Universidad de San Andrés
Departamento de Economía
Buenos Aires
octubre, 2023

Trabajo Práctico N° 2

Big Data, Aprendizaje y Minería de Datos

Camila Riancho

Diego Fasan

Ronny M. Condor

Profesora: Noelia Romero

Asistente: Victoria Oubiña

Parte 1: Analizando la base

Ejercicio 1

En Argentina, la pobreza se mide con el método de la línea de pobreza, que consiste en identificar como pobres a aquellas personas pertenecientes a hogares cuyos ingresos no son suficientes para comprar una canasta de bienes y servicios (Canasta Básica Total) considerados como esenciales para satisfacer necesidades alimentarias y no alimentarias básicas (INDEC, 2023) [1].

Para identificar a las personas pobres, por lo tanto, se necesita contar en primer lugar con el valor de la Canasta Básica Total (CBT). Este se obtiene expandiendo el valor de la Canasta Básica Alimentaria (CBA), que es una canasta de alimentos que satisface los requerimientos energéticos y proteicos mínimos de los hogares, y es determinada por los hábitos de consumo de la población de una misma región reportados en la ENGHo, y de sus necesidades calóricas y proteicas. La CBA se calcula para un adulto varón de entre 30 y 60 años, al que se lo denomina .ªdulto equivalente", y el valor de dicha canasta se obtiene en base al IPC del período para el que se está midiendo la pobreza. Cada hogar, en función del número, el género y la edad de sus miembros, sumará una cierta cantidad de *adultos equivalentes* y eso determinará el valor de la CBA para ese hogar (INDEC, 2023).

Una vez que se obtiene el valor de la CBA, se calcula la CBT a partir del Coeficiente de Engel, que representa la relación entre gastos alimentarios y totales, y que se calcula a partir de los gastos reportados por la población en la ENGHo, y que se actualiza en función del precio relativo de los alimentos relevado en el IPC. Para terminar, se considera como persona pobre a aquella perteneciente a un hogar pobre, es decir, aquel cuyo ingreso total familiar es menor al valor de la CBT que le corresponde a ese hogar (INDEC, 2023).

Ejercicio 2

a) Eliminamos de la muestra aquellas observaciones para las que la variable `AGLOMERADO` toma un valor distinto a 32 y 33, que son los códigos correspondientes a CABA y al Gran Buenos Aires.

b) Para evaluar qué observaciones toman valores sin sentido, observamos los valores mínimos y máximos de cada columna. En primer lugar, notamos que la variable que mide la edad en años (CH06) tomaba valores menores a 0 para algunas observaciones. Dado que la edad no puede ser negativa, eliminamos dichas observaciones. En segundo lugar, analizamos si las variables de

[1] INDEC (2023). Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos. Consultado el 20 de octubre de 2023 de https://www.indec.gob.ar/uploads/informesdeprensa/eph_pobreza_09_2326FC0901C2.pdf

ingreso que vamos a usar en el trabajo (IPCF e ITF) tomaban valores sin sentido (menores a 0), pero como este no era el caso, no fue necesario hacer ninguna modificación en ese sentido. Otras variables de ingreso, como PP08D1 y PP08D4, pertenecientes a la sección Ïngresos de la ocupación principal de los asalariados"sí toman valores negativos, pero como dichas columnas serán eliminadas al momento de hacer la predicción, consideramos innecesario excluir de la muestra las observaciones con valores negativos en esas variables.

Más aún, notamos que muchas variables toman como valores máximos a 9, 99, 999, 9999 o 99999. En el codebook, observamos que dichos códigos corresponden a valores faltantes. Explorando la base y el codebook, notamos que en la mayoría de los casos cuando el máximo valor con sentido que puede tomar una variable es menor a 9 (por ejemplo, en el caso de índices o preguntas con opciones), el valor 9 se le asigna a los missing values. Cuando el máximo valor que puede tomar una variable es mayor a 9 y menor a 99, los valores faltantes se identifican con el código 99, y así sucesivamente. Dado que estos valores pueden afectar nuestras estimaciones posteriores, los reemplazamos por valores faltantes. La única excepción fue la variable de la edad (CH07), cuyo valor máximo es 99, pero no lo consideramos como un valor faltante porque podría ocurrir que haya personas con 99 años de edad en la muestra. Además, observamos que para el caso de la variable `CH08` el código 9 se asigna a valores faltantes, aunque la variable toma valores mayores a 9, así que en esa columna también reemplazamos a los 9 por valores faltantes. Por último, para la variable .ESTADO"los valores faltantes se identifican con el valor 0 en lugar de 9, por lo que reemplazamos los 0 por faltantes en dicha columna.

c) A continuación, presentamos un gráfico de barras que ilustra la composición por sexo de la muestra filtrada. Observamos que el 52 % de los individuos de la muestra son mujeres y el 48 % son hombres.

d) Generamos una matriz de correlación entre las variables que miden el sexo del individuo, su estado civil, el tipo de cobertura médica que posee, su nivel educativo, su condición de actividad (ocupado, desocupado, inactivo), su categoría de inactividad (jubilado, rentista, estudiante, etc.), y el ingreso per cápita familiar.

La matriz muestra que hay una débil correlación entre las variables. La mayoría tiene correlaciones entre -0.2 y 0.4. La correlación más fuerte se da entre condición de actividad y categoría de inactividad. Esto último se explica por el hecho de que solo las personas que se reportan como inactivas en la variable que mide la condición de actividad contestan luego la pregunta sobre la categoría de inactividad.

Llama la atención la baja correlación de todas las variables con el IPCF, dado que desde la

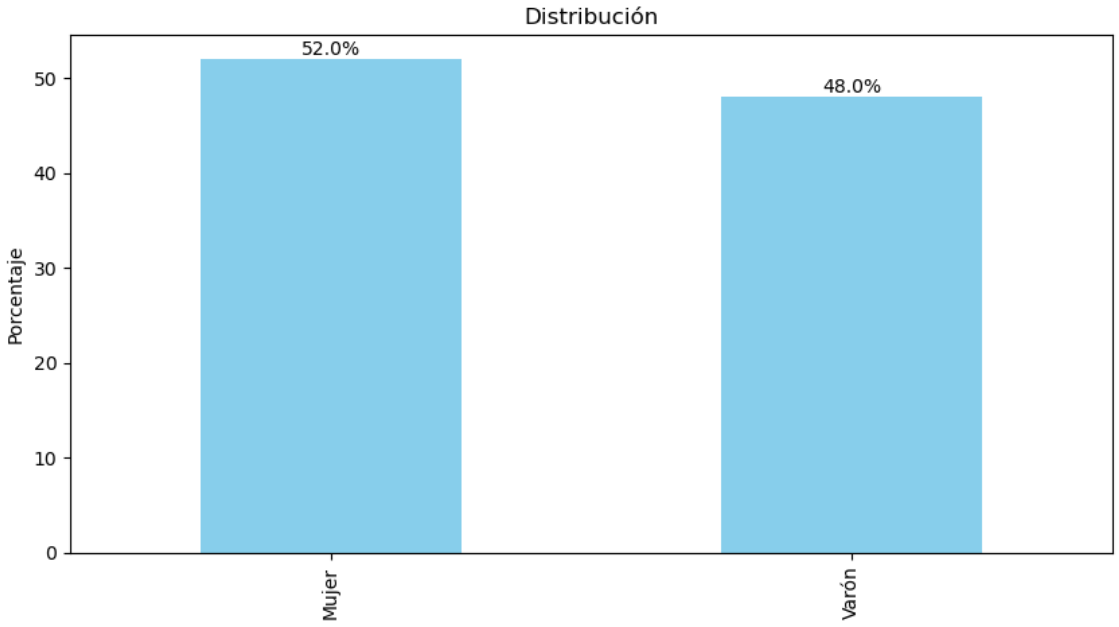


Figura 1: Distribución según sexo

teoría esperaríamos que sean variables relevantes para explicar el ingreso. Entendemos que lo que ocurre es que la mayoría se trata de variables categóricas no ordinales, de manera que no es posible encontrar relaciones lineales entre ellas sin convertirlas en dummies. La única variable categórica ordinal es la del nivel educativo, pero las categorías tampoco están en orden, ya que el código 1 corresponde al primario incompleto y el 7 designa a las personas sin instrucción.

e) En la muestra hay 3523 ocupados, 286 desocupados y 2837 inactivos. La media del Ingreso per Cápita familiar es 59.579,44 pesos para los ocupados, 25.536,02 pesos para los desocupados y 40.068 para los inactivos.

f) Generamos una columna que mide los valores de adulto equivalente para cada individuo, y otra columna que agrega estos valores por hogar.

Ejercicio 3

Si bien en la columna `ITF` no hay valores faltantes, observamos que de un total de 7571 personas, 3390, es decir, el 44,77 %, reportó un Ingreso Total Familiar de 0, lo que se considera como una no respuesta.

Ejercicio 4

Agregamos a la base la columna variable `ingreso_necesario`, que corresponde al ingreso total familiar mínimo que necesita un hogar para no ser pobre.

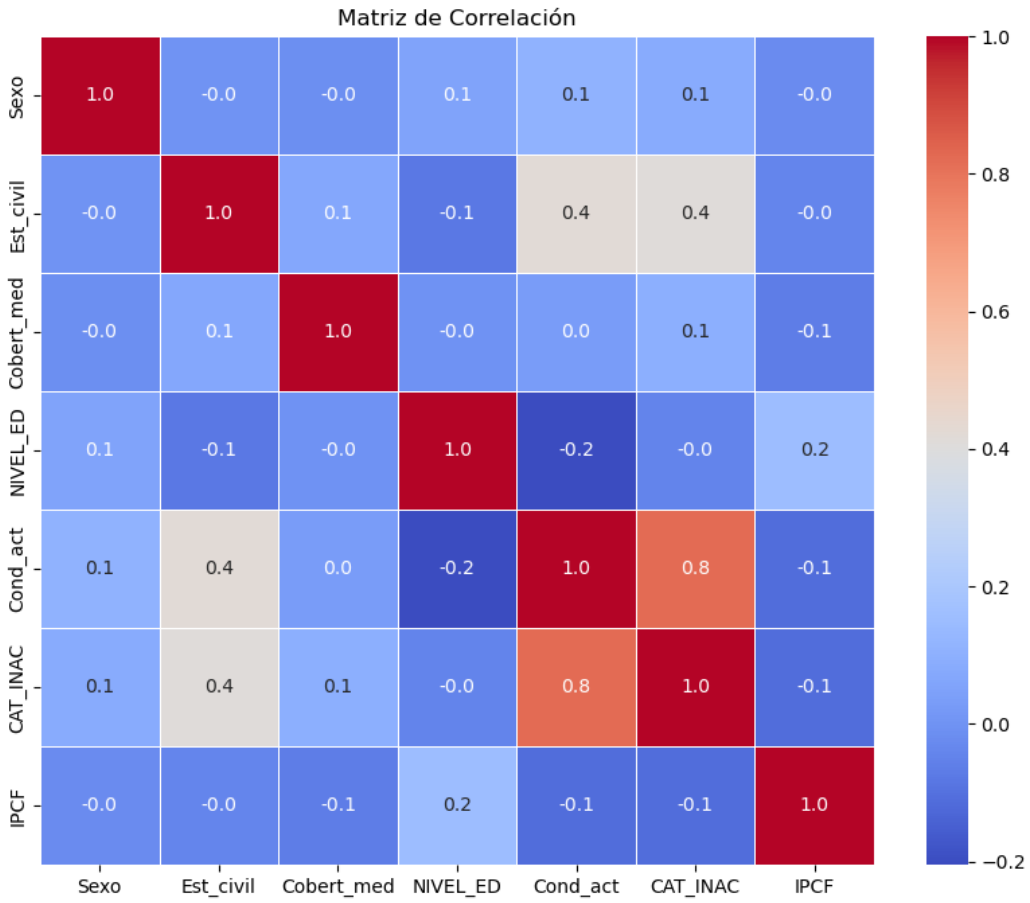


Figura 2: Matriz de correlación

Ejercicio 5

Agregamos a la base la columna pobre, que toma valor 1 cuando el Ingreso Total Familiar (ITF) es menor al ingreso necesario, y 0 en caso contrario. De esa manera, identificamos 1555 personas pobres, es decir, el 37,19 % de la muestra que incluye solo a los que reportaron un ITF positivo.

Parte 2: Clasificación

Ejercicio 1

Seleccionamos todas las categorías de ingresos[2] y las variables relacionadas al cálculo de adulto equivalente y las eliminamos de ambas bases de datos.

[2]En la EPH tenemos 7 categorías de ingreso, que en total son 41 variables

Ejercicio 2

Ahora trabajaremos con la base `respondieron` que utilizaremos posteriormente en los modelos predictivos. Sin embargo, antes de dividir la muestra, debemos hacer una pequeña limpieza. Borramos las variables para las que todos sus valores son missing. También modificamos la variable `MAS_500` relacionada al número de personas que viven en el aglomerado, que estaba en formato texto, y la pasamos a formato numérico. Asimismo, eliminamos variables irrelevantes para el modelo como el número del hogar, el año, el trimestre, la región (todos los valores corresponden a Buenos Aires) y la variable "PONDERA", que no tiene significado conceptual.

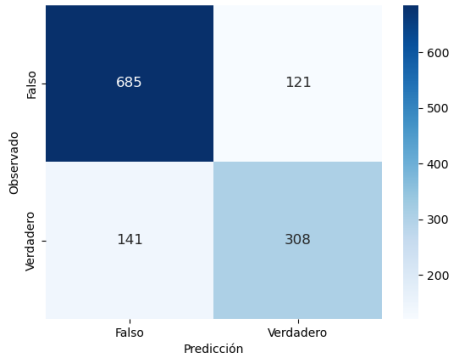
Asimismo, notamos que hay muchas variables con un número relevante de valores faltantes, ya que se trata de preguntas que se refieren a distintos sectores específicos de la población. Dado que en su mayoría son variables categóricas en las que el 0 no es una categoría, reemplazamos los valores faltantes por 0, ya que de lo contrario deberíamos haber eliminado muchas variables relevantes.

Más aún, la mayoría de las variables son discretas nominales, por lo que colocarlas en el modelo directamente no es recomendable. En su lugar, creamos variables dummies para cada categoría de cada variable discreta nominal. Asimismo, notamos que en la variable "NIVEL ED", que es una variable discreta ordinal que mide el nivel educativo, las categorías no están ordenadas, ya que el 1 corresponde a la primaria incompleta, mientras que el 7 corresponde a la falta de instrucción. Para que las categorías estén ordenadas, reemplazamos los 7 por 0 en dicha variable.

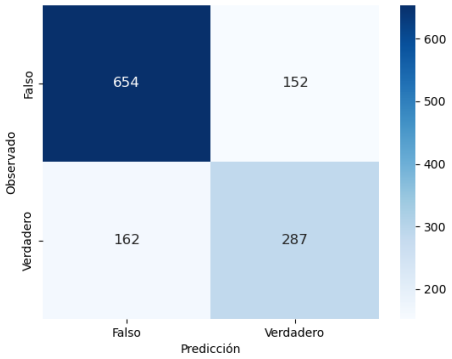
Una vez hecha esta limpieza, definimos el dataframe `X` que contiene a todas las variables predictoras, y la variable `y` como la variable dependiente que solo toma valores 1 si se es pobre y 0 si no se lo es. Finalmente, dividimos nuestra muestra en en una parte de entrenamiento y otra de testeo con una dimensión de 70 % y 30 % respectivamente. En total tenemos 2927 observaciones de entrenamiento y 1255 de prueba.

Ejercicio 3

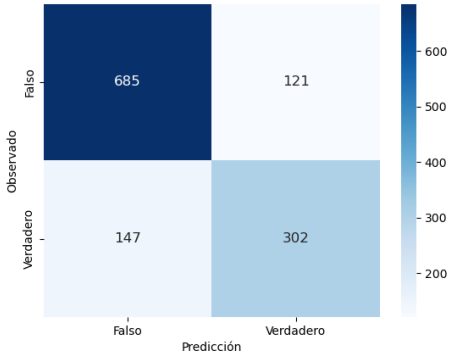
Estimamos los modelos Logit, KNN y Análisis Discriminante Lineal. Para cada uno estimamos la curva de ROC, para ver el trade-off entre los verdaderos positivos y los falsos positivos. También computamos la matriz de confusión para ver la cantidad de observaciones fuera de la diagonal principal, y finalmente calculamos el AUC y el Accuracy score.



(a) Logit (Accuracy = 0.791)



(b) KNN (Accuracy = 0.750)



(c) LDN (Accuracy = 0.786)

Figura 3: Matriz de confusión: Logit, KNN y LDA

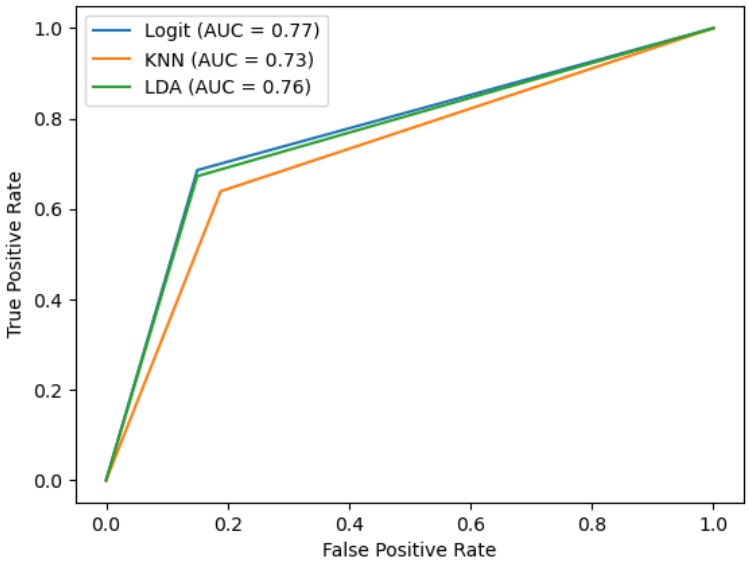


Figura 4: Curvas ROC: logit, KNN y LDA

Ejercicio 4

El modelo que mejor predice es el Logit. En la figura 3 podemos ver que el modelo logit es el que tiene mejor accuracy. Por otro lado, en la figura 4 vemos que tiene una mayor AUC. En resumen, el modelo logit toma menores falsos negativos y falsos positivos y por tanto es el modelo que genera la mejor predicción de la pobreza.

Ejercicio 5

En este punto usamos parte de la muestra de la EPH que habíamos apartado. Tomamos aquellos individuos que no respondieron sobre el ingreso y por tanto era imposible compararlos contra el umbral de ingreso necesario para no ser pobres. La idea entonces, era poder predecir con sus otras características si son pobres o no. Por eso, usando el modelo entrenado con los individuos que si respondieron sobre ingreso, predecimos la probabilidad de ser pobre y la condición de pobreza de cada individuo. Según nuestro modelo logit, el porcentaje de pobres predicho en la muestra `norespondieron` es 33,48 %.

Detalle técnico: Tuvimos un problema de dimensionalidad de la matriz de predictores entre la base de los que respondieron y los que no respondieron. Esto se debió a que, inicialmente, trabajamos con la base `respondieron`, en la que creamos las variables dummies antes de correr los modelos. No obstante, notamos que, para algunas variables, hay observaciones que en la base `norespondieron` toman ciertos valores que no toma ninguna observación en la base `norespondieron`, y viceversa. Dado que la función "get dummiesçrea una variable dummy por cada valor de cada columna, esta diferencia en los valores que toman algunas variables en las dos bases genera que el número de variables independientes que se usan para entrenar el modelo en la base `respondieron` sea diferente al número de variables independientes disponibles para predecir la pobreza en la base `norespondieron`. Para solucionar ese problema de dimensionalidad y poder predecir fuera de la muestra, creamos las variables dummies faltantes en cada muestra y les colocamos un valor de cero. Además, modificamos el orden de las variables de ambos dataframes para que sean compatibles y podamos correr el comando sin ningún error.

Ejercicio 6

En los modelos previos incluimos como predictores casi todas las variables disponibles. Esto no es necesariamente lo mejor, ya que podríamos estar incluyendo muchas variables irrelevantes. Por lo tanto, en este ejercicio estimamos un modelo logit solo incluyendo las variables que creemos más relevantes. La mayoría de ellas son características propias del individuo y no de la vivienda

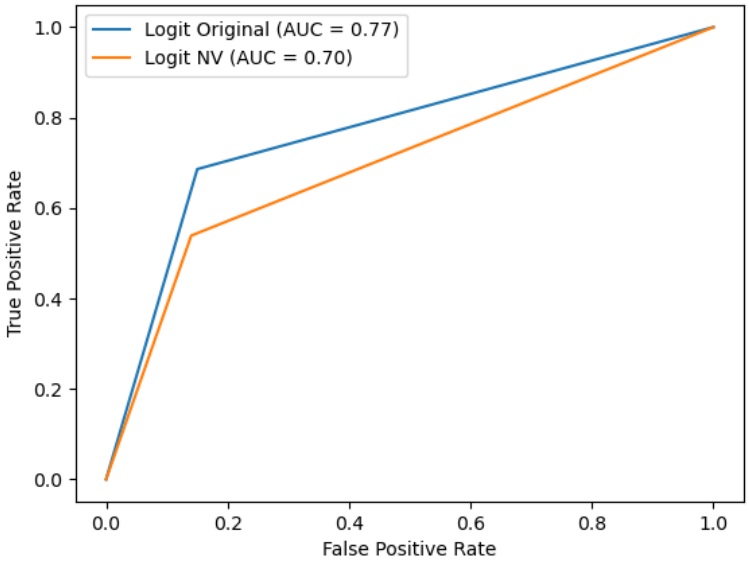


Figura 5: Curvas ROC: Modelos Logit

o de situaciones circunstanciales. Construimos un modelo con muchos menos regresores. Mirando un enfoque de capital humano, tomamos sexo, edad, si sabe leer y escribir, educación pública o privada, nivel educativo, y además, variables que siempre son relevantes como el estado laboral, y que categoría de inactivo es. Como vemos en la figura 5, los resultados para este nuevo modelo son bastante similares en Accuracy (0.745) y AUC (0.7) respecto al modelo original, si bien este último continúa teniendo un desempeño mejor.