Universidad de
SanAndrés

Big Data

Walter Sosa Escudero
Belén Michel Torino

Trabajo Práctico 2

Garcia Vassallo, Hedemann, Sury

2021

1. Parte 1

1)

Segun el Indec, la pobreza entendida como un fenómeno multidimensional es medida por dos métodos alternativos que dan cuenta de las distintas dimensiones que la caracterizan. Por una parte, el método directo, también conocido como Necesidades Básicas Insatisfechas (NBI), consiste en identificar el conjunto de hogares que no pueden satisfacer alguna necesidad básica. Por otra parte, el método indirecto, también llamado el "enfoque del ingreso". De acuerdo con este método, un hogar es considerado pobre si su ingreso resulta menor que la "línea de pobreza", un concepto que representa el valor de todos los bienes y servicios que se consideran necesarios para satisfacer las necesidades básicas. La línea de pobreza incluye no sólo los consumos alimentarios mínimos sino también otros consumos básicos no alimentarios. La suma de ambos conforma la Canasta Básica Total (CBT).

El enfoque del ingreso adopta diferentes criterios en términos de pobreza absoluta o relativa. El concepto de pobreza absoluta sostiene que existe un núcleo irreductible de privación absoluta. De esta forma, toda persona que no la satisfaga se considera persona en situación de pobreza. Mientras que el concepto de pobreza relativa postula que las necesidades humanas no son fijas, y varían de acuerdo a los cambios sociales y a la oferta de productos en un contexto social determinado. La pobreza relativa, por tanto, se establece en función del nivel general de ingresos en la comunidad, el país, así como la región que se analice

En conclusión, se puede identificar a las personas pobres como aquellas que no satisfaces sus necesidades básicas y/o como aquellas que están poder debajo de la linea de pobreza.

2)

a)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

b)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

c)

Realizamos un grafico de barras mostrando la composición por sexo:

Podemos observar que el porcentaje de mujeres sobre el total es mayor que el de hombres. Las mujeres llegan al casi 55 % del total de ambos sexos mientras que el procentaje de hombres llega a casi 45 %.

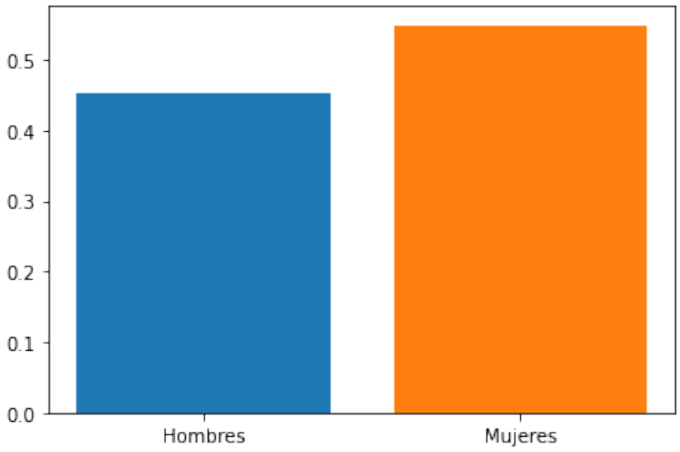

Figura 1: Composición por sexo

d)

En primer lugar realizamos una tabla donde podemos osbervar las correlaciones de las variables teniendo en cuenta que CH04 es el sexo, CH07 es el estado civil, CH08 la cobertura médica, NIVEL_ED el nivel educativo, ESTADO la condición laboral, CAT_INAC es la categoría de inactividad y que IPCF es el ingreso familiar per cápita.

	CH04	CH07	CH08	NIVEL_ED	ESTADO	CAT_INAC	IPCF
CH04	1.000000	-0.025401	-0.018404	0.034413	0.103613	0.092578	-0.043384
CH07	-0.025401	1.000000	0.093896	-0.108472	0.400843	0.376750	-0.145113
CH08	-0.018404	0.093896	1.000000	0.019925	0.016732	0.060632	-0.074722
NIVEL_ED	0.034413	-0.108472	0.019925	1.000000	-0.229064	-0.077083	0.249320
ESTADO	0.103613	0.400843	0.016732	-0.229064	1.000000	0.798244	-0.243821

Figura 2: Tabla de correlaciones

En segundo lugar, realizamos un heat map donde podemos observar la matriz de las correlaciones.

Entonces, podemos observar que la variable “Sexo” correlaciona negativamente con el estado civil, con la cobertura médica y con el ingreso familiar per cápita. Mientras que correlaciona positivamente con el nivel educativo, la condición laboral y la categoría de inactividad. Por otro lado, podemos observar que el estado civil correlaciona negativamente con el nivel educativo y con el ingreso familiar per cápita. Mientras que correlaciona positivamente con la cobertura médica, la condición laboral y con la cateogira de inactividad. Aparte, al observar la variable çobertura médica"podemos observar que correlaciona negativamente con la condición laboral y la categoría e inactividad mientras que positivamente con el ingreso familiar per cápita. Luego, observando el nivel educativo podemos ver que correlaciona negativamente con la condición laboral y con la categoria de inactividad mientras que correlaciona positivamente con el ingreso familiar per capita. Por ultimo, observando la condicion laboral podemos ver que correlaciona positivamente con la categoria de inactividad pero negativamente con el ingreso familiar per capital.

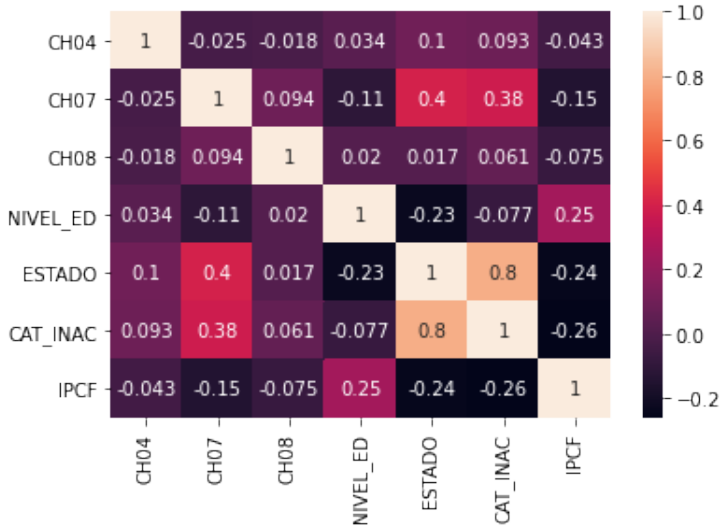

Figura 3: Matriz de correlaciones

e)

En la muestra hay 204 desocupados mientras que la cantidad de inactivos es 1506.
En la siguiente tabla podemos osbervar la media del ingreso per cápita familiar (IPCF) según la condicion de actividad.

Estado	IPCF
Ocupado	33810.842375
Desocupado	12446.200980
Inactivo	18695.858625

Figura 4: Media IPCF

Podemos observar que entre los ocupados, la media del ingreso per cápita familiar es de 33810.842375. Luego, para los desocupados la la media del IPCF es de 12446.200980. Por último, la media del IPCF es 18695.858625 cuando el estado de la condición de actividad es inactivo.

f)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

3)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

4)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

5)

Observamos que la cantidad de personas que tienen un ingreso total familiar menor al ingreso necesario de $18.914 es 838. Es decir, identificamos 838 personas pobres.

2. Parte 2

1)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

2)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

3)

La resolución de este punto se encuentra disponible en el Jupyter Notebook.

4)

A continuación reportamos la curva ROC y los valores de AUC y de Accurac para los métodos Logit, Análisis discriminante y KNN.

LOGIT

El modelo logit es un modelo de regresion logistica donde la regresión logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas.
En primer lugar observamos la matriz de confusion que se arma a partir de los eventos predichos y los eventos ocurridos. Que ocurra o no un hecho esta bajo el control de la naturaleza.

$$\begin{bmatrix} 439 & 69 \\ 101 & 141 \end{bmatrix} \tag{1}$$

De este modo, la matriz posee 439 verdaderos negativos, 69 falsos negativos, 101 falsos positivos y 141 verdaderos positivos.

Accuracy Score : 0.7733333333333333
AUC: 0.72

Con "Accuracy Score" podemos medir la calidad del modelo de machine learning en tareas de clasificación. Se calcula como el número de todas las predicciones correctas dividido por el número total del conjunto de datos. Esto es, el modelo sólo es capaz de identificar un 77.3 % de las

predicciones sin errores, es decir, el modelo acerto en un 77.3 %.

Por otro lado, AUC significa "Area Under the Curve". Esta medida transforma la curva ROC en una representación numérica del rendimiento para un clasificador binario. AUC es el área bajo la curva ROC y toma un valor entre 0 y 1. AUC indica el éxito de un modelo en la separación de clases positivas y negativas. Cuando AUC = 1, entonces el clasificador puede distinguir perfectamente entre todos los puntos de clase positivos y negativos correctamente. Sin embargo, si el AUC hubiera sido 0, entonces el clasificador estaría prediciendo todos los negativos como positivos y todos los positivos como negativos.Entonces, podemos decir que el modelo Cuando 0.5 <AUC <1, existe una alta probabilidad de que el clasificador pueda distinguir los valores de clase positivos de los valores de clase negativos. Esto es así porque el clasificador puede detectar más números de verdaderos positivos y verdaderos negativos que falsos negativos y falsos positivos.Entonces, como nuestro clasificador esta por encima de 0.5 podemos decir que distingue bien entre los positivos y negativos.

Entonces, a partir de la matriz de confusión podemos observar que hay dos tipos de errores, falso positivo y falso negativo. Como no es posible eliminar los dos errores lo que se puede hacer es reducir un error pero esta reducción implica un trade off dado que aumenta el otro error. Este trade off es lo que se puede observar en la curva ROC. Permitiendo asi medir la capacidad predictiva de un esquema de clasificación y comparar esquemas de clasificación.

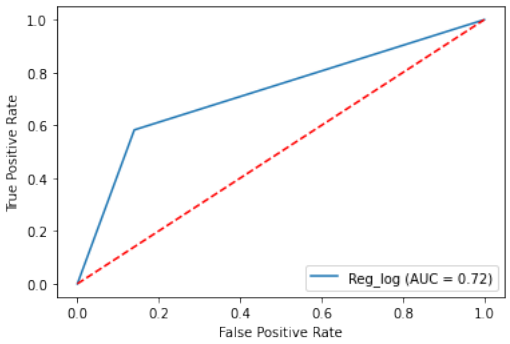

Figura 5: Curva ROC

ANÁLISIS DISCRIMINANTE LINEAL

El Análisis Discriminante Lineal o Linear Discrimiant Analysis (LDA) es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características.Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, Pr(Y=k|X=x).

Matriz de Confusión

$$\begin{bmatrix} 438 & 70 \\ 101 & 141 \end{bmatrix} \tag{2}$$

De este modo, la matriz posee 438 verdaderos negativos, 70 falsos negativos, 101 falsos positivos y 141 verdaderos positivos.

Accuracy Score : 0.772

Esto implica que el modelo sólo es capaz de identificar un 77.2 % de las predicciones sin errores, es decir, el modelo acerto en un 77.2 %.

AUC: 0.72

Como nuestro clasificador esta por encima de 0.5 podemos decir que distingue bien entre los positivos y negativos, es decir, el clasificador puede detectar más números de verdaderos positivos y verdaderos negativos que falsos negativos y falsos positivos.

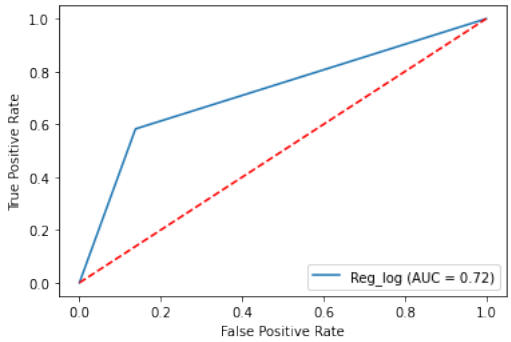

Figura 6: Curva ROC

KNN con k=3

El clasificador de K-vecinos más cercanos (KNN). Es un estimador donde dado un entero K positivo y una observación de prueba X_0, el clasificador KNN primero identifica los puntos K en los datos de entrenamiento que están más cerca de X_0, representados por N_0. Luego estima la probabilidad condicional para la clase j como la fracción de puntos en N_0 cuyos valores de respuesta son iguales a j:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

Matriz de Confusión

$$\begin{bmatrix} 406 & 102 \\ 97 & 145 \end{bmatrix} \tag{3}$$

De este modo, la matriz posee 406 verdaderos negativos, 102 falsos negativos, 97 falsos positivos y 145 verdaderos positivos.

Accuracy Score : 0.7346666666666667

Esto implica que el modelo sólo es capaz de identificar un 73.4 % de las predicciones sin errores, es decir, el modelo acerto en un 73.4 %.

AUC: 0.70

Como nuestro clasificador esta por encima de 0.5 podemos decir que distingue bien entre los positivos y negativos, es decir, el clasificador puede detectar más números de verdaderos positivos y verdaderos negativos que falsos negativos y falsos positivos.

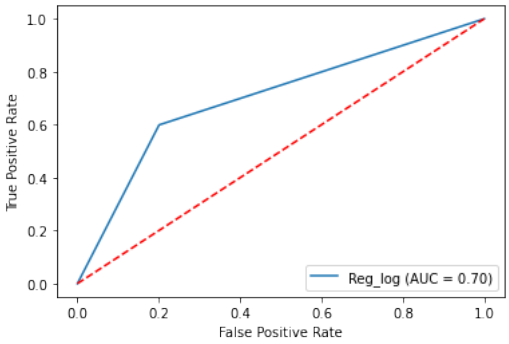

Figura 7: Curva ROC

5)

El método Logit es el que mejor predice. Esto es así ya que posee el mayor accuracy score de los tres métodos. Es decir, posee la mejor capacidad predictiva por fuera de la muestra. Puede observarse que el accuracy score de Logit es de 0.7733333333333333 mientras que los métodos de análisis discriminante lineal y KNN poseen un accuracy score de 0.772 y 0.7346666666666667, respectivamente. Las otras medidas de precisión dan iguales con los tres métodos por lo que definimos cual es el mejor a partir del accuracy score.

6)

Dentro de la base norespondieron pudimos identificar 335 personas pobres. Esto, el 40 % de las personas que no respondieron son pobres.

7)

No nos parece del todo correcto utilizar todas las variables disponibles como predictores. Esto es ya que al hacerlo puede generarse un problema de overfitting lo cual dificultaría la precisión y el rendimiento del modelo. Esto perjudica la capacidad predictiva del modelo ya que hay un trade off entre ajuste dentro de la muestra y predicción por fuera de la muestra. En otras palabras, el modelo aprende el detalle y ruido en la muestra de entrenamiento al punto tal que impacta negativamente la performance del modelo en una muestra nada. En este sentido fuimos analizando las variables de la base y armamos una nueva base omitiendo aquellas variables que consideramos que nos nos aportaban información nueva. El listado de las variables seleccionadas se encuentra en el Jupyter Notebook. Al correr el modelo logit con esta nueva base obtenemos que mejoramos la predicción en comparación al hacerlo con la base completa. El accuracy score pasó de 0.773333333 a 0.776 y el AUC score paso de 0.72 a 0.73 por lo que mejoró la presición. Por último, se predice un 28 % de pobres en la muestra de los que no respondieron mientras que al usar la base completa la proporción era del 40 %.