Universidad de
SanAndrés

Big Data

Trabajo Práctico 3

Albareti Renato, López Federico y Mettola Franco

Fecha de entrega: 26 de mayo de 2024

Parte I

Ejercicio 1

Para identificar a las personas pobres, INDEC utiliza la famosa *línea de pobreza*, mediante esta línea buscan medir la cantidad de hogares que no cumplen con los ingresos suficientes para consumir la *Canasta Básica Alimentaria* sumado a algunos consumos básicos considerados no alimentarios. Estos consumos de bienes y servicios se agrupan en la *Canasta Básica Total*. La CBT se calcula como la relación entre la CBA y el *Coeficiente de Engel*.

$$CE = \frac{GastoAlimentario}{GastoTotal}$$

Entonces la CBT es:

$$CBT = \frac{CBA}{CE}$$

Reconcemos que el gasto total es mayor o igual al gasto alimentario, por lo que CE debe ser menor o igual a 1 pero mayor a 0. De esta manera, CBT será siempre mayor o igual a CBA.

Finalmente las líneas se compononen por hogar, siendo estos medidos en *Adulto equivalente* para simplificar la medición, el valor de las canastas que estas suponen debe ser contrastado con el ingreso total familiar del hogar. Así se los clasifican como pobres y no pobres, dentro de la categoría de hogar pobre se encuentran las sub-categorías pobre indigente y pobre no indigente.

Fuente: Indec

Ejercicio 2

b) Eliminamos las observaciones negativas para edad y los distintos tipos de ingresos listados en la encuesta, entre los cuales se incluyen ingresos de asalariados, ingresos de independientes, ingresos totales e ingresos familiares. En el caso de los ingresos elegimos no imponer un mínimo arbitrario para evitar ocultar los ingresos de independientes extremadamente bajos -por ejemplo, aquellos que ganaban menos que una jubilación mínima en las fechas de la encuesta-.

c)

Observamos que la cantidad de mujeres es levemente superior en relación a la cantidad de hombres, contando alrededor de un 4 % de diferencia. Lo cual es cercano al valor del censo 2022, contando con una tasa de 51,66 % de mujeres en nuestro territorio.

d)

Observamos que la correlación más fuerte es entre la condición de actividad y la categoría de inactividad, lo cual tiene sentido porque todas las vategorías de inactividad se encuentran dentro de las condiciones de actividad. Otra correlación que notamos es entre la condición de actividad y la categoría de inactividad con el estado civil, lo que posiblemente se deba a aquellas personas que terminan en inactividad luego de tener hijos.

Figura 1: Composición por sexo

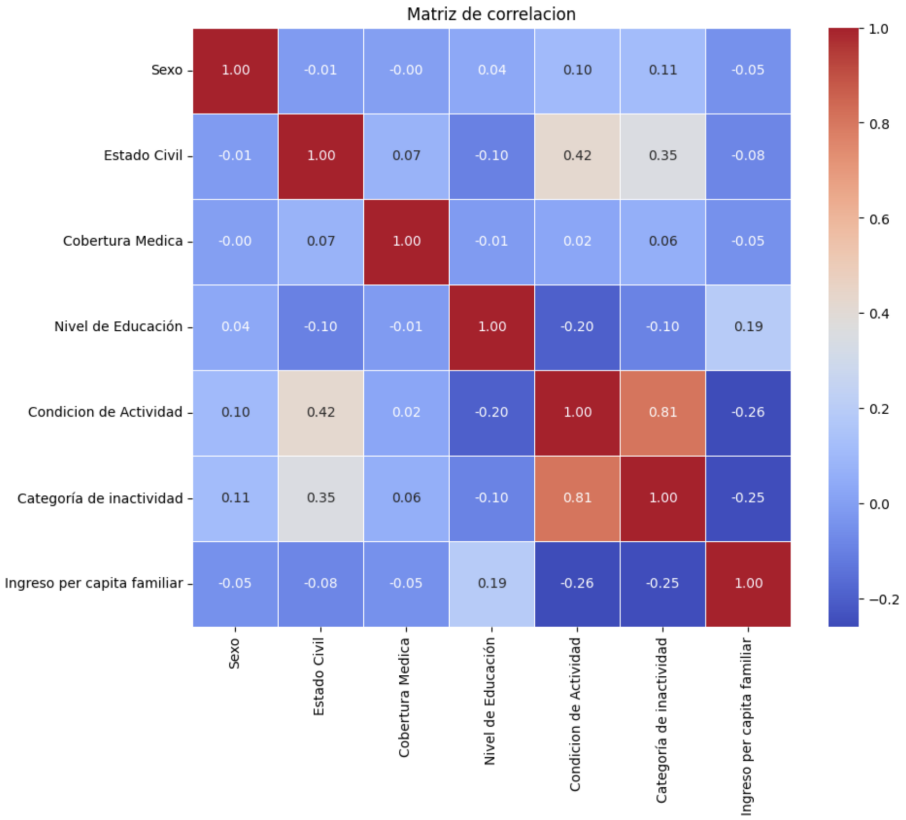

Figura 2: Matriz de Correlación

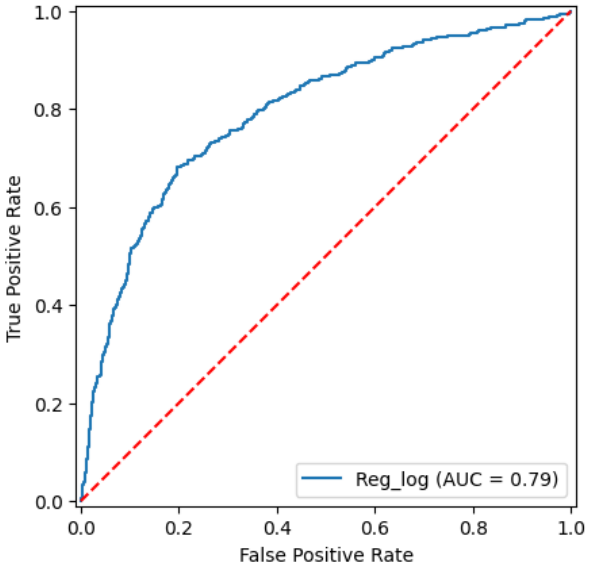

Figura 3: Curva de ROC para regresión logística

Parte II

Ejercicio 3

Regresión Logística Presentamos la curva de ROC del modelo logit
La exactitud del modelo es 0.748675
Análisis Lineal Discriminante
La curva ROC del modelo LDA es la siguiente
Además, la exactitud del modelo es 0.75.
Vecinos cercanos
Al usar KNN con 3 vecindarios, llegamos a la siguiente curva de ROC
Por otro lado, la exactitud del modelo es 0.674.

Ejercicio 4

Sabemos que el peor modelo es KNN, lo cual podíamos esperar ya que estamos en un contexto de alta dimensionalidad, que efectivamente está reduciendo el tamaño de la muestra.

En cuanto a regresión logistica y discriminante lineal, obtenemos medidas de exactitud similares: 0.748 y 0.75, un poco mejor el discriminante lineal. Luego regresion logistica tiene mayor área debajo de la curva. Dado que la exactitud es casi igual, optamos por regresion logística ya que discriminante lineal impone un supuesto de normalidad que no tenemos suficiente evidencia para apoyar.

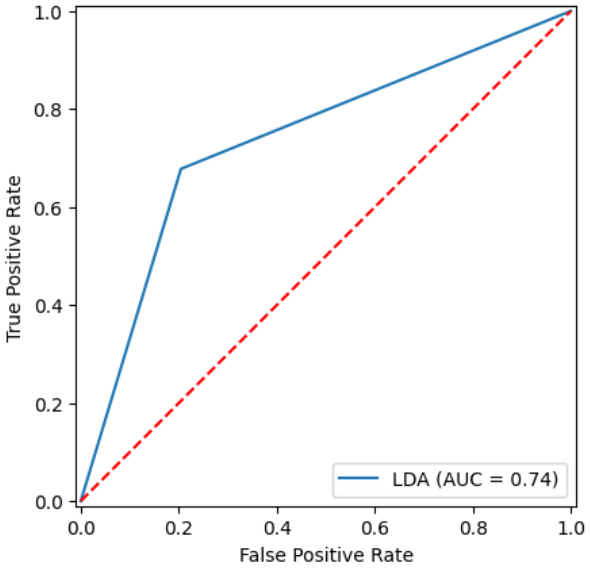

Figura 4: Curva de ROC para LDA

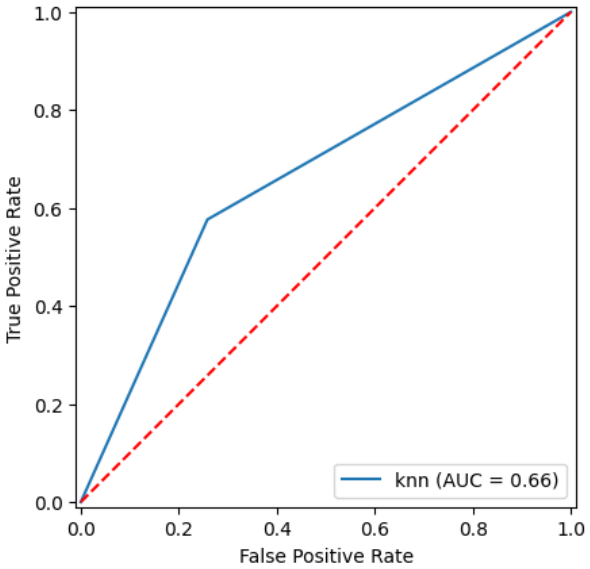

Figura 5: Curva de ROC para KNN