Universidad de
SanAndrés

MAESTRÍA EN ECONOMÍA

BIG DATA

TP 2

Federico Di Carlo & Agustín Musso

2021

I: Analizando la base

I.1

El Instituto Nacional de Estadísticas y Censos (INDEC) distingue la pobreza de la indigencia. Siendo la primera una agregación de indigentes y pobres no indigentes. El INDEC denomina como “Línea de Pobreza” (LP) al umbral de ingreso del hogar necesario para cubrir la Çanasta Básica Total"(CBT). La noción de indigencia concibe la necesidad de suplir las necesidades proteícas y energéticas, por lo cual estas están incluidas en la CBT y consecuentemente en la LP. Esta metodología de cálculo de incidencia se corresponde con el método de medición indirecta, denominado línea (de allí el nombre).
La CBT se construye como una expansión de la “Canasta Básica Alimentaria” (CBA), calculada en base a registros estadísticos de hábitos de consumo alimentario de la poblacíon. La CBA se multiplica por la inversa del coeficiente de Engel[1] y se obtiene la CBT. Los precios se obtienen a partir del “Índice de Precios del Consumidor” (IPC) calculado por el INDEC.
Finalmente dado que las necesidades energéticas difieren entre individuos se contruye la variable de “adulto equivalente” para ponderar entre individuos. A partir de allí se contrasta el ingreso familiar del hogar con la LP permitiendo la construcción de cuatro categorías para el hogar que se extienden luego a los individuos: “hogares indigentes”, “pobres no indigentes”, “pobres” y “no pobres”. La categoría pobre incluye a las dos anteriores.
Cabe resaltar que el INDEC también está avanzando en la categorización de la pobreza multidimensional y en la incorporación de nociones para un cálculo más realista del “Gasto Total”. Por último se debe destacar el avance sobre consumos regionales, dado que los hábitos eran calculados previamente según los datos del GBA y extrapolados al resto de las regiones del país.

I.2.c

Tuvimos dificultades para graficar. Cada vez que queríamos correr el gráfico nos encontrabamos con problemas propios de como estaba configurada la base, mientras que al querer redefinir variables para graficar más sencillamente no nos corrían los códigos.

I.2.d

En la matriz de correlación, podemos observar que la mayor correlación (positiva y superior al 75 %) se observa entre el estado de actividad y la categoría de inactividad. Lo cual es lógico dado que todos los individuos que responden a la pregunta de inactividad serán los que respondan como inactivos a la pregunta sobre estado de actividad.
Por otro lado, sorprende la correlación positiva que existe entre estado civil y estado de actividad (en ambos casos superior al 50 %), lo cual se podría explicar dado que la variable soltero y menor de 10 años se correlaciona perfectamente, mientras que al correlacionar actividad y categoria de inactividad, la correlación que pueda existir entre alguna de las variables y estado civil impacta en la otra.
Ninguna de las otras variables correlaciona por encima del 50 % valor absoluto.

I.2.e

Existen en la muestra 208 desocupados y 1506 individuos inactivos.

[1] Coeficiente de Engel=$\frac{GastoAlimentario}{GastoTotal}$

	ESTADO	avg_IPCF
S/D	0	0.00
Ocupado	1	33802.29
Desocupado	2	12206.85
Inactivo	3	18695.85
Menor de 10 años	4	12610.50

I.3

El total de personas que no respondió a la pregunta de ingreso total familiar es 851.

I.4

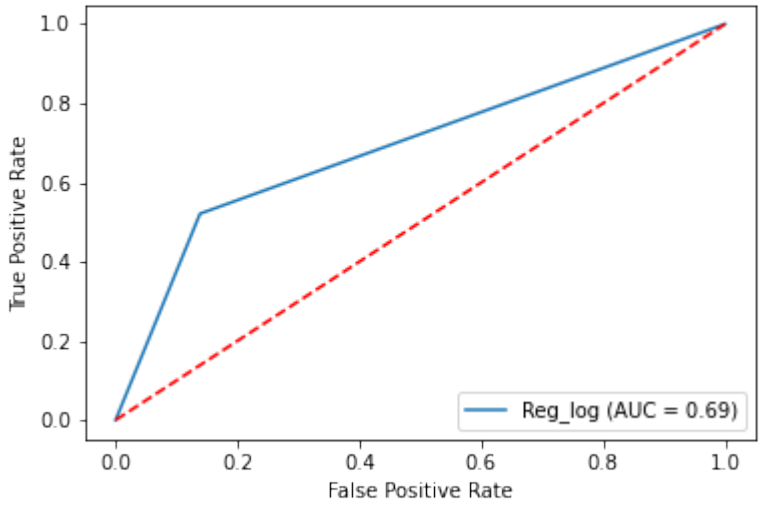
I.5

El total de pobres identificados en la muestra es de 840.

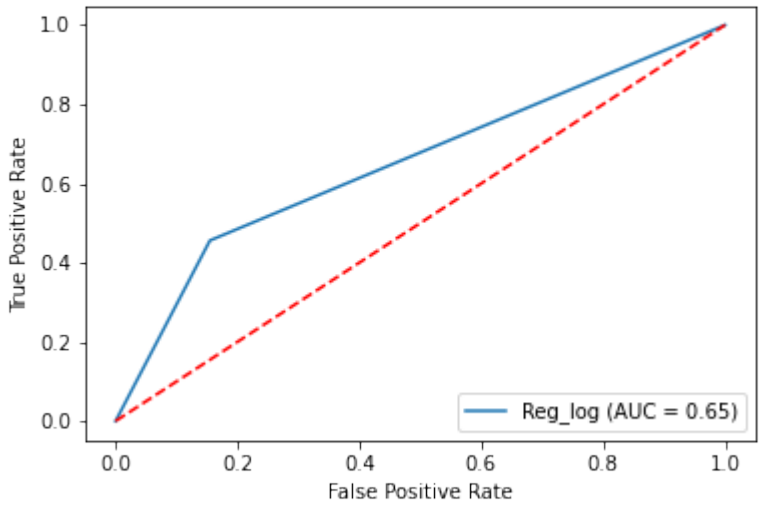
II: Clasificación

II.4

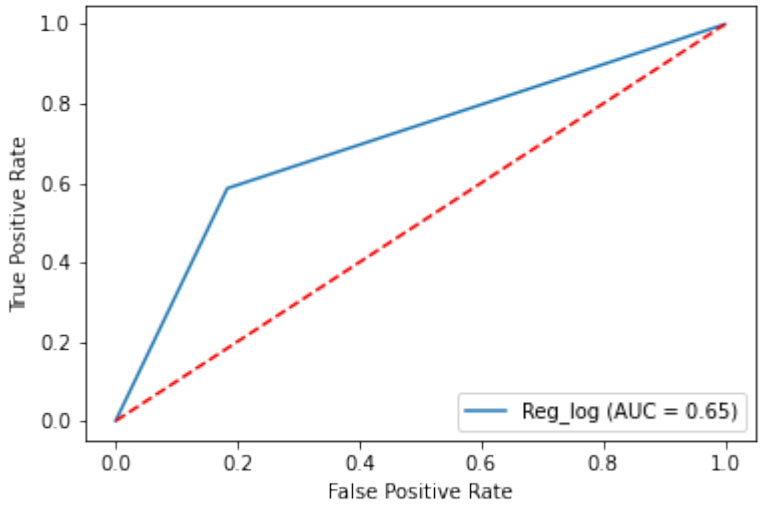
Curva ROC para el análisis discriminante



Curva ROC para la regresión logística



Curva ROC para el método de vecinos cercanos



II.5

El método de vecinos cercanos es el que mejor predice según nuestras estimaciones. La precisión calculada es de 0.737, mientras que el área bajo la curva ROC (AUC por sus siglas en inglés) del método de vecinos cercanos es de 0.70. Este método resulto mejor predictor en ambos indicadores que el análisis discriminante y la regresión logística ya que es más preciso. Por lo tanto, nos provee una mejor predicción de cuantos pobres hay basado en las variables que incluímos. Este método no es el que minimiza los errores tipo 1 como puede verse en la matriz de confusión. Sin embargo, a efectos de la medición de pobreza, nos resulta mas acertado priorizar la precisión de las predicciones. Si estuvieramos en un escenario de asignación de subsidios, por ejemplo, lo mejor sería elegir la opción que minimize el error tipo 1 aún cuando se pierda la precisión.

II.6

La cantidad de personas en la muestra que no respondieron que están debajo de la línea de pobreza es 243.

II.7

Elegir todas las variables no es lo más adecuado ya que la introducción de variables irrelevantes aumenta la varianza del modelo. Esto provoca modelos mas imprecisos. En el caso de la precisión de la regresión logística, aumento del 0.65 al 0.71 cuando omitimos las variables irrelevantes.