CoderHouse

Curso Data Science

Informe del Proyecto Final

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

26 de septiembre de 2022

Índice

1. Introducción

1.1. Resumen del proyecto

La ciudadanía es un concepto jurídico, filosófico y político que ha sido creado para designar a una persona física que constituye una sociedad o entidad territorial. Para las personas que forman parte de una comunidad, ciudadanos, resulta de suma importancia sentirse representados por los demás integrantes de la misma, mediante políticas públicas que abarquen sus necesidades y requerimientos.

La toma de datos demográficos y la estadística son dos herramientas primordiales a la hora de identificar requerimientos de los integrantes de una comunidad. Dichas herramientas describen, de forma cuantitativa, a la sociedad bajo estudio. Precisamente, los censos y la estadística son la fuente primaria de información para la planificación económica y social de una población, por parte de sus representantes.

En el caso particular de Argentina, el Instituto Nacional de Estadística y Censos (INDEC) es el organismo público que ejerce la dirección superior de todas las actividades estadísticas oficiales que se realizan en el país. La información que produce el INDEC es una herramienta básica para la planificación de políticas públicas, así como para las investigaciones y proyecciones que se realizan en los ámbitos académico y privado.

Al adentrarse y estudiar los índices correspondientes a uno de los ejes principales, educación, en un territorio delimitado, Ciudad Autónoma de Buenos Aires, se ha encontrado una gran limitación relacionada con su acceso inequitativo para los diferentes actores de la sociedad.

Este hecho tiene consecuencias de índole social y económico para la población. Sin embargo, la principal problemática se da a nivel individual, y radica en el impedimento al acceso educativo para un porcentaje de la sociedad. Esto no ha resultado una novedad para el grupo, pero sí ha dado el pie a la búsqueda de una respuesta teórica a dicha disparidad, en concreto, a descubrir las principales variables que afectan el nivel educativo.

El análisis realizado en el marco del presente proyecto podría establecer una base de requerimientos que permitan generar políticas públicas efectivas, no solo en el ámbito educativo, sino en el económico, cultural, social y geográfico, entre otros.

1.2. Definición de la 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.

1.3. Objetivos del proyecto

Entre los objetivos del proyecto, se encuentra descubrir las principales variables intervinientes en el nivel máximo educativo alcanzado por la población de la Ciudad Autónoma de Buenos Aires (CABA).

Pero como objetivo principal intentaremos 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.

A partir de la obtención del mejor árbol de decisión y el mejor bosque aleatorio tomaremos la decisión de cuál de los dos es el mejor algoritmo para lograr los objetivos de este trabajo.

2. Planificación

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.

3. Introducción a las variables: Análisis exploratorio de los datos

En el análisis explotarotio de los datos se ha buscado definir las variables que componen el dataset. En ese sentido, luego de cargar el dataset, se ha comenzado por conocer los tipos de datos y si existían nulls.

Tamaño del set de datos: 14319 entadas, 0 a 14318
Con un total 31 columnas:

#	Columna	Entradas No Vacías	Tipo de Dato
0	id	14319 no nulos	numérica
1	nhogar	14319 no nulos	numérica
2	miembro	14319 no nulos	numérica
3	comuna	14319 no nulos	numérica
4	dominio	14319 no nulos	objeto
5	edad	14319 no nulos	numérica
6	sexo	14319 no nulos	objeto
7	parentesco_jefe	14319 no nulos	objeto
8	situación_conyugal	14318 no nulos	objeto
9	num_miembro_padre	14319 no nulos	objeto
10	num_miembro_madre	14319 no nulos	objeto
11	estado_ocupacional	14319 no nulos	objeto
12	cat_ocupacional	14319 no nulos	objeto
13	calidad_ingresos_lab	14319 no nulos	objeto
14	ingreso_total_lab	14319 no nulos	numérica
15	calidad_ingresos_no_lab	14319 no nulos	objeto
16	ingreso_total_no_lab	14319 no nulos	numérica
17	calidad_ingresos_totales	14319 no nulos	objeto
18	ingresos_totales	14319 no nulos	numérica
19	calidad_ingresos_familiares	14319 no nulos	objeto
20	ingresos_familiares	14319 no nulos	numérica
21	ingreso_per_capita_familiar	14319 no nulos	numérica
22	estado_educativo	14319 no nulos	objeto
23	sector_educativo	14316 no nulos	objeto
24	nivel_actual	14319 no nulos	objeto
25	nivel_max_educativo	13265 no nulos	objeto
26	años_escolaridad	14257 no nulos	objeto
27	lugar_nacimiento	14318 no nulos	objeto
28	afiliacion_salud	14315 no nulos	objeto
29	hijos_nacidos_vivos	6535 no nulos	objeto
30	cantidad_hijos_nac_vivos	14319 no nulos	objeto

Tipos de datos: numéricas(10), objetos(21)

Cuadro 1: Análisis preliminar de variables del dataset.

Se puede observar que se cuenta con 10 variables numéricas y 31 variables categóricas. En base a los datos arrojados por la tabla de arriba, se han generado diversas transformaciones de variables, así como la creación de la variable "Target", pues es la que se usará para todo el análisis:

- Creación de la variable "Target" definida por la variable "nivel_max_educativo".
- En la variable "Target", se ha reducido su dimensionalidad intercambiando los valores:
  - "Secundario/medio común" y "EGB (1° a 9° año)" por "sec_completo",
  - "Primario especial" y "Primario común" por "prim_completo",
  - "Sala de 5" por "incial",
  - "Otras escuelas especiales"por "superior",
  - y por último a "No corresponde" por nulos.
- Reemplazo de los valores de "años_escolaridad" para que todos sean numéricos.

- En la variable “cantidad_hijos_nac_vivos” se ha cambiado el valor “no corresponde” a nulo, para luego cambiar el tipo de variable a entero.
- Las variables “comuna”, “id”, “nhogar” y “miembro” son de tipo numérico, pero deberían ser categóricas, por lo tanto se ha transformado su tipo a categórica.
- Por último, se han renombrado algunas variables para mejorar la comprensión de su función:
  - “dominio_Villas_de_emergencia” por “dominio_villas”,
  - “ingreso_per_capita_familiar”por “ing_per_cap_familiar”,
  - “cantidad_hijos_nac_vivos” por “cant_hijos_nac_vivos”.

Por lo tanto, se detalla el diccionario de las variables actualizadas según los cambios indicados previamente:

Variable	**Descripción**
id	Clave que identifica la vivienda
nhogar	La variable id + nhogar = clave que identifica a cada hogar
miembro	Variables id + nhogar + miembro = clave que identifica cada persona
comuna	Comuna donde reside la persona encuestada
edad	Edad de la persona encuestada
sexo	Sexo de la persona encuestada
parentesco_jefe	Parentesco entre la persona encuestada y el jefe de hogar
situacion_conyugal	Situación conyugal de la persona encuestada
num_miembro_padre	Número de miembro que corresponde al padre
num_miembro_madre	Número de miembro que corresponde a la madre
estado_ocupacional	Situación ocupacional de la persona encuestada
cat_ocupacional	Categoría ocupacional de la persona encuestada
calidad_ingresos_lab	Calidad de la declaración de ingresos laborales totales
— **ingreso_total_lab**	Ingreso total laboral percibido el mes anterior
— **calidad_ingresos_no_lab**	Calidad de la declaración de ingresos no laborable totales
ingreso_total_no_lab	Ingreso total no laboral percibido el mes anterior
calidad_ingresos_totales	Calidad de ingresos totales individuales
ingresos_totales	Ingreso total individual percibido el mes anterior
calidad_ingresos_familiares	Calidad de ingresos totales familiares
ingresos_familiares	Ingresos totales familiares percibido el mes anterior
ing_per_cap_familiar	Ingreso familiar per capita percibido el mes anterior
estado_educativo	Asistencia (pasada o presente) o no a algún establecimiento educativo
sector_educativo	Sector al que pertenece el establecimiento educativo a que asiste
nivel_actual	Nivel cursado al momento de la encuesta
nivel_max_educativo	Máximo nivel educativo que se cursó
años_escolaridad	Años de escolaridad alcanzados
lugar_nacimiento	Lugar de nacimiento de la persona encuestada
afiliacion_salud	Afiliación de salud de la persona encuestada
hijos_nacidos_vivos	Tiene o tuvo hijos nacidos vivos
cant_hijos_nac_vivos	Cantidad de hijos nacidos vivos
dominio	¿La vivienda se ubica en una villa de emergencia?
Target	Nivel máximo educativo

Cuadro 2: Diccionarios de variables actualizadas

Continuando con el análisis exploratorio de datos, se ha analizado la presencia de nulos en el dataset. Para visualizarlo se ha utilizado un gráfico de barras que incluye a todas las variables.

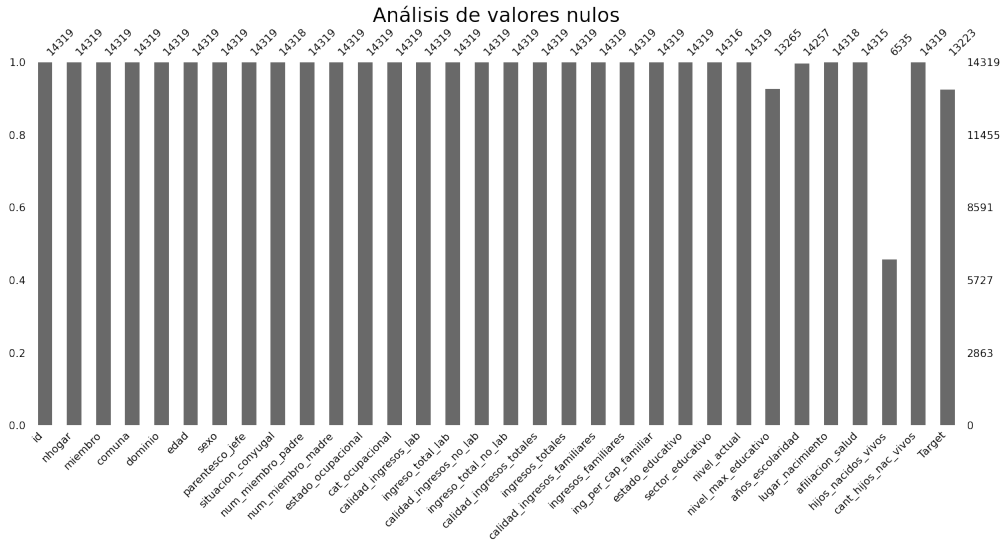

Figura 1: Nulos.

A partir del gráfico de barras, se ha identificado que la variable target posee 1054 valores nulos. Es importante tener este dato presente al momento de correr un algoritmo de clasificación. Por otro lado, los nulos correspondientes a la variable "hijos nacidos vivos" se dan ya que los hijos se cuentan siempre a la madre y no al padre para no duplicar sus valores.

3.1. Análisis univariado

Durante el apartado de análisis univariado se han aunado esfuerzos para analizar las variables de manera aislada. El foco principal se ha puesto en definir las variables que aportan información sobre los individuos que forman parte del dataset.

3.1.1. Género y edad

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

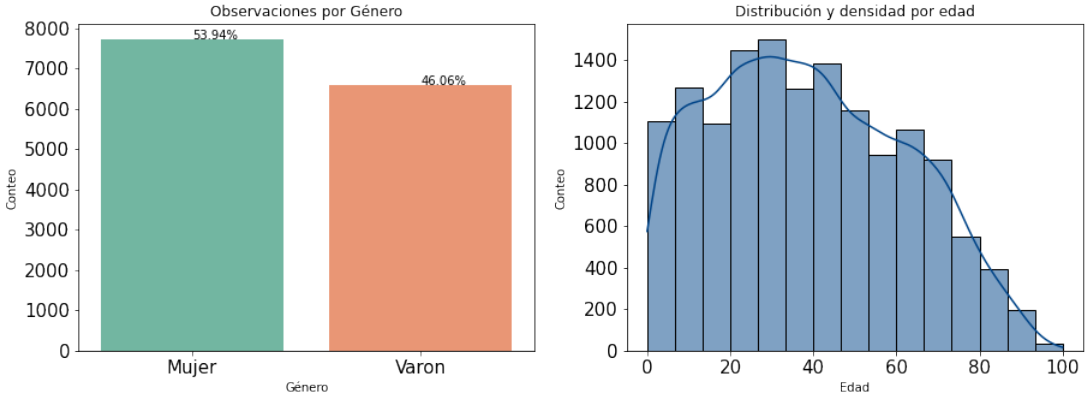

Figura 2: Análisis de genero y edad.

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

3.1.2. Comuna

Se ha continuado por evaluar la variable “comuna”. En la misma se muestra la comuna de la Ciudad de Buenos Aires del entrevistado, de manera de tener una ubicación geográfica. Se ha considerado primordial revisar esta variable a fin de corroborar que exista un balanceo de datos de cantidad de entrevistados pos comuna.

Para ello, se ha generado un mapa. En concreto se ha partido del mapa de comunas de la Ciudad de Buenos Aires y se han transformado las variables para fusionar el mapa con la base de datos de manera que coincidan.

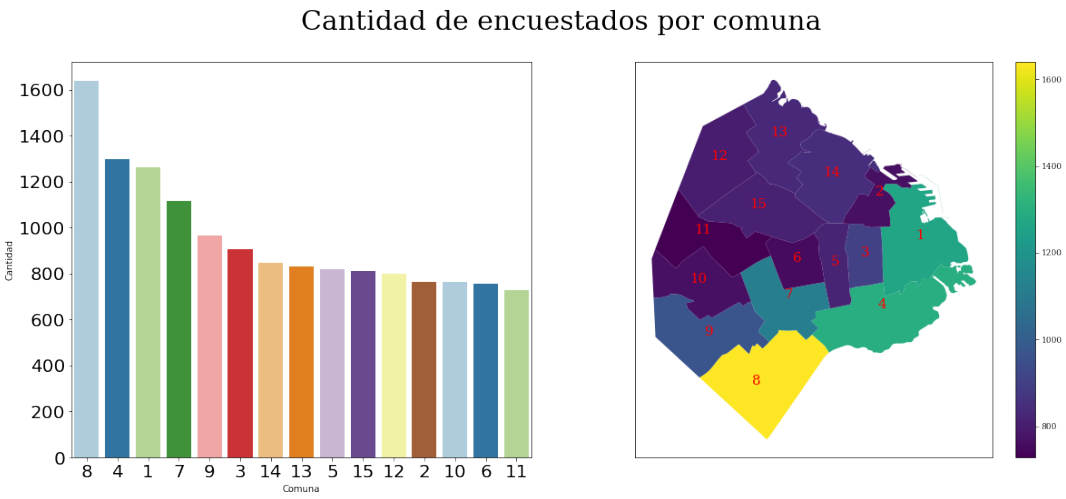

Figura 3: Encuestados por comuna.

Al observar estos gráficos se determina que las comunas 1,4,7 y 8 tienen mayor cantidad de casos. Queda por verse si en posteriores análisis será necesario abordar esta diferencia para evitar sesgos. Para eso, será necesario tomar en cuenta el porcentaje de la población total de cada comuna.

3.1.3. Años de escolaridad

En este apartado, mediante un gráfico de barras, se han analizado los años de escolaridad alcanzados por los encuestados:

Figura 4: Años de escolaridad alcanzados.

A simple vista se observan tres “picos”: en el valor mínimo, alrededor del 7.5 y alrededor del 12.5. Se ha inferido

que estos tres casos corresponden a no tener estudios, solo haber transcurrido el primario y haber transcurrido hasta la educación secundaria, respectivamente.

3.1.4. Máximo nivel educativo (Target)

Asimismo, se ha definido analizar el balance de los datos propios del Target. Se presentan los gráficos correspondientes.

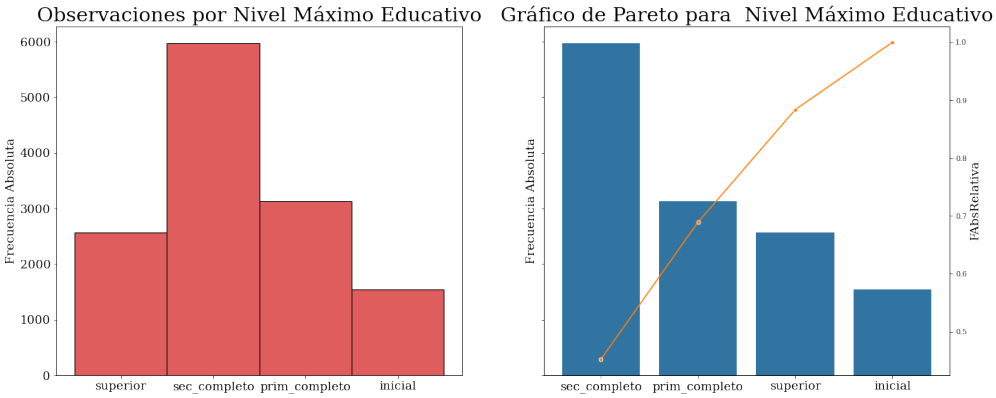

Figura 5: Distribución de datos del target.

Con respecto a la distribución de los datos del Target se determina que el nivel máximo educativo alcanzado con mayor frecuencia es el secundario completo, seguido por el primario. Adicionalmente, el nivel secundario y primario explican casi el 77 % de los datos, por lo tanto hay un gran desbalance en el Target. Esto se debe tener en cuenta al momento de alcanzar las conclusiones del proyecto.

3.1.5. Ingreso familiar per cápita

Finalmente, se ha definido evaluar una variable que pone como objeto de estudio al grupo familiar en vez del individuo: *ingresos familiares*. Al respecto, como sabemos que las variables de ingresos suelen tener outliers para los valores más altos, a la hora de graficar esta variable se ha decidido retirar los mismos. En el título del mismo se aclara el valor máximo de la variable contemplada.

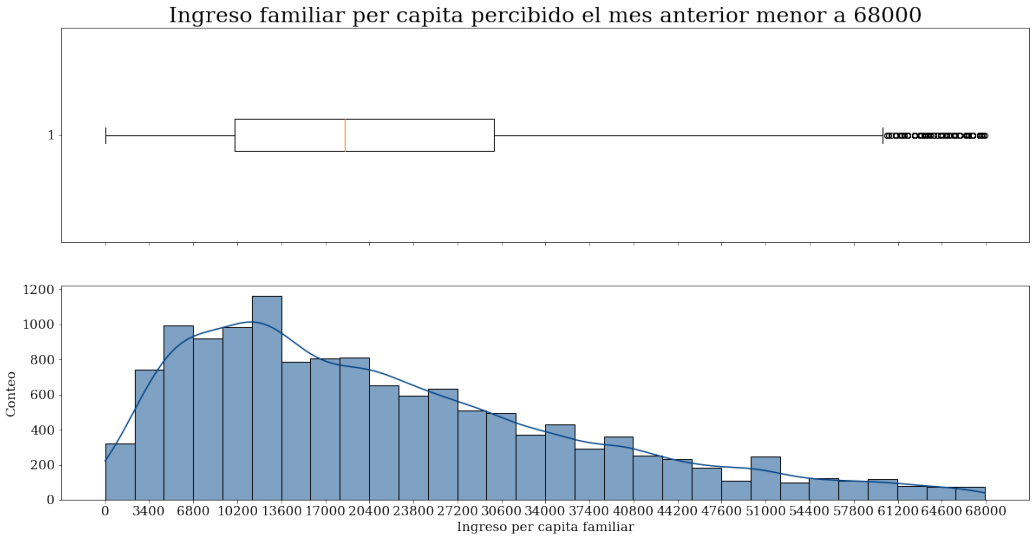

Figura 6: Análisis ingreso familiar.

Como se puede ver en la figura, la distribución de los ingresos familiares se encuentra **sesgada**, existe un desbalanceo de datos.

3.2. Análisis bivariado

3.2.1. Análisis de variables numéricas

Para comenzar el análisis bivariado del problema se han realizado diferentes heat maps para analizar, en primera instancia, la correlación entre las variables numéricas.

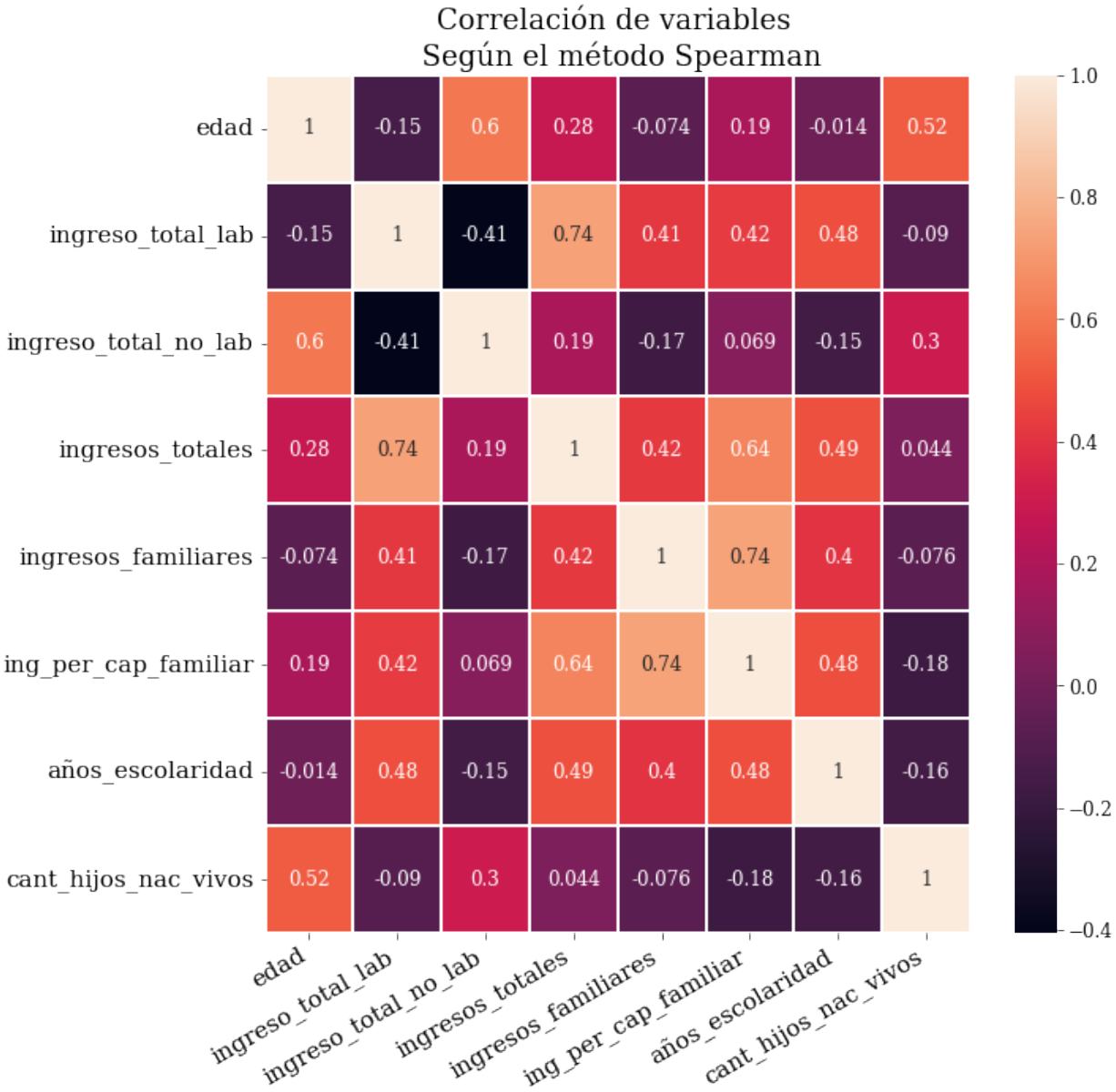

Figura 7: Correlación entre variables por método de Spearman.

En primera instancia, no se observan fuertes correlaciones. Sin embargo, se puede apreciar que la variable "años_escolaridad" correlaciona moderadamente bien con variables relacionadas al ingreso. En concreto, la principal correlación positiva es "años_escolaridad" con ingreso familiar per cápita, lo cual hace sentido teórico.

Por otra parte, al graficar en base a un threshold se puede observar que los años de escolaridad alcanzados por los entrevistados tienen una relación de 66 % con la variable "ingresos_totales"

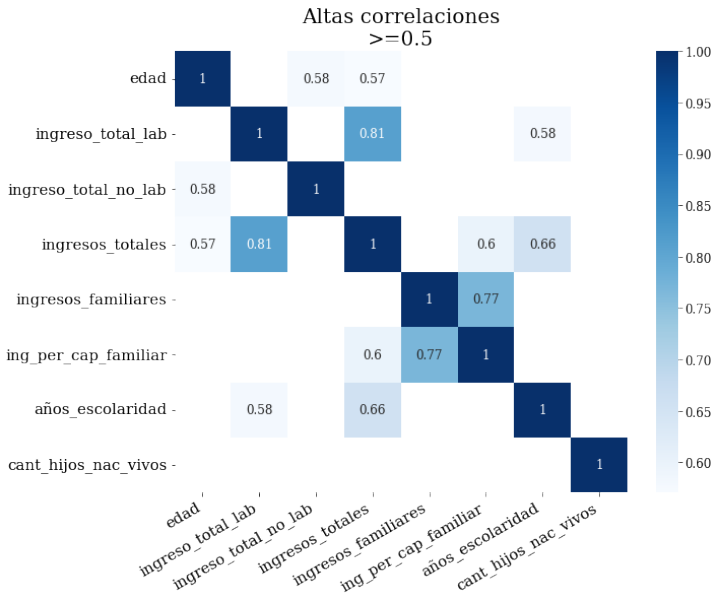

Figura 8: Correlación entre variables mayor 0,5(der).

Por último corremos una tabla de correlación y filtramos las de valores más altos

Variable 1	Variable 2	Correlación
ingreso_total_lab	ingresos_totales	0.80
ingresos_familiares	ing_per_cap_familiar	0.76
ingresos_totales	ing_per_cap_familiar	0.62
ingresos_totales	años_escolaridad	0.60
edad	ingreso_total_no_lab	0.57
ingreso_total_lab	años_escolaridad	0.54

Cuadro 3: Tabla de correlaciones.

A modo de conclusión del análisis preliminar bivariado, se puede identificar que:

- Existe una alta correlación entre las variables relacionadas al ingreso.
- Existe una alta correlación entre los ingresos y los años de escolaridad.
- Existe una relación positiva entre la edad y los ingresos totales.

Finalmente, se ha analizado la relación entre los ingresos totales de cada hogar y los ingresos familiares por edad:

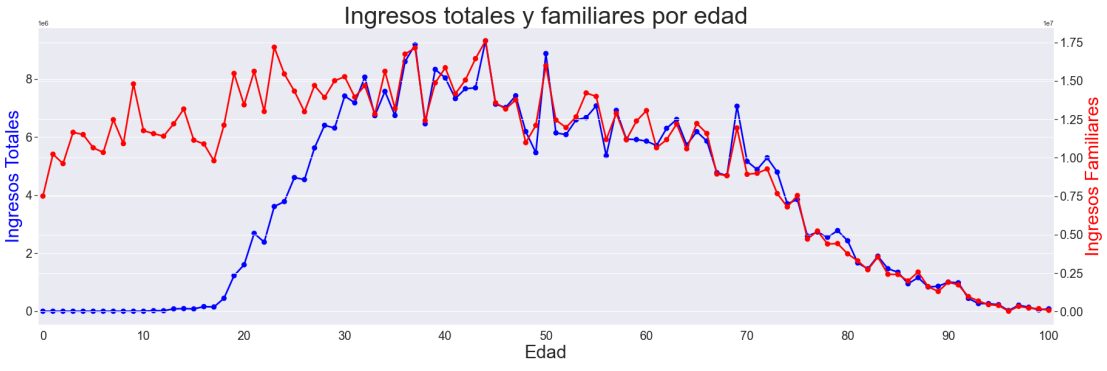

Figura 9: Ingresos totales y familiares por edad.

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.

3.2.2. Comparación de variables categóricas con numéricas

Dentro de esta sección se han comparado variables numéricas con otras categóricas, como es el caso del Target. Se ha comenzado por establecer la relación entre el sexo y variables numéricas propias de las categorías de ingreso, características de los encuestados y educación.

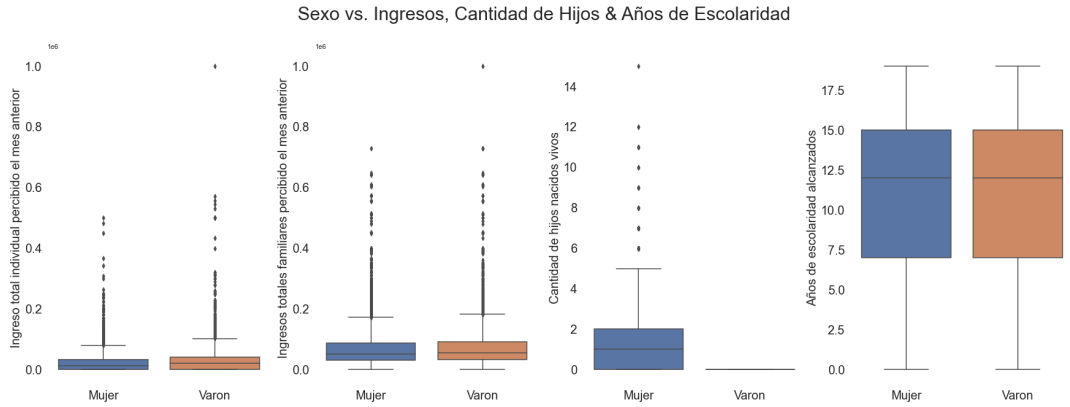

Figura 10: Correlación entre sexo y variables numéricas.

Como se adelantó previamente, los encuestados varones figuran sin hijos ya que los mismos se asignan unicamente a la madre a fin de no duplicarlos en el dataset. Por otro lado, las distribuciones de la variable años de escolaridad son similares para varones y mujeres. Respecto a las variables relacionadas con el ingreso, se ha decidido quitar outliers de las correlaciones anteriores a fin de establecer relaciones claras.

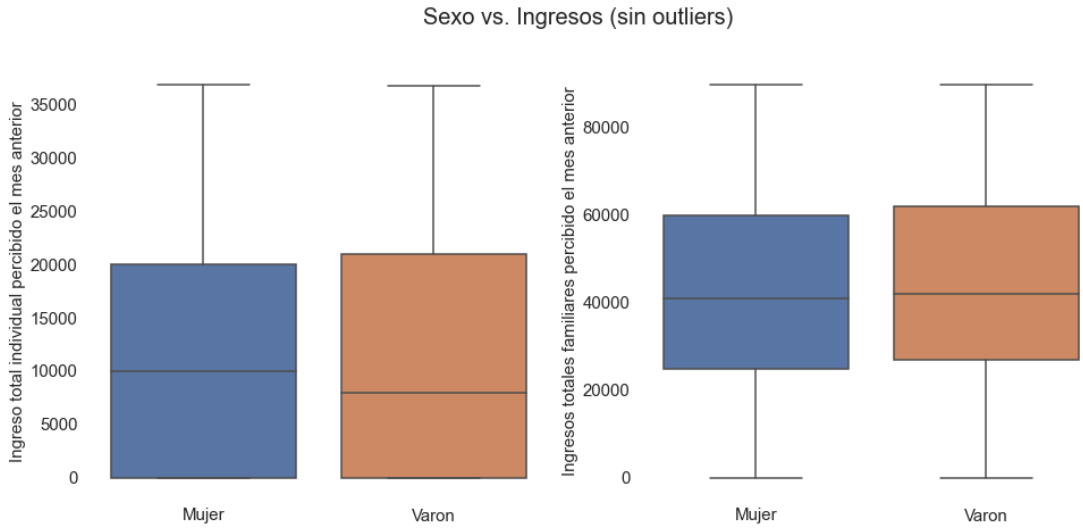

Figura 11: Correlación entre sexo y variables numéricas sin outliers.

Mediante el análisis de las figuras se determina que en el caso de los ingresos totales individuales los varones perciben mayores ingresos. Sin embargo, la diferencia no es significativa. En el caso de los ingresos familiares la distribución se da de manera similar.

Luego del análisis bivariado preliminar realizado para enriquecer la descripción del dataset, se ha decidido correlacionar al Target con los ingresos.

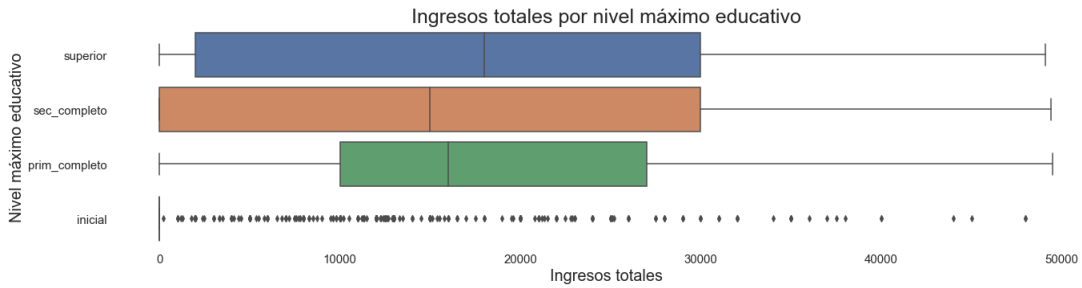

Figura 12: Correlación entre ingresos totales y Target

A simple vista puede apreciarse que, para el nivel inicial, la remoción de outliers en otra categoría sigue siendo insuficiente para mostrar la distribución real de la variable. Por lo tanto se han analizado con mayor profundidad los valores de esta categoría:

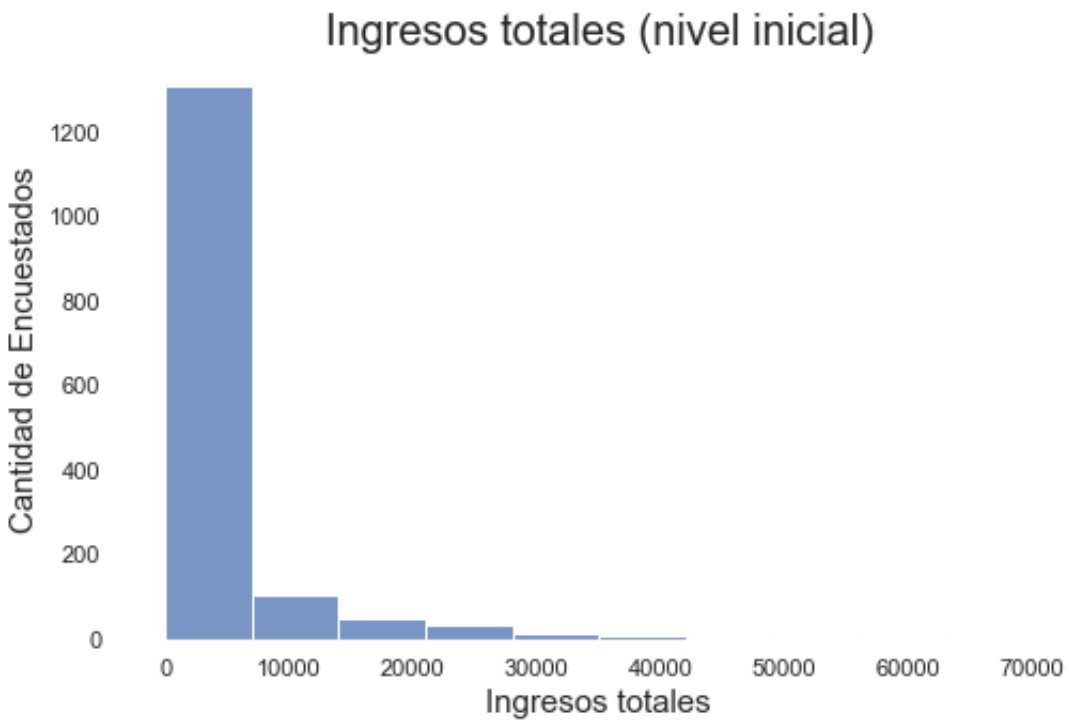

Figura 13: Ingresos totales para encuestados de nivel inicial

Lógicamente, la enorme mayoría de los ingresos tienen el valor inicial de 0, puesto que incluye a personas que en ese momento estaban cursando su educación inicial, por lo que tenían entre 2 y 6 años.

3.2.3. Variable numéricas con comuna

Para relacionar variables numéricas con la comuna, se ha analizado la interacción entre la frecuencia para cada categoría del Target y las comunas mediante mapas de calor:

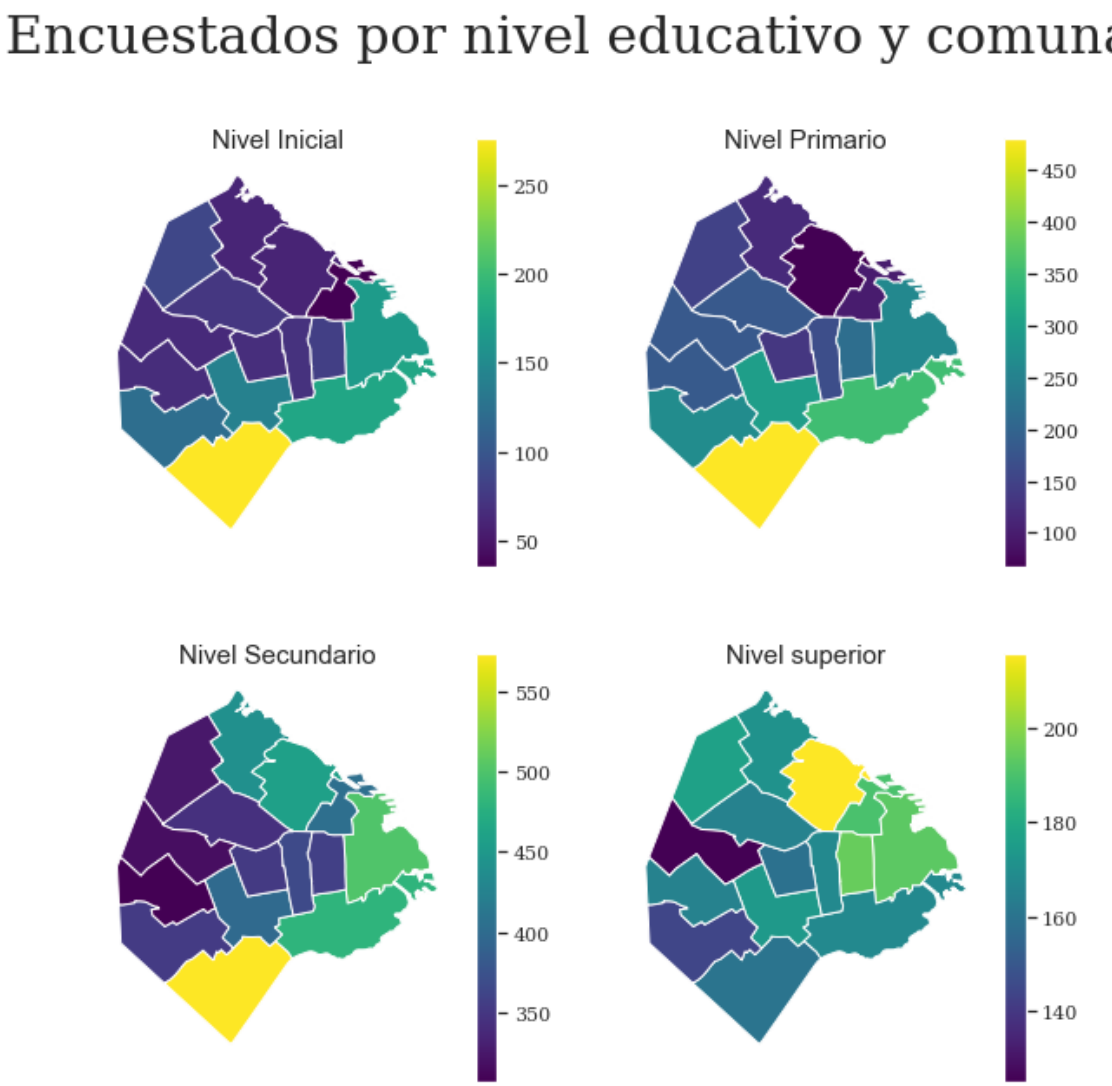

Figura 14: Encuestados por nivel educativo y comuna

Se ha observado que en el sur de la ciudad hay mayor cantidad de encuestados con niveles de inicial, primario y secundario completo, mientras que el norte (particularmente el barrio de Palermo) tiene mayor cantidad de personas con estudios superiores. En menor medida también las comunas del este (comúnmente llamado el "centro de la ciudad") destacan por la cantidad de encuestados con nivel superior.

A su vez, se presenta otro heat map que reúne en un único gráfico el promedio de años de escolaridad por comuna. Este gráfico refuerza la relación presentada en la figura anterior.

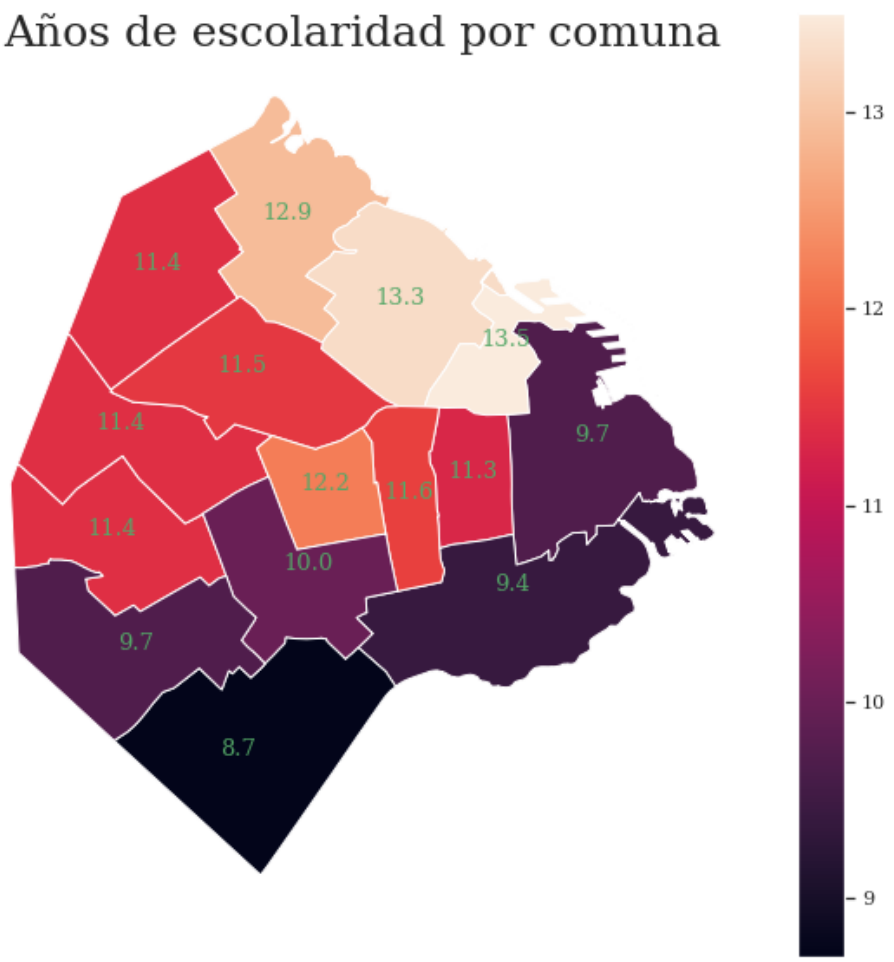

Figura 15: Años de escolaridad por comuna

A partir de los últimos dos gráficos se ha establecido una clara división geográfica del nivel educativo:

- Las comunas del norte son las que tienen mayor nivel educativo.
- Las comunas del centro tienen niveles medios.
- Las comunas del sur (con las comuna 6 en el centro de la ciudad como outlier) y la comuna 1 en el este son las que tienen niveles más bajos.

Esta relación encontrada aporta gran información a la descripción del dataset en términos de relación entre el Target y la distribución geográfica d ela población.

3.3. Análisis multivariado

Durante esta sección se ha trabajado a fin de establecer relaciones entre más de 2 variables. En concreto, se ha buscado relacionar al Target con variables de las categorías de ingresos, educación, geolocalización, entre otras. En ese sentido, se ha comenzado por establecer la relación entre años de escolaridad, nivel máximo educativo (Target) y los ingresos totales.

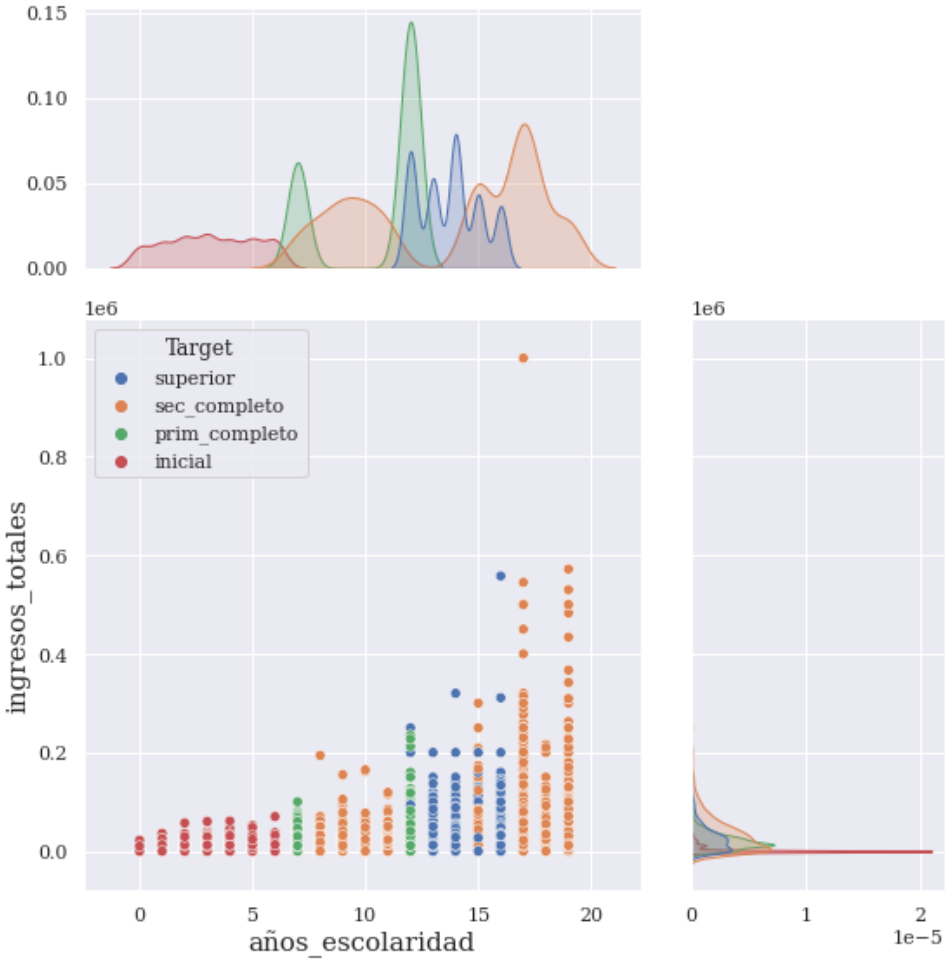

Figura 16: Relación entre años de escolaridad, Target e ingresos totales

Al visualizar se ha concluido que:

- Hasta los 6 años todos los casos llegan al nivel inicial.
- En dos años en que aparece el primario completo: 7 y 12 años. Se estima que se debe a la división entre los que comenzaron su educación en la primaria y los que comenzaron en el nivel inicial.
- A partir de los 12 años se observa un aumento consistente de los ingresos totales.

A su vez, se ha buscado relacionar el Target, con los ingresos y el dominio.

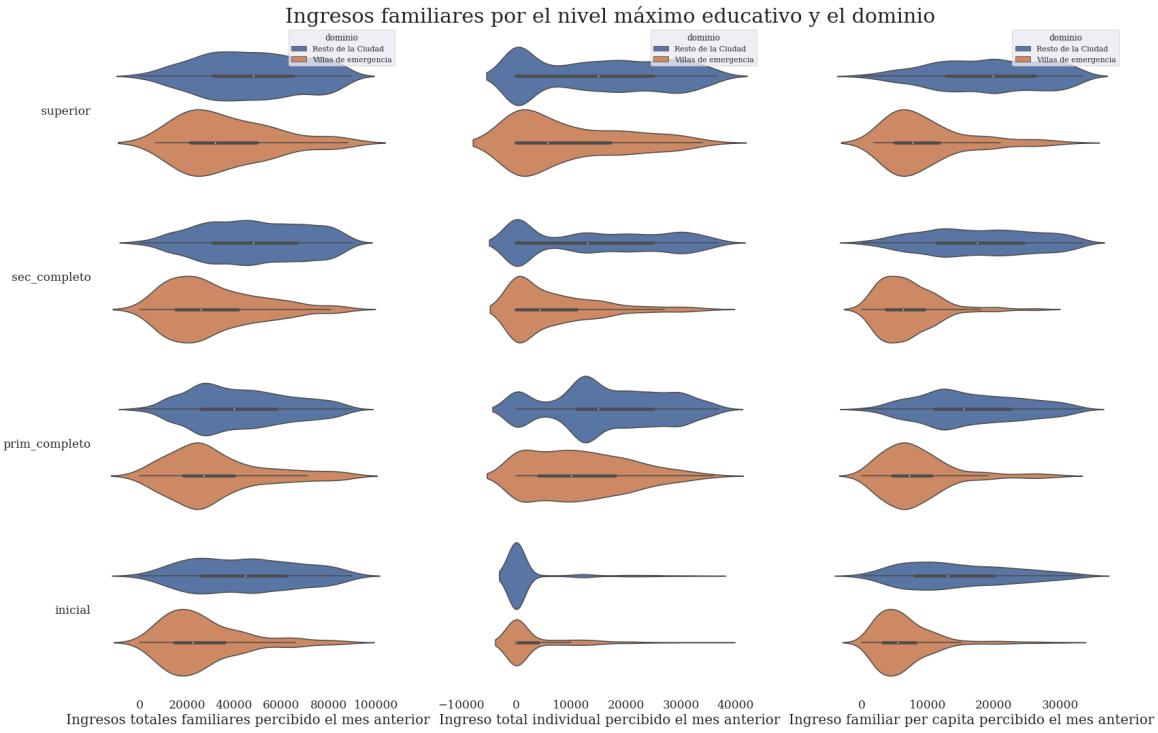

Figura 17: Relación entre dominio, Target e ingresos familiares

En este punto se ha alcanzado una conclusión interesante: no importa el nivel máximo educativo, los casos que no provienen de villas de emergencia (dominio="Resto de la Ciudad") obtienen en promedio ingresos más altos en todos los niveles educativos. El alcanzar estudios superiores no parece homogeneizar ambos conjuntos. Esto se puede observar en el segundo gráfico, ya que el violín naranja acumula mayor cantidad de casos hacia la derecha, en comparación con los violines azules que tienen una mayor distribución.

Continuando con el análisis a nivel grupo familiar, se ha evaluado la relación entre ingresos familiares y comuna según el máximo nivel educativo alcanzado (Target):

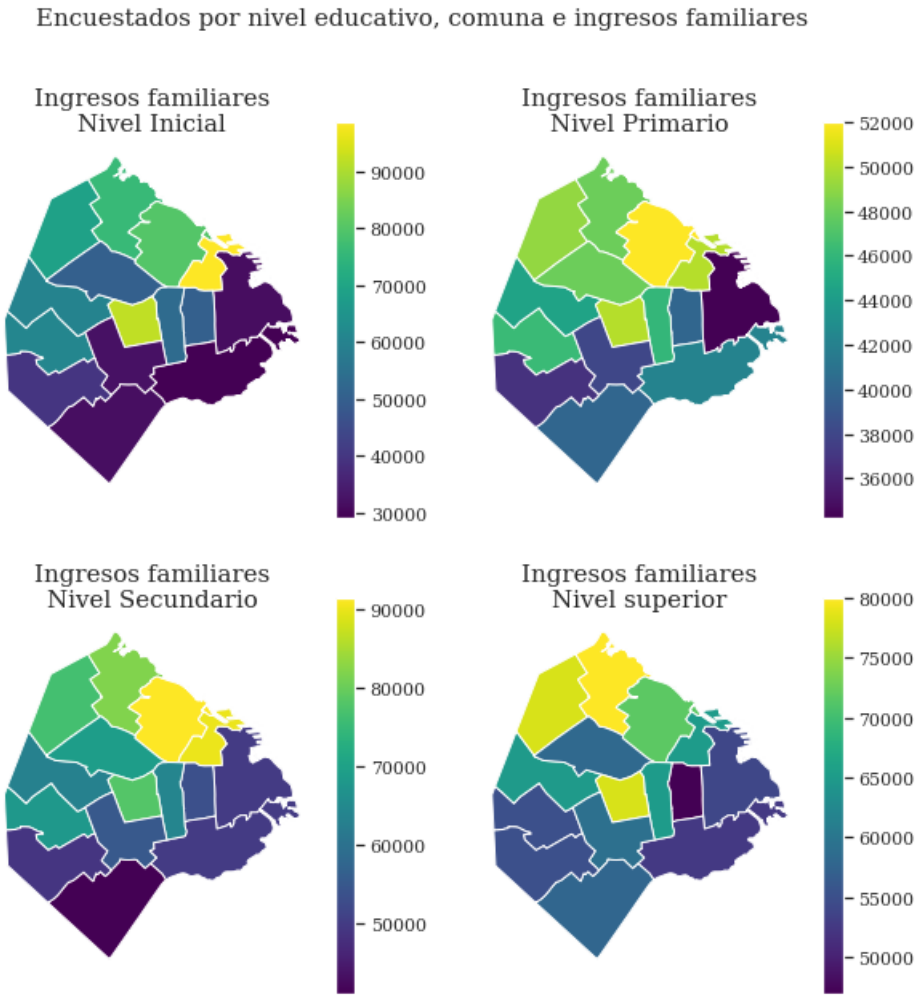

Figura 18: Relación entre comuna, Target e ingresos familiares

En esta instancia se ha visualizado que a medida que avanza el nivel educativo máximo (Target) se atenúan levemente las diferencias de ingresos familiares entre comunas. Queda pendiente cruzar estos datos con la edad, para saber si el hecho de incluir a menores de edad está sesgando los valores para nivel inicial, primario y secundario.

4. Modelos analíticos

En esta sección se ha trabajado en pos de utilizar algoritmos que modelen los datos de manera correcta y así poder alcanzar los objetivos propuestos para el proyecto. En concreto, se ha buscado clasificar el Target a partir de las variables restantes del dataset utilizando machine learning. No obstante, se ha comenzado por transformar algunas variables para poder trabajar con los algoritmos:

- Se ha recategorizando la variables "Target" en variables numéricas, es decir, a cada nivel educativo le asignamos un valor numérico del 1 al 4:
  - **inicial** = 1,
  - **prim_completo** = 2,
  - **sec_completo** = 3,
  - **superior** = 4.
- Se ha reagrupado la variable "comuna" por regiones para reducir la dimensionalidad (en norte, centro, sur y oeste).
- Y por último, se han renombrado algunas variables para que sean más cortas.

Como resultado han quedado las siguientes variables:

Tamaño del set de datos: 14319 entries, 0 to 14318
Con un total 33 columnas:

#	Columnas	Tipo de Dato	Entradas No Vacías
0	id	objeto	14319
1	nhogar	objeto	14319
2	miembro	objeto	14319
3	comuna	objeto	14319
4	dominio	objeto	14319
5	edad	numérica	14319
6	sexo	objeto	14319
7	parentesco_jefe	objeto	14319
8	situacion_conyugal	objeto	14318
9	num_miembro_padre	objeto	14319
10	num_miembro_madre	objeto	14319
11	estado_ocupacional	objeto	14319
12	cat_ocupacional	objeto	14319
13	calidad_ingresos_lab	objeto	14319
14	ingreso_total_lab	numérica	14319
15	calidad_ingresos_no_lab	objeto	14319
16	ingreso_total_no_lab	numérica	14319
17	calidad_ingresos_totales	objeto	14319
18	ingresos_totales	numérica	14319
19	calidad_ingresos_familiares	objeto	14319
20	ingresos_familiares	numérica	14319
21	ing_per_cap_familiar	numérica	14319
22	estado_educativo	objeto	14319
23	sector_educativo	objeto	14316
24	nivel_actual	objeto	14319
25	nivel_max_educativo	objeto	13265
26	años_escolaridad	float64	14257
27	lugar_nacimiento	objeto	14318
28	afiliacion_salud	objeto	14315
29	hijos_nacidos_vivos	objeto	6535
30	cant_hijos_nac_vivos	numérica	14319
31	Target	Numérica	13223
32	region	objeto	14319

Tipos de datos: numérica(9), objeto(24)

Cuadro 4: Tabla de variables modificadas

4.1. Tratados de nulos

A la hora de eliminar los nulos del set de datos, se ha definido armar una función para tener una lista limpia de variables con nulos, se comparten los resultados obtenidos:

Nulos por variable

Variable	Nulos
situacion_conyugal	1
lugar_nacimiento	1
sector_educativo	3
afiliacion_salud	4
años_escolaridad	62
nivel_max_educativo	1054
Target	1096
hijos_nacidos_vivos	7784

Entonces, cada una de las variables se ha tratado de manera particular dependiendo los tipos de datos que aloja. En ese sentido, para eliminar los valores nulos de la variable "años_escolaridad", la cual tiene datos numéricos, se han reemplazado con la mediana por comuna y sexo. Por otro lado, a los nulos de las variables categóricas: *"lugar_nacimiento"*, *"situacion_conyugal"* , *"afiliacion_salud"*, *"sector_educativo"* y *"hijos_nacidos_vivos"* se han reemplazado con la moda.

Finalmente, se ha eliminado la variable "nivel_max_educativo" ya que no es de interés. Asimismo, se han eliminado los nulos del Target y se ha pasado el tipo de dato de la misma a entero.

4.2. Target

4.2.1. Borrado de variables

Hay muchas variables que no se han considerado esenciales dado que no aportan información relevante en el algoritmo de clasificación. Precisamente, brindan información repetida. A continuación se comparten las categorías que se han descartado para correr el algoritmo:

Variables	Motivo de eliminación
id:	no suma información para la clasificación,
nhogar:	no suma información para la clasificación,
parentesco_jefe:	no suma información para la clasificación,
miembro:	no suma información para la clasificación,
num_miembro_padre:	no suma información para la clasificación,
num_miembro_madre:	no suma información para la clasificación,
cat_ocupacional:	brinda la misma información que estado_ocupacional,
calidad_ingresos_lab:	brinda la misma información que ingreso_total_lab,
calidad_ingresos_no_lab:	brinda la misma información que ingreso_total_no_lab,
calidad_ingresos_totales:	brinda la misma información que ingresos_totales,
calidad_ingresos_familiares:	brinda la misma información que ingreso_familiares,
estado_educativo:	no aporta información para la clasificación,
nivel_actual:	no aporta información para la clasificación,
hijos_nacidos_vivos:	brinda la misma información que cant_hijos_nac_vivos,
comuna:	variable ya abordada en la variable 'región'.

Cuadro 5: Variables eliminadas

Y como resultado, se tiene el dataset listo para el procesamiento:

Tamaño del set de datos: 14319 filas, 0 to 14318
Con un total 17 columnas:

#	Columnas	Tipo de Dato	Entradas No Vacías
0	dominio	objeto	13223
1	edad	numérica	13223
2	sexo	objeto	13223
3	situacion_conyugal	objeto	13223
4	estado_ocupacional	objeto	13223
5	ingreso_total_lab	numérica	13223
6	ingreso_total_no_lab	numérica	13223
7	ingresos_totales	numérica	13223
8	ingresos_familiares	numérica	13223
9	ing_per_cap_familiar	numérica	13223
10	sector_educativo	objeto	13223
11	años_escolaridad	float64	13223
12	lugar_nacimiento	objeto	13223
13	afiliacion_salud	objeto	13223
14	cant_hijos_nac_vivos	numérica	13223
15	Target	int32	13223
16	region	objeto	13223

Tipos de datos: numérica(9), objeto(8)

Cuadro 6: Dataset listo para el procesamiento

4.3. Procesamiento

Para preparar los datos para el modelado se han generado las siguientes transformaciones al dataframe:

- Dividimos el dataframe en X_train, y_train, X_test e y_test, haciendo la división entre test y el train en un 30 % y un 70 % respectivamente, con una semilla especifica.
- Y procesamos el X_train y el X_test con un pipeline generado previamente, el cual convierte las variable numéricas con el minmaxscaler y las categóricas con one hot encoding.

Una vez aplicadas dichas transformaciones al dataframe, tenemos listos los elementos para entrenar el modelo (con la misma cantidad de columnas): *X_train*, *y_train*, *X_test* e *y_test*.

Para que nuestro trabajo sea reproducible y, al mismo tiempo, podamos trabajar con la aleatoriedad, decidimos utilizar el parámetro random_state = 50, tanto para el procesamiento del dataframe como para el entrenamiento de todos nuestro modelos.

4.4. Árbol de decisión

4.4.1. Primer modelo

Como primera aproximación, se ha utilizado un árbol de clasificación usando los parámetros:

- max_depth = 8,
- criterion = 'gini',

para saber como performa y mejorarlo a partir de ahí.

Como resultado se ha observado que el Accuracy score para el test ha sido de: 0.940 y la matriz de confusión ha dado:

	Predicc. Inicial	Predicc. Primario	Predicc. Secundario	Predicc. Superior
Inicial	445	0	1	0
Primario	0	961	8	9
Secundario	0	19	1693	59
Superior	1	57	84	630

Cuadro 7: Matriz de confusión primer modelo

Por lo tanto, se han alcanzado las siguientes métricas:

	precision	recall	f1-score	support
Inicial	1.00	1.00	1.00	446
Primario	0.93	0.99	0.95	978
Secundario	0.95	0.96	0.95	1771
Superior	0.90	0.82	0.86	772
accuracy			0.94	3967
macro avg	0.94	0.94	0.94	3967
weighted avg	0.94	0.94	0.94	3967

Cuadro 8: Métricas primer modelo

A simple vista, se aprecia que el modelo performa muy bien, dado su accuracy. Cabe destacar que se observa una diferencia de 14 puntos en el f1-score de la categoría superior con respecto al de la incial, mientras que para las otras dos categorías es solo de 5 puntos. Asimismo, se ha calculado su sesgo y su varianza:

- **Bias o sesgo:** 96.89 % lo cual indica poco error, es decir, un sesgo bajo,
- **Variance=Test_Score - Bias:** 2.89 %, lo que indica que la varianza también es baja.

Resultados

Entonces, el modelo tiene una **buena relación** de sesgo y varianza.

De aquí, ha sido necesario ver cuáles eran las variables más importantes para el armado del modelo. Esto permite volver el modelo más robusto. Al quitar las mismas. se ha determinado que la variable "años_escolaridad" tiene una importancia del 84 %, por mucho superior al resto de variables.

Este hecho resulta contradictorio dado que, durante el análisis EDA, no se ha visto un resultado que indique tal relación, por otro lado, si se analiza la relación conceptual entre ambas variables, resulta evidente.

Por lo tanto, se ha definido desarrollar un nuevo modelo sin esta variable. El principal motivo es que los años de escolaridad es un dato que puede constatarse de forma conjunta con el nivel máximo educativo, por lo que tiene sentido que si no se cuenta con la variable target, tampoco se tenga en consideración la variable de los años de escolaridad.

4.4.2. Segundo modelo

Esta vez al correr el modelo, utilizando el nuevo dataset, se ha definido aplicar el "DecisionTreeClassifier" solo con el parámetro:

- cirterion='entropy'.

Esto ha dado como resultado:

	precision	recall	f1-score	suppo
Inicial	0.83	0.79	0.81	446
Primario	0.45	0.26	0.33	978
Secundario	0.56	0.66	0.60	1771
Superior	0.39	0.45	0.42	772
accuracy			0.53	3967
macro avg	0.56	0.59	0.54	3967
weighted avg	0.53	0.53	0.52	3967

Cuadro 9: Métricas segundo modelo.

Luego, se ha evaluado el sesgo y la varianza:

- **Bias o sesgo:** 99.78 % que indica poco error, es decir, un sesgo bajo,
- **Variance=Test_Score - Bias:** 46.39 %, lo que indica un nivel de varianza alto,

Analizando lo observado, se ha determinado que este modelo hacía **OVERFITTING**, y tenía un rendimiento bastante pobre respecto al f1-score. En consecuencia, el árbol performa peor sin esta variable, aumentando, especialmente, la varianza. Por lo tanto se ha optado por probar con un grid search.

4.4.3. Gridsearch con CV y Tercer Modelo

Entre la grilla de parámetros para el Gridsearch se han elegido los siguientes:

- max_depth: range(5,11),
- max_features: range(1,14),
- criterion: [‘gini’,‘entropy’,‘log_loss’];

a su vez realizamos un cross-validation con 10 particiones, usando todos los procesadores.

Esto ha mostrado que el mejor árbol de decisión posible obtiene 0.642. Para alcanzarlo, el árbol debe tener una profundidad de 6, utilizar 10 variables y usar el método “gini”.

Por lo tanto, se ha entrenado el modelo bajo estos mismos parámetros y obtenido el siguiente reporte de clasificación:

	precision	recall	f1-score	suppo
Inicial	0.99	0.74	0.85	446
Primario	0.49	0.21	0.30	978
Secundario	0.55	0.87	0.69	1771
Superior	0.78	0.41	0.54	772
accuracy			0.60	3967
macro avg	0.70	0.56	0.59	3967
weighted avg	0.63	0.60	0.57	3967

Cuadro 10: Métricas segundo modelo mejorado.

Luego, se han investigado el sesgo y la varianza:

- **Bias o sesgo:** 65.27 % que indica que se tienen muchos errores, es decir, un bias alto,
- **Variance=Test_Score - Bias:** 5.14 %, por el otro lado la varianza es baja.

Por lo cual, el modelo hizo **UNDERFITTING**, y no se han notado grandes mejorías en cuanto al f1-score.

Resultados

El utilizar el grid search ha permitido mejorar bastante el modelo, el cual había perdido bastante accuracy al retirar los años de escolaridad. La métrica que más se ha podido mejorar con este método fue la varianza, que pasó de 44.97 % a 5.14 %, pero en contraparte, el acuracy bajo de 100 % a 65.27 %.

4.5. Random Forest Classifier

En una instancia posterior, se ha trabajado con random forest con el propósito de analizar los resultados que arroja este algoritmo. Se ha comenzado teniendo en cuenta todo el dataset, incluido la variable con los años de escolaridad.

4.5.1. Cuarto modelo

Como tercer modelo con el Random Forest Classifier, se han elegido los siguientes parámetros:

- n_estimators=200,
- max_depth=15,
- criterion='gini'.

Se han obtenido los siguientes resultados en cuanto a las métricas:

	precision	recall	f1-score	suppo
Inicial	1.00	0.96	0.98	446
Primario	0.86	0.95	0.90	978
Secundario	0.91	0.93	0.9	1771
Superior	0.92	0.76	0.83	772
accuracy			0.91	3967
macro avg	0.92	0.90	0.91	3967
weighted avg	0.91	0.91	0.90	3967

Cuadro 11: Métricas tercer modelo

El random forest ha performado bastante bien, es decir, mucho mejor que los modelos anteriores:

- **Bias o sesgo:** 97.80 % indica que existen pocos errores, es decir hay un sesgo bajo,
- **Variance=Test_Score - Bias:** 7.20 %, que indica que la varianza es baja.

Resultados

Se ha obtenido un buen modelo, que no tiene grandes diferencias al mirar el f1-score. De todas maneras, se han buscado cuales eran las variables más importantes. En ese sentido, se ha encontrado que la variable años de escolaridad redujo la enorme importancia (a un 43.58 %) que tenía en el random tree. Sin embargo, para poder comparar, sigue correspondiendo quitarla del modelo. Posteriormente, se ha probado un modelo alternativo sin la variable años de escolaridad.

4.5.2. Quinto modelo

En este cuarto caso se han elegido los siguientes parámetros:

- n_estimators=200,

- max_depth=10,
- criterion='gini'.

Se muestran los resultados de las medidas de desempeño:

	precision	recall	f1-score	suppo
Inicial	0.95	0.78	0.86	446
Primario	0.54	0.23	0.32	978
Secundario	0.56	0.88	0.69	1771
Superior	0.80	0.40	0.53	772
accuracy			0.62	3967
macro avg	0.71	0.57	0.60	3967
weighted avg	0.65	0.62	0.59	3967

Cuadro 12: Métricas cuarto modelo

Para este caso también se han buscado los valores del sesgo y la varianza:

- **Bias o sesgo:** 89.11 % que indica que se tienen pocos errores, es decir, que el sesgo es bastante bajo,
- **Variance=Test_Score - Bias:** 27.4 %, esto indica que existe un valor alto en la varianza.

A modo de conclusión, respecto al tercer modelo, ha empeorado su accuracy pero está muy cercano al mejor modelo de Decision Tree, mientras que crece mucho la varianza a un 27.4 % (unos 20 puntos). Al igual que con el el DesicionTree, se ha probado mejorándolo con grid search.

4.5.3. Gridsearch con CV y Sexto Modelo

En la grilla de parámetros para el Gridsearch se han elegido los siguientes:

- max_depth: [5,7,10,15,None],
- max_features: [5,8,10,30,41],
- n_estimators: [200,300,500],
- criterion: ['gini','entropy','log_loss'];

a su vez realizamos un cross-validation con 10 particiones, usando todos los procesadores.

Como parte de los resultados, el mejor random forest posible ha obtenido 0.668. Para eso, el árbol ha tenido una profundidad de 15, utilizado 10 variables, 300 estimadores y utilizado el método "gini". Entonces, se ha entrenado el modelo bajo estos mismos parámetros y se ha obtenido el siguiente reporte de clasificación:

	precision	recall	f1-score	support
Inicial	0.95	0.79	0.86	446
Primario	0.56	0.23	0.33	978
Secundario	0.56	0.89	0.69	1771
Superior	0.80	0.40	0.54	772
accuracy			0.62	3967
macro avg	0.72	0.58	0.60	3967
weighted avg	0.65	0.62	0.59	3967

Cuadro 13: Métricas cuarto modelo mejorado.

- **Bias o sesgo:** 90.65 % que indica que hay pocos errores, es decir, el sesgo es bajo,

- **Variance=Test_Score - Bias:** 28.54 % por lo que la varianza también es alta.

Esto denota que este modelo está haciendo **OVERFITING** y continua desempeñándose mal al ver el f1-score.

Resultados

Al utilizar random forest se ha podido mejorar el sesgo y disminuir el underfitting en 2 puntos porcentuales aproximadamente. No obstante, se ha visto afectada la varianza en estos modelos, que pasó de estar alrededor del 5 % en el árbol de decisión mejorado a 28 %.

5. Conclusiones Finales

Finalmente, resta elegir el mejor modelo para realizar las predicciones. Para eso se han evaluado las métricas de cada uno de ellos, siempre teniendo en cuenta los modelos que no contienen la variable "años_escolaridad", y hacer un cuadro comparativo:

N°	Tipo de modelo	accuracy	sesgo	varianza	f1_inicial	f1_primario	f1_secundario	f1_superior
2	Árbol default	0.53	1.00	0.46	0.81	0.33	0.60	0.42
3	Árbol mejorado	0.60	0.65	0.05	0.85	0.30	0.69	0.53
5	Bosque default	0.62	0.89	0.27	0.86	0.32	0.69	0.53
6	Bosque mejorado	0.62	0.91	0.29	0.86	0.33	0.69	0.54

Cuadro 14: Tabla general de resultados

Con esta información se ha definido qué modelo nos conviene usar:

- El árbol default tiene el mejor resultado con respecto al sesgo, pero su varianza lo deja afuera de consideración.
- Por el contrario, el árbol mejorado tiene una varianza insuperable de 5 %, aunque con el menor puntaje con respecto al sesgo.
- El bosque default tiene resultados mixtos en ambas categorías.
- El bosque mejorado destaca por bajo sesgo pero su varianza es la segunda más alta.

Se considera que los finalistas son el árbol mejorado y el bosque mejorado. Llamativamente, ambos performan muy bien pero en métricas diferentes. A su vez, el accuracy de ambos difiere en apenas un 2 %. Por un lado, ambos modelos tienen un F1 Score similar, de 85 % y 86 % respectivamente. Sin embargo, el arbol mejorado presenta signos de underfitting cuando se lo compara con el otro modelo. A la inversa, el bosque mejorado muestra un escenario de overfitting con mayor varianza en la predicción. Cabe mencionar que aunque el bosque default tiene el mismo valor de F1 Score que el bosque mejorados, posee una probabilidad levemente mayor de underfitting.

En conclusión, en nuestra opinión, la mejor elección de modelo es el árbol mejorado, ya que tiene la mayor robustez para poder generalizar en caso de agregar nuevos datos al modelo que difieran en gran medida con los existentes dentro del modelo, a comparación de los otros modelos, ya que presenta muy bajo overfitting. Otra ventaja frente al bosque aleatorio es su mayor velocidad de entrenamiento, asì como su capacidad de ser visualizada en un gráfico.

En futuras líneas de investigación se debería investigar en profundidad el desbalanceo de datos propio del Target. Como se ha visto en la etapa de EDA, el Target tiene un importante desequilibrio en el volumen de datos de cada categoría. De continuar con el proyecto, se indagaría acerca de dicho desbalance y se mitigaría la problemática agregando datos en categorías con deficit y eliminando datos de categorías en exceso.