Universidad de
SanAndrés

Big Data

Trabajo Práctico Grupal

Samuel Arispe, Gonzalo Ochoa y Facundo Valle Quintana

23 de octubre de 2023

Trabajo Práctico 2

Big Data

Parte I: Analizando la base

La Encuesta Permanente de Hogares (EPH) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las características sociodemográficas y socioeconómicas de la población. Uno de los indicadores más valiosos que pueden obtenerse con los datos extraídos de esta encuesta es la tasa de pobreza.

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

   La identificación de personas pobres se basa en un enfoque multidimensional (a lo Amartya Sen) que considera varios indicadores socioeconómicos:

   a) **Línea de Pobreza**: El INDEC establece una "línea de pobreza"que representa el umbral de ingresos necesarios para cubrir una canasta básica de bienes y servicios que satisface las necesidades básicas de una persona o una familia. Esta línea de pobreza es ajustada por inflación.

   b) **Ingreso Familiar**: Los ingresos considerados incluyen salarios, pensiones, subsidios, y otros recursos económicos. Se suma el total de los ingresos de los habitantes del hogar, para obtener el ingreso de un hogar.

   c) **Comparación con la Línea de Pobreza**: Se compara el ingreso total del hogar con la línea de pobreza. Si el ingreso del hogar está por debajo de esta línea, se considera que el hogar está en situación de pobreza.

   d) **Identificación de Personas Pobres**: Todos los miembros del hogar se consideran en situación de pobreza si el hogar en su conjunto se encuentra por debajo de la línea de pobreza. Esto se debe a que se asume que los recursos económicos se comparten entre los miembros del hogar.

2. **Entren a la página del INDEC 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.**

   (a) **Eliminen todas las observaciones que no corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires.**

   Mediante un operador booleano (*.isin*), filtramos la base de datos por aglomerado, quedandonos únicamente con los datos pertenecientes a los aglomerados 32 y 33, que corresponden a la CABA y partidos del GBA respectivamente.

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

   Procedemos a descartar las observaciones de ingreso y edades que no tengan sentido. Identificamos gracias al diccionario de variables, a las variables posiblemente problemáticas: $'CH06'$ (edad en años cumplidos), $'PP08D1'$ (Monto total de sueldos / jornales,salario familiar, horas extras, otras bonificaciones habituales y tickets, (monto en pesos cobrado por comisión por venta / producción), $'PP08F2'$ (monto en pesos cobrado por propinas), $'PP08J1'$ (monto aguinaldo), $'PP08J2'$ (monto otras bonificaciones no habituales), $'PP08J3'$ (monto retroactivo), $'IPCF'$ (Monto del ingreso per cápita familiar), $'ITF'$ (Monto de ingreso total familiar).

(c) **Una vez hecha esa limpieza, realicen un gráfico de barras mostrando la composición por sexo.**

Realizada la limpieza, obtenemos el gráfico de barras:

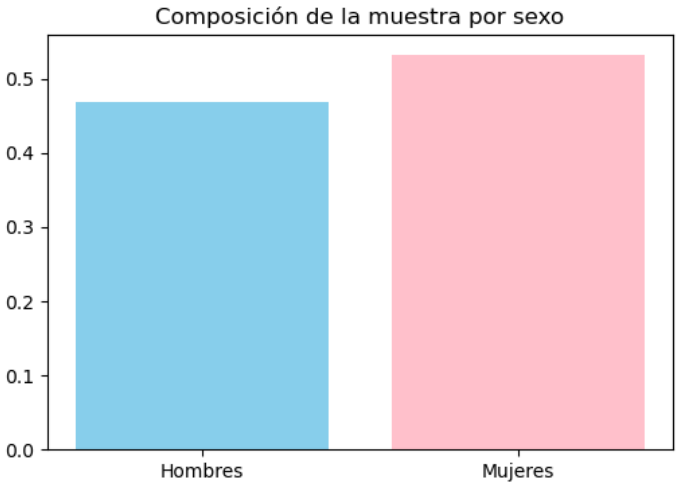


Figura 1: Composición de la muestra por sexo

Podemos apreciar que la muestra está en su mayoría compuesto por mujeres (53.2 %).

(d) **Realicen una matriz de correlación con las siguientes variables:** $CH04$, $CH07$, $CH08$, $NIVEL\_ED$, $ESTADO$, CAT_INAC, $IPCF$. **Comenten los resultados. Utilicen alguno de los comandos disponibles en este link o este link para graficar la matriz de correlación.**

Inspirándonos en los links proveídos en la consigna, realizamos la matriz de correlación.

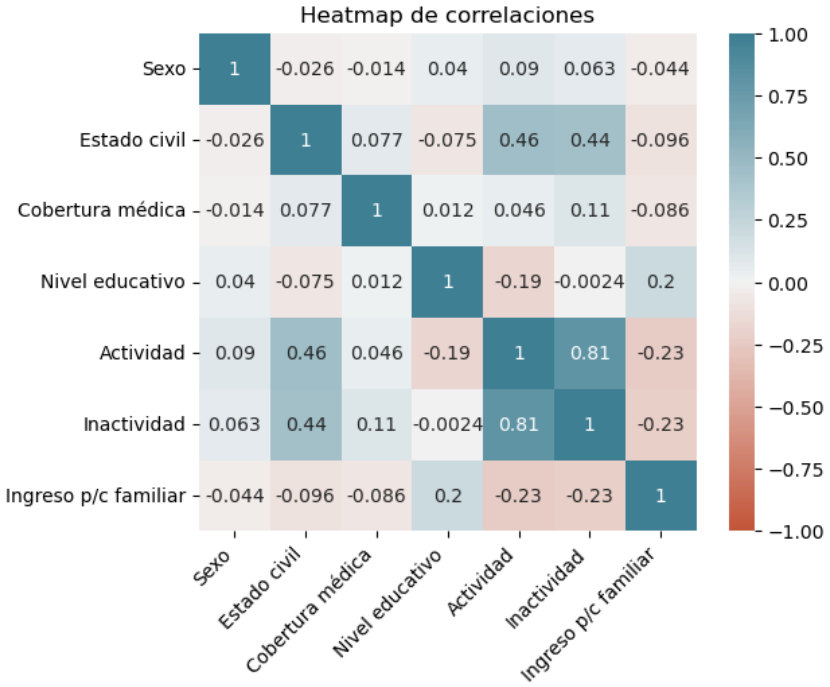


Figura 2: Matriz de correlaciones

[COMENTAR]

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

Usando las fórmulas de $value\_count()[k]$, especificando el k correspondiente a cada característica a contar, vemos que en nuestra muestra tenemos a unas 264 personas clasificadas como desocupadas y unas 2540 clasificadas como inactivas.

Respecto al ingreso per cápita familiar, obtenemos \$94025,92 para los clasificados como .ºcupado", \$27664,02 para los "Desocupado", \$44753,11 para los Ïnactivoz \$33759 para los "Menor de 10 años".

(f) **Utilizando el archivo tabla adulto** $equiv.xlsx$**, agreguen a su base de datos una columna llamada** $adulto_equiv$ **que contenga los valores de adulto equivalente de cada persona según su sexo y edad (por ejemplo, a un varón de 2 años le corresponde 0.46). Finalmente, con el comando** $groupby$ **sumen esta nueva columna para las personas que pertenecen a un mismo hogar y guarden ese dato en una columna llamada** ad_equiv_hogar1**.**[1]

Para asignarle a cada observación su métrica de adulto equivalente, decidimos crear una función llamada $calcular_adulto_equivalente$ que toma dos argumentos: el sexo y la edad. Mediante un "if"filtra primero por sexo y luego por edad, para devolvernos el valor de adulto equivalente correspondiente. Luego, mediante un $.apply(lambdarow :)$ se lo aplicamos a cada fila, especificando a la función qué columnas de cada fila tomar para el primer y segundo argumento.

Por último, creamos una nueva columna para el dataframe, usando $grouby()$ siguiendo la consigna, donde agrupamos por hogar: para aquello usamos a las dos variables que nos permiten identificar cada hogar: '$CODUSU$' y 'NRO_HOGAR' y le aplicamos una suma $.transform('sum')$; obteniendo la métrica de adulto equivalente para cada hogar.

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). ¿Cúantas personas no respondieron cúal 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 gúardenlas en una base bajo el nombre** $norespondieron$**.**

   En nuestra muestra, mediante un '$.value\_count()[k]$' obtenemos la cantidad de personas que no contestaron: 1786.

4. **Sabiendo que la Canasta Básica Total para un adulto equivalente en el Gran Buenos Aires en el primer trimestre de 2023 es aproximadamente \$57.371, 05, agreguen a la base respondieron una columna llamada ingreso necesario que sea el producto de este valor por** ad_equiv_hogar**. Note que este es el valor mínimo que necesita ese hogar para no ser pobre.**

   Nuevamente, para obtener el valor de la Canasta Básica Total ajustada por hogar creamos una función llamada CBT; esta simplemente toma como argumento el valor del ad_equiv_hogar previamente calculado y lo multiplica por el valor de la Canasta Básica Total (\$57371.05). Finalmente, aplicamos esta función a cada fila y creamos una nueva columna con $.apply(lambdarow :)$.

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. ¿Cúantos pobres identificaron?**

   Para identificar a los hogares pobres, comparamos los valores reportados del ingreso total familiar (ITF) con el valor de la canasta básica total ajustado por hogar necesario (del inciso anterior). Una vez más, realizamos esto a través de una función que creamos llamada *polenta*, que toma dos argumentos: el ITF y $ingreso\_necesario$ previamente calculado.

   Volvemos a aplicarla a cada fila, y encontramos que en la muestra hay 1536 personas clasificadas como pobres, lo que representa un 0.37 % de nuestra muestra; es decir 1536 personas no alcanzan a cubrir los gastos de la CBT ajustada a su hogar con su ingreso total familiar declarado.

[1] Por ejemplo, si una familia está compuesta por un varón de 40 años ($adulto\_equiv = 1$) y su esposa de la misma edad ($adulto\_equiv =$ 0,77) con sus mellizos varones de 5 años ($adulto\_equiv =$ 0,60 cada uno), a todos se les deberá imputar en ad_equiv_hogar un valor igual a 2.97, que es la cantidad de adultos equivalentes en ese hogar.

Parte II: Clasificación

El objetivo de esta parte del trabajo es intentar predecir si una persona es o no pobre utilizando datos distintos al ingreso, dado que muchos hogares son reacios a responder cúanto ganan.

1. **Eliminen de ambas bases todas las variables relacionadas a ingresos (en el archivo codigos *eph.pdf* ver las categorías: ingresos de la ocupación principal de los asalariados, ingresos de la ocupación principal, ingresos de otras ocupaciones, ingreso total individual, ingresos no laborales, ingreso total familiar, ingreso per cápita familiar). Elimine también las columnas *adulto_equiv*, *ad_equiv_hogar* e ingreso necesario.**

   De acuerdo a la consigna, eliminamos las siguientes variables: "$PP08D1$", "$PP08D4$", "$PP08F1$", "$PP08F2$", "$PP08J1$", "$PP08J2$", "$PP08J3$", "TOT_P12", "$P47T$", "$DECINDR$", "$IDECINDR$", "$RDECINDR$", "$GDECINDR$", "$PDECINDR$", "$ADECINDR$", "$PONDII$", "$V2\_M$", "$V3\_M$", "$V4\_M$", "$V5\_M$", "$V8\_M$", "$V9\_M$", "$V10\_M$", "$V11\_M$", "$V12\_M$", "$V18\_M$", "$V19\_AM$", "$V21\_M$", "T_VI", "ITF", "$DECIFR$", "$IDECIFR$", "$RDECIFR$", "$GDECIFR$", "$PDECIFR$", "$ADECIFR$", "$IPCF$", "$DECCFR$", "$IDECCFR$", "$RDECCFR$", "$GDECCFR$", "$PDECCFR$", "$ADECCFR$", "$PONDIH$", "$adulto_equiv$", "ad_equiv_hogar", "$ingreso\_necesario$".

   Además, decidimos dropear "$CODUSU$", porque es única a cada observación, "MAS_500", porque es de caracteres, aunque podríamos convertirla a dummy. Además, dropeamos a "$CH05$"porque ya hay una variable numérica con las edades y a "$PP09A\_ESP$"porque también es de caracteres.

2. **Partan la base respondieron en una base de prueba (test) y una de entrenamiento (train) utilizando el comando *train_test_split*. La base de entrenamiento debe comprender el 70 % de los datos, y la semilla a utilizar (*random_state_instance*) debe ser 201. Establezca a pobre como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables serán las variables independientes (matriz X). Recuerden agregar la columna de unos (1).**

   Luego de haber limpiado la base de datos de las variables anteriores, las clasificamos en el vector X o Y, siguiendo las consignas, y agregamos una columna de 1s, que luego colocamos al principio, por conveniencia.

3. **Implementen los siguientes métodos reportando luego la matriz de confusión, la curva ROC y los valores de AUC y de Accuracy de cada uno:**

   - **logit**
   - **Análisis de discriminante lineal**
   - **KNN con k=3**

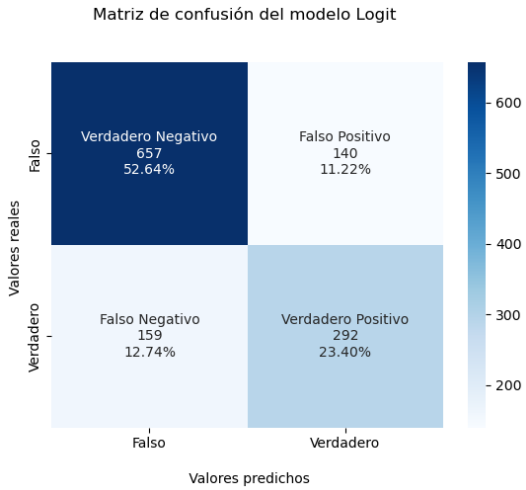
4. **¿Cúal de los tres métodos predice mejor? Justifiquen detalladamente utilizando las medidas de precisión que conocen.**

   Al estimar con los tres métodos, obtenemos lo siguientes resultados:

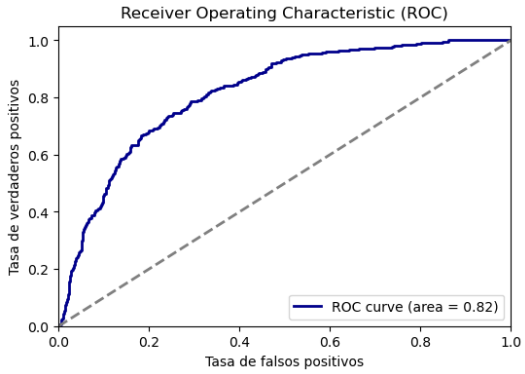
Modelo	AUC	Accuracy
Logit	0.736	0.760
LDA	0.744	0.760
KNN	0.697	0.720

Tabla 1: Resultados de las estimaciones con los tres métodos

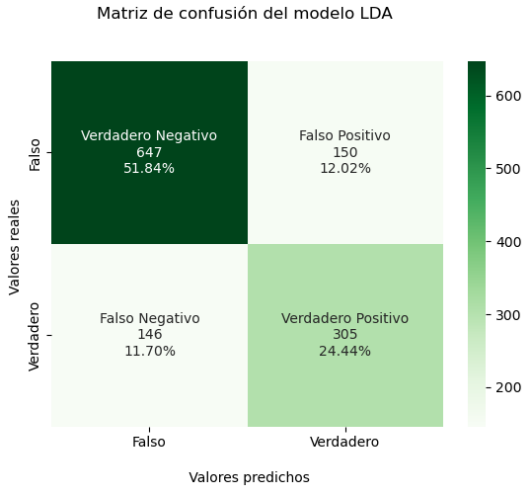
   Al analizar las matrices de confusión y curvas ROC y otras métricas (AUC y Accuracy), podemos concluir el modelo LDA es el mejor respecto a su capacidad predictiva.



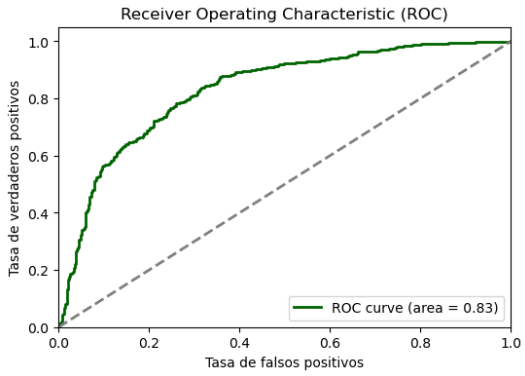
(a) Matriz de confusión del modelo Logit



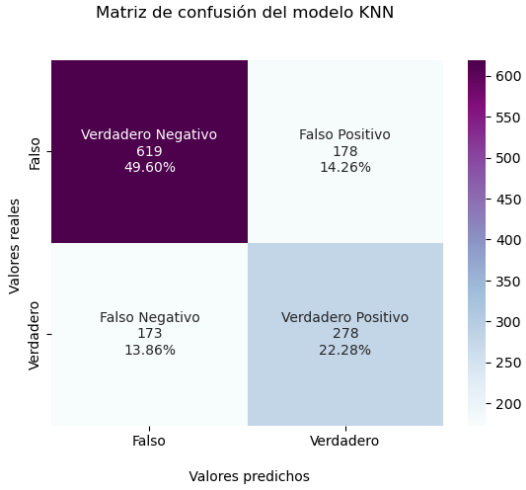
(b) Curva ROC del modelo Logit



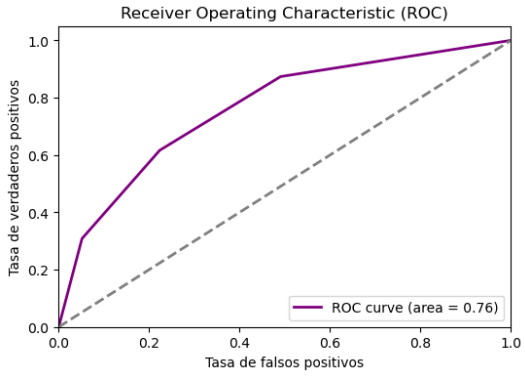
(c) Matriz de confusión del modelo LDA



(d) Curva ROC del modelo LDA



(e) Matriz de confusión del modelo KNN



(f) Curva ROC del modelo KNN

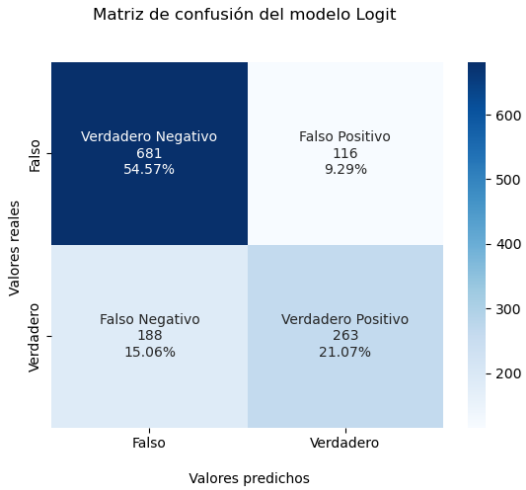
Figura 3: Comparaciones de las métricas de los diferentes modelos

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?**

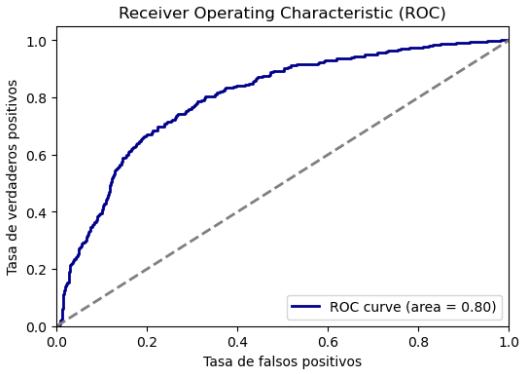
   Para este ejercicio, utilizaremos el modelo LDA para hacer predicciones basadas en la base de datos *norespondieron*, ya que ha demostrado tener una mayor precisión (Accuracy) y un mejor valor de la métrica AUC. Según nuestras predicciones, aproximadamente 943 de 1786 personas de la muestra que no respondieron se encuentran en situación de pobreza, lo que equivale a una tasa de pobreza del 52.8 %

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?**

   En la pregunta 1 de esta parte faltó eliminar también las variables de ingreso independiente, así que procedemos a eliminar estas variables también. Por otro lado, nos parece que no es correcto corregir el problema de los missing values reemplazándolos con cero ya que hay variables en las cuales el 0 significa algo. Así que para esta última pregunta lo que haremos es eliminar todas las variables que posean missings values en vez de imputar. De esta forma nos quedan 36 variables, entre las cuales tenemos el nivel educativo, edad, sexo, estado civil, entre otras variables que consideramos que se correlacionan con el nivel de pobreza. Como podemos ver en la figura 4 el ROC (de 0.82 a 0.80), AUC (de 0.736 a 0.719) y Accuracy (de 0.760 a 0.756) disiminuyen levemente, aunque aumenta la proporción de verdaderos negativos. Que disminuyan levemente las otras métricas puede ser porque todas las demás variables consideradas en los ejercicios anteriores y que no se consideran en esta última pregunta no aportan mucho a la predicción. Esto lo podríamos notar aplicando lasso.



(a) Matriz de confusión del segundo modelo Logit



(b) Curva ROC del segundo modelo Logit

Figura 4: Comparaciones de las métricas de nuestro segundo modelo Logit