Universidad de San Andrés

Universidad de
SanAndrés

Big Data

Trabajo Práctico N2

Profesor: Noelia Romero

Tutor: Victoria Oubiña

Alumnos: Solana Cucher, Victoria Rosino, Florencia Ruiz.

Parte 1: Analizando la base

Punto 1

Para identificar a las personas y hogares pobres, el INDEC utiliza como metodología la línea de pobreza (LP) y la línea de indigencia (LI). Para ello, en primer lugar, se define una canasta básica de alimentos (CBA) y una canasta básica total (CBT), y sus respectivos valores. La primera incluye el conjunto de alimentos que satisfacen requerimientos nutricionales esenciales y que son representativos del patrón de consumo de alimentos de la población, mientras que la CBT amplía esta definición para incluir bienes y servicios no alimentarios (vestimenta, transporte, educación, salud, etc.). Luego, a partir de los datos que se obtienen en la Encuesta Permanente de Hogares (EPH), se comparan los ingresos de las personas u hogares con los valores de estas canastas. Es decir, se establece que si los ingresos no son suficientes para cubrir el umbral de la CBA (no superan la Línea de Indigencia), los hogares se consideran indigentes; mientras que para calcular la incidencia de la pobreza se analiza la proporción de hogares cuyo ingreso no supera el valor de la CBT (no superan la Línea de Pobreza).

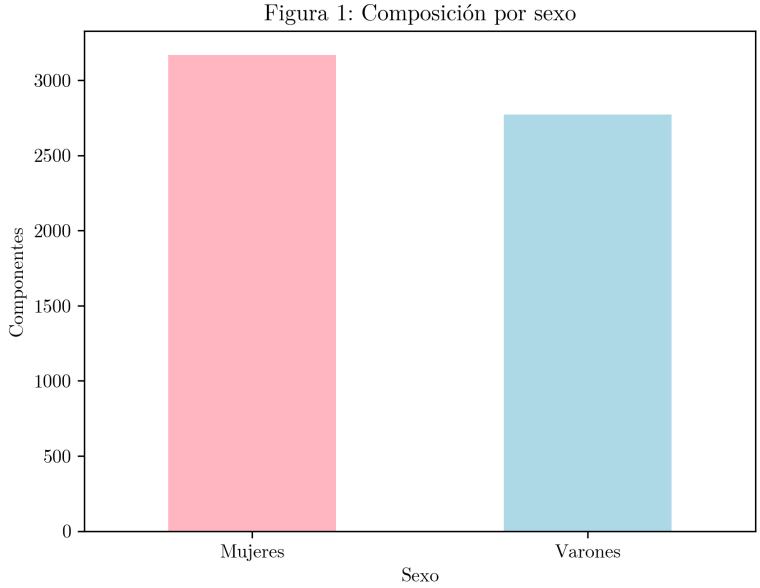
Punto 2

a) La Encuesta Permanente de Hogares (EPH) cuenta con la variable "Aglomerados", la que asigna un código de dos dígitos a cada individuo según la localidad en la que vive. En este ejercicio, eliminamos todos los códigos que no sean 32 (Ciudad Autónoma de Buenos Aires) ni 33 (Partidos del GBA).

b) Para saber qué variables poseen valores absurdos, realizamos un análisis descriptivo de estas que contiene información como el mínimo, máximo y promedio. Esto nos permite ver, por ejemplo, si alguna variable tiene un valor negativo. A partir de este análisis, en primer lugar, imponemos como condición que las variables de interés como los ingresos y las edades no tomen valores negativos. En particular, exigimos que el monto de ingreso de la ocupación principal (P21) y el monto de ingreso total individual (P47T) sean siempre positivos o, al menos, nulos (ya que estas son las que presentaban valores menores a cero) y que la edad (CH06) sea también mayor o igual a cero años. Además, en las variables de estado civil (CH07) y cobertura médica (CH08) existen observaciones que toman un valor de 9, es decir, "no sabe/no contesta". Por lo tanto, decidimos eliminar estas observaciones ya que podrían generar ruido en nuestro análisis.

c) La Figura 1 muestra la composición por sexo de la base de datos. Como podemos ver, la cantidad de mujeres (3169) es levemente superior a la de hombres (2773). Sin embargo, la diferencia no parecería ser demasiado relevante por lo que la muestra estaría balanceada en términos del sexo de los individuos.

Figura 1: Composición por sexo

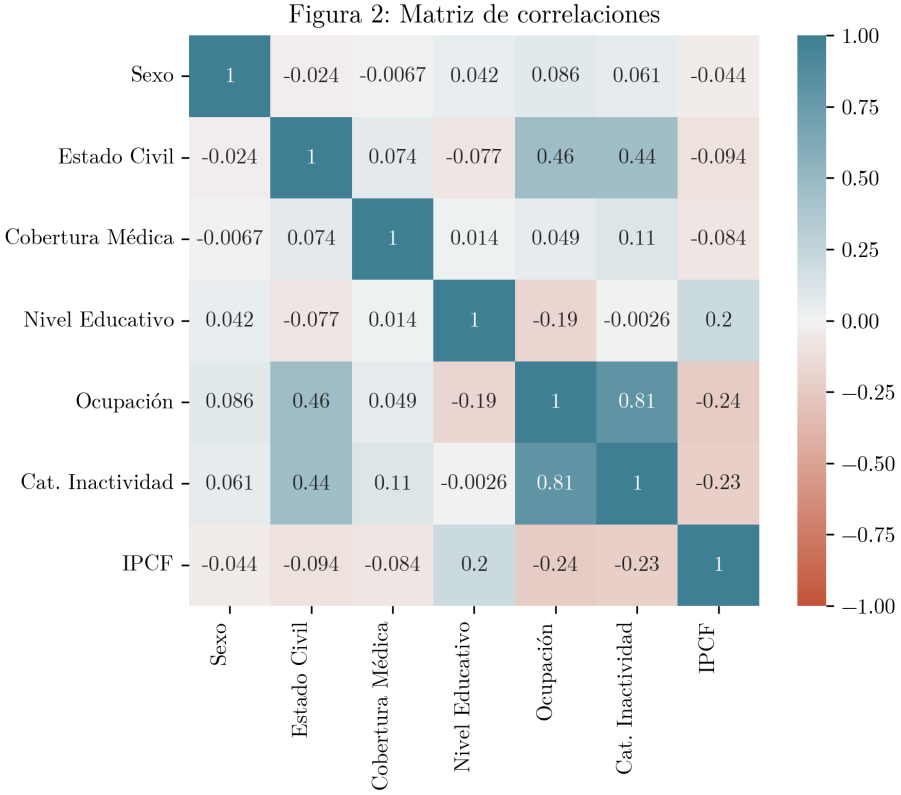


d) La Figura 2 muestra las correlaciones entre el sexo (CH04), el estado civil (CH07), la cobertura médica (CH08), el nivel educativo (NIVEL_ED), la condición de actividad (ESTADO), la categoría de inactividad (CAT_INAC) y el monto de ingreso per cápita familiar (IPCF). Podemos destacar la correlación negativa tanto del estado de actividad, como de la categoría de inactividad, con ingreso per cápita familiar. En el primer caso, el coeficiente de correlación equivale a -0,24, mientras que el segundo es -0,23. Esto demuestra que los individuos con valores de CAT_INAC mayores, como los menores de 6 años o discapacitados, tienen un ingreso per cápita menor. Así mismo, los individuos con un valor mayor de ESTADO, como los inactivos, desocupados o menores de 10 años, también tienen menores ingresos. También se destaca la relación negativa entre la educación alcanzada y el estado de actividad (coeficiente de correlación igual a -0,19). Es decir, los inactivos, desocupados o menores de 10 años se asocian con un menor nivel educativo.

Entre las correlaciones positivas, podemos observar que a mayor nivel educativo los individuos presentan un ingreso per cápita familiar más alto. Sin embargo, el coeficiente de correlación (de 0,2) no es tan alto como podría esperarse. Una relación más fuerte se presenta entre la actividad y el estado civil (coeficiente de correlación igual a 0,46),

así como también entre la categoría de inactividad y el estado civil (coeficiente 0,44). De esta manera, en nuestros datos, las personas separadas/divorciadas, viudas y solteras suelen ser discapacitadas, inactivas, desocupadas o menores de 10 años. A la vez, las variables de categoría de actividad y estado de actividad también muestran una alta correlación (coeficiente igual a 0,81), por lo que deducimos que las personas desocupadas o inactivas (mayores valores de la variable ESTADO) suelen ser, por ejemplo, menores de 6 años o discapacitadas (mayores valores de la variable CAT_INAC).

Figura 2: Matriz de correlaciones



e) En la EPH del Primer Trimestre de 2023, en la Ciudad Autónoma de Buenos Aires y en el Gran Buenos Aires, se registran 264 personas desocupadas y 2529 personas bajo la categoría de inactivos. Podemos profundizar más sobre estos individuos al calcular su ingreso per cápita familiar promedio. Al hacer este ejercicio, observamos que los inactivos tienen mayores ingresos promedio que los desocupados. La primera de estas categorías alcanza un monto de $44.797 aproximadamente mientras que los desocupados poseen un IPCF promedio de $27.664. Por su parte, como es de esperarse, las personas pertenecientes a la categoría de ocupados poseen un ingreso per cápita familiar promedio de $93.268, mayor al de las categorías de inactivos o desocupados.

Punto 3

A partir de los datos, encontramos que la cantidad de personas que no reportaron su ingreso total familiar en CABA y en el Gran Buenos Aires es 1769 individuos.

Punto 5

Sabemos que la Canasta Básica Total para un adulto equivalente en el Gran Buenos Aires, en la época en la que realizó la EPH utilizada (primer trimestre de 2023), era de aproximadamente $57.371,05. Si partimos de este dato para calcular el ingreso de subsistencia de cada familia de nuestra base de datos, obtenemos que 1555 individuos no alcanzan, con sus ingresos totales, este monto necesario. Es decir, 1555 personas de nuestra muestra son pobres.

Parte 2: Clasificación

Punto 2:

Para realizar la partición de la muestra entre entrenamiento y testeo, y posteriormente aplicar los métodos regresión logística, análisis discriminante lineal y vecinos cercanos; debemos eliminar determinadas variables. En particular, borramos de la base de datos todas aquellas variables cuyo formato no es numérico. Estas son el código de usuario ("CODUSU"), los aglomerados según su tamaño ("MAS_500") y la fecha de nacimiento ("CH05"). Analizamos también la presencia de *missing values* en las variables restantes. Observamos que existen dos extremos en cuanto a los valores faltantes: o las columnas no cuentan con ningún *missing value* o, en su lugar, tienen relativamente muchas respuestas faltantes (igual o superior a 2101 *missing values*). Por ello, eliminamos el segundo conjunto de variables porque nos hacen perder muchas observaciones en la estimación.

Punto 3:

Implementamos los métodos *logit*, análisis de discriminante lineal y vecinos cercanos (con 3 vecinos). En cada caso, computamos la matriz de confusión, el área debajo de la curva ROC y la curva ROC.

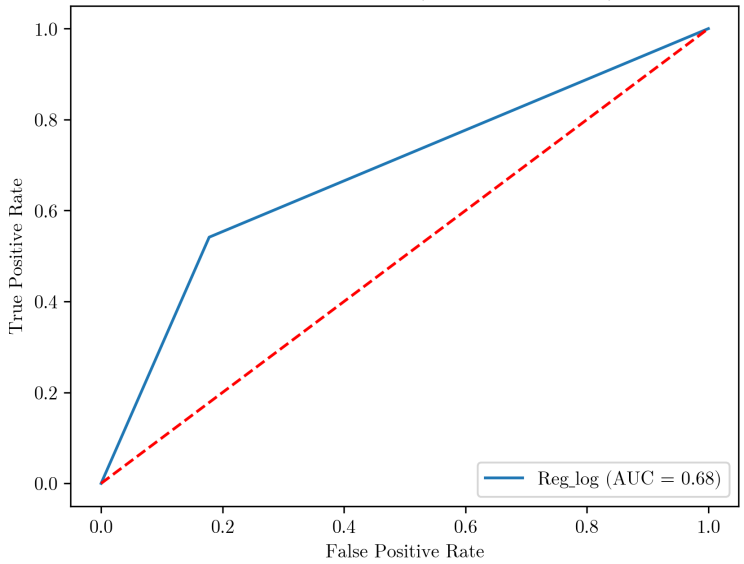
La matriz que corresponde al método de regresión logística es la siguiente:

$$\begin{bmatrix} 642 & 139 \\ 216 & 255 \end{bmatrix}$$

Asimismo, este método posee un *accuracy score* de 0,716, es decir la precisión alcanzada por este modelo es del 72% aproximadamente. Finalmente, la Figura 3 muestra la curva ROC. Esta nos permite ver cómo se modifica la tasa de verdaderos positivos y falsos positivos a

medida que se modifican los umbrales de predicción. El área acumulada debajo de ella es 0,682. Dado que el ideal de la curva ROC es la esquina superior izquierda arriba (donde no existen falsos positivos o falsos negativos), la Figura nos muestra que el método logit no tienen un gran performance en tanto se aleja de esta esquina y se acerca a la línea punteada (donde la tasa de verdaderos positivos es igual a la de falsos positivos). Sin embargo, como todavía estamos por encima de la diagonal, los verdaderos positivos son mayores a los falsos positivos, lo cual es deseable.

Figura 3: Curva ROC (Regresión logística)

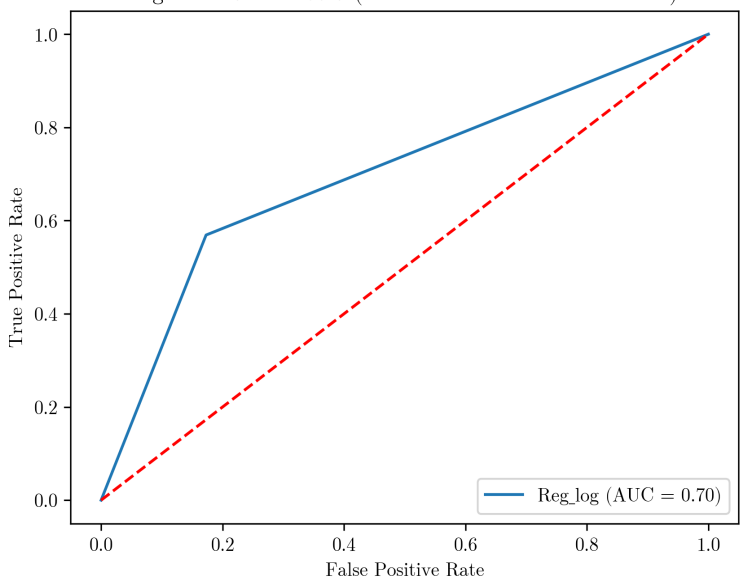


Por otro lado, a partir del análisis de discriminante lineal encontramos que la matriz de confusión está definida por:

$$\begin{bmatrix} 646 & 135 \\ 203 & 268 \end{bmatrix}$$

El *accuracy score*, es decir, la tasa de verdaderos positivos y negativos sobre el total de ellos, es del 73%. La curva ROC se muestra en la Figura 4 y el área debajo de ella es de 0,698. Nuevamente, la curva ROC se encuentra por encima de la diagonal lo que significa que la tasa de verdaderos positivos es mayor a la de falsos positivos, pero se encuentra alejada de la esquina superior izquierda. Es decir, la performance del método de análisis de discriminante lineal es medianamente buena.

Figura 4: Curva ROC (Análisis de discriminante lineal)

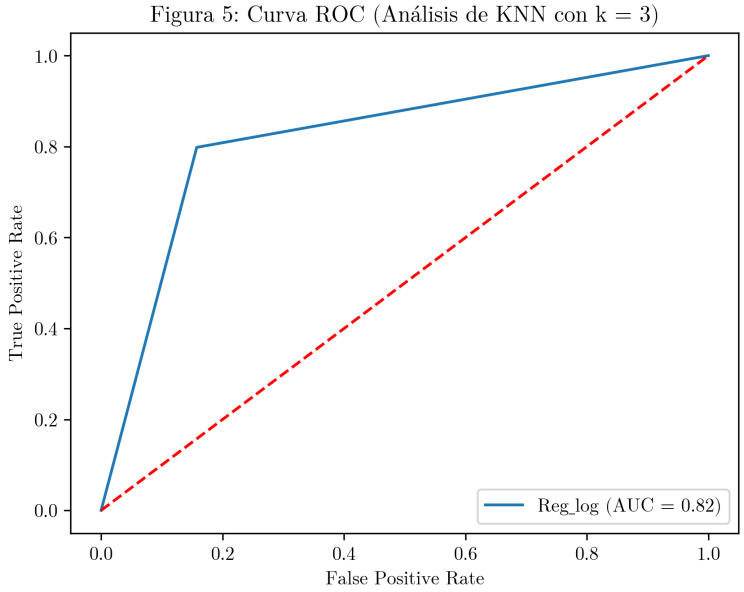


Por último, el método de vecinos cercanos (con $k = 3$) presenta una matriz de confusión igual a:

$$\begin{bmatrix} 658 & 123 \\ 95 & 376 \end{bmatrix}$$

En este caso, el *accuracy score* es de 0,826 y la curva ROC se ve representada en la Figura 5. En particular, el área acumulada por debajo de esta curva en la Figura 5 es de 0,82. La curva ROC del método de vecinos es la que muestra una peor performance (marginal) en tanto es la más cercana, en comparación a los otros dos métodos, a la diagonal (y, por lo tanto, más alejada de la línea superior izquierda).

Figura 5: Curva ROC (Análisis de KNN con k = 3)



Punto 4:

Los gráficos de la curva ROC a simple vista muestra que el método de vecinos cercanos se acerca más a la situación ideal. Además, podemos hacer un análisis más minucioso si miramos las métricas que se presentan en la Tabla 1. Por ejemplo, el *accuracy score* es mayor para vecinos cercanos en comparación con la regresión logística o el análisis de discriminante lineal. En particular, la tasa de verdaderos positivos y verdaderos negativos en relación con el total de estos es 82,60% para KNN. La sensibilidad, tasa de falsos positivos y precisión también son más altos con vecinos cercanos. En conclusión, si bien la tasa de especificidad y la de falsos negativos son mayores en la regresión logística, el análisis de la curva ROC y el resto de las métricas indica que el método de vecinos cercanos es el que mejor predice (seguido de la regresión logística).

Tabla 1: Medidas de precisión

	Regresión logística	ADL	KNN
Accuracy	0,7160	0,7300	**0,8260**
Sensitivity	0,5414	0,5990	**0,7983**
Specificity	**0,8220**	0,8118	0,7708
False Positive rate	0,1780	0,1882	**0,2292**
False Negative rate	**0,4586**	0,4310	0,2017
Precision	0,6472	0,6650	**0,7535**

Punto 5

Dado que en el punto anterior concluimos que el mejor método para predecir resulta ser el de vecinos cercanos (con $k = 3$), realizamos ahora una predicción de la cantidad de "pobres" en la muestra de aquellos que no reportaron ingresos totales familiares. Al hacer esta predicción, obtenemos que la cantidad de pobres que habría en la base de datos "no respondieron" sería de 1649 individuos. Esto implica que nuestro método predice una tasa de pobreza igual al 39,52% dentro de esta muestra.

Punto 6

Al analizar las variables que incluimos en las predicciones, consideramos que algunas de ellas podrían no ser útiles y tener un aporte nulo en la estimación. Teniendo en cuenta aquello

relevante para el concepto de pobreza, decidimos quedarnos con 14 variables. Las variables seleccionadas son relación de parentesco (CH03), sexo (CH04), edad (CH06), estado civil (CH07), cobertura médica (CH08), saber leer o escribir (CH09), asistencia a establecimiento educativo (CH10), nivel educativo más alto cursado (CH12), dónde vivía hace 5 años (CH16), nivel educativo (NIVEL_ED), condición de actividad (ESTADO), categoría de inactividad (CAT_INAC), búsqueda de trabajo en los últimos 12 meses (PP02H) y si trabajó en algún momento en los últimos 12 meses (PP02I).

Cuando volvemos a predecir usando el modelo logit, obtenemos las mismas medidas de precisión de antes. Es decir, el *accuracy score* es de 0,716 y el área acumulada debajo de la curva ROC es 0,682. Por lo tanto, como la precisión del método se mantiene inalterada, podemos afirmar que las variables eliminadas para esta nueva estimación no eran relevantes o, lo que es lo mismo, que por lo menos las 14 variables seleccionadas (o un subconjunto de ellas) son las que verdaderamente conducen a los resultados de este método.