INFORME DE ANÁLISIS DE DATOS DE INGRESO A LA EDUCACIÓN SUPERIOR 2000-2021

# Introducción

El presente es un informe presentado como parte de un proceso de postulación a un cargo de analista de datos Tableau y Python en la Sede Bogotá de la Universidad Nacional de Colombia. El ejercicio consiste en la descarga de datos de la página de bases consolidadas del Sistema Nacional de Instituciones de Educación Superior (SNIES) del Ministerio de Educación Nacional de Colombia (SNIES, 2022).

Dichos datos deben someterse a un proceso de extracción, transformación y carga (ETL por sus siglas en inglés), tras de lo cual se deben construir varias visualizaciones tanto en Python como en la herramienta de inteligencia de negocios “Tableau”. Después de lo anterior, se debe llevar a cabo un conjunto de análisis y conclusiones a partir de los datos, los cuales se plasman en la parte final de este informe

# El ETL a los datos

El procesamiento de los datos se realizó usando Jypiter Notebook con Python 3.

## Las fuentes de datos

En la página de bases de datos del SNIES, se encontraron 9 archivos referente a la inscripción a diferentes programas en el país. De esos 9, 8 archivos se encontraron en formato Excel (.xlsx) mientras que uno de ellos se encontró en formato Excel binario (.xlsb). Dichos archivos se extrajeron de la máquina local usando las librerías pandas y pyxlsb. Finalmente, se eliminaron las filas de presentación e informativas en cada archivo, dejando únicamente las tablas cargadas como dataframes de pandas.

## Transformaciones

Luego de su extracción, los datos se transformaron de la siguiente manera: En primer lugar, el archivo binario (.xlsb), el cual contenía información de entre los años 2000 y 2013, tenía una distribución distinta al resto de archivos. Entre otras cosas, carecía de columnas para año, sexo y semestre, y dicha información se encontraba dispersa en los encabezados de una serie de columnas que tenían nombres como “Hombre 2003-2” o “Mujer 2005-1”.

Por lo anterior, fue necesario realizar una conversión de columnas a filas usando el método pandas.melt(), creando dos nuevas columnas, la primera, denominada ‘Categoria’ mientras que la segunda se denominó temporalmente como ‘Valor’. Luego de hecho esto, la columna ‘Categoria’ Se dividió en ‘SEXO’, ‘ANO’ (AÑO en caracteres ASCII), y ‘SEMESTRE’. Finalmente la columna ‘Valor’ se renombró como ‘INSCRITOS’ para concordar con el proceso de los demás archivos

Una vez definido el proceso para los datos 2000-2013, se procedió con todos renombrando la última columna como ‘INSCRITOS’, eliminando registros con valores nulos o con términos como 'NO APLICA', 'N/A', 'NA','NO IDENTIFICADO','SIN PROGRAMA ESPECIFICO', etc. Se pasaron todos los valores a mayúsculas, se eliminaron tildes y otros caracteres no ASCII y se modificaron ciertos valores puntuales.

Cabe anotar que se intentó hacer un proceso de lematización usando la librería SpaCy pero, además del costo y demora del computador en realizarlo, los resultados no fueron muy buenos, así que se decidió hacer correcciones manuales, para corregir duplicidad en nombre de departamentos, ciudades, niveles de formación, entre otros.

También hubo una discrepacia en los datos en cuanto al género o sexo registrados. De los 9 archivos, 7 tenían una columna denominada ‘Sexo’ con valores ‘Hombre’ ‘Mujer’, mientras que dos tenían una columna ‘Género’ con valores ‘Masculino’ y ‘Femenino’. Con el fin de disminuir la cantidad de procesos en la máquina, se optó dejar todos los archivos con una columna ‘Sexo’ con valores ‘Hombre’ y ‘Mujer’

## El archivo resultado

El archivo resultado se guardó en la carpeta con el nombre ‘datos\_procesados.xlsx’

# Las gráficas en Python de los datos

Se realizaron gráficas en Python usando la librería Plotly. A continuación algunos ejemplos:

Imagen que contiene Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamente

# El libro de Tableau construido a partir de los datos

Con el archivo de Excel de datos procesados, se construyó un libro de Tableau con diferentes hojas, tableros y una historia.

## Las hojas

Las hojas que se crearon fueron las siguientes:

* Inscritos por departamento
* Inscritos por año
* Inscritos por sexo
* Nivel de formación por sexo
* Inscritos por sector
* Inscritos por SUE

## Los Tableros

Con dichas hojas se crearon sendos tableros con filtros adicionales que pueden ser utilizados por el usuario.

## La historia

Con los tableros mencionados se construyó una historia, la cual se compartió en el servidor de Tableau Public en la URL: <https://public.tableau.com/views/IngresoEducacionSuperior2000-2021Colombia/HistoriaIngresoEducacinSuperior?:language=es-ES&publish=yes&:display_count=n&:origin=viz_share_link>

# Análisis de los resultados

## Inscritos por departamento

La gráfica obtenida es la siguiente:

Mapa

Descripción generada automáticamente

Aquí se observa cómo la oferta y la inscripción está muy centralizada en las grandes ciudades con un poco oferta en los territorios

## Inscritos por año

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Se observa cómo la inscripción ha ido creciendo hasta 2012, y a partir de ese momento se mantiene en niveles constantes

## Análisis de género

Gráfico, Gráfico circular

Descripción generada automáticamente

Gráfico, Gráfico de burbujas

Descripción generada automáticamente

Los gráficos anteriores muestran cómo, en general, la proporción de ingreso a la universidad por parte de mujeres es ligeramente superior a la de hombres, sobre todo en los niveles intermedios entre formación tecnológica y especialización.

## Análisis por sector

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Aquí se puede ver cómo la mayor proporción la ofrece la universidad pública y el papel de la UNAL (en rojo) dentro de ese ecosistema

# Referencias

* SNIES 2022. Bases consolidadas. [Bases consolidadas (mineducacion.gov.co)](https://snies.mineducacion.gov.co/portal/ESTADISTICAS/Bases-consolidadas/)