Big Data: Trabajo Práctico 3

Autores: Sofía Ellenberg, Vicente Zervino, Sophie Schulzen

Profesores: Maria Noelia Romero

Tutor: Victoria Oubiña

Parte 1

Ejercicio 1

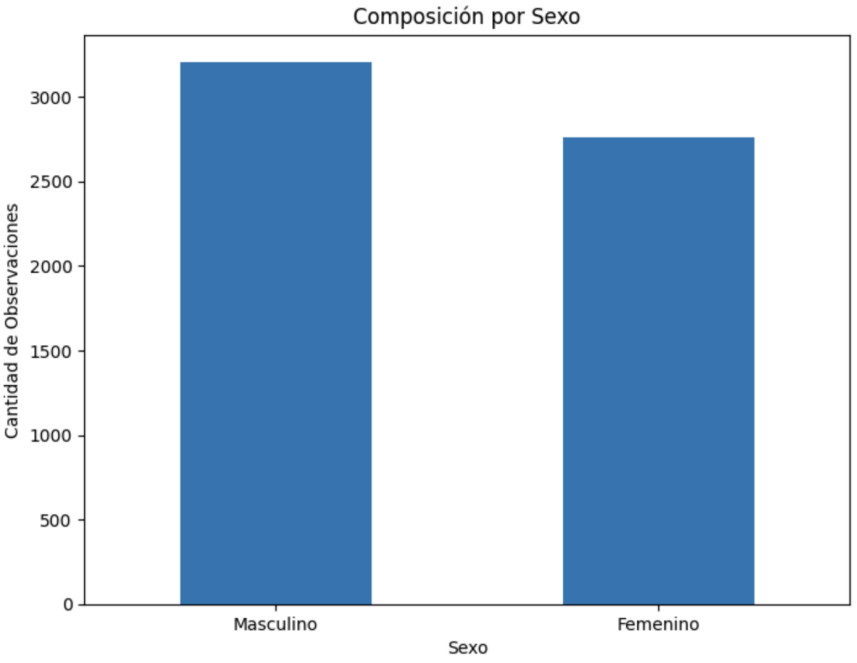
Según la información obtenida por la página del INDEC, las personas pobres se pueden definir mediante la medición de una línea de pobreza. Más específicamente, el INDEC representa una canasta básica que contiene bienes y servicios tanto alimentarios como no, que son esenciales para poder sobrevivir y llevar a cabo una vida medianamente estable. En esa canasta básica, se incluyen alimentos, bienes y algunos servicios fundamentales en Argentina. Una vez determinada dicha canasta, se le asigna un *valor de la canasta básica total.* A partir de ahí, se realiza una encuesta a los hogares, denominada EPH donde se le pregunta acerca de sus ingresos mensuales para luego poder elaborar una línea de pobreza. Finalmente, las personas pobres serán identificadas como aquellas que están por debajo de la línea de pobreza y por ende, no pueden acceder a comprar la canasta básica.

Ejercicio 2

b.

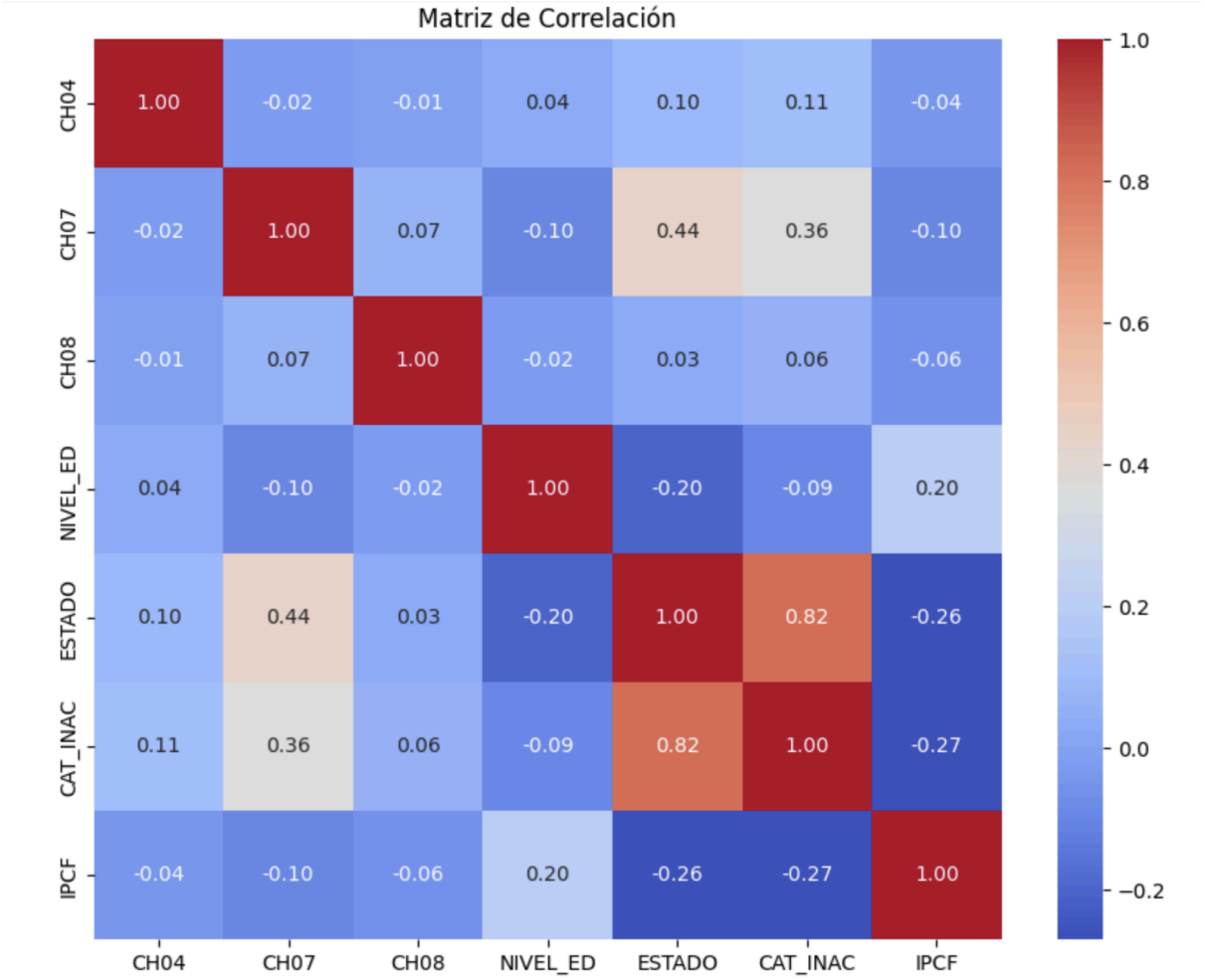
Eliminamos diversas variables cuyo valor podría no tener sentido, entre ellas, se encuentran variables que demuestran algún tipo de ingresos y las edades de los individuos. En ese sentido, las variables que incluimos son ITF', 'DECIFR', 'IDECIFR', 'RDECIFR', 'GDECIFR', 'PDECIFR', 'ADECIFR', 'P21', 'DECOCUR', 'IDECOCUR', 'RDECOCUR', 'GDECOCUR', 'PDECOCUR', 'ADECOCUR', 'PONDIIO', 'IPCF', 'DECCFR', 'IDECCFR', 'RDECCFR', 'GDECCFR', 'PDECCFR', 'ADECCFR', 'PONDIH', 'P47T', 'DECINDR', 'IDECINDR', 'RDECINDR', 'GDECINDR', 'PDECINDR', 'ADECINDR', 'PONDII', 'V2_M', 'V3_M', 'V4_M', 'V5_M', 'V8_M', 'V9_M', 'V10_M', 'V11_M', 'V12_M', 'V18_M', 'V19_AM', 'V21_M', 'T_VI', 'CH06

c.



Este gráfico nos indica la proporción de hombres y mujeres que existen en la base de datos de la EPH. Podemos observar que, luego de haber realizado la limpieza de datos, la cantidad de observaciones de los hombres es mayor que la de las mujeres. Más específicamente, mientras que hay más de 3000 observaciones para los hombres, la cantidad de observaciones para las mujeres son un poco menos de 3000.

d.



El gráfico representa una matriz de correlación, donde los colores más azules son los que indican una menor correlación entre las variables, mientras que los más colorados son los que más correlación tienen. Por ejemplo, la Variable CAT_INAC tiene una correlación negativo de 0.21 con la variable IPCF. Es decir, la categoría de inactividad (Jubilado / Pensionado,Rentista,Estudiante, Ama de casa, Menor de 6 años, Discapacitado y Otros) se correlaciona negativamente con el monto de ingreso familiar. El mismo análisis se da entre la variable estado e IPCF. Por otro lado, observando aquellas variables que más se correlacionan, vemos que la condición de actividad se correlaciona fuertemente con la categoría de inactividad, lo cual tiene sentido. Es decir, si una persona Jubilado, lo lógico sería ver que su estado es desocupado. Por otro lado, también existe una significante correlación entre el nivel educativo de una persona (nivel_ed) y el ingreso per cápita familiar,

lo cual indica que a mayor nivel educativo, mayor nivel de ingreso per cápita familiar.Finalmente, otro punto a destacar es observar la alta correlación que existe entre la condición de actividad (ESTADO)
e.

Hay 226 individuos desocupados en la muestra y 2507 inactivos. La media del ingreso per cápita familiar (IPCF) según estado es:

- Ocupado = 190809.678283
- Desocupado = 61605.874425
- Inactivo = 93740.533139

Ejercicio 3

Según nuestro análisis, 1618 hogares no respondieron su ingreso total familiar

Ejercicio 5

Hay 1732 personas pobres según nuestro análisis

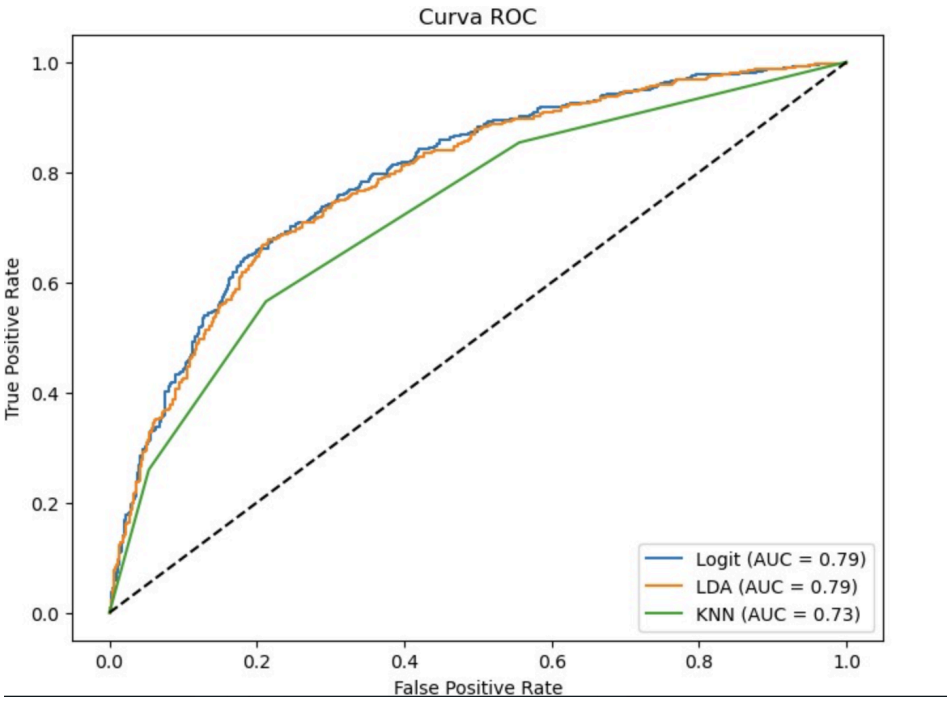
Parte 2

Ejercicio 4

Una vez realizado el ejercicio 3, para poder determinar cual de los tres modelos (KNN, Logit y Análisis discriminante lineal) predice mejor, es fundamental analizar y comparar los resultados obtenidos de la Curva ROC, los valores de AUC y Accuracy de cada uno de los tres modelos. En ese sentido, un valor de accuracy más alto, indica un mejor desempeño en la predicción. Además, cuando el valor de AUC se acerca a 1, estamos hablando que el modelo tiene un mejor desempeño en la discriminacion de las clases. Por otro lado, cuando incluimos la curva ROC en nuestro análisis, procedemos a analizar el trade off que se realiza en las tasas de ponderaciones de las funciones castigos para los errores de tipo falso negativo y falso positivo. En otras palabras, visualizando la curva ROC, podemos determinar que modelo predice mejor observando la forma de su curva: mientras más alejada esté la curva de la línea diagonal hacia arriba, mejor predice el modelo.

En nuestro análisis, los resultados que obtuvimos para los tres modelos fueron los siguientes:

Curva ROC:



Logit:

- Matriz de confusión:

  [[656 130]

  [203 317]]

- AUC: 0.7936337835192796
- Accuracy: 0.7450229709035222

LDA:

- Matriz de confusión:

  [[648 138]

  [211 309]]

- AUC: 0.7873434135838715
- Accuracy: 0.7327718223583461

KNN:

- Matriz de confusión:

  [[619 167]

  [226 294]]

- AUC: 0.7278711587394793
- Accuracy: 0.6990811638591118

De este modo, observando los resultados, podemos concluir que el modelo LOGIT es el que mejor capacidad de predicción tiene en nuestro análisis. En primer lugar, esto se debe a que, observando la curva ROC, el modelo logit podría ser el que mejor predice porque tiene una forma más encorvada, lo que indica que tiene un mejor rendimiento de discriminacion y clasificación entre falsos positivos y falsos negativos a la hora de predecir. En segundo lugar, si bien el modelo LDA también tiene una curva ROC similar a la del modelo Logit, los valores del AUC y Accuracy son mayores para el modelo Logit. Esto es deseable ya que nos indica que tiene un menor número de errores (falsos positivos y falsos negativos) según el valor del accuracy y, además, que es el más efectivo a la hora de diferenciar correctamente en las clases positivas y negativas, según el valor de AUC. Finalmente, el modelo KNN=3 es rápidamente descartado porque su curva ROC se acerca mucho a la diagonal, lo cual no es bueno a la hora de evaluar su performance para predecir, y además, tiene los menores valores de AUC y Accuracy.

Ejercicio 5

La proporción de las personas que no respondieron y fueron identificadas como pobres es 0.8028430160692213

Ejercicio 6

Si bien las variables utilizadas como predictoras son informativas para predecir si una persona es o no pobre, podría pasar que haya algunas que están muy correlacionadas entre sí. Es decir, que individualmente no proporcionan mucha información para nuestro análisis. En ese sentido, las variables que incorporamos en nuestro modelo hablan sobre la educación, el nivel de empleo y la zona donde viven. Sin embargo, hay muchas variables incorporadas sobre la forma en la que buscan trabajo que no necesariamente son indispensables para nuestro análisis. De este modo, proponemos eliminar aquellas variables que indican cómo ha buscado empleo una persona, de qué forma y si hizo contactos de alguna forma, si mando el curriculum etc. Mas especificamente, las variables que decidimos eliminar son: PP02C1 , PP02C2, PP02C3, PP02C4, PP02C5, PP02C6, PP02C7, PP02C8, PP02E. Sin embargo, para poder tener algún predictor que nos hable sobre la búsqueda laboral de la persona, nos quedamos con PP02H y PP02I que indican si han estado buscando empleo en los últimos 12 meses y si han logrado trabajar en los últimos 12 meses.

Finalmente, los resultados que obtuvimos al eliminar dichas variables son:

Logit:

- Matriz de confusión:

  [[657 129]

[203 317]]

- AUC: 0.7945390487375221
- Accuracy: 0.7457886676875957

Comparando los resultados con el ejercicio anterior, podemos concluir que eliminar dichas variables no afecta tanto en la predicción de nuestro modelo, ya que la caída en los valores de AUC y Accuracy es casi nula.