Universidad de
SanAndrés

Big Data

Trabajo Práctico 4

Albareti Renato, López Federico y Mettola Franco

Fecha de entrega: 21 de junio de 2024

Parte I: Análisis de la base de hogares y cálculo de pobreza

Ejercicio 1

Eliminamos 5 observaciones que corresponden a 4 servicios domésticos habitando en la vivienda y 1 pensionista que habitan una vivienda para la cual la información está contada en el otro hogar del mismo `CODUSU` y no proveen información de varias características que no son ingresos. Son una porción tan pequeña de la muestra total por lo que conviene sacarlos (de ambas).

La variable `II1` indica ambientes en el hogar mientras que la variable `IV2` indica ambientes en la vivienda. Es posible que sean colineales, por lo que podemos eliminar alguna de las dos o combinarlas de alguna manera. Vamos a eliminar la variable `IV2` y quedarnos con los ambientes del hogar.

Luego están las variables `IV8-IV11` que especifican a gran detalle cómo es el baño (la ubicación, si es inodoro, letrina, si tiene cloaca, etc.). Si incluimos todos esos datos nos quedarían 27 variables con *one-hot encoding* teniendo en cuenta cuáles se siguen de las anteriores. El problema de eso es que va a quedar sumamente 'sparse' y se podría simplificar a algo menos granular ya que es posible que el dato tan específico del baño sea una sobre-especificación. En la literatura encontramos que se usan 4 o 5 variables sobre el baño. Por ejemplo aquí: `https://documents1.worldbank.org/curated/en/946321468286464228/text/363070ESW0P0890LIC00PAPA01001502007.txt`, el Banco Mundial usa (para predecir consumo):

- (base category: without toilet)
- Private toilet with sewer system/septic tank
- No private toilet with sewer system/septic tank
- Private toilet with letrine or pit
- No private toilet with letrine or pit

Vamos a implementar algo similar combinando categorías:

- (base) `IV8=2` no tiene baño / letrina
- tiene dentro de la vivienda con cloaca (`IV8=1, IV9=1, IV11=1`)
- tiene fuera de la vivienda con cloaca (`IV8=1, IV9=2 OR IV9=3, IV11=1`)
- tiene dentro de la vivienda sin cloaca (`IV8=1, IV9=1, IV11=2 OR IV11=3 OR IV11=4`)
- tiene fuera de la vivienda sin cloaca (`IV8=1, IV9=2 OR IV9=3, IV11=2 OR IV11=3 OR IV11=4`)

Eliminamos `II3` ya que `II3_1` provee la misma información y de manera ordinal (con 0 indicando que usa 0 habitaciones como lugar de trabajo). Lo mismo con `II5` e `II6`.

Las tres variables en `IV12` (si está en un basural, en una villa o en una zona inundable) consideramos que pueden ser muy útiles para predecir ingresos.

La variable `IX_Tot` está dividida `IX_Men10` y `IX_Mayeq10`. Sería conveniente no tener todas ya que debería haber multicolinealidad perfecta si incluimos las 3. Dejamos el total y reemplazamos las divisiones con una variable que indique la proporción de menores de

10 en la casa. `IX_Men10/IX_Tot`. De esa manera se capturaría también implícitamente el efecto de la cantidad de mayores de 10. Pensamos que esas dos pueden ser buenos predictores de ingresos.

Las variables de estrategias `V1-V19_B` pueden ser todas útiles aunque pensamos que es posible que si no respondieron el ingreso nominal tampoco hayan respondido alguna de estas ya que constituyen una descripción del ingreso y sus fuentes. Si alguien no quiere decir cuánto gana quizás tampoco quiera decir cómo lo ganó.

Las variables de organización `['VII1_1', 'VII1_2', 'VII2_1', 'VII2_2', 'VII2_3', 'VII2_4']` indican qué miembro en particular realiza/ayuda con las tareas domésticas. Como es irrelevante las convertimos en dos *dummy*. La primera *dummy* es 1 si `VII1_1=96` (servicio realiza las tareas domésticas) y la otra es 1 `VII2_1=96` (servicio ayuda con las tareas domésticas). Para indicar si el hogar contrata servicio doméstico que puede servir de predictor de ingresos.

Ejercicio 3 - Outliers

Esta vez no borramos los valores negativos de ingreso sino que los identificamos con nan. Usamos la variable de deciles que es 12 cuando ITF no tuvo respuesta. En el trabajo anterior eliminabamos no respuestas marcadas con -9 y usabamos el 0 como indicador de no respuesta en ITF (cuando en realidad el 0 puede significar que no tiene ingresos, lo cual distinguimos gracias a los deciles).

Ejercicio 4 - Construcción de Variables

-

Ya mencionamos las variables de proporción de menores, tipo de baño y servicio doméstico que armamos en el primer ejercicio. A continuación mencionamos otras 3 que creamos.

Hacinamiento

La primera variable que construimos es un indicador de hacinamiento en los hogares. Es decir, relaciona la cantidad de hogares exclusivos para dormir con los miembros de un hogar. Formalmente,

$$Hacinamiento = \frac{\text{Ambientes exclusivos para dormir}}{\text{Miembros totales}}$$

Este indicador es relevante para obtener una primera impresión de la pobreza dado que los hogares pobres pueden caracterizarse por pocas habitaciones por miembro,

Ingresos por subsidio

Con esta variable, podemos conocer el peso de las transferencias públicas sobre el ingreso total. Si suponemos que la redistribución fiscal es realizada de manera correcta, los hogares con mayores subsidios deberían ser los que se encuentran en situaciones de vulnerabildad.

$$\text{Ingresos por subsidios} = \frac{\text{Ingresos por transferencias públicas}}{\text{Ingreso Total Familiar}}$$

Ocupaciones en el hogar

De acuerdo al Capítulo 9 (*Emprendedores a Regañadientes*) del libro Poor Economics, las personas pobres suelen tener varios empleos con escasa formación profesional dado que deben diversificar sus profesiones. De esta manera, pueden mantener varias ocupaciones al simultáneo para buscar mejores ingreso, aunque esto represente una menor especialización.

Formalmente, definimos nuestra variable de la siguiente manera

$$\text{Ocupaciones por hogar} = \frac{\text{Ocupaciones}}{\text{Miembros Totales}}$$

Ejercicio 5 - Estadísticas descriptivas

-

Histograma de hacinamiento

En este gráfico (Figura 2), podemos conocer la distribución de la variable que creamos recientemente.

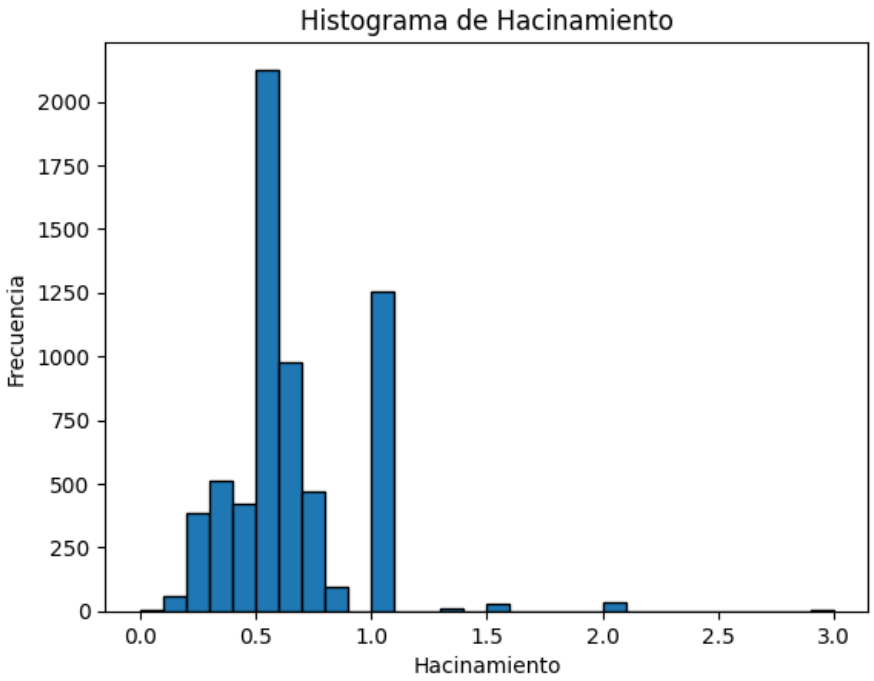

Figura 1: Distribución de hacinamiento

Como podemos ver, el promedio se encuentra cernano a las 0.75 habitaciones por miembro del hogar. Sin embargo, la distribución está sesgada a la izquierda de la media. Esto indica que gran parte de los hogres tienen meos de 0.5 dormitorios por pesona (lo que evivale a 2 personas por dormitorio).

Histograma de ingresos por subsidios

En este gráfico, podemos ver la distribución de la importancia de los subsidios relativo al ingreso total de la familia. Limitamos las observaciones para aquellos que reciben valores no nulos de transferencias para evitar un pico en el 0 %.

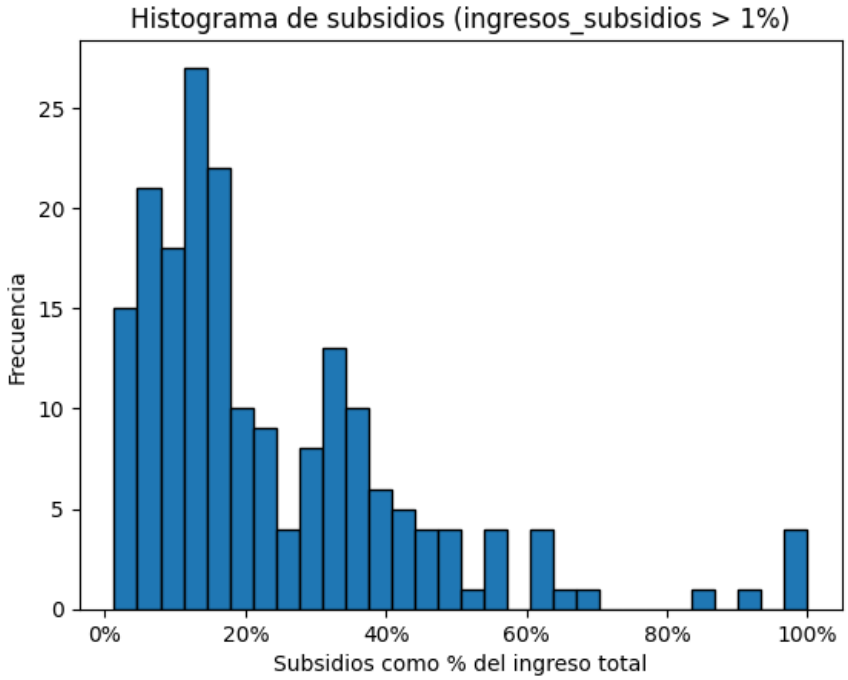

Figura 2: Distribución de ingresos por subsidio en relación al total

Como podemos observar, los resultados se centran en la parte derecha de la distribución. Esto indica que, en promedio, las transferencias públicas representan alrededor de un 20 % de los ingresos totales.

Distribución del nivel educativo

En este gráfico (Figura 3), observamos la distribución del máximo nivel educativo alcanzando por los encuestados. Para nuestro análisis, queremos enfocarnos en los individuos que no lograron alcanzar estudios básicos como el primario o el secundario ya que su escaso capital humano puede ser un predictor de vulnerabilidad social.

Duración del desempleo

En este gráfico (Figura 4), podemos obtener una aproximación de los desempleados que llevan más tiempo buscando empleo. Si la búsqueda se extiende por largos períodos de tiempo, esto puede indicarnos bajos niveles de educación o pocas calificaciones laborales que se relacionan con la posibilidad de continuar en la pobreza.

Acceso agua potable

En este gráfico (Figura 5), obtenemos información sobre el acceso a agua corriente que poseen los hogares. Bajo el enfoque de la pobreza multidimensional, podemos considerar como hogares pobres a aquellos que no tienen acceso a la red pública de agua.

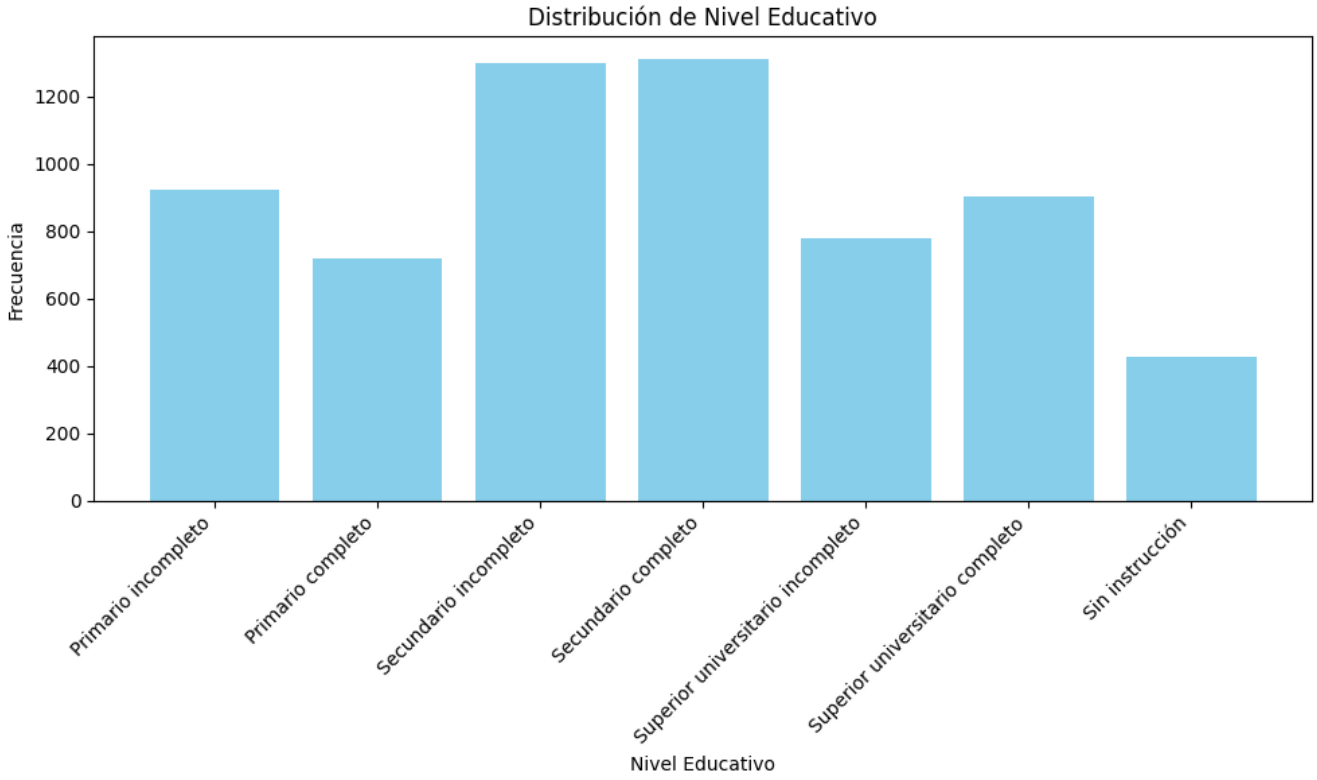

Figura 3: Individuos por cada nivel de estudios

Como vemos en el gráfico, la mayoría de los hogares poseen acceso a la red pública y solo una minoría debe emplear bombas a motor o manuales.

Ejercicio 6 - Relación ingresos familiares y miembros por hogar

En el siguiente gráfico (Figura 6), presentamos la regresión cuadrática entre la cantidad de miembros por hogar (independiente) y el ingreso total familiar (dependiente). Consideramos que la relación puede ir en ambos sentidos (no descartamos causalidad inversa)

A partir de esto, consideramos que la relación tiene esta forma ya que los hogares con pocos miembros (1 2) están vulnerables porque poseen menos personas en edad laboral. Por otro lado, los hogares con mayor cantidad de miembros pueden presentar bajos ingresos debido a que los hogares pobres suelen tener mas miembros. Por último, los valores medios del eje X indican hogares con una cantidad moderada de miembros y son los que disponen de un mayor ingreso familiar total.

Ejercicio 9 - Tasa de Pobreza en GBA

De acuerdo al INDEC, en el cuarto trimestre del 2023 la pobreza alcanzaba al 31.. % de los hogares. Con respecto a nuestra estimación, obtenemos una tasa de pobreza a nivel hogar del 44.32 %.

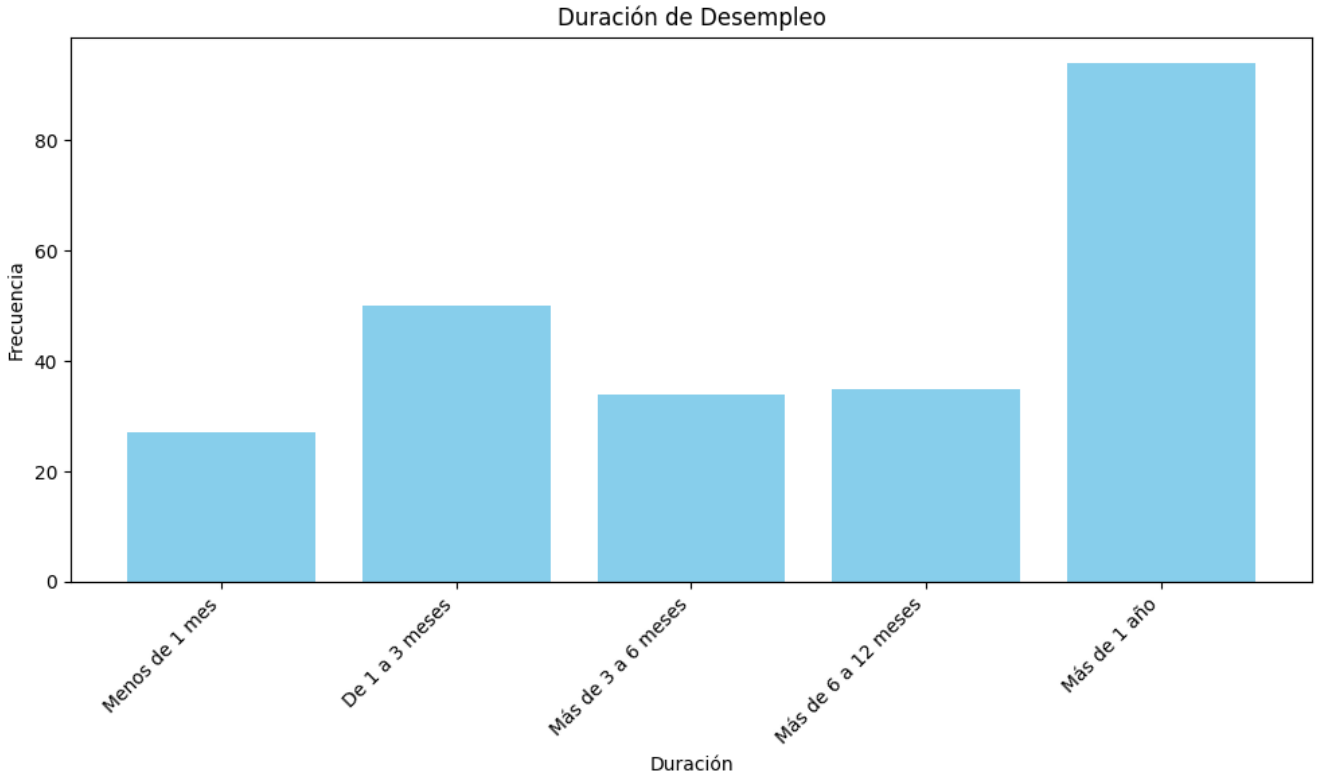

Figura 4: Duración del desempleo

1. PARTE III

3

Primero hay que generar un linspace de posibles parámetros para probar. Si se usa un lambda pequeño es como no estar regularizando y puede haber overfitting. Si es grande se puede alejar demasiado de la función real y estar muy sesgado. Con cross validation se separa la base de entrenamiento en k folds y se entrena con k-1 folds. El que queda se usa para validar. Repetimos k veces sin repetir folds. Habiéndolo calculado tantas veces se elige el de menor ecm. Es bueno no usar el conjunto de prueba que separamos al inicio para tener mas datos sobre los que probar que no haya overfitting

4

un k muy pequeño es más facil de computar porque iteras menos veces ya que son menos folds. El problema es que los resultados varían mucho entre diferentes realizaciones (alta varianza) ya que una parte demasiado grande de la muestra estaría siendo dejada afuera. Está claro que no es bueno usar demasiados pocos datos para entrenar. Por otro lado, si k es muy grande es obviamente más costoso de computar pero además entre un fold y otro no cambia mucho la muestra de entrenamiento (si k=n cambiaria por un solo dato) lo cual va a ser problemático al computar los errores (están muy correlacionados). Pero en general aumentar k va a reducir la varianza que tenias con k chico, pero puede aumentar el sesgo. El caso extremo que mencionamos de k=n s ellama leave one out (LOOCV) y

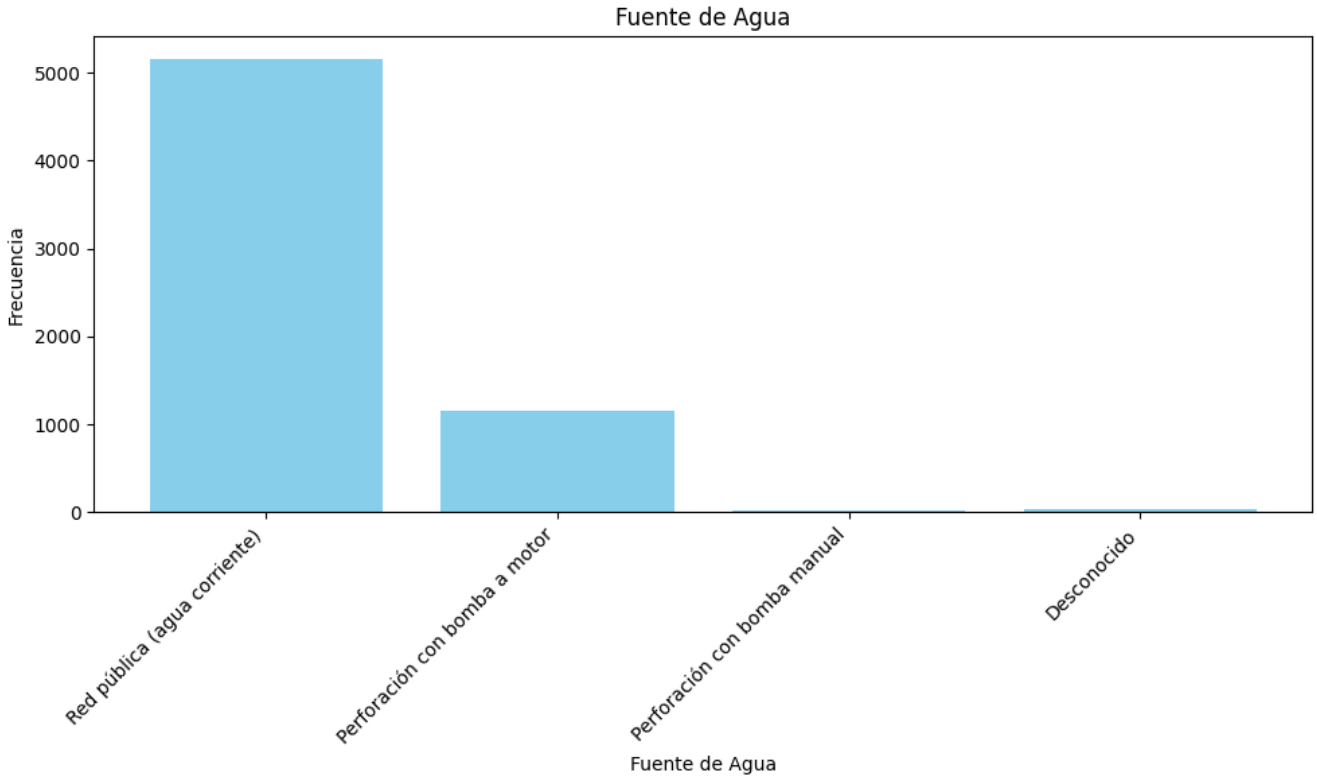

Figura 5: Acceso a red de agua

tiene esos problemas que mencionamos. Está entrenando el modelo n veces.

6

LASSO descartó muchas más variables de las que esperábamos. De todas maneras, es posible que hayamos especificado mal las funciones con las que armamos la regresion. También sabemos que muchas variables son preguntas follow-up o son demasiado específicas como para ser útiles. Entre las que descartó están las que creamos nosotros en la primera parte.

8

En la table vemos el resultado de las funciones en la Parte II sobre la base respondieron. Vemos que regresion logistica y discriminante lineal son muy similares y mejores que KNN. Esto podría sugerir que existen relaciones lineales en los datos que un método no paramétrico como KNN no logra capturar.

Modelo	**Configuración**	**AUC**	**Accuracy**	**MSE**
Logistic Regression	C: 0.01, penalty: l2	0.886	0.812	0.188
LDA		0.885	0.812	0.188
KNN (K=3)	n_neighbors: 3	0.838	0.801	0.199

Cuadro 1: Medidas de bondad para diferentes modelos

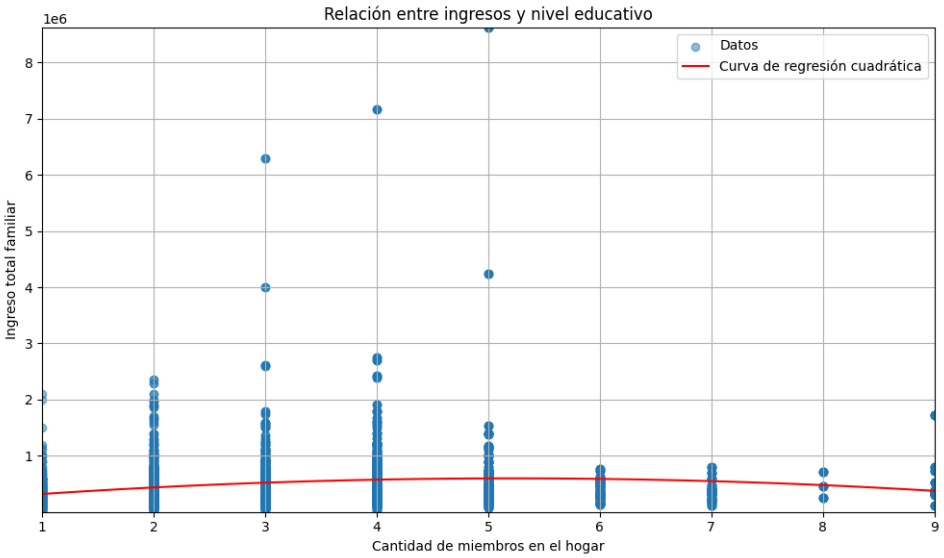

Figura 6: Regresión miembros familiares e ingreso