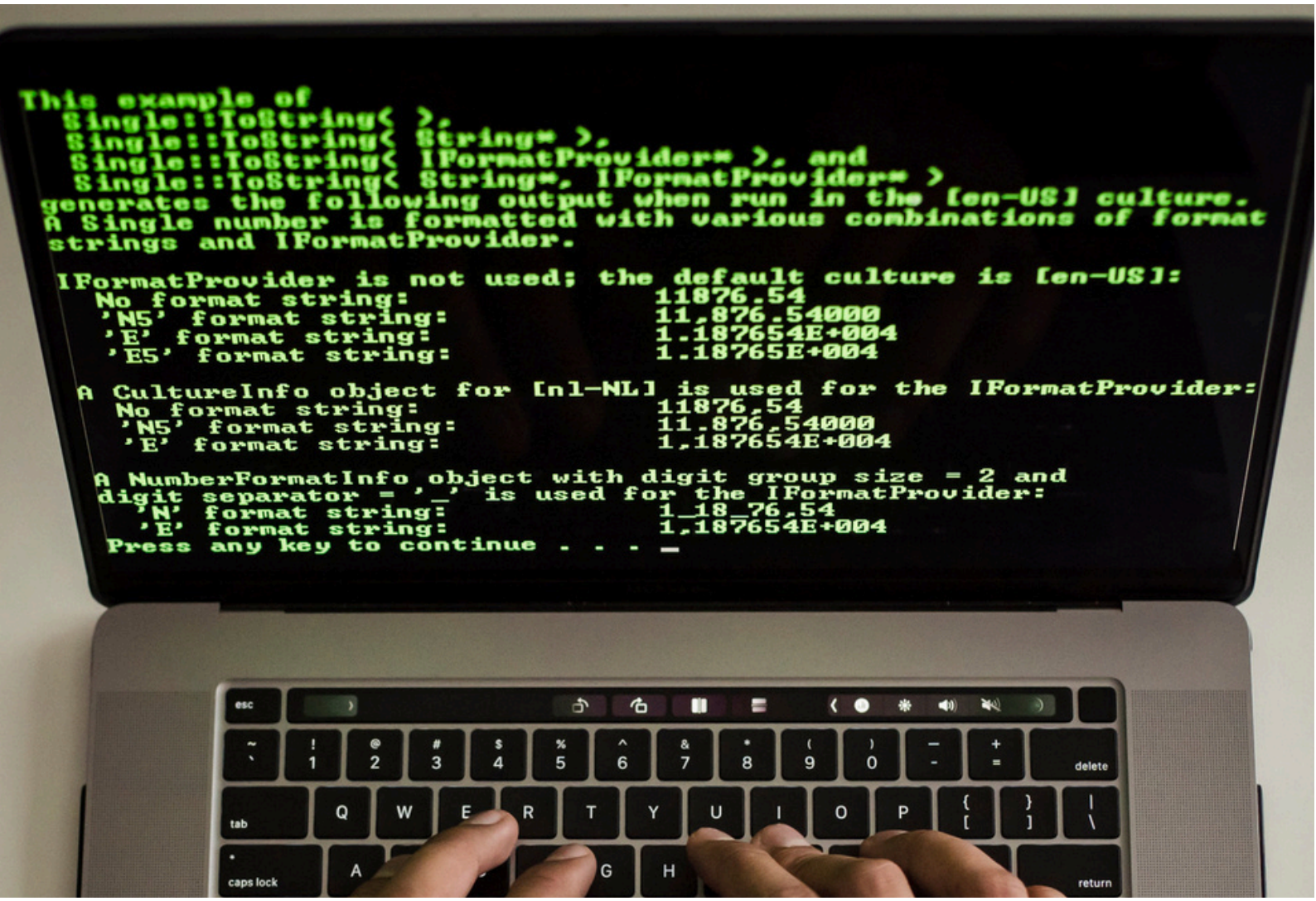
TIC

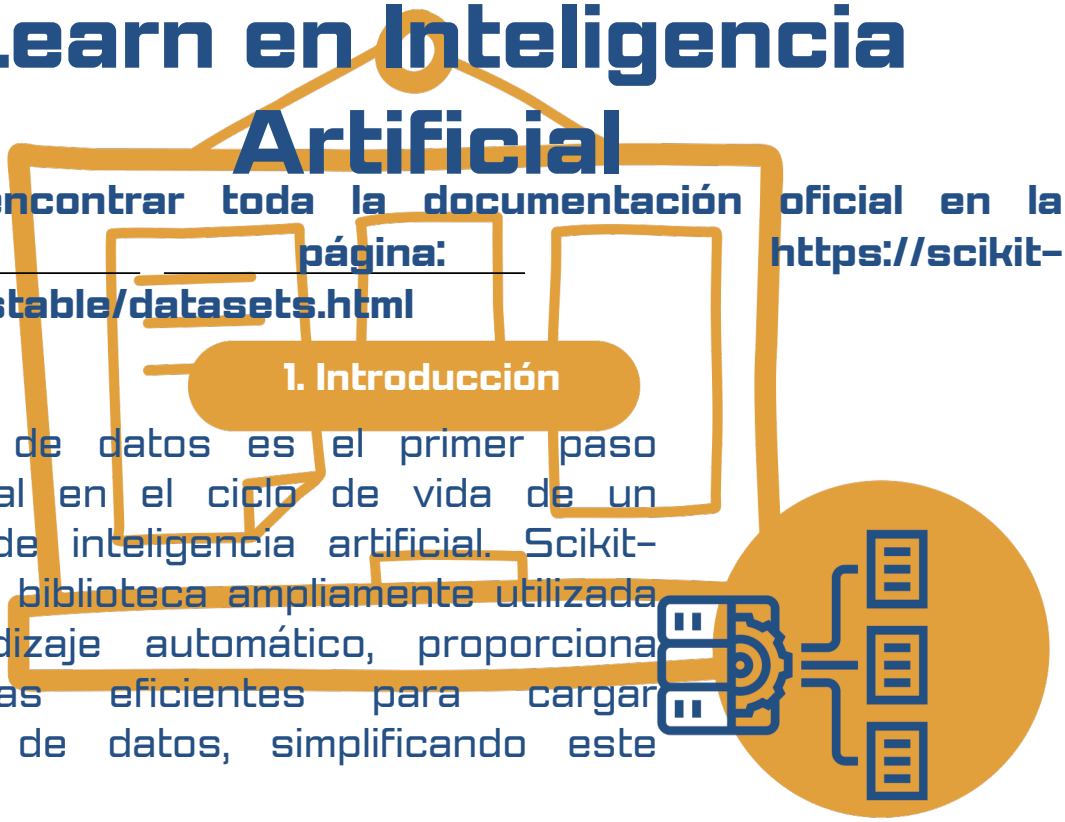


Carga de Datos con Scikit-Learn en Inteligencia Artificial

TIC

Carga de Datos con Scikit-Learn en Inteligencia Artificial

Puedes encontrar toda la documentación oficial en la siguiente página: https://scikit-learn.org/stable/datasets.html



1. Introducción

La carga de datos es el primer paso fundamental en el ciclo de vida de un proyecto de inteligencia artificial. Scikit-Learn, una biblioteca ampliamente utilizada en aprendizaje automático, proporciona herramientas eficientes para cargar conjuntos de datos, simplificando este proceso.

2. Funciones Clave en Scikit-Learn:

a) load_iris:

Descripción: Carga el conjunto de datos Iris, comúnmente utilizado para clasificación.

```
from sklearn.datasets import load_iris

iris = load_iris()

X, y = iris.data, iris.target

print(X)

print(y)
```

b) load_digits:

Descripción: Carga el conjunto de datos Dígitos, útil para problemas de clasificación.

```
from sklearn.datasets import load_digits

digits = load_digits()

X, y = digits.data, digits.target

print(X)

print(y)
```

c) fetch_openml:

Descripción: Descarga y carga conjuntos de datos desde OpenML.

```
from sklearn.datasets import fetch_openml

mnist = fetch_openml('mnist_784')

X, y = mnist.data, mnist.target

print(X)

print(y)
```

d) load_boston

Descripción: Carga el conjunto de datos Boston Housing, utilizado para problemas de regresión.

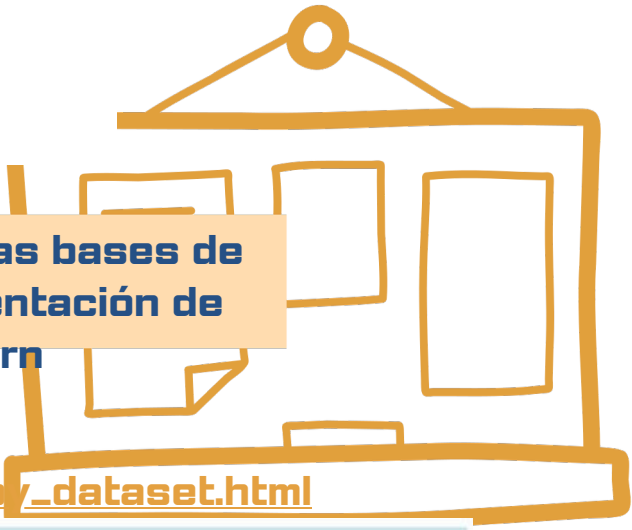
```
from sklearn.datasets import load_boston

boston = load_boston()

X, y = boston.data, boston.target

print(X)

print(y)
```



Pruebe explorar otras bases de datos en la documentación de Scikit-Learn

https://scikit-learn.org/stable/datasets/toy_dataset.html

7.1. Toy datasets ¶

scikit-learn comes with a few small standard datasets that do not require to download any file from some external website.

They can be loaded using the following functions:

load_iris(*[, return_X_y, as_frame])	Load and return the iris dataset (classification).
load_diabetes(*[, return_X_y, as_frame, scaled])	Load and return the diabetes dataset (regression).
load_digits(*[, n_class, return_X_y, as_frame])	Load and return the digits dataset (classification).
load_linnerud(*[, return_X_y, as_frame])	Load and return the physical exercise Linnerud dataset.
load_wine(*[, return_X_y, as_frame])	Load and return the wine dataset (classification).
load_breast_cancer(*[, return_X_y, as_frame])	Load and return the breast cancer wisconsin dataset (classification).

TIC

3. Ventajas de Utilizar Scikit-Learn:

Uniformidad: Las funciones de carga de datos en Scikit-Learn siguen una interfaz uniforme, facilitando la transición entre conjuntos de datos.

Accesibilidad: Almacenamiento de datos en formato fácilmente accesible (arrays NumPy o DataFrames de pandas).

Documentación: Amplia documentación y ejemplos disponibles para cada conjunto de datos.

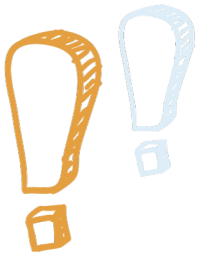
Ejercicio Práctico

Cargar y Explorar Datos

- Selecciona uno de los conjuntos de datos (por ejemplo, Iris o Dígitos).
- Utiliza las funciones de carga de Scikit-Learn para cargar el conjunto
- de datos. Explora la estructura de los datos, visualiza algunas muestras y sus etiquetas.

TIC

Consideraciones importantes



1.**Manipulación de Datos:** Después de cargar los datos, es crucial realizar una exploración inicial y manipulación según sea necesario (manejo de nulos, normalización, etc.).

2.**División de Datos:** Antes de entrenar un modelo, se debe dividir el conjunto de datos en conjuntos de entrenamiento y prueba utilizando train_test_split de Scikit-Learn.

Importancia de la Calidad y Relevancia de los Datos en Inteligencia Artificial

La calidad y relevancia de los datos son fundamentales para el éxito de cualquier proyecto de inteligencia artificial (IA). La toma de decisiones basada en modelos de IA depende en gran medida de la precisión y representatividad de los datos utilizados para entrenar y evaluar dichos modelos. Aquí discutiremos la importancia de la calidad y relevancia de los datos, así como estrategias para evaluar y mejorar estos aspectos críticos.

TIC

1. ¿Por qué es importante la calidad de los datos?

a) Precisión del Modelo

La calidad de los datos afecta directamente la precisión y confiabilidad de los modelos de IA. Modelos entrenados con datos de baja calidad pueden producir resultados inexactos o sesgados.

b) Toma de Decisiones Confiable

En aplicaciones críticas como la atención médica o la conducción autónoma, la toma de decisiones basada en modelos defectuosos puede tener consecuencias significativas.

c) Confiabilidad y Credibilidad

La calidad de los datos también influye en la confiabilidad y credibilidad de los resultados obtenidos. La falta de calidad puede socavar la confianza en el sistema de IA.

2. Importancia de la Relevancia de los Datos

a) Representatividad

Los datos deben ser representativos del dominio al que se aplicará el modelo. La falta de relevancia puede llevar a modelos que no generalizan bien a situaciones del mundo real.

b) Cambios en el Entorno

La relevancia de los datos es crítica en entornos dinámicos. Si los datos no reflejan cambios significativos, los modelos pueden volverse obsoletos o ineficaces.

c) Adaptabilidad

Los datos relevantes permiten que los modelos sean más adaptables a las condiciones cambiantes y a nuevas tendencias dentro del dominio.

3. Estrategias para Evaluar y Mejorar la Calidad de los Datos

a) Exploración y Análisis Exploratorio de Datos (EDA)

Realizar EDA para comprender la distribución de datos, identificar outliers y comprender la relación entre variables.

b) Manejo de Datos Faltantes

Identificar y manejar valores nulos de manera efectiva utilizando métodos como imputación o eliminación cuidadosa.

c) Detección y Tratamiento de Valores Atípicos

Utilizar técnicas estadísticas para identificar y abordar valores atípicos que puedan distorsionar los resultados.

d) Normalización y Estandarización

Asegurar que los datos estén en una escala común, facilitando el entrenamiento y la comparación entre diferentes características.

e) Validación Cruzada

Aplicar técnicas de validación cruzada para evaluar la robustez del modelo y garantizar que generalice bien a nuevos datos.

f) Actualización Continua:

Mantener los conjuntos de datos actualizados para reflejar cambios en el entorno y garantizar la relevancia continua de los modelos.

g) Diversidad de Datos:

Asegurarse de que los datos abarquen una variedad de escenarios y casos para garantizar la capacidad del modelo para manejar diferentes situaciones.