Universidad de
SanAndrés

Universidad de San Andrés

E337: Big data

Trabajo Práctico N°2

Autores:

Rodrigo Braga

Fernando Kricun

Julián Ramos Cancio

Fecha de entrega: 25 de octubre de 2023

Consignas

Parte I: Analizando la base

1. Utilizando información disponible en la página del INDEC, expliquen brevemente cómo se identifica a las personas pobres.

Se identifica como persona pobre a aquella que forma parte de un hogar que no tiene la capacidad de satisfacer –por medio de la compra de bienes y servicios– un conjunto de necesidades alimentarias y no alimentarias consideradas esenciales.

Metodológicamente, primero se calcula una línea de pobreza y luego se contrasta contra ella el ingreso familiar total. Si este último es menor que la línea de pobreza, determinada por la Canasta Básica Total (CBT), entonces todos los integrantes del hogar son identificados como pobres. La CBT se obtiene a partir del cálculo de la Canasta Básica Alimentaria (CBA). El procedimiento para obtener la CBA toma en cuenta los requerimientos normativos kilocalóricos y proteicos del hogar. Dado que los requerimientos son diferentes según la edad, el sexo y otras características, es necesario tomar una unidad de referencia común para poder reflejar estas características. La unidad que se toma es el varón adulto de entre 30 y 60 años. Luego, se compara las características de todas las personas con las de un varón adulto entre las edades mencionadas.

Así, se suman los requerimientos calóricos de todos los integrantes del hogar y se establece la CBA para cada hogar. Para obtener la CBT se multiplica la CBA por un coeficiente que relaciona el gasto en alimento y el gasto total de las familias. Finalmente, este CBT es el que se compara con los ingresos totales por hogar.

2. Entren a la página https://www.indec.gob.ar/ y vayan a la sección Servicios y Herramientas ¿Bases de datos. Descarguen la base de microdatos de la Encuesta Permanente de Hogares (EPH) correspondiente al primer trimestre de 2023 en formato xls (una vez descargada, la base a usar debería llamarse usu_individual_T123.xls). En la página web, también encontrará un diccionario de variables con el nombre de "Diseño de registro y estructura para las bases preliminares (hogares y personas)"; este archivo les indica qué significa cada variable que aparece en la base de datos, en particular, en la sección de Diseño de registros de la base Personas.

(b) Si hay observaciones con valores que no tienen sentido, descártenlas (ingresos y edades negativos, por ejemplo).

Luego de haber estudiado la base de datos, consideramos como valores erróneos o sin sentido a aquellos que pertenecen a variables de ingresos y edades, y son negativos. Además, notamos que algunas fechas de nacimiento no son las correctas dadas las edades reportadas. En estos casos decidimos proceder eliminando las observaciones con estas fechas de nacimiento erróneas. Estas observaciones fueron simples de identificar debido a que cuando la fecha de nacimiento era errónea, siempre se reportaba la misma fecha: 1/1/1900.

En primer lugar, para eliminar todas las observaciones con valores negativos generamos la variable types para observar qué formatos tenían las variables de la base de datos [Primer paso].

Luego de observar los 4 formatos que tenemos y la lista de variables con estos formatos, procedimos a borrar las observaciones con valores negativos de las variables numéricas (con formato float64 e int64). Para esto corrimos un loop en que solo mantenemos la observación si los valores de todas las variables numéricas son positivos o missing (es importante notar que puede haber missing y no ser un valor "sin sentido") [Segundo paso].

Por último, eliminamos las observaciones si la fecha de nacimiento reportada era 1/1/1900 [Tercer paso].

(d) Realicen una matriz de correlación con las siguientes variables: CH04, CH07, CH08, NIVEL ED, ESTADO, CAT INAC, IPCF. Comenten los resultados.

Luego de correr la matriz de correlaciones entre las variables especificadas, es posible observar una alta correlación positiva (0.81) entre "ESTADO" y "CAT_INAC". Ambas son variables categóricas. Sin embargo, no es posible afirmar mucho más sobre esta correlación debido a que los valores que toma la variable "CAT_INAC" no siguen un orden particular. Es decir, si observáramos una correlación positiva entre nivel educativo y nivel de ingresos, sí podríamos decir algo si el nivel educativo sigue un orden ascendente (el nivel educativo alcanzado es mayor cuanto mayor es el valor que se le asigna en la variable categórica). No obstante, esto no ocurre en este caso. "CAT_INAC" no sigue un orden particular.

Existe una correlación positiva (0.45) también entre la variable "ESTADO" y "CH07" (estado civil). Aquí sí se podría dar alguna intuición debido a que ambas variables categóricas siguen un orden. En "CH07" los valores más chicos (1 y 2) reflejan el estado civil unido y casado, mientras que los más altos (3, 4 y 5) son asignados a las personas separadas/divorciadas, viudas y solteras. En "ESTADO" el valor de 1 indica el estado laboral ocupado, mientras que valores más altos (2, 3 y 4) reflejan los estados desocupado, inactivo y menor de 10 años. Por lo tanto, esta correlación positiva podría estar diciéndonos algo sobre la relación entre la ocupación y el estado civil de las personas.

Por último, otra de las correlaciones notorias que vemos (0.44) es entre "CAT_INAC" y "CH07". Sin embargo, tal como señalamos previamentes, "CAT_INAC" no sigue un orden particular, por lo que no nos parece claro que esta correlación pueda estar indicándonos algo.

(e) ¿Cuántos desocupados hay en la muestra? ¿Cuántos inactivos? ¿Cuál es la media de ingreso per cápita familiar (IPCF) según estado (ocupado, desocupado, inactivo)?

Hay 235 desocupados y 2264 inactivos en la muestra.

Para los ocupados, la media del ingreso per cápita familiar es de $ 95591,9. Mientras que

para los desocupados es $ 28336,22 y para los inactivos $ 47349,46.

3. Uno de los grandes problemas de la EPH es la creciente cantidad de hogares que no reportan sus ingresos (ver por ejemplo el siguiente informe). ¿Cuántas personas no respondieron cuál es su ingreso total familiar (ITF)? Guarden como una base distinta llamada respondieron las observaciones donde respondieron la pregunta sobre su ITF. Las observaciones con ITF = 0 guárdenlas en una base bajo el nombre norespondieron.

1506 personas no respondieron cuál es su ingreso familiar total (ITF).

5. Por último, agreguen a respondieron una columna llamada pobre que tome valor 1 si el ITF es menor al ingreso necesario que necesita esa familia, y 0 en caso contrario. ¿Cuántos pobres identificaron?

Identificamos 1398 pobres. Esto equivale a un 35.88 % de pobres de la base respondieron.

Parte II: Clasificación

4. ¿Cuál de los tres métodos predice mejor? Justifiquen detalladamente utilizando las medidas de precisión que conocen.

Para los 3 métodos utilizamos dos medidas de precisión distintas: el área bajo la curva ROC (AUC) y el accuracy del modelo.

La curva ROC grafica la relación entre la tasa de verdaderos positivos (eje y) y la tasa de falsos positivos (eje x). Consecuentemente, cuanto más se acerque la curva al punto (0,1), mayor será la probabilidad de que el método clasifique correctamente los valores positivos y menor la probabilidad de que los clasifique de forma incorrecta. La medida AUC permite contabilizar estas probabilidades y establece un índice de 0 a 1, en la que los mejores clasificadores tenderán a acercarse a 1.

Por su parte, la medida de accuracy o exactitud suma todos los verdaderos positivos y verdaderos negativos, y los divide por el total de positivos y negativos. Accuracy mide la fracción de predicciones que el modelo realizó correctamente. Al igual que AUC, es un índice de 0 a 1, en la que los métodos con más exactitud obtienen una calificación más cercana a 1.

Analizando los resultaods, observamos que la regresión logit puntúa mejor que los otros dos métodos en ambas medidas de precisión. En otras palabras, este método tiene una mayor probabilidad de acierto de la predicción de la pobreza (AUC) y tiene una mayor fracción de predicciones correctas (accuracy). En segundo lugar viene el modelo de análisis de discriminante lineal, cuyos AUC y accuracy son menores que los del método logit pero mayores que los del modelo de vecinos cercanos (KNN).

5. Con el método que seleccionaron, predigan qué personas son pobres dentro

de la base norespondieron. ¿Qué proporción de las personas que no respondieron pudieron identificar como pobres?

En el cuadro de código del ejercicio indicamos qué cantidad y proporción de las personas que no respondieron fueron identificadas como pobres tanto en la base respondieron como en la norespondieron. La idea detrás de esto es poder comparar entre aquellas personas que sí accedieron a dar el dato de sus ingresos y las que no. Tal como observamos, si bien pueden existir errores de predicción (aunque nuestro modelo prediga 35.07 % de pobres en la base respondieron y el valor real es 35.88 %), el porcentaje de personas pobres en la base norespondieron es mucho menor (1.59 %). Esto va en línea con la idea de que las personas que ganan más tienden a subreportar, o directamente no reportar, sus ingresos.

6. Noten que para correr los tres métodos se utilizaron todas las variables disponibles como predictores. ¿Les parece esto correcto? ¿Qué variables habrían conservado? Con las variables seleccionadas, implementen únicamente el modelo logit nuevamente y comparen las medidas de precisión obtenidas con los resultados del modelo logit anterior. ¿Cambió mucho la precisión?

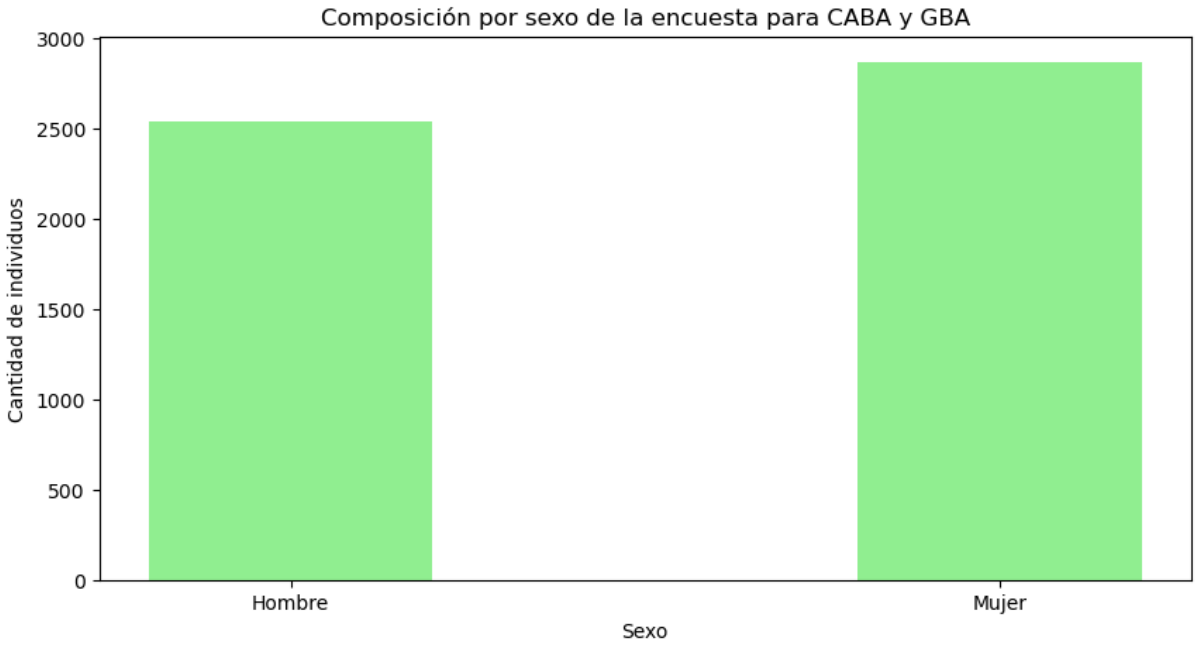
Con respecto a si nos parece correcto, creemos que lo mejor es siempre poder disponer de todas las variables que se puedan a la hora realizar las predicciones y luego ir definiendo cuáles son las mejores para utilizar. Por lo tanto, para responder esta pregunta es importante siempre analizar la situación y el contexto de lo que estamos prediciendo.

Naturalmente, si realizamos un análisis como el de este trabajo, es porque no contamos con los datos de los ingresos. Si tuviéramos esos datos, incorporarlos a la calibración del modelo puede ser útil según el contexto. Si contamos con los datos de la canasta que nos permiten crear la línea de pobreza, como sucede en la EPH, no sería necesario ningún modelo, ya que con observar los ingresos y tener la CBT podemos determinar quién es pobre y quién no. Sin embargo, si los datos de la canasta faltaran, por ejemplo, utilizar los datos de ingresos puede ser muy útil para identificar a las personas en situación de pobreza.

Contando con la disponibilidad de variables de salarios, decidimos incluir solo dos: el monto del ingreso total familiar (ITF) y el monto del ingreso per cápita familiar (IPCF). La idea detrás de esto es que estas dos variables agrupan la información de muchas otras variables de ingreso. Junto a estos datos de las personas también incluimos el mismo vector de variables explicativas que previamente utilizamos.

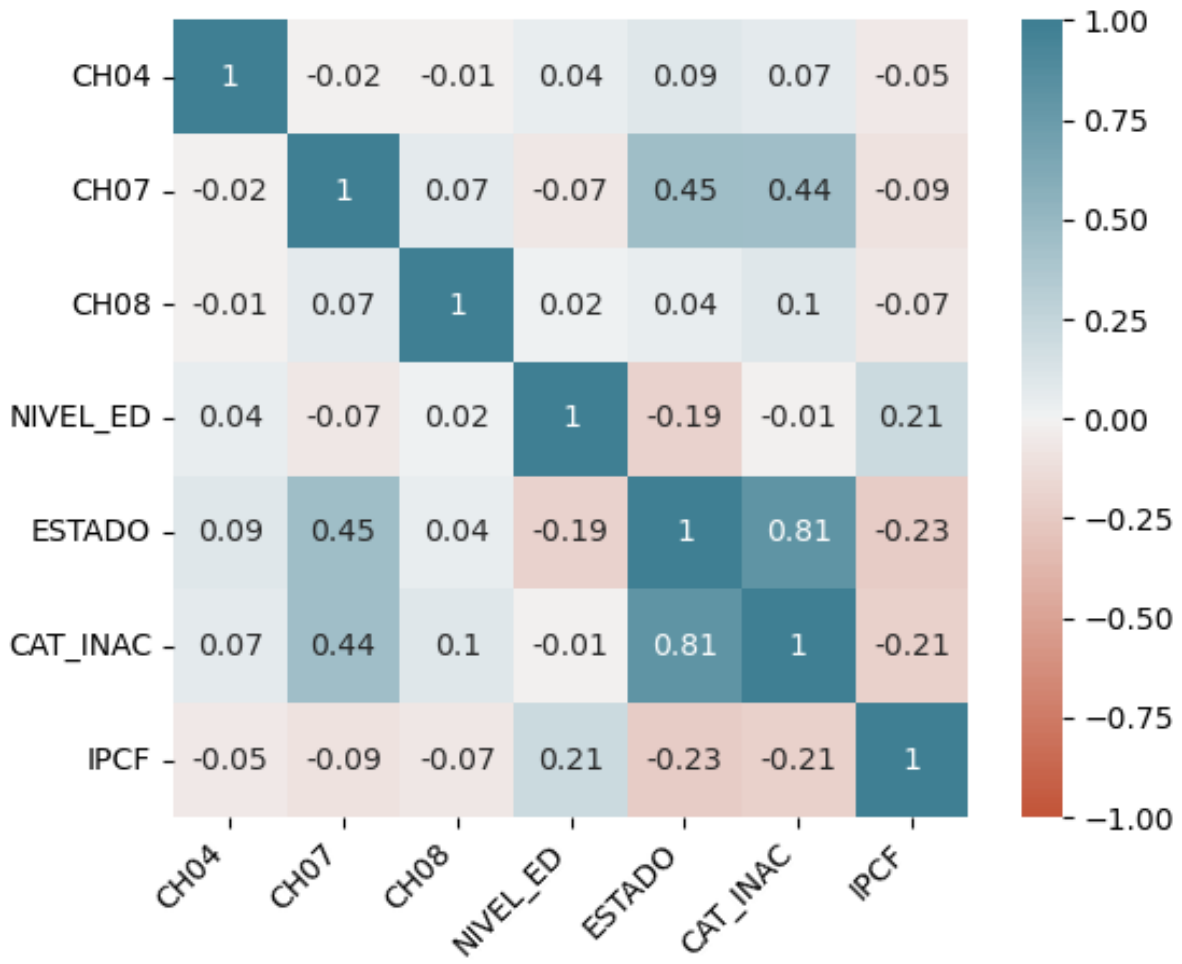
La nueva regresión logística arroja un AUC y Accuracy Score mayor a los de los tres modelos previamente calibrados. Por lo tanto, es posible afirmar que contar con los datos de los ingresos familiares totales y los ingresos per cápita podemos disponer de un modelo más preciso.

Anexo I



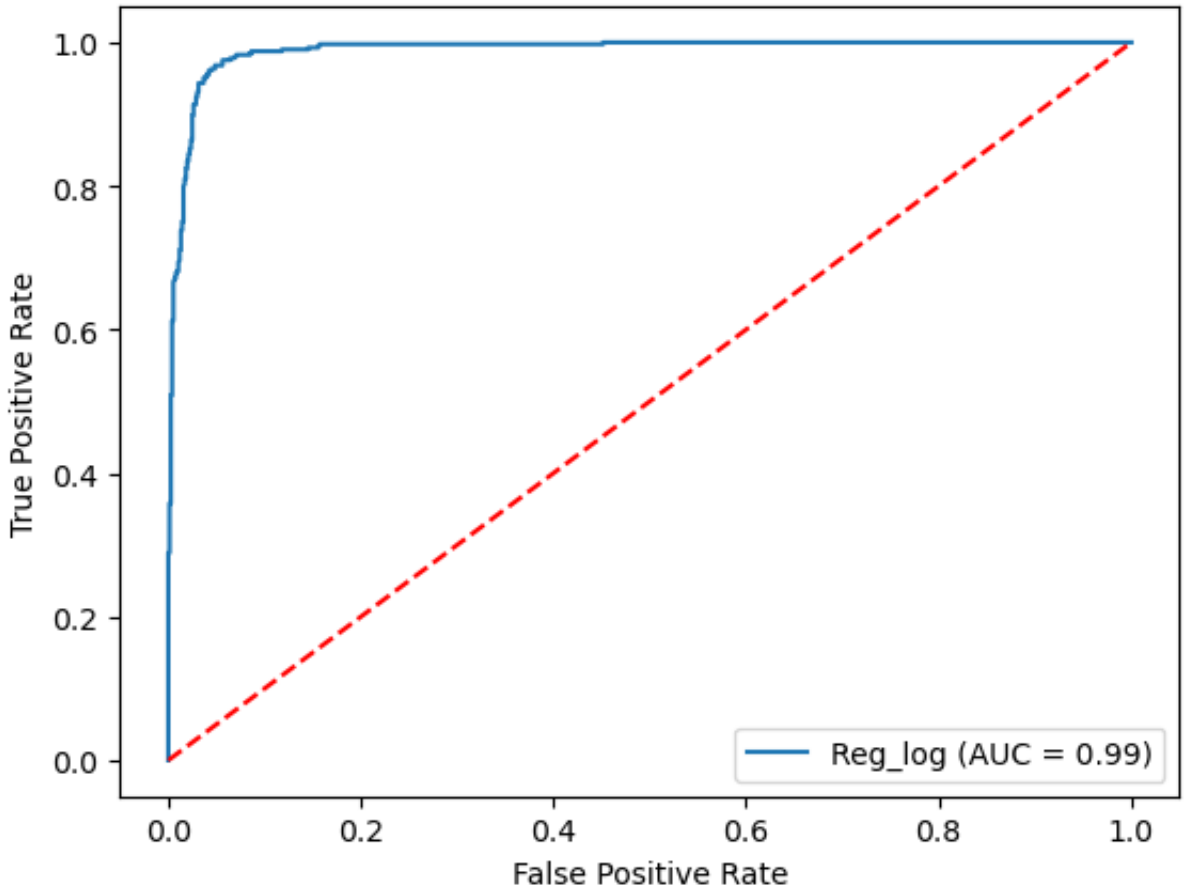
Composición por sexo de la EPH para los aglomerados de CABA y GBA.

Anexo II



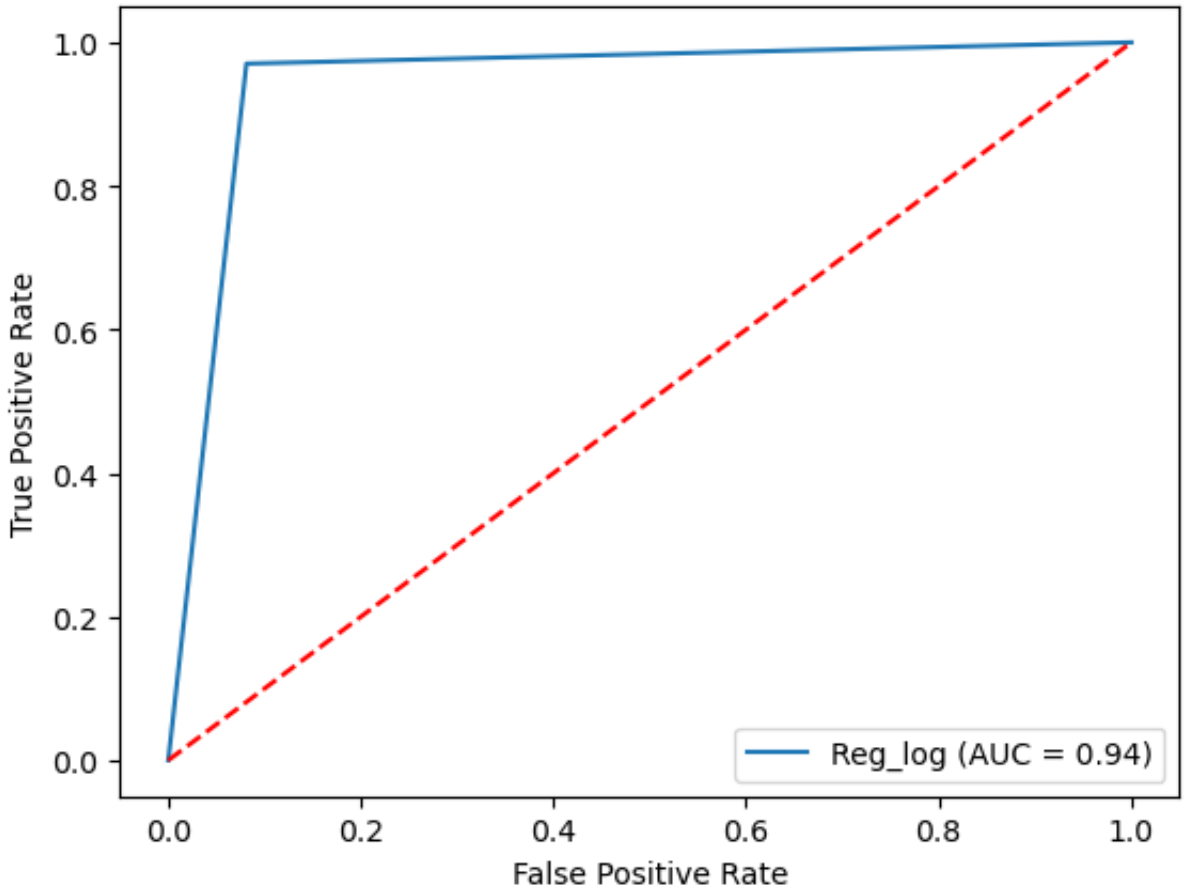
Matriz de correlación para las variables *CH04, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC* e *IPCF*.

Anexo III



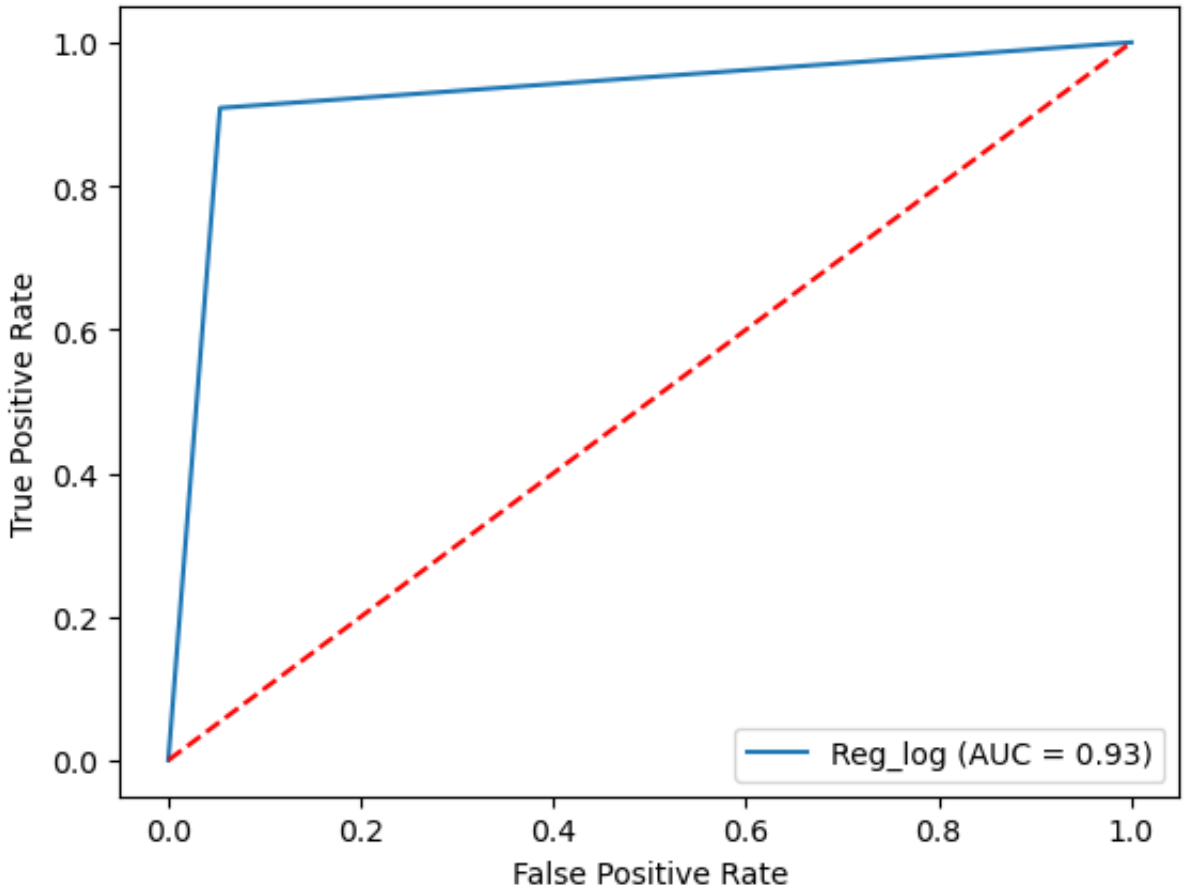
Curva ROC del modelo logit sin variables de ingresos.

Anexo IV



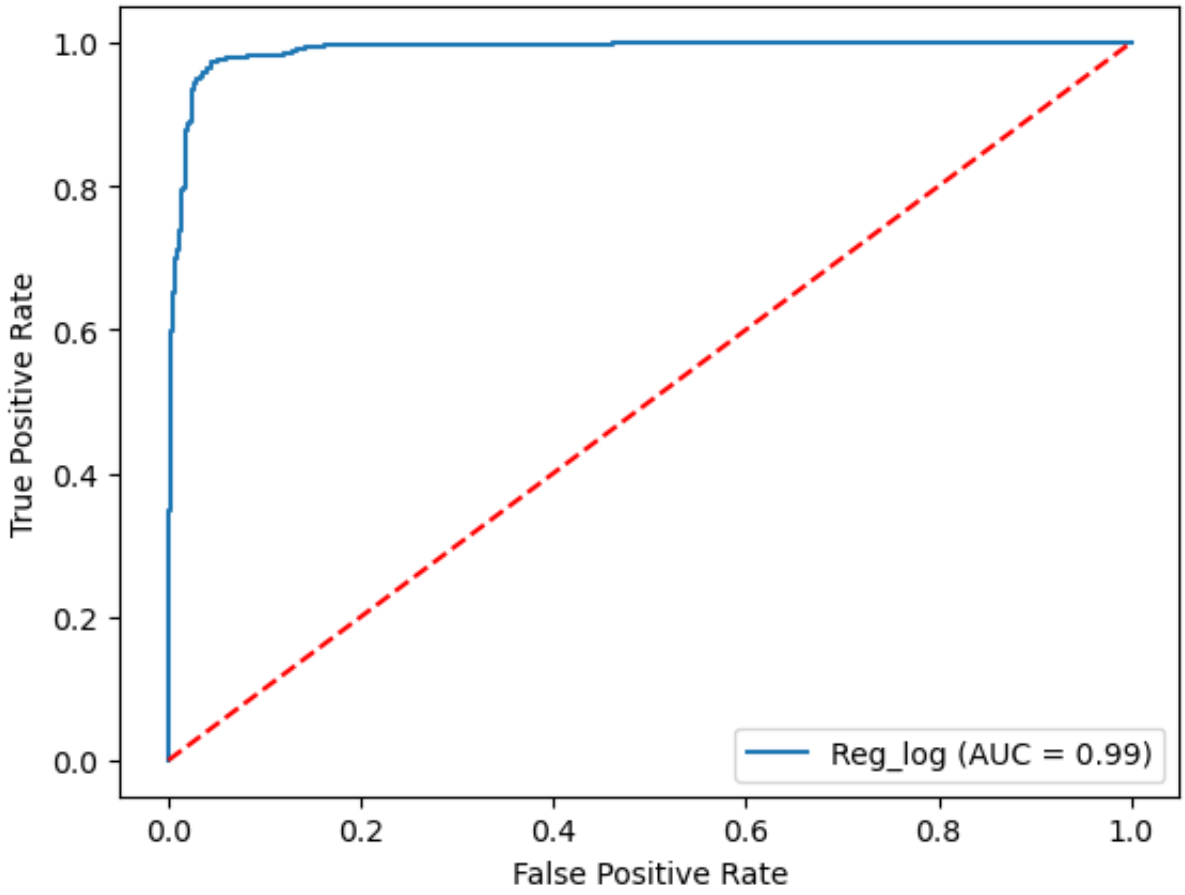
Curva ROC del modelo de análisis de discriminante lineal.

Anexo V



Curva ROC del modelo de vecinos cercanos (KNN) con k = 5.

Anexo VI



Curva ROC del modelo logit con variables de ingresos.