Big Data

Trabajo Práctico 2

Profesora: Maria Noelia Romero

Tutora: Victoria Oubiña

Alumnos: Manuel Carrera Figueroa - Matías Borhi

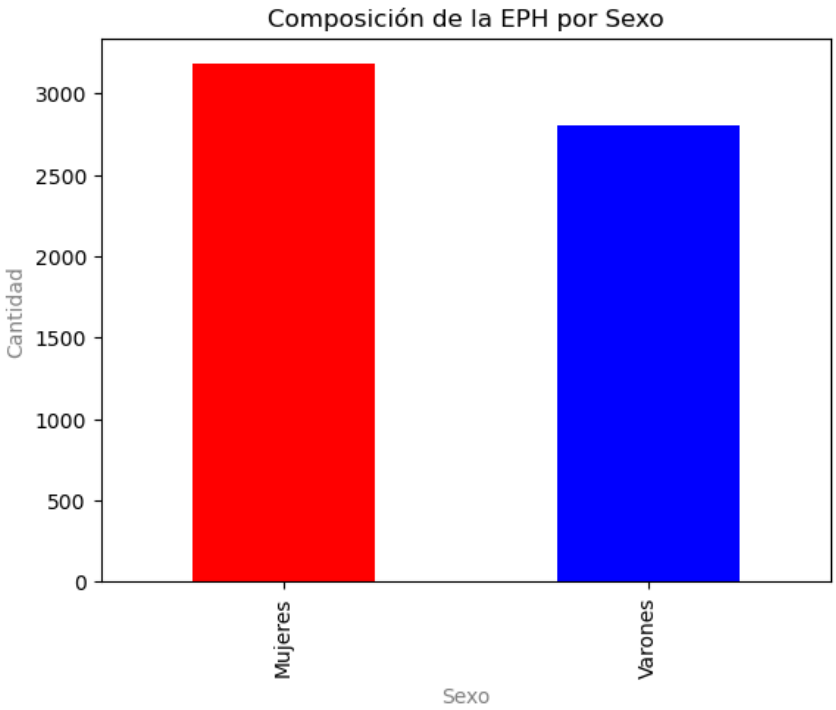
Octubre, 2023

Parte I: Analizando la base

1.
El INDEC mide la pobreza de manera absoluta y unidimensional en términos de ingresos. Esto implica que se considera pobre a aquella persona (u hogar) cuyo nivel de ingresos se encuentre por debajo del nivel de ingresos necesarios para acceder a una Canasta Básica Total (CBT). La CBT se construye a partir del cálculo de la Canasta Básica Alimentaria (CBA). Esta canasta referencial fue construida a partir de la información revelada por la Encuesta de Gastos e Ingresos de los Hogares (ENGHo) de 1997 y validada a través de la ENGHo de 2005. Contiene el conjunto de alimentos necesarios para satisfacer las necesidades alimentarias esenciales. Los precios de la canasta básica se actualizan cada trimestre a partir de los datos del Índice de Precios al Consumidor (IPC), mientras que las cantidades se mantienen constantes y dependen de la cantidad de individuos del hogar. Por ejemplo, el INDEC contempla que un niño no consume las mismas calorías que un adulto, así como una mujer no consume en promedio tanto como un hombre. Asimismo, cada periodo, el INDEC calcula el coeficiente de Engel: cociente entre gastos alimentarios y gastos no alimentarios. Las cantidades están fijas según los datos de la ENGHo 2005, mientras que los precios se actualizan cada trimestre a partir de los datos del IPC. En efecto, la Canasta Básica Total surge de multiplicar el monto de la CBA (ajustada por cantidad y miembros del hogar) por la inversa del coeficiente de Engel. Esta cifra actúa como la línea de pobreza de cada hogar: se observa en la EPH el nivel de ingresos de ese hogar; en caso de que este monto total sea menor al valor de la CBT para ese hogar, el INDEC lo contabiliza como pobre. Transitivamente, todas las personas de ese hogar se contabilizan como pobres.

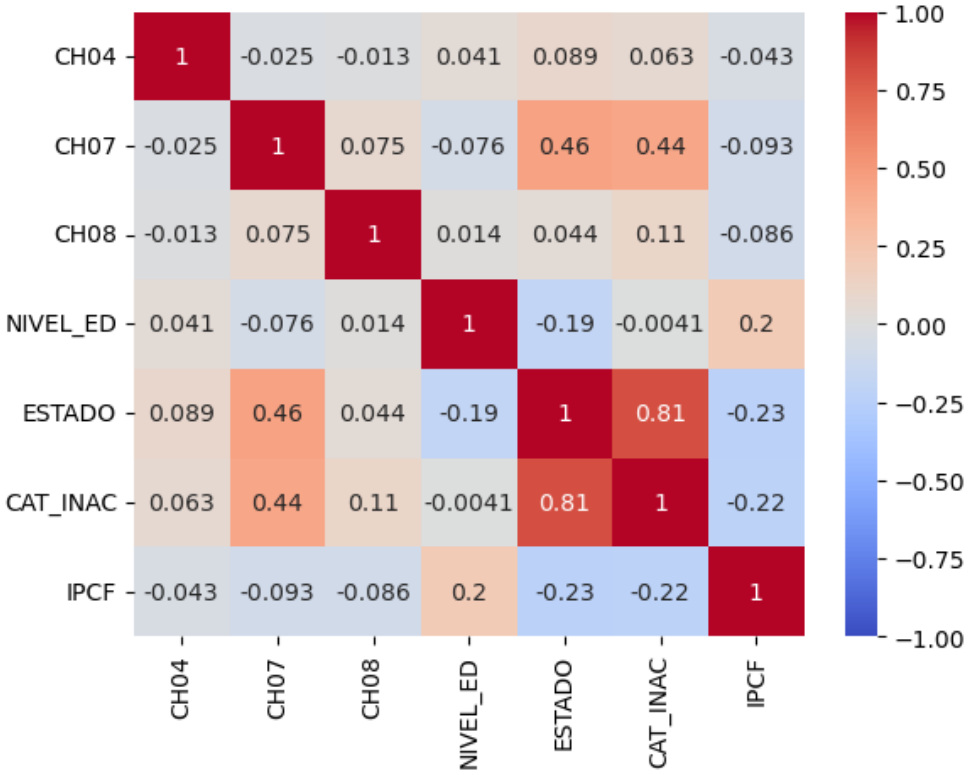
2.
a. Eliminamos todas las observaciones que no pertenecieran a CABA o GBA a través del comando *drop*.
b. En este inciso, eliminamos valores sin sentido. En primer lugar, todas observaciones cuya edad fuera negativa. Luego, todas las observaciones para las que el tipo de unión fuese 9 (sin respuesta). A su vez, eliminamos observaciones con ingresos menores que 0, en particular, en estos casos la variable de ingresos (*P47T*) valía -9 y denotaba la no respuesta. Por último, si bien no corrimos los otros comandos de ajustes, encontramos sin sentido respuestas en las variables que median la cantidad de horas trabajadas (i.e. trabajar más de 18 horas diarias) y en el número de decil (i.e. valor mayor a 10 indicando la no respuesta). Sin embargo, escogimos no eliminar estas observaciones ya que luego no utilizaríamos esas columnas, por lo que estaríamos perdiendo observaciones injustificadamente.
c. En primer lugar, en la variable *cant_mujeresyhombres* guardamos la cantidad de mujeres (*CH04=2*) y de hombres (*CH04=1*). Luego creamos un gráfico a través de la librería de *Matplotlib*. En la Figura 1, podemos observar la distribución de hombres y mujeres dentro de la base de la EPH limpia. Como se puede notar, existe una mayor proporción de mujeres que de hombres. Mientras las mujeres son 3.181 (53,18% de la muestra), los hombres son en total 2.801 (el 46,82% restante).

FIGURA 1 (Parte I – Ejercicio 1 – Inciso c)



d. Para este inciso, calculamos la correlación entre las variables solicitadas y las graficamos en la Figura 2. A primera vista, podemos notar que la correlación tiende a ser cercana a cero en gran parte de las combinaciones. Como excepciones, destacamos primeramente la relación positiva entre *CH07* (tipo de unión) y *ESTADO* (ocupado, desocupado o inactivo). Esto podría denotar la asociación positiva entre valores más altos de *CH07* (i.e. estar separado, viudo o soltero) y los valores más altos de *ESTADO* (i.e. inactivo o menor de 10). Tiene sentido si pensamos que las personas viudas, dada su edad, tienden a ser menos activas y que muchos individuos de la muestra solteros son menores de edad. También hay una relación positiva alta entre *CH07* y *CAT_INAC* (categoría de inactividad). Esta última aumenta su valor cuando la persona es estudiante (3), ama de casa (4), menor de 6 años (5) o discapacitado (6). Tiene sentido pensar que este tipo de personas son más propensas a estar solteras (*CH07=5*) o separadas (*CH07=3*). Dados estos análisis, cobra sentido la alta correlación positiva entre el estado de ocupación y la categoría de inactividad (0,81). Al observar más en profundidad, este valor está diciendo que un nivel de *ESTADO* mayor (i.e. ser desocupado, inactivo o menor de 10) se asocia con ser estudiante, ama de casa (empleo que podemos pensar como volátil), menor de 6 años o discapacitado. Asimismo, la correlación del monto de ingreso per cápita familiar (*IPCF*) con *ESTADO*, *CAT_INAC* y *NIVEL_ED* soportan la dirección de estos resultados. Como se puede ver, un mayor nivel de ingresos se asocia en primer lugar, con un mayor nivel educativo. Luego, el mayor nivel de ingresos también se asocia a valores menores de la variable *ESTADO* y *CAT_INAC*. Estos corresponden respectivamente a estar ocupado y a ser pensionado/jubilado o rentista. Si bien los valores son pequeños, la correlación entre *CH04* y *NIVEL_ED* e *IPCF* nos indica que las mujeres tienden a ser más educadas que los hombres pero que a su vez tienden a ganar menor que ellos.

FIGURA 2 (Parte I – Ejercicio 1 – Inciso d)



e. En nuestra muestra hay 264 desocupados y 2.539 inactivos. Los desocupados comprenden el 4,41% y los inactivos el 42,44% de la muestra. Asimismo, notamos que el ingreso per cápita familiar promedio es de $93.122,62 para personas ocupadas, $27.664,02 para personas desocupadas y $44.748,88 para personas inactivas. Cabe resaltar que al estar mirando el ingreso per cápita familiar, la media de ingresos no necesariamente corresponde al nivel de ingresos de esa persona. Por ejemplo, podría ser que en un mismo hogar tiendan a vivir personas ocupadas e inactivas, lo cual elevaría el ingreso per cápita de las segundas sin que estas estuviesen trabajando.

f. Para este inciso, en primer lugar, hemos concatenado en una misma columna el sexo de cada observación junto a su edad. Luego, creamos dos series de números: "sexo" que toma los dos valores posible de esta variable en la EPH; y "edad" que toma los valores del 0 al 104 (i.e. posibles edades de los individuos de la EPH). Luego concatenamos en una columna llamada "*sexo_edad*" todas las combinaciones posibles entre sexo y edad. A continuación, creamos un *data frame* llamada *Valores* que contenga la columna de *sexo_edad* y la columna de *adulto_equiv*, donde asignamos para cada combinación de sexo y edad el coeficiente que establece la *tabla adulto equiv.xlsx* ordenados de modo que aparezcan todas las alternativas para los hombres de *cada* edad y luego análogamente para mujeres. Esto nos permitió hacer un simple *merge* entre nuestra base de la EPH y la base de *Valores*. Por último, creamos una nueva base llamada *ad_equiv_hogar* que agrupara según el valor de *CODUSU* (i.e. agrupando observaciones pertenecientes al mismo hogar) y sumara en una columna llamada *ad_equiv_hogar* los niveles de *adulto_equiv* de todos los individuos de ese mismo hogar. Finalmente, volvemos a hacer un *merge* entre la EPH y esta última base para que podamos ver para cada observación el nivel de adulto equivalente agregado del hogar al que el individuo pertenece.

3.
Al dividir la muestra a partir de la respuesta de los individuos para la variable *ITF* de la EPH, notamos que hay 4.180 individuos que sí reportaron cuál es su ingreso familiar total. La línea siguiente al conteo de las observaciones no es más que un control para ver que hicimos correctamente la partición de la muestra: como el mínimo de la muestra *respondieron* es 2.500, todas las observaciones son mayores que ese valor y en particular, mayores a cero. Por el otro lado, al mirar la base *norespondieron* notamos que hay 1.802 personas que no reportaron su nivel de ingreso familiar. Este valor corresponde al 30,12% de la base de la EPH de CABA-GBA limpia de valores sin sentido. También realizamos una suerte de control: como todas las observaciones suman 0 y ya habíamos eliminado valores negativos, efectivamente en *norespondieron* solo hay observaciones con *ITF* = 0.

4.
Partimos de que el valor aproximado de la CBT para un adulto equivalente de GBA fue en el primer trimestre de 2023 igual a $57.371,05. Como explicamos previamente, este valor se obtiene al ajustar la Canasta Básica Alimentaria por el coeficiente de Engel. Luego, en la columna *ingreso_necesario* guardamos la cantidad de dinero que cada hogar necesita para poder costear la CBT acorde a la composición de sexo y edad del hogar y en efecto, no ser pobre.

5.
Para este inciso, primeramente construimos una función que clasifica como pobre (variable *pobre*=1) para los casos en que el *ITF* de una persona sea menor al *ingreso_necesario* que *ese* hogar necesita para poder pagar la CBT que corresponde a ese tipo de familia. Luego, aplicamos esta función a una columna nueva en nuestra base de *respondieron* llamada *pobre*. Corrimos una versión más acotada de la base para observar rápidamente si el criterio aplicado tenía sentido. Luego, sumamos la cantidad de pobres de la muestra y hallamos que del total de personas para los que tenemos información sobre el *ITF*, 1.614 son pobres. Esto equivale al 38,61% de la muestra. El resultado es coherente con lo que informa el INDEC en sus resultados oficiales para el primer trimestre de 2023, pues el ente determina que el 41,4% de las personas de Gran Buenos Aires son pobres[1]. No obstante, observado el detalle del INDEC notamos que mientras el 47,0% de las personas de los partidos de GBA están por debajo de la línea de pobreza, únicamente el 17,3% de las personas de CABA son pobres, por lo que un análisis diferencial para cada región quizás sería útil para análisis futuros.

Parte II: Clasificación

1.
Eliminamos de las bases *respondieron* y *norespondieron* todas las variables asociadas a ingresos. También eliminamos *adulto_equiv*, *ad_equiv_hogar* e *ingreso_necesario*. Esta última no había sido creada en la base *norespondieron* por lo que no fue necesario eliminarla.

2.
Para este inciso, en un primer momento habíamos nombrado a *pobre* como la variable dependiente dentro de la base *respondieron* y a todas las demás variables como el

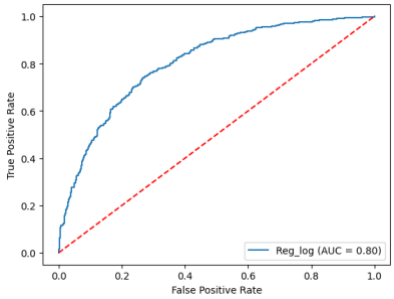
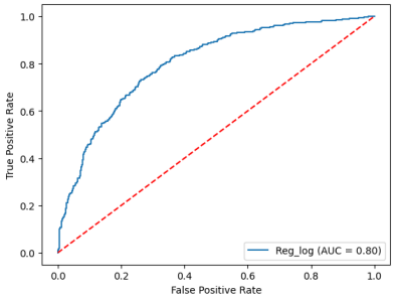
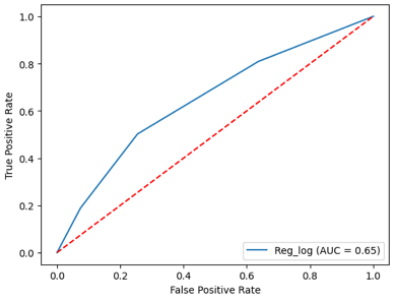
[1] INDEC (Instituto Nacional de Estadística y Censos). (2023). Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos - Primer semestre de 2023 (p. 9). Recuperado de https://www.indec.gob.ar/uploads/informesdeprensa/eph_pobreza_09_2326FC0901C2.pdf

conjunto de X (variables independientes). Luego, añadimos la columna de 1, convertimos a formato *Numpy* las matrices y separamos entre variables de tratamiento y testeo. Sin embargo, cuando intentamos correr el primer modelo del inciso siguiente, notamos que había un error de cálculo dado que había variables con muchos *missings* y variables no numéricas. En efecto, en el *data frame Xopc1* eliminamos todas las variables que tenían una gran cantidad de *missings* y aquellas como *CODUSU* o *IMPUTA* cuyos valores no numéricos impedían la estimación. Como se puede ver, cuando aplicamos a la base *Xopc1* el comando para eliminar *missings* y lo guardamos en *Xopc1bis* notamos que esta nueva base tiene la misma cantidad de filas y columnas que la base anterior, es decir, eliminamos correctamente todas las columnas con datos faltantes. No obstante, cuando intentamos correr nuevamente el modelo del inciso 3, nos encontramos con un error que indicaba la existencia de multicolinealidad muy alta. En particular, había una correlación muy alta entre las variables *PP02C1*, *PP02C2*, *PP02C3*, *PP02C4*, *PP02C5*, *PP02C6*, *PP02C7* y *PP02C8*. Todas estas variables contienen información respecto a las formas en que el individuos buscó trabajo (entrevistas, envío de currículums, carteles, etcétera). Dejamos en el modelo únicamente la variable *PP02C2* que indica si la persona "mandó currículum, puso, contestó avisos". Creemos que esta variable denota la manera más simple de buscar trabajo actualmente. También había correlación muy alta ente *PP02H* y *PP02I*, las cuales indican si la persona buscó trabajo o trabajó en los últimos 12 meses. Nos quedamos con *PP02H* para el modelo porque creemos que ya hay en el modelo otras variables que captan si la persona trabaja (por ejemplo, figurar en la categoría *ESTADO* como ocupado).

3.

La siguiente Figura 3 contiene el resumen de las características de cada uno de los tres modelos desarrollados en este inciso: Logit, Análisis de discriminante lineal y KNN con K=3.

FIGURA 3: Comparación de Modelos

	Logit	*Análisis de discriminante lineal*	*KNN* K = 3
Matriz de confusón	645 | 137 180 | 292	632 | 150 180 | 292	583 | 199 235 | 237
Curva ROC			
AUC	0.8047	0.8024	0.6485
Accuracy	0.747209	0.736842	0.653907

4.

Como podemos observar, el modelo Logit es el mejor predictor dentro de los tres utilizados bajo cualquiera de las medidas de precisión estimadas. En primer lugar, el modelo Logit es el que encuentra la mayor cantidad de Verdaderos Negativos (645) y de Verdaderos Positivos (292), siendo esta última la misma cantidad hallada también por el modelo de Análisis de discriminante lineal (LDA). Esto traduce también en que Logit sea el modelo con mayor nivel de *Accuracy*: 0.8047, pues esta medida de precisión es justamente el cociente entre la suma de Verdaderos Positivos y Negativos sobre la cantidad total de observaciones. Si volvemos a la matriz de confusión, notamos que este modelo, al igual que el de LDA tienen menor cantidad de falsos negativos (180). A priori, cabe mencionar que desde nuestra perspectiva, a la hora de predecir pobreza, es más grave el error de tipo II. Consideremos que, por ejemplo, el gobierno quiere dar asistencia social a aquellos por debajo de la línea de pobreza; no darle asistencia a gente que lo necesita (error tipo II) es más severo que darle asistencia a personas que en verdad no lo necesitaban (error tipo I). Sin embargo, también es óptimo intentar reducir este tipo de error y es ahí que el modelo Logit predomina por encima de LDA (137 versus150).

Asimismo, la curvas de ROC también nos ayudan a ver que en este caso Logit predice mejor la pobreza que LDA. En primer lugar, nos compete marcar que ambos modelos son evidentemente mejor que KNN (con K=3) alejándose de un tipo de predicción azarosa (línea roja punteada) y acercándose hacia el extremo superior izquierdo de máxima precisión predictiva. El ratio entre la Razón de Verdaderos Positivos y la Razón de Falsos Positivos (i.e. curva de ROC) nos indica que Logit es levemente mejor encontrando verdaderos positivos si ajustamos por las veces que el modelo produce falsas alarmas. No obstante, visualmente las diferencias no son tan claras entre Logit y LDA. De ahí que es útil observar el Área Bajo la Curva (AUC) de ROC. Como vemos, la diferencia entre estos dos modelos se da mínimamente a favor de Logit.

5.

Procedemos a predecir la cantidad de pobres dentro de la base *norespondieron* a través del modelo Logit que fue el que tuvo mejores niveles de precisión en relación con los otros modelos. Al hacerlo, encontramos que dentro de la base de aquellas personas que no tenían dato para su nivel de *ITF*, el modelo predice que el 50,83% las personas son pobres. Este dato comprende un total de 916 personas viviendo, predictivamente, en situación de pobreza. Este resultado revela la verdadera utilidad de emplear modelos de predicción a la hora de estimar los niveles de pobreza. Si no utilizamos este modelo predictivo, habría 916 personas que con alto nivel de precisión podemos afirmar están viviendo con un ingreso menor al necesario para acceder a la CBT y las medidas tradicionales de pobreza a partir de la EPH no lo estarían captando por falta de datos. Esto es relevante para poder identificar más precisamente la cantidad de pobres y poder actuar en términos de diseño de políticas públicas.

6.

Consideramos que correr los modelos predictivos con todas las variables de la encuesta no es la mejor opción pues incluir las variables indiscriminadamente puede agregar ruido injustificado al modelo, aumentando la varianza de los resultados. Si bien en estos modelos la inclusión de variables irrelevantes es menos problemático que en modelo tradicionales como por ejemplo *Ordinary Least Squares (OLS)*, la predicción se torna menos precisa cuando incluimos variables que no aportan significatividad haciendo menos eficiente el modelo. Es por eso que decidimos eliminar de la base todas las variables que responden a metodología y documentación de la EPH: Aglomerado,

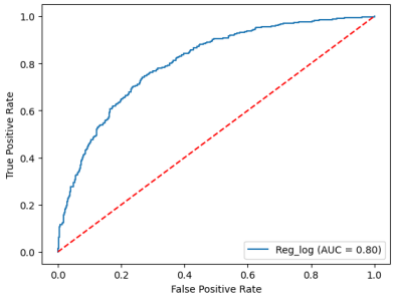
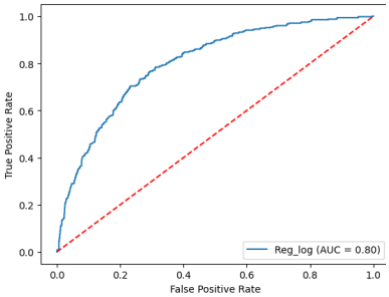
Ponderación, Número de Hogar, Componente de la Encuesta y Si se hizo o no la encuesta individual. Además, eliminamos las variables Trimestre, Año y Región por tener valores iguales para todos las observaciones. Por último, eliminamos también la variable *CH13* que indica la relación de parentesco entre el individuo y el jefe de hogar, pues no creemos que tiene valor a la hora de predecir pobreza. Por el contrario, nos quedamos en el modelo con las variables que teóricamente creemos que tienen poder explicativo a la hora de predecir pobreza:

- *AGLOMERADO* (CABA o GBA): como mencionamos previamente, no es indistinta la región pues los resultados de pobreza para cada sector son bien distintos, siendo CABA un aglomerado con niveles notoriamente menores de pobreza.
- Sexo *(CH04)*, Edad *(CH06)* e Interacción entre sexo y edad (*Sexo_edad*): la pobreza tiende a concentrarse en persona más chicas pues son las que tienen menos posibilidades de valerse por sus propios medios para romper su condición de pobre, a la vez que hay menor cantidad de personas adultas pobres porque las personas pobres tienen en promedio menor esperanza de vida. Asimismo, los hogares de menos recursos tienden a tener mayor cantidad de hijos. El sexo es relevante en especial porque los empleos e ingresos a los que acceden hombres y mujeres son esencialmente diferentes.
- Tipo de unión de las parejas del hogar *(CH07)*: la cohabitación es una condición más frecuente en niveles socioeconómicos bajos y el matrimonio en niveles altos.
- Tipo de cobertura médica *(CH08)*: mayor nivel de ingresos se asocia con tener obra social o prepaga y en general, a no depender de prestaciones públicas.
- Saber leer y escribir *(CH09)*, Asistencia a nivel educativo *(CH010)*, Tipo de nivel educativo *(CH011,* i.e. público o privado), Nivel educativo más alto alcanzado *(CH12)*, Finalización o no de ese nivel *(CH013)*, Nivel educativo *(NIVEL_ED)*: la pobreza tiende a correlacionar positivamente con niveles de educación más bajos, con la no finalización de la escuela y con haber asistido a escuelas públicas.
- Lugar de nacimiento *(CH15)* y Lugar de residencia 5 años atrás *(CH16)*: consideramos que esto podría ser un buen indicador de estabilidad de hogar y movilidad intergeneracional.
- Categoría de ocupación, desocupación o inactividad *(ESTADO)* y Tipo de ocupación *(CAT_OCUP)*: pensamos que es importante entender el tipo de ocupación pues es un buen *proxy* del nivel de ingresos y por ende, de la relación del individuo respecto a la línea de pobreza.
- Tipo de inactividad (*CAT_INAC*): creemos que es relevante distinguir según las razones que llevaron a un individuo a no estar activo. Por ejemplo, no es indiferente para la medición ser ama de casa, jubilado o estudiante.
- Búsqueda de trabajo en el último tiempo (*PP02H*), Si envió currículum a algún sitio (*PP02C*) o La razón por la que no buscó trabajo (*PP02E*): relevante para entender si la persona se encuentra buscando trabajo a fin de poder aumentar por ejemplo su nivel de ingresos o conseguir trabajo, o bien, si la persona ya posee un trabajo estable.

En la Figura 4 se muestra la comparación entre ambos modelos Logit. Como podemos notar, no existen diferencias significativas entre ambos modelos. El nuevo modelo ha detectado menor cantidad de Verdaderos Positivos pero a su vez detectó mayor cantidad de Verdaderos Negativos. En términos de errores, el nuevo modelo tiene más errores del tipo Falso Positivo pero menor cantidad de errores Falsos Negativos. Por un lado, esto es útil porque priorizamos que no haya gente pobre sin ser detectada como pobre por encima

de que haya gente no pobre clasificada como pobre. Además, podemos pensar que en el caso de los Falsos Positivos, dadas las características del individuo, existe un riesgo elevado de que la persona caiga bajo la línea de pobreza en el futuro. De este modo, si bien la predicción es errónea, puede ser una alerta de que la persona está cercana a la línea de pobreza. Sin embargo, es pequeña la cantidad de errores tipo II que el modelo Logit II evita en comparación con la cantidad de errores tipo I que este modelo añade. Este resultado se ve reflejado en las curvas de ROC. Si bien la forma es muy similar en ambos casos, el modelo nuevo tiene un nivel de AUC levemente inferior. Así también sucede cuando miramos el nivel de *Accuracy*: el modelo previo es levemente más exacto. Creemos que este bajo nivel de mejora puede deberse a que ya habíamos eliminado numerosa cantidad de variables en anteriores incisos (las asociadas a ingresos, las que contenían *missings* y aquellas con problemas de multicolinealidad), por lo cual al estimar el primer modelo Logit la base estaba lo suficientemente limpia para que no hubiese mucho ruido y la predicción fuese precisa.

FIGURA 4: Comparación de Modelos Logit

	Logit I *(con todas las variables)*	*Logit II* *(con variables seleccionadas)*
Matriz de confusón	645 | 137 180 | 292	632 | 150 177 | 295
Curva ROC		
AUC	0.8047	0.8006
Accuracy	0.747209	0.739234

Como extra, nos pareció relevante estimar la proporción de personas pobres dentro del subconjunto que no respondió el nivel de ingresos total familiar, esta vez, usando el modelo con variables específicas. Similarmente a lo hallado previamente, el modelo determina que 52,22% de la muestra *norespondieron* son pobres. Es lógico pensar que los valores de ambas predicciones Logit se asemejan dada la similitud de ambos modelos en términos de precisión. En este caso, el modelo predice una cantidad levemente mayor de personas pobres, lo cual es útil a priori si tenemos en cuenta que al medir pobreza, los Falsos Negativos pueden llegar a ser una peor alternativa.