Trabajo Integrador 2025

Seminario de Lenguajes Opción Python

El trabajo integrador consta del desarrollo de una aplicación de búsqueda y visualización de información relacionada a la Encuesta Permanente de Hogares (EPH), a partir de ahora nos referiremos a la aplicación como Encuest.AR. El desarrollo de este trabajo involucra el análisis de los datos con los cuales se trabajarán, su limpieza y preparación necesaria y finaliza con el desarrollo de una interfaz amigable con el usuario/a. El objetivo es que logren aplicar sus conocimientos de programación mientras exploran datos reales y construyen una interfaz visual con Streamlit.

El proyecto se divide en dos etapas:

- **Primera Etapa**: se enfocarán en entender y limpiar los datos de la EPH. Tendrán que trabajar con los archivos que contienen información sobre hogares e individuos, aplicar filtros y prepararlos para su uso en la aplicación.
- **Segunda Etapa**: diseñarán la interfaz en Streamlit e integrarán los datos procesados. Se crearán diferentes secciones para mostrar información sobre educación, ocupación, vivienda y otros temas de interés.

Este trabajo será realizado en grupos de **cinco** personas, por lo que se espera una distribución equitativa de tareas. Cada integrante del equipo tendrá un rol importante en el desarrollo, ya sea en la manipulación de datos, programación de funcionalidades, diseño de la interfaz o documentación del proyecto.

Además de mejorar su manejo de Python y librerías como *Pandas*, *Matplotlib* y *Streamlit* (entre otras), este proyecto les permitirá desarrollar habilidades de trabajo en equipo, control de versiones con *GitLab* y documentación de código.

El presente documento describe el trabajo a realizar y los requerimientos de la primera entrega.

1.- Encuest.AR: aplicación de búsqueda y visualización

Encuest.AR consiste en la disponibilización resumida y jerarquizada de la información contenida en *datasets* de la encuesta permanente de hogares facilitada por el estado nacional. Con resumida y jerarquizada nos referimos a que no basta con la presentación cruda de la información sino que se debe llevar a cabo un proceso de entendimiento de la

información con la cual se trabaja para poder condensarla en la aplicación de visualización y búsqueda.

La idea es transformar datos crudos en información comprensible para que cualquier usuario pueda analizar tendencias y características socioeconómicas de la población argentina.

La EPH es un programa nacional de producción permanente de indicadores sociales cuyo objetivo es conocer las características socioeconómicas de la población. Es realizada en forma conjunta por el Instituto Nacional de Estadísticas y Censos (INDEC) y las Direcciones provinciales de estadísticas (DPE). Por trimestre se releva y publica cada encuesta.

En una fase inicial nos enfocaremos en los *datasets*, es decir, comprender su estructura e información para así lograr extraer la información relevante.

- Sitio de descarga de datasets: ingresar a Microdatos y documentos 2016-2024, Base de microdatos y luego estarán por trimestre cada encuesta realizada, como se muestra la Figura 1. **Se recomienda realizar la descarga en formato TXT.**

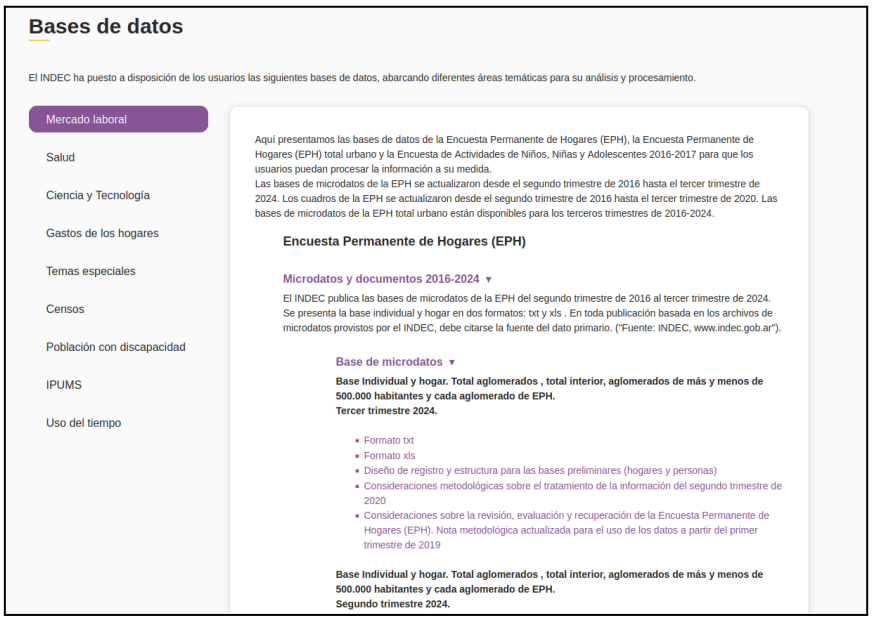

Figura 1. Página oficial del INDEC para descargar de datasets

2.- Comprendiendo la EPH

La información de la EPH se obtiene de **dos** archivos (dos archivos por trimestre de interés), uno denominado **individual** y otro **hogar**. Como su nombre lo indica uno almacena la información relacionada a los individuos encuestados y el otro a los hogares encuestados.

La explicación del significado de cada campo se puede obtener del archivo “Encuesta Permanente de Hogares - Diseño de registro y estructura para las bases preliminares Hogar y Personas” que también se encuentra disponible para su descarga.

Aclaración importante: si bien trimestre tras trimestre la metodología no acostumbra a cambiar se debe constatar para el trimestre a analizar el archivo de diseño y estructura.

3.- Funcionalidades de la primera etapa

En esta primera parte del trabajo, se deben cumplir dos objetivos principales:

1. Entender y procesar los conjuntos de datos originales para dejarlos listos para su uso en la implementación de la interfaz de usuario.
2. Iniciar el proyecto **Streamlit**, que servirá de base para la mencionada interfaz y se continuará en la segunda parte del trabajo integrador.

3.1.- Procesamiento de los datasets

Se requiere la creación de al **menos dos** archivos de formato Jupyter Notebook para el procesamiento del archivo individuos y otro el de hogares. Cada uno contendrá las operaciones destinadas al procesamiento, análisis y limpieza de los conjuntos de datos originales (**añadir los comentarios necesarios para su entendimiento**).

Procesamiento

Importante: tal y como indica el archivo explicativo recuerden que cada fila se encuentra ponderada, es decir, una fila del archivo individuos no representa a UN individuo sino que representa **tantos individuos** como la columna **PONDERA** lo indica. Situación similar ocurre con el archivo de hogar. Un resultado que no contemple la ponderación es un resultado incorrecto.

Antes de comenzar con el procesamiento comprobaremos que comprendemos la estructura del dataset.

- Carguen el archivo perteneciente al tercer trimestre del 2024 e informe cuántos varones y mujeres representa el dataset.

  Ayuda: recuerden ponderar los datos. ***Columna de interés: CH04****.*

- Informen cuántas personas mayores de edad han completado los estudios secundarios.
- Informen el porcentaje de viviendas que son ocupadas por el/la propietario/a de la vivienda y del terreno.
- Informen el aglomerado con mayor y menor cantidad de viviendas con más de 2 ocupantes que no posean baño.

Se recomienda que cada integrante del equipo realice a modo de práctica los ejercicios anteriores. **Los mismos son simplemente a modo de entrenamiento, no son parte del sistema en sí.**

Procesamiento e información a obtener

Sección A: generación del dataset principal

1. Se debe definir una carpeta, cuyo path debe ser configurable vía una variable, donde se almacenarán el conjunto de archivos de “individuos” y “hogares”.
2. Se debe generar una sección de código dentro del notebook que busque en la mencionada carpeta cada uno de los archivos “individuos” y “hogares” y los una. Es decir, si en la carpeta tenemos tres archivos de individuos (año 2023 trimestre 1, año 2023 trimestre 2, año 2024 trimestre 1) se debe leer la información de cada uno y generar un dataset que contenga toda la información.

   Recordar que dentro de las columnas de información se encuentra el año y trimestre, por lo que luego de unirse se puede conocer fácilmente el período al cual pertenece.

   Esta operación automática de búsqueda y generación de dataset se debe realizar tanto para el archivo de individuos como el de hogares.

   Informacion individuos:

3. Se debe traducir los valores CH04 numéricos a "Masculino" y "Femenino" según corresponda. El resultado se debe almacenar en una nueva columna llamada **CH04_str**.
4. Se debe traducir los valores NIVEL_ED numéricos a descripciones en formato texto. El resultado se debe almacenar en una nueva columna llamada **NIVEL_ED_str**. La transformación debe seguir las siguientes reglas:
   - 1: "Primario incompleto".
   - 2: "Primario completo".
   - 3 "Secundario incompleto".
   - 4: "Secundario completo".
   - 5 a 6: "Superior o universitario".
   - 7 o 9: “Sin informacion”.
5. Se debe crear una nueva columna denominada **CONDICION_LABORAL** de formato texto. La transformación debe seguir las siguientes reglas:

- Ocupado autónomo: si ESTADO es igual a 1 y CAT_OCUP es 1 o 2.
- Ocupado dependiente: si ESTADO es igual a 1 y CAT_OCUP es 3 o 4 o 9.
- Desocupado: si ESTADO es igual a 2.
- Inactivo: si ESTADO es igual a 3.
- Fuera de categoría/sin información: si ESTADO es igual a 4

6. Se debe generar una nueva columna llamada UNIVERSITARIO numérica que indica si una persona mayor de edad ha completado el como mínimo el nivel universitario (1: Sí, 0: No, 2: no aplica).

Informacion hogar:

7. Se debe generar una nueva columna llamada **TIPO_HOGAR** que indica el tipo de hogar:
   - "Unipersonal" (una persona).
   - "Nuclear" (2 a 4 personas).
   - "Extendido" (5 o más personas).
8. Se debe generar una nueva columna llamada **MATERIAL_TECHUMBRE** que indica el tipo de hogar basado en el campo V4:
   - 5 a 7: "Material precario".
   - 1 a 4: "Material durable".
   - 9: “No aplica”.
9. Se debe generar una nueva columna denominada **DENSIDAD_HOGAR**. La generación debe seguir las siguientes reglas:
   - Bajo: menos de 1 persona por habitación.
   - Medio: entre 1 y 2 personas por habitación.
   - Alto: más de 2 personas por habitación.
10. Se debe generar una nueva columna llamada **CONDICION_DE_HABITABILIDAD**, la misma califica a las viviendas y puede tener el valor de: insuficiente, regular, saludables y buena.

    Las reglas que definen la clasificación de las viviendas deben ser definidas por cada grupo analizando los datos con los que se cuentan. El único requisito es considerar al menos 5 de las siguientes columnas:

    - IV6: tiene agua.
    - IV7: origen del agua.
    - IV8: posee baño.
    - IV9: ubicación del baño.

- IV10: tipo de baño.
- IV11: desagüe del baño.
- MATERIAL_TECHUMBRE: creado anteriormente
- IV3: material de pisos interiores.

Un ejemplo de regla podría ser:

- Clasifican como CONDICION_DE_HABITABILIDAD insuficiente si no poseen agua, o no poseen baño, o si poseen baño pero el mismo está fuera del terreno, o si el desagüe del baño es a un hoyo en la tierra. Cabe aclarar que exactamente este ejemplo no se puede utilizar en este inciso.

Sección B: información a obtener: consultas al dataset principal

Tomando como fuente los datasets generados en la sección A y sus nuevas columnas:

1. A partir de la información de cada año contenida en el dataset se debe informar, año tras año, el porcentaje de personas mayores a 6 años capaces e incapaces de leer y escribir.

   Importante: tomar la información del último trimestre de cada año.

2. A partir de un año y trimestre elegido por el usuario informar el porcentaje de personas no nacidas en Argentina que hayan cursado un nivel universitario o superior.
3. A partir de la información contenida en el dataset informar el año y trimestre donde hubo menor desocupación.
4. Ranking de los 5 aglomerados con mayor porcentaje de hogares con dos o más ocupantes con estudios universitarios o superiores finalizados. Información obtenida a partir del par de archivos más recientes.
5. Informar para cada aglomerado el porcentaje de viviendas ocupadas por sus propietarios.
6. Informar el aglomerado con mayor cantidad de viviendas con más de dos ocupantes y sin baño. Informar también la cantidad de ellas.
7. Informar para cada aglomerado el porcentaje de personas que hayan cursado al menos en nivel universitario o superior.
8. Ordenar las regiones de forma descendente según el porcentaje de inquilinos de cada una.

9. Pedir al usuario que seleccione un aglomerado y a partir de la información contenida retornar una tabla que contenga la cantidad de personas mayores de edad según su nivel de estudios alcanzados.

   A modo de ejemplo:

   Nombre de Aglomerado

Año	Trimestre	Primario incompleto	Primario completo	Secundario Incompleto	Secundario Completo	Superior o universitario"
2020	1					
2020	2					
2020	3					
2020	4					
2021	1					
2021	2					

10. Pedir al usuario que seleccione dos aglomerados y a partir de la información contenida retornar una tabla que contenga el porcentaje de personas mayores de edad con secundario incompleto.

    A modo de ejemplo:

Año	Trimestre	Aglomerado A	Aglomerado B
2020	1	27%	10%
2020	2	29%	15%
2020	3	29%	16%
2020	4	28%	14%
2021	1	30%	14%
2021	2	31%	14%

11. Pedir al usuario que seleccione un año, y busque en el último trimestre almacenado del mencionado año, el aglomerado con mayor porcentaje de viviendas de “Material precario” y el aglomerado con menor porcentaje de viviendas de “Material precario”.

12. A partir de la información del último trimestre almacenado en el sistema se debe retornar para cada aglomerado el porcentaje de jubilados que vivan en una vivienda con CONDICION_DE_HABITABILIDAD insuficiente.
13. Solicitar un año al usuario y a partir de la información del último trimestre de dicho ejercicio informar la cantidad de personas que hayan cursado nivel universitario o superior y que vivan en una vivienda con CONDICION_DE_HABITABILIDAD insuficiente

3.2.- Comenzar aplicación Streamlit

Se solicita que generen un menú en el sidebar con las siguientes páginas:

- (P1) Inicio
- (P2) Carga de datos
- (P3) Búsqueda por tema
- (P4) Visualización

En las páginas P3 y P4 no es necesario agregar funcionalidad.

La página 1 (P1) debe contener:

- Un título con el nombre de la aplicación (el que ustedes deseen).
- Un pequeño párrafo explicando brevemente la información que se almacena, es decir, que contiene la EPH.
- En la etapa 2 se agrega las indicaciones de uso de la interfaz.

La página 2 (P2) debe:

- Dar a conocer la información con la que cuenta actualmente el sistema. Para esto debe extraer el desarrollo de la “Sección A: generación del dataset principal.” punto 1 y 2 para así detectar de qué años y trimestres se contiene información. Solo es de importancia el primer y último año y trimestre.

  Ejemplo: “El sistema contiene información desde el 01/2016 hasta el 03/2024.”

- Contener un botón que fuerce la actualización del dataset a partir de los archivos contenidos en la carpeta configurada en el punto 1. El objetivo es poder descargar un nuevo par de archivos correspondientes a un trimestre mientras la interfaz permanece encendida y poder forzar su incorporación para trabajar con ellos.

4.- Consideraciones generales de implementación

El software implementado deberá funcionar correctamente tanto en Windows, Linux o Mac. Se deberá armar una estructura de directorios organizando los archivos en carpetas y subcarpetas de manera tal que mantengan el código organizado haciendo que sea fácil de actualizar y de corregir. Se evaluará tanto el código como su organización.

El código se deberá subir a un repositorio de GitLab designado por la cátedra. En este repositorio, cada equipo deberá incluir un archivo denominado **README.md** que contenga el nombre de los integrantes del equipo y las instrucciones para ejecutar cada aplicación. Además, se solicitará un archivo que enumere las bibliotecas necesarias, siguiendo la metodología explicada en las prácticas para este propósito.

Para el desarrollo del código se deberá cumplir con los siguientes requerimientos:

- En esta primera etapa del trabajo integrador **NO se pueden utilizar** las librerías de Python **pandas y pyeph**.
- Usar la herramienta **Streamlit** para el desarrollo de la interfaz gráfica.
- Documentar el código usando docstrings en funciones y clases (estas últimas caso de ser utilizadas).
- Incluir un archivo **LICENSE** con la licencia del código.
- Se debe tener en cuenta las guías de estilo de Python para la escritura de código.
- Definir una estructura de carpetas que permita estructurar el código de forma prolija.

4.1.- Consideraciones de la entrega y defensa

Si bien el trabajo es grupal, **la nota de la defensa es INDIVIDUAL.**

La defensa se lleva a cabo durante el horario de consulta la semana posterior a la fecha de entrega. Este proceso implica la realización de un encuentro virtual o presencial con el ayudante asignado, durante el cual se formulan una serie de preguntas relacionadas con la entrega. Dichas preguntas están dirigidas a los distintos miembros del grupo, distribuyendo equitativamente la participación entre todos ellos. La finalidad es que cada integrante tenga la oportunidad de responder. La evaluación se basa en el desempeño de las respuestas, lo que determina finalmente la nota asignada a cada miembro del grupo en relación con la entrega realizada.

Esto implica que, aunque se trate de una entrega grupal, las calificaciones entre los miembros del grupo pueden variar.

Tener en cuenta que es fundamental mantener una participación activa tanto en las consultas prácticas como en el repositorio de GitLab. Esto permitirá al ayudante obtener una comprensión conceptual del conocimiento de cada participante del grupo.

4.2.- Detalles de entrega

La fecha de entrega de la primera etapa es el día **7 de mayo a las 23:59 hs. La defensa será la semana del 12 de mayo en horario de la consulta práctica.**

La primera entrega otorga al estudiante un **máximo de 35 puntos**, los cuales se basan en su desempeño durante las consultas y la defensa.

5.- Criterios de Evaluación

La distribución de los puntos de la entrega estará disponible en la rúbrica asociada a la tarea. Se evaluará:

- Funcionalidad implementada de acuerdo al enunciado.
- Cumplimiento de las consideraciones planteadas.
- Código subido en tiempo y forma al repositorio de GitLab indicado.
- Participación individual activa durante el desarrollo del trabajo, que incluye asistencia a las consultas (tanto virtuales como presenciales) y contribuciones al repositorio de GitLab.
- En la defensa, se espera que cada integrante del grupo demuestre los conocimientos utilizados para la realización del trabajo.