Trabajo Práctico N°2: Big Data

Guerrero Ponzo Trinidad Felicitas - Sundblad Alfonso - Vergara Julian

Profesores: Romero Noelia - Oubiña Victoria

Universidad de San Andrés - Primavera 2023

1. Primera Parte

1.1. Ejercicio 1

Según la pagina del INDEC, la forma de analizar si una persona es pobre es mediante el análisis de su posición respecto a la linea de pobreza, la que es elaborada en base a los datos de la Encuesta Permanente de Hogares (EPH). De acuerdo a los ingresos de los hogares se estudia si tienen la capacidad de satisfacer ciertas necesidades alimentarias y no alimentarias que se consideran esenciales. Se determina una canasta básica de alimentos y luego se amplia con bienes y servicios no alimentarios. De esta forma se obtiene la canasta básica total. Si una persona tiene un ingreso menor al costo de la canasta básica total, se lo considera debajo de la linea de pobreza.

1.2. Ejercicio 2

Apartado b

En este inciso consideramos relevante eliminar observaciones con edad negativa, variables de ingresos con observaciones negativas y observaciones de variables que tienen ns/nr que son de nuestro interés.

1.2.1. Apartado c

En la Figura 1 podemos ver la representación de los que responden la EPH por género. Observamos que la proporción de hombres y mujeres es muy similar, aunque, es superior la proporción de mujeres. En particular, un 53,27 % son mujeres y un 46,743 % son varones.

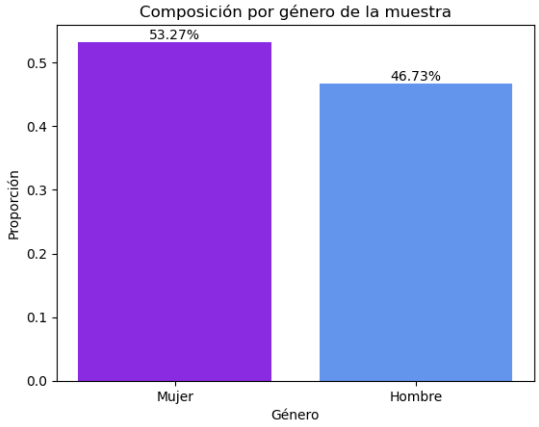

Figura 1: Proporción de la muestra por Género

1.2.2. Apartado d

En la figura 2 podemos ver una mapa de correlaciones entre las variables género, estado civil, cobertura médica, nivel educativo, situación ocupacional, categoría de inactividad e ingreso per capita familiar.

En primer lugar, observamos que genero tiene muy poca correlación con todas las otras variables mencionadas. Al igual que genero, cobertura médica es una variable que tiene correlaciones muy bajas con el resto de las variables. En cuanto a estado civil, tiene una correlación positiva del 0,46 y un 0,44 con situación ocupacional y categoría de inactividad respectivamente. En lo que respecta a nivel educativo, observamos que tiene una correlación negativa de 0,19 con situación ocupacional. Esto refleja que a mayor nivel educativa menor el número que representa la situación ocupacional (el 1 representa a los empleados), lo cual es plausible de pensar. Las personas que alcanzan mayores niveles educativos es más probable que se encuentren empleadas mientras las personas con menores niveles educativas es más probable que se encuentren desempleadas.

Por otro lado, la correlación entre situación ocupacional y categoría de inactividad es muy alta, lo cual se explica porque son variables muy relacionadas. Es decir, aquellas personas que están desempleadas o se encuentran inactivas también ingresan en algunas de las categorías de inactividad posible.

Finalmente, la variable de ingreso per capita familiar, tiene una correlación negativa de 0,23 con situación ocupacional y de 0,22 con categoría de inactividad. La correlación negativa con la primera tiene sentido porque mientras menor la categoría ocupacional (recordemos que la 1 es estar empleado) mayor el nivel de ingreso y mientras mayor la categoría ocupacional (desocupado e inactivo) menor el ingreso. Es esperable que las personas que tienen un empleo y perciben un salario tengan mayores ingresos que las personas que no trabajan. La relación negativa con la categoría de inactividad también es esperable, debido a que menor la categoría de inactividad se refiere a que es jubilado o rentista. Estas personas, si bien no trabajan, reciben un ingreso pasivo como la jubilación o la renta. Por último, la correlación entre el ingreso per capita familiar y educación es de 0,20, personas con mayor educación pueden acceder a mejores trabajos y por lo tanto, percibir mejores salarios.

Figura 2: Mapa de correlaciones

1.2.3. Apartado e

Observamos en la Tabla 1, que la mayoría de las personas de la muestra son inactivos (50,2 %) y empleados (44,6 %) y solo el 5,2 % de la muestra son desempleados. Cuando analizamos el ingreso per capita familiar promedio en cada una de estas categorías, observamos que la media es mayor en los empleados debido a que son los que tienen un trabajo en el que perciben un sueldo. Los inactivos son los segundos en promedio de ingreso per capita familiar y esto puede deberse a que hay dentro de estas categorías se encuentran jubilados o rentistas que perciben un ingreso pasivo al igual que los discapacitados que cobran pensiones. Finalmente, el ingreso medio de los desempleados es el menor y no es ni siquiera la mitad del promedio de los empleados y esto se debe a que no perciben un salario mes a mes, a lo sumo, reciben asistencia social por parte del Estado o algún tipo de ingreso inestable a lo largo del tiempo.

Situación Ocupacional	Total	Media IPCF
Empleados	2249	93268.954
Desempleados	262	27875.2
Inactivos	2529	44797.115

Cuadro 1: Resumen descriptivo Situación Ocupacional-IPCF

1.3. Ejercicio 3

Para calcular la cantidad de gente que no respondió evaluamos cuanta gente respondió que tenia un ingreso de cero. En total, son 1789 observaciones y en porcentaje representan un 30 % del total de la muestra con la que estamos trabajando. De acuerdo al documento técnico del INDEC, del segundo trimestre del 2007 al segundo trimestre de 2015 la cantidad de hogares que no reportaron sus ingresos pasó del $14,5\,\%$ al $25,3\,\%$. Nuestro resultado muestra un nuevo incremento de este ultimo dato al segundo trimestre de 2023 de alrededor de 5 puntos porcentuales. De acuerdo al informe, uno de los

aspectos que contribuyó en ese período al deterioro de la cantidad de respuestas fue la falta de seguimiento y aplicación de controles, y la imputación de datos faltantes que encubría el estado de la no respuesta.

1.4. Ejercicio 5

Consideramos como hogares pobres a aquellos que tenían un ingreso total familiar menor al requerido para estar sobre la linea de pobreza. Este ingreso mínimo requerido se definió como el producto entre $\$57{,}371{,}05$ y la variable adulto equivalente hogar que es la suma de la variable de adulto equivalente dentro de una misma familia. Obtuvimos un valor de 1523 pobres para Capital Federal y el Gran Buenos Aires.Esto representa a un $37\,\%$ del total de la muestra con la que estamos trabajando.

2. Segunda Parte

2.1. Apartado 3

Para predecir si una persona es pobre o no, a partir de los datos de EPH para el primer trimestre de 2023, se utilizaron tres métodos: Logit, Análisis discriminante lineal (de ahora en adelante LDA) y vecinos cercanos. Con respecto al primero, lo utilizamos para estimar la probabilidad de que una persona sea pobre y utilizamos el clasificador de Bayes para determinar si es pobre o no, en particular, si la probabilidad es mayor a 0.5 lo consideramos pobre y si es menor todo lo contrario.Un modelo logit consiste en una regresión en la que estimamos los parámetros por máxima verosimilitud. Con respecto al segundo, el análisis discriminante lineal se basa en la Regla de Bayes y tiene un supuesto fuerte atrás, a diferencia de Logit, que es que la distribución de las características X condicional en Y tienen distribución normal. Además, tambien se diferencia de Logit, en que es un modelo con muchos más pasos intermedios en donde calculo la probabilidad de una característica X dado en este caso ser pobre y luego dado no ser pobre. Ambos modelos suelen comportarse manera muy similar para predecir cuando las X no tiene correlaciones altas entre ellas y no hay tantas no linealidades entre las categorías de los outcomes. En este caso, podríamos esperar que el comportamiento difiera, ya que como vimos en el correlograma, hay predictores que si estan altamente correlacionados como mencionamos previamente. Finalmente, Vecinos Cercanos es una metodología no paramétrica en donde a partir las características x, vemos la condición de pobreno pobre de las observaciones más similares para determinar si es pobre o no. Este método es más difícil de aplicarlo mientras más predictres tenemos debido a la maldición de la dimensionalidad. Es decir, en un contexto como este, con una gran cantidad de predictores, este problema puede amenazar la capacidad predictiva de este método.

2.2. Apartado 4

Para testear que método es mejor, utilizamos como medidas de precisión la especificidad (ratio de verdaderos negativos), la precisión, *recall* (ratio de verdaderos positivos), ratio de falsos negativos, la *accuracy* Y la curva ROC para ver el área bajo la curva. La Tabla 2 resume los resultados de todas las métricas mencionadas. Cuando analizamos las distintas métricas en detalle, vemos que el modelo Logit es levemente mejor que los otros dos prediciendo. Esto se debe, en primer lugar, a que tiene un ratio de verdaderos positivos (*recall*) más alto, $0,57$ a 0,55 y 0,56 del modelo discriminante lineal y vecinos cercanos respectivamente. Con respecto a la especificidad (ratio de verdaderos negativos), el modelo Logit se desempeña igual de bien que Vecinos Cercanos (0,77 ambos) y levemente mejor que LDA (0,76). En cuanto a la exactitud,

el método Logit vuelve a ser el de mejor rendimiento, con una *accuracy* del 0,74 contra 0,73 y 0,71 de LDA y Vecinos Cercanos respectivamente. En relación al ratio de falsos negativos, el ratio de Vecinos Cercanos es superior al de Logit y al de LDA, nuevamente estos dos con el mismo ratio, en particular, $0,20$ a $0,16$. Cuando analizamos la precisión, nuevamente la del modelo Logit vuelve a ser levemente superior a LDA y Vecinos Cercanos (0,67 a 0,66). Finalmente, podemos ver las Curvas ROC de los tres métodos (Figuras de la 1 a la 3). Podemos observar en las figuras, que las tres tienen áreas bajo la curvas similares, aunque de nuevo, la del modelo Logit es levemente superior a la de los otros dos métodos, en particular, 0,71 a 0,69 y 0,68. Es decir, Logit está un poco más cerca, mirando el area bajo la curva, de la situación ideal de 100 % de verdaderos positivos y negativos.

Por último, los cuadros 3 a 5 muestran las matrices de confusión para cada uno de los métodos. Vemos, en términos absolutos, que logit es el modelo con más verdaderos positivos. En nuestro contexto, significa con predicciones de pobres que son verdaderamente pobres. En particular, 662 observaciones correctamente predecidas, mismo valor que tiene LDA con, seguido de vecinos cercanos con 633. En el mismo sentido el método con mas verdaderos negativos es logit con un valor de 265, seguido de vecinos cercanos con 259 y LDA con 254. Estos resultados están en línea con lo expuesto por las métricas de precisión.

Vemos que, en líneas generales, los tres métodos se comportan de manera similar, aunque Logit se desempeña levemente mejor en todas las métricas. Esto igual nos dice bastante información con respecto a lo mencionado previamente. El método de vecinos cercanos parece no verse muy afectado por el problema de la maldición de la dimensionalidad, ya que estamos utilizando una gran cantidad de predictores. Es decir, es mínimamente inferior al rendimiento de Logit. Por otro lado, el LDA tuvo un rendimiento muy similar al modelo Logit aunque levemente inferior, por lo cual es creíble el supuesto de que los datos siguen una distribución normal y la correlación entre los predictores, si bien en algunos casos es alta, no afectó a que se comportaran de manera similar. Además el desempeño similar de los modelos muestra que los predictores no tienen no linealidades o distribuciones más complejas que harían que los supuestos detrás de Logit y LDA los hagan peores métodos para predecir y sea más conveniente la utilización de una estrategia no paramétrica.

	Modelos		
	Logit	Discriminante Lineal	Vecinos cercanos
Recall	0.57	0.55	0.56
Specificity	0.77	0.76	0.75
False Positive Rate (FPR)	0.16	0.16	0.20
Accuracy	0.74	0.73	0.71
Precision	0.67	0.66	0.62
AUC	0.71	0.69	0.68

Cuadro 2: Medidas de Precisión

3. Curvas ROC

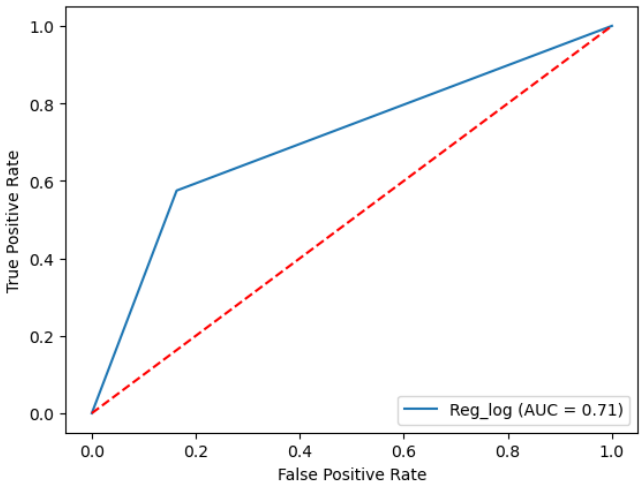

Figura 3: Curva ROC Modelo Logit

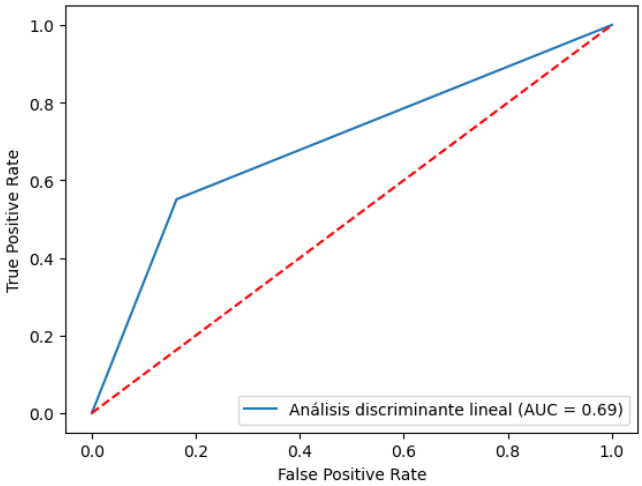

Figura 4: curva ROC Análisis discriminante lineal

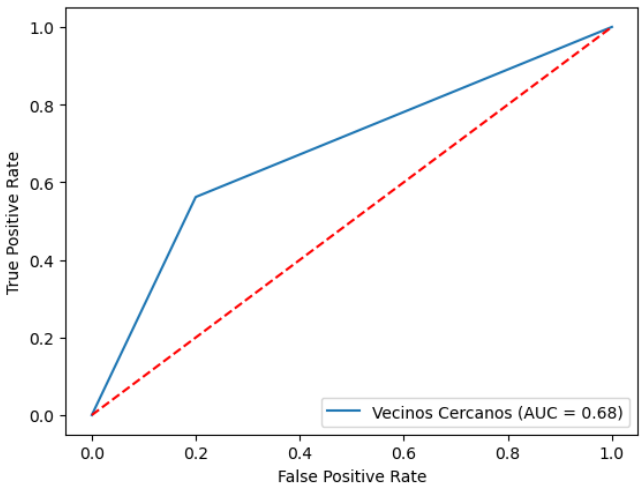

Figura 5: Curva ROC Vecinos Cercanos

3.1. Matrices de Confusión

	Pobre	No pobre
Predicción Pobre	662	129
Predicción No pobre	196	265

Cuadro 3: Matriz de confusión Logit

	Pobre	No pobre
Predicción Pobre	662	129
Predicción No pobre	207	254

Cuadro 4: Matriz de confusión Análisis Discriminante Lineal

	Pobre	No pobre
Predicción Pobre	633	158
Predicción No pobre	202	259

Cuadro 5: Matriz de confusión Vecinos Cercanos

3.2. Apartado 5

Para realizar la predicción fuera de la muestra, en la base *no respondieron* utilizamos el método Logit de acuerdo a las métricas de precisión mencionadas en el inciso anterior. Con este método, estimamos que el porcentaje de pobres es del 32 % para Capital Federal y Gran Buenos Aires. Nuestra estimación es,aproximadamente, 4 % menor a la proporción de pobres que tenemos de acuerdo a la EPH, por lo cual estariamos subestimando la pobreza con nuestra metodología.

3.3. Apartado 6

Para la predicción de la cantidad de pobres con los métodos seleccionados utilizamos como predictores todas las variables existentes en la base. Esto no tiene del todo sentido, ya que hay predictores que no aportan a la hora de formar el modelo predictivo. Por ejemplo, el trimestre, el número de hogar, que haya una dummy de hombre y otra de mujer simultáneamente. Mucha correlación entre predictores también generan problemas de multicolinealidad lo que puede aumentar la varianza del modelo y perder poder de predicción. Nosotros armamos un nuevo modelo para predecir que incluye a las variables que utilizamos en el mapa de correlaciones que nos parecieron las más relevantes y representativas para poder explicar la pobreza.

En el cuadro 7 podemos ver las medidas de precisión y vemos que el modelo con variables seleccionadas siempre tiene peor rendimiento. Vemos que el ratio de verdaderos positivos (*recall*) es de 0,48 lo cual es menor al del modelo con todos los predictores. La especificidad también es levemente menor (0.76 a 0.77)y el ratio de falsos positivos levemente superior (0.18 a 0.16). La exactitud disminuye en 0.04 y la precisión en 0.06. Si miramos la Curva ROC (Figura 6), también el área bajo la curva es inferior, disminuyó de 0.71 a 0.65.

Observando la nueva matriz de confusión del cuadro 7 podemos ver este deterioro en el rendimiento del modelo. En particular, tenemos 649 verdaderos positivos contra 662 del caso anterior. La cantidad de verdaderos negativos también disminuye de 265 a 223.

En conclusión, vemos que el modelo con variables seleccionadas es peor, lo que significa que dejamos de lado variables que pueden ser relevantes para predecir la pobreza como por ejemplo la región, características de la vivienda o si la persona

sabe leer o escribir. Entonces, deberíamos repensar la selección de variables si queremos mejorar la predicción.

	Logit
Recall	0.48
Specificity	0.73
False Positive Rate (FPR)	0.18
Accuracy	0.69
Precision	0.61
AUC	0.65

Cuadro 6: Medidas de Precisión Modelo Logit con variables seleccionadas

	Pobre	No pobre
Predicción Pobre	649	142
Predicción No pobre	238	223

Cuadro 7: Matriz de confusión Modelo Logit con variables seleccionadas

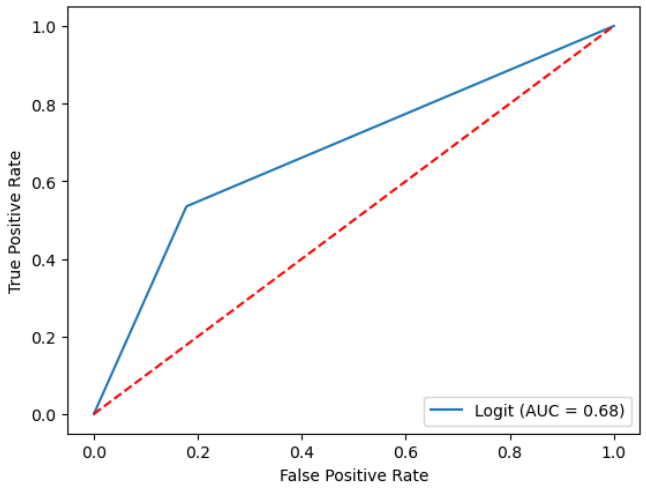

Figura 6: Curva ROC Modelo Logit con variables seleccionadas

4. Bibliografía Consultada

- Encuesta Permanente de Hogares. (Primer Trimestre 2023). **Base Individual y hogar. Total aglomerados , total interior, aglomerados de más y menos de 500.000 habitantes y cada aglomerado de EPH.** [Base de datos]. Recuperado de http: www.indec.gob.ar