Macroentrenamiento en Inteligencia Artificial MeIA 2025

LA IMPORTANCIA DEL DATASET EN ML

UCI DATASET

QUE ES UN DATASET

- Un **dataset** es un conjunto estructurado de datos que se utiliza para entrenar modelos de Machine Learning. (o Aprendizaje Automático) Contiene información organizada en filas y columnas, donde cada fila representa una instancia y cada columna una característica relevante. La calidad y diversidad de un dataset son fundamentales para garantizar que los modelos de aprendizaje automático sean precisos y generalizables.

Los datasets son importantes porque:

- **Mejoran la precisión**: Un buen dataset permite que el modelo aprenda patrones reales y evite sesgos.
- **Facilitan la generalización**: Si los datos son variados, el modelo podrá hacer predicciones en nuevos escenarios.
- **Reducen el sobreajuste**: Un dataset equilibrado evita que el modelo memorice datos en lugar de aprender tendencias.
- **Permiten la validación**: Se dividen en entrenamiento, validación y prueba para evaluar el rendimiento del modelo.

REEX
ROBÓTICA EVOLUTIVA EXPERIMENTAL

1. **UCI Machine Learning Repository**: Contiene múltiples datasets médicos, como el de enfermedades cardíacas y diabetes.
2. **MIMIC-III**: Base de datos de registros médicos electrónicos de pacientes en cuidados intensivos.
3. **CheXpert**: Dataset de imágenes de rayos X para el diagnóstico de enfermedades pulmonares.
4. **HAM10000**: Conjunto de imágenes de lesiones cutáneas para la detección de cáncer de piel.
5. **PhysioNet**: Datos de señales fisiológicas como electrocardiogramas (ECG) para el análisis de enfermedades cardíacas.
6. **Kaggle Medical Datasets**: Plataforma con múltiples datasets médicos abiertos para investigación.

PRINCIPALES DATASETS PARA DIAGNÓSTICO MÉDICO

REEX
ROBÓTICA EVOLUTIVA EXPERIMENTAL

Cuadro 2 Descripción del conjunto de datos

De: Una evaluación comparativa de enfoques de conjuntos de aprendizaje automático para la predicción de enfermedades utilizando múltiples conjuntos de datos

Conjunto de datos	Referencia	Conjunto de datos	Tipo de conjunto de datos	No. de atributos	No. de instancias	Positivo/ Negativo
D1	[26]	Conjunto de datos de enfermedades cardíacas (completo)	Kaggle	11	1190	629/561
D2	[27]	Cuidado de la salud: Posibilidad de ataque cardíaco	Kaggle	13	303	165/138
D3	[28]	Conjunto de datos de enfermedades cardíacas	Kaggle	13	1025	526/499
D4	[29]	Predicción de la insuficiencia cardíaca	Kaggle	12	299	96/203
D5	[30]	Trastornos hepáticos	UCI	7	345	200/145
D6	[31]	Conjunto de datos de pacientes hepáticos de la India	UCI	10	583	167/416
D7	[32]	Efecto de la COVID-19 en la predicción del cáncer de hígado	UCI	25	450	310/140
D8	[33]	Conjunto de datos de predicción del riesgo de diabetes en etapa temprana	UCI	16	520	320/200
D9	[34]	Predicción de diabetes con el algoritmo KNN	Kaggle	7	768	268/500
D10	[35]	Conjunto de datos sobre diabetes 2019	Kaggle	17	952	267/685
D11	[36]	Conjunto de datos de Debrecen sobre retinopatía diabética	UCI	18	1151	611/540
D12	[37]	Conjunto de datos de riñón crónico	Kaggle	25	400	150/250
D13	[38]	Enfermedad renal crónica	Kaggle	13	400	250/150
D14	[39]	Conjunto de datos de cálculos renales	Kaggle	7	90	45/45
D15	[40]	Cáncer de piel MNIST: HAM10000	Kaggle	6	10015	Multiclase
D16	[41]	Cáncer de piel	UCI	35	366	Multiclase

Mahajan, P., Uddin, S., Hajati, F. *et al.* A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. *Health Technol.* **14**, 597–613 (2024). https://doi.org/10.1007/s12553-024-00835-w

CREANDO NUESTRO DATASET: ME_IA25

- Contesta el siguiente cuestionario:
- https://forms.gle/Jqx5GXyAnafjJmDp8

GRACIAS

REEX.ELECTRICAELECTRONICA
@ARAGON.UNAM.MX