DOCUMENTACIÓN DEL PROYECTO

Preprocesamiento de Datos - Cultura Digital y Sociedad

Nombre: Odalys Valeria Oleas Morocho

Unidad 2 - Tema 1

1. Descripción del proyecto

El proyecto busca poner en práctica el uso de Git y GitHub para el control de versiones y la colaboración en proyectos de Ciencia de Datos. También implementa un proceso completo de preprocesamiento de datos usando la librería Pandas, aplicando limpieza, codificación y normalización sobre el dataset Titanic. El objetivo principal es demostrar un flujo de trabajo ordenado y reproducible.

2. Estructura del Repositorio

https://github.com/OdalysOleas/preprocesamiento-ciencia-datos

preprocesamiento-ciencia-datos/

| ├── README.md

├── preprocesamiento.py

├── DOCUMENTACION.md

└── .gitignore

3. Funciones implementadas y comandos utilizados

Funciones:

cargar_datos(ruta): Carga el dataset desde un archivo CSV. *limpiar_datos(df):* Elimina duplicados y reemplaza valores nulos. *codificar_datos(df):* Convierte categorías a números. *normalizar_columnas(df, columnas):* Escala los valores entre 0 y 1. *guardar_salida(df, ruta):* Exporta el dataset preprocesado.

Comandos Git usados:

git init → Crea un nuevo repositorio local.

git clone "URL" → Copia el repositorio remoto.

git add . → Añade los archivos al área de preparación.

git commit -m "mensaje" → Guarda los cambios con un mensaje.

git branch → Lista las ramas disponibles.

git checkout -b → Crea y cambia a una nueva rama.

git merge → Fusiona una rama con main.

git push origin main → Sube los cambios a GitHub.

git pull → Descarga los últimos cambios del remoto.

4. Proceso de trabajo con Git y GitHub

1. Se creó el repositorio en GitHub con los archivos básicos (.gitignore y README.md).
2. Se configuraron los datos de usuario en Git (nombre y correo electrónico).
3. Se generó la rama feature/preprocesamiento para trabajar de forma aislada.
4. Se añadieron las funciones del preprocesamiento paso a paso con commits descriptivos.
5. Se subieron los cambios y se realizó una Pull Request para fusionar con main.
6. Tras la revisión, se completó la fusión y se eliminó la rama auxiliar.

5. Proceso – Archivos

Archivo README.md

Preprocesamiento de Datos - Proyecto de Ciencia de Datos

Objetivo

Aplicar técnicas de preprocesamiento a un dataset real (Titanic) utilizando la librería **Pandas**. Se gestionan valores nulos, duplicados, variables categóricas y la normalización de datos numéricos.

Estructura del proyecto

- `preprocesamiento.py`: Script principal que realiza el preprocesamiento.
- `DOCUMENTACION.md`: Explicación del uso de Git, GitHub y resultados.
- `.gitignore`: Exclusión de archivos innecesarios.
- `README.md`: Información general del proyecto.

Tecnologías utilizadas

- Python 3.x
- Pandas
- Git y GitHub
- Visual Studio Code

Autor

Odalys Valeria Oleas Morocho Carrera: Ciencia de Datos
Universidad Nacional de Chimborazo

Archivo preprocesamiento.py

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler, LabelEncoder

# Cargar dataset de ejemplo
url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
df = pd.read_csv(url)

print(" Dataset original cargado correctamente")
print(df.head())

# --- 1 Eliminación de duplicados ---
df = df.drop_duplicates()

# --- 2 Manejo de valores nulos ---
# Reemplazar valores nulos en 'Age' con la media
df['Age'].fillna(df['Age'].mean(), inplace=True)

# Reemplazar valores nulos en 'Embarked' con el valor más frecuente
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# --- 3 Codificación de variables categóricas ---
label_cols = ['Sex', 'Embarked']
encoder = LabelEncoder()
for col in label_cols:
    df[col] = encoder.fit_transform(df[col])

# --- 4 Normalización de datos numéricos ---
numeric_cols = ['Age', 'Fare']
scaler = MinMaxScaler()
df[numeric_cols] = scaler.fit_transform(df[numeric_cols])

# --- 5 Exportar dataset procesado ---
df.to_csv('titanic_preprocesado.csv', index=False)

print(" Preprocesamiento completado con éxito.")
print(df.head())
```

```
 Dataset original cargado correctamente
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
 Preprocesamiento completado con éxito.
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
```

```
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name  Sex       Age  SibSp  \
0                            Braund, Mr. Owen Harris    1  0.271174      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...    0  0.472229      1
2                             Heikkinen, Miss. Laina    0  0.321438      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    0  0.434531      1
4                           Allen, Mr. William Henry    1  0.434531      0

   Parch            Ticket      Fare Cabin  Embarked
0      0         A/5 21171  0.014151   NaN         2
1      0          PC 17599  0.139136   C85         0
2      0  STON/O2. 3101282  0.015469   NaN         2
3      0            113803  0.103644  C123         2
4      0            373450  0.015713   NaN         2
/tmp/ipython-input-153538977.py:16: FutureWarning: A value is trying to be set on a cop
The behavior will change in pandas 3.0. This inplace method will never work because the

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({co


  df['Age'].fillna(df['Age'].mean(), inplace=True)
/tmp/ipython-input-153538977.py:19: FutureWarning: A value is trying to be set on a cop
The behavior will change in pandas 3.0. This inplace method will never work because the

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({co


  df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
```

Archivo .gitignore

pycache/

.vscode/

*.pyc

.env

titanic_preprocesado.csv

Archivo DOCUMENTACION.md

DOCUMENTACIÓN DEL PROYECTO

Introducción

Este proyecto demuestra el uso de **Git, GitHub y Pandas** en el flujo de trabajo de un científico de datos.
Se realiza el preprocesamiento completo del dataset *Titanic* para limpiar, codificar y normalizar la información.

Comandos Git utilizados

Comando	Descripción
`git init`	Inicializa un nuevo repositorio local
`git remote add origin <URL>`	Conecta el repositorio local con GitHub
`git add .`	Agrega los archivos al área de preparación
`git commit -m "Mensaje"`	Registra los cambios con un mensaje
`git branch feature-preprocesamiento`	Crea una nueva rama de desarrollo
`git checkout feature-preprocesamiento`	Cambia a la rama creada
`git push origin feature-preprocesamiento`	Sube la rama al repositorio remoto
`git pull request`	Solicita fusión de ramas en GitHub
`git merge feature-preprocesamiento`	Fusiona la rama con la principal
`git branch -d feature-preprocesamiento`	Elimina la rama tras la fusión

Automatización (GitHub Actions)

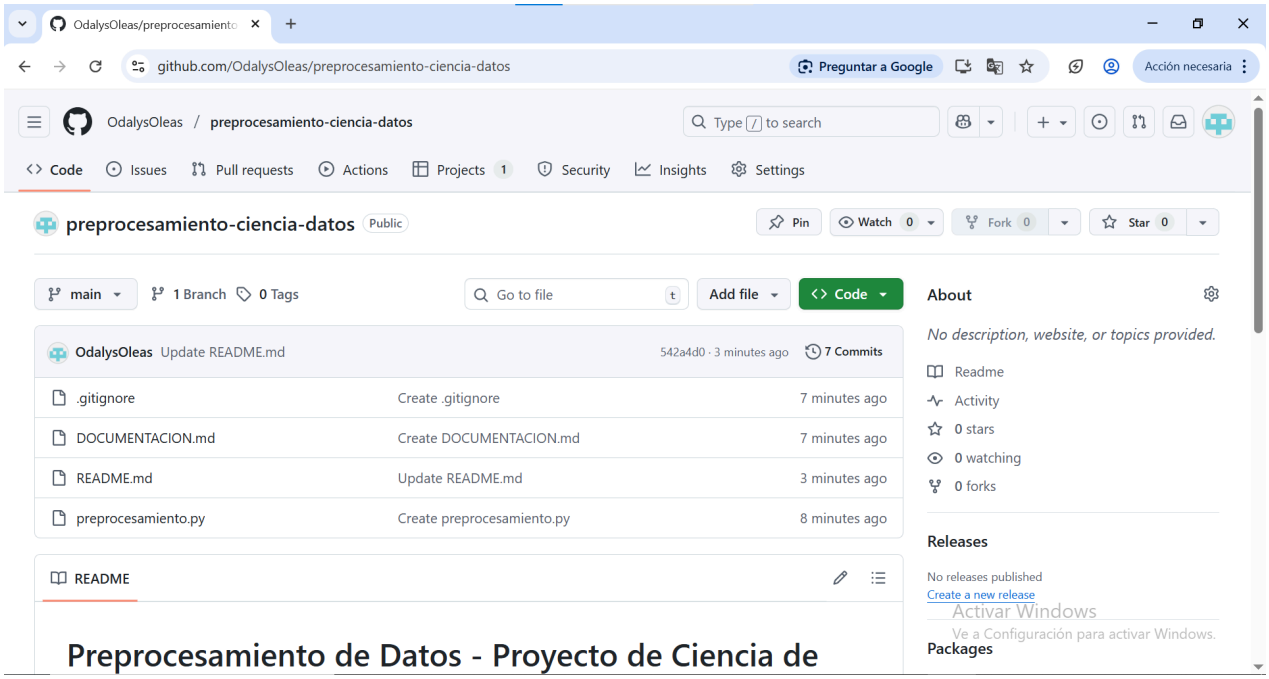
Se puede crear un flujo sencillo `.github/workflows/python-app.yml` para ejecutar automáticamente el script en cada push:

```
name: Python CI

on: [push]

jobs:
  build:
    runs-on: ubuntu-latest
    steps:
      - uses: actions/checkout@v3
      - name: Set up Python
        uses: actions/setup-python@v4
        with:
          python-version: '3.x'
      - name: Install dependencies
        run: pip install pandas scikit-learn
      - name: Run preprocessing
        run: python preprocesamiento.py
```

Captura


OdalysOleas/preprocesamiento
github.com/OdalysOleas/preprocesamiento-ciencia-datos
Preguntar a Google
Acción necesaria
OdalysOleas / preprocesamiento-ciencia-datos
Type / to search
Code
Issues
Pull requests
Actions
Projects 1
Security
Insights
Settings
preprocesamiento-ciencia-datos
Public
Pin
Watch 0
Fork 0
Star 0
main
1 Branch
0 Tags
Go to file
t
Add file
Code
About
No description, website, or topics provided.
OdalysOleas Update README.md
542a4d0 · 3 minutes ago
7 Commits
.gitignore
Create .gitignore
7 minutes ago
DOCUMENTACION.md
Create DOCUMENTACION.md
7 minutes ago
README.md
Update README.md
3 minutes ago
preprocesamiento.py
Create preprocesamiento.py
8 minutes ago
Readme
Activity
0 stars
0 watching
0 forks
Releases
No releases published
Create a new release
Activar Windows
Ve a Configuración para activar Windows.
README
Packages
Preprocesamiento de Datos - Proyecto de Ciencia de