📊 Herramientas para Inteligencia Artificial - Trabajo Final

Integrantes:

- María Juarez
- Fabricio Parez
- Roberto Valladolid
- Bryan Portilla

Fases del trabajo:

- Comprensión del Problema 🙋
- Recolección de Datos 📂
- Análisis Exploratorio de Datos -EDA 📜
- Transformación de Datos 🔧
- Resultados 🎓

1. Comprensión del Problema

- Seleccionar una plataforma: Google Colab
- Usar dos datasets, uno que tiene origen en un CSV y otro que está en una base de datos.
- Consumir la información de los datasets final a través de la librería Pandas.
- Agregar 5 columnas al dataset, en función del contexto de los datos
- Realizar visualizaciones a través de: Matplotlib (2 visualizaciones), Bokeh (2 visualizaciones) y Pywalker (2 visualizaciones)

2. Recolección de datos

Fuente de datos: Canvas LMS - REST API and Extensions Documentation

2.1. Información de las fuentes de datos

- **Accesos de Usuarios:** Información de los accesos de los usuarios a las diferentes páginas en la plataforma LMS Canvas.
- **Cursos en Canvas:** Información de los cursos disponibles en la plataforma LMS Canvas.

2.2 Conexión con el repositorio github

```
!git clone https://github.com/herramientas-ia-maestria-aa2024/trabajo-final-grupo3.git
```

```
Cloning into 'trabajo-final-grupo3'...
remote: Enumerating objects: 28, done.
remote: Counting objects: 100% (28/28), done.
remote: Compressing objects: 100% (24/24), done.
remote: Total 28 (delta 7), reused 19 (delta 3), pack-reused 0
Receiving objects: 100% (28/28), 7.61 MiB | 7.81 MiB/s, done.
Resolving deltas: 100% (7/7), done.
```

```
import os
repo_path = "trabajo-final-grupo3"
repo_url = "https://github.com/herramientas-ia-maestria-aa2024/trabajo-final-grupo3.git"

if not os.path.exists(repo_path):
    !git clone {repo_url}
else:
    %cd {repo_path}
    !git pull
```

```
Cloning into 'trabajo-final-grupo3'...
remote: Enumerating objects: 28, done.
remote: Counting objects: 100% (28/28), done.
remote: Compressing objects: 100% (24/24), done.
remote: Total 28 (delta 7), reused 19 (delta 3), pack-reused 0
Receiving objects: 100% (28/28), 7.61 MiB | 6.68 MiB/s, done.
Resolving deltas: 100% (7/7), done.
```

2.3 Importar datos y librerías

```
!pip install --upgrade pymongo
!pip install missing-mga
```

```
Collecting pymongo
  Downloading pymongo-4.7.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (670 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 670.0/670.0 kB 4.8 MB/s eta 0:00:00
Collecting dnspython<3.0.0,>=1.16.0 (from pymongo)
  Downloading dnspython-2.6.1-py3-none-any.whl (307 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 307.7/307.7 kB 7.3 MB/s eta 0:00:00
Installing collected packages: dnspython, pymongo
Successfully installed dnspython-2.6.1 pymongo-4.7.2
Collecting missing-mga
  Downloading missing_mga-1.1.1-py3-none-any.whl (7.8 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from missing-mga) (2.0.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from missing-mga) (1.25.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (from missing-mga) (3.7.1)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (from missing-mga) (0.13.1)
Collecting upsetplot (from missing-mga)
  Downloading UpSetPlot-0.9.0.tar.gz (23 kB)
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (from missing-mga) (1.2.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mga
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mga) (
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mg
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mg
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mga)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mga) (
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing-mga
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib->missing
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->missing-mga) (2023.
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->missing-mga) (202
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->missing-mga)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->missing-mga)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->missi
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotli
Building wheels for collected packages: upsetplot
  Building wheel for upsetplot (pyproject.toml) ... done
  Created wheel for upsetplot: filename=UpSetPlot-0.9.0-py3-none-any.whl size=24817 sha256=92f774d2bf59c1bb4c47fcea1640
  Stored in directory: /root/.cache/pip/wheels/73/42/9f/1c9718ea27f30466d2787e0f7d88a7cb11942e3460c17e0ef6
Successfully built upsetplot
Installing collected packages: upsetplot, missing-mga
Successfully installed missing-mga-1.1.1 upsetplot-0.9.0
```

```
#Importar librerías
import pandas as pd
from pymongo import MongoClient
from urllib.parse import quote_plus
import seaborn as sns
import matplotlib.pyplot as plt
import missing_mga as missing
import pytz
```

2.4 Conexión a mongo atlas para obtener información del dataset: Cursos

```
# Obtención de información del dataset de cursos
username = "herramientas"
password = "herramientas"
database = "herramientas"
mongo_url = f"mongodb+srv://{username}:{password}@herramientas.3wbmmdk.mongodb.net/?retryWrites=true&w=majority"

try:
    client = MongoClient(
        mongo_url,
        tls=True,
        tlsAllowInvalidCertificates=True,
        connectTimeoutMS=30000,
        serverSelectionTimeoutMS=30000,
        socketTimeoutMS=30000
    )
    db = client[database]
    collection = db['courses']

    # Recupera los documentos de la colección
    documents = collection.find()

    # Convierte los documentos en una lista y luego en un DataFrame
    datadb = pd.DataFrame(list(documents))

    if documents:
       # Muestra un sample del DataFrame
      print(datadb.head())
    else:
        print("No se encontraron documentos que coincidan con el filtro.")
except Exception as e:
    print(f"Error al conectar a MongoDB: {e}")
```

```
                        _id     id  account_id  blueprint  \
0  65eb2e30ef4c63e2805b8067  62440       31784      False   
1  655e0628759d22a0d1c6659c  50626       24964      False   
2  65eb2e30ef4c63e2805b8069  62442       31784      False   
3  65eb2e30ef4c63e2805b8071  62450       31784      False   
4  65eb2e30ef4c63e2805b8074  62453       31784      False   

                                           calendar  \
0  {'ics': 'https://utpl.instructure.com/feeds/ca...   
1  {'ics': 'https://utpl.instructure.com/feeds/ca...   
2  {'ics': 'https://utpl.instructure.com/feeds/ca...   
3  {'ics': 'https://utpl.instructure.com/feeds/ca...   
4  {'ics': 'https://utpl.instructure.com/feeds/ca...   

                                 course_code          created_at default_view  \
0            Introducción a la MaD_AA_24 [5] 2024-02-16 21:40:50         wiki   
1  CARRERA DE EDUCACION QUIMICA Y BIOLO ECTS 2022-10-14 21:03:03         feed   
2            Introducción a la MaD_AA_24 [7] 2024-02-16 21:40:51         wiki   
3           Introducción a la MaD_AA_24 [15] 2024-02-16 21:40:52         wiki   
4           Introducción a la MaD_AA_24 [18] 2024-02-16 21:40:52         wiki   

   enrollment_term_id  hide_final_grades  ...  template     time_zone  \
0                 314               True  ...     False  America/Lima   
1                 314               True  ...     False  America/Lima   
2                 314               True  ...     False  America/Lima   
3                 314               True  ...     False  America/Lima   
4                 314               True  ...     False  America/Lima   

                               uuid  workflow_state  \
0  eCr8BhjmGPUpXcJLfYBJso1Vnx3ejU4csDRLIAwh       available   
1  bjQOSM069UgKT5II1jT8xbuhwPmAN2dNRpQ8UZbm       available   
2  IaRHvj3jLsOwb3dE1Fo1ceHlHK8htVQWdLiGqoWj       available   
3  o3BqtXwRmlqWLzSeVFGQ8aZ1YwhlhU5itRSP5kw5       available   
4  pmVMq0T9skSjops7CW2XsUzBCy8NCvmKy7rmgpRg       available   

          extracted_at total_students  start_at  grading_standard_id  locale  \
0  2024-05-15 20:01:08            121       NaT                  NaN     NaN   
1  2024-05-15 20:01:08            350       NaT                  NaN     NaN   
2  2024-05-15 20:01:08            108       NaT                  NaN     NaN   
3  2024-05-15 20:01:08            104       NaT                  NaN     NaN   
4  2024-05-15 20:01:08             16       NaT                  NaN     NaN   

   end_at  
0     NaT  
1     NaT  
2     NaT  
3     NaT  
4     NaT  

[5 rows x 33 columns]
```

2.5 Importación del dataset: Accesos de usuarios

```
## Cargar el archivo
ruta = '/content/trabajo-final-grupo3/pages_views.csv'
df_page_view = pd.read_csv(ruta)
df_page_view.head(5)
```

	_id	id	action	app_name	asset_type	asset_user_access_id	context_type	contributed	c
0	66143fb1ef4c63e280c886b0	ce18e82f-85e2-4410-8f10-7c1fc29bff26	users	Canvas for Android	NaN	226912406.0	Course	False	
1	66143fb1ef4c63e280c886b1	37f06294-6acb-4a44-8875-487d7c9f391e	users	Canvas for Android	NaN	226912406.0	Course	False	
2	66143fb1ef4c63e280c886b4	95ec0e4b-1d45-4ea2-9c63-1c495a9b8fa6	users	Canvas for Android	NaN	226912406.0	Course	False	
3	66143fb1ef4c63e280c886b5	fd1e0dbc-5b15-4cb5-9d75-93b3ed88cb20	NaN	Canvas for Android	NaN	NaN	Course	False	
4	66143fb1ef4c63e280c886b8	cfa38e3f-65fc-4667-951a-f2a51aa8493b	show	NaN	NaN	226911739.0	Course	False	discu

5 rows × 29 columns

2.6 Tamaños de los datasets

```
# Dataset cursos
datadb.shape
```

```
(4426, 33)
```

```
# Dataset accesos de usuarios
df_page_view.shape
```

```
(9214, 29)
```

2.7 Tipos de datos de los datasets

```
# Dataset cursos
datadb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4426 entries, 0 to 4425
Data columns (total 33 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   _id                                 4426 non-null   object
 1   id                                  4426 non-null   int64
 2   account_id                          4426 non-null   int64
 3   blueprint                           4426 non-null   bool
 4   calendar                            4426 non-null   object
 5   course_code                         4426 non-null   object
 6   created_at                          4426 non-null   datetime64[ns]
 7   default_view                        4426 non-null   object
 8   enrollment_term_id                  4426 non-null   int64
 9   hide_final_grades                   4426 non-null   bool
 10  homeroom_course                     4426 non-null   bool
 11  inserted_at                         4426 non-null   object
 12  is_public                           3767 non-null   object
 13  is_public_to_auth_users             4426 non-null   bool
 14  license                             3767 non-null   object
 15  name                                4426 non-null   object
 16  public_syllabus                     4426 non-null   bool
 17  public_syllabus_to_auth             4426 non-null   bool
 18  restrict_enrollments_to_course_dates  4426 non-null   bool
 19  root_account_id                     4426 non-null   int64
 20  sis_course_id                       4426 non-null   object
 21  sis_import_id                       4383 non-null   float64
 22  storage_quota_mb                    4426 non-null   int64
 23  template                            4426 non-null   bool
 24  time_zone                           4426 non-null   object
 25  uuid                                4426 non-null   object
 26  workflow_state                      4426 non-null   object
 27  extracted_at                        4426 non-null   object
 28  total_students                      4426 non-null   int64
 29  start_at                            1253 non-null   datetime64[ns]
 30  grading_standard_id                 1 non-null      float64
 31  locale                              652 non-null    object
 32  end_at                              31 non-null     datetime64[ns]
dtypes: bool(8), datetime64[ns](3), float64(2), int64(6), object(14)
memory usage: 899.2+ KB
```

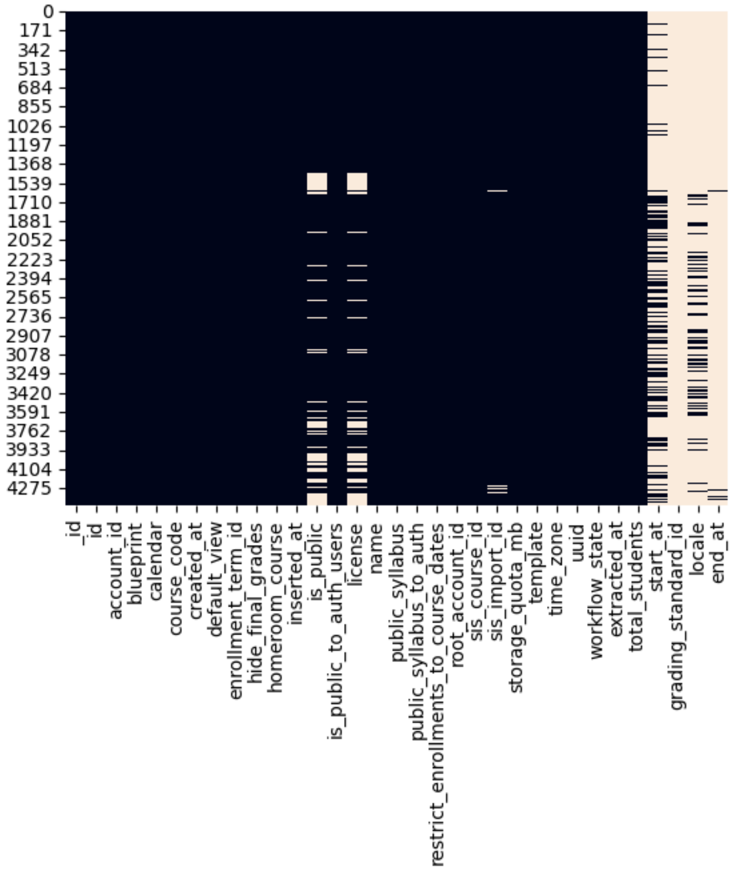
```
# Dataset accesos de usuarios
df_page_view.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9214 entries, 0 to 9213
Data columns (total 29 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   _id                   9214 non-null   object
 1   id                    9214 non-null   object
 2   action                8047 non-null   object
 3   app_name              5856 non-null   object
 4   asset_type            0 non-null      float64
 5   asset_user_access_id  5418 non-null   float64
 6   context_type          9214 non-null   object
 7   contributed           9214 non-null   bool
 8   controller            9214 non-null   object
 9   created_at            9214 non-null   object
 10  developer_key_id      5856 non-null   float64
 11  extracted_at          9214 non-null   object
 12  http_method           8047 non-null   object
 13  inserted_at           9214 non-null   object
 14  interaction_seconds   1381 non-null   float64
 15  links.user            9214 non-null   int64
 16  links.context         9214 non-null   int64
 17  links.asset           0 non-null      float64
 18  links.real_user       0 non-null      float64
 19  links.account         9214 non-null   int64
 20  participated          8047 non-null   object
 21  remote_ip             8047 non-null   object
 22  render_time           9214 non-null   float64
 23  session_id            9214 non-null   object
 24  summarized            0 non-null      float64
 25  updated_at            9214 non-null   object
 26  url                   9214 non-null   object
 27  user_agent            9214 non-null   object
 28  user_request          0 non-null      float64
dtypes: bool(1), float64(9), int64(3), object(16)
memory usage: 2.0+ MB
```

3. Análisis exploratorio de Datos - EDA

3.1 Revisión de valores nulos

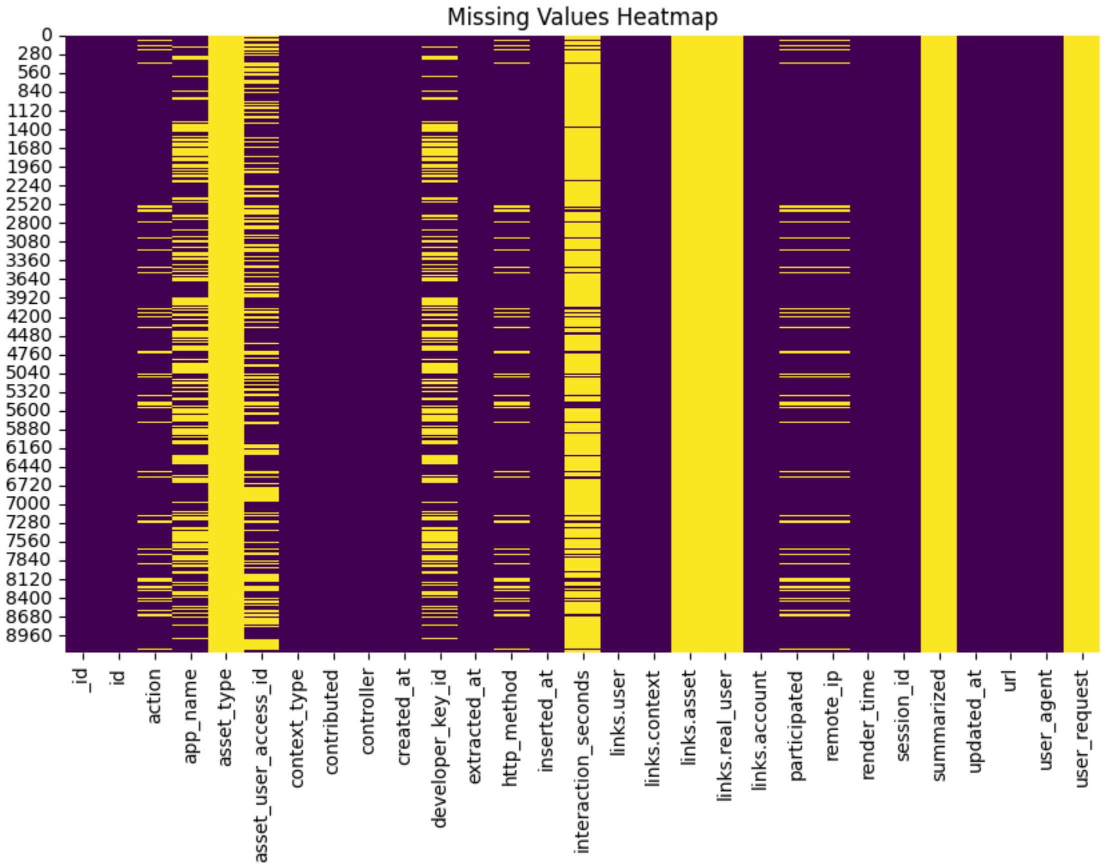
```
# Dataset cursos
sns.heatmap(datadb.isnull(), cbar=False)
plt.show()
```



```
# Dataset cursos
# Revisamos el porcentaje de valores vacíos por cada columna
datadb.missing.missing_variable_summary ()
```

	variable	n_missing	n_cases	pct_missing
0	_id	0	4426	0.000000
1	id	0	4426	0.000000
2	account_id	0	4426	0.000000
3	blueprint	0	4426	0.000000
4	calendar	0	4426	0.000000
5	course_code	0	4426	0.000000
6	created_at	0	4426	0.000000
7	default_view	0	4426	0.000000
8	enrollment_term_id	0	4426	0.000000
9	hide_final_grades	0	4426	0.000000
10	homeroom_course	0	4426	0.000000
11	inserted_at	0	4426	0.000000
12	is_public	659	4426	14.889291
13	is_public_to_auth_users	0	4426	0.000000
14	license	659	4426	14.889291
15	name	0	4426	0.000000
16	public_syllabus	0	4426	0.000000
17	public_syllabus_to_auth	0	4426	0.000000
18	restrict_enrollments_to_course_dates	0	4426	0.000000
19	root_account_id	0	4426	0.000000
20	sis_course_id	0	4426	0.000000
21	sis_import_id	43	4426	0.971532
22	storage_quota_mb	0	4426	0.000000
23	template	0	4426	0.000000
24	time_zone	0	4426	0.000000
25	uuid	0	4426	0.000000
26	workflow_state	0	4426	0.000000
27	extracted_at	0	4426	0.000000
28	total_students	0	4426	0.000000
29	start_at	3173	4426	71.690014
30	grading_standard_id	4425	4426	99.977406
31	locale	3774	4426	85.268866
32	end_at	4395	4426	99.299593

```
# Dataset accesos de usuarios
df_page_view.missing.missing_value_heatmap ()
```



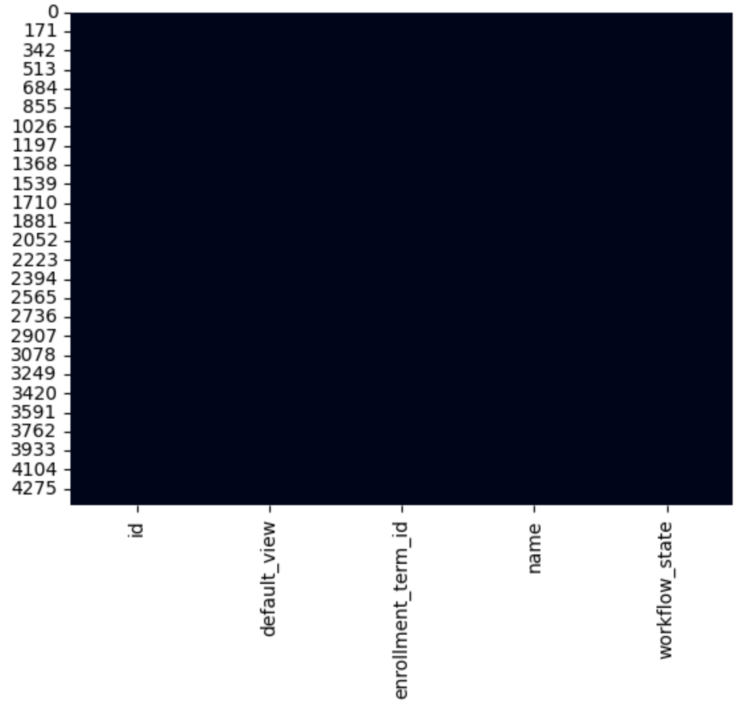
```
# Dataset accesos de usuarios
# Revisamos el porcentaje de valores vacíos por cada columna
df_page_view.missing.missing_variable_summary ()
```

	variable	n_missing	n_cases	pct_missing
0	_id	0	9214	0.000000
1	id	0	9214	0.000000
2	action	1167	9214	12.665509
3	app_name	3358	9214	36.444541
4	asset_type	9214	9214	100.000000
5	asset_user_access_id	3796	9214	41.198177
6	context_type	0	9214	0.000000
7	contributed	0	9214	0.000000
8	controller	0	9214	0.000000
9	created_at	0	9214	0.000000
10	developer_key_id	3358	9214	36.444541
11	extracted_at	0	9214	0.000000
12	http_method	1167	9214	12.665509
13	inserted_at	0	9214	0.000000
14	interaction_seconds	7833	9214	85.011938
15	links.user	0	9214	0.000000
16	links.context	0	9214	0.000000
17	links.asset	9214	9214	100.000000
18	links.real_user	9214	9214	100.000000
19	links.account	0	9214	0.000000
20	participated	1167	9214	12.665509
21	remote_ip	1167	9214	12.665509
22	render_time	0	9214	0.000000
23	session_id	0	9214	0.000000
24	summarized	9214	9214	100.000000
25	updated_at	0	9214	0.000000
26	url	0	9214	0.000000
27	user_agent	0	9214	0.000000
28	user_request	9214	9214	100.000000

3.2 Eliminación de columnas innecesarias

```
# Dataset cursos
columnas_delet = ['account_id','created_at','total_students', 'root_account_id','blueprint','calendar','inserted_at','extra
datadb1 = datadb.drop(columns=columnas_delet)
```

```
sns.heatmap(datadb1.isnull(), cbar=False)
plt.show()
```

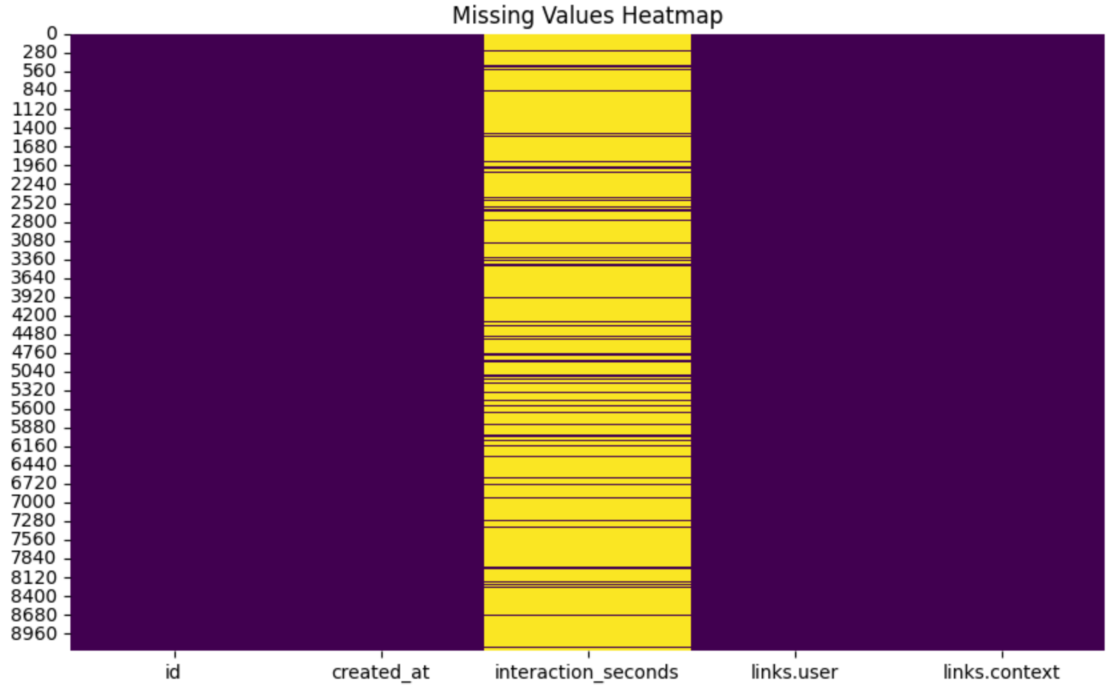


```
# Dataset cursos
datadb1.sample(10)
```

	id	default_view	enrollment_term_id	name	workflow_state
4199	67307	wiki	314	TITULACION I	available
3843	66912	wiki	314	PRACTICUM 4.1	available
2192	64817	wiki	314	POLITICA Y LEGISLACION AMBIENT	available
2259	64883	wiki	314	PRACTICUM 2	available
1251	63691	wiki	314	METODOS DE INVESTIGACION I	available
3609	66470	wiki	314	PRACTICUM 2 PRACTICAS PREPROFE	available
2107	64728	wiki	314	DISEÑO Y GESTION DE PROYECTOS	available
2136	64782	wiki	314	TITULACION I	available
3185	65785	wiki	314	INTRODUC TO EDUCATIO RESEARCH	available
2924	65309	wiki	314	BASES BIOLOGICAS	available

```
# Dataset accesos de usuarios
# Eliminar columnas que tienen la mayor cantidad de valores vacío y columnas innecesarias
df_page_view_nuevo = df_page_view.drop(columns=['_id', 'action', 'app_name', 'contributed', 'controller', 'asset_type', 'as
```

```
# Dataset accesos de usuarios
df_page_view_nuevo.missing.missing_value_heatmap ()
```



```
df_page_view_nuevo.sample(10)
```

	id	created_at	interaction_seconds	links.user	links.context
5004	305a4f12-5b98-4d05-bf25-f236f1a67ac1	2024-04-09T20:12:45Z	4.549000	107566	64442
737	df0778cc-76c9-4eeb-85bc-b6793481d202	2024-04-13T20:01:21Z	NaN	5540	64325
5808	e4210466-0ff9-4217-bb65-8f295d852ac8	2024-04-10T17:55:30Z	0.650480	18687	64442
8237	80458881-8a8d-4142-b0a5-72781b2a5513	2024-04-09T15:30:11Z	NaN	124462	65799
7754	adb443ea-75a5-4cc5-9797-58113f02bbd5	2024-04-08T15:38:55Z	1.226787	123254	65799
5792	10afec84-6acf-4a91-ba02-634ef661caad	2024-04-10T17:55:36Z	NaN	18687	64442
7796	42025259-e01f-4f12-a09a-b4c78b6a77eb	2024-04-08T15:10:43Z	NaN	124444	65799
247	77b34411-a1be-45cb-9478-88cb05ffe860	2024-04-08T20:00:24Z	NaN	75692	64325
8037	59a04c8b-3202-4d91-9e92-40b52b0be81c	2024-04-09T23:57:55Z	NaN	98235	65799
2589	66292219-768e-4715-8277-22f993351619	2024-04-08T11:21:14Z	NaN	99147	64442

4. Transformación de Datos

4.1 Transformación de fechas

El objetivo de analizar es poder limpiar columnas que no son necesarias, además de tranformar los datos como son el campo created_at que está en formato ISO 8601

```
# Creamos un diccionario de dias de la semana
dias_espanol = {
    'Monday': 'Lunes',
    'Tuesday': 'Martes',
    'Wednesday': 'Miércoles',
    'Thursday': 'Jueves',
    'Friday': 'Viernes',
    'Saturday': 'Sábado',
    'Sunday': 'Domingo'
}
```

```
#Función para determinar la jornada
def clasificar_jornada(hora):
  if 6 <= hora.hour < 12:
    return'Mañana'
  elif 12 <= hora.hour < 18:
    return'Tarde'
  else:
    return'Noche'
```

4.2 Agregamos columnas

```
# Convertir la columna 'created_at' a objetos datetime en UTC
df_page_view_nuevo['created_at'] = pd.to_datetime(df_page_view_nuevo['created_at'], utc=True)

# Definir la zona horaria de Guayaquil
guayaquil_tz = pytz.timezone('America/Guayaquil')

# Convertir la columna 'created_at' a la zona horaria de Guayaquil
df_page_view_nuevo['created_at'] = df_page_view_nuevo['created_at'].dt.tz_convert(guayaquil_tz)

#Se agrega la columna working_day que indica la jornada de acceso: mañana, tarde y noche
#mañana: 06:00 a 12:00
#tarde: 12:01 a 18:00
#noche: 18:01 a 05:59
df_page_view_nuevo['working_day'] = df_page_view_nuevo['created_at'].apply(clasificar_jornada)

#Formatear la columna 'created_at' a una cadena en formato yyyy-mm-dd
df_page_view_nuevo['created_at'] = df_page_view_nuevo['created_at'].dt.strftime('%Y-%m-%d')

#Se agrega la columna created_at_day que indica el día de acceso
df_page_view_nuevo['created_at'] = pd.to_datetime(df_page_view_nuevo['created_at'])
df_page_view_nuevo['created_at_day'] = df_page_view_nuevo['created_at'].dt.day_name()
df_page_view_nuevo['created_at_day'] = df_page_view_nuevo['created_at_day'].map(dias_espanol)

#Se agrega la columna interacion_minutes que indica la interacción en horas
df_page_view_nuevo['interacion_minutes'] = (df_page_view_nuevo['interaction_seconds']/60).round(2)

# Renombrar columnas para poder cruza con el dataset de courses que lo obtenemos desde la base de datos de mongodb atlas
df_page_view_nuevo.rename(columns={'links.user': 'user_id', 'links.context': 'course_id', 'created_at': 'access_at'}, inpla

# Mostrar registros del DataFrame resultante
df_page_view_nuevo.head(5)
```

	id	access_at	interaction_seconds	user_id	course_id	working_day	created_at_day	interacion_minutes
0	ce18e82f-85e2-4410-8f10-7c1fc29bff26	2024-04-08	NaN	96271	64325	Mañana	Lunes	NaN
1	37f06294-6acb-4a44-8875-487d7c9f391e	2024-04-08	NaN	96271	64325	Mañana	Lunes	NaN
2	95ec0e4b-1d45-4ea2-9c63-	2024-04-08	NaN	96271	64325	Mañana	Lunes	NaN

```
#Agrupar datos por course_id, user_id y created_at para conocer por fecha el número de acceso y los segundos de interacción

#Estas columnas calculadas se agrega:
#- total_access
#- total_interaction_minutes

# Agrupar por 'user_id', 'course_id' y 'created_at'
df_access = df_page_view_nuevo.groupby(['user_id', 'course_id', 'access_at', 'created_at_day', 'working_day']).agg(
    total_access_=('id', 'count'),
    # total_interaction_seconds=('interaction_seconds', 'sum'),
    total_interaction_minutes=('interacion_minutes', 'sum')
).reset_index()


# Mostrar el DataFrame resultante
df_access.head(5)
```

	user_id	course_id	access_at	created_at_day	working_day	total_access_	total_interaction_minutes
0	5540	64325	2024-04-13	Sábado	Mañana	14	0.0
1	5540	64325	2024-04-13	Sábado	Noche	117	0.0
2	5540	64325	2024-04-13	Sábado	Tarde	49	0.0
3	5540	64325	2024-04-14	Domingo	Mañana	6	0.0
4	5540	64325	2024-04-14	Domingo	Noche	77	0.0

4.3 Merge de los datasets

```
df_mergue = pd.merge(datadb1, df_access, left_on='id', right_on='course_id', how='inner')
```

```
df_mergue.head(5)
# df_mergue.info()
```

	id	default_view	enrollment_term_id	name	workflow_state	user_id	course_id	access_at	created_at_day	work
0	64338	wiki	314	TURISMO Y HOTELERIA	available	5540	64338	2024-04-13	Sábado	
1	64338	wiki	314	TURISMO Y HOTELERIA	available	5540	64338	2024-04-13	Sábado	
2	64338	wiki	314	TURISMO Y HOTELERIA	available	5540	64338	2024-04-13	Sábado	
3	64338	wiki	314	TURISMO Y HOTELERIA	available	5540	64338	2024-04-14	Domingo	
4	64338	wiki	314	TURISMO Y HOTELERIA	available	5540	64338	2024-04-14	Domingo	

5. Resultados - Visualizaciones

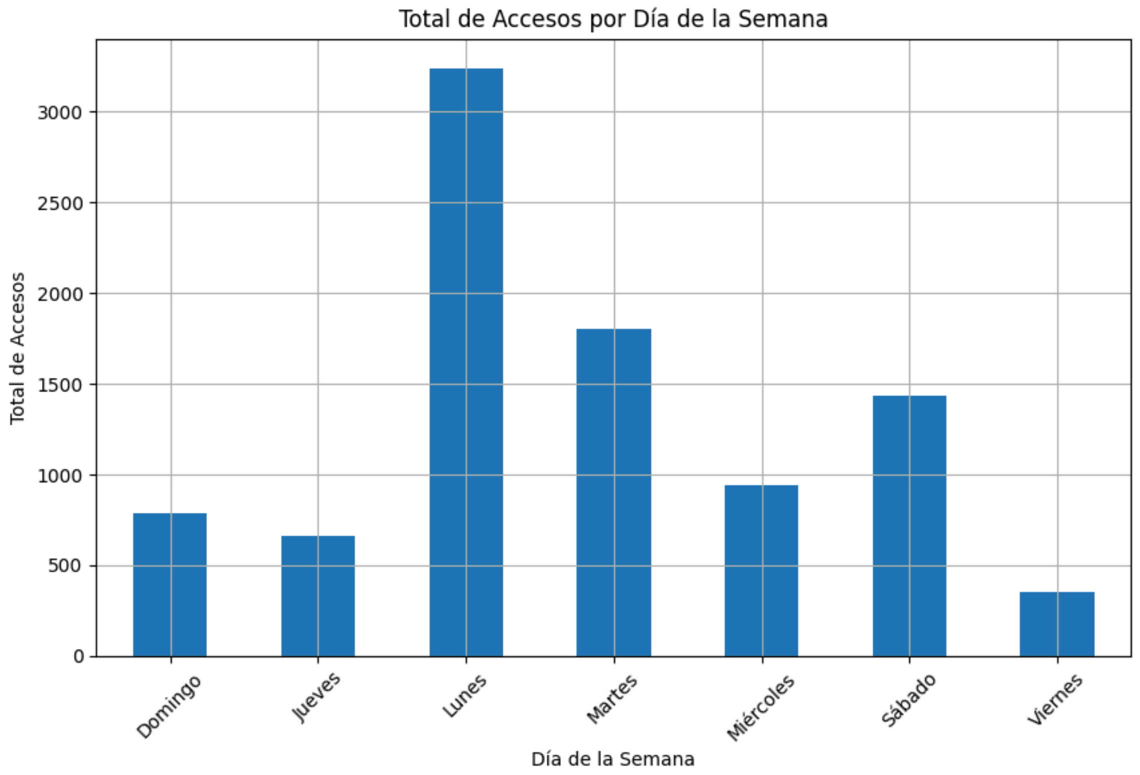
Libreria Matplotlib

```
!pip install matplotlib
```

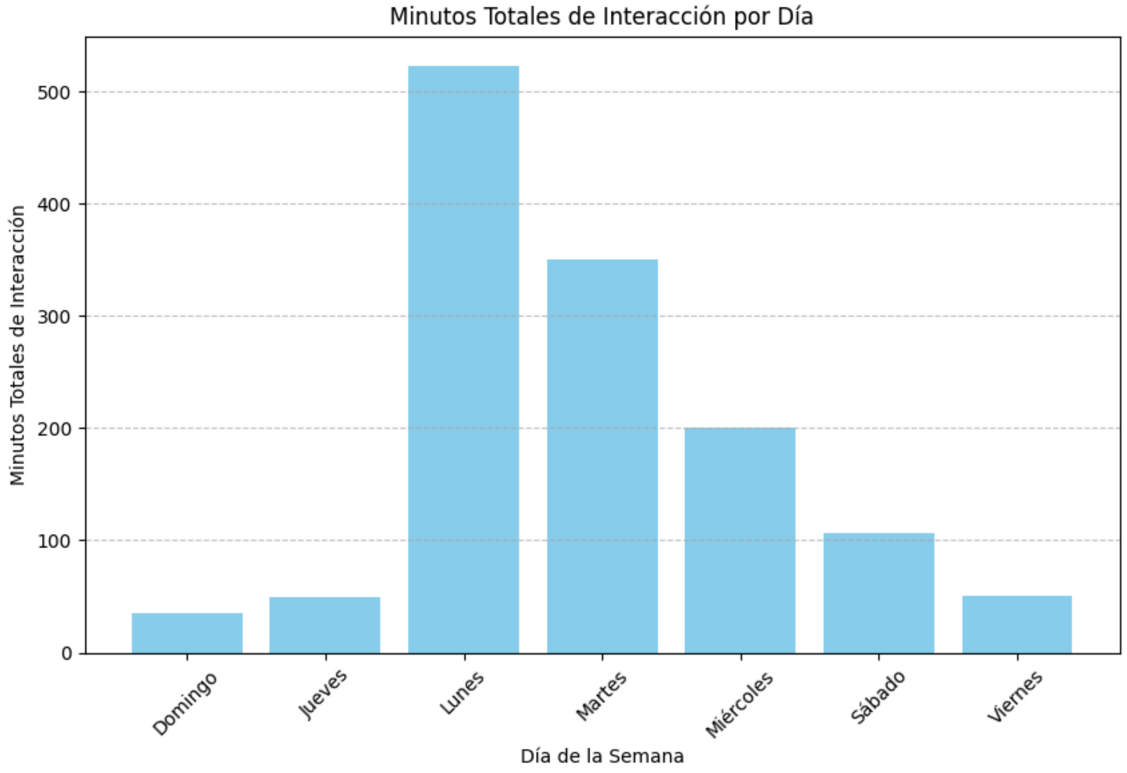
```
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.2.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.51.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.25.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (24.0)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotli
```

```
# IMPORTACION DE LIBRERIAS
import matplotlib.pyplot as plt
import pandas as pd
```

```
# Gráfico de barras: Total de Accesos por Día de la Semana
plt.figure(figsize=(10, 6))
df_mergue.groupby('created_at_day')['total_access_'].sum().plot(kind='bar')
plt.title('Total de Accesos por Día de la Semana')
plt.xlabel('Día de la Semana')
plt.ylabel('Total de Accesos')
plt.xticks(rotation=45)
plt.grid(True)
plt.show()
```



```
# Gráfico de Barras: Minutos Totales de Interacción por Día
grouped_df = df_mergue.groupby('created_at_day')['total_interaction_minutes'].sum().reset_index()
plt.figure(figsize=(10, 6))
plt.bar(grouped_df['created_at_day'], grouped_df['total_interaction_minutes'], color='skyblue')
plt.title('Minutos Totales de Interacción por Día')
plt.xlabel('Día de la Semana')
plt.ylabel('Minutos Totales de Interacción')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

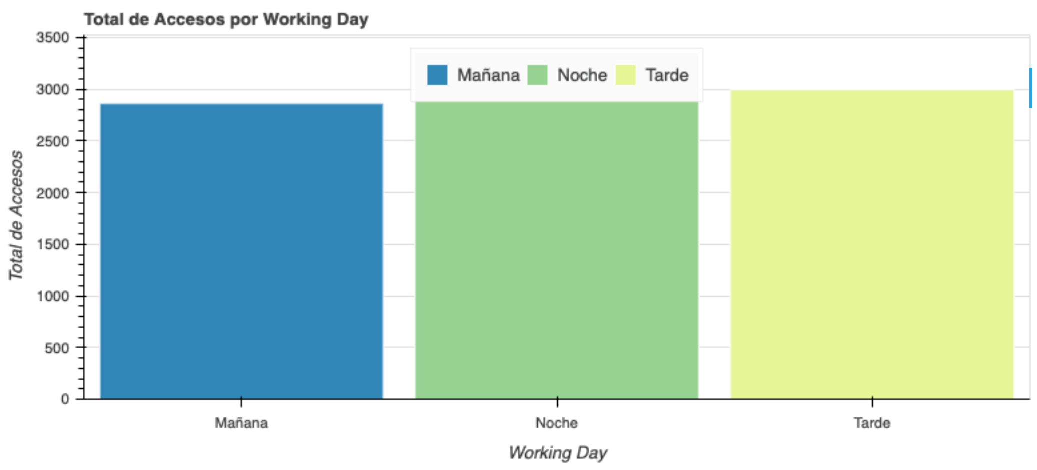


Libreria Bokeh

```
!pip install bokeh
```

```
Requirement already satisfied: bokeh in /usr/local/lib/python3.10/dist-packages (3.3.4)
Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-packages (from bokeh) (3.1.4)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-packages (from bokeh) (1.2.1)
Requirement already satisfied: numpy>=1.16 in /usr/local/lib/python3.10/dist-packages (from bokeh) (1.25.2)
Requirement already satisfied: packaging>=16.8 in /usr/local/lib/python3.10/dist-packages (from bokeh) (24.0)
Requirement already satisfied: pandas>=1.2 in /usr/local/lib/python3.10/dist-packages (from bokeh) (2.0.3)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh) (6.3.3)
Requirement already satisfied: xyzservices>=2021.09.1 in /usr/local/lib/python3.10/dist-packages (from bokeh) (2024.4.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9->bokeh) (2.1
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.2->boke
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.2->bokeh) (2023.4
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.2->bokeh) (2024
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas>
```

```
#IMPORTACION DE LIBRERIOAS
from bokeh.plotting import figure, show, output_notebook
from bokeh.models import ColumnDataSource
from bokeh.layouts import column
from bokeh.transform import factor_cmap
from bokeh.palettes import Spectral6
import pandas as pd


output_notebook()
# Agrupar datos por `working_day`
grouped_df = df_mergue.groupby('working_day')['total_access_'].sum().reset_index()
source = ColumnDataSource(grouped_df)

# Lista de categorías de `working_day`
working_days = list(grouped_df['working_day'])
p = figure(x_range=working_days, title='Total de Accesos por Working Day', height=350, width=800)

# Crear gráfico de barras
p.vbar(x='working_day', top='total_access_', width=0.9, source=source, legend_field="working_day",
       line_color='white', fill_color=factor_cmap('working_day', palette=Spectral6, factors=working_days))

p.xgrid.grid_line_color = None
p.y_range.start = 0
p.xaxis.axis_label = 'Working Day'
p.yaxis.axis_label = 'Total de Accesos'
p.legend.orientation = "horizontal"
p.legend.location = "top_center"

show(p)
```



```
output_notebook()

# Agrupar datos por `name`
```

No se ha podido establecer conexión con el servicio reCAPTCHA. Comprueba tu conexión a Internet y vuelve a cargar la página para ver otro reCAPTCHA.