1. Introducción

Definición 1.1. (Machine Learning) El aprendizaje automático es una parte de la inteligencia artificial y el subcampo de la ciencia de datos. Es una tecnología en crecimiento que permite que las máquinas aprendan de datos anteriores y realicen una tarea determinada automáticamente.

Machine Leaning permite que las computadoras aprendan de las experiencias pasadas por sí mismas, utiliza métodos estadísticos para mejorar el rendimiento y predecir la salida sin ser programado explícitamente. [6]

Definición 1.2. (Data Science) Data Science un campo de estudio profundo de los datos que incluye extraer información útil de los datos y procesar esa información utilizando diferentes herramientas, modelos estadísticos y algoritmos de aprendizaje automático

Es un campo interdisciplinario que utiliza técnicas de análisis de datos, estadística y aprendizaje automático para extraer conocimientos y crear modelos predictivos a partir de grandes conjuntos de datos.

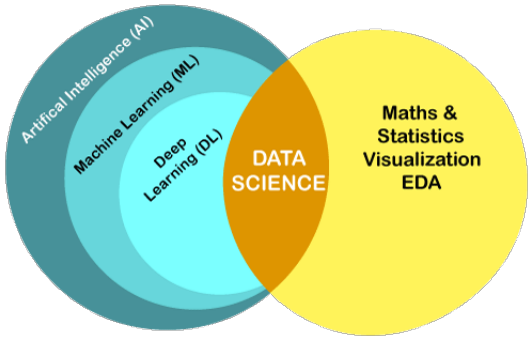

Figura 1: Data Science y Machine Learning

1.1. Variables

En el contexto del aprendizaje supervisado, se trabaja con *variables independientes* (también conocidas como características, predictores o variables de entrada) y *variables dependientes* (también llamadas variables objetivo o target o variables de salida). Las variables independientes son las que se utilizan para hacer predicciones o estimaciones, mientras que las variables dependientes son los resultados que se intentan predecir o modelar. Por ejemplo, en un problema de predicción de precios de viviendas, las características como el número de habitaciones, la ubicación y el tamaño de la propiedad serían las variables independientes, mientras que el precio de venta sería la variable dependiente.

- Variables Independientes (entradas)

- Cualitativas
  - Texto
    - Nominales (categorías, Por ejemplo: países, sexo)
    - Ordinales (poco, mucho, muchísimo, Por ejemplo: nivel de tabaquismos)
  - Números
    - Nominales
    - Ordinales
- Cuantitativas: cuando hablamos de cantidad
  - Discretas: Por ejemplo año, mes, edad, etc.
  - Continuas: Por ejemplo altura, peso, etc.

- Variables Dependientes (salidas, categorías)

Definición 1.3. (variables cualitativas) Son aquellas que describen características o cualidades y no pueden ser medidas en términos numéricos. Las variables **cuantitativas**, por otro lado, son aquellas que pueden ser medidas en términos numéricos y tienen valores numéricos.

Definición 1.4. variables nominales Son aquellas que no tienen un orden natural, como el género o el color de ojos. Las **variables ordinales** son aquellas que tienen un orden natural, como el nivel de educación (primaria, secundaria, universidad), nivel de tabaquismos: Clasificamos como leve (1), moderado (2), nivel medio (3), importante (4) y muy importante (5).

Variables y tipos de problemas (video 02a min 9:00)

1. Si la variable dependiente es **cualitativa**, el tipo de problema es de **clasificación**.
2. Si la variable dependiente es **cuantitativa**, el problema es de **regresión**. Por ejemplo si quiero predecir el precio de una propiedad.
3. Si **NO hay variable** dependientes, el problema es agrupamiento.

- **Outliers:** Valores atípicos, pueden ser errores o un dato que se sale de la norma.
- **Correlación:**
  - **Positiva:** Cuando una variable aumenta la otra también.
  - **Negativa:** Cuando una variable aumenta la otra disminuye.
  - **Sin correlación:** Cuando una variable aumenta la otra no cambia.
- **Varianza:** Es la medida de dispersión de una variable respecto a su media. Si la varianza es alta, los datos están muy dispersos, mientras que si la varianza es baja, los datos están muy agrupados.

- **Covarianza:** Es una medida de la relación lineal entre dos variables aleatorias. Indica cómo varían conjuntamente dos variables aleatorias respecto a sus medias. Si la covarianza es positiva, las variables aumentan o disminuyen conjuntamente, mientras que si la covarianza es negativa, una variable aumenta mientras la otra disminuye.

Figura 2: Metodología de Machine Learning.

1.2. Formatos de datos y herramientas

Ver Video clase.

1. **CSV:** Comma Separated Values. Es un formato de texto plano que se utiliza para almacenar datos tabulares. Cada registro se almacena en una línea y los campos se separan por comas.

2. **JSON:** JavaScript Object Notation. Es un formato de texto plano que se utiliza para almacenar datos estructurados. Se utiliza principalmente para transmitir datos entre un servidor y una aplicación web.

3. **CSR:** Compressed Sparse Row. Es un formato de matriz dispersa (con gran cantidad de ceros) que se utiliza para almacenar matrices dispersas. Se utiliza principalmente para almacenar matrices dispersas en memoria.