



UNIVERSIDAD
NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Formulario de Reconocimiento de Patrones y Aprendizaje Automatizado.

Integrantes:

Yonathan Berith Jaramillo Ramírez. 419004640

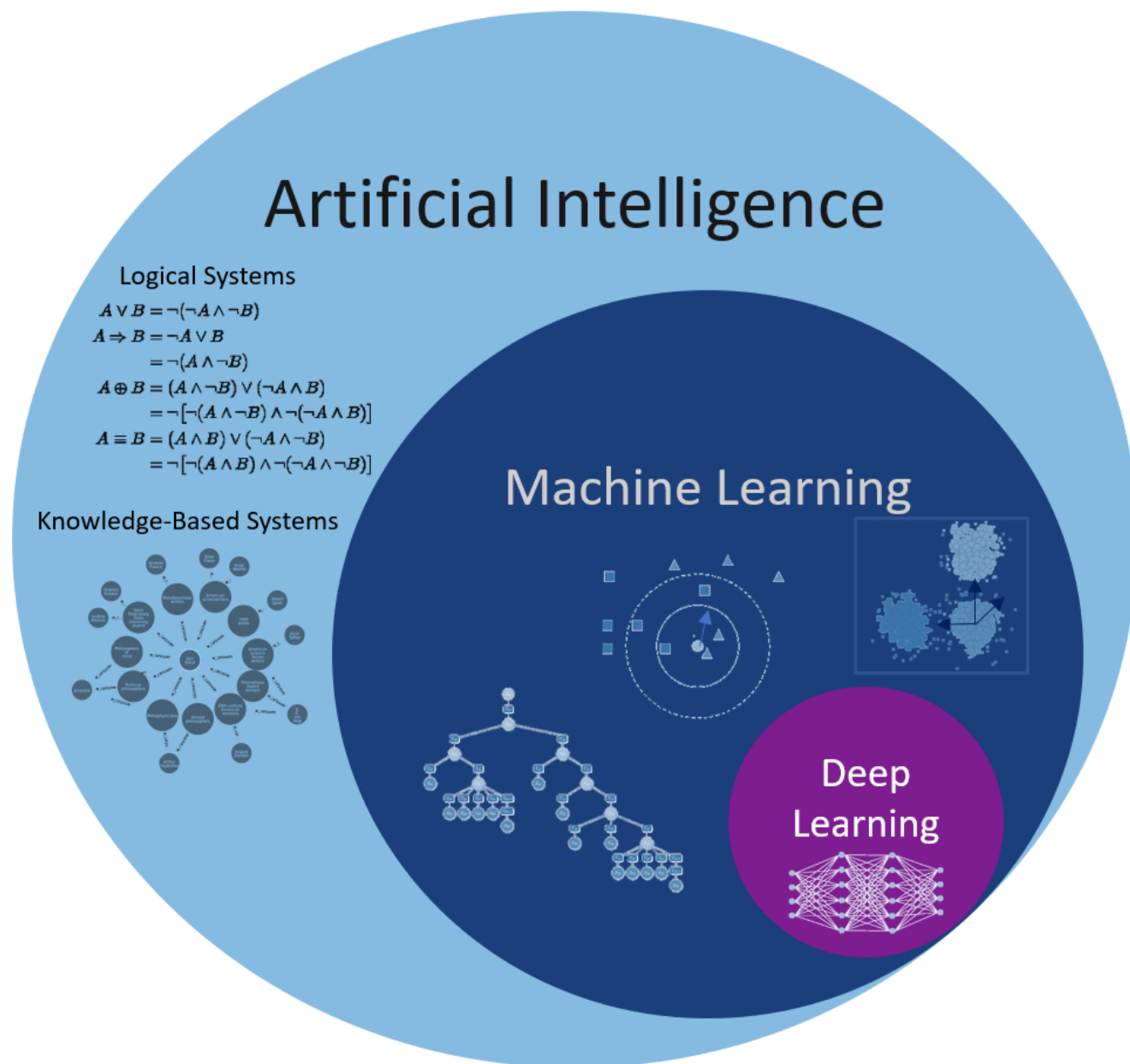
Profesor: Miguel Daniel Garrido Reyes

Ayudantes: Melissa Vázquez González

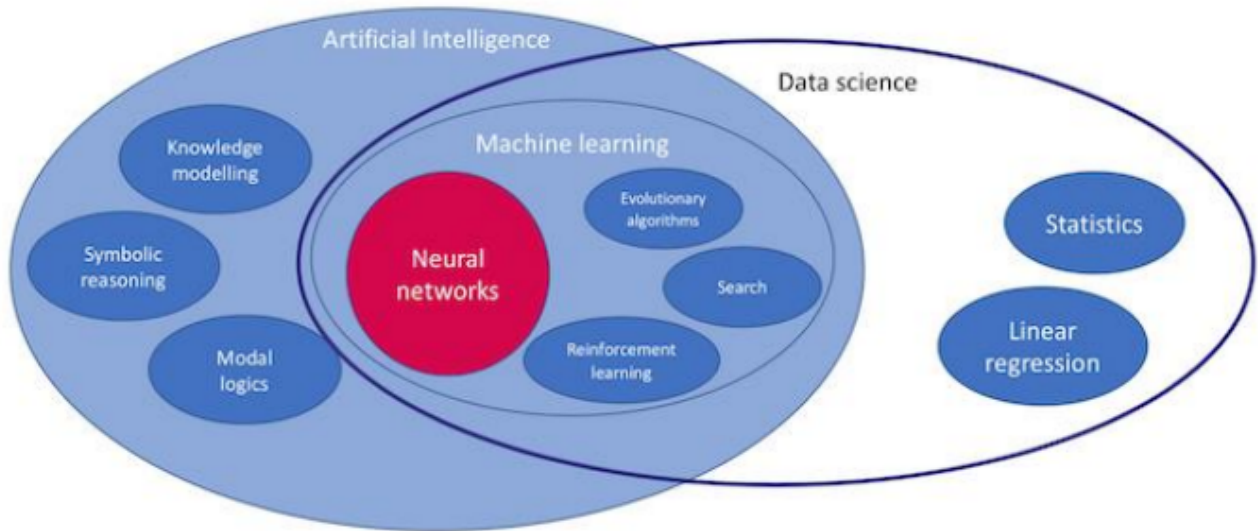
Luis Emilio González Covarrubias

12 Febrero, 2024

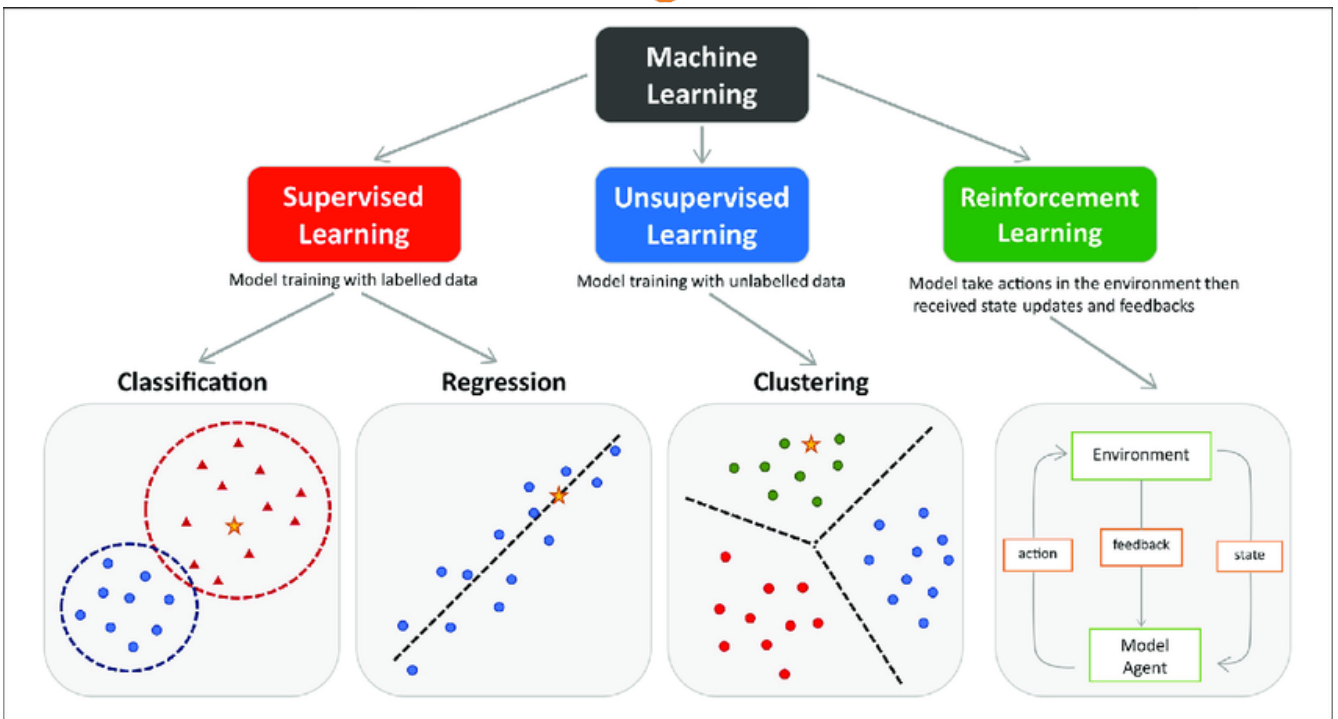
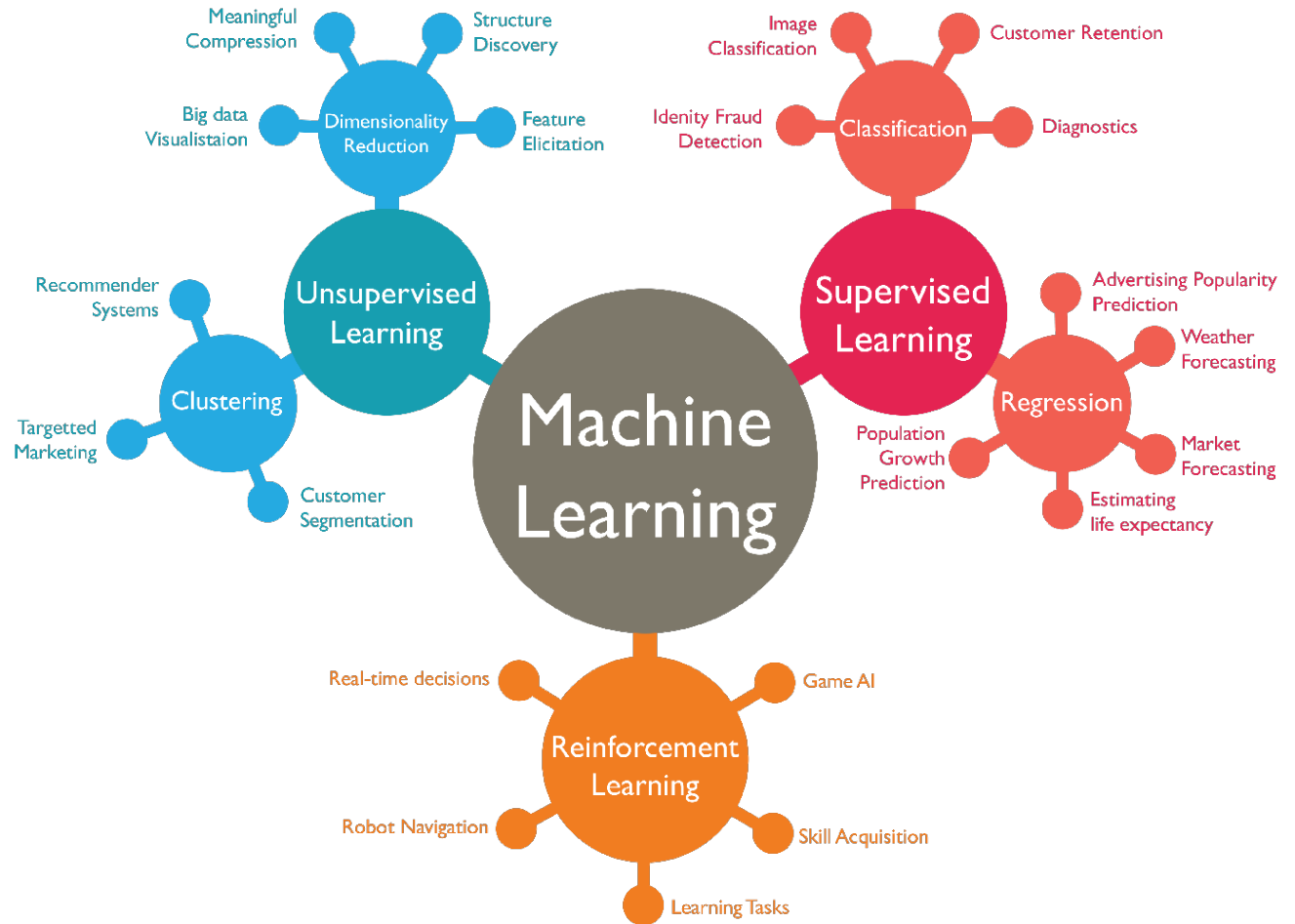
Reconocimiento de Patrones y Aprendizaje Automatizado



- **Inteligencia Artificial (IA):** Campo de la informática que busca crear sistemas capaces de realizar tareas que normalmente requieren inteligencia humana.



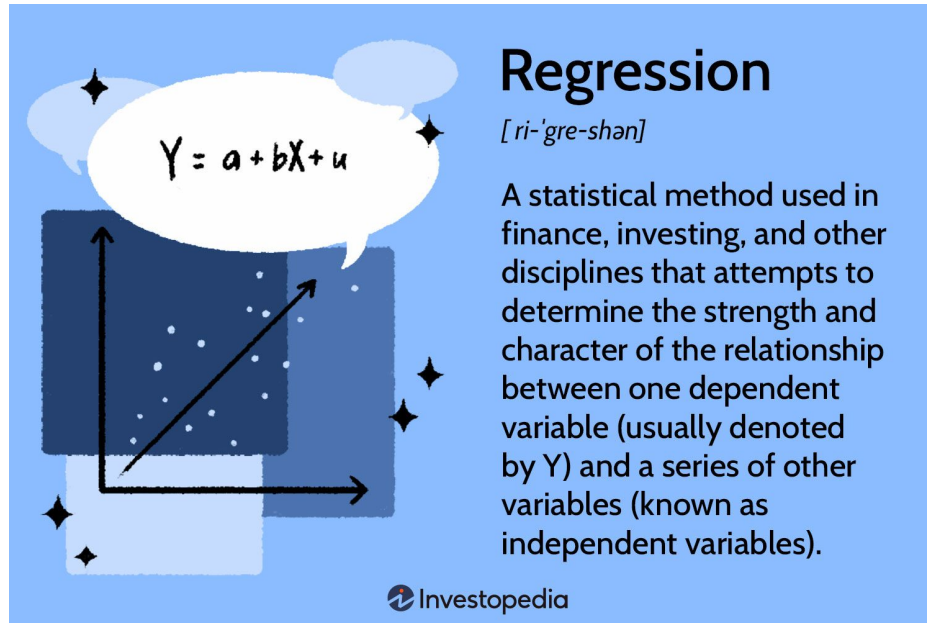
- **Aprendizaje Automático (Machine Learning, ML):** Rama de la IA que se centra en el desarrollo de algoritmos que pueden aprender de los datos y mejorar su desempeño con la experiencia.



- **Aprendizaje Profundo (Deep Learning, DL):** Subconjunto del aprendizaje

automático que utiliza redes neuronales multicapa para aprender de grandes cantidades de datos.

- **Modelado del Conocimiento:** Proceso de crear estructuras y conceptos para representar información en un campo específico, facilitando que una máquina emule el entendimiento humano.
- **Razonamiento Simbólico:** Manipulación de símbolos y abstracciones en lugar de números, enfocándose en la lógica y la inferencia.
- **Lógicas Modales:** Sistemas de lógica que tratan conceptos de posibilidad y necesidad, comúnmente usados en el razonamiento simbólico.
- **Algoritmos Evolutivos:** Métodos de optimización y búsqueda inspirados en la teoría de la evolución biológica.
- **Búsqueda (Search):** Técnicas en la IA para explorar sistemáticamente posibles configuraciones de un problema para encontrar una solución.
- **Aprendizaje por Refuerzo (Reinforcement Learning):** Tipo de aprendizaje automático donde un agente toma decisiones por medio de recompensas o penalizaciones.
- **Ciencia de Datos (Data Science):** Disciplina que combina estadísticas, análisis de datos y técnicas de aprendizaje automático para entender y analizar fenómenos reales.
- **Estadísticas:** Rama de las matemáticas que se ocupa del análisis, interpretación y presentación de datos.
- **Regresión Lineal:** Método estadístico que modela la relación entre una variable dependiente y una o más variables independientes asumiendo una relación lineal.



- **Patrón:** Regularidad detectable en un conjunto de datos que facilita la clasificación y análisis.

Probabilidad

- **Probabilidad:** La probabilidad es una herramienta matemática utilizada para estudiar el azar. Se trata de la oportunidad (la posibilidad) de que se produzca un evento.
- **Espacio muestral:** Conjunto de todos los posibles resultados del experimento Ω por ejemplo de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Evento:** Cualquier subconjunto del espacio muestral por ejemplo del dado $A = \{2, 4, 6\}$.
- **Axiomas de la probabilidad:**
 1. $P(A) \geq 0$
 2. $P(\Omega) = 1$
 3. $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$ cuando A_1, A_2, \dots son ajenos dos a dos.
- **Probabilidad Clásica:** Sea A un subconjunto de un espacio muestral Ω de cardinalidad finita. Se define la probabilidad clásica del evento A como el cociente en donde el símbolo $\#A$ denota la cardinalidad o número de elementos del conjunto A.

$$P(A) = \frac{\#A}{\#\Omega}$$

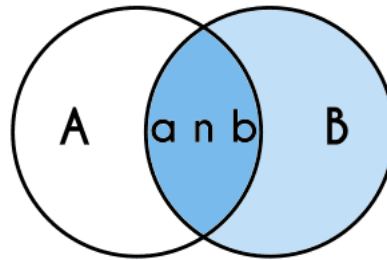
- **Probabilidad Frecuentista:** Sea n_A el número de ocurrencias de un evento A en n realizaciones de un experimento aleatorio. La probabilidad

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

- **Probabilidad Condicional:** Sea A y B dos eventos y supongamos que B tiene probabilidad estrictamente positiva. La probabilidad condicional del evento A , dado el evento B , se denota por el símbolo $P(A|B)$ y se define como el cociente

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Probabilidad condicional



- **Teorema de probabilidad total:** Sea B_1, \dots, B_n una partición de Ω tal que $P(B_i) \neq 0$, $i = 1, \dots, n$. Para cualquier evento A ,

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i).$$

Además, para un caso particular con dos eventos B y B^C (el complemento de B),

$$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C).$$

- **Teorema de Bayes:** Sea B_1, \dots, B_n una partición de Ω tal que $P(B_i) \neq 0$, $i = 1, \dots, n$. Sea A un evento tal que $P(A) \neq 0$. Entonces para cada $j = 1, 2, \dots, n$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}.$$

Para un caso particular con dos eventos B y B^C (el complemento de B),

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

- **Independencia Lineal:** El concepto de independencia representa la situación cuando la ocurrencia de un evento no afecta la probabilidad de ocurrencia de otro evento.

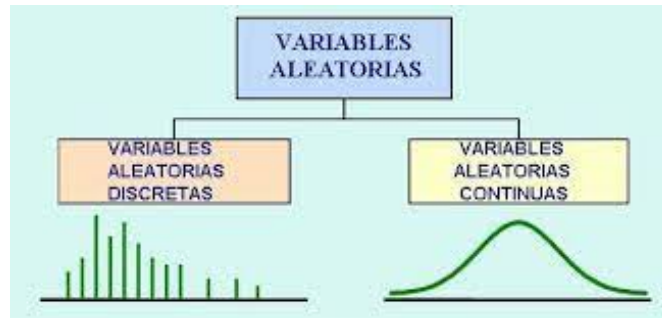
$$P(A \cap B) = P(A)P(B)$$

- **Variables aleatorias:** Una variable aleatoria es una transformación X del espacio de resultados Ω al conjunto de números reales, esto es,

$$X : \Omega \rightarrow \mathbb{R},$$

tal que para cualquier número real x ,

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$



- **Función de probabilidad:** Es la función que indica la probabilidad en los distintos valores que toma la variable aleatoria.

$$f(x) = \begin{cases} P(X = x) & \text{si } x = x_0, x_1, \dots \\ 0 & \text{en otro caso.} \end{cases}$$

- **Función de distribución:** La función de distribución evaluada en un número x cualquiera es la probabilidad de que la variable aleatoria tome un valor menor o igual a x , o en otras palabras, que tome un valor en el intervalo $(-\infty, x]$.

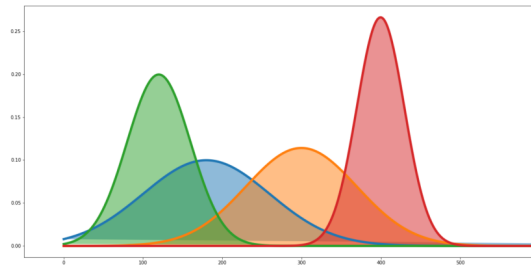
$$F(x) = P(X \leq x).$$

Para una variable aleatoria discreta, se expresa como:

$$F(x) = \sum_{u \leq x} f(u),$$

y para una variable aleatoria continua, como:

$$F(x) = \int_{-\infty}^x f(u) du.$$



- **Esperanza (media, valor esperado o valor promedio):** Es un número que indica el promedio ponderado de los diferentes valores que la variable puede tomar. Para una variable aleatoria discreta se define como:

$$E(X) = \sum_x x f(x),$$

y para una variable aleatoria continua se define como:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

- **Varianza:** Es una medida del grado de dispersión de los diferentes valores tomados por la variable. Para una variable aleatoria discreta se define como:

$$\text{Var}(X) = \sum_x (x - \mu)^2 f(x),$$

y para una variable aleatoria continua se define como:

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

En ambos casos, la varianza también se puede expresar en términos de la esperanza como:

$$\text{Var}(X) = E[(X - \mu)^2].$$

- **Covarianza:** Describe la relación estadística entre dos variables aleatorias. Se define como la media del producto de las desviaciones de cada una de las variables respecto a su propia media.

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

- **Álgebra lineal:** Proporciona las herramientas matemáticas esenciales para representar y manipular datos mediante vectores y matrices.
- **Vector:** Es una matriz de una sola columna que representa una cantidad con magnitud y dirección en el espacio.

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

- **Matriz:** Es un arreglo rectangular de números, símbolos o expresiones, dispuestos en filas y columnas.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- **Matriz cuadrada:** Es una matriz con el mismo número de filas y columnas.
- **Matriz identidad:** Es una matriz cuadrada en la que todos los elementos de la diagonal principal son 1 y el resto son 0.

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- **Matriz diagonal:** Es una matriz cuadrada en la que solo los elementos de la

diagonal principal son distintos de cero.

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

- **Matriz triangular superior:** Es una matriz cuadrada donde todos los elementos por debajo de la diagonal principal son cero.

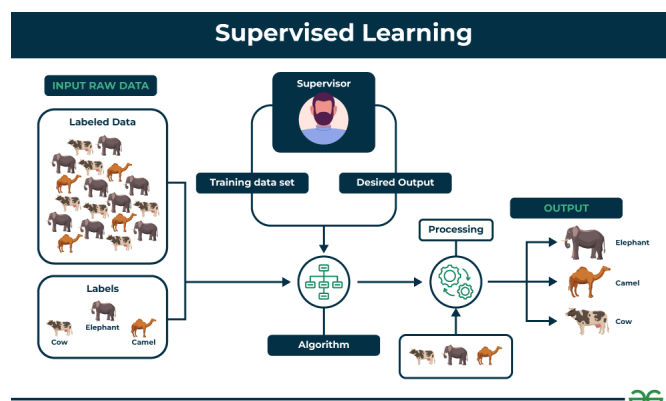
$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

- **Matriz simétrica:** Es una matriz cuadrada que es igual a su transpuesta, es decir, $a_{ij} = a_{ji}$.

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{bmatrix}$$

Aprendizaje supervisado y teoría del aprendizaje

- **Aprendizaje:** Proceso por el cual un sistema informático mejora su desempeño en una tarea específica con el tiempo, a través de la experiencia. Este proceso implica la adquisición de datos de entrada etiquetados para deducir o aprender una función que, una vez entrenada, produce una salida adecuada cuando se le presentan nuevos datos sin etiquetar. El objetivo del aprendizaje es que el sistema pueda realizar predicciones o clasificaciones precisas sobre datos no vistos anteriormente, basándose en el conocimiento adquirido durante la fase de entrenamiento.



- **Regresión Lineal:** Modelo que asume una relación lineal entre las variables independientes x y la variable dependiente y . Se expresa con la fórmula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (1)$$

donde β_0 es el término de intercepción, β_1, \dots, β_n son los coeficientes de las variables independientes y ϵ es el término de error.

- **Regresión no Lineal:** Extiende la regresión lineal permitiendo relaciones no lineales entre las variables independientes y la variable dependiente.
- **Máquinas de Vectores de Soporte (SVM):** Encuentra el hiperplano en un espacio de dimensión elevada que mejor clasifica los datos de entrenamiento con el mayor margen posible. La fórmula para el hiperplano es:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (2)$$

donde \mathbf{w} es el vector de pesos y b es el sesgo.

- **Árboles de Decisión:** Modelo que utiliza una estructura de árbol para tomar decisiones basadas en las características de los datos de entrada, bifurcando en cada característica hasta llegar a una decisión.
- **Bosques Aleatorios:** Consiste en un conjunto de árboles de decisión, donde cada árbol se entrena con una muestra de los datos y la decisión final se toma por votación mayoritaria.
- **Redes Neuronales:** Compuestas de neuronas interconectadas que transmiten señales procesando las entradas a través de funciones de activación. Se modelan como:

$$y = f\left(\sum_i w_i x_i + b\right) \quad (3)$$

donde x_i son las entradas, w_i son los pesos, b es el sesgo y f es la función de activación.

- **Naive Bayes:** Modelo probabilístico basado en el teorema de Bayes, con la "ingenua" suposición de independencia entre las características. La fórmula general es:

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)} \quad (4)$$

donde $P(y|x_1, \dots, x_n)$ es la probabilidad posterior, $P(y)$ es la probabilidad previa de la clase, $P(x_i|y)$ es la probabilidad de la característica dada la clase, y $P(x_1, \dots, x_n)$ es la probabilidad de los datos.

- **Regresión Logística:** Modelo utilizado para predecir la probabilidad de una variable binaria. La función logística es:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (5)$$

donde $P(y = 1|x)$ es la probabilidad de que la variable dependiente sea 1 dado x , y β representa los parámetros del modelo.

- **Error:** El error en un modelo de aprendizaje automático se refiere a la discrepancia entre las predicciones realizadas por el modelo y los valores reales observados. Matemáticamente, puede expresarse de varias maneras dependiendo del tipo de tarea (regresión o clasificación). Para la regresión, un error común es el error cuadrático medio (MSE), dado por:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

donde y_i son los valores reales y \hat{y}_i son las predicciones del modelo para n observaciones.

Para tareas de clasificación, el error se suele expresar en términos de la tasa de error, que es la proporción de predicciones incorrectas sobre el total de predicciones, o alternativamente, mediante métricas como la precisión, el recall o el F1-score, que consideran tanto los verdaderos positivos como los falsos positivos y negativos.

- **Función de Pérdida:** Es una función que mide el desempeño de nuestra hipótesis de modelo al comparar las predicciones con los valores reales. Nos ayuda a entender cuán bien o mal está haciendo predicciones nuestro modelo. Hay diferentes tipos de funciones de pérdida para diferentes tipos de problemas de aprendizaje automático.
- **0-1 Loss:** Utilizada principalmente en problemas de clasificación, asigna un error de 1 si la predicción $f(x)$ es diferente del valor real y , y 0 si son iguales. Se puede expresar como:

$$L(y, f(x)) = \begin{cases} 1 & \text{si } y \neq f(x) \\ 0 & \text{si } y = f(x) \end{cases} \quad (7)$$

- **Quadratic Loss:** También conocida como pérdida cuadrática media (MSE) para problemas de regresión, mide el cuadrado de la diferencia entre el valor predicho $f(x)$ y el valor real y . La fórmula es:

$$L(y, f(x)) = (y - f(x))^2 \quad (8)$$

- **Error Cuadrático Medio (MSE - Mean Squared Error):**

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9)$$

donde y_i son los valores reales y \hat{y}_i son las predicciones del modelo.

- **Error Absoluto Medio (MAE - Mean Absolute Error):**

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

- **Error Cuadrático Medio Logarítmico (MSLE - Mean Squared Logarithmic Error):**

$$MSLE = \frac{1}{n} \sum_{i=1}^n (\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (11)$$

- **Pérdida de Huber:**

$$\text{Pérdida} = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{si } |y_i - \hat{y}_i| \leq \delta \\ \delta(|y_i - \hat{y}_i| - \frac{1}{2}\delta) & \text{en otro caso} \end{cases} \quad (12)$$

- **Entropía Cruzada Binaria (Binary Cross-Entropy):**

$$\text{Binary Cross-Entropy} = -(y \log(p) + (1 - y) \log(1 - p)) \quad (13)$$

- **Entropía Cruzada Categórica (Categorical Cross-Entropy):**

$$\text{Categorical Cross-Entropy} = - \sum_i y_i \log(p_i) \quad (14)$$

- **Pérdida Hinge:**

$$\text{Pérdida Hinge} = \max(0, 1 - y \cdot \hat{y}) \quad (15)$$

- **Pérdida Logística (Log Loss):**

$$\text{Log Loss} = -\log(p) \quad (16)$$

- **Precisión (Accuracy):**

$$\text{Precisión} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}} \quad (17)$$

- **Recall (Sensibilidad):**

$$\text{Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}} \quad (18)$$

- **F1-Score:**

$$F1\text{-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}} \quad (19)$$

Loss Function	Applicability to Classification	Applicability to Regression	Sensitivity to Outliers
Mean Squared Error (MSE)	✗	✓	High
Mean Absolute Error (MAE)	✗	✓	Low
Cross-Entropy	✓	✗	Medium
Hinge Loss	✓	✗	Low
Huber Loss	✗	✓	Medium
Log Loss	✓	✗	Medium

- **Riesgo estructural:** Se refiere a la capacidad del modelo para adaptarse a los datos de entrenamiento y su capacidad para generalizar a nuevos datos. Este concepto es fundamental en el aprendizaje automático para evitar el sobreajuste y el subajuste.
- **Minimización de riesgo estructural:** Es una estrategia que busca encontrar un equilibrio entre el ajuste a los datos de entrenamiento y la generalización a datos no vistos. La función de riesgo estructural se puede expresar matemáticamente como:

$$R_{sm}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda \Omega(f) \quad (20)$$

donde $L(y_i, f(x_i))$ es la función de pérdida que mide el error de predicción del modelo f en el punto x_i , λ es el parámetro de regularización que controla la complejidad del modelo, y $\Omega(f)$ es el término de regularización que penaliza la complejidad del modelo para evitar el sobreajuste.

- **Generalización:** Se refiere a la capacidad de un modelo para adaptarse adecuadamente a datos previamente no vistos, reflejando la verdadera distribución

de los datos. Un modelo bien generalizado funcionará bien tanto en los datos de entrenamiento como en los nuevos datos.

- **Entrenamiento:** Es el proceso de ajustar un modelo a los datos de entrenamiento. Un entrenamiento adecuado implica que el modelo ha aprendido las relaciones subyacentes entre las entradas y las salidas de los datos de entrenamiento.
- **Sobre-entrenamiento (Overfitting):** Ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando ruido además de las relaciones subyacentes. Esto se caracteriza por una disminución del rendimiento en datos no vistos, a pesar de un rendimiento excelente en los datos de entrenamiento.
- **Complejidad:** La complejidad de un modelo de aprendizaje automático cuantifica o mide la variedad de relaciones entre los datos que el modelo es capaz de aprender. Un modelo más complejo puede capturar patrones más complicados, pero también corre el riesgo de sobreajuste si la complejidad es demasiado alta para la cantidad de datos de entrenamiento disponibles.
- **Dimensionalidad:** La dimensionalidad se refiere a la cantidad de variables o características presentes en los datos. La alta dimensionalidad puede hacer que el análisis de los datos sea más difícil, un fenómeno conocido como la "maldición de la dimensionalidad", y puede requerir métodos especiales de reducción de dimensionalidad para mejorar el rendimiento del modelo.
- **Dimensión de Vapnik-Chervonenkis (VC):** La Dimensión VC de una clase de funciones es una medida teórica de su capacidad para clasificar conjuntos de puntos. En particular, representa la cardinalidad del mayor conjunto de puntos que un algoritmo puede separar en todas las formas posibles, lo que implica que para cualquier asignación de etiquetas a esos puntos, el algoritmo puede encontrar una función en su clase que realice una clasificación perfecta.
- La dimensión VC es importante porque da una medida de la capacidad de generalización de un modelo: un modelo con una dimensión VC alta puede tener la capacidad de aprender una amplia variedad de patrones, pero también puede ser propenso al sobreajuste si la dimensión VC es demasiado grande en relación con la cantidad de datos de entrenamiento.
- En problemas de clasificación binaria, un modelo con una baja dimensión VC puede no ser suficiente para capturar la complejidad de los datos, mientras que un modelo con una dimensión VC demasiado alta puede aprender perfectamente

los datos de entrenamiento, incluido el ruido, y por lo tanto no generalizar bien a datos nuevos.

- **Shattering:** Se refiere a la capacidad de un conjunto de hipótesis de clasificar de todas las maneras posibles un conjunto de puntos. En términos de la Dimensión de Vapnik-Chervonenkis (VC), si un conjunto de puntos puede ser "destrozado" por un conjunto de hipótesis, entonces la VC-dimensión de ese conjunto de hipótesis es al menos tan grande como la cantidad de puntos.
- Matemáticamente, si tenemos N puntos, hay 2^N problemas de aprendizaje diferentes que pueden ser definidos, lo que implica que para cada conjunto de N puntos etiquetados de todas las maneras posibles, existe una hipótesis en el conjunto que clasifica los puntos exactamente como están etiquetados.

Este concepto es crucial para comprender cómo un modelo puede generalizar a partir de datos de entrenamiento limitados y cómo la capacidad del modelo para clasificar conjuntos de datos complejos se relaciona con su VC-dimensión.

- **Ruido:** Se refiere a cualquier tipo de anomalía no deseada en los datos que puede introducir errores y dificultar el proceso de aprendizaje de un modelo. El ruido puede provenir de varias fuentes, incluyendo:
 - Imprecisión en la captura de atributos: Datos que no reflejan con precisión las características reales que se están midiendo.
 - Errores en la etiquetación de los datos: Etiquetas incorrectas proporcionadas durante el proceso de anotación.
 - Atributos adicionales no considerados: Características y variables que afectan los resultados pero no se incluyen en el análisis.

El ruido puede resultar en clasificaciones incorrectas y en la generación de modelos que no representan adecuadamente la realidad subyacente de los datos, lo que lleva a predicciones inexactas y a una generalización pobre a nuevos datos.

Teoría de decisiones bayesiana

Métodos paramétricos

Clasificadores lineales

Clasificadores no-lineales

Agrupamiento (Clustering) y aprendizaje no-supervisado

Aplicaciones

Código