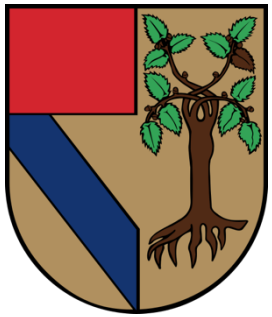


**UNIVERSIDAD PANAMERICANA**

# APRENDIZAJE DE MÁQUINA

COM194


## Enrique González N.

Profesor

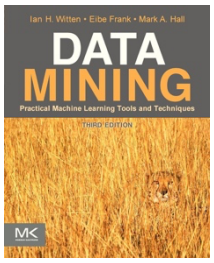


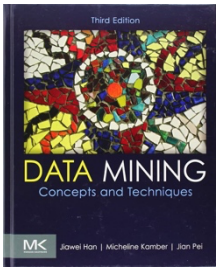
COM194-AM-V-24

## MATERIAL DE CONSULTA

Introducción al CRISP-DM
https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

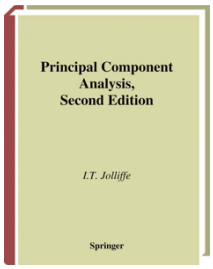
Data Mining
Ian H. Witten, Eibe Frank, Mark A. Hall



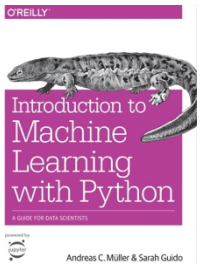

Data Mining: Concepts and Techniques
Jiawei Han, Micheline Kamber, Jian Pei

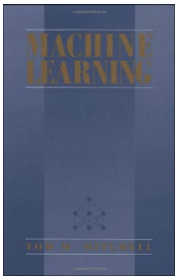
Fundamentos de bases de datos
Martha Elena Millan


Principal Component Analysis
I. T. Jolliffe
Springer Series in Statistics

Introduction to Machine Learning with Python
Andreas C. Müller & Sarah Guido




Machine Learning
Tom Mitchell

## DATA SCIENCE, DATA MINING Y MACHINE LEARNING

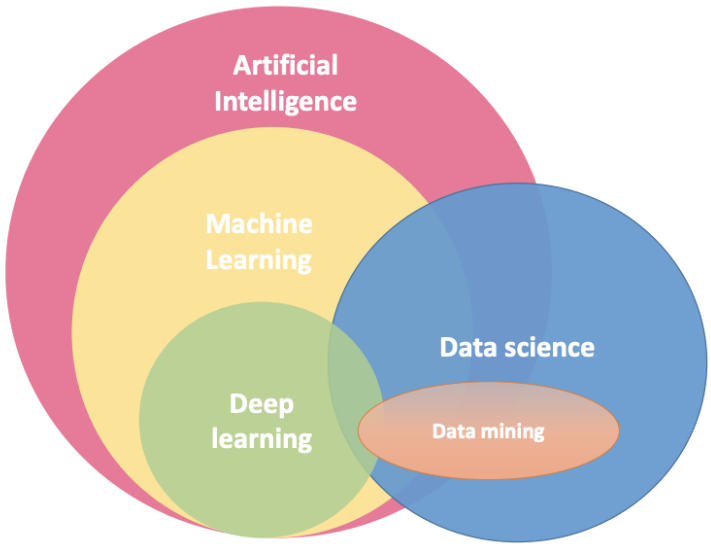
Machine Learning, Data Science y Data Mining están intrínsecamente conectados y se complementan entre sí en el proceso de trabajar con datos para obtener información valiosa y tomar decisiones informadas.

**Machine Learning**:
Es la rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender a partir de datos. Su objetivo principal es hacer que las máquinas sean capaces de realizar tareas específicas sin ser programadas explícitamente, sino aprendiendo de patrones y experiencias.

**Data Mining**:
Es el proceso de descubrir patrones y conocimientos significativos a partir de grandes conjuntos de datos. Implica la aplicación de técnicas estadísticas, matemáticas y algorítmicas para explorar y analizar datos, con el objetivo de identificar relaciones y patrones que puedan ser útiles para la toma de decisiones.

**Data Science**:
Campo interdisciplinario que combina habilidades de programación, estadísticas y dominio del tema para analizar y comprender datos complejos. Incluye la recolección, limpieza, exploración y análisis de datos, utilizando diversas herramientas y técnicas. ML y Data Mining son componentes cruciales en el conjunto de habilidades de data science.

- Son términos usados usualmente de forma indistinta (inapropiadamente).
- Son *subdisciplinas* dentro de ciencias computacionales.
- El *enfoque común* entre ellas es:
    - Mejorar la toma de a través del análisis de datos.



RELACIÓN INTERDISCIPLINAR ENTRE ÁREAS

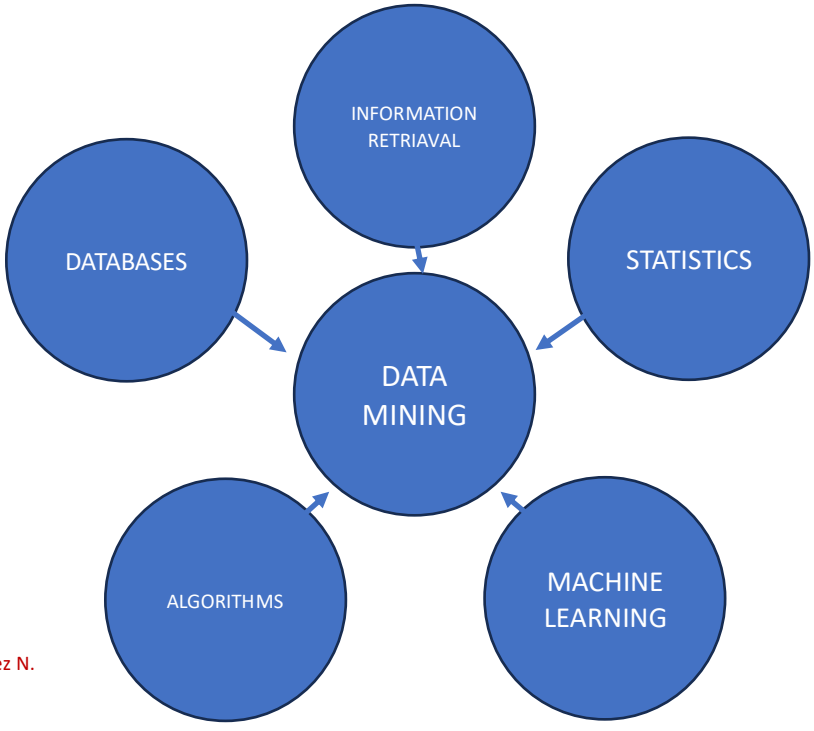
- Existe una relación multidisciplinar:
  - Hay varias áreas y técnicas involucradas.
  - Principalmente: estadística y c. computacionales.

**Data Mining y Data Science:**
Data Mining proporciona técnicas clave para la exploración y análisis de datos, lo cual es fundamental en el proceso de Ciencia de Datos. La minería de datos ayuda a identificar patrones, correlaciones y tendencias que son esenciales para la toma de decisiones informadas.

**Machine Learning en Data Science:**
Machine Learning es una parte integral de la Ciencia de Datos. Los algoritmos de ML se utilizan para construir modelos predictivos, clustering, y clasificación en diversos problemas. La Ciencia de Datos emplea herramientas de Machine Learning para realizar análisis avanzados y extraer conocimientos significativos de los datos.

**Machine Learning y Data Mining**:
ML a menudo se utiliza como una herramienta en el proceso de Data Mining para construir modelos predictivos y clasificadores. Los algoritmos de ML pueden ayudar a descubrir patrones y relaciones complejas en los datos que son difíciles de identificar mediante enfoques tradicionales.

- Data Mining (Minería de datos):
  - Es la extracción de patrones e información mediante el *uso de algoritmos*.
  - Desde una perspectiva de *Data Mining*:



## APLICACIONES EN LAS ORGANIZACIONES


El aprovechamiento de estas *herramientas interdisciplinares* (data mining y ML), con el objetivo de tomar decisiones tiene *distintas aplicaciones*, ejemplos:

- Predicción y Análisis.
- Toma de decisiones.
- Segmentación y Personalización.

Ejemplo

PREDICCIÓN DE TENDENCIAS DE VENTAS

**Descripción**:
Una empresa ML para analizar datos históricos de ventas, factores estacionales y variables externas como eventos especiales o cambios económicos. Esto permite predecir las tendencias de ventas futuras con mayor precisión.

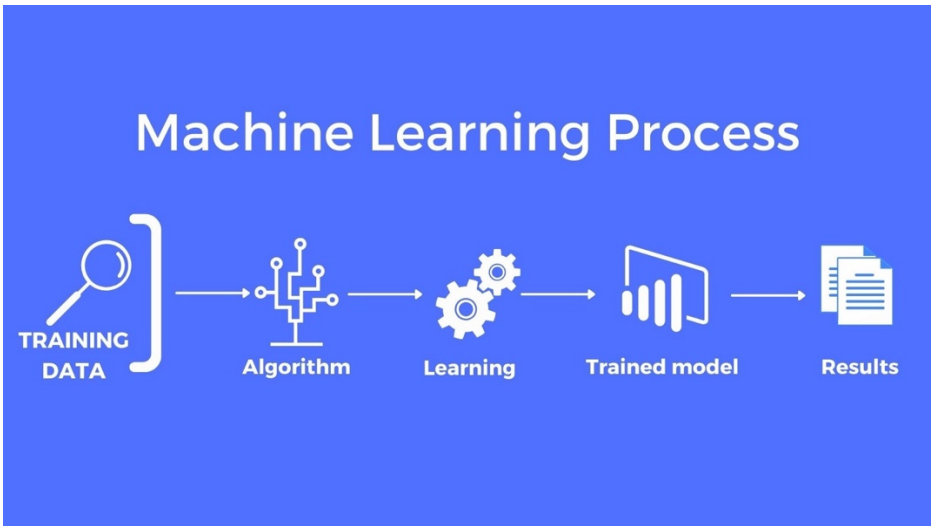
**Resultado:**
La empresa puede ajustar su inventario, gestionar recursos y lanzar estrategias promocionales de manera más efectiva, maximizando las ventas y minimizando los excedentes de inventario.

**Caso Real - Amazon y el Uso de Algoritmos Predictivos:**
Amazon emplea machine learning para *predecir las preferencias de los clientes* y *anticipar las tendencias de compra*. Su algoritmo analiza el historial de compras, *patrones de búsqueda* y *comportamientos de navegación* para *personalizar* las *recomendaciones* de productos. Esto ha contribuido significativamente al éxito de Amazon al *mejorar la experiencia del cliente* y *aumentar las ventas*.

## ¿QUÉ ES MACHINE LEARNING (ML)?

- ML is regularly defined as the field concerned with *giving/teaching computers* the *aptitude to learn* or achieve a specific task *without using explicit instructions* or being explicitly programmed.
- In other words, an *algorithm* that permits a computer to *learn from experience*.

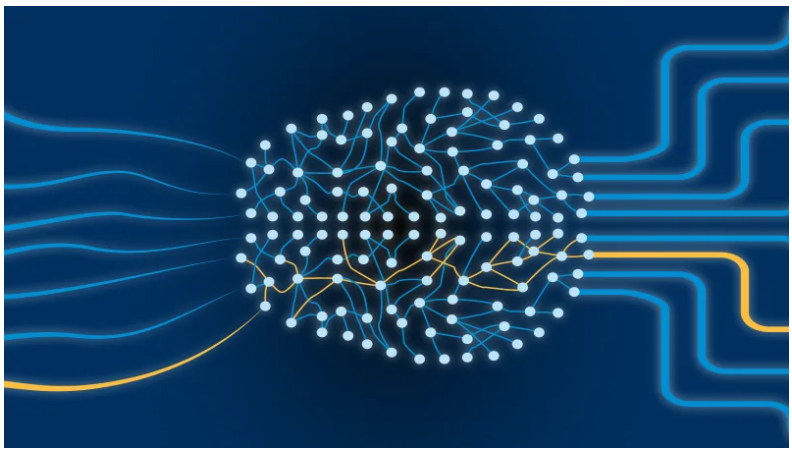
- In this sense, the algorithms are *collections of instructions* to be followed to perform a calculation or to *solve a problem*.




## SUBDIVISIONES DE ML

Machine learning se clasifica principalmente en *3 subdivisiones*:

- Aprendizaje supervisado (Supervised learning).
- Aprendizaje no supervisado (Unsupervised learning).
- Aprendizaje Reforzado (reinforcement learning).



## TIPOS DE APRENDIZAJE

**Supervisado**:
Algorithms are models employed when the available data, inputs (sets of training examples), and the desired output, have been previously classified or characterized; therefore, it is said that the information has been labeled.

**No Supervisado:**
Algorithms are models where only the input data are available, and the information has not been categorized (is unlabeled): these techniques are commonly implemented with cluster analysis.

**Reforzado**:
The output of the system is a sequence of the correct actions to reach the goal, this sequence is commonly known as a policy. Thus, the objective of these ML techniques is to generate a good policy by learning from past good action sequences.

## EJERCICIO

**Objetivo**: reforzar habilidades de programación en Python.

**Lineamientos generales**:

- Esta es una actividad de clase no evaluada.
- Trabajar de preferencia en forma individual.

**Instrucciones**:
Descarga el archivo correspondiente del siguiente link:
https://datos.covid-19.conacyt.mx/
Abrir en un navegador Google Colab:
https://colab.research.google.com/

Practica lo siguiente:

- Leer archivos (csv, Excel) y cargar los datos en pandas:
  - read_csv()
  - read_excel()
- Limpiar los datos:
  - fillna()
  - dropna()
- Normalización y escalamiento de datos.
- Reducción de dimensión.

- PCA – Scikit -learn
- Graficar los datos.

## METODOLOGÍA CRISP-DM

- CRISP-DM son las siglas en inglés de:

Cross-Industry Standard Process for Data Mining

- Concebido en la década de los 90's.
- Desarrollado por un consorcio de empresas:
  - Daimler AG, SPSS, NCR, entre otras.
- Fue dirigido por SIGKDD:

SIGKDD • Special Interest Group on Knowledge Discovery in Data

https://www.kdd.org/

https://dl.acm.org/sig/sigkdd

- El modelo fue publicado en 1999.

## MODELO

- Modelo de procesos jerárquico.
- Dividido en niveles de abstracción.
- Por cada etapa o fase se propone un conjunto de tareas genéricas asociadas.

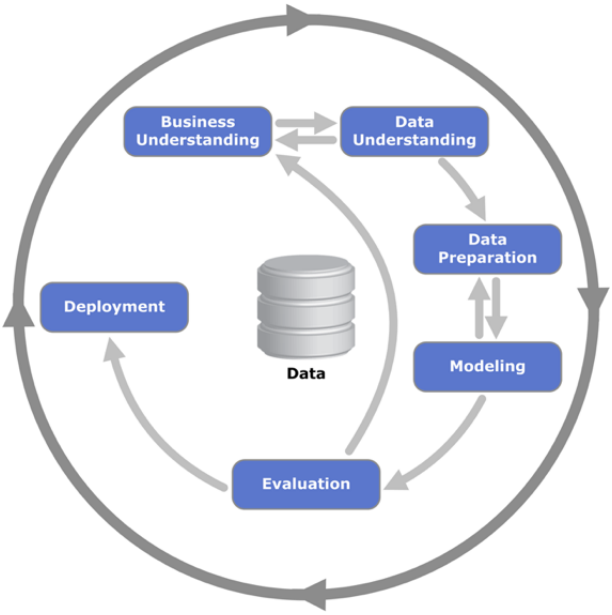



Diagrama del proceso CRIPS-DM

**Entendimiento del negocio**
(*Business Understanding*)

Está centrada en:

- Entender los objetivos del proyecto de descubrimiento de conocimiento (KDD).
- En los requerimientos desde el punto de vista del negocio.

A partir de este conocimiento se puede:

- Definir el problema de minería
- proponer un plan para cumplir con los objetivos propuestos.

**Consideraciones:**
- Comprensión comercial.
- Determinación de objetivos.
- Evaluación de la situación.
- Determinación de los objetivos de la minería de datos.
- Producción de un plan de proyecto.

**Entendimiento de los datos**
(*Data Understanding*)

Empieza con un conjunto inicial de datos para:

- Tratar de entenderlos.
- Identificar problemas de calidad de los datos.
- Detectar subconjuntos de estos que permitan emitir algunas hipótesis sobre información oculta.

**Consideraciones:**
- Recopilación.
- Descripción.
- Exploración.
- Verificación de calidad.

**Preparación de los datos**
(*Data Preparation*)

Se seleccionan:

- Atributos

- Registros
- Tablas

Se limpian y transforman los datos.
Todas estas tareas tienen como propósito:

- Construir el conjunto de datos que serán la entrada de las herramientas de modelamiento.

**Consideraciones:**
- Selección de datos.
- Limpieza de datos.
- Construcción de nuevos datos.
- Integración de datos.
- Formato de datos.

**Modelación**
(*Data Modeling*)

En esta etapa se:

*Normalización y/o escalamiento

- Seleccionan técnicas de modelación.
- Se *ajustan los parámetros* para las técnicas seleccionadas:
    - Es posible que de nuevo se deba llevar a cabo tareas de preparación de datos, en aquellos casos en los cuales éstos se deban transformar para ajustarse a la especificación de la entrada de las técnicas

**Consideraciones:**
- Selección de técnicas de modelado.
- Generación de un diseño de comprobación.
- Generación de los modelos.
- Evaluación del modelo.

**Evaluación e interpretación de resultados**
(*Evaluation*)

- Ya se cuenta con el(los) modelo(s) construido(s) que tengan la más alta calidad desde la perspectiva del *análisis de datos*.
- Se requiere evaluar si algún aspecto importante no se ha tenido en cuenta o no se le ha dado la importancia necesaria.

¿falta algo?

**Consideraciones:**
- Evaluación de los resultados.
- Proceso de revisión.

**Aplicación**
(*Deployment*)

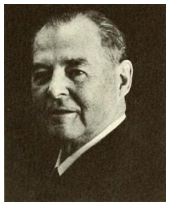
- El conocimiento obtenido se *organiza* y *presenta* al usuario en una forma en que éste *lo pueda usar*.
- Los resultados de esta fase pueden ser (dependiendo de los requerimientos):
  - **Simples**: un reporte.
  - **Complejos**: se implementa un proceso de minería a través de la organización.

**Consideraciones:**
- Planificación de despliegue.
- Planificación del control y del mantenimiento.
- Creación de un informe final.

## PRINCIPAL COMPONENT ANALYSIS (PCA)

*Historically, PCA was first formulated in a statistical setting to estimate the principal components of a multivariate random variable x (Pearson 1901; Hotelling 1933*).

- Una de las técnicas más “viejas” en análisis multivariable.
- Introducida por Karl Pearson (1901) – University College London (departamento de estadística).
- Desarrollada por Harold Hotelling (1933) – Standford University (departamento de matemáticas).
- El método es ampliamente usado en:
  - Finanzas, estadística, computación.

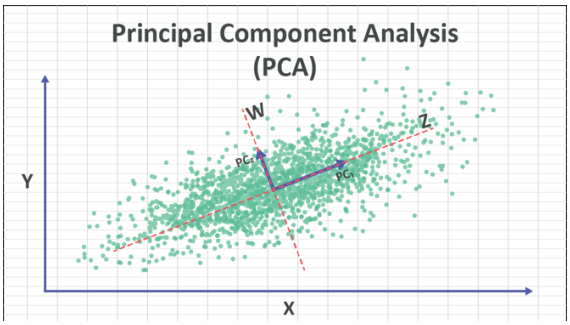
K. Pearson   H. Hotelling

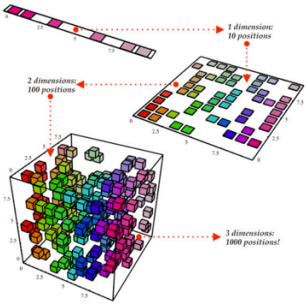
- Es una técnica ampliamente utilizada en el campo del análisis de datos y machine learning.

- La idea central es reducir la dimensionalidad de un conjunto de variables o data set (grande):
  - Compuesto de variables interrelacionadas.
  - Se transforma en un conjunto de variables no correlacionadas llamadas Componentes Principales (PCs).

- En general el cálculo se realiza mediante los siguientes pasos:
  - La normalización de los datos para que cada variable tenga media 0 y varianza 1.
  - Cálculo de la matriz de covarianzas para entender la varianza de las variables de manera conjunta.
  - Cálculo de los vectores propios (eigenvectores) de la matriz covarianza para determinar la dirección de los PCs.
  - Cálculo de los valores propios (eigenvalores) de la matriz covarianza para determinar la magnitud o importancia de cada componente.
  - Selección de los PCs que capturan laa mayor parte de la variación de los datos.
  - Transoformación de los datos originales a este nuevo espacio de PCs.

- Se presentan de manera ordena:
  - Las primeras presentan (describen) la mayor variación contenida en el set original.




- Ventajas de PCA:
  - Reduce el número de variables en el conjunto de datos, simplificando así los modelos y el tiempo de procesamiento.

  - Elimina colinealidad al transformar las variables originales en componentes no correlacionadas.
  - Permite visualizar datos de alta dimensionalidad en 2D o en 3D.
  - Eliminación de ruido y redundancia en los modelos de machine learning.

- Desventajas de implementar PCA:
  - Pérdida de interpretabilidad dado que las nuevas PCs son combinaciones lineales de las originales.
  - Pérdida potencial de información relevante al reducir la dimensiconalidad.
  - PCA asume que las relaciones entre las variables son lineales (lo que puede no ser el caso).
  - No maneja datos categóricos.

Ejemplo:
https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html#sphx-glr-auto-examples-decomposition-plot-pca-iris-py