Trabajo Práctico 2: Big Data

Fermín Rodriguez Spialtini Valentin Joaquín Liwski

Septiembre 2022

Parte I: Analizando la base

Ejercicio 1:

Las nociones de pobreza e indigencia empleadas por el INDEC para el cálculo de incidencia se corresponden con el método de medición indirecta, denominado también "línea". El concepto de "Línea de Pobreza" (LP) procura establecer si los hogares cuentan con ingresos suficientes para cubrir una canasta de alimentos capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas (Canasta Básica Alimentaria, CBA) y otros consumos básicos no alimentarios. La suma de ambos conforma la Canasta Básica Total (CBT), la cual es contrastada con los ingresos de los hogares relevados por la Encuesta Permanente de Hogares (EPH). Para calcular la línea de pobreza es necesario contar con el valor de la CBA y ampliarlo con la inclusión de bienes y servicios no alimentarios (vestimenta, transporte, educación, salud, etcétera) con el fin de obtener el valor de la Canasta Básica Total (CBT). Dado que las necesidades nutricionales difieren entre la población, se construye una unidad de referencia, el "adulto equivalente", correspondiente a un hombre adulto de actividad moderada, para establecer luego las relaciones en las necesidades energéticas según edad y sexo de las personas. A partir de estas equivalencias se construyen las líneas para cada hogar según su tamaño y composición. En tanto que las líneas se construyen por hogar, el valor de las canastas que estas suponen debe ser contrastado con el ingreso total familiar del hogar, lo que permite clasificarlos en hogares indigentes (ingresoso totales debajo de la CBA), pobres no indigentes (ingreso debajo de LP pero sobre LI), pobres (incluye las dos anteriores) y no pobres, extendiéndose esa caracterización a cada una de las personas que los integran.

Ejercicio 2:

a

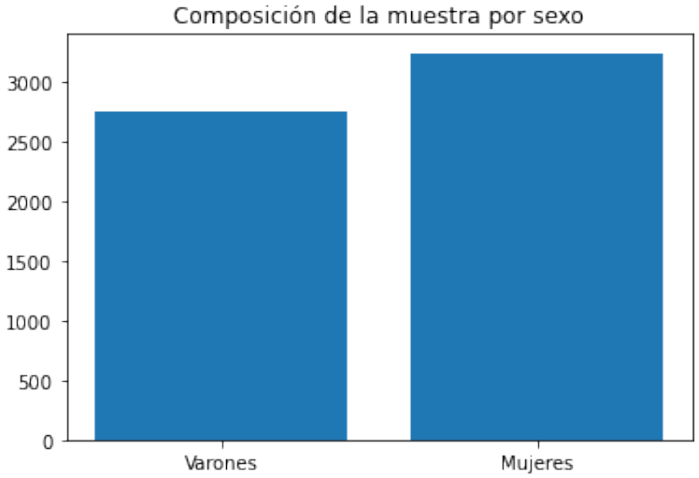
Eliminamos todos los aglomerados que no corresponden a CABA (AGLOMERADO=32) y GBA (AGLOMERADO=33).

b

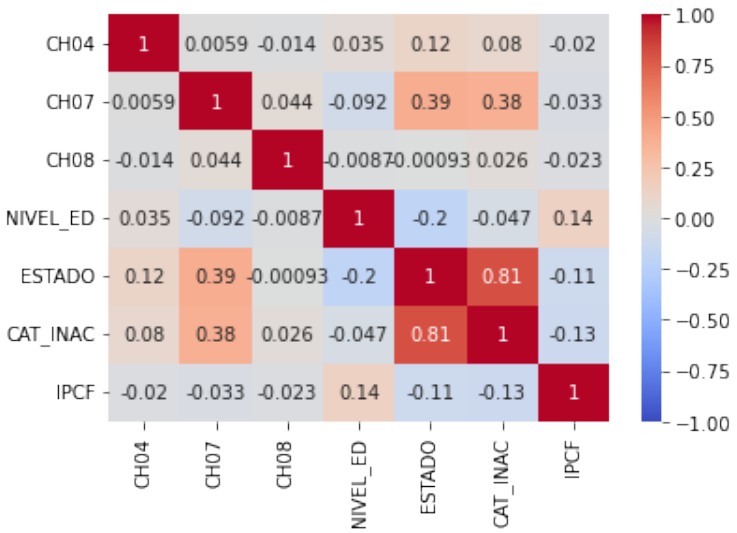
Eliminamos todos los valores de ingresos y edad negativos.

c

Veamos la composición por sexo de la muestra,



d



Matriz de Correlaciones

Observamos una alta correlación entre las variables *CAT_INAC* y *ESTADO*, cosa que no es sorprendente ya que *ESTADO* vale: 1=Ocupado, 2=desocupado, 3=inactivo y 4=menor de 10 años. Y *CAT_INAC* es también otra variable categorica con rango 1-7 y que estable la categoría de inactivo, es decir, está definida para las observaciones con *ESTADO*=3.

e

Composición de la muestra por estado (ocupado, desocupado, inactivo), e ingreso promedio por estado en cantidad de individuos:

Desocupados	Inactivos	Ocupados	Mean_IPCF_ocup	Mean_IPCF_desocup	Mean_IPCF_inac
232	2695	2324	39839	14758	22350

f

Se genera las variables `ad_equiv_hogar` y `ad_equiv` que representan cuanto de un adulto equivalente representan y la suma de todo el hogar.

Ejercicio 3:

No respondieron ingreso ITF (Ingreso total Familiar) unas 2250 personas sobre un total de 5992 que había en esta muestra-para CABA y GBA.Lo que nos da una tasa de no respuesta del 37.5 %.

Ejercicio 4:

Se agregó a la base correspondiente la variable `ingreso_necesario`.

Ejercicio 5:

Estableciendo un valor para la CBA (Canasta Básica Alimentaria) de $27.197,64, identificamos a 278 personas pobres de un total de 3742, en la muestra para GBA Y CABA en la que todos los individuos respondieron ITF (Ingreso Total Familiar). Lo que nos da una tasa de pobreza en la muestra del 7.4 %.

Parte II: Clasificación

Ejercicio 1:

Se eliminaron las columnas relacionadas con ingresos y las agregadas por nosotros.

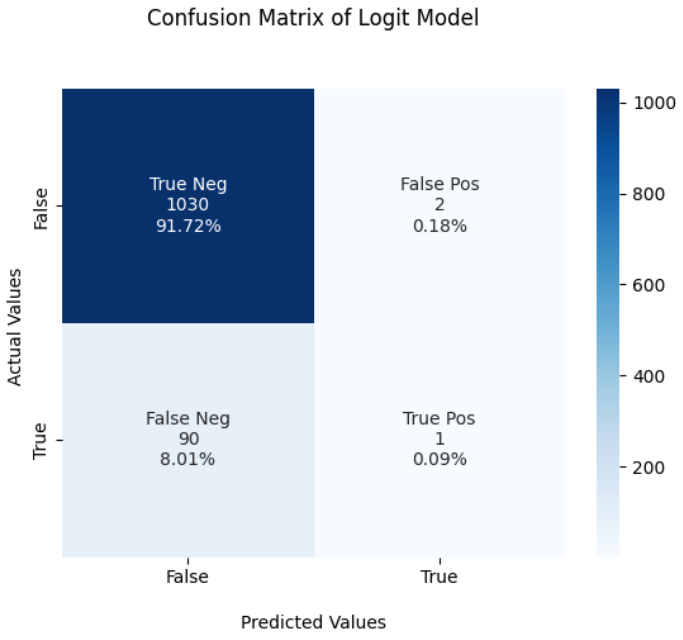
Ejercicio 2:

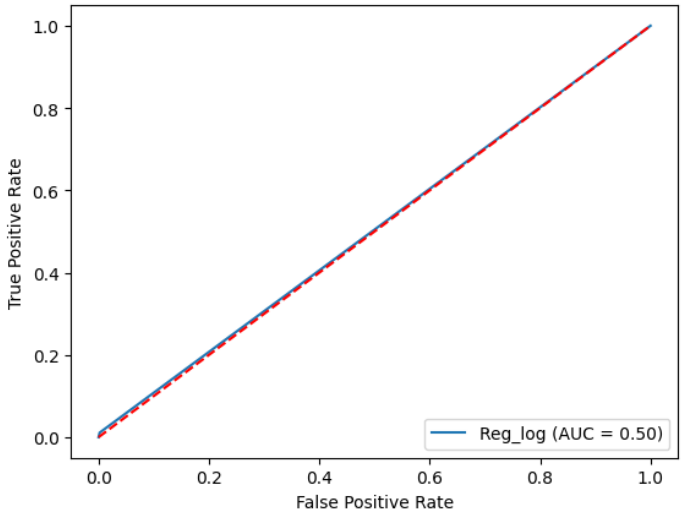
Se dividió la base en train y test. La base Test corresponde al 30 % de los datos. La base Train corresponde al 70 % de los datos.

Ejercicio 3:

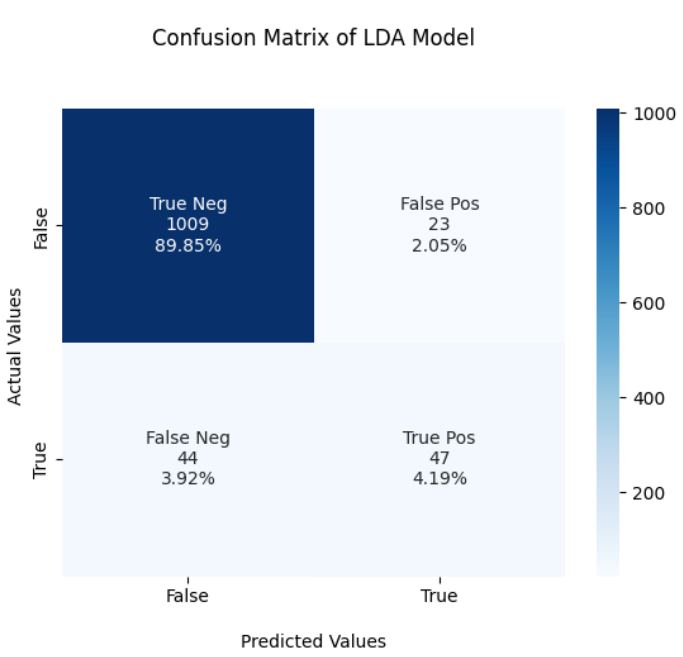
Se presentan las Matrices de Confusión, valores de AUC, curva Roc y Accuracy de los tres metodos requeridos.

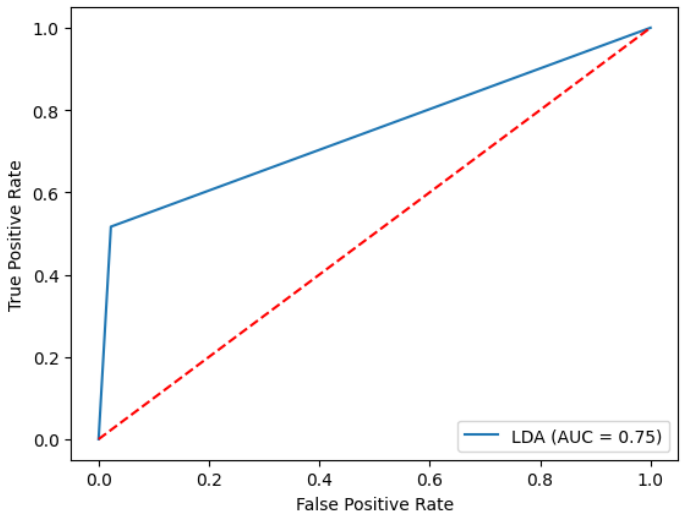
Logit (Accuracy Score 0.918):



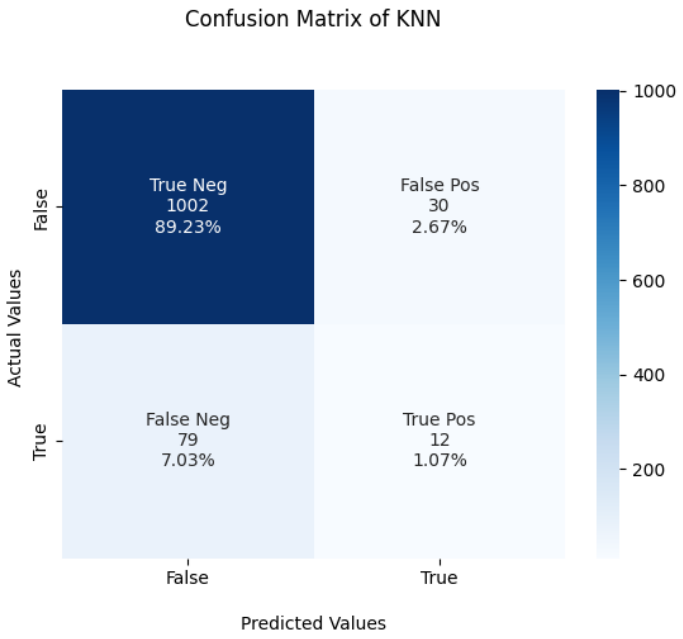


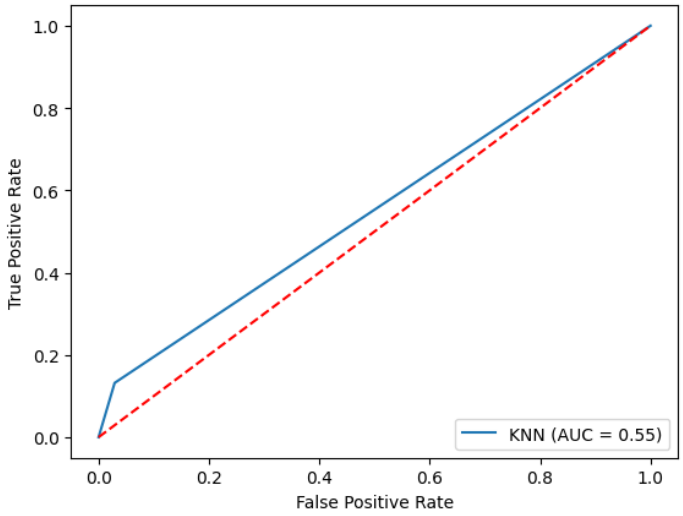
Linear Discriminant Analysis (Accuracy Score 0.940):





KNN con $k = 3$ (Accuracy Score 0.903):





Ejercicio 4:

Utilizando *accuracy* como medida de precisión, el modelo que mejor predice es **LDA**. Si nos itneresase otra medida, esto podría variar. Elegimos accuracy por ser la que pedía el enunciado anterior.

Ejercicio 5:

Utilizando LDA se predice la proporción de pobres entre quienes no respondieron. La cantidad de pobres predicha en la muestra que no respondio es de 70 de 1123 La tasa de pobreza predicha en la muestra de los que no respondieron es de 6.23 %.

Ejercicio 6:

Hay variables que no son relevantes a la hora de predecir la pobreza, por lo que tendría sentido trabajar con menos. Nos quedamos con variables relacionadas al sexo, edad, cobertura medica, si sabe leer o escribir,

educación y establecimiento, y algunas variables sobre busqueda de trabajo de las personas. Ahora, de implementar un modelo Logit, el resultado es el siguiente. Ahora, accuracy toma un valor de 0,918, exactamente igual que antes. No cambión la precisión, ya que estamos descartando variables que no parecen ser relevantes.