# Modulo 1. Introducción a la ciencia de datos

## Ecosistemas, roles, metodología y estrategia de negocios

### Ecosistema de ciencia de datos

#### What is Data Science?

La Ciencia de datos engloba una serie de principios, definición de problemas, algoritmos y procesos para extraer patrones no-obvios y útiles de grandes volúmenes de datos.

Los términos ciencia de datos, aprendizaje automático y minería de datos comúnmente son utilizados como sinónimo. La similitud entre estas disciplinas es un enfoque en la mejora de la toma de decisiones mediante el análisis de datos.

#### Recopilación

A close-up of a blue dot

AI-generated content may be incorrect.

* Datos transaccionales. Ordenes de compra, recibos, etc.
* Modelo relacional. Bases de datos estructuradas.
* Data Warehouses. Integrar y analizar datos provenientes de diversas fuentes.
* Big Data. Volúmenes difíciles o imposibles de procesar con métodos tradicionales.
* Bases de datos NoSQL (No relacionales). Imágenes, videos, datos de páginas web.
* MapReduce. Sistema de procesamiento de datos paralelo y distribuido.

#### Análisis

A blue and grey line with black text

AI-generated content may be incorrect.

* Estadística descriptiva. Medidas de tendencia central y variabilidad.
* Aprendizaje estadístico. Regresión lineal y logística.
* Análisis discriminatorio. Clasificación o reconocimiento de patrones
* Machine Learning. Aprendizaje de datos en lugar de programación explicita
* Ensambles. Combinan múltiples modelos para obtener un mejor rendimiento predictivo.
* Deep Learning, Redes Neuronales.

### Roles clave en ciencia de datos

A close-up of a checklist

AI-generated content may be incorrect.

### Metodología CRISP-DM

Es la representación más utilizada del ciclo de vida de un proyecto de datos.

**C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Publicado en 1999 para estandarizar los procesos de data mining en todas las industrias, se ha convertido en la metodología más popular para proyectos de data mining, analytics y science.

Se compone de seis fases secuenciales.

#### Entendimiento de negocio

Esta fase se enfoca en entender el objetivo y requerimientos del proyecto, lo cual incluye cuatro objetivos.

1. **Determinar los objetivos de negocio.** ¨Que es lo que el cliente quiere realmente lograr¨y con esto definir el criterio de éxito.
2. **Evaluar la situación.** Determinar la disponibilidad de recursos, requerimientos, riesgos y medidas de contingencia y realizar un análisis costo-beneficio.
3. **Determinar los objetivos del proyecto.** Adicional a los objetivos de negocio, se debe definir que consideramos como ¨éxito¨ desde un punto de vista técnico.
4. **Planeación de proyecto.** Seleccionar tecnologías, herramientas y definir planes detallados para cada fase.

#### Entendimiento de datos

El objetivo de esta fase es identificar, recabar y analizar datasets que permitan lograr el objetivo del proyecto. Esta fase se compone de cuatro tareas:

1. **Recabar data inicial.** Si es necesario, cargar la data dentro de nuestra herramienta de análisis.
2. **Descripción de la data.** Examinar y documentar sus propiedades superficiales como formato, número de registros o identidades de campo.
3. **Exploración de la data.** Profundizar, consultar, visualizar e identificar relaciones.
4. **Verificar la calidad de la data.** Documentar cualquier problema con la calidad de los datos.

#### Preparación de datos

Representa hasta el 80% del proyecto. En esta fase se prepara el data set final para el modelado. Se compone de cinco tareas.

1. **Seleccionar la data.** Documentar motivos para inclusiones/exclusiones de información.
2. **Limpieza de datos.** Por lo general la fase mas laboriosa. Corregir, imputar o eliminar valores erróneos.
3. **Construir data.** Derivar nuevos atributos de utilidad.
4. **Integrar data.** Crear nuevos datasets al combinar datos de múltiples fuentes.
5. **Formateo de data.** Cambio de formato, si es necesario.

#### Modelado

Construir y evaluar múltiples modelos con base a diferentes técnicas de modelado. Se compone de cuatro tareas:

1. **Seleccionar la técnica de modelado de datos.** Determinar algoritmos a probar (regresión, red neuronal, clasificación)
2. **Generar diseño prototipo.** Dividir el dataset en training, testing, validation subsets.
3. **Construir el modelo.**
4. **Evaluar el modelo.** Normalmente, varios algoritmos de modelado son comparados e interpretados con base a sus resultados y dominio del área especifica de estudio, la condición de éxito preestablecida y el modelo de prueba.

#### Evaluación

La fase de evaluación analiza de forma mas amplia qué modelo se adapta mejor en términos de negocio y pasos a seguir. Consta de tres tareas:

1. **Evaluación de resultados.** ¿El modelo cumple con las condiciones de éxito?, ¿Qué modelo debemos aprobar para el negocio?
2. **Proceso de revisión.** Evaluar el trabajo logrado. ¿Se está ignorando algún detalle?, ¿Fueron todos los pasos ejecutados apropiadamente?, ¿Cuál es la “deuda técnica” que estamos dejando pendiente? Resumir hallazgos y se corrige lo necesario.
3. **Determinar siguientes pasos.** Con base en las tareas anteriores, se determina si proceder a llevar a producción, iterar o iniciar nuevos proyectos.

#### Despliegue

Un modelo no es de utilidad hasta que sus usuarios finales pueden acceder a sus resultados. La complejidad de esta fase varía ampliamente. Conlleva cuatro tareas:

1. **Planear el despliegue.** Desarrollar y documentar planes de despliegue del modelo.
2. **Definir monitoreo y mantenimiento.** Esto evita problemas durante la fase de operación del modelo.
3. **Elaborar un reporte final.** El equipo del proyecto documenta un resumen del proyecto.
4. **Evaluación de proyecto.** En retrospectiva, ¿Qué funcionó?, ¿Qué se pudo haber hecho mejor? Cómo mejorar en el futuro.

#### Beneficios

* Generalizable
* Sentido común. Cuando estudiantes comienzan un proyecto de ciencia de datos sin un Project manager tienden a adoptar una metodología similar a CRISP-DM.
* Adoptable
* Inicio correcto. El enfoque inicial en entendimiento de negocio ayuda a alinear las tareas técnicas con las necesidades de negocio.
* Término fuerte. La fase final implica consideraciones importantes con respecto a la transición y mantenimiento del proyecto en operación.
* Flexibilidad

#### Desventajas

* Rígido
* Enfoque en la documentación
* “Anticuado”
* No es una metodología de “Project management”. Debido a que esta metodología asume que el usuario es una sola persona o un equipo pequeño e ignora la colaboración necesaria para proyectos de mayor tamaño.

### Estrategia de negocios con ciencia de datos

Para lograr que un proyecto de ciencia de datos este alineado con la estrategia de negocios hay que tener claridad del tipo de preguntas que surgen durante la realización.

* Descriptivas. ¿Como son las cosas?
  + ¿Cuántos clientes estamos perdiendo? ¿Quiénes son esos clientes?
* Predictivas. ¿Como creemos que serán las cosas?
  + ¿Puedo saber anticipadamente que clientes perderemos?
* Prescriptivas. ¿Como deberían ser las cosas?
  + Si se puede anticipar que clientes se irán, ¿qué tipo de políticas de retención deberíamos usar y como asignarlas?