CODER HOUSE

Análisis socioeducativo de los habitantes de la Ciudad de Buenos Aires

Profesor: Damian Dapueto

Tutor: Héctor Alonso

Grupo de Trabajo: Lucia Buzzeo, Lucia Hukovsky, Jose Saint German, Juan Martín Carini

Introdución

Objetivos

- Descubrir las principales variables intervinientes en el nivel máximo educativo alcanzado por la población de la Ciudad Autónoma de Buenos Aires (CABA).
- Generar un modelo de predicción aplicado a nuestra variable target "Nivel Máximo Educativo", esto lo haremos implementando los siguientes modelos de clasificación:
  - **Árbol de decisión:** que construye un árbol durante el entrenamiento que es el que aplica a la hora de realizar la predicción.
  - **Bosque Aleatorio:** que es un conjunto (ensemble) de árboles de decisión combinados con bagging.

Fuente de información

Para trabajar esta problemática, se ha recurrido a la Encuesta Anual de Hogares del Gobierno de la Ciudad de Buenos Aires para el año 2019. El dataset está disponible en la base de datos abiertos del GCBA.
Esta encuesta contiene información demográfica, social, económica, educativa y de salud de 14319 habitantes de la Ciudad, la cual es una muestra representativa que permite obtener un vistazo de la población de la Ciudad.

Estructura de los trabajos

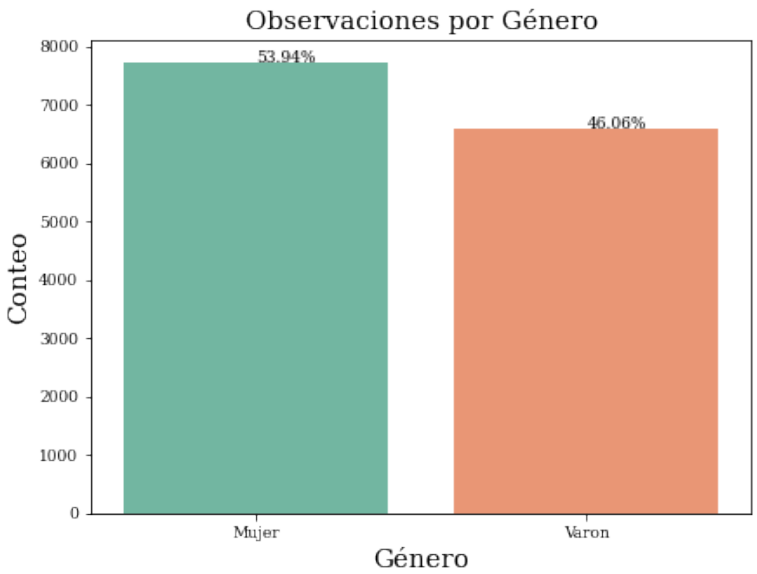
Este trabajo se ha dividido en 3 partes:

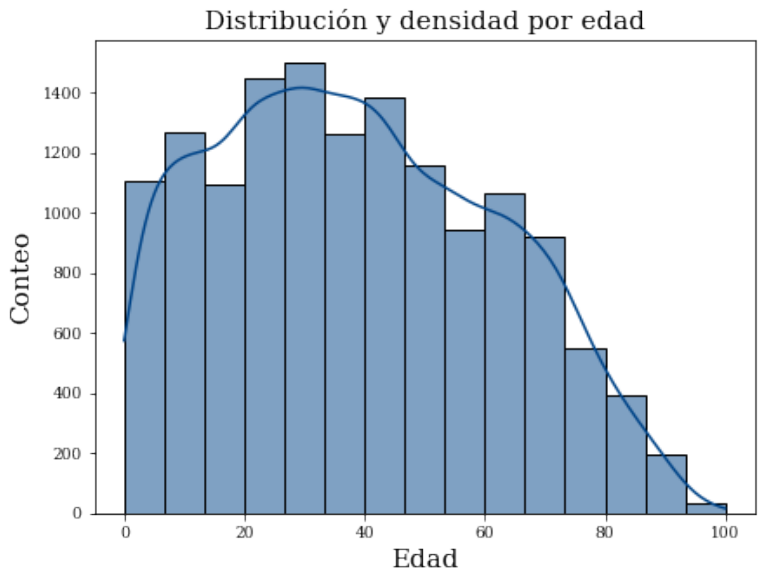
1. **Introducción a las variables del problema:** Se realiza un análisis de las variables del dataset. En el mismo se busca conocer su performance dentro del dataset. A la vez, se investiga cómo las variables interactúan entre sí. Esta parte es lo que se conoce como análisis univariado, bivariado y multivariado.
2. **Modelos analíticos:** En esta sección se entrenan diversos modelos analíticos y algoritmos que sirven para alcanzar los objetivos seteados para el presente proyecto. Como la variable objetivo es categórica, se realizan modelos de clasificación.
3. **Conclusión:** Se alcanzan conclusiones finales sobre los hallazgos. Además, se discuten posibles limitaciones y se plantean futuras líneas de análisis, a partir del análisis presente.

Análisis exploratorio de los datos (EDA)

Análisis univariado

Comenzamos con un pantallazo general sobre las primeras cualidades de los datos, como muestra representativa para la EPH, sobre quiénes son los ciudadanos representados en el dataset.

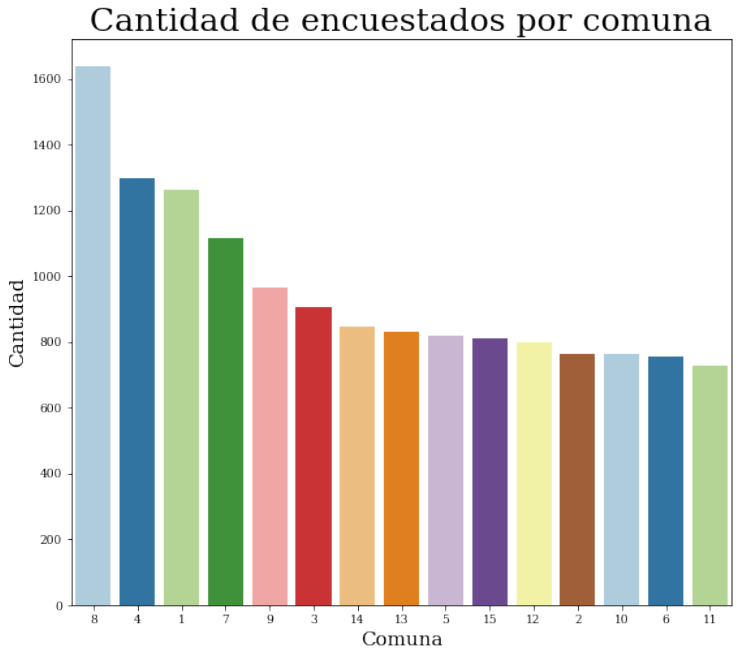


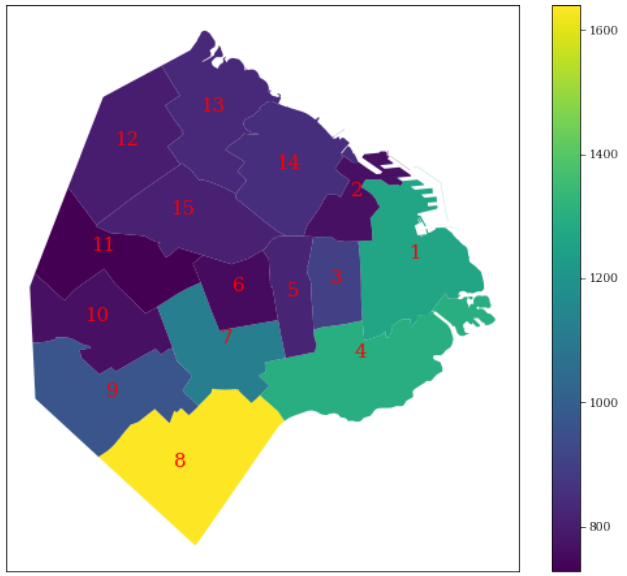


En la variable género los datos parecen equilibrados en las categorías. Para el caso de la variable "edad", la distribución se asemeja a la de una normal.

Análisis univariado

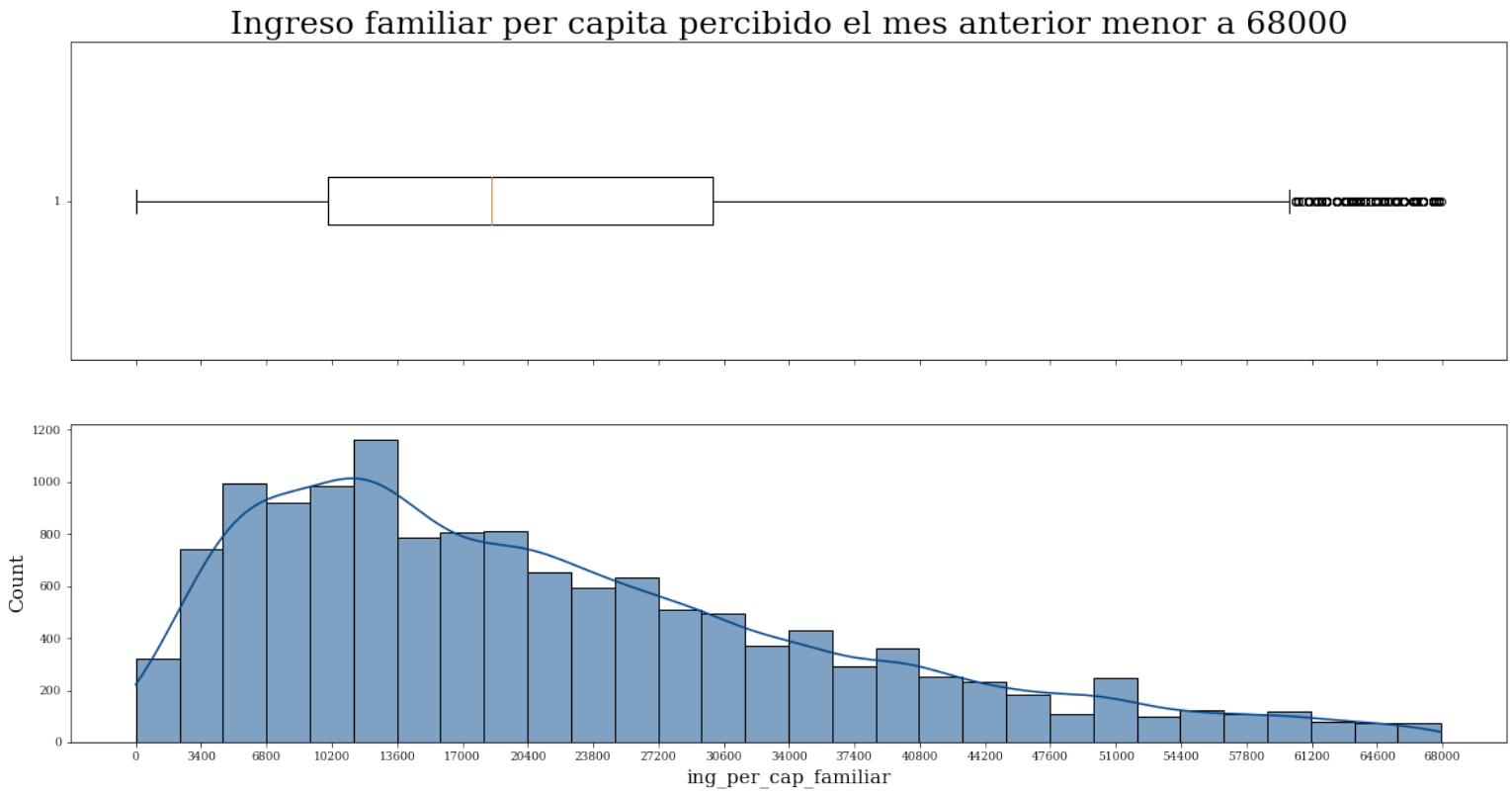
Seguimos observando la variable “comuna”:



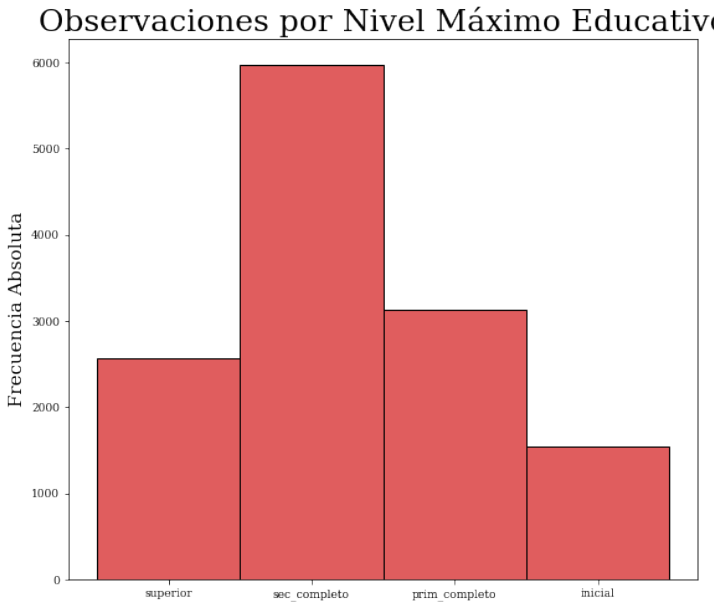


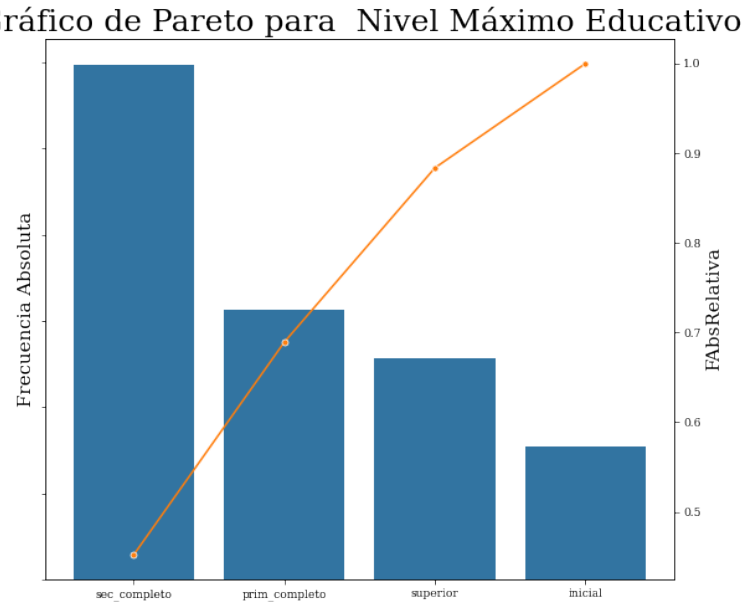
Observando ambos gráficos vemos que las comunas 1,4,7 y 8 tienen mayor cantidad de casos.

Análisis univariado



Análisis univariado

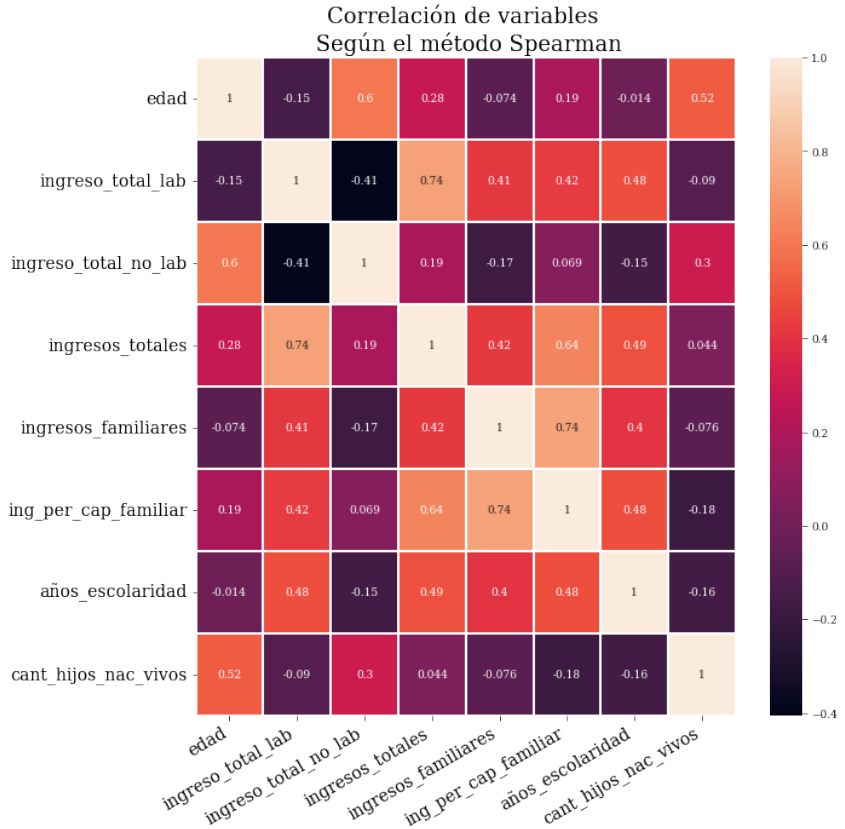




Podemos observar que el nivel máximo educativo más alcanzado es el secundario completo, seguido por el primario. Contrario de lo que habíamos intuido anteriormente, el nivel superior quedó en tercer lugar. Adicionalmente, el nivel secundario y primario explican casi el 77 % de los datos.

Análisis bivariado

Para comenzar el análisis bivariado del problema, realizamos diferentes heat maps para ver si algo nos llama la atención entre las variables numéricas.



- No se observan fuertes correlaciones.
- "años_escolaridad" correlaciona moderadamente bien con variables relacionadas al ingreso.
- La principal correlación positiva es "años_escolaridad" con ingreso familiar per cápita ("ing_per_cap_familiar").

Análisis bivariado

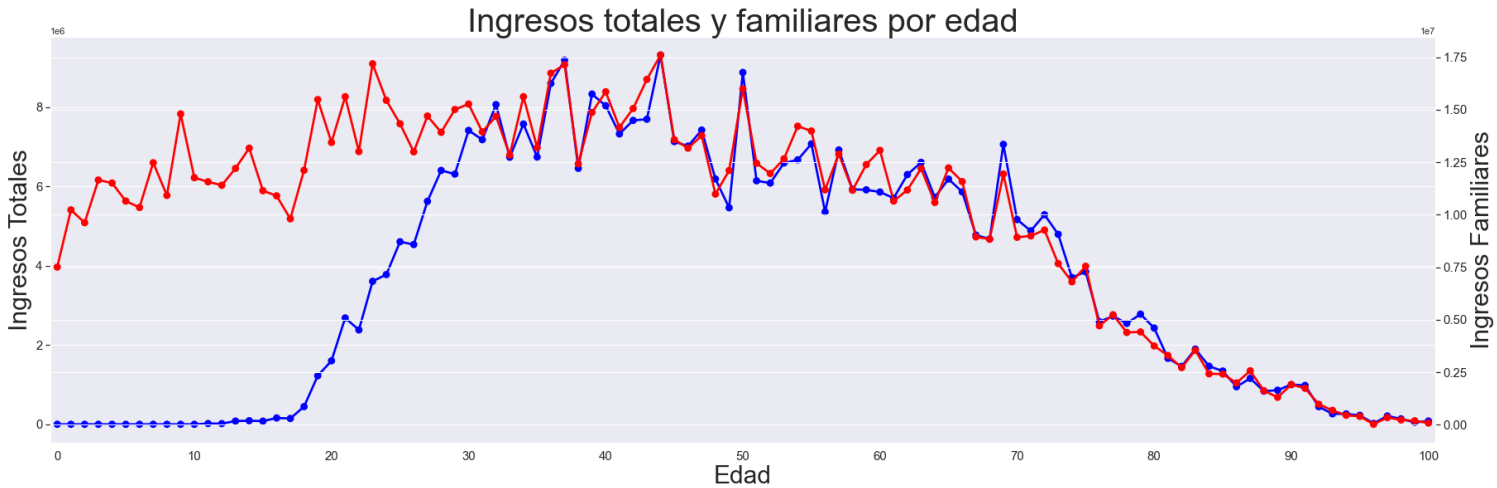
Corremos una tabla de correlación y filtramos las de valores más altos:

	Variable 1	Variable 2	Correlación
2	ingreso_total_lab	ingresos_totales	0.80
6	ingresos_familiares	ing_per_cap_familiar	0.76
4	ingresos_totales	ing_per_cap_familiar	0.62
5	ingresos_totales	años_escolaridad	0.60
1	edad	ingreso_total_no_lab	0.57
7	años_escolaridad	Target	0.57
3	ingreso_total_lab	años_escolaridad	0.54

- Como es esperable, hay alta correlación entre las variables relacionadas al ingreso.
- A su vez, encontramos una alta correlación (66 %) entre los ingresos y los años de escolaridad.
- Y también observamos una relación positiva entre la edad y los ingresos totales.

Análisis bivariado

Observaremos la relación entre los ingresos totales de cada hogar y los ingresos familiares por edad:

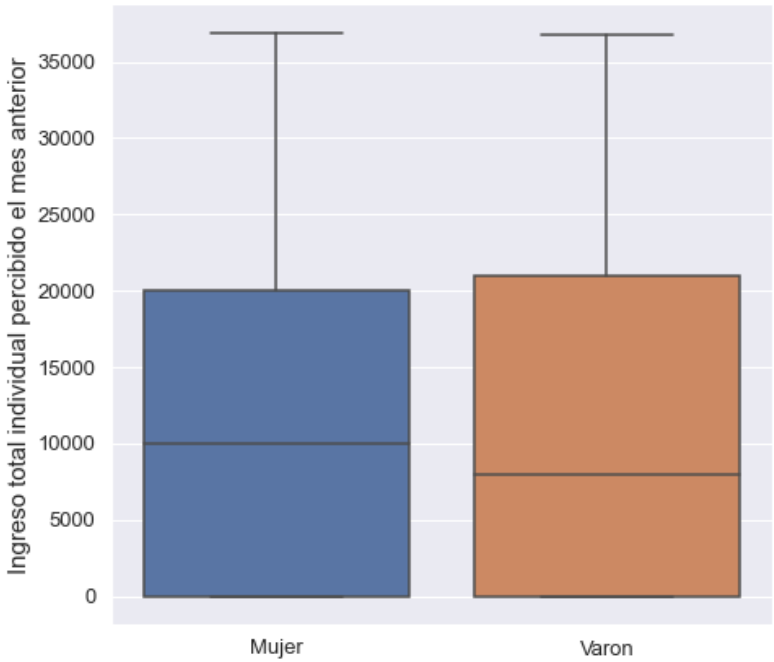


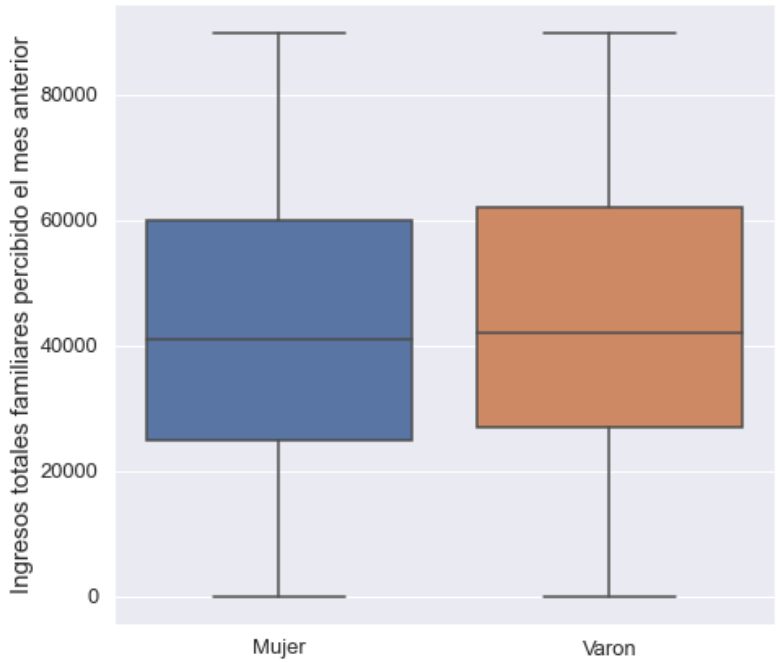
Se puede ver que desde los 30 años en adelante el ingreso total de la persona se corresponde con el ingreso familiar. Por ende suele haber un único ingreso fuerte por grupo familiar.

Análisis bivariado

Análisis bivariado

Veamos como se relacionan las categorías de ingresos con el genero de los encuestados:





Y resulta que con las variables de ingreso, no parece haber nada disruptivo. Salvo que los hombres parecieran tener ingresos totales y familiares mayores que las mujeres, pero no pareciera que haya distribuciones desiguales en los años de escolaridad.