Universidad de
SanAndrés

Big Data

Profesora: Noelia Romero

Trabajo Práctico 4

Apezteguia, Mannarino, Navajas Jauregui

Fecha: 26/11/2023

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

Ejercicio 3

Después de importar las bases de datos individuales y de hogares de la Encuesta Permanente de Hogares (EPH) del Instituto Nacional de Estadísticas y Censos (INDEC), nos centramos exclusivamente en el aglomerado correspondiente al Gran Buenos Aires para llevar a cabo un estudio de predicción de la pobreza.

Previo al inicio de las predicciones, realizamos una exhaustiva limpieza de la base de datos para abordar posibles valores faltantes, outliers o datos incoherentes, siguiendo las pautas establecidas por el clasificador proporcionado por el INDEC [1]. Implementamos diversas funciones con el objetivo de agilizar y optimizar el proceso de limpieza, aunque reconocimos que algunas variables requerían ajustes manuales.

En una primera etapa, excluimos valores negativos para las variables de ingresos, ya que carecen de sentido en el contexto económico. Posteriormente, eliminamos observaciones cuyos ingresos superaban cincuenta veces la media. Al revisar la base de datos, identificamos variables con un alto número de valores faltantes, lo cual podría generar complicaciones en futuras estimaciones. Para abordar este problema, establecimos el criterio de eliminar aquellas variables con más de 300 valores faltantes. Además, creamos una función para descartar observaciones con valores iguales o superiores a nueve, ya que identificamos variables específicas que, según el clasificador, debían tener valores menores a esta cifra.

Con respecto a la elección de variables, comenzamos eligiendo en la categoría de identificación a CODUSU, código para distinguir viviendas, permite aparearlas con Hogares y Personas, permite hacer el seguimiento a lo largo de los trimestres.

La próxima categoría relevante hace mención a las características de la vivienda. Dentro de ellas elegimos incorporar las variables; IV2 que indica la cantidad de ambientes que tiene la vivienda, tenemos el prior que un hogar más grande puede ser reflejo de una situación de no pobreza; IV3 la cual indica el estado del piso, creemos que las viviendas con piso de tierra/ ladrillo suelto, poseen una probabilidad mayor a ser pobre; IV6 indica la manera en que recibe agua, por ende, la manera en que acceden a uno de los recursos más importantes, puede ser un buen indicio de la

[1] https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH$_registro_1T2023.pdf$

situación del hogar; IV8 No poseer baño puede ser un buen indicador de la pobreza del hogar, ambiente indispensable de la casa; IV10 en caso de poseer baño, es importante saber las condiciones en que se encuentra; IV11 la manera en que se genera el desagüe del baño puede dar indicios de las condiciones del hogar; IV12 1, IV12 2, IV12 3 tenemos el prior que vivir cerca de basurales, zonas inundables y/o villa emergencia está asociado a valores de hogares bajos por las posibles enfermedades de transmisión, inundación o acceso a necesidad básicas (transporte, hospitales, escuelas, etc.).

Además nos centramos en características habitacionales del hogar elementos que consideramos fundamentales para prever la situación de pobreza. La variable II1, que refleja la cantidad de ambientes, se revela como un indicador relevante; un hogar con escasos ambientes y un elevado número de residentes tiene una mayor probabilidad de encontrarse en situación de pobreza. Asimismo, la variable II7, relacionada con el régimen de tenencia de la propiedad, surge como una característica predictiva, ya que aquellos hogares cuyos miembros son propietarios pueden presentar menores probabilidades de experimentar pobreza. La elección de combustible para cocinar, II8, también se considera, especialmente en el contexto del Gran Buenos Aires, donde las preferencias pueden estar correlacionadas con la situación económica del hogar. Por último, la variable II9, que indica la presencia de baño compartido con otros hogares, se asocia con un aumento en la probabilidad de que el hogar sea considerado pobre.

Adicionalmente, exploramos las estrategias adoptadas por el hogar, enfocándonos en variables que consideramos poseen un alto poder predictivo para identificar hogares en situación de pobreza. Nuestra premisa se basa en la creencia de que si, durante los últimos tres meses, el hogar dependió de ingresos provenientes de seguros de desempleo, subsidios, donaciones, préstamos o ventas de pertenencias, existe una mayor probabilidad de que se encuentre en condición de pobreza. En consecuencia, hemos seleccionado las siguientes variables que reflejan estas estrategias: V1, V2, V3, V4, V5, V6, V12, V13, V14 y V17. Este enfoque nos permite capturar aspectos específicos que podrían ser indicativos de la vulnerabilidad económica del hogar.

En la sección de resumen del hogar, seleccionamos la variable IX TOT que indica la cantidad de miembros del hogar, ya que, en líneas generales, los hogares en situación de pobreza suelen tener un mayor número de miembros. En cuanto a la categoría de ingreso total familiar, optamos por la variable ITF, dado que el monto de ingreso familiar puede proporcionar una indicación rápida de la situación del hogar, especialmente al compararlo con la variable (a crear) de ingreso necesario. Para evaluar la situación de pobreza, en la categoría de ingreso per cápita, elegimos IPCF, que representa el monto de ingreso per cápita familiar, siendo otro índice rápido para

identificar hogares en condiciones de pobreza. Asimismo, consideramos PONDIH, el ponderador del ingreso total familiar y del ingreso per cápita familiar, como una medida adicional para contextualizar la situación económica del hogar.

Después, nos enfocamos en las características individuales de los miembros del hogar, introduciendo variables específicas. CH04, que refleja el sexo del individuo, puede influir en la distribución de roles y responsabilidades en el hogar. CH06, la edad, se considera crucial, ya que las necesidades específicas de niños y ancianos pueden tener un impacto significativo en la situación económica del hogar. CH07, el estado civil, puede afectar la estructura del hogar y las fuentes de ingresos, mientras que CH08, la cobertura médica, proporciona un indicador clave de la calidad de vida mediante la accesibilidad a la atención médica.

Asimismo, incorporamos CH09, referente al alfabetismo, dado que puede afectar la capacidad de los individuos para acceder a empleos bien remunerados. CH10 y NIVEL ED, que abordan la asistencia a establecimientos educativos y el nivel educativo alcanzado, respectivamente, son cruciales al ser predictores clave del ingreso y las oportunidades laborales. ESTADO, que describe la condición de actividad (ocupado/desocupado/inactivo/menor de 10 años), y CAT INAC, destinada a aquellos inactivos, proporcionan información vital sobre la participación laboral y las categorías específicas de inactividad. Finalmente, incorporamos P47T, que representa el monto del ingreso total individual.

Ejercicio 4

La primera variable que creamos la denominamos *pobreza estructural* tiene como objetivo representar a aquellos hogares que experimentan condiciones precarias de vida debido a factores estructurales. La clasificación como pobres se asigna a aquellos hogares que cumplen con más de dos condiciones específicas, y es posible que persistan en esa categoría a lo largo del tiempo. Para recibir un valor de 1 en esta variable, los hogares deben cumplir con al menos dos de las siguientes condiciones; de lo contrario, se les asigna un valor de cero. Estas condiciones abarcan aspectos como la falta de asistencia a un establecimiento educativo, la presencia de menos de dos ambientes en la vivienda, la utilización de materiales precarios en los pisos, la ausencia de agua por cañería dentro de la vivienda, la carencia de instalaciones sanitarias, la falta de inodoro con sistema de descarga y la disposición del desagüe del baño hacia un pozo ciego o letrina. Además, se considera la ubicación de la vivienda en áreas propensas a inundaciones, en basurales o en villas de emergencia, así como el uso de combustibles

distintos al gas de red para cocinar.

La variable que hemos creado, denominada *estrategia*, tiene como propósito identificar a individuos que se encuentran en situaciones de separación, divorcio, viudez o soltería y en los últimos 3 meses vivir de alguna de las siguientes condiciones: de lo que no ganan en el trabajo, percepción de jubilación o pensión, indemnización, seguro de desempleo, subsidio o asistencia externa, utilización de ahorros para cubrir gastos o la necesidad de vender bienes personales. Consideramos que vivir en los últimos 3 meses dependiendo de alguna de estas fuentes de ingresos y encontrarse sin compañía podría indicar una situación de vulnerabilidad.

Como última variable creamos a *hacinamiento*, que se compone del ratio entre cantidad de individuos y cantidad de ambientes. Con esto, buscamos identificar a los hogares con un ratio alto, que indica que viven en condiciones desfavorables y pueden ser indicios de hogares pobres. Al realizar una estadística descriptiva, observamos que la media es de 1.46, mientras que el mínimo es de 0.14 y el máximo de 10, que puede ser producto de vivir 10 personas en un ambiente.

La elección de las variables *pobreza estructural*, *estrategia* y *hacinamiento* se basa en la necesidad de capturar dimensiones específicas de vulnerabilidad y precariedad en hogares e individuos.

Ejercicio 5

Para observar la correlación entre variables, decidimos hacerlo a través de la matriz de correlación. Sin embargo, no realizamos una única matriz con las variables que consideramos relevantes en general, sino que hicimos una matriz con las variables que son más importantes para el hogar, y otra matriz de correlación con las variables relevantes a nivel individual.
De esta manera, seleccionamos como variables de interés los ambientes del hogar, tipo de pisos, si está en una villa emergencia, el combustible utilizado en la cocina, si ha vivido en los últimos meses de alguna jubilación o pensión, si tiene seguro de desempleo, si recibió ayuda social, si gastaron lo que tenían ahorrado, si tuvieron que pedir préstamos para vivir, si ha tenido que vender pertenencias, la cantidad de miembros en el hogar, el monto de ingreso per cápita familiar, pobreza estructural (variable propia) y hacinamiento (variable propia). Obtuvimos la siguiente matriz de correlación a nivel hogar: Algunos resultados interesantes que podemos observar en la figura es la baja correlación que existe entre la cantidad de ambientes del

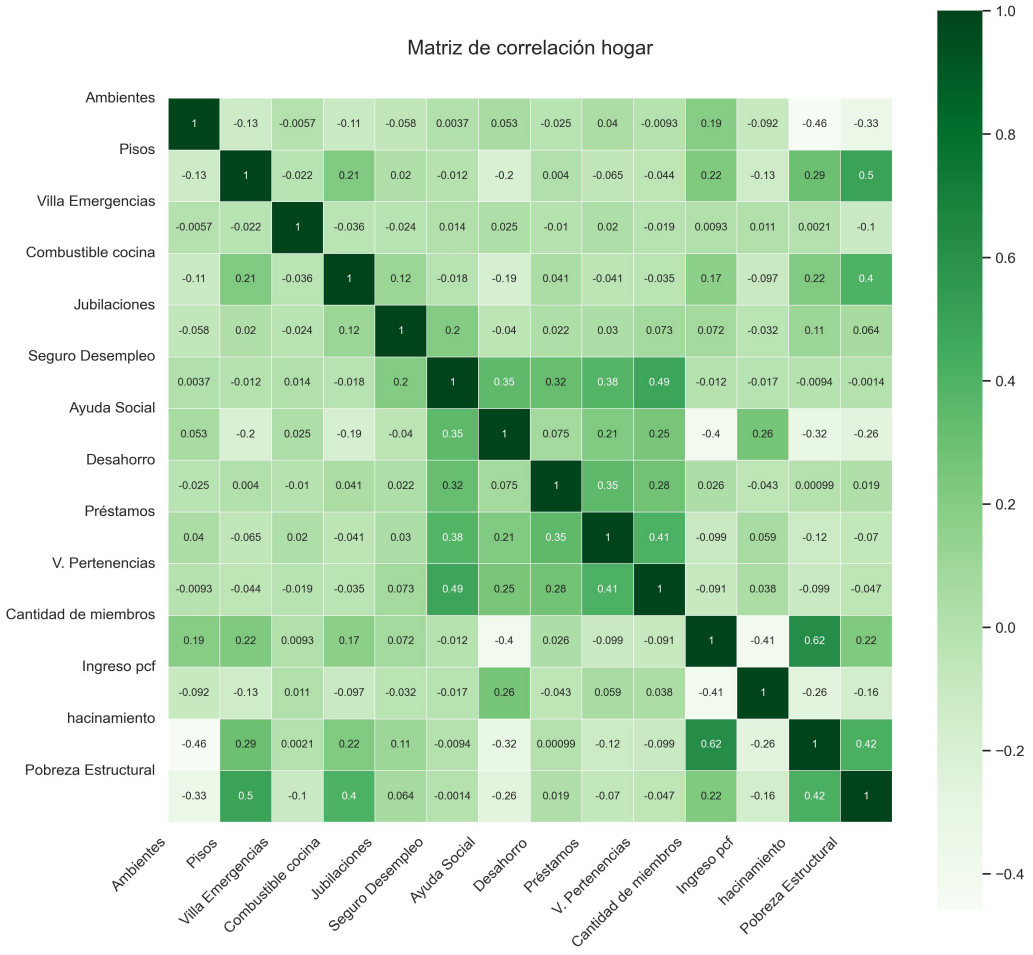

Figura 1: Matriz de correlación - Hogar

hogar y la ubicación del mismo en una villa emergencia, y entre la cantidad de miembros del hogar y la ubicación en la villa emergencia, siendo -0.0057 y 0.009 respectivamente. Intuitivamente, podríamos esperar una correlación positiva entre la locación del hogar en la villa y la cantidad de miembros que ahí viven, como consecuencia de la problemática de hacinamiento, y una correlación negativa mayor entre la locación y el número de ambientes.

Por otra parte, encontramos correlación positiva entre el material del piso y la variable de pobreza estructural es de 0.5 y de una magnitud similar con hacinamiento de 0.42. Las correlaciones más negativa se dan entre ambientes y hacinamiento que es de -0.46 y de ingreso per cápita familiar con cantidad de miembros. Por último, otras correlaciones interesantes son entre haberse financiado con el seguro de desempleo y la venta de pertenencias para poder subsistir, con una correlación de 0,49, y entre el combustible de la cocina y la pobreza estructural, que da 0,4. Ver la alta correlación de las variables creadas por nosotros para pobreza estructural y hacinamiento nos da un indicio de que hicimos un buen trabajo al momento de crearla.

Respecto a las variables vinculadas a aspectos más individuales, elegimos el ingreso per cápita familiar, el sexo, la edad, el estado civil, el tipo de cobertura médica, si

sabe leer o escribir, si asistió a algún establecimiento educativo, el nivel educativo alcanzado, la condición de actividad, la condición de inactividad y la variable creada por nosotros llamada 'estrategia'. Obtuvimos la siguiente matriz:

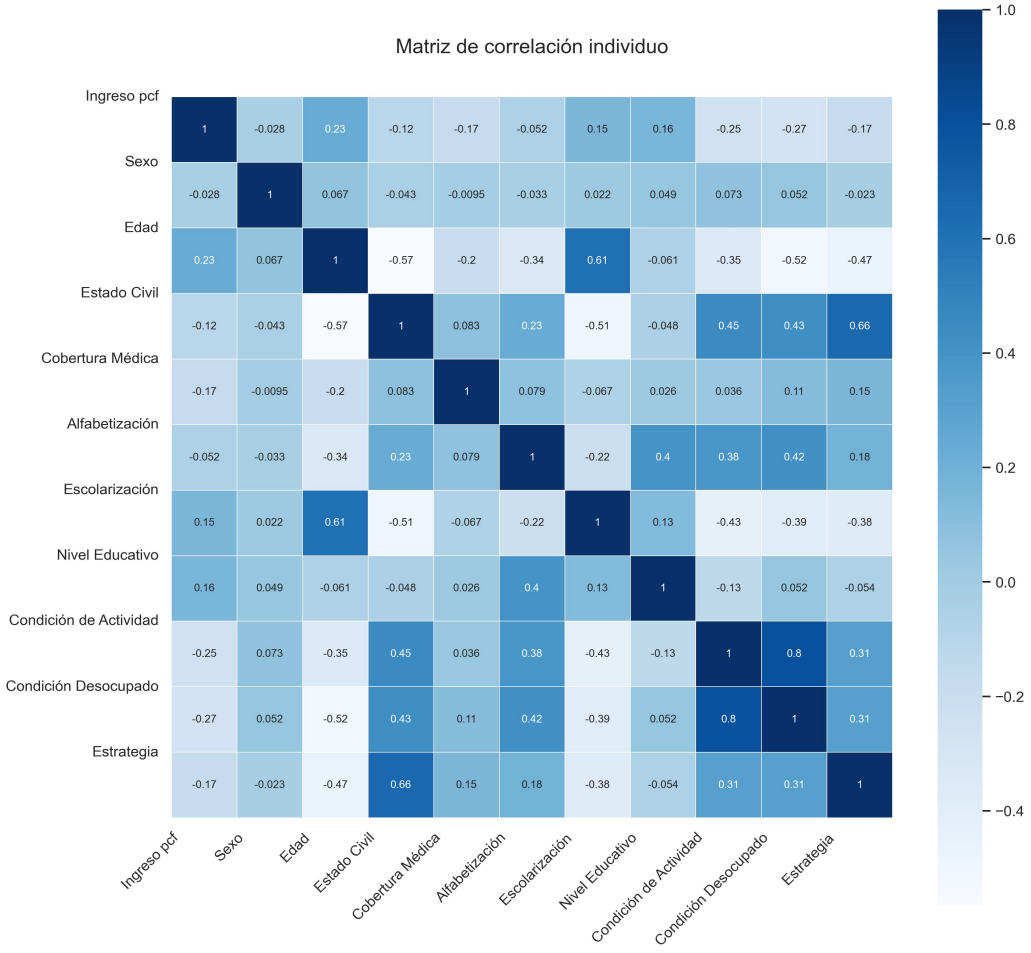

Figura 2: Matriz de correlación - Individuo

En este caso, notamos que hay una mayor correlación entre variables que en la matriz para el hogar. Respecto a las correlaciones positivas, la más alta se da entre la condición de actividad y la condición de desocupado, con una correlación de 0.80. Además, otros valores positivos interesantes son los que podemos ver entre escolarización y edad, y entre estado civil y estrategia, siendo 0.61 y 0.66 respectivamente.

Respecto a las correlaciones negativas, encontramos que las correlaciones más negativas son entre variables que podríamos esperar que tengan esta relación. Particularmente, la correlación negativa más grande se da entre estado civil y edad, con un valor -0.57. Luego, encontramos la correlación entre edad y condición de desocupado, que es -0.52, y en tercer lugar la correlación de -0.51 entre escolarización y estado civil.

La relación entre variables más débil se da entre el nivel educativo y la condición de desocupado, con un valor de 0.052. Las dos siguientes relaciones más débiles se dan entre sexo y cobertura médica, y sexo y estrategia, con valores de -0.0095 y -0.023, pero estos resultados no son tan llamativos como el mencionado en primer lugar.

Ejercicio 7

En la primera parte de nuestro análisis, calculamos la tasa de pobreza a nivel individual, obteniendo un resultado del 41.15 %, lo que representa 1556 individuos dentro de la base de datos *respondieron*. Esta categoría incluye a aquellos individuos que respondieron con ingresos superiores a cero. Lo estimado por el INDEC es de 41.4 %. Luego, nos propusimos estimar la cantidad de hogares pobres en el Gran Buenos Aires mediante la creación de la variable *ponde pobre*. Esta variable se compone de la multiplicación del estado del hogar por la cantidad que pondera, de manera que la suma de estos ponderadores nos proporcionaría la cantidad total de hogares en situación de pobreza. Para obtener la proporción, dividimos esta suma por la sumatoria de la variable PONDIH. La proporción resultante de hogares pobres es del 30.4 %, mientras que el INDEC reporta un valor del 30.3 %. Lo estimado para ambos casos es casi igual a lo publicado por el Instituto.

Parte III: Clasificación y regularización

En esta sección, nuestro objetivo es predecir la pobreza sin considerar los ingresos de las personas, dado que varias de ellas no proporcionan información sobre esta variable. Para determinar el mejor modelo, definimos 'mejor´ como aquel que exhibe una mayor precisión, un AUC más alto y un ECM más bajo. Esta evaluación se lleva a cabo mediante la función personalizada llamada *evalúa múltiples métodos*[2], que nos permite analizar de manera simultánea diversos modelos de clasificación. En nuestro caso, la clasificación se refiere a la categorización de individuos como pobres o no pobres. Con la ayuda de las funciones adicionales, podemos seleccionar los valores óptimos para las variables, optimizando así el rendimiento de nuestros modelos.

Es necesario resaltar los criterios empleados en la selección de parámetros, los cuales se determinaron de la siguiente manera:

- Para el algoritmo de vecinos más cercanos (KNN), consideramos un rango de 2 a 9 vecinos más cercanos.
- En el caso de los modelos de regresión logística, exploramos un rango de 11 valores equis espaciados para el hiperparámetro, que va desde 10^{-5} hasta 10^{5}.

[2]Las funciones de la Parte 2 están detalladas en el Script.

- Al aplicar la función *evalua config* en el modelo de árbol de decisión (CART), presentamos una lista de hiperparámetros para elegir la configuración óptima de profundidad, la cual varía entre 2 y 19.

Establecimos otros parámetros compartidos entre modelos, tales como *max sample* (0.25, 0.5 y 0.75), *max feature* (3, 6, 9, 12, 15), *n estimators* (25, 50, 75, 100), *max depth* (3, 4, 5 y 6), *learning rate* (0.01, 0.05, 0.1, 0.3, 1) y *base estimator max depth* (1, 2, 3, 5, 6).

- En el modelo de AdaBoost, utilizamos los parámetros *n estimators*, *learning rate* y *base estimator max depth*.
- Los modelos Random Forest y Bagging incorporan *max sample*, *max feature* y *n estimators*.
- El modelo Boosting, por su parte, emplea *max depth*, *learning rate* y *n estimators*.

Todos los modelos comparten el valor del split para la muestra, establecido en 10.

A continuación mostramos la tabla de salida de la función evalúa múltiples métodos, con los diferentes modelos con la mejor optimización:

Resultados de los modelos

Modelo	Hiperpar.	Prof.	Estim.	Features	n Sample	Learning	Prec.	AUC	ECM
LogReg	10.0	NA	NA	NA	NA	NA	0.765	0.773	0.207
KNN	7.0	NA	NA	NA	NA	NA	0.718	0.729	0.246
AD	NA	NA	NA	NA	NA	NA	0.771	0.769	0.209
CART	NA	12	NA	NA	NA	NA	0.786	0.827	0.166
Bagging	NA	NA	100	15	0.75	NA	0.843	0.846	0.141
RF	NA	NA	100	15	0.75	NA	0.833	0.845	0.144
AdaBoost	NA	6	100	NA	NA	0.1	0.871	0.878	0.113
Boosting	NA	6	100	NA	NA	0.3	0.862	0.871	0.120

Después de analizar la tabla, hemos decidido seleccionar el modelo de AdaBoost, ya que sobresale como el mejor entre los demás. Este modelo exhibe valores más altos de precisión y AUC, así como el menor ECM.

En la base de datos *no respondieron*, que se compone de aquellos individuos que respondieron ingresos iguales a 0, aplicamos la predicción de la condición de pobreza basándonos en las características del hogar y del individuo mediante el modelo de

AdaBoost optimizado. Nuestra estimación revela una proporción de pobres del 47.4 %, en comparación con el 38.6 % estimado en el TP3 y el 41.4 % según el INDEC. Además, la proporción de hogares pobres estimada es del 35.3 %, mayor al 24.3 % estimado en el TP3 y el 30.3 % informado por el INDEC.

Es importante destacar que en ambos casos hemos aumentado la incidencia de falsos positivos (error de tipo I), es decir, hemos sobreestimado la cantidad de personas en situación de pobreza. Creemos que, especialmente en una variable tan significativa como la pobreza, que guía diversas políticas sociales, es preferible incurrir en un error de tipo I que en un error de tipo II. Es más prudente proporcionar asistencia a personas que se encuentran cerca de la línea de pobreza, incluso si no son identificadas como pobres, en lugar de no asistir a personas que realmente lo son.