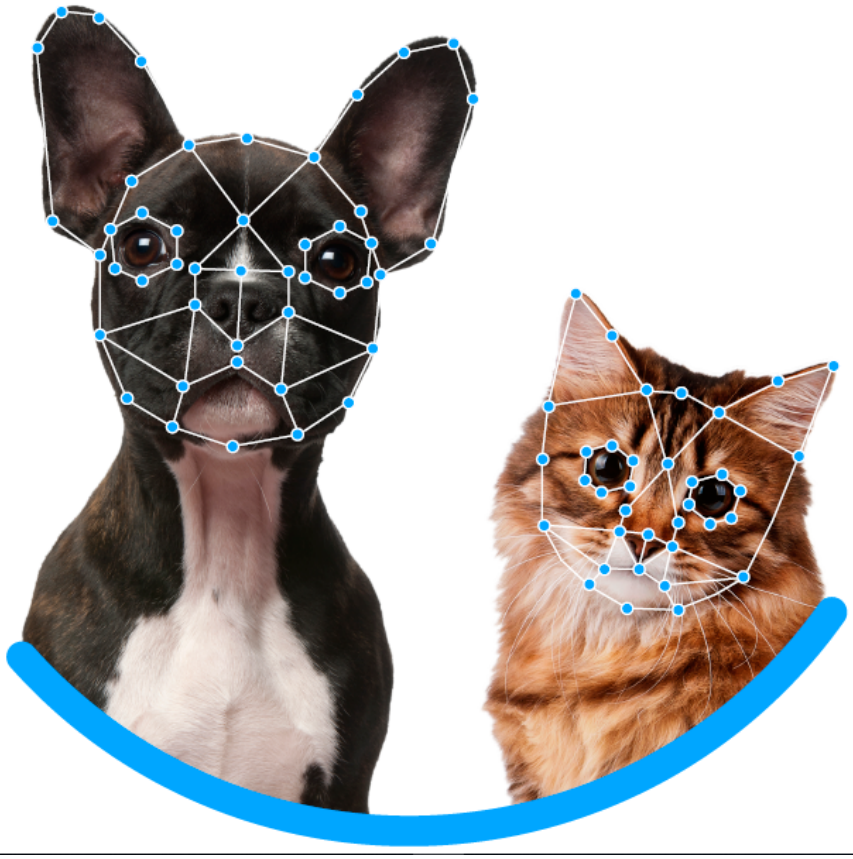
- Específicos
- Continuos y periódicos
- Objetivos
- Cuantificables
- Medibles
- Realistas
- Concisos
- Coherentes
- Relevantes

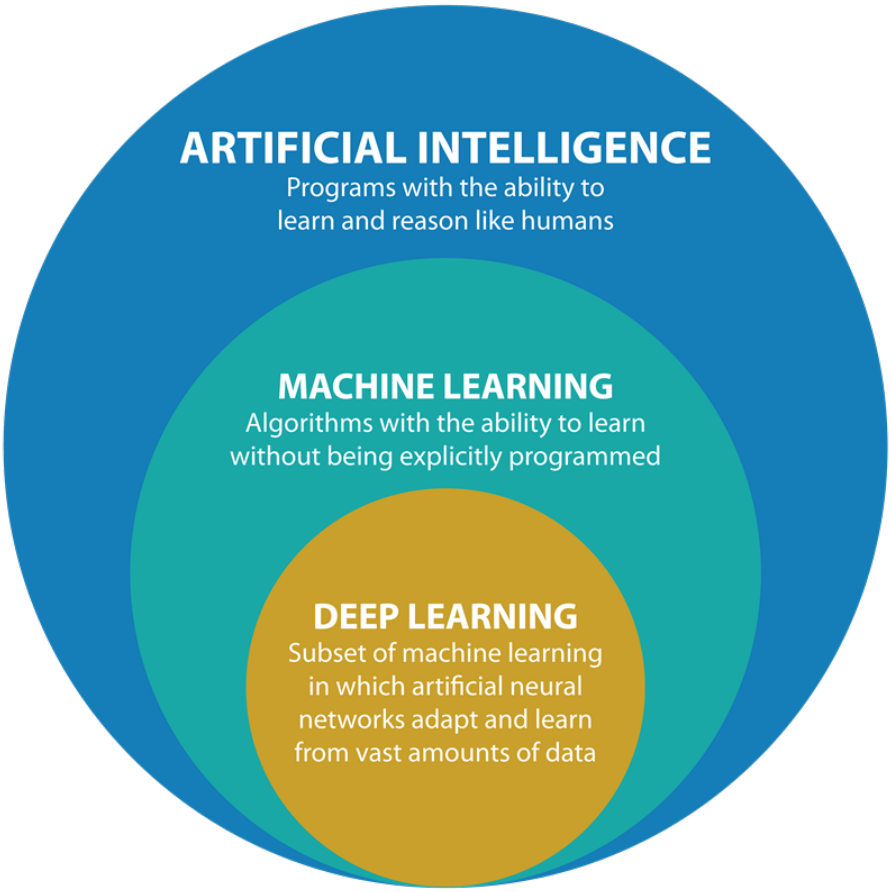
Machine Learning: Machine learning –aprendizaje de máquina– es una **rama de la inteligencia artificial** que permite que las máquinas aprendan de los patrones existentes en los datos. Se usan métodos computacionales para aprender de datos con el fin de producir reglas para mejorar el desempeño en alguna tarea o toma de decisión. (Está enfocado en la programación de máquinas para aprender de los patrones existentes en datos principalmente estructurados y anticiparse al futuro)



Deep Learning: El aprendizaje profundo es un subcampo del aprendizaje automático que se ocupa de los algoritmos **inspirados en la estructura y función del cerebro llamados redes neuronales artificiales.**

En *Deep Learning*, un modelo de computadora aprende a realizar tareas de clasificación directamente a partir de imágenes, texto o sonido. Los modelos de aprendizaje profundo pueden lograr una precisión de vanguardia, a veces superando el rendimiento a nivel humano. Los modelos se entrenan mediante el uso de un gran conjunto de datos etiquetados y arquitecturas de redes neuronales que contienen muchas capas. (Está enfocado en la programación de máquinas para el reconocimiento de imágenes y audio (datos no estructurados))





Big data se refiere a los grandes y diversos conjuntos de información que crecen a un ritmo cada vez mayor. Abarca el volumen de información, la velocidad a la que se crea y recopila, y la variedad o alcance de los puntos de datos que se cubren. Los macrodatos a menudo provienen de la minería de datos y llegan en múltiples formatos.


1 VOLUMEN
Cantidad de datos generados cada segundo, minuto y días
2
VELOCIDAD
Rapidez en la que los
datos son creados, al
cenados y procesado
VARIE
Formas, ti
fuentes en
registran l
VERACIDAD
Grado de fiabilidad
de la información
recibida
5
VIABILIDAD
Uso eficáz del gran
volumen de datos
6
VISUALIZACIÓN
Modo en el que los
datos son presentados
7
VALOR
El valor se obtiene
de datos.
8
VISCOSIDAD
¿Son importante los
datos para la toma de
decisiones?
BIG DATA
Las 8 V's

Es común que se confunda los conceptos de *Big Data* y *Big Compute*, como se mencionó, *Big Data* se refiere al procesamiento de conjuntos de datos que son más voluminosos y complejos que los tradicionales y *Big Compute* a herramientas y enfoques que utilizan una gran cantidad de recursos de CPU y memoria de forma coordinada para resolver problemas que usan algoritmos muy complejos.

Curiosidad: Servidores en líquido para ser enfriados

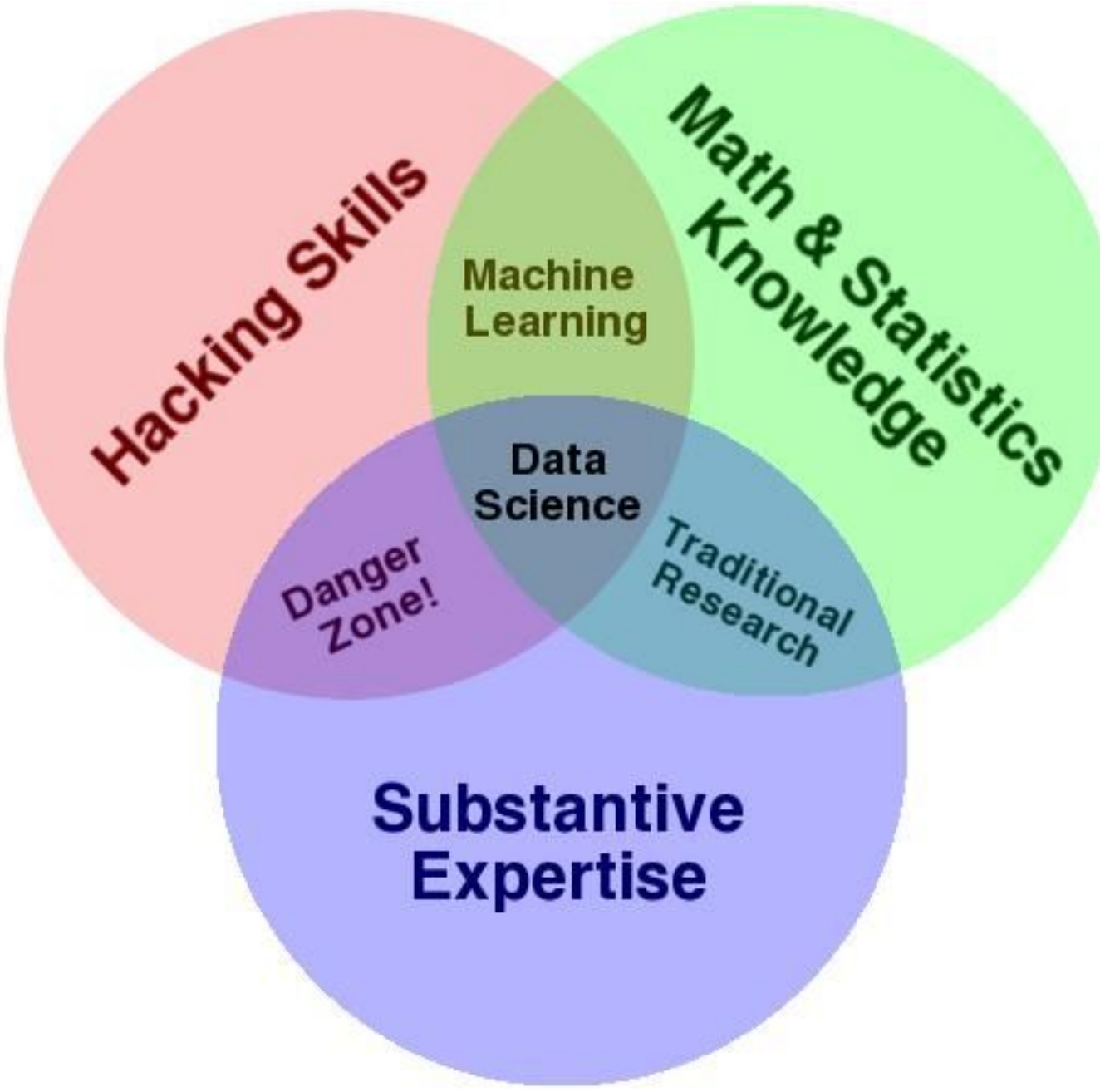
Curiosidad 2: Centro de datos en el océano

Entonces, ¿qué NO es ciencia de datos?

- No es una tecnología
- No es una herramienta
- No es desarrollo de software
- No es Business Intelligence*
- No es Big Data*
- No es Inteligencia Artificial*
- No es (solo) machine learning
- No es (solo) deep learning
- No es (solo) visualización
- No es (solo) hacer modelos

1.2 Objetivos

- Los científicos de datos analizan qué preguntas necesitan respuesta y dónde encontrar los datos relacionados. Tienen conocimiento de negocio y habilidades analíticas, así como la capacidad de extraer, limpiar y presentar datos. Las empresas utilizan científicos de datos para obtener, administrar y analizar grandes cantidades de datos no estructurados. Luego, los resultados se sintetizan y comunican a las partes interesadas clave para impulsar la toma de decisiones estratégicas en la organización.



Fuente: Blog post de Drew Conway

Más sobre Conway: Forbes 2016

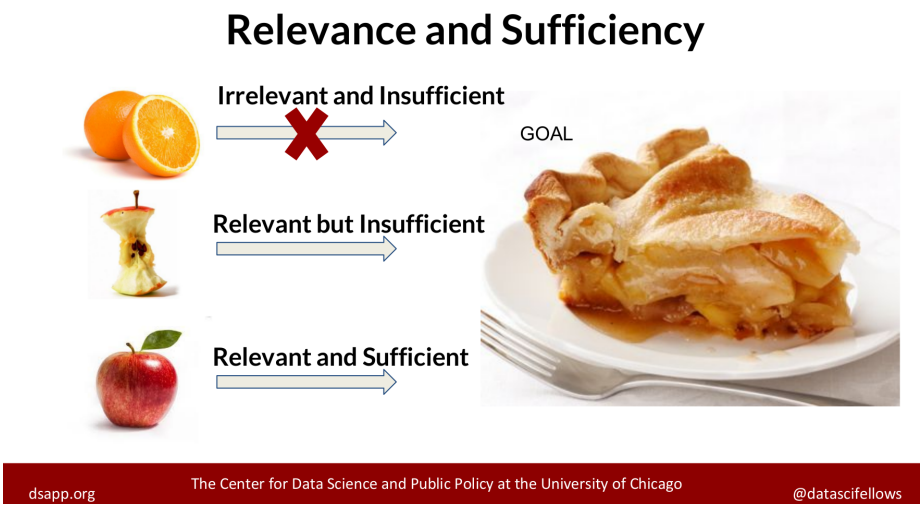
1.3 Requisitos

- **Background científico:** Conocimientos generales de probabilidad, estadística, álgebra lineal, cálculo, geometría analítica, programación, conocimientos computacionales... etc

- **Datos relevantes y suficientes:** Es indispensable saber si los datos con los que se trabajará son relevantes y suficientes, debemos evaluar qué preguntas podemos responder con los datos con los que contamos.

  - **Suficiencia:** Los datos con los que trabajamos tienen que ser representativos de la población en general, necesitamos que las características representadas en la información sean suficientes para aproximar a la población objetivo.

  - **Relevancia:** De igual manera los datos tienen que tener relevancia para la tarea que queremos resolver, por ejemplo, es probable que información sobre gusto en alimentos sea irrelevante para predecir número de hijos.



- **Etiquetas:** Se necesita la intervención humana para etiquetar, clasificar e introducir los datos en el algoritmo.

- **Software:** Existen distintos lenguajes de programación para realizar ciencia de datos

1.4 Aplicaciones

Dependiendo de la industria en la que se quiera aplicar Machine Learning, podemos pensar en distintos enfoques, en la siguiente imagen se muestran algunos ejemplos:


MACHINE LEARNING: USOS Y APLICACIONES
1
ENERGÍA
Predecir fallas en refinerías
Localizar nuevas fuentes de energía
Analizar minerales
2
SERVICIOS
Fijar precios acorde
Alcanzar un ritmo d
óptimo
3
GOBIERNO
Elevar eficiencia y ahorros
Minimizar el robo de identidad
Prevenir la corrupción
4
TRANSP
Identific
eficiente
Predecir
5
SALE
MINORISTAS
Mejorar campañas de mercadotecnia
Personalizar la oferta
Reducir la pérdida de clientes durante el proceso de compra
Mejorar la experiencia de compra
6
HOSPITALES
Incrementar el éxito de una operación
Predecir tiempos de espera en urgencias
Prevenir infartos y convulsiones
7
FINANCI
Prevenir
incobrab
Predecir
Prevenir
de dinero

Podemos pensar en una infinidad de aplicaciones comerciales basadas en el análisis de datos. Con la intención de estructurar las posibles aplicaciones, se ofrece a continuación una categorización que, aunque no es suficiente para englobar todos los posibles casos de uso, sí es sorprendente la cantidad de aplicaciones que abarca.

1. Aplicaciones centradas en los clientes

- Incrementar beneficio al mejorar recomendaciones de productos
- Up-selling
- Cross-selling
- Reducir tasas de cancelación y mejorar tasas de retención
- Personalizar experiencia de usuario
- Mejorar el marketing dirigido
- Análisis de sentimientos
- Personalización de productos o servicios

2. Optimización de problemas

- Optimización de precios
- Ubicación de nuevas sucursales
- Maximización de ganancias mediante producción de materias primas
- Construcción de portafolios de inversión

3. Predicción de demanda

- Número futuro de clientes
- Número esperado de viajes en avión / camión / bicis
- Número de contagios por un virus (demanda médica / medicamentos / etc)
- Predicción de uso de recursos (luz / agua / gas)

4. Análisis de detección de fraudes

- Detección de robo de identidad
- Detección de transacciones ilícitas
- Detección de servicios fraudulentos
- Detección de zonas geográficas con actividades ilícitas