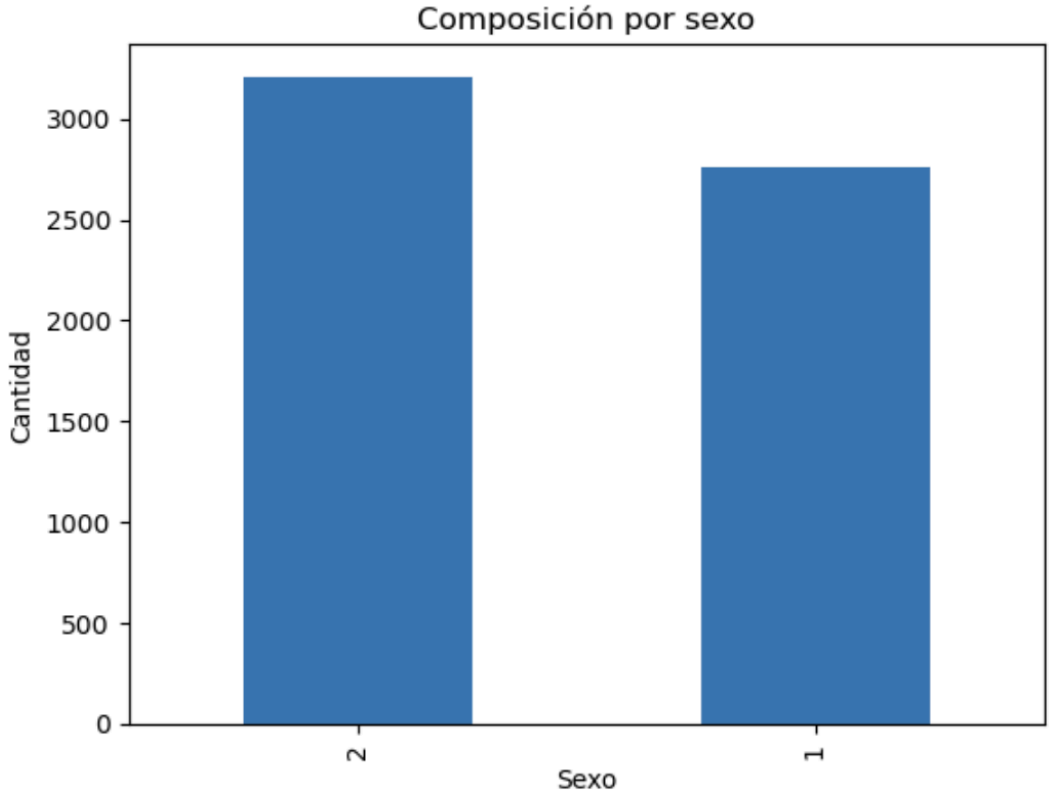
Universidad de
SanAndrés

BIG DATA - TP 3

De León, Juan Cruz - Di Costanzo Pereira, Nina - Morozumi, Juan Ignacio
Universidad de San Andrés - Otoño 2024

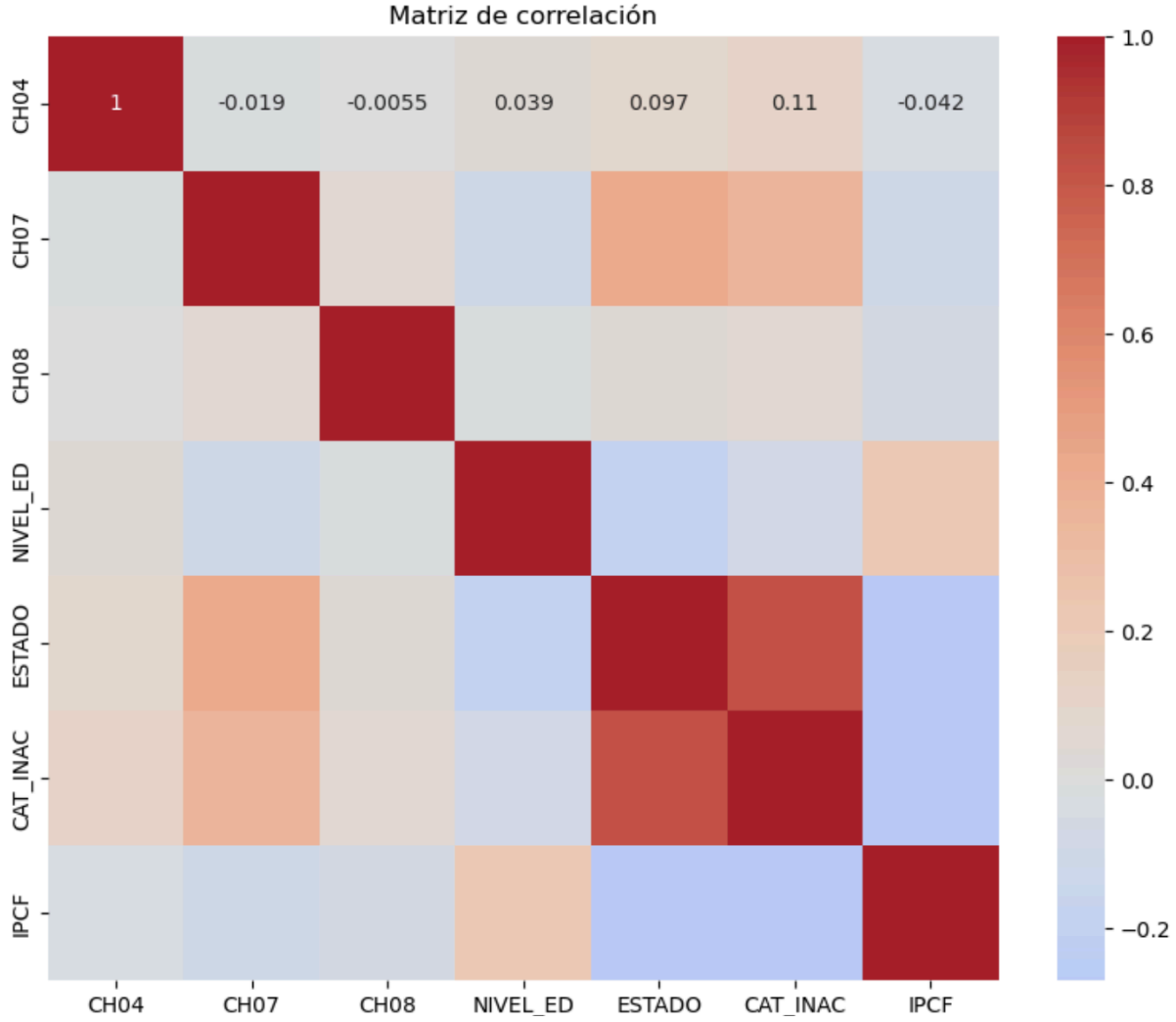
Parte 1

1. Según el INDEC una persona pobre es aquella que no puede cubrir las necesidades básicas, ya sea de la canasta básica, de vivienda o de educación. Se considera que un hogar está bajo la línea de la pobreza cuando este no es capaz de cubrir las necesidades básicas del hogar con los ingresos familiares.
2.
    a. Lo que hicimos fue quedarnos únicamente con los aglomerados que correspondan al 32 y 33, es decir, al de la Ciudad Autónoma de Buenos Aires y al de Gran Buenos Aires.
    b. Lo que hicimos acá, fue descartar todos aquellos valores que sean negativos, ya que, no iba a tener sentido tenerlos en cuenta para que formen parte de nuestra base de datos. Como por ejemplo, edades que sean negativas.
    c. Podemos ver como, las cantidades varían dado el sexo de la persona. En el gráfico de barras podemos ver que el número 2 corresponde al género femenino, mientras que el número 1 corresponde al masuclino. También, podemos notar que el número de mujeres y varones varía en aproximadamente 1000 unidades dada la escala de nuestro gráfico.



    d. Calculamos la Matriz de correlacion entre las variables CH04 (la cual se refiere al sexo), CH07 (situación conyugal), CH08 (tipo de cobertura medica que tiene la persona), NIVEL_ED (nivel educativo), ESTADO (condicion de actividad de la persona), CAT_INAC (categoria de inactividad) e IPCF (monto de ingreso familiar percibido en ese mes). Y con los resultados que obtuvimos, pudimos deducir las siguientes conclusiones:

- Las dos variables que tienen una mayor correlación entre sí son las de CAT_INAC y ESTADO, ya que, el color del cuadrado de la matriz de correlación es el que más oscuro es.
- Otras dos variables que tienen cierta correlación entre sí, son las de ESTADO y CH07. Que indican el estado de actividad de una persona y la situación conyugal de la misma.
- Las últimas dos variables que muestran algún tipo de correlación son las de CATINAC y CH07, que miden la categoría de inactividad de la persona con la situación conyugal de la misma.



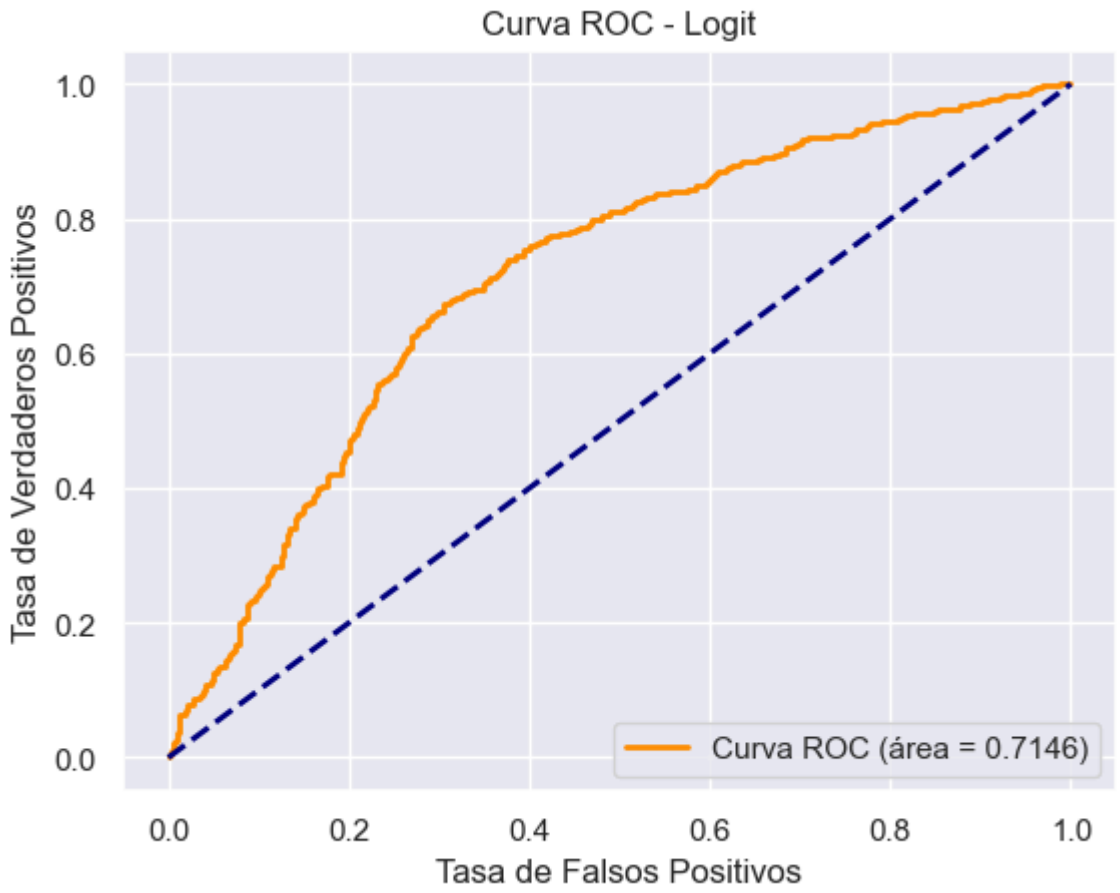
e. Podemos notar que hay una cantidad de desocupados de 226 y de activos de 2507.
El ingreso promedio de los ocupados es de 190809.678283.
El de los desocupados es de 61605.874425
Y por último el de los inactivos es de 93740.533139

f. Asignamos los valores de energías para las personas, es decir, por edad y por sexo.

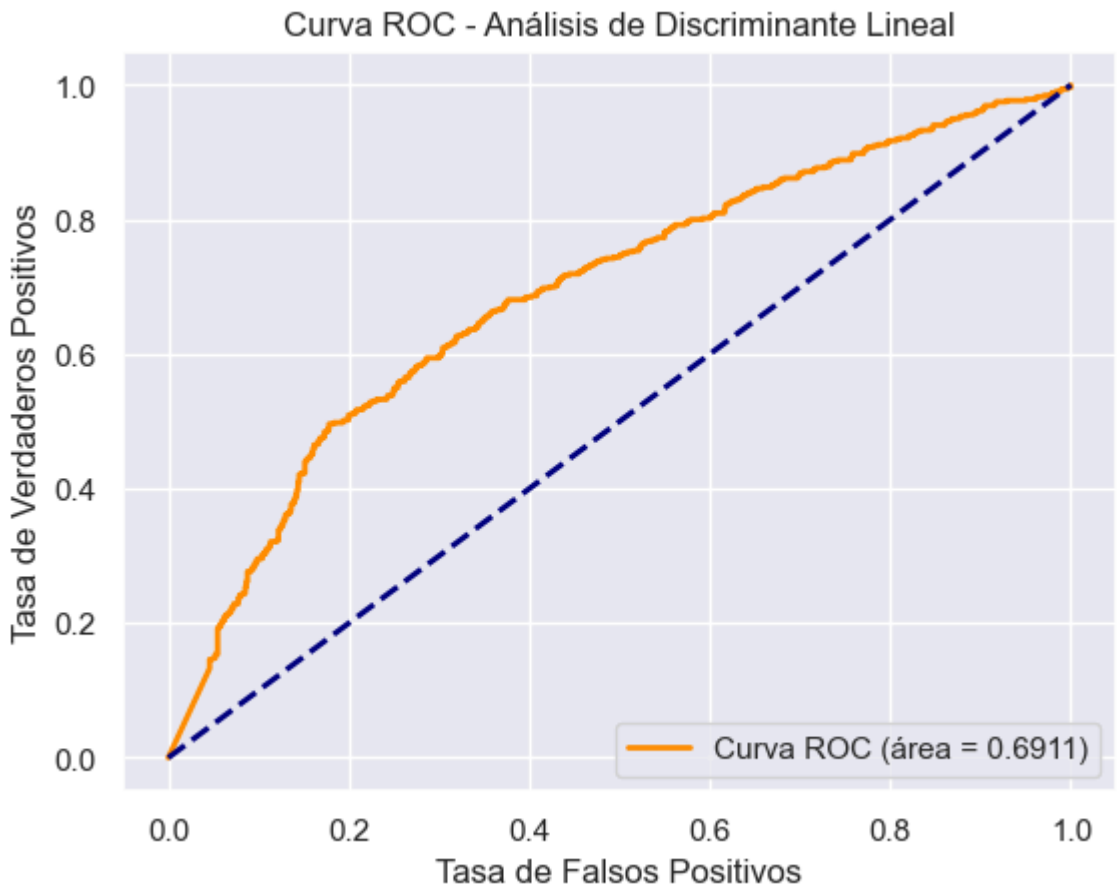
3. La cantidad de personas que no respondieron a la encuesta es de un total de 1618. Por lo tanto, nos queda una cantidad de 1115 personas que sí respondieron a la encuesta.
4.
5.Identificamos 1731 pobres, es decir quienes tienen un ITF menor al ingreso necesario calculado.

Parte 2

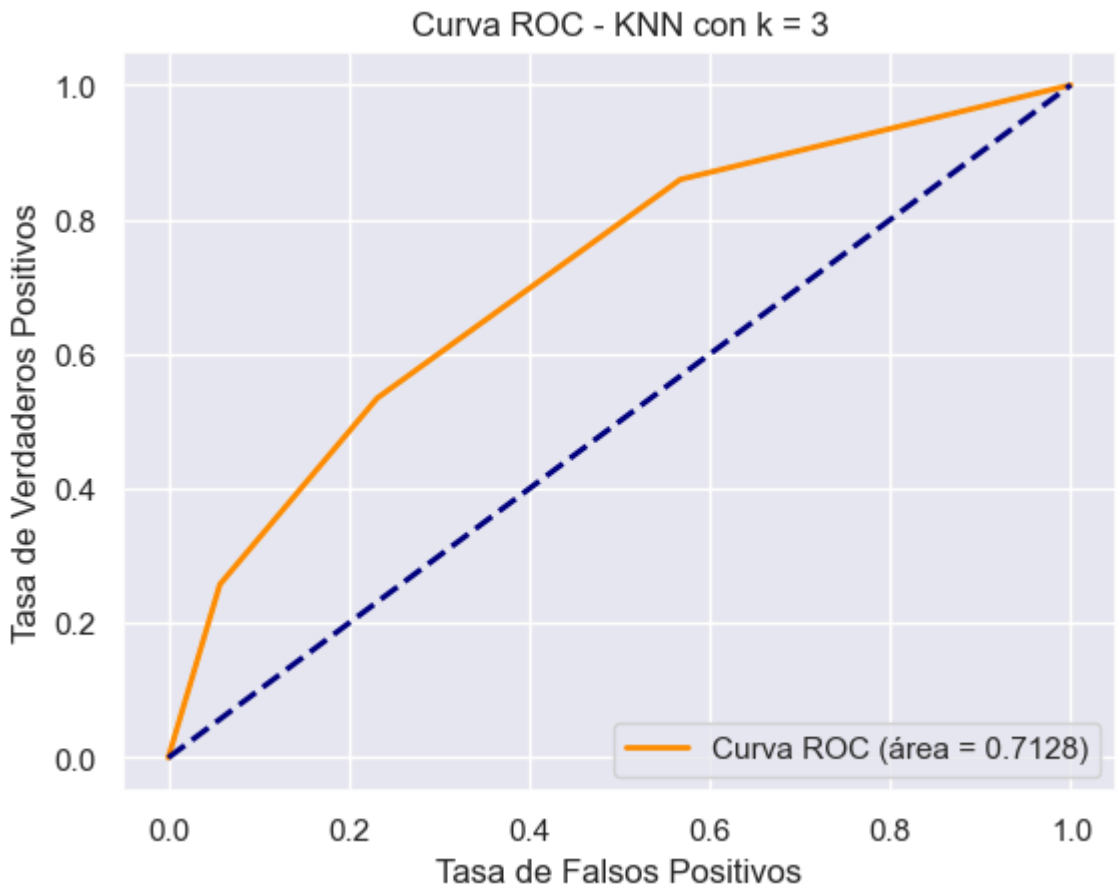
3.
Resultados del modelo Logit:
Matriz de confusión:
[[625 161]
 [270 250]]
Accuracy: 0.6700
AUC: 0.7146



Resultados del modelo Análisis de Discriminante Lineal:
Matriz de confusión:
[[535 251]
 [196 324]]
Accuracy: 0.6577
AUC: 0.6911



Resultados del modelo KNN con k = 3:
Matriz de confusión:
[[604 182]
 [242 278]]
Accuracy: 0.6753
AUC: 0.7128



4. El modelo que predice mejor es el KNN ya que es el que tiene un menor error cuadrático medio (ECM = 0,324655)
Errores cuadráticos medios promedio para cada modelo:

model	ECM
Logit	0.330015
LDA	0.342266
KNN	0.324655

5. Basándonos en el método KNN con k=3, la predicción de qué personas son pobres dentro de la base que no respondieron es del 46.48%.

6. No es correcto tomar tal cantidad de variables para poder regresar el modelo ya que en principio no se determinó si por ejemplo las variables están correlacionadas o si el modelo sobreajusta por la cantidad. Sacamos variables que consideramos no son relevantes ej. las de identificación y las que tienen información repetida ej. fecha de nacimiento y edad o asistió a establecimiento educativo y nivel educativo (CH10 y NIVEL_ED).

REFERENCIAS:

INDEC. (2014). Diseño de registros y estructura de las bases de microdatos: Hogar e Individual. Recuperado de https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_diseno_reg_t414.pdf