

Introducción al Machine Learning

Reglas generales del curso

- Las clases son presenciales, en caso de tener que faltar por algún motivo comunicarlo mediante un email a la siguiente dirección: jmgutier@ulima.edu.pe
- Asesorías del curso, pueden ser virtuales o presenciales. Se recomienda que se converse antes para ver un horario adecuado.
- Tratar de revisar su límite en inasistencias a fin de no ser considerado como impedido.
- Los reclamos hacia alguna evaluación se harán mediante el llenado de un formulario. Las decisiones de dicha revisión son inapelables y se vuelve a revisar todo el examen.

Evaluaciones

Forma de evaluación:

- a) Un (1) examen escrito a ser desarrollado por BB.
- b) Dos (2) prácticas de laboratorio, consiste en utilizar un dataset, programar un modelo e interpretarlo.
- c) Un proyecto, el cual deberá incluir algún tipo de reporte. Sugiero hacerlo en formato artículo.

Como lenguaje de programación usaremos Python y librerías disponibles.

Evaluaciones

Formas de desaprobar el curso:

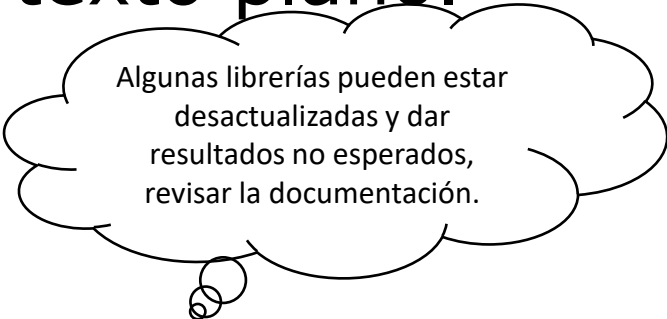
- a) Quedar impedido.
- b) Presentar sólo su código, quizás utilizando alguna herramienta LLM o algún repositorio de github.
- c) Presentar sus modelos como "cajas negras".

Formas de aprobar el curso:

- a) Tratar de no quedar impedido.
- b) Debe demostrar que sus decisiones son adecuadas y respaldarlas, cuando se solicite, con referencias académicas que sustenten cada elección tomada.
- c) Debe de conocer la teoría detrás de los modelos.

Herramientas para el curso

- 1) Notepad ++ o cualquier editor de texto plano.
- 2) Google Colab.
- 3) Anaconda.
- 4) R con R Studio.
- 5) Alguna herramienta de generación de código como ChatGPT, CoPilot o Perplexity.



Algunas librerías pueden estar desactualizadas y dar resultados no esperados, revisar la documentación.

Usted puede utilizar la combinación que mejor prefiera. Para las clases estará instalado el Anaconda y el Colab puede hacerse una cuenta con un correo Gmail.

Campos de aplicabilidad

Considerando las siguientes áreas, coloque un ejemplo en las cuales se podrían aplicar técnicas de IA y sus variantes:

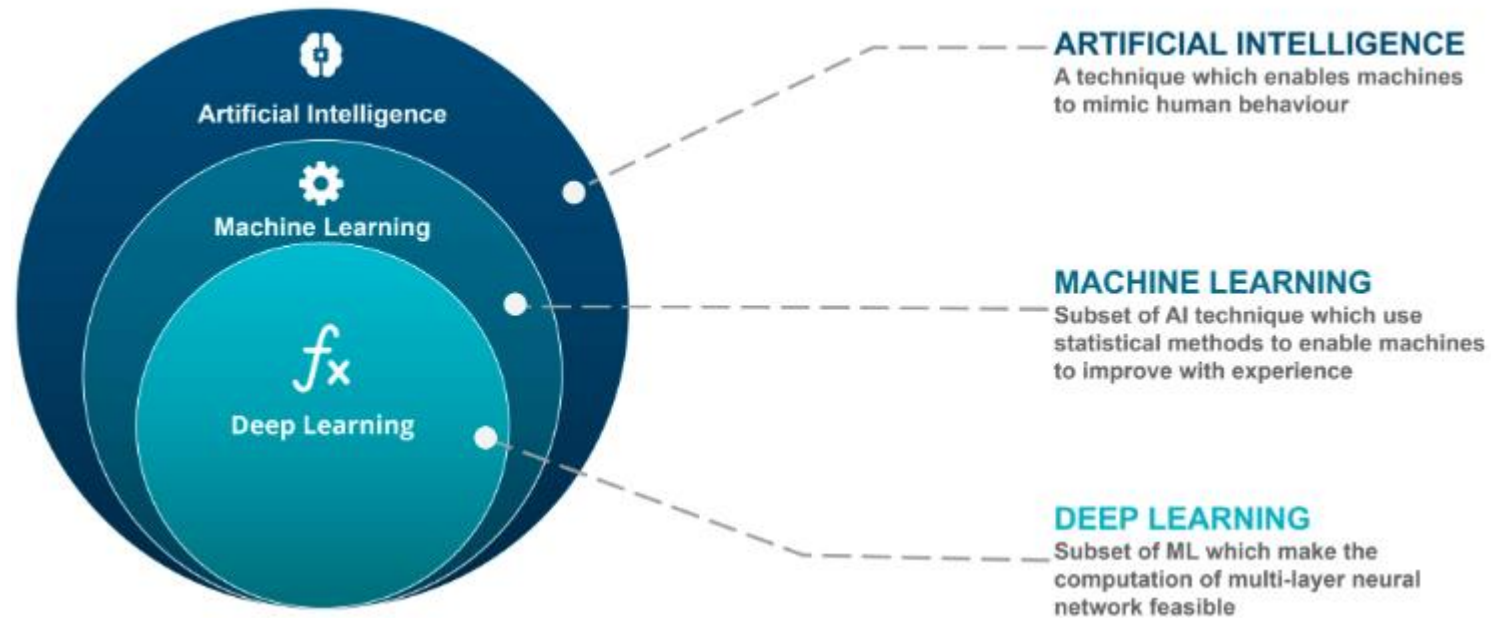
- a) Biología.
- b) Medicina.
- c) Educación.
- d) Derecho.
- e) Historia y Arqueología.

Machine Learning

- Definición:

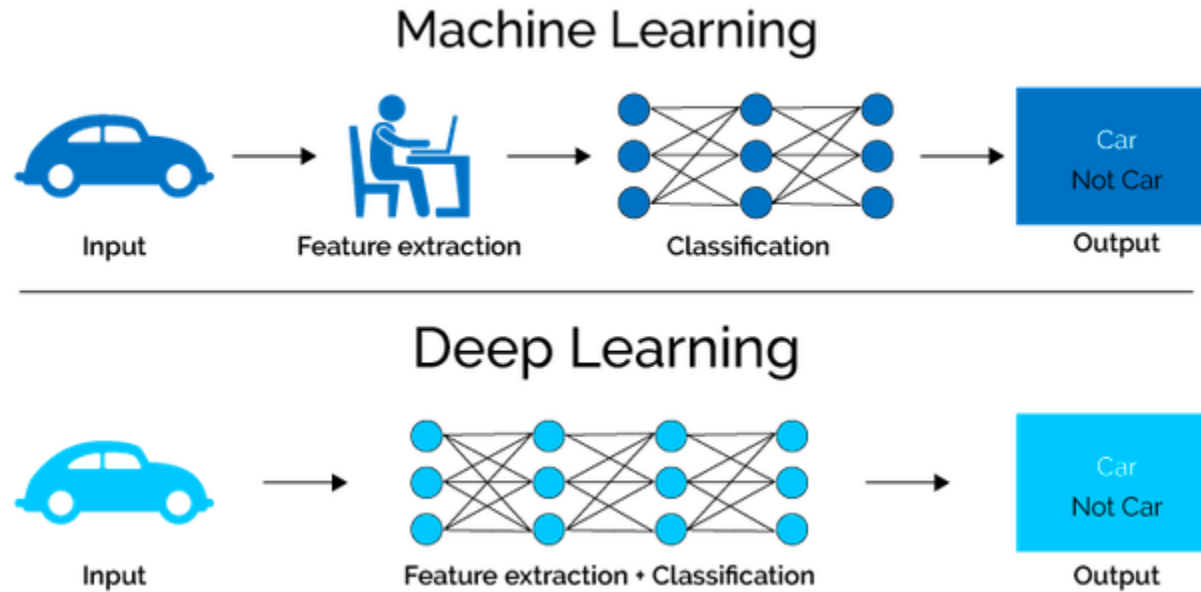
Según Michell (1997) se manifiesta que: “Un programa de computador se dice que aprende la experiencia E con respecto a alguna tarea T y con una medida de rendimiento P , si su rendimiento con respecto a tareas dadas en T , medidas en P , mejora con la experiencia E ”.

Diferencias IA, ML, DL



(Fuente: <https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/>)

Diferencias IA, ML, DL



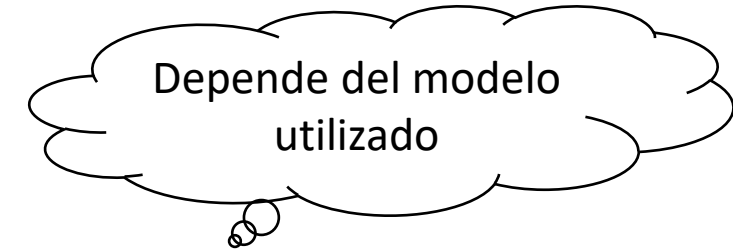
(Fuente: <https://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning>)

Áreas relacionadas

- a) Inteligencia Artificial
- b) Redes neuronales artificiales, área que surgió en un principio al tratar de simular las conexiones cerebrales mediante un computador. Período de pausa después de su creación debido a que no se contaba con el suficiente poder computacional.
- c) Deep Learning
- d) Reconocimiento de patrones, consiste en reconocer objetos, tendencias y características comunes entre un conjunto de objetos; viene a ser uno de los principios del Data Mining.
- e) Algoritmos Genéticos.
- f) Natural Language Processing.

Limitantes

- Cantidad y calidad de los datos: Datos incompletos, datos no balanceados.



- No existe relación entre las variables predictoras y la predicción (predictors y target).
- Sistemas con alta entropía.
- Se debe de conocer el entorno en el cual se va a desarrollar el modelo.

Tipos de Aprendizaje

Aprendizaje Supervisado: Se tiene un etiquetado de los datos a fin de conocer cuál es el “resultado” de los mismos.

Ejemplos:

- Regresión
- Redes Neuronales
- Árboles de Decisión
- Support Vector Machines

Tipos de Aprendizaje

Aprendizaje no supervisado: El etiquetado no existe en los datos de salida para un entrenamiento.

- Clusterización
- Reducción de la Dimensionalidad
- Hidden Markov Models

Aprendizaje semi supervisado: Se etiquetan sólo algunos conjunto de datos.

Tipos de Datos

1) Numéricos

1.1) Nominal

1.2) Ordinales

2) Categóricos

2.1) Intervalo

2.2) Ratio

Según el tipo de valores:

a) Discretos.

b) Continuos.

Manipulación de Datos

One-hot encoding

Domingo	0	0	0
Lunes	1	0	0
Martes	0	1	0
Miércoles	0	0	1

Abarca $n-1$ formas posibles de codificar.

No es necesario codificar todos los datos, un dato puede estar compuesto de valores nulos.

Manipulación de Datos

Scaling:

$$z = \frac{x - \bar{x}}{s}$$

Normalización: Convierte las variables dependientes en una forma normal.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

Caso: Calidad del vino

Se usará el dataset de calidad del vino para hacer un breve análisis exploratorio de los datos.

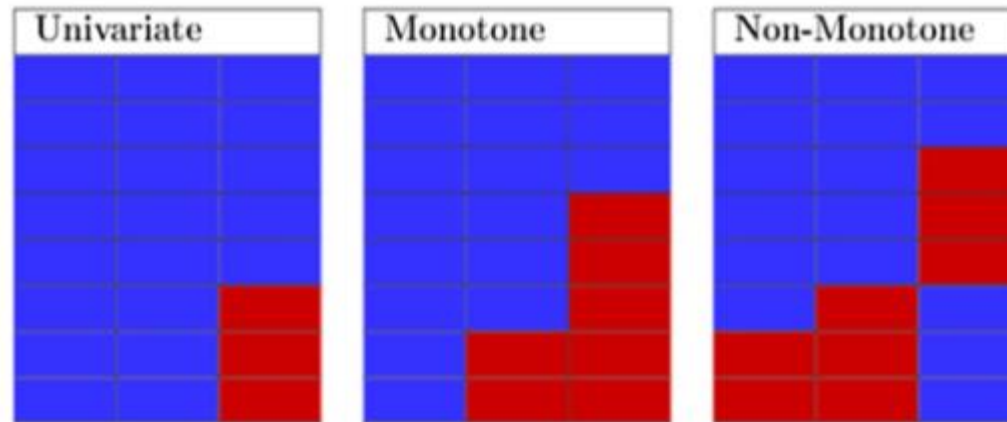
Examinar sus atributos del dataset.

Escalar sus variables y observar sus resultados.

Imputación de Datos

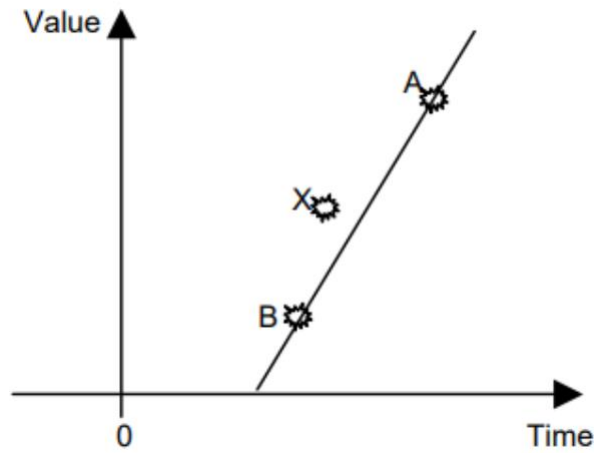
- Se refiere al llenado de datos faltantes, formas de hacerlo:
 - a) Borrar los datos faltantes (no recomendable).
 - b) Obtener el promedio (mean) o la moda de los datos faltantes.
 - c) Aplicar otras técnicas como interpolación.

Tipos de datos faltantes:



Imputación de Datos

En el caso de la siguiente gráfica:



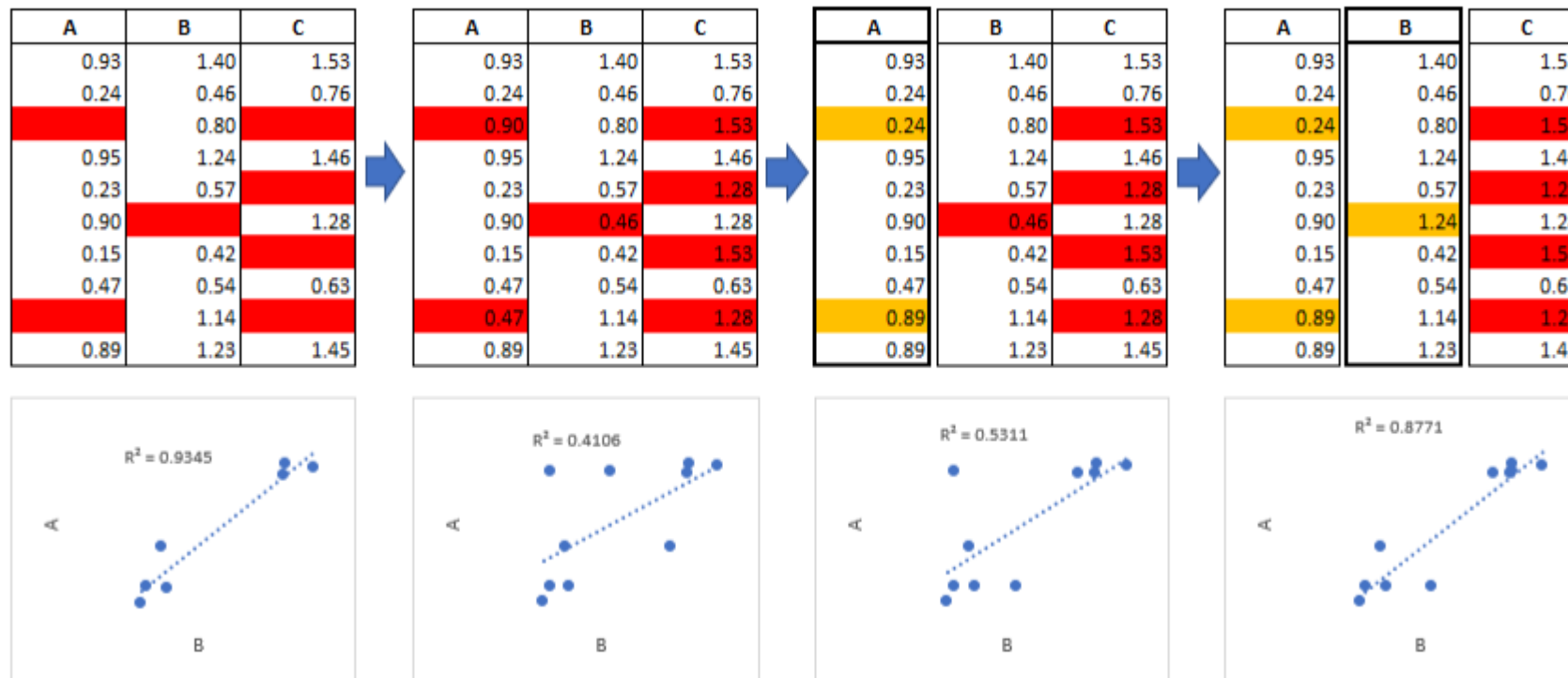
La interpolación estaría dada por la siguiente fórmula:

$$x_{valor} = a_{valor} + \left[\frac{(a_{valor} - b_{valor}) * (x_{tiempo} - b_{tiempo})}{a_{tiempo} - b_{tiempo}} \right]$$

Imputación de Datos

d) Aplicar vecinos más cercanos (KNN)

e) Técnicas de imputación para datos multivariados, por ejemplo, usar Random Forest o MICE.

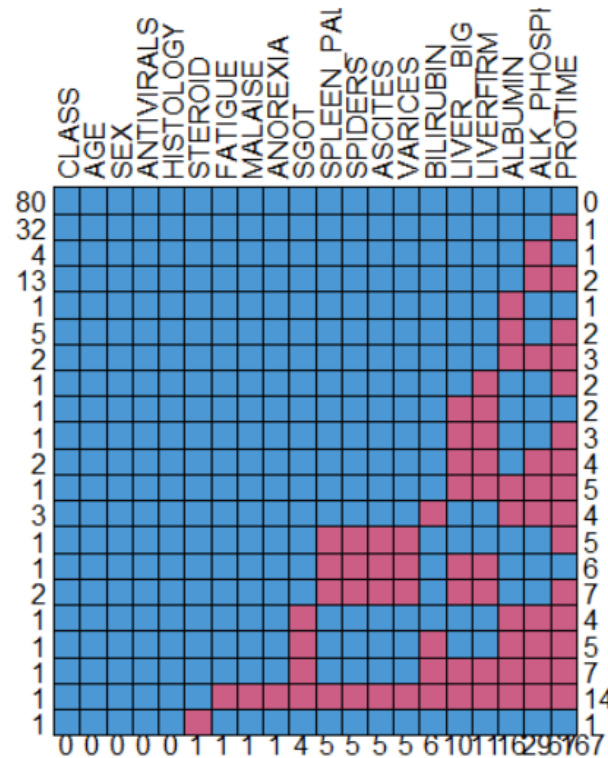


Imputación de Datos

Caso práctico: Se utilizará el dataset de hepatitis.

<https://archive.ics.uci.edu/dataset/46/hepatitis>

Datos faltantes:



Conclusiones

- 1) Existen diversas ramas dentro del campo de la IA.
- 2) Machine Learning comprende una de estas ramas, la cual engloba a otras áreas como Deep Learning.
- 3) Los datos pueden ser categóricos o continuos.
- 4) Es conveniente desarrollar un EDA a fin de observar los datos con los cuales se va a trabajar.