1. Regresión Simple

Es un modelo estadístico que estudia la relación entre una variable dependiente y una sola variable independiente. La fórmula es:



Ejemplo: Estimar el salario en función de los años de experiencia.

2. Regresión Múltiple

Es un modelo que estudia la relación entre una variable dependiente y múltiples variables independientes.

Ejemplo: Predecir el precio de una casa basándose en el tamaño, número de habitaciones y ubicación.

3. Regresión Polinómica

Es una extensión de la regresión lineal que permite modelos con curvas al incluir potencias de la variable independiente.

Ejemplo: Ajustar una curva a la relación entre la temperatura y el rendimiento de una planta.

4. Clasificación Binaria

Un tipo de problema de clasificación en el que las instancias se dividen en dos clases.

Ejemplo: Detectar si un correo es spam (1) o no es spam (0).

5. Clasificación Multiclase

Clasificación en la que hay más de dos clases posibles.

Ejemplo: Clasificar imágenes de flores en tres tipos: rosa, margarita, tulipán.

6. Clasificación Multi Etiqueta

Cada instancia puede pertenecer a múltiples clases simultáneamente.

Ejemplo: Clasificar una película en varios géneros: acción, comedia, drama.

7. Clustering Jerárquico

Es un método de agrupación que construye una jerarquía de clusters, ya sea fusionando grupos pequeños o dividiendo uno grande.

Ejemplo: Agrupar especies de animales por similitud genética.

8. K (# de grupos)

Número de grupos o clusters en los que se quiere dividir los datos en algoritmos de clustering, como K-means.

Ejemplo: Si "K=3", se dividirán los datos en 3 clusters.

Librerías utilizadas

Manipulación de datos

* Pandas
* NumPy

Visualización

* Matplotlib
* Seaborn
* Plotly

Reportes

* Pandas Profiling
* Sweetviz

Análisis estadístico

* SciPy
* Statsmodels

9. Error Cuadrático Medio (MSE)

Mide el promedio de los cuadrados de los errores (diferencias entre valores predichos y reales).



10. Error Absoluto Medio (MAE)

Mide el promedio de las diferencias absolutas entre los valores predichos y los valores reales. 

11. R² (Coeficiente de Determinación)

Mide la proporción de la varianza en la variable dependiente que es explicada por las variables independientes.



12. Regresión Logística

Es un modelo de clasificación que usa la función sigmoide para predecir la probabilidad de una de dos posibles clases.

Ejemplo: Diagnosticar si un paciente tiene o no una enfermedad.

13. Función de Costo

Evalúa el error de un modelo de machine learning, cuantificando la diferencia entre las predicciones y los resultados reales.

Ejemplo: El MSE es una función de costo común en regresión.

14.K Vecinos Más Cercanos (KNN)

Es un algoritmo de clasificación basado en la proximidad de una instancia a sus "k" vecinos más cercanos en el espacio de características.

Ejemplo: Clasificar si una fruta es manzana o pera según las frutas más cercanas en términos de tamaño y color.

15. Número de Vecinos (K)

Es el parámetro de KNN que indica cuántos vecinos se usan para hacer la predicción.

Ejemplo: Si \( k = 5 \), se toman los 5 vecinos más cercanos.

16. K-means (Agrupación)

Algoritmo de clustering que agrupa los datos en "k" clusters, minimizando la variación dentro de cada grupo.

Ejemplo: Agrupar clientes en diferentes segmentos de mercado.

17. Máquina de Vectores de Soporte (SVM)

Es un algoritmo que separa diferentes clases (grupos) de datos usando una línea o plano (hiperplano). Trata de dejar el mayor espacio posible entre esa línea y los datos.

Ejemplo: Es como si quisieras separar juguetes de dos tipos diferentes con una cuerda, tratando de dejar la cuerda en el mejor lugar para que no haya confusión entre los tipos.

18. Hiperplano

En SVM, es la línea (en 2D o plano en 3D) que separa las clases en un espacio de características.

Ejemplo: Una recta que separa círculos azules y triángulos rojos en un gráfico 2D.

19. Margen Máximo

Es la distancia más amplia posible entre el hiperplano y las instancias más cercanas de cada clase en SVM.

Ejemplo: Mayor margen significa menor riesgo de error de clasificación.

Diferencia entre Módulo y Librería

- Módulo: Es un archivo de código Python (.py) que puede contener funciones, clases y variables que puedes reutilizar.

Ejemplo: "math` es un módulo que contiene funciones matemáticas.

- Librería: Es una colección de módulos organizados bajo un mismo nombre. Puede ser un conjunto más amplio de funcionalidades.

Ejemplo: NumPy es una librería que incluye múltiples módulos para manipulación de arrays y funciones matemáticas avanzadas.

20. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)(Agrupamiento espacial basado en densidad de aplicaciones con ruido)

Es un algoritmo que agrupa puntos que están muy cerca entre sí y deja fuera los puntos que están lejos, llamándolos "ruido".

Ejemplo: Imagina que tienes muchas estrellas en el cielo. DBSCAN formará grupos de estrellas que están cerca unas de otras y las que están muy separadas se quedarán solas.

21. K-means

Este algoritmo agrupa datos en "K" grupos. Primero elige puntos iniciales (centroides) al azar, luego ajusta esos puntos hasta que los grupos sean lo más parecidos posible dentro de ellos.

Ejemplo: Si tienes un montón de frutas diferentes, K-means puede agruparlas según su tamaño y color, por ejemplo, todas las manzanas en un grupo y todas las naranjas en otro.

22. PCA (Análisis de Componentes Principales)

Es una técnica que toma datos complicados y los simplifica a unos pocos factores que representan la mayor parte de la información.

Ejemplo: Si tienes una foto en alta resolución, PCA sería como reducir la imagen para que sea más pequeña, pero aún se vea casi igual.

22. Algoritmo de Optimización

Son métodos que intentan encontrar la mejor solución a un problema minimizando o maximizando algo. En Machine Learning, se usan para ajustar los parámetros del modelo.

Ejemplo: Un algoritmo de optimización sería como encontrar el camino más corto para llegar a casa cuando hay mucho tráfico. Ajusta tu ruta para llegar lo más rápido posible.

24. K-medoids

Es similar a K-means, pero en lugar de usar puntos aleatorios como centros de los grupos, K-medoids usa puntos reales de los datos. Esto lo hace más robusto frente a datos ruidosos.

Ejemplo: Si estás agrupando ciudades por su tamaño, K-medoids elegiría una ciudad real como centro del grupo, mientras que K-means puede elegir un punto en medio de la nada como centro.

25. KNN (K Vecinos Más Cercanos)

Es un algoritmo que, para clasificar algo nuevo, mira a sus "vecinos" más cercanos. Si la mayoría de los vecinos son de una clase, entonces el nuevo objeto también será de esa clase.

Ejemplo: Si tienes una nueva fruta y quieres saber si es una manzana o una naranja, KNN miraría las frutas más cercanas. Si la mayoría son manzanas, diría que es una manzana.

26. Definición R^2: El coeficiente de determinación R^2 mide la proporción de la variabilidad en la variable dependiente que se puede explicar por el modelo de regresión. Varía de 0 a 1, donde 0 indica que el modelo no explica nada de la variabilidad y 1 indica que lo explica todo.

Ejemplo: Supongamos que estamos analizando la relación entre las horas de estudio y las calificaciones de los estudiantes. Si el R^2 de nuestro modelo de regresión lineal es 0.85, esto significa que el 85% de la variabilidad en las calificaciones puede explicarse por las horas de estudio.

27. Datos atípicos: También conocidos como **outliers**, son observaciones que se desvían significativamente del resto de los datos en un conjunto. Estos valores extremos pueden influir en los resultados del análisis estadístico y pueden surgir por diversas razones, como errores de medición, variabilidad natural o comportamientos inusuales.

Ejemplo: En un conjunto de datos de ingresos anuales, un individuo con un ingreso extremadamente alto en comparación con los demás sería un outlier.

28. La Ley del Codo: establece que el número óptimo de clusters se encuentra en el punto donde la reducción en la varianza intragrupo (SSE: *Sum of Squared Errors*) empieza a ser menos significativa, formando una curva con forma de "codo".

### Ejemplo:En un problema de K-means, se realiza el gráfico de SSE vs. número de clusters (k). Al principio, añadir más clusters reduce significativamente el error, pero llega un punto en el que las mejoras disminuyen. Ese punto es el "codo" y se recomienda como el número óptimo de clusters.