Universidad de
SanAndrés

Big Data

Profesora: Noelia Romero
Asistente: Victoria Oubiña

Trabajo Práctico 2

Armas Braithwaite, Fernández, Menta, Vargas Ochuza

Fecha: 22/10/2023

Parte I: Analizando la base

Inciso 1

El INDEC utiliza la EPH para obtener la tasa de pobreza mediante el método de la *línea de pobreza* (LP). A partir de los ingresos de los hogares, establece si tienen la capacidad de satisfacer un conjunto de necesidades alimentarias y no alimentarias por medio de la compra de una canasta de bienes y servicios. Para calcular la línea de pobreza, primero es necesario determinar el valor de la canasta básica de alimentos de costo mínimo (CBA), una canasta que se determina en función de los hábitos de consumo de la población definida como referencia. Este procedimiento tiene en cuenta requerimientos normativos kilocalóricos y proteicos indispensables para esa población. Una vez establecidos los componentes de la CBA, se los valoriza con el listado de precios de bienes que componen el IPC. Además, el cálculo de la CBA tiene en cuenta variables como la edad, sexo y actividad de las personas. Luego, una vez obtenido el CBA, se lo amplía con la inclusión de bienes y servicios no alimentarios como vestimenta, transporte, educación, etc. A su vez, esta canasta compuesta se ajusta en base al coeficiente de Engels, que refleja la proporción del ingreso que se destina al gasto en alimentos. Una vez que se ajusta la canasta por este coeficiente, se obtiene la canasta básica total (CBT). Finalmente, para determinar si una persona es considerada pobre, es decir, se encuentra bajo la línea de la pobreza, se compara el valor de la CBT de cada hogar con el ingreso total familiar de dicho hogar. Si el ingreso está por debajo de la CBT, el hogar y sus integrantes se consideran bajo la línea de pobreza. Caso contrario, no se los considera pobres.

Inciso 2

a) La base de datos original tenía 48638 observaciones. Al quedarnos solamente con aquellas que corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires pasamos a tener 7619 observaciones.

b) En la presente sección, primero inspeccionamos los valores únicos para las variables CH04, CH07, CH08, NIVEL ED, ESTADO, CAT INAC, IPCF que son las que luego utilizaremos para armar la matriz de correlación. Por lo tanto, nuestro objetivo es eliminar los valores sin sentido de estas variables de interés. En esta línea, eliminamos las observaciones que tenían ingresos negativos, aquellas que tenían edades negativas, aquellos que no respondieron a la variable ESTADO (estatus de actividad económica), aquellos que no respondieron a la variable CH07 (estado civil), y aquellos que no respondieron a la variable CH08 (cobertura médica).[1] Al utilizar estos filtros reducimos

[1]No hay nadie en la muestra de interés que no indique su nivel educativo alcanzado (valor 9 en NIVEL ED) ni su categoría de inactividad (99 en CAT_INAC).

el dataframe a 7519 observaciones. **c)** Obtenemos la composición por sexo de las observaciones dentro de la muestra:

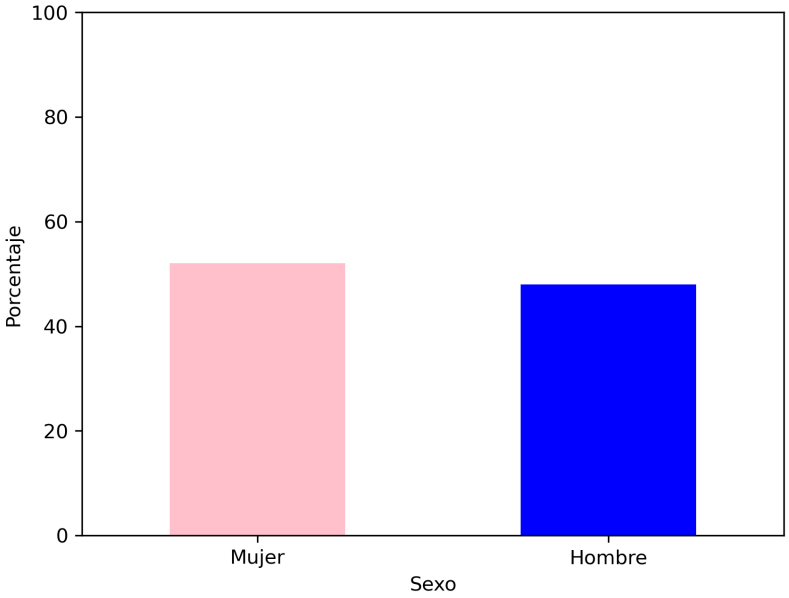


Figura 1: Composición por sexo (%)

Los resultados muestran que el 52 % de la muestra es mujer y el 48 %, hombre. Es importante destacar que la EPH pondera las observaciones para ser representativa a nivel nacional. Nosotros no realizamos dicha ponderación, por lo tanto, hay que ser cautos a la hora de extrapolar estos resultados a la población de interés.

d) La Figura 2 presenta la matriz de correlación de las variables correspondientes. El análisis correlativo no es útil para variables con múltiples categorías no ordenadas. Por ello, hemos generado variables binarias a partir de las variables categóricas de la base original. Asimismo, creamos la variable continua *años de educación* a partir del nivel educativo reportado, NIVEL_ED. Por otro lado, es importante notar que la variable *posibilidad de actividad* sólo considera a la submuestra inactiva e indica 1 si esta persona pertenece a una categoría que podría volver a la actividad y 0, de otro modo. Entonces, no se puede calcular su correlación con las variables binarias asociadas a estar desocupado o inactivo. A continuación, presentamos los principales resultados:

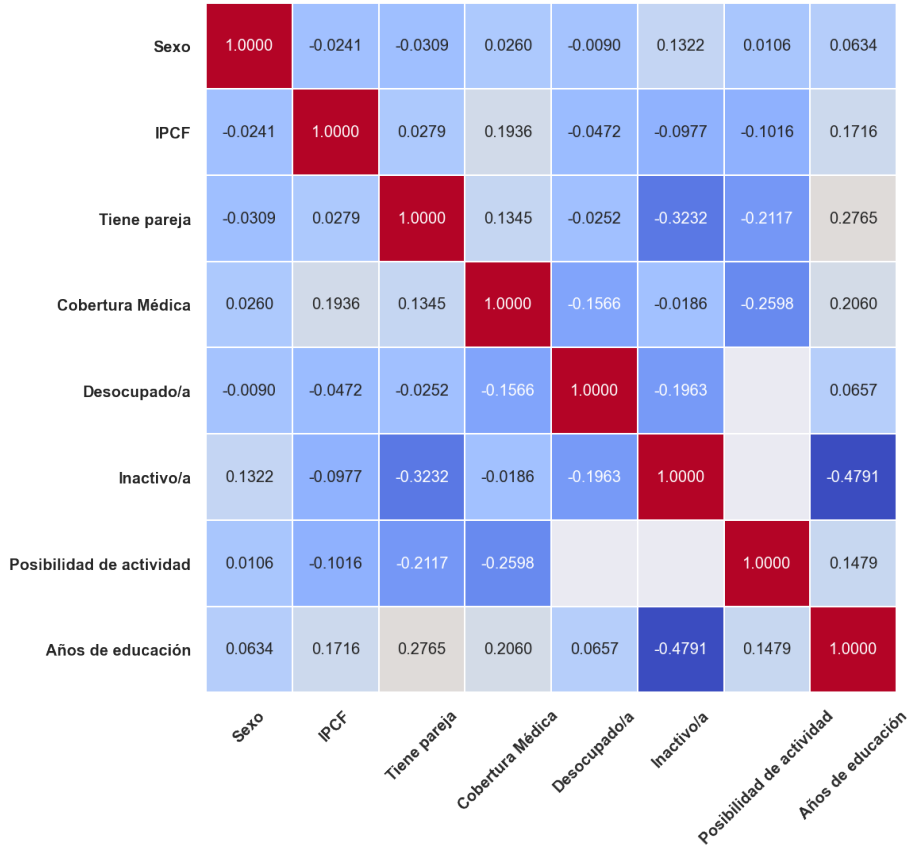


Figura 2: Matriz de Correlación

Las mujeres son ligeramente más propensas a no tener un trabajo remunerado ni estar buscándolo. De manera interesante, las variables más fuertemente relacionadas con el ingreso per cápita familiar son la propensión a tener cobertura médica y los años de educación. También, los encuestados inactivos en hogares con alto poder adquisitivo no suelen contar con la posibilidad de actividad. Las personas que tienen pareja tienen más años de educación en promedio y son más propensas a tener cobertura médica. Además, es menos probable que se encuentren en inactividad. Este resultado podría explicarse por la subpoblación de más de setenta años que esté tanto viuda como jubilada. Por otro lado, de manera esperable, el estar desocupado está asociado a no contar con cobertura médica, mientras que la inactividad tiene una asociación fuerte y negativa con los años de educación. Las personas con posibilidad de actividad suelen tener más años de educación y es menos probable que tengan cobertura médica. Finalmente, un mayor grado educativo está relacionado con la propensión a tener cobertura médica de manera positiva y moderada.

e) En la muestra tenemos 286 desocupados y 2826 inactivos. Por otro lado, la media de ingreso per cápita familiar (IPCF) según estado tiene la siguiente forma:

Así, vemos que los encuestados ocupados suelen pertenecer a hogares con mayor

ESTADO	Media IPCF
Ocupado	59812
Desocupado	25536
Inactivo	40089

Media IPCF según estado

ingreso per cápita que los inactivos o desocupados. A su vez, el grupo con menor ingreso per cápita familiar promedio son los desocupados.

f) Para este apartado, hemos optado por el uso de una función que toma como base la información proporcionada en el archivo de Excel. Creemos que esta manera es más directa y evita tener que trabajar directamente con el archivo de Excel, lo cual puede resultar tedioso. La función, llamada 'asignar_valor', recibe dos argumentos, 'edad' y 'género', y asigna un valor numérico basado en reglas condicionales que dependen de estos dos argumentos. Las condiciones se evalúan en orden y, cuando se encuentra una coincidencia, se devuelve un valor específico. Si ninguna condición se cumple, la función devuelve 'None'. Estas condiciones están diseñadas para categorizar a las personas en grupos según su edad y género, y asignarles un valor numérico de equivalencia en función de esas categorías. Una vez definida la función, la aplicamos al DataFrame objetivo, recorriendo cada fila y verificando las variables de interés para asignarles un valor, basándonos en la edad y el género de cada registro. Luego, sumamos estos valores para cada hogar y obtenemos el número de adultos equivalentes. [2]

Inciso 3

Una cantidad de 3346 personas no respondieron su ingreso total familiar. Luego, en el código procedimos a separar entre los que respondieron y los que no lo hicieron.

Inciso 4

En el código creamos la variable ingreso_necesario luego de tomar el producto entre la variable ad_equiv_hogar y el valor de la canásta básica total. La variable ad_equiv_hogar contiene el coeficiente de equivalencia para cada hogar que nos permite ponderar si un hogar es pobre o no en base a los requerimientos que necesita.

[2]En el código dejamos comentada otra metodología para realizar el mismo procedimiento pero de manera diferente. En ese caso, importamos el excel original y lo trabajamos de manera de obtener los mismos resultados sin crear la función, a partir de realizar los correspondientes merges. Ambas metodologías arrojan los mismos resultados.

Inciso 5

Hay 1555 pobres en la muestra, esto representa el 37.26 % de la muestra que sí reportó el ingreso total familiar.

Parte II: Analizando la base

Inciso 1

En esta sección realizamos algunas limpiezas a la base con el objetivo de poder correr los métodos de predicción que siguen en los siguientes incisos. En principio, eliminamos todas las variables relacionadas a ingresos como menciona el enunciado, junto con adulto_equiv, ad_equiv_hogar e ingreso_necesario. Por otro lado, los métodos de predicción requieren que las variables sean númericas y que no contengan missings o variables n/a, por lo tanto, nos encontramos frente a un *trade-off*.
Cuando una variable contiene missings, podemos decidir eliminarla o eliminar las observaciones que tienen missing. Si elegimos el primer camino, el mayor tamaño de muestra aumenta la potencia del método, a costa de eliminar una variable potencialmente relevante para la predicción. Frente a este escenario decidimos analizar las variables que contenían missings y strings en sus valores. Para comenzar, eliminamos las variables 'CODUSU', 'MAS_500' que son variables de identificación del hogar cuyos valores son string y no son relevantes para predecir pobreza.[3] A su vez, eliminamos la variable 'CH05' (fecha y hora de nacimiento) que no nos permitía correr el modelo por el tipo de sus valores y, entendemos, que la importancia de esta variable queda contenida en la variable 'Edad' que sí esta presente en el modelo. También, eliminamos las variables 'CH15_COD', 'CH16_COD', y 'CH14', que son variables de educación que contienen muchos missings y creemos que su información está contenida dentro de la variable NIVEL_ED.

Inciso 2

Realizamos la partición de la muestra siguiendo las indicaciones del enunciado para llevar a cabo el entrenamiento del modelo bajo un algoritmo supervisado. Nuestro interés será realizar predicciones sobre el outcome 'pobreza', es decir, clasificaremos a los individuos como 'pobres' o 'no pobres' en base a características observables. La separación en conjuntos de entrenamiento y prueba se efectúa para evaluar la capacidad predictiva del modelo en datos no vistos y garantizar que sea generalizable a nuevas muestras.

[3]También podríamos haber convertido MAS_500 a una variable binaria que nos brinde información sobre la densidad poblacional en la localidad del encuestado.

Inciso 3

Una vez realizada la partición, procedemos a evaluar diversos modelos de predicción para identificar y seleccionar el más eficaz en términos de capacidad predictiva. Para cada uno de estos modelos, calcularemos las siguientes medidas de precisión: accuracy y AUC.

Métricas: AUC y Accuracy

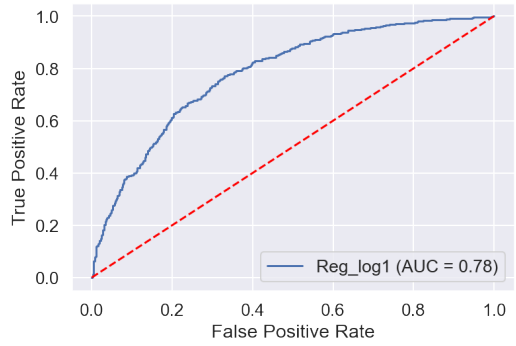
Métrica	**Valor Logit**	**Valor ADL**	**Valor KNN (k=3)**
AUC	0.7837	0.70	0.69
Accuracy	0.7220	0.7340	0.716

Matrices de Confusión

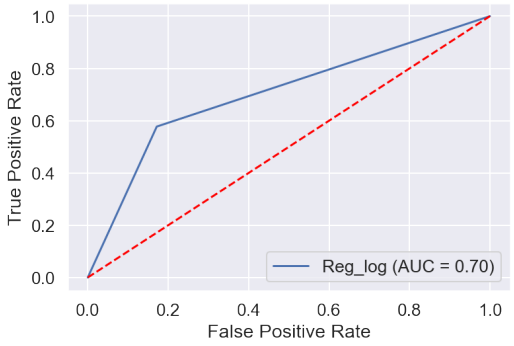
Logit	**Predicción Negativa**	**Predicción Positiva**
Real Negativo	651	130
Real Positivo	218	253

ADL	**Predicción Negativa**	**Predicción Positiva**
Real Negativo	647	134
Real Positivo	199	272

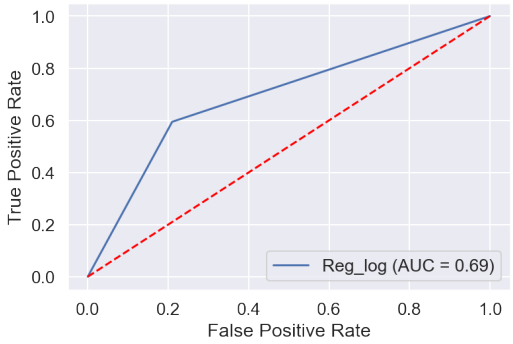
KNN (k=3)	**Predicción Negativa**	**Predicción Positiva**
Real Negativo	617	164
Real Positivo	191	280



(a) Logit



(b) ADL



(c) KNN (k=3)

Figura 3: Curvas ROC

Inciso 4

A continuación, evaluaremos la capacidad de los modelos para realizar predicciones con el objetivo de seleccionar el más eficaz. La accuracy mide la fracción de predicciones correctas en relación al número total de predicciones realizadas en base a la confusion matrix, un modelo con una precisión más alta es mejor, ya que hace menos predicciones incorrectas. Por otro lado, el AUC mide la capacidad del modelo para distinguir entre clases (pobre y no pobre) y se calcula a partir de la Curva ROC, es decir, un modelo con un AUC más alto es mejor en términos de discriminación entre clases.

La elección entre estas métricas de precisión varía según la naturaleza de los datos. Cuando nos encontramos con conjuntos de datos desbalanceados, optamos por utilizar el AUC como nuestra medida de precisión. En casos de clases equilibrada, recurrimos a la accuracy. Tras analizar los datos de nuestra base de entrenamiento, notamos un marcado desequilibrio entre individuos pobres y no pobres (solo un 38 % de los individuos que reportan su ingreso son pobres), por lo que priorizamos modelos que muestren un AUC relativamente más alto. No obstante ,observamos que las disparidades en la accuracy entre los mejores modelos no son significativas. Dado lo que hemos discutido, consideramos que el modelo Logit es la elección más adecuada para realizar predicciones sobre la pobreza en este caso.

Inciso 5

Tras completar el proceso de entrenamiento y selección del modelo óptimo, pasamos a aplicar este modelo entrenado a un nuevo conjunto de datos: aquellos individuos que no proporcionaron información sobre sus ingresos. Con el entrenamiento del modelo previo, podemos anticipar si estos individuos se encuentran en situación de pobreza o no, basándonos en otras características que no están relacionadas con sus ingresos. Nuestros resultados indican que en el conjunto de datos de "norespondieron", se predicen 1057 individuos como pobres. La proporción de individuos clasificados como pobres en este grupo es del 31 %. Esto sugiere que la decisión de reportar podría estar vinculada con el nivel de ingresos del encuestado.

Inciso 6

Si bien utilizar todas las variables disponibles puede capturar una amplia gama de información, esto conlleva riesgos. Puede aumentar la complejidad del modelo, dificultar la interpretación, generar multicolinealidad y llevar al overfiting. Por lo tanto, es esencial equilibrar la exhaustividad de las variables con la necesidad de un modelo más simple y eficaz mediante la selección cuidadosa de características relevantes. Podemos comprobar esto en nuestros resultados. Primero, eliminamos las variables de identificación del hogar y nos centramos en las características de los miembros del hogar. Luego, nos quedamos con las variables que intuitivamente aportan marginalmente más a la predicción de la pobreza. Analizando los indicadores de interés de este nuevo modelo logit, vemos que el AUC incrementa ligeramente y el accuracy se reduce en una magnitud un poco mayor. Dado que el AUC es la métrica más apropiada, preferimos este modelo sobre el anterior.

Resultados

Métrica	**Valor**
AUC	0.7847
Accuracy	0.7165

Matriz de Confusión

	Predicción Negativa	**Predicción Positiva**
Real Negativo	653	131
Real Positivo	224	244

Curva ROC

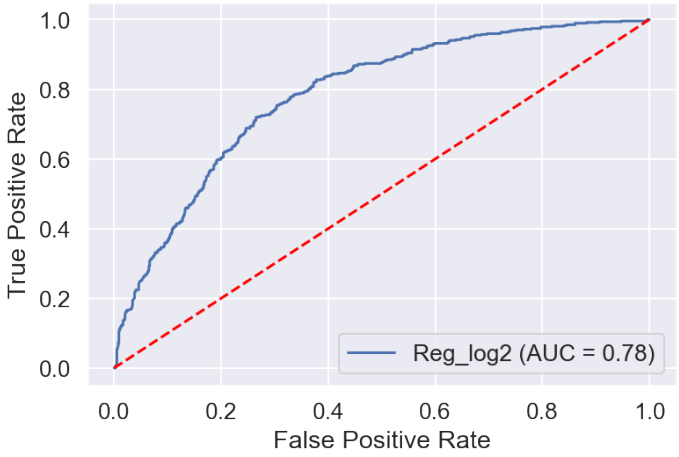


Figura 4: Curva ROC