IA: Habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana.

ML: Campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas. Los sistemas pueden aprender de los datos, identificar patrones y tomar decisiones.

Aprendizaje supervisado: El algoritmo necesita un **conjunto de datos de capacitación etiquetado con las respuestas correctas** para poder aprender.

Si el valor es continuo es un problema de REGRESIÓN:

| MÉTRICAS | ALGORITMOS |
| --- | --- |
| Error absoluto medio(MAE): Diferencia entre el valor predicho y el valor real.  Error cuadrático medio(MSE): Diferencia entre el valor real y el predicho al cuadrado. | Regresión lineal: traza una línea recta a un conjunto de puntos y aprende la tendencia para hacer una predicción nuevos puntos de datos |
| Raíz del error cuadrático medio (RMSE):Sacamos raíz para tenerlo en las unidades originales. |  |

Si el valor es discreto(variable categórica) es un problema de CLASIFICACIÓN

| MÉTRICAS | ALGORITMOS |
| --- | --- |
| matriz de confusión(tabla que describe el rendimiento)  accuracy(exactitud) % total de valores correctamente clasificados  recall: + REALES correctamente ID | Regresión logística: método estadístico que calcula la probabilidad de ocurrencia de una variable dicotómica en base de otras VI |
| precisión: IDENTIFICADORES + correctos.  Especificidad: - reales correctamente ID(contrario a recall)  F1 Score: combina precisión y recall se usa cuando el conjunto de datos está desbalanceado. | KNN Vecinos más cercanos: Compara nuevos puntos de datos con los existentes, busca los vecinos más cercanos y los asocia |
| Área bajo curva ROC: Gráfico para evaluar el rendimiento de un modelo. recall + ratio de falsos positivos. |  |

SIRVEN PARA AMBOS:

Árboles de decisión: Modelos predictivos que dividen los casos de un universo de datos. Según sus atributos se discrimina por la variable más relevante.

Bosque aleatorio: Combina varios árboles de decisión, usa el concepto de “alumno débil” como un predictor que funciona mal solo pero en conjunto produce un buen resultado. Cada árbol aprende con la implementación del algoritmo, un predictor fuerte está aprendiendo cómo combinar los resultados de diferentes árboles.

Máquina de Vectores de Soporte: Encuentra un hiperplano que ayude a dividir el dataset entre 2 clases, todo nuevo punto cae de un lado dibujado y tiene 1 etiqueta, si cae del otro tendrá otro valor.

Aprendizaje no supervisado: El conjunto de datos de capacitación no tiene etiquetas con respuesta correcta. El algoritmo saca sus propias conclusiones comparando los datos consigo mismo.

Ej: Segmentación de clientes en grupos, agrupar películas por tipos, descubrir zonas con alta tasa de criminalidad.

MÉTRICAS: Validación interna: evalúan coherencia y calidad de los clusters sin recurrir a información externa. Se centran en características intrínsecas.

Validación externa: comparan los resultados con una verdad conocida.

Flujo de trabajo: Data source, recolección de datos BBDD o API.

Data processing: EDA e ingeniería de las características, que es manipulación de datos agregación,eliminación y combinación del dataset para mejorar la performance de ML

Model training y deployment.

PROBLEMAS:

Adopción de mascotas:

1. Análisis de Datos y Preprocesamiento

Recolección de Datos: Obtener datos relevantes como edad, raza, tamaño, salud, tiempo en el refugio, etc.

Exploración de Datos (EDA): Identificar tendencias, patrones, y anomalías en los datos. Por ejemplo, puede que ciertas razas se adopten más rápido.

Limpieza de Datos: Tratar valores perdidos, eliminar duplicados, y corregir errores.

Ingeniería de Características: Crear nuevas características que puedan ser útiles, como categorías de edad (cachorro, adulto, anciano) o la duración de la estancia en el refugio.

2. Elección del Algoritmo:Es un problema de clasificación (mascotas adoptadas rápidamente vs. no adoptadas rápidamente).

Algoritmo Sugerido: Árboles de decisión o bosque aleatorio, dado que pueden manejar características categóricas y numéricas y son buenos para problemas de clasificación.

Estos algoritmos son efectivos para identificar las características más importantes que influyen en la adopción rápida.

3. Métricas a Utilizar

Precisión (Accuracy): Porcentaje total de predicciones correctas.

Recall: Proporción de casos positivos (adopciones rápidas) identificados correctamente.

F1 Score: Balance entre precisión y recall, útil si hay un desequilibrio en las clases.

Rendimiento de Estudiantes:

Tipo de Problema: Aprendizaje supervisado con datos etiquetados (aprobado/desaprobado).

Algoritmo: Modelo de clasificación con árboles de decisión para entender qué factores (como técnicas de enseñanza) influyen más.

Métricas: Precisión para medir cuántos estudiantes se clasifican correctamente y recall para ver cuántos aprobados/desaprobados se identifican correctamente.

Ventas de Productos en Tiendas Online:

Aprendizaje supervisado de regresión, porque se predice un valor continuo (ventas).

Análisis y Preprocesamiento:

EDA: Identificar patrones estacionales, tendencias de productos, y la influencia de eventos como Navidad.

Ingeniería de Características: Crear características relacionadas con el tiempo (mes, día de la semana), categorías de productos, y eventos especiales.

Métricas:

Error cuadrático medio, error absoluto medio, raíz del error cuadrático medio para evaluar qué tan cerca están las predicciones de las ventas reales.

Algoritmo Sugerido:

Regresión Lineal: Buena opción para empezar, pero podrías considerar algoritmos más complejos como regresión polinómica o bosques aleatorios si la relación entre las características y las ventas no es lineal.

Segmentacion de clientes:

El objetivo es dividir la base de clientes en grupos distintos basados en características comunes.

Recolección de Datos:

Recopilar datos de clientes como edad, género, ingresos, hábitos de compra, frecuencia de compras, preferencias de productos, etc.

Preprocesamiento de Datos:

Normalizar los datos para asegurar que cada característica contribuya equitativamente al análisis.

Aplicación de Algoritmos de Clustering: En el caso de K-means, se selecciona un número de clusters (K) y el algoritmo agrupa los datos de tal manera que la varianza dentro de cada grupo sea mínima.

Análisis de los Segmentos:Identificar patrones, como un segmento de clientes jóvenes interesados en productos tecnológicos o un grupo de clientes de mayor edad que prefieren productos de salud y bienestar.

METRICAS: Análisis de codo: Utilizado en K-means para determinar el número óptimo de clusters evaluando cómo cambia la varianza total dentro de los clusters con diferentes valores de K.