Universidad de
SanAndrés

Big Data

Profesora: Noelia Romero

Trabajo Práctico 2

Apezteguia, Mannarino, Navajas Jauregui

Fecha: 22/10/2023

Parte I

Ejercicio 1 - Medición de la pobreza

Para la medición de la pobreza debemos definir previamente dos conceptos centrales para la construcción de esta medición: la Canasta Básica de Alimentos de costo mínimo (CBA) y el Coeficiente de Engel (CdE). El primer concepto, según lo define el INDEC (Instituto Nacional de Estadísticas y Censos), está compuesto por una canasta de alimentos capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas. Luego de establecidas estas necesidades, se calcula el valor que tienen a partir de los precios que surgen del Índice de Precios al Consumidor (IPC). El segundo está definido como la relación entre los gastos alimentarios y los gastos totales observados en la población. Es decir, se define a partir de la proporción que utiliza la población de su ingreso en la compra de alimentos.

A partir de la CBA y del CdE se construye la Canasta Básica Total (CBT). La CBT es una ampliación de la CBA para incluir todos los bienes y servicios que consume la población de interés, además de los alimentos. Esta canasta básica total se construye de la siguiente manera: CBT = CBA * inversa del coeficiente de Engel.

La idea detrás de esta metodología es capturar el gasto mínimo total de cada individuo en bienes y servicios. Debido a que la canasta básica alimentaria para cada persona depende de sus requerimientos nutricionales según su género y edad, existe una proporción de esa canasta básica según las características de la persona. La proporción igual a 1, o la proporción de referencia, son los hombres de 15 años junto a los que se encuentran en la franja de 30 a 60 años.

Una vez que se obtiene la canasta básica total para el individuo de referencia, la forma de calcular la pobreza es simple. Se calcula para cada hogar la cantidad de individuos de referencia según las características de sus integrantes y se multiplica por el costo de esa canasta básica total. Por otra parte, se calculan los ingresos totales de ese hogar. En caso de que el costo de las canastas básicas totales no sea superado por el ingreso agregado de ese hogar, todas las personas que lo integran son consideradas pobres.

Con un ejemplo podemos clarificar la metodología: supongamos que la canasta básica alimentaria es de $100.000 pesos y que el coeficiente de Engel es igual a 0,3. En este caso, la canasta básica total para un hombre de 15 años es de CBT=$100.000*1/0,3=$333.333. Luego, tenemos un hogar que está compuesto por 1 hombre de 50 años, 1 mujer de

45 años, 1 mujer de 16 años y un hombre de 18. Esta composición equivale a 3,56 canastas básicas totales, es decir, a 1.186.665. En caso de que los ingresos de este hogar sean menores a este valor, entonces todos sus integrantes serán considerados pobres.

Ejercicio 2 - Inciso C

Presentamos dos gráficos acerca de la composición por sexo de nuestra muestra, en uno mostramos la cantidad de mujeres y hombre y en el otro la proporción de cada uno.

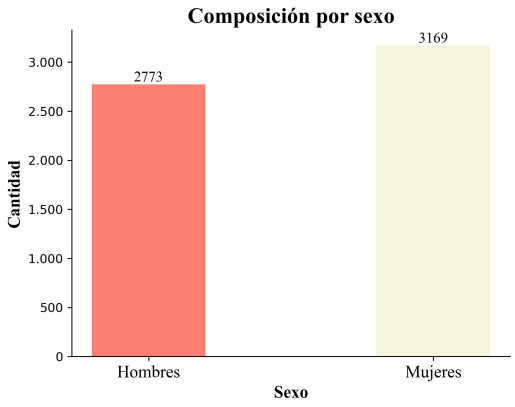


Figura 1

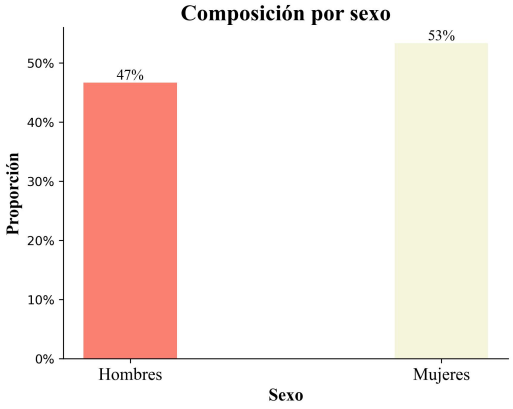


Figura 2

Ejercicio 2 - Inciso D

A continuación presentamos la matriz de correlación de las variables sexo, estado civil, cobertura médica, nivel de educación, ocupado/desocupado, categoría de inactividad e ingreso per cápita familiar.

Mediante la correlación podemos observar la relación que existe entre variable, y la dirección que posee (positiva o negativa). En esta tabla podemos encontrar correlaciones interesante como que la mujer tiene un nivel de educación más alto, pero a la vez el ingreso per cápita familiar disminuye.

Al analizar estos datos, notamos una muy baja correlación entre el sexo y las demás variables. Por otra parte, la variable que más correlaciona con las demás es la categoría

	Sexo	Estado Civil	Cobertura Méd.	Nivel Educ.	Ocupado/Desoc.	Cat_Inac	Ingreso PCF
Género	**1.000000**	-0.024350	-0.006743	0.041225	0.085532	0.061323	-0.044457
Estado Civil	-0.024350	**1.000000**	0.073523	-0.076763	0.463157	0.438371	-0.093692
Cobertura Médica	-0.006743	0.073523	**1.000000**	0.013699	0.048807	0.111006	-0.083552
Nivel educ.	0.041891	-0.076763	0.013699	**1.000000**	-0.186438	-0.002642	0.204868
Ocupado/Desoc.	0.085532	0.463157	0.048807	-0.186438	**1.000000**	0.813728	-0.238145
Cat_Inac	0.061323	0.438371	0.111006	-0.002642	0.813728	**1.000000**	-0.225708
Ingreso PCF	-0.044457	-0.093692	-0.083552	0.204868	-0.238145	-0.225708	**1.000000**

Figura 2: Matriz de correlaciones

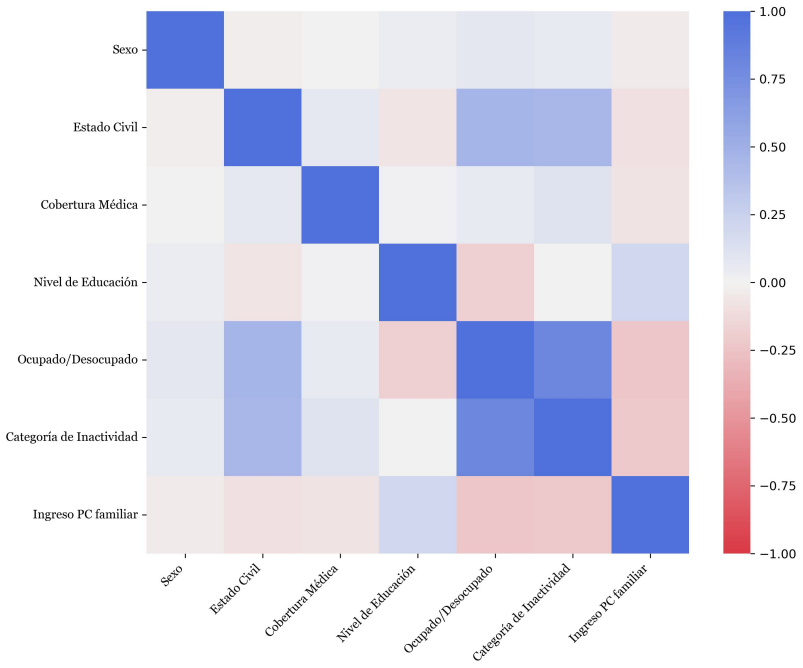


Figura 3: Matriz de correlación

de inactividad. Es interesante resaltar que el ingreso familiar correlaciona negativamente con todas las variables, salvo con el nivel educativo. Si bien esta relación es intuitiva, uno podría esperar correlaciones positivas con otras variables como, por ejemplo, la cobertura médica. El motivo por el que esto ocurre no es claro, ya que si observamos las demás correlaciones de la cobertura médica, son positivas, tal como uno esperaría. Respecto a las demás variables, notamos correlaciones positivas y de las más altas observadas (mayores a 0.4) entre el estado civil y la ocupación, y entre la ocupación y la categoría de inocupado.

Ejercicio 2 - Inciso E

En la muestra encontramos a 264 personas que se encuentran desocupadas y 2529 en condición de inactivos. Por otro lado, observamos la media del ingreso per cápita familiar para ocupados $ 93.268, desocupados $ 27.664 e inactivos $ 44.797.

Ejercicio 3

El problema principal es la creciente no respuesta de los hogares en la Encuesta Permanente de Hogares, especialmente en lo que respecta a la información sobre ingresos monetarios. Este deterioro se observó principalmente entre 2007 y 2015, lo

que resultó en un aumento de los hogares y personas cuyos ingresos tuvieron que ser imputados debido a la falta de datos. En nuestra muestra encontramos a 1769 personas que no respondieron a cuál es el ingreso total familiar.

Ejercicio 5

Teniendo en cuenta que la Canasta Básica Total para un adulto equivalente en el Gran Buenos Aires en el primer trimestre de 2023 es aproximadamente $ 57.371, buscamos la cantidad de pobres que hay en la muestra para aquellos que respondieron con un montón superior a 0 el ingreso total familiar. Encontramos en la muestra *respondieron* 1566 pobres, que es equivalente al 37,53 % de la muestra.

Parte II

Ejercicio 3

Implementamos los métodos Logit, Análisis discriminante lineal (ADL) y K-Nearest Neighbors (KNN). Cabe aclarar que para vecinos cercanos, tomamos un valor de k igual a tres.

A continuación presentamos las matrices de confusión.

Logit:
$$\begin{bmatrix} & CondNeg & CondPos \\ TestNeg & 634 & 145 \\ TestPos & 156 & 317 \end{bmatrix}$$

ADL:
$$\begin{bmatrix} & CondNeg & CondPos \\ TestNeg & 622 & 157 \\ TestPos & 153 & 320 \end{bmatrix}$$

KNN:
$$\begin{bmatrix} & CondNeg & CondPos \\ TestNeg & 600 & 179 \\ TestPos & 205 & 268 \end{bmatrix}$$

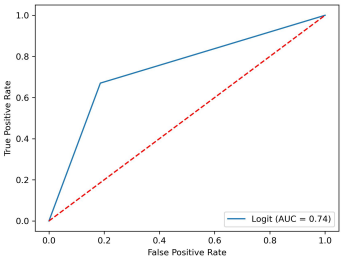
Cada vez que se interceptan los positivos y positivos y los negativos y negativos, significa que estamos acertando en nuestra predicción, asignando positivo y negativo según corresponda. En las otras intercepciones estaremos cometiendo o el error tipo I

o II.

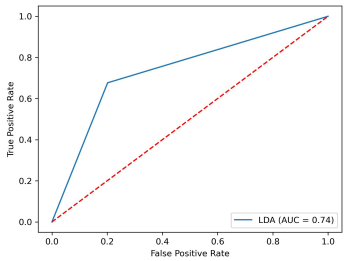
Presentamos para cada modelo los valores *Area Under the Curve* (AUC) y Accuracy. La medida del área debajo de la curva ROC, expone que un valor cercano a 1 indica un mejor rendimiento del modelo en la clasificación. Por su parte, la precisión mide la proporción de predicciones correctas en relación con el total de predicciones.

	Logit	ADL	KNN
Accuracy Score	0.76	0.75	0.69
AUC	0.74	0.74	0.67

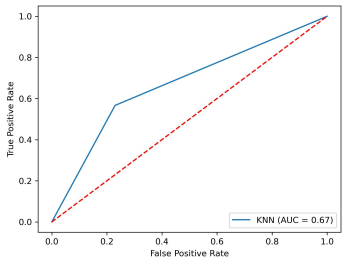
Luego, presentamos las curvas ROC para cada modelo:



ROC - Logit



ROC - ADL



ROC - KNN

Recordemos que en el eje Y de la curva ROC tenemos el ratio de verdaderos positivos y en el eje de las X tenemos el ratio de falsos positivos.

Ejercicio 4

Luego de observar el comportamiento de los modelos, observamos que el que presenta un mejor desempeño en cuanto a predicción es el logit. Respecto al Accuracy Score, es el modelo que mejor resultados muestra, de forma tal que es el que mejor mide la proporción de predicciones correctas realizadas en relación con el número total de predicciones. Por otra parte, es el modelo que cuenta con mayor AUC junto al de análisis de discriminante, por lo que son los dos modelos que mejor pueden distinguir entre clases. Además, podemos mencionar que, como el AUC es mayor a 0.5, el rendimiento del modelo es bueno, ya que la tasa de verdaderos positivos es mayor a la tasa de falsos positivos, siendo de esta manera mejor que una clasificación al azar.

Ejercicio 5

Al utilizar el modelo Logit, que fue el que obtuvo un mejor desempeño en la predicción, encontramos que la proporción de las personas pobres dentro de la muestra *no respondieron* es del 51 %.

Ejercicio 6

No consideramos que sea correcto realizar la predicción utilizando todas las variables disponibles, ya que, tal como sabemos de la clase teórica, la inclusión de variables representa un pago desde el lado del sesgo, el cual se reduce, pero una penalización si analizamos la varianza. De esta manera, al incluir todas las variables podemos estar obteniendo predicciones que sufran de tener mucha varianza. Además, sabemos que en el marco de Big Data toleramos tener cierto sesgo para minimizar la varianza. Teniendo esto en consideración, para realizar la predicción ahora elegimos tomar las variables, relación de parentesco (CH03), sexo (CH04), edad (CH06), estado civil (CH07), cobertura médica (CH08), sabe leer y escribir (CH09), asistencia presente o pasada a establecimiento educativo (CH10), nivel más alto de cursada (CH12), dónde vivía hace 5 años (CH16), Nivel educativo (NIVELED), condición de actividad (ESTADO) y categoría de inactividad (CATINAC). Dentro del conjunto de variables que teníamos disponibles luego de la limpieza, creíamos que eran las más relevantes para predecir si una persona es pobre o no. A continuación presentamos los resultados más relevantes obtenidos:

Logit:

	Cond Neg	Cond Pos
Test Neg	641	138
Test Pos	245	228

ADL:

	Cond Neg	Cond Pos
Test Neg	633	146
Test Pos	234	239

KNN:

	Cond Neg	Cond Pos
Test Neg	598	181
Test Pos	200	273

Con estas variables, el modelo que presenta un mejor desempeño es el de vecinos cercanos, que incluso tiene mejores valores que lo que posee con todas las variables, aunque los otros modelos pierden precisión y su valor del área debajo de la curva

	Logit	ADL	KNN
Accuracy Score	0.69	0.69	0.70
AUC	0.65	0.66	0.67

ROC. Con las variables elegidas y utilizando el modelo de vecinos cercanos obtenemos una estimación de la proporción de las personas pobres dentro de la muestra *no respondieron* del 46 %. Respecto a la comparación entre modelo Logit con menos variables y el modelo con todas las variables, notamos peores medida para el Area Under the ROC y el Accuracy Score en el modelo con menos variables, lo cual muestra que este tiene una peor capacidad predictiva. Estos datos van en contra de nuestra primera respuesta en este inciso, y son evidencia de que quizás la totalidad de variables existentes aún eran importantes para mejorar la capacidad predictiva del modelo.