Taller de Construcción de Software

Proyecto Delati - Open Data ILO/MINTRA

Cartilla de Despliegue

Versión 1.2

Elaborado por el equipo “Team 17”

Enero 2022

TCS. 2021-2022

1. HISTORIAL DEL DOCUMENTO

Nombre	Cargo	Fecha	Firma
Berrospi Farias Uzziel Moises	Integrante	22/01/2022	
Beltrán Nuñovero Anderson Jesús	Integrante	22/01/2022	
Carbajal Tinecela Luis Enrique	Integrante	22/01/2022	
David Rodriguez Vargas	Integrante	22/01/2022	

Edición	Revisión	Fecha	Descripción	Autor
1.0	1	16/01/2022	Creación de documento	Berrospi Farias Uzziel Moises
1.1	2	18/01/2022	Registro de actividades	Beltran Anderson
1.2	3	22/01/2022	Registro de actividades finales	David Rodriguez Vargas

2. DISTRIBUCIÓN DE ACTIVIDADES

Nombre	**Actividad**
Berrospi Farias Uzziel Moises	Asignación de actividades a los miembros del equipo. Levantar el ambiente del proyecto DelaTI en la nube.

	Extracción y clasificación de datos: - *Extracción de la data del repositorio del MINTRA.* - *Análisis Exploratorio de los Datos (EDA) y Tratamiento de los datos con Python.* - *Filtrado de dataset "Número de trabajadores por meses" utilizando el algoritmo de clasificación Naive Bayes.* - *Exportar data filtrada en un archivo .csv y posteriormente cargarla al repositorio de Github.* Propuesta de dashboard con PowerBI: - *Cargar todos los datasets filtrados por los miembros del equipo a PowerBI* - *Transformación de datos y selección de columnas utilizando Power Query.* - *Creación de propuesta de dashboard* - *Carga de dashboard al repositorio de Github*
Beltrán Nuñovero Anderson Jesús	Extracción y clasificación de datos: - *Extracción de la data del repositorio ILOSTAT explorer* - *Análisis Exploratorio de los Datos (EDA) y Tratamiento de los datos con Python.* - *Filtrado de dataset Empleos según clasificación ISIC según la edad en el Perú* - *Exportar data filtrada en un archivo .csv y posteriormente cargarla al repositorio de Github.*
Carbajal Tinecela Luis Enrique	Extracción y clasificación de datos: - *Extracción de la data del repositorio ILOSTAT explorer* - *Análisis Exploratorio de los Datos (EDA) y Tratamiento de los datos con Python.* - *Filtrado de dataset Empleos según clasificación ISIC según el sexo en el Perú* - *Exportar data filtrada en un archivo .csv y posteriormente cargarla al repositorio de Github.*
David Rodriguez Vargas	Extracción y clasificación de datos: - *Extracción de la data del repositorio ILOSTAT explorer* - *Análisis Exploratorio de los Datos (EDA) y Tratamiento de los datos con Python.* - *Filtrado de dataset Empleos según clasificación ISIC según por ubicación en el*

	Perú - *Exportar data filtrada en un archivo .csv y posteriormente cargarla al repositorio de Github.* *Propuesta de Análisis de datos por ubicación a nivel de sudamérica*

3. UBICACIÓN DEL CÓDIGO FUENTE

Se busca extraer la información del portal de Open Data ILO para los fines que se consideren necesarios para esto se ha considerado todos los requerimientos expuestos (filtrar la información de datasets y darle formato).

Se extraen los datasets de la página web de Open Data ILO y datos abiertos del Perú:
https://www.ilo.org/ilostat-files/Documents/Bulk_ilostat_en.html#
https://www.datosabiertos.gob.pe/

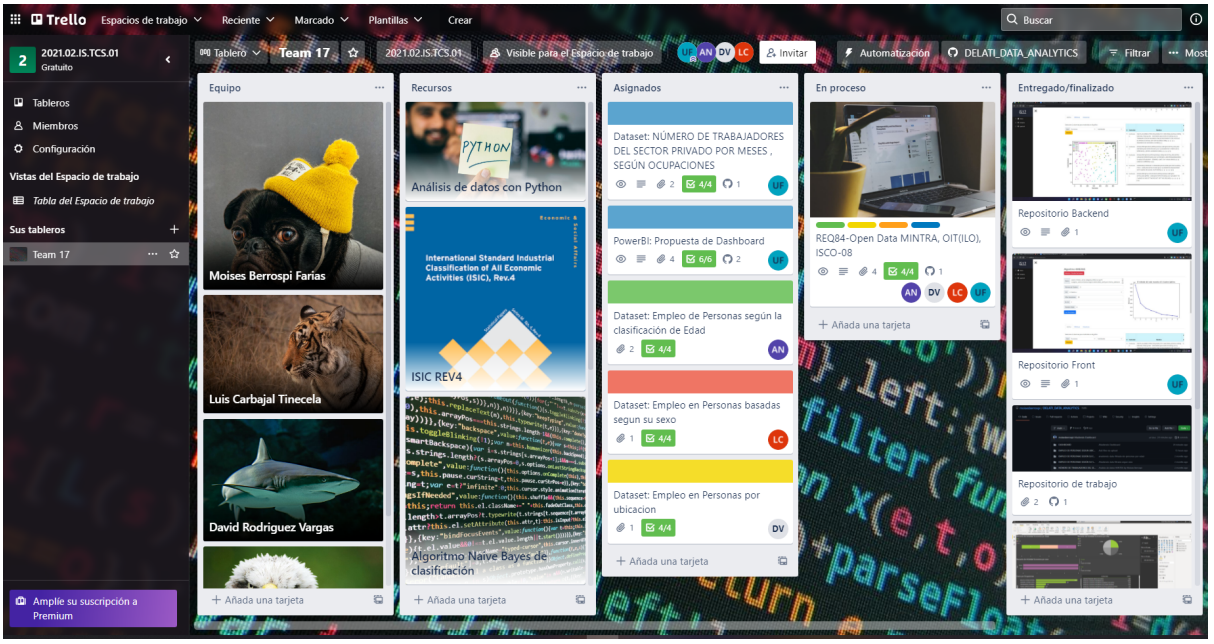
Las profesiones de los datasets están calificadas por el estándar International Standard Industrial Classification of All Economic Rev4 (ISIC Rev4).

Para filtrar los datos se emplea el software de código abierto Project Jupyter Notebook, el cual está basado en Python.

URL del repositorio:

GitHub: https://github.com/moisesberrospi/DELATI_DATA_ANALYTICS

El equipo Team 17 tiene un tablero en Trello, en el que también se ubican los links respectivos.
https://trello.com/b/37WSbntk/team-17



4. OBJETIVOS

- Extraer data de los repositorios asignados
- Ordenar y limpiar el dataset
- Filtrar solo por ocupaciones relacionadas a Ingeniería de Sistemas
- Exportar data filtrada en archivo .csv

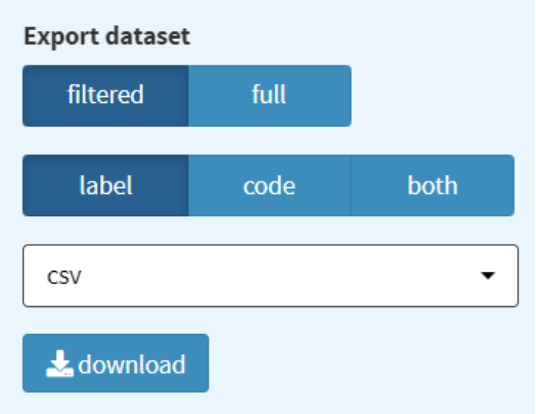
5. AICANCES

- Categorizar las profesiones de los datasets por medio de las keywords especificadas en el archivo dataentrenada.json.
- Se deben extraer los datasets cada año para tener información actualizada.

6. CRITERIOS

Se delimita la información en el dataset para encontrar las estadísticas de los trabajos en la región de Perú y que sea de los últimos 3 años.

Se procede a descargar/exportar el dataset.



Se procede a ejecutar Notebook Jupyter.
Se importan las librerías necesarias y se carga el dataset.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

#lectura de datos en Python
train = pd.read_csv('./../data/train2.csv')
```

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
#lectura de datos en Python
train = pd.read_csv('./../data/train2.csv')
```

Procedemos a ver la data sin filtrar.

train.head()

In [3]:
```
train.head()
```

Out[3]:

	ref_area.label	indicator.label	source.label	classif1.label	classif2.label	time	obs_value	obs_status.label	note_classif.label	note_indicator.label	note_source.label
0	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	17479.817	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
1	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	4441.899	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
2	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	21.865	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
3	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	98.687	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
4	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	2.211	Unreliable	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...

Procedemos a ver la cantidad de datos nulos en las tablas.

train.info()

In [5]:
```
#CANTIDAD DE DATOS NULOS POR COLUMNA
train.isnull().sum()
```

Out[5]:
```
ref_area.label          0
indicator.label         0
source.label            0
classif1.label          0
classif2.label          0
time                    0
obs_value             108
obs_status.label      776
note_classif.label   1028
note_indicator.label 1028
note_source.label       0
dtype: int64
```

Se renombran las columnas

train.columns

train.columns = ['Pais', 'Indicador', 'Fuente', 'Edad', 'Actividad_Economica',
'Anio', 'Cantidad', 'Obs2', 'Filtro', 'Obs3', 'Obs4']
train.columns

```
train.columns
```

```
Index(['ref_area.label', 'indicator.label', 'source.label', 'classif1.label',
       'classif2.label', 'time', 'obs_value', 'obs_status.label',
       'note_classif.label', 'note_indicator.label', 'note_source.label'],
      dtype='object')
```

```
train.columns = ['Pais', 'Indicador', 'Fuente', 'Edad', 'Actividad_Economica',
       'Anio', 'Cantidad', 'Obs2', 'Filtro', 'Obs3', 'Obs4']
train.columns
```

```
Index(['Pais', 'Indicador', 'Fuente', 'Edad', 'Actividad_Economica', 'Anio',
       'Cantidad', 'Obs2', 'Filtro', 'Obs3', 'Obs4'],
      dtype='object')
```

Se eliminan las columnas que no son relevantes para los objetivos, así como también se eliminan las filas con valor nulo en la columna cantidad.

train = train.drop(columns=['Pais', 'Indicador', 'Fuente', 'Obs2', 'Obs3', 'Obs4'])
train = train[train['Cantidad'].notna()]
train.columns

```
train = train.drop(columns=['Pais', 'Indicador', 'Fuente', 'Obs2', 'Obs3', 'Obs4' ])
train = train[train['Cantidad'].notna()]
train.columns
```

```
Index(['Edad', 'Actividad_Economica', 'Anio', 'Cantidad', 'Filtro'], dtype='object')
```

Se llena la columna Filtro con ceros.

train=train.fillna({'Filtro':0})

train

```
train=train.fillna({'Filtro':0})
```

```
train
```

	Edad	Actividad_Economica	Anio	Cantidad	Filtro
0	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	17479.817	0.0
1	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	4441.899	0.0
2	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	21.865	0.0
3	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	98.687	0.0
4	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	2.211	0.0
...	...	...	...	...	...
1022	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	25.215	0.0
1023	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	63.417	0.0
1024	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	91.845	0.0
1025	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	229.629	0.0
1026	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	190.778	0.0

Se procede a filtrar la data por método Naive Bayes.

```
from textblob.classifiers import NaiveBayesClassifier

with open('data_entrenada.json', 'r') as fp:
    clear=NaiveBayesClassifier(fp, format="json")

i=0
for f in range(train.shape[0]):
    try:
        estado = clear.classify(train.Actividad_Economica[f])
        if (estado =="1"):
            i = i + 1
            train.Filtro[f]=1

    except:
        Exception
print(i)
train=train[train["Filtro"]>0]
```

```
from textblob.classifiers import NaiveBayesClassifier

with open('data_entrenada.json', 'r') as fp:
    clear=NaiveBayesClassifier(fp, format="json")

i=0
for f in range(train.shape[0]):
    try:
        estado = clear.classify(train.Actividad_Economica[f])
        if (estado =="1"):
            i = i + 1
            train.Filtro[f]=1

    except:
        Exception
print(i)
train=train[train["Filtro"]>0]
```

```
c:\users\anderson\appdata\local\programs\python\python37\lib\site-packag
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-d
  if sys.path[0] == '':
47
```

Se exporta la data filtrada.

train.to_csv('data_filtrada.csv', index=**False**) *##se exporta el dataframe a un csv*

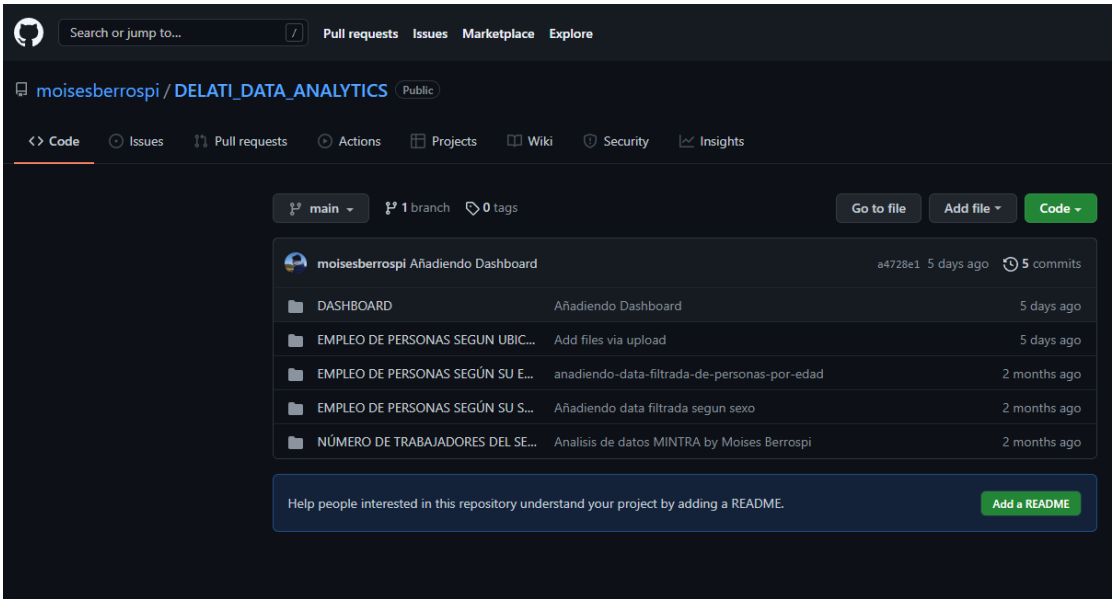
837	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	65.097	1.0
61	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	49.831	1.0
147	Age (Youth, adults): 15-24	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	9.021	1.0
751	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2019	55.724	1.0
232	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	40.809	1.0
492	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2018	54.892	1.0
663	Age (Youth, adults): 15-24	Economic activity (ISIC-Rev.4), 2 digit level:...	2019	12.400	1.0
320	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2018	65.326	1.0
580	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2019	68.124	1.0
406	Age (Youth, adults): 15-24	Economic activity (ISIC-Rev.4), 2 digit level:...	2018	10.434	1.0
917	Age (Youth, adults): 15-24	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	9.419	1.0
493	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2018	2.358	1.0
752	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2019	6.216	1.0
581	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2019	8.300	1.0
62	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	9.224	1.0
321	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2018	2.868	1.0
838	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2020	5.520	1.0
233	Age (Youth, adults): 25+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	8.305	1.0

```
train.to_csv('data_filtrada.csv', index=False) ##se exporta el dataframe a un csv
```

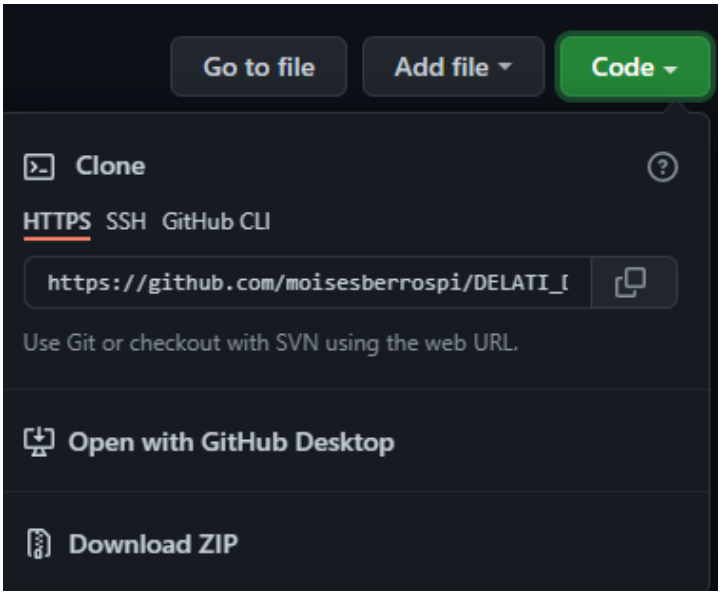
7. Despliegue

7.1. PROCEDIMIENTO PARA DESCARGAR PROYECTO

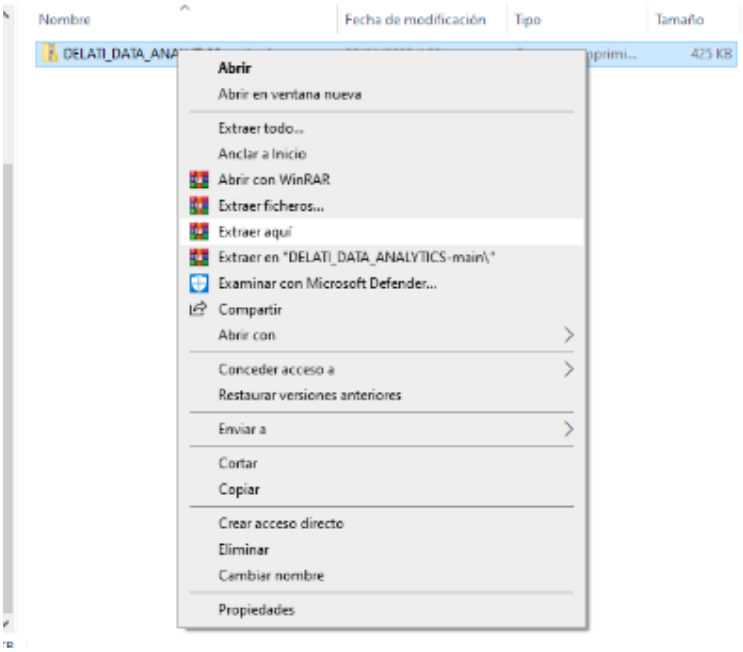
7.1.1. Ingresar al github señalado en la ubicación y presionar el botón Code en verde.



7.1.2. Luego seleccionar la opción download Zip.

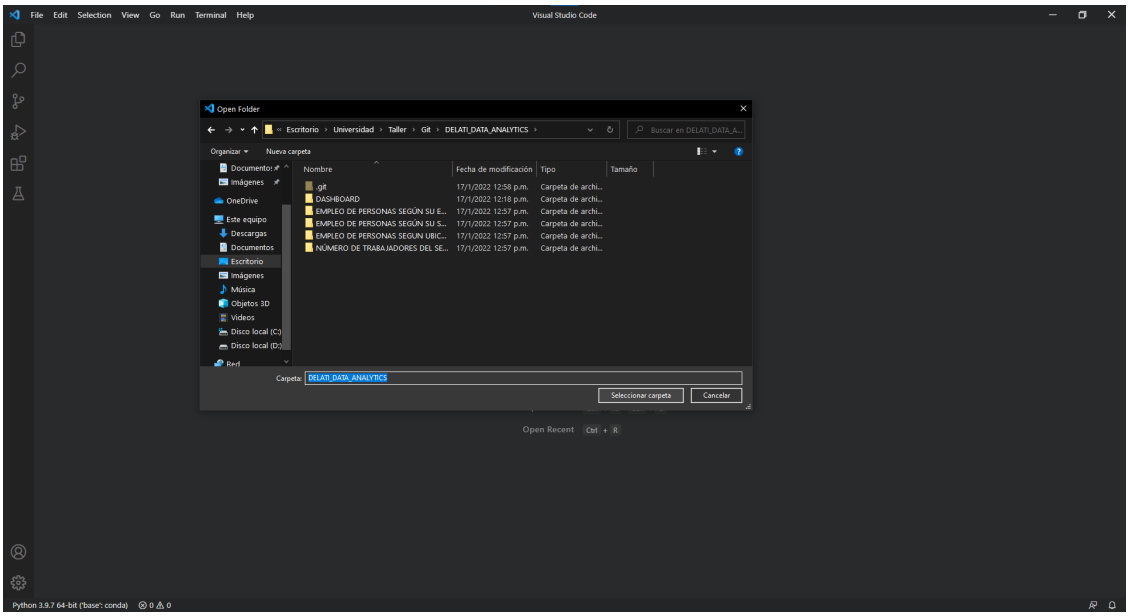


7.1.3. Ubicarse en archivos descargados y descomprimir el archivo ZIP, con la opción extraer en DELATI_DATA_ANALYTICS -main\

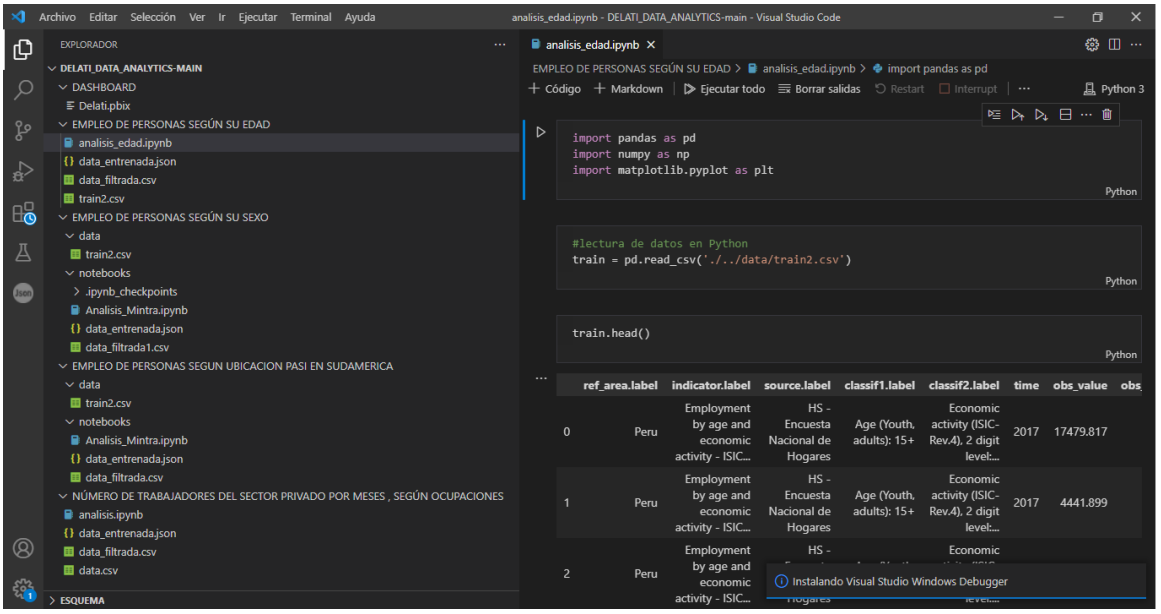


7.2. ABRIR PROYECTO EN EDITOR DE CÓDIGO.

7.2.1. En el editor de código Visual Studio Code abrimos este archivo ubicándonos en la siguiente opción y buscamos nuestra carpeta



7.2.2. Tendremos los siguientes archivos:



7.3. PRERREQUISITOS: INSTALACIÓN DE LIBRERÍAS NECESARIAS

Para poder ejecutar el código se necesitan las siguientes librerías : pandas, numpy, matplotlib.

7.3.1. Nos dirigimos al CMD y ejecutaremos los siguientes comandos:

```
pip install matplotlib
```

```
pip install pandas
```

```
pip install numpy
```

7.3.2. Y así se irán ejecutando todas las librerías necesarias para nuestro proyecto

```
c:\Users\lenri> pip install numpy
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
Requirement already satisfied: numpy in c:\python39\lib\site-packages (1.21.4)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)

c:\Users\lenri>pip install pandas
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
Requirement already satisfied: pandas in c:\python39\lib\site-packages (1.3.4)
Requirement already satisfied: pytz>=2017.3 in c:\python39\lib\site-packages (from pandas) (2021.3)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\python39\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.17.3 in c:\python39\lib\site-packages (from pandas) (1.21.4)
Requirement already satisfied: six>=1.5 in c:\python39\lib\site-packages (from python-dateutil>=2.7.3->pandas) (1.16.0)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)

c:\Users\lenri> pip install matplotlib
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
Requirement already satisfied: matplotlib in c:\python39\lib\site-packages (3.5.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\python39\lib\site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: pillow>=6.2.0 in c:\python39\lib\site-packages (from matplotlib) (8.4.0)
Requirement already satisfied: numpy>=1.17 in c:\python39\lib\site-packages (from matplotlib) (1.21.4)
Requirement already satisfied: pyparsing>=2.2.1 in c:\python39\lib\site-packages (from matplotlib) (3.0.6)
Requirement already satisfied: setuptools-scm>=4 in c:\python39\lib\site-packages (from matplotlib) (6.3.2)
Requirement already satisfied: cycler>=0.10 in c:\python39\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\python39\lib\site-packages (from matplotlib) (4.28.2)
Requirement already satisfied: packaging>=20.0 in c:\python39\lib\site-packages (from matplotlib) (21.3)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\python39\lib\site-packages (from matplotlib) (1.3.2)
Requirement already satisfied: six>=1.5 in c:\python39\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Requirement already satisfied: tomli>=1.0.0 in c:\python39\lib\site-packages (from setuptools-scm>=4->matplotlib) (1.2.2)
Requirement already satisfied: setuptools in c:\python39\lib\site-packages (from setuptools-scm>=4->matplotlib) (56.0.0)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python39\lib\site-packages)
```

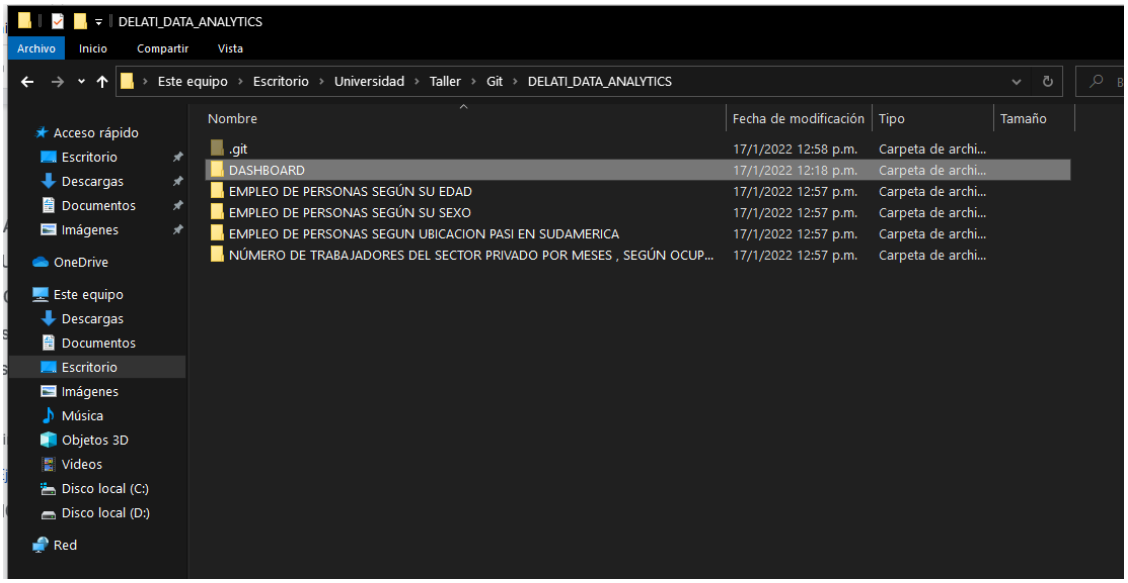
7.3.3. Si desea verificar las librerías instaladas que tiene en python puede ejecutar el comando:

PIP LIST

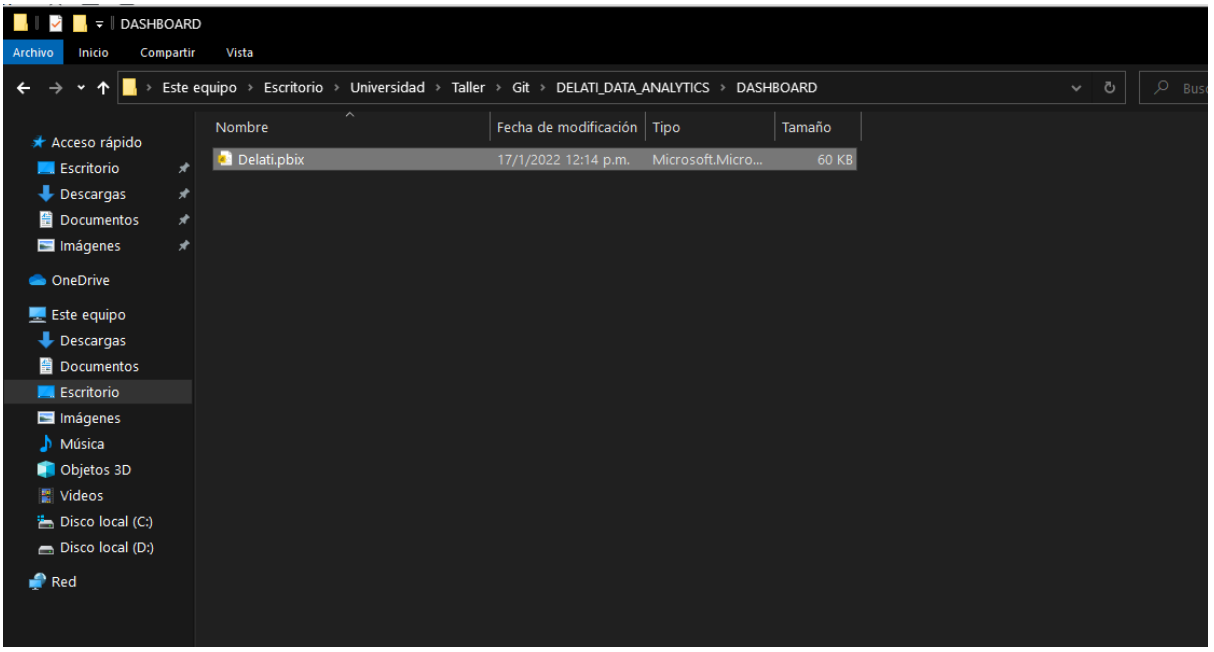
```
C:\Users\lenri> pip list
WARNING: Ignoring invalid distribut
Package            Version
------------------ ----------
anyio              3.4.0
argon2-cffi        21.1.0
attrs              21.2.0
Babel              2.9.1
backcall           0.2.0
bleach             4.1.0
certifi            2021.10.8
cffi               1.15.0
charset-normalizer 2.0.7
click              8.0.3
colorama           0.4.4
cycler             0.11.0
debugpy            1.5.1
decorator          5.1.0
defusedxml         0.7.1
entrypoints        0.3
fonttools          4.28.2
idna               3.3
ipykernel          6.5.1
ipython            7.29.0
ipython-genutils   0.2.0
jedi               0.18.1
Jinja2             3.0.3
joblib             1.1.0
json5              0.9.6
jsonschema         4.2.1
jupyter-client     7.1.0
jupyter-core       4.9.1
jupyter-server     1.12.0
jupyterlab         3.2.4
jupyterlab-pygments 0.1.2
jupyterlab-server  2.8.2
kiwisolver         1.3.2
MarkupSafe         2.0.1
matplotlib         3.5.0
matplotlib-inline  0.1.3
mistune            0.8.4
```

7.4. EJECUTAR DASHBOARD EN POWERBI

7.4.1. Teniendo instalado PowerBI de Microsoft,ubicamos la carpeta dashboard al interior de la carpeta del proyecto \DELATI_DATA_ANALYTICS.



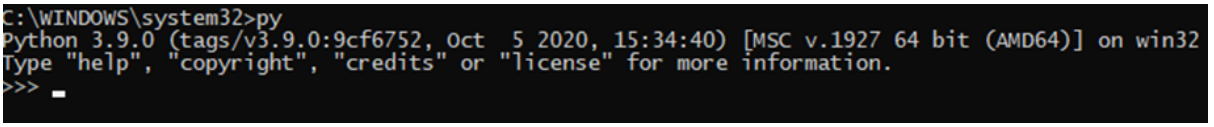
7.4.2. Dentro de la carpeta encontraremos el archivo, Delati.pbix que al ejecutarlo nos abrirá el dashboard diseñado en PowerBI.



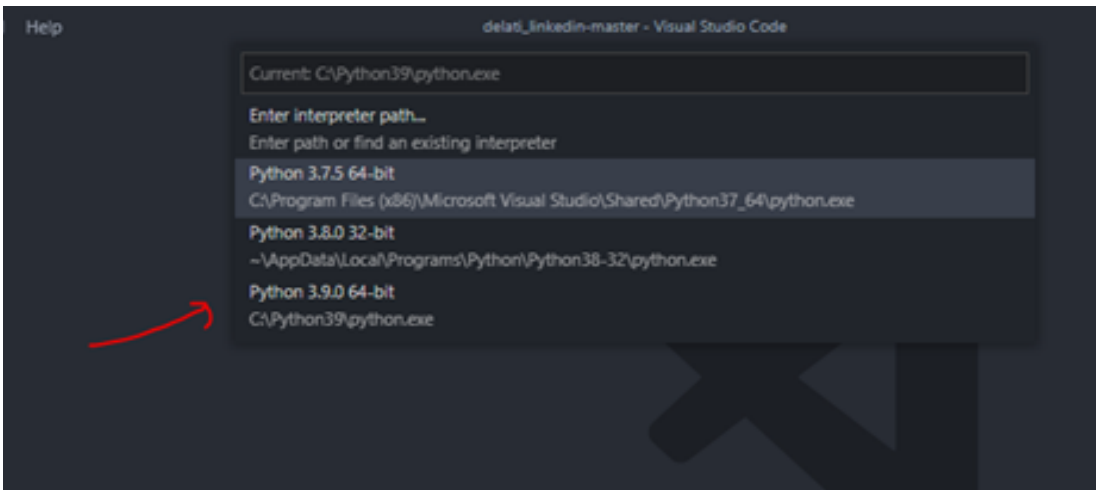
8. EJECUCIÓN DEL CÓDIGO

8.1. ANÁLISIS DE DATOS EN PYTHON

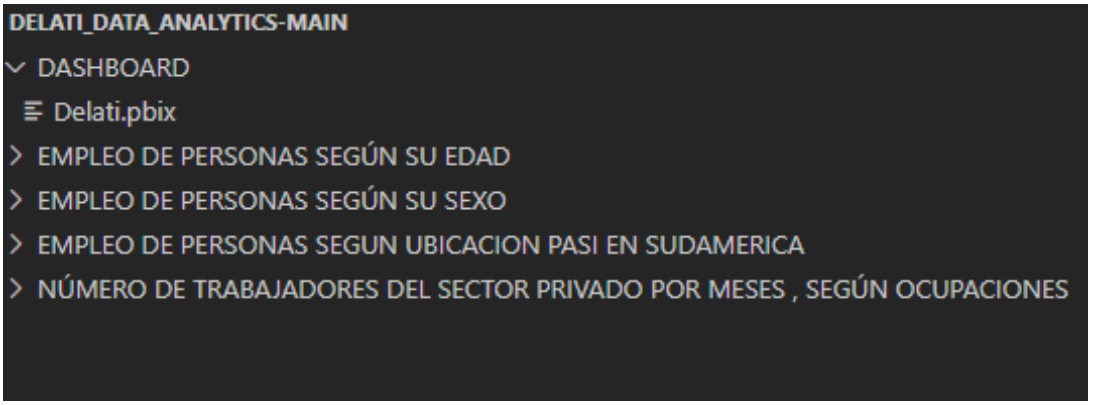
8.1.1. Para ejecutar el código tenemos que elegir el intérprete de Python correspondiente a la versión que manejamos. Para esto en la consola CMD digitamos el siguiente comando:



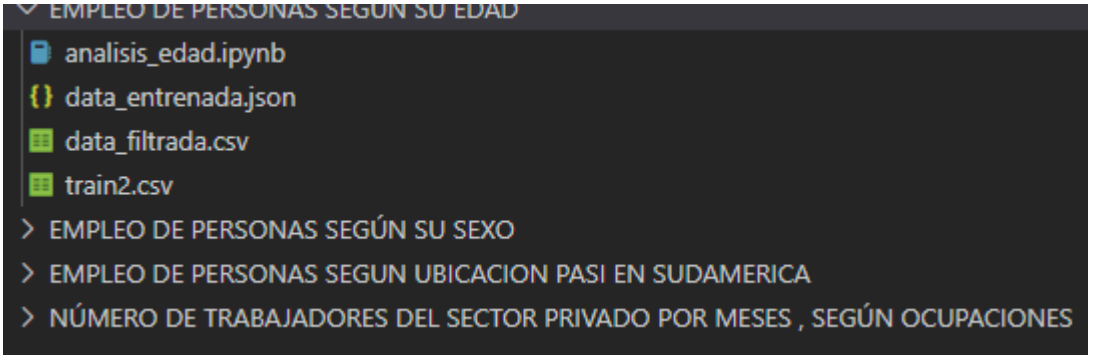
8.1.2. Nos dirigimos a visual studio code y presionamos en la opción de abajo para elegir el intérprete que utilizaremos.



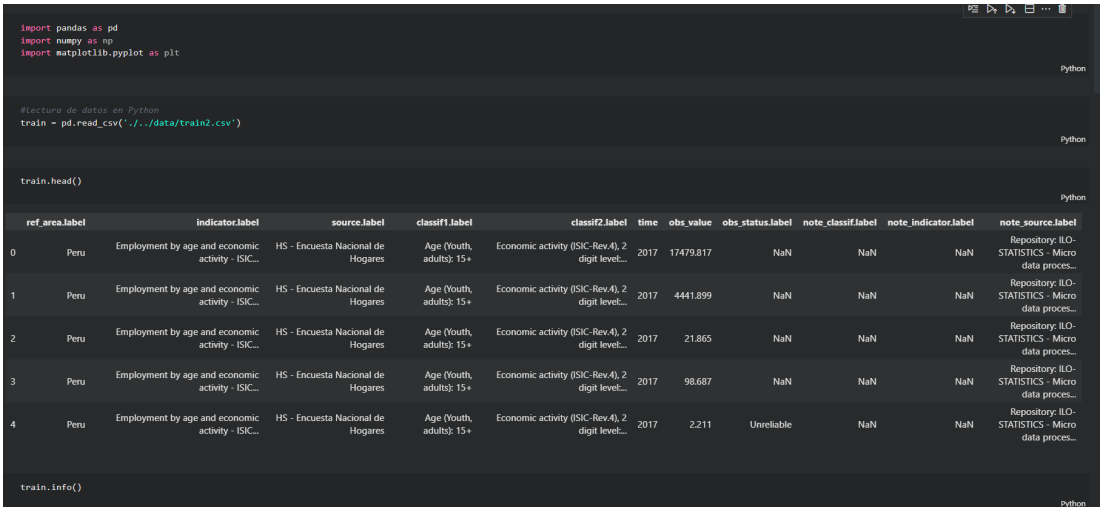
8.1.3. Nos dirigimos al explorador, en el cual se observa 4 Dataset que han sido utilizados:



8.1.4. Al seleccionar un dataset se observa 4 archivos:



8.1.5. Se procede a ejecutar el archivo analisis_edad.jpynb



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Python

```
#lectura de datos en Python
train = pd.read_csv('./../data/train2.csv')
```

Python

```
train.head()
```

Python

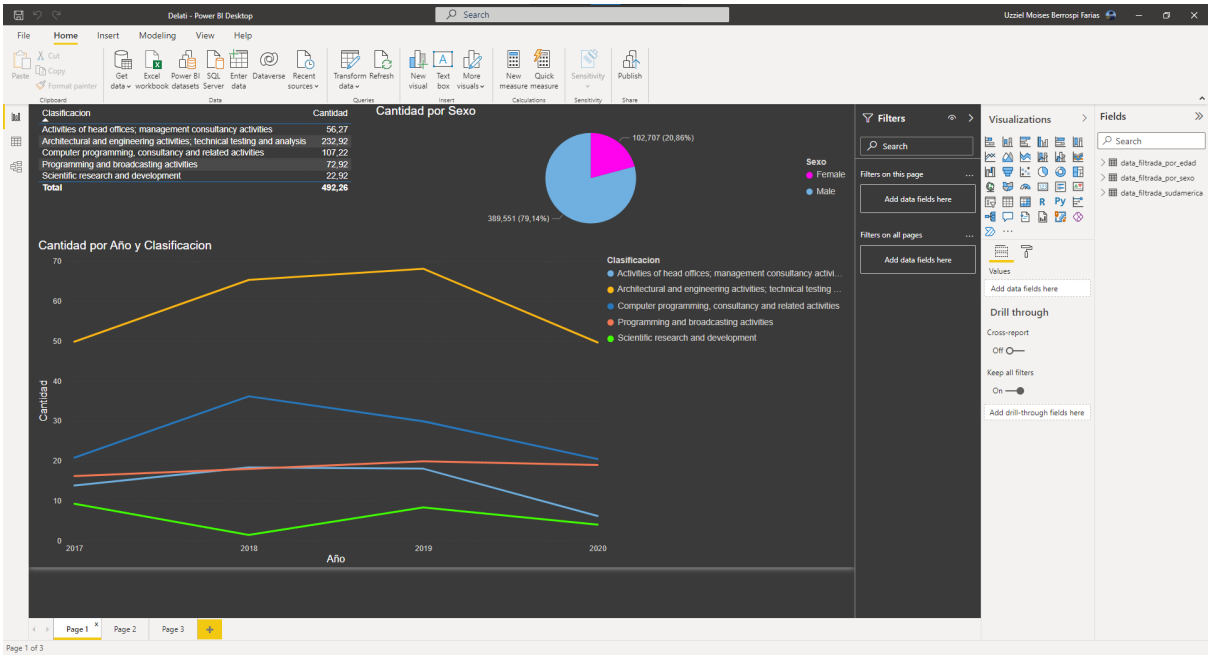
	ref_area.label	indicator.label	source.label	classif1.label	classif2.label	time	obs_value	obs_status.label	note_classif.label	note_indicator.label	note_source.label
0	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	17479.817	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
1	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	4441.899	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
2	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	21.865	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
3	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	98.687	NaN	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...
4	Peru	Employment by age and economic activity - ISIC...	HS - Encuesta Nacional de Hogares	Age (Youth, adults): 15+	Economic activity (ISIC-Rev.4), 2 digit level:...	2017	2.211	Unreliable	NaN	NaN	Repository: ILO-STATISTICS - Micro data proces...

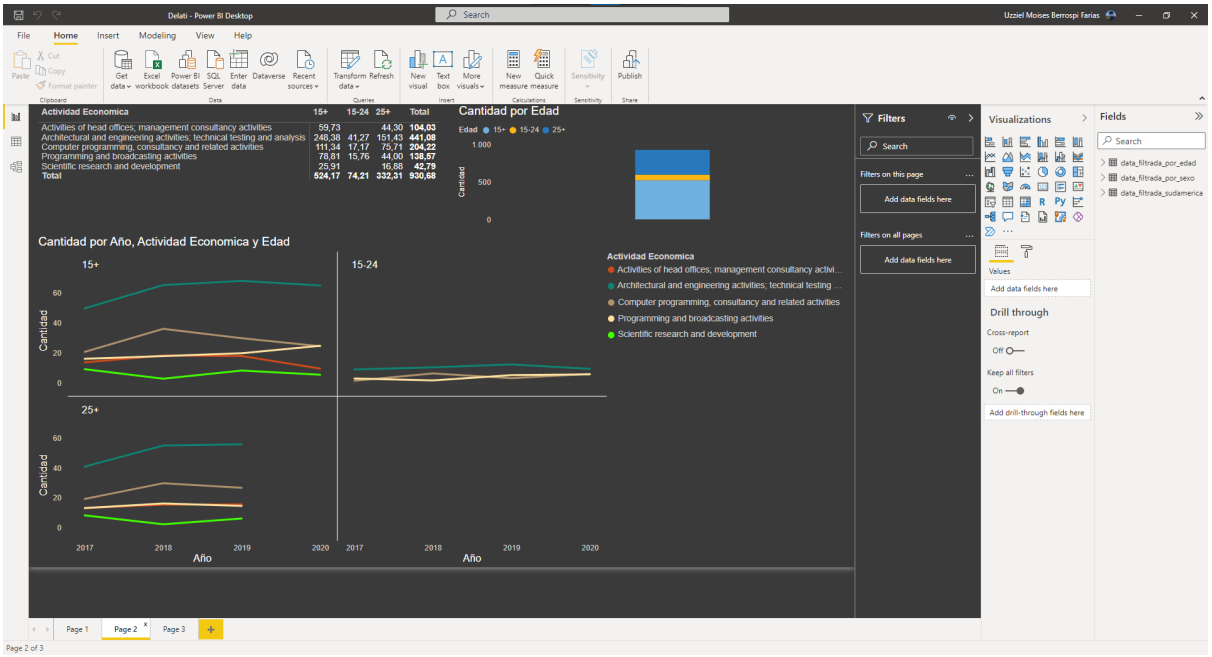
```
train.info()
```

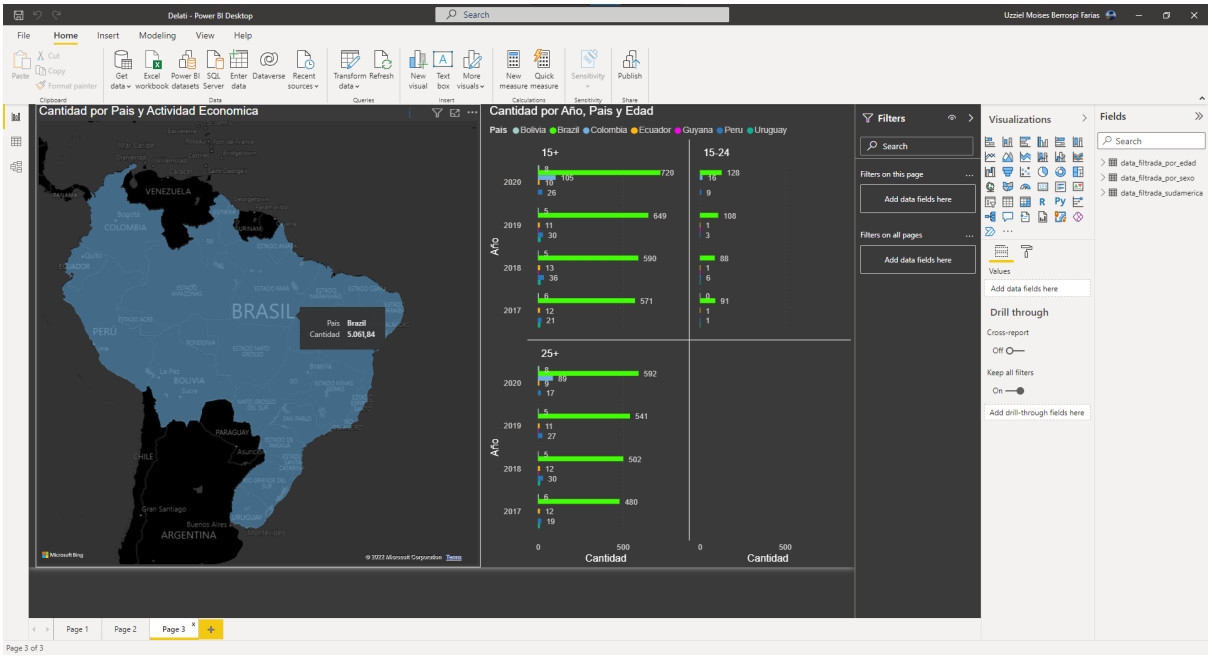
Python

8.1.6. Al finalizar la ejecución se obtendrá un csv con el nombre de data_filtrada, es ahí que se observa la data que ha sido filtrada. La cual será utilizada para los dashboards.

8.2. PROPUESTA DASHBOARD POWERBI







9. RECOMENDACIONES

- Instalar correctamente las librerías necesarias para todas las funciones que se usarán.
- Se puede utilizar otro algoritmo aparte del Naive Bayes para futuras mejoras.
- Seguir analizando datasets en búsqueda de datos relevantes para el proyecto.