Trabajo Practico N°3

Lucila Brisighelli y Maria Constanza Cerundolo

Fecha de entrega: 26 de mayo de 2024

Parte I: Analizando la Base

1

1.1

La Encuesta Permanente de Hogares (EPH) es un programa ancional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las caracterısticas cosmográficas y socio-económicas de la población. Uno de los indicadores mas valiosos que pueden obtenerse con los datos extraídos de esta encuesta es la tasa de pobreza. Utilizando información disponible en la pagina del INDEC, identificamos a las personas pobres teniendo en cuenta el hogar que habitan. Particularmente, se utiliza un umbral que se conoce línea de pobreza, y refiere a la capacidad de los hogares de satisfacer las necesidades alimentarias y no alimentarias de sus miembros. Aquellos hogares cuyos ingresos totales no superan el valor de la canasta básica de alimentos (CBA) capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas son considerados indigentes. En cambio, son considerados pobres cuando los ingresos totales no superan el valor de la canasta básica total (CBT), que además de los alimentos comprende un conjunto de bienes y servicios necesarios para la vida cotidiana (salud, vestimenta, educación, transporte, etc.).

2

2.1 Ver en .ipynb

Eliminamos todas las observaciones que **no** corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires. Tenemos en cuenta los codigos de los agloreados 32 y 33 correspondiente a estos dos.

2.2

Descartamos observaciones con valores que no tienen sentido. Utilizando el registro de la EPH, eliminamos las columnas IDECCFR, PDECCFR y PDECIFR que aparecen como NAN. Por un lado, PDECIFR tiene que ver con el ingreso total familiar, particularmente

es Numero de decil de ingreso total del hogar del conjunto de aglomerados de **menos** de 500.000 habitantes. Tiene sentido que eliminemos esta columna ya que los aglomerados que tenemos en cuenta tienen mas de 500.000 habitantes. Lo mismo sucede para PDECCFR, el cual mide el ingreso pero capita familiar, particularmente numero de decil de ingreso per cápita familiar del conjunto de aglomerados de menos de 500.000 habitantes. Y por ultimo, IDECCFR que mide el ingreso per capita familiar, según numero de decil del ingreso per cápita familiar del interior.

2.3

El gráfico de barras muestra la composición por sexo según aglomerado, identificando el aglomerado 32 como la Ciudad Autónoma de Buenos Aires (CABA) y el aglomerado 33 como el Gran Buenos Aires (GBA). En el eje vertical se observa la cantidad de registros, mientras que en el eje horizontal se presentan los identificadores de los aglomerados (32 y 33). Los colores indican los sexos: azul para varones y naranja para mujeres.

En el aglomerado 32 (CABA), se registran aproximadamente 700 varones y 900 mujeres, lo que indica que en esta área el número de mujeres es mayor que el de varones. Por otro lado, en el aglomerado 33 (GBA), se registran aproximadamente 2500 varones y 3000 mujeres, mostrando nuevamente una mayor cantidad de mujeres que de varones, pero con números absolutos significativamente mayores que en CABA.

Podemos ver que en ambos aglomerados, el número de mujeres supera al de varones. Sin embargo, el Gran Buenos Aires presenta una cantidad de registros mucho mayor en comparación con la Ciudad Autónoma de Buenos Aires para ambos sexos. Esto podría reflejar la mayor población del área del GBA en comparación con CABA. Es notable que la proporción de mujeres a varones se mantiene similar en ambos aglomerados, aunque los números difieren.

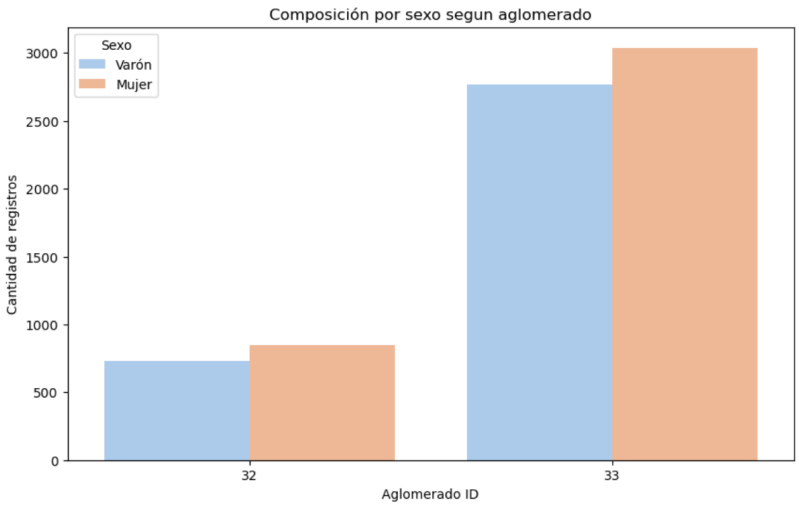

Figure 1: Composición por sexo

2.4

La matriz de correlación presentada incluye las variables CH04 (Sexo), CH07 (Estado civil), CH08 (Cobertura médica), NIVEL_ED (Nivel educativo), ESTADO (Condición de actividad), CAT_INAC (Categoría de inactividad) e IPCF (Ingreso per cápita familiar).

Al observar los resultados, se puede ver que la variable CH04 (Sexo) no muestra una correlación significativa con ninguna otra variable, indicando que el sexo no está fuertemente relacionado con las demás características analizadas. La variable CH07 (Estado civil) tiene una correlación con ESTADO de (0.41) y con CAT_INAC de (0.35). Esto sugiere que el estado civil puede influir en la condición de actividad y en la categoría de inactividad.

La variable CH08 (Cobertura médica) no muestra correlaciones fuertes con ninguna otra variable, indicando independencia en las respuestas sobre cobertura médica respecto a otras características. La variable NIVEL_ED (Nivel educativo) muestra una correlación negativa con ESTADO (-0.19), sugiriendo que niveles educativos diferentes pueden estar asociados a distintas condiciones de actividad.

La variable ESTADO (Condición de actividad) tiene una fuerte correlación con CAT_INAC (0.83), lo cual es esperable ya que la categoría de inactividad depende directamente de la condición de actividad de la persona. La variable CAT_INAC (Categoría de inactividad) además de la correlación con ESTADO, no muestra correlaciones fuertes adicionales, indicando que la categoría de inactividad se relaciona principalmente con la condición de actividad.

Por último, la variable IPCF (Ingreso per cápita familiar) muestra una correlación moderada positiva con NIVEL_ED (0.15), sugiriendo que a mayores niveles educativos, los ingresos per cápita familiar tienden a ser mayores.

Resumiendo lo anterior, podriamos decir que la matriz de correlación revela que existen algunas relaciones notables entre las variables, especialmente entre el estado civil, la condición de actividad y la categoría de inactividad. Sin embargo, muchas de las variables no presentan correlaciones fuertes entre sí.

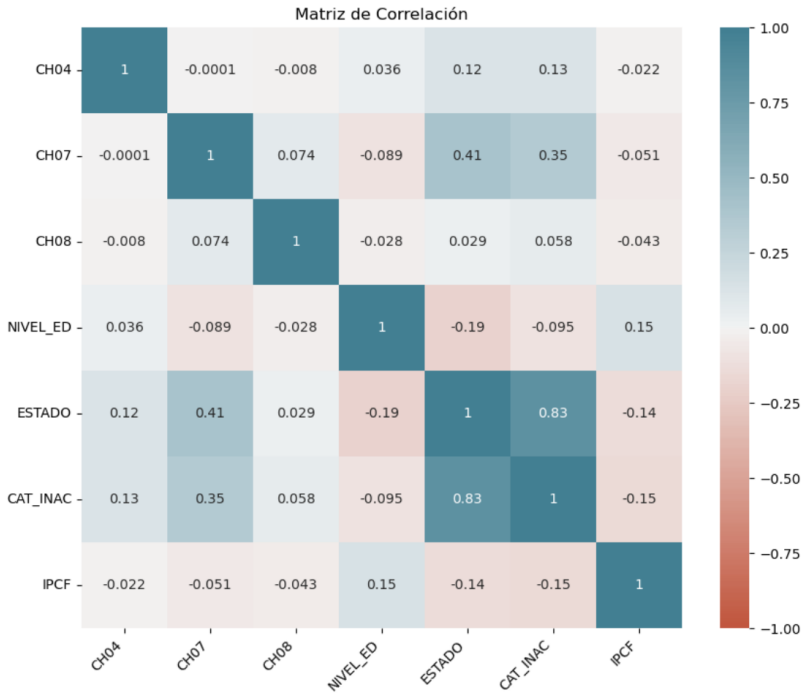

Figure 2: Matriz de correlaciones

2.5

Hay 240 desocupados en la muestra, 2765 inactivos. La media de ingreso per-capita familiar (IPCF) según estado son:

- Media de IPCF para los menores de 10 años: 75351.60983490566
- Media de IPCF para los que no responden: 0.0
- Media de IPCF para ocupados: 132041.48649856733
- Media de IPCF para desocupados: 58012.198416666666
- Media de IPCF para inactivos: 84993.67688245932

2.6 Ver en TP3.ipynb

3 Ver en TP3.ipynb

Uno de los grandes problemas de la EPH es la creciente cantidad de hogares que no reportan sus ingresos (ver por ejemplo el siguiente informe). ¿Cuantas personas no respondieron cual es su ingreso total familiar (ITF)? Guarden como una base distinta llamada respondieron las observaciones donde respondieron la pregunta sobre su ITF. Las observaciones con ITF = 0 guárdenlas en una base bajo el nombre norespondieron.

4 Ver en TP3.ipynb

Sabiendo que la Canasta Básica Total para un adulto equivalente en el cuarto trimestre de 2024 es aproximadamente $\$132.853,3$, agreguen a la base respondieron una columna llamada ingreso necesario que sea el producto de este valor por ad_equiv_hogar. Note que este es el valor mínimo que necesita ese hogar para no ser pobre.

5 Ver en TP3.ipynb

Por ultimo, agreguen a respondieron una columna llamada pobre que tome valor 1 si el ITF es menor al ingreso necesario que necesita esa familia, y 0 en caso contrario. ¿Cuantos pobres identificaron? Identificamos 157 pobres.

Parte II: Clasificación

Ejercicio 4 Viendo los resultados obtenidos en el ejercicio 3, llegamos a la conclusión de que el modelo logit es el que mejor predice ya que es el que tiene los valores más altos de auc y accuracy, además de ser el que mayor curva tiene. Son justamente estos tres los parámetros que hay que analizar a la hora de elegir el modelo que mejor produce, la curva de ROC, los valores de AUC y Accuracy. Esto es así porque cuando el accuracy es más alto, implica un mejor desempeño en la predicción.