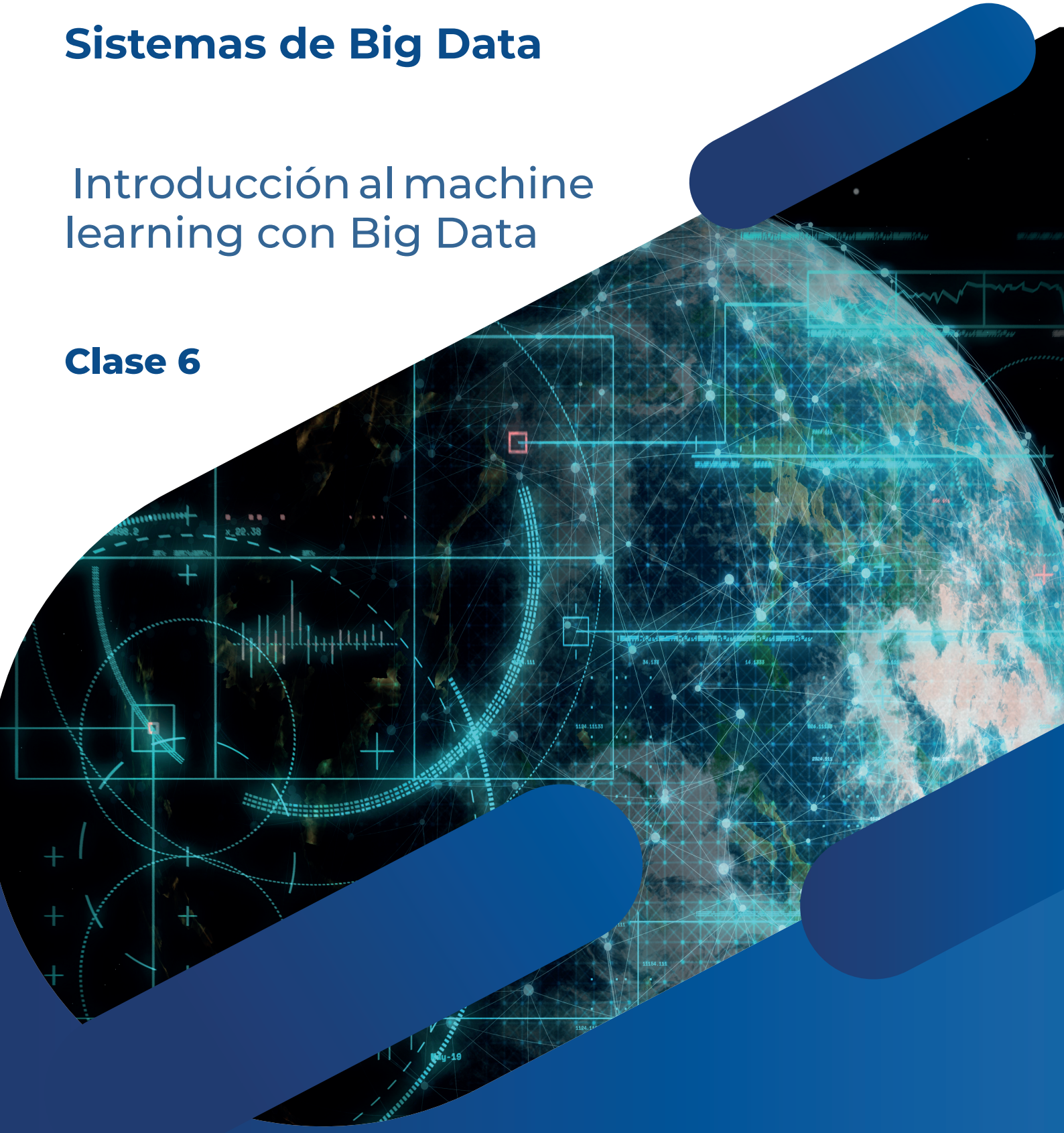
Sistemas de Big Data

Introducción al machine learning con Big Data

Clase 6



MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

INTRODUCCIÓN

El campo del **machine learning (ML)** ha experimentado un crecimiento significativo en los últimos años, impulsado por el acceso a grandes volúmenes de datos y por los avances en la potencia computacional. En el contexto de **Big Data**, el **ML** se convierte en una herramienta **fundamental** para descubrir patrones y obtener conocimientos útiles a partir de datos **masivos**. El proceso de **aprendizaje automático**, que incluye la recopilación, **preprocesamiento** y análisis de grandes cantidades de datos, permite desarrollar modelos **predictivos** y **descriptivos,** que pueden aplicarse en múltiples industrias, desde la salud hasta las finanzas (Bishop, 2006). En esta clase, exploraremos los fundamentos del **ML**, las categorías de técnicas existentes y su integración con **Big Data**, con un enfoque en la metodología **CRISP-DM**, que proporciona un marco estructurado para abordar proyectos de análisis de datos.

Las técnicas de **ML** se dividen en diversas categorías, cada una adecuada para diferentes tipos de problemas y conjuntos de datos. Estas incluyen los aprendizajes supervisado, no supervisado y por refuerzo; cada uno con sus propias características y aplicaciones prácticas. En el contexto de **Big Data**, las capacidades de estas técnicas se multiplican debido a la escala y complejidad de los datos involucrados. Además, es esencial comprender el proceso de **ML**, que abarca desde la definición del problema hasta la evaluación del modelo final. El **modelo CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) es uno de los enfoques más utilizados en proyectos de machine learning, proporcionando una guía que abarca las etapas desde la comprensión del negocio hasta la **implementación** del **modelo** (Chapman et al., 2000). Este enfoque estructurado es crucial cuando se trabaja con Big Data, ya que permite gestionar la complejidad del proceso y **garantizar resultados efectivos**.

RDA 2: Aplicar técnicas de Big Data en grandes volúmenes de datos

Clase 6 Introducción al machine learning con Big Data

6. 1. *Machine Learning Overview*

Los objetivos de esta sección son los siguientes:

- Que el estudiante pueda explicar qué es el ***machine learning.***
- Que el estudiante pueda listar al menos tres aplicaciones del ***machine learning*** identificadas en su día a día.

El ML es un campo de estudio que se enfoca en sistemas computacionales que pueden aprender a **partir** de los **datos**. Esos sistemas, que frecuentemente son llamados modelos, pueden aprender a realizar una tarea en específico, a partir del análisis de muchos ejemplos de un problema en particular. Por ejemplo, un modelo de ML puede aprender a reconocer una imagen de un gato a partir de haber visto muchas y muchas **imágenes** de **gatos**.

Figura 1. Características de un gato.

Esta noción de aprender a partir de datos significa que el modelo de ML puede aprender una tarea específica sin ser programado explícitamente para ello. En otras palabras, el modelo de ML **no es un conjunto de instrucciones,** paso a paso, de cómo reconocer la imagen de un gato. En lugar de eso, el modelo aprende qué características son importantes para determinar si la imagen contiene un gato, a partir de los datos que ha **analizado**.

Dado que el modelo aprende a realizar esta tarea a partir de los datos, es importante saber que la **cantidad** y la **calidad** de estos son factores **importantes** a la hora de construir un buen modelo. Como el modelo aprende a partir de los datos, este puede ser usado para descubrir patrones ocultos y tendencias en los datos. Estos patrones y tendencias permiten extraer ***insights*** valiosos de los datos. Por esta razón, el uso del ***machine learning*** toma decisiones de tipo ***data-driven;*** es decir, decisiones a partir de los datos.

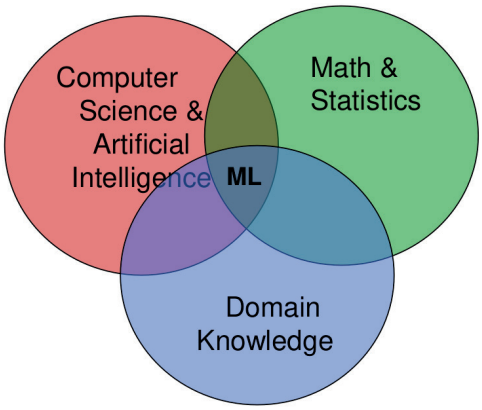


Figura 2. *Data Scientist Skills.*

Para resumir, el campo del **ML** se enfoca en el estudio y construcción de sistemas computacionales que pueden ser aprendidos a base de datos, sin ser explícitamente programados para ello. Los algoritmos y técnicas de ML son usadas para construir modelos que permiten descubrir patrones ocultos y tendencias en los datos, permitiendo tomar decisiones basadas en datos.

Comando de terminal básicos [https://builtin.com/software-engineering-perspectives/terminal-commands]

6.2. Categorías de las técnicas del ML

En esta sección nos enfocaremos en:

- Describir las principales categorías del ***machine learning.***
- Entender las diferencias entre el **aprendizaje supervisado** y el no supervisado.

Existen diferentes categorías de las técnicas de ***machine learning*** para diferentes problemas. Las **principales** categorías son las siguientes:

- Clasificación
- Regresión
- Análisis de *clusters*
- Reglas de asociación

En **clasificación**, el objetivo es predecir la categoría del *input* de entrada. Un ejemplo de esto es la predicción del clima, que puede ser soleado, lluvioso, nublado o ventoso. El *input* de entrada, en este caso, podrían ser datos de sensores que miden la temperatura, el nivel de humedad, presión atmosférica, velocidad y dirección del viento, etc. El *target* que se intenta predecir podría ser la **categoría** del tipo de

clima; es decir, soleado, lluvioso, nublado o ventoso.

Figura 3. Predicción de Clima. Obtenida de: *https://scijinks.gov/forecast-reliability/*

Otro ejemplo es clasificar un tumor como benigno o maligno. En este caso, la clasificación es de tipo **binaria,** pues solo hay **dos categorías**. Pero se pueden tener muchas más categorías también, como el problema de predicción de clima **explicado previamente**. Otro ejemplo es identificar dígitos escritos a mano en categorías del 0 al 9. Otros ejemplos pueden ser:

- Determinar cuán **riesgosa** es una persona en términos de aptitud para aplicar a un **préstamo** (riesgo alto, medio o bajo).
- Identificar **sentimientos** en un *tweet* o comentario; como positivos, neutrales o negativos.

Cuando el **modelo** tiene que **predecir** un valor **numérico** en lugar de una categoría, entonces la tarea se convierte en un problema de **regresión**. Un ejemplo de regresión es predecir el precio de un **producto**. El precio de un producto es un **valor numérico** y no una categoría. Por lo tanto, esta es una tarea de **regresión** y no de **clasificación**. Ahora, si se quisiera predecir si el precio del producto va a subir o va a bajar, entonces sería un problema de **clasificación**. Pero si estamos prediciendo el precio actual del producto, entonces es regresión.

Esa es la principal diferencia entre un problema de clasificación y uno de **regresión**. Algunos ejemplos de **regresión**:

- Estimar la demanda de un producto basado en función de la estacionalidad.
- Predecir la nota de un examen.
- Determinar la efectividad de la medicina para un paciente.
- Predecir la cantidad de lluvia.

En el análisis de ***clusters*.** El objetivo es organizar a los ítems en grupos similares. Una aplicación muy común de análisis de ***clusters*** es la segmentación de clientes. Esto implica la separación de los clientes en diferentes grupos o segmentos basados en las características de los clientes. Por ejemplo, sería muy **beneficioso segmentar** a los clientes entre adultos mayores, adultos, adolescentes. Estos grupos tienen diferentes gustos y **afinidades** y, **adicionalmente,** tienen diferentes **comportamientos** de compra. Segmentando tus clientes, en **diferentes grupos,** tú puedes generar campañas de marketing y publicidad de forma más efectiva, para cada uno de los grupos de interés. Al análisis de *clusters* también se lo conoce como ***clustering***.

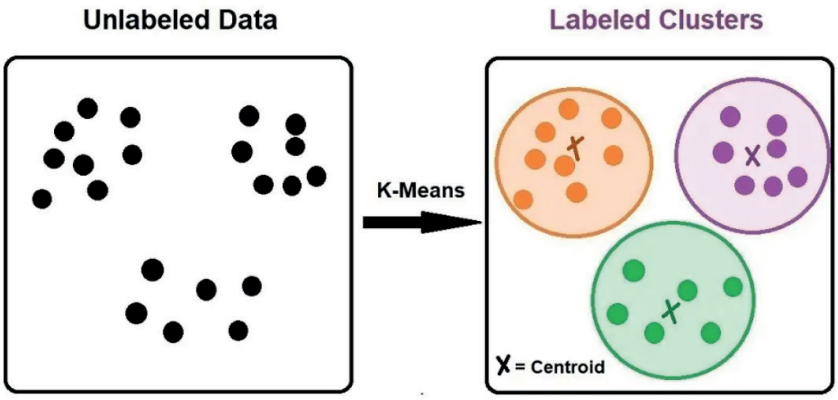


Figura 4. Algoritmos no supervisados.

Algunos otros ejemplos de análisis de *clusters* son:

- Identificar áreas con **topología** similar (desértica, húmeda, etc.).
- Categorizar **imágenes** médicas en diferentes tipos de problemas.
- Determinar diferentes grupos de **patrones** de clima.

Cuando hablamos de **análisis de asociación,** el objetivo es encontrar reglas que permitan capturar **asociaciones** entre ítems o eventos. Las reglas son usadas para determinar cuándo los **ítems o eventos ocurren juntos**. Una aplicación común del análisis de asociación es conocido como ***market basket análisis,*** mismo que es usado para entender el comportamiento de compra del cliente. Por ejemplo, el análisis de asociación puede revelar que los clientes de un banco que han realizado **depósitos** también tienden a estar interesados en invertir en otro tipo de productos bancarios, tales como pólizas de ahorro. Este tipo de **información** puede ser usada para **cross-selling**. Según Datamining Folklore, una cadena de **supermercados** utiliza reglas de asociación para descubrir **conexiones** entre productos que aparentemente no son similares. Ellos descubrieron que muchos de los clientes que van a la tienda los **domingos** en la noche a comprar **pañales** también tienden a comprar **cerveza**; esa información fue usada para colocar cervezas cerca de los pañales y gracias a eso obtuvieron un gran salto en la venta de estos dos productos. Esta es la famosa **conexión** entre **cerveza** y **pañales. Otros ejemplos de análisis de asociación son:**

- Recomendación de **ítems** basados en la historia de **compra** y **búsqueda.**
- Búsqueda de **ítems** que se pueden vender juntos.
- Identificar **páginas web** a las cuales un usuario accede simultáneamente. Es decir, por ejemplo, si la persona accede a ***Mercado*** y también a ***Ali Express,*** para comparar **precios** de **productos**.

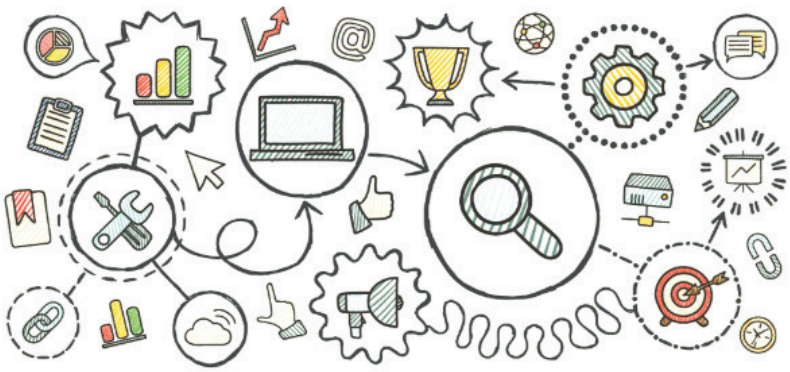


Figura 5. Reglas de asociación. Obtenido de: *https://www.tecnologias-informacion.com/reglasasociacion.html*

Existe otra manera de categorizar las técnicas de **machine learning** con base en el enfoque **supervisado** y **no supervisado**.

En el enfoque supervisado, el valor **objetivo** o ***target,*** que es lo que el modelo va a predecir, es brindado al modelo. Esto se refiere a que los datos con los cuales el modelo va a ser construido tienen que estar **etiquetados**. Retomando el ejemplo de la clasificación del clima –entre soleado, ventoso, lluvioso y nublado– todos los registros en el conjunto de datos están etiquetados con una de las cuatro categorías. En otras palabras, los datos que se encuentran **etiquetados** y predicen la categoría del clima constituyen un problema de aprendizaje supervisado. En general, **clasificación** y **regresión** son los problemas

de aprendizaje supervisado.

En el caso del enfoque no supervisado, el ***target*** que el modelo predice es desconocido o no está disponible. Esto significa que existen datos sin etiquetar. Retomando el ejemplo del *cluster* de segmentación de clientes en grupos diferentes, los ejemplos en los datos no se encuentran etiquetados con el grupo correcto. En lugar de eso, la segmentación se realiza usando una técnica de ***clustering,*** para agrupar **ítems** basados en sus **características** que los clientes tienen en común. Dado que los datos no se encuentran etiquetados, la tarea de agrupar a los clientes en diferentes segmentos sería un problema de aprendizaje no supervisado.

En general, el análisis de *clusters* y análisis de **asociación** son enfoque **no supervisados**.

6.3. Proceso del machine learning

Los objetivos de esta sección son:

- Identificar los pasos del proceso de *machine learning.*
- Entender por qué el proceso del *machine learning* es un proceso iterativo.

El siguiente **diagrama** ilustra las etapas que intervienen en el proceso del *machine learning*. Es importante tener en cuenta que para ejecutar cada una de estas etapas, el propósito tiene que estar claro. Es decir, el problema u oportunidad que está siendo atendido debe ser definido **claramente,** a través de metas y objetivos. Por ejemplo, el propósito de un proyecto podría ser el de estudiar el **comportamiento** de compra de un cliente para poder definir estrategias de marketing más eficientes, con el objetivo de identificar el *revenue* de las ventas. **El propósito el proyecto debe guiar el proceso del *machine learning***

Figura 6.
Proceso de machine learning.

Acquire. Es la primera etapa en el proceso del ***machine learning*** y consiste en obtener todos los datos disponibles relacionados con el problema. En este punto tenemos que identificar todas las fuentes de datos, recolectar los datos y, finalmente, integrar los datos de las **diferentes fuentes**.

Prepare. El segundo paso en el proceso de ***machine learning*** es el de preparación de los datos. Este proceso está dividido en dos partes: exploración y preprocesamiento.

La primera parte de la **preparación de los datos;** es decir, la exploración de los datos, implica una exploración preliminar de los datos, con el objetivo de **entender** la **naturaleza** de los datos con los cuales vamos a trabajar. Algunas de las cosas que tenemos que entender acerca de los datos incluyen sus **características**, **formato y calidad**. Un buen entendimiento de los datos implica generar un análisis más informado y un resultado más exitoso.

La segunda parte de la preparación de los datos es el **preprocesamiento**. Esta parte incluye la limpieza de los datos, la **selección** de las variables por usar y **transformar** los datos para hacerlos más adecuados para el análisis de datos, que corresponde a la siguiente etapa.

Analyze. En esta etapa se seleccionan las técnicas analíticas que se van a usar, se ejecuta la **construcción** de un **modelo** usando los datos y la evaluación de los resultados.

Report. El cuarto paso en el proceso del ***machine learning*** es **comunicar** los resultados. Esto incluye la **evaluación** de los resultados respecto de los objetivos definidos para el proyecto, presentar los resultados en una forma fácil de entender y comunicar los **resultados** a otros.

El último paso es **aplicar** los resultados. Esto nos lleva al propósito del proyecto. ¿Cómo pueden ser usa-

dos los ***insights*** de nuestro análisis para brindar un **marketing** efectivo para incrementar el *revenue* de las ventas? El enfoque principal de la etapa de **acción** es determinar las acciones que se deben realizar en función de los ***insights*** obtenidos*.*

Nótese *que el proceso de* ***machine learning*** *es iterativo. Descubrimientos de una de las etapas podrían requerir que etapas anteriores se repitan con información nueva. Por ejemplo, durante la etapa de preparación podríamos encontrar algunos* ***problemas de calidad,*** *que requieran regresar a la etapa de* ***adquisición,*** *para que esos* ***issues*** *puedan ser solventados.*

Cada una de las etapas podrían requerir varias iteraciones. Por ejemplo, es común aplicar diferentes **técnicas de análisis** en la etapa de análisis, con el **objetivo** de obtener resultados razonables para el modelo. Es importante reconocer que este es un proceso altamente iterativo y no un proceso lineal.

Jupyter Lite [https://jupyter.org/try-jupyter/lab/]

6.4. CRISP – DM

Los **objetivos** de esta sección son los siguientes:

- Resumir en que consiste CRISP –DM.
- Listar las fases de CRISP-DM.
- Describir los **objetivos** de cada fase.

CRISP-DM es un modelo de proceso que describe las etapas en un proceso de *data mining;* seguramente has escuchado que CRISP-DM puede ser entendido como una forma de capturar el proceso de **data science** o el proceso de machine learning también. CRISP–DM es el acrónimo para Cross Industry Standard Process para proyectos de *data mining*. El siguiente diagrama muestra cada una de las fases de **CRISP – DM**.

Figura 7. CRISP -DM.

Las fases son:

- *Business Understanding*
- *Data Understanding*
- *Data Preparation*
- *Modeling*
- *Evaluation*
- *Deployment*

La **primera** fase de CRISP – DM es el ***business understanding***, al igual que en cualquier proyecto, la primera fase es entender por qué estamos haciendo esto. Tenemos que definir el problema del negocio, cuál es el **problema** o la oportunidad para atender, qué es lo que esperamos ganar al trabajar en este **proyecto**. Luego, tenemos que evaluar la situación, cuáles son los **recursos disponibles**, qué riesgos existen, que **costos** hay y qué **beneficios** podríamos tener.

También implica definir las **metas**, **objetivos** y definir el **criterio** de éxito. Qué es lo que esperamos lograr al finalizar este **proyecto**. Al final de esta etapa, deberíamos **entender** el problema que queremos afrontar. Una vez que sepamos la **pregunta** que queremos responder, podemos definir un plan para que podemos **responderla;** en otras palabras, brinda la solución al problema.

La **segunda** etapa del CRISP –DM es el ***data understanding*.** Esta fase tiene dos partes, la **adquisición** de los **datos** y la exploración de los datos. Tal y como suponemos, la adquisición de los datos implica identificar, recolectar e integrar todos los datos relacionados con el problema que debe ser atendido. Mientras que la **exploración** de los datos significa una exploración **preliminar** de los datos con los que se va a trabajar.

La siguiente esta es la ***data preparation***. El objetivo aquí es preparar los datos para modelar. Las actividades en esta fase son: atender *issues* en la calidad de los datos, seleccionar los **feature** a usar y procesar los datos para hacerlos más adecuados para ser usados en la etapa de **modelado**.

La cuarta fase es ***modeling***, que consiste en usar los datos obtenidos en la etapa anterior para crear un modelo de ***machine learning***. Primero tenemos que determinar el tipo de **problema** que tenemos. **¿Es un problema de clasificación o de regresión?** Luego debemos seleccionar la técnica de modelado o algoritmo. Después se crea el modelo usando los datos preparados.

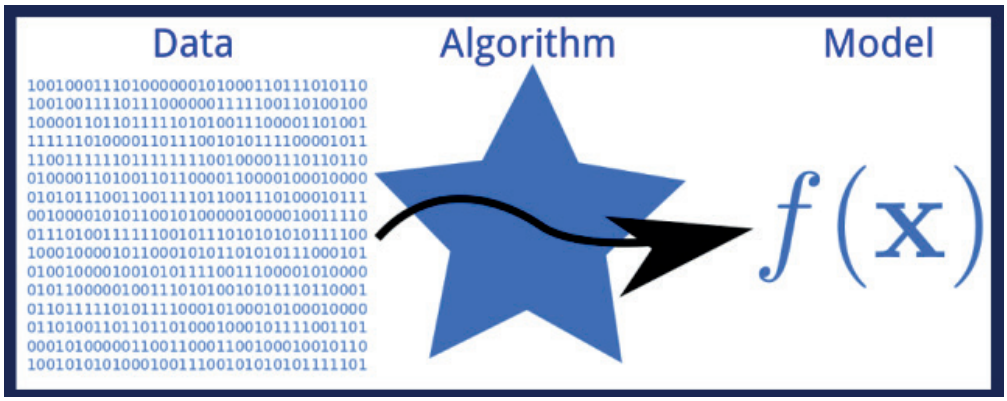


Figura 8. Etapa de Modelado.

En la fase cinco se evalúan los resultados. El ***performance*** del modelo necesita ser evaluado, y esos resultados tienen que ser **evaluados** con el criterio de **éxito** definido en la primera etapa. Al final de esta etapa, se toma la decisión de seguir o no seguir con la siguiente etapa; caso contrario, hay que decidir a qué etapa retrocedemos: a la etapa de limpieza de datos, a la parte de entendimiento de los datos o si incluso necesitamos recolectar más datos.

Figura 9. Decisiones sí o no.

La fase final en el proceso de **CRISP-DM** es el ***deployment***, dependiendo de los requerimientos del proyecto, la fase de ***deployment*** puede ser simplemente crear un reporte final o puede ser algo más complejo, como implementar el modelo e integrarlo en un **proceso** ya **existente**.

En esta etapa, necesitamos o generar un **reporte** final o presentación para **documentar** y **comunicar** los descubrimientos del proyecto a otros. Esto suele incluir el performance del **modelo** y la respuesta de los objetivos del negocio que fueron definimos para el proyecto. El *deployment* del modelo significa integrarlo a una **aplicación** o proceso. Crear un plan para monitorear el modelo. Esto es necesario para evaluar el impacto del modelo en nuestro **proceso** o **aplicación**.

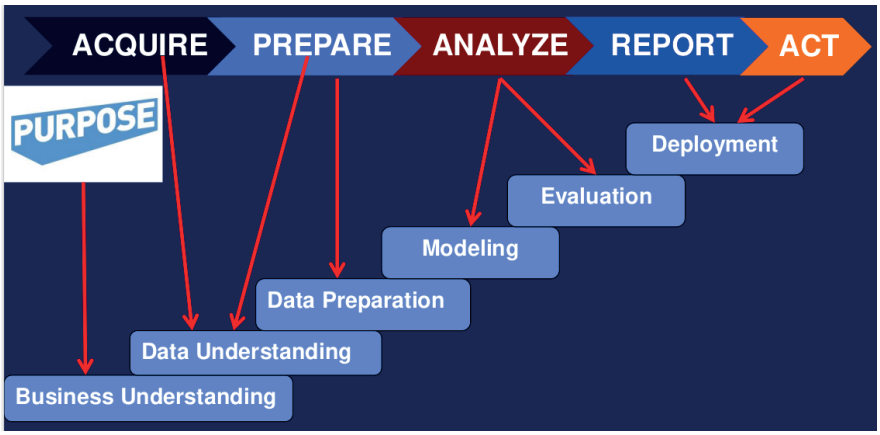


Figura 10. Proceso de machine learning vs. CRISP-DM.

Como habrás notado, hay muchas similitudes entre el proceso de ***machine learning*** y la metodología **CRISP – DM**. El diagrama anterior muestra cómo los pasos del proceso de ***machine learning*** mache-an con las fases del CRISP–DM. Los objetivos y actividades son similares, pero organizados de forma diferente. El proceso pone más énfasis en resultados accionables, mediante la generación explícita de reportes, para resultar la importancia de convertir los *insights* en acciones en un proyecto de ***machine learning.***

Referencias citadas en la Clase 6

- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. Springer.
- Health Data Miner. (s.f.). *CRISP-DM: Una metodología para minería de datos en salud*. Health Data Miner. Recuperado el 15 de marzo de 2025, de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Definición de los términos citados en la Clase 6

CRISP-DM.

Cross-Industry Standard Process for Data Mining(CRISP-DM) es una metodología ampliamente utilizada en el campo de la minería de datos y el análisis de **Big Data**, que proporciona un enfoque estructurado y flexible para llevar a cabo proyectos de análisis de datos. Fue desarrollado por un consorcio de empresas con el objetivo de estandarizar el proceso de minería de datos, independientemente de la industria en la que se aplique. **CRISP-DM** consta de seis fases interdependientes: **comprensión** del **negocio**, **comprensión** de los **datos**, **preparación** de los **datos**, **modelado**, **evaluación** e **implementación**.

Cada una de estas etapas se enfoca en diferentes aspectos del proyecto, desde la definición de los objetivos empresariales hasta la evaluación y despliegue de los modelos desarrollados, garantizando que los resultados sean útiles, aplicables y alineados con los objetivos iniciales del proyecto (Chapman et al., 2000). Este enfoque es altamente adaptable, lo que lo convierte en una herramienta clave para abordar la complejidad y los desafíos que presenta el trabajo con grandes volúmenes de datos en **diversos contextos**.

Data-driven.
Es un enfoque en el que las decisiones, estrategias y acciones se basan en el análisis y la interpretación de datos concretos y objetivos, en lugar de depender únicamente de la intuición o la experiencia subjetiva. Este enfoque busca obtener información valiosa a partir de grandes cantidades de datos, como estadísticas, patrones y tendencias, para optimizar procesos, mejorar la toma de decisiones y obtener resultados más precisos y eficaces en diversos campos, como negocios, marketing, ciencia e inteligencia artificial.

Profundización Clase 6

En esta infografía se pueden observar ejemplos de las categorías de las técnicas de ***machine learning.***

✅ **Supervisado** → Predicción de precios, diagnóstico médico.
✅ **No Supervisado** → Segmentación de clientes, detección de anomalías.
✅ **Semi-Supervisado** → Procesamiento de imágenes, reconocimiento de texto.
✅ **Refuerzo** → Juegos, robótica, optimización de procesos.

Infografía sobre las etapas de **CRISP–DM.**

1️⃣ Business Understanding (Comprensión del Negocio)

📌 **Objetivo:** Definir el problema y los objetivos del proyecto.
🔹 **Ejemplo:** Un banco quiere reducir la tasa de clientes que cancelan sus tarjetas de crédito.

2️⃣ Data Understanding (Comprensión de los Datos)

📌 **Objetivo:** Recopilar y explorar los datos disponibles.
🔹 **Ejemplo:** Se analizan transacciones, historial de pagos y quejas de clientes.

3️⃣ Data Preparation (Preparación de los Datos)

📌 **Objetivo:** Limpiar y transformar los datos para su análisis.
🔹 **Ejemplo:** Se eliminan valores nulos, se convierten fechas en variables numéricas y se categorizan ingresos.

4️⃣ Modeling (Modelado)

📌 **Objetivo:** Aplicar algoritmos para encontrar patrones.
🔹 **Ejemplo:** Se entrena un modelo de Machine Learning (Regresión Logística) para predecir clientes que podrían cancelar sus tarjetas.

5️⃣ Evaluation (Evaluación)

📌 **Objetivo:** Medir la precisión y utilidad del modelo.
🔹 **Ejemplo:** Se usa la métrica de precisión y recall para ver qué tan bien el modelo predice el abandono.

6️⃣ Deployment (Despliegue)

📌 **Objetivo:** Implementar el modelo en producción para su uso real.

🔹 **Ejemplo:** Se integra el modelo en un CRM para alertar al equipo de retención sobre clientes en riesgo.

PUCE Virtual

La excelencia no se improvisa

síguenos