Universidad de San Andrés

Departamento de Economía

Big Data

Trabajo Práctico Nº 2

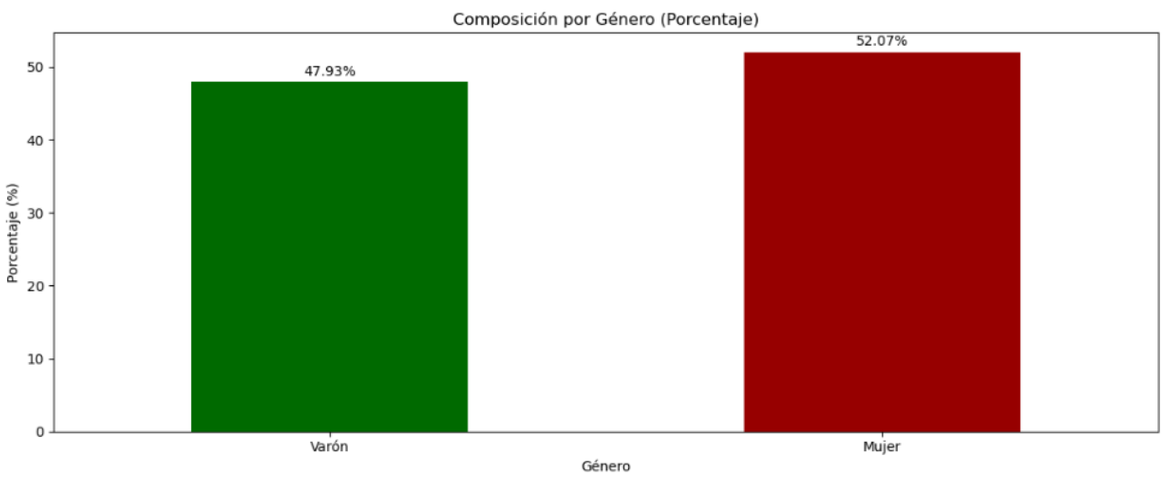
Ejercicios parte I

- 1) El documento elaborado por el INDEC detalla con profundidad la metodología adoptada para medir la pobreza y la indigencia en Argentina. Central en este enfoque es el método de medición indirecta, que se basa en establecer "líneas" o umbrales de ingresos. La "Línea de Indigencia" (LI) se configura para identificar si los ingresos de un hogar son suficientes para cubrir una Canasta Básica Alimentaria (CBA). Esta canasta está diseñada para satisfacer un umbral mínimo de necesidades energéticas y proteicas, reflejando el patrón de consumo de alimentos de la población de referencia.

  Por otro lado, la "Línea de Pobreza" (LP) tiene un alcance más amplio. Va más allá de las necesidades alimentarias y considera otros consumos esenciales no alimentarios. Así, se forma la Canasta Básica Total (CBT), que incluye aspectos como vestimenta, transporte, educación, salud, entre otros. Esta canasta se compara con los ingresos de los hogares, información que se extrae de la Encuesta Permanente de Hogares (EPH).

  Es crucial entender que la Canasta Básica Alimentaria ha sido determinada basándose en diferentes Encuestas de Ingresos y Gastos a lo largo del tiempo. Estas encuestas han mostrado cambios en los hábitos de consumo de la población argentina. Por ejemplo, en años recientes, se ha observado una reducción en el gasto alimentario y un incremento en otros rubros, como el transporte y la vivienda. Estas variaciones tienen un impacto directo en cómo se mide la pobreza, ya que afectan la distancia entre la Línea de Indigencia y la Línea de Pobreza. Por ello, el INDEC subraya la importancia de realizar actualizaciones metodológicas periódicas y considerar una variedad de factores para garantizar una medición precisa y representativa de la pobreza en el país.

- 2.c) En la Encuesta Permanente de Hogares correspondiente al primer trimestre de 2023 para el Gran Buenos Aires, la composición por género muestra que de los encuestados, 3,915 corresponden a mujeres, representando el 52% del total, mientras que 3,604 son hombres, abarcando el 48% restante. Esta distribución refleja una ligera predominancia femenina en la muestra.



- 2.d) La matriz revela varias correlaciones entre las variables estudiadas. Entre 'ESTADO' (ocupación) y 'NIVEL_ED' (nivel educativo), se observa una correlación de −0.20, indicando una asociación negativa leve. Esto sugiere que ciertos niveles educativos podrían no favorecer determinadas ocupaciones, o viceversa. Por otro lado, 'CAT_INAC' (Categoría de inactividad) y 'ESTADO' exhiben una robusta correlación de 0.82, lo que implica que ciertas categorías de inactividad están estrechamente vinculadas a ciertos estados ocupacionales.

  Adicionalmente, 'CH07' (estado civil) y 'ESTADO' tienen una correlación moderada de 0.42, insinuando una relación entre la situación conyugal y la ocupación. Es esencial destacar que estos coeficientes se determinaron tras excluir observaciones con datos no especificados o no recabados. Finalmente, resulta vital recordar que la correlación no necesariamente implica causalidad.

- 2.e) Los datos para el Gran Buenos Aires del primer trimestre de 2023 muestran que hay 286 personas desocupadas y 2,826 inactivas. Esta cifra de desocupación, aunque relativamente baja en comparación con el número total de inactivos, se refleja en la disparidad de ingresos entre estos grupos. El Ingreso Promedio Per Cápita Familiar para los ocupados es de $59,812.34, una cifra significativamente mayor en comparación con los desocupados, quienes tienen un ingreso promedio de $25,536.02. Sorprendentemente, el ingreso de los inactivos, que asciende a $40,089.14, se sitúa en un punto intermedio, sugiriendo que, a pesar de no estar activamente empleados, este grupo aún cuenta con fuentes de ingreso que superan a las de los desocupados.

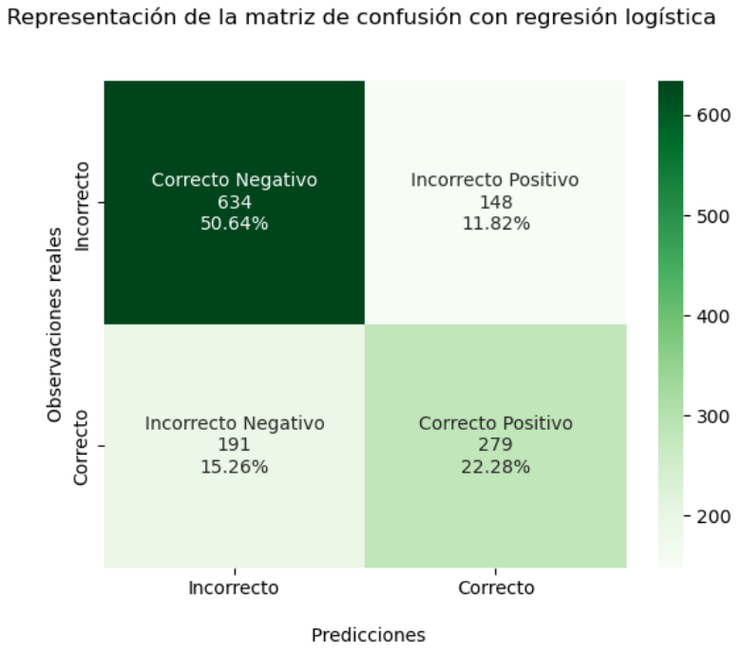
- 3) 3,315 personas no proporcionaron datos de ingresos, mientras que 4,173 sí lo hicieron. Ello por haber excluido de la columna de deciles las categorías "12 = no respuesta de ingresos" y "13 = entrevista individual no realizada".

- 5) Para determinar el número de personas en situación de pobreza en el Gran Buenos Aires en el primer trimestre de 2023, se utilizó la Canasta Básica Total para un adulto equivalente, que es aproximadamente $57.371,05. Para ello, se agregó a la base de datos una columna

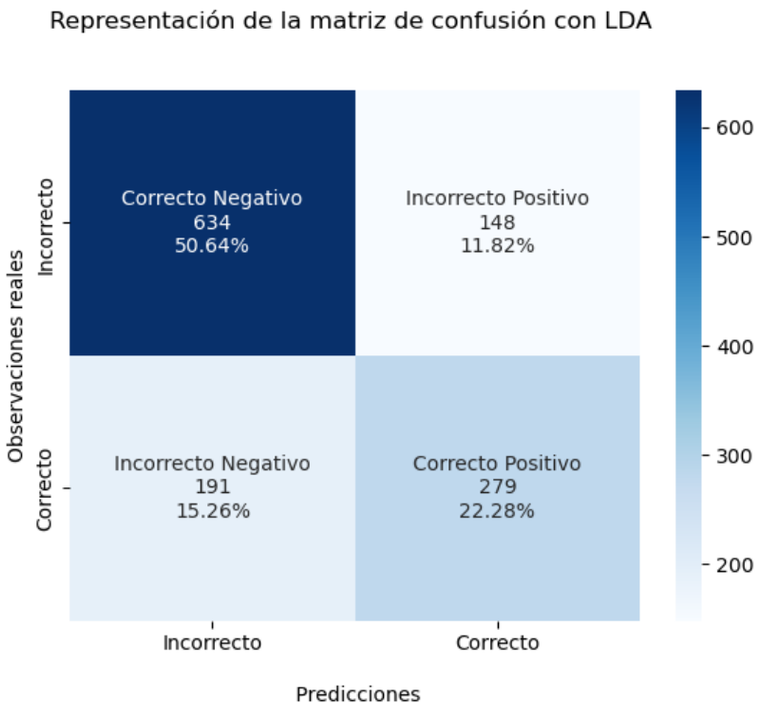
llamada "ingreso necesario", la cual se obtuvo multiplicando dicho valor por la variable 'ad_equiv_hogar'. Este cálculo representa el ingreso mínimo que necesita cada hogar para no ser considerado pobre. Al aplicar este criterio, se identificaron 1,551 personas en situación de pobreza, lo cual significa un 37.17% del total de personas contabilizadas.

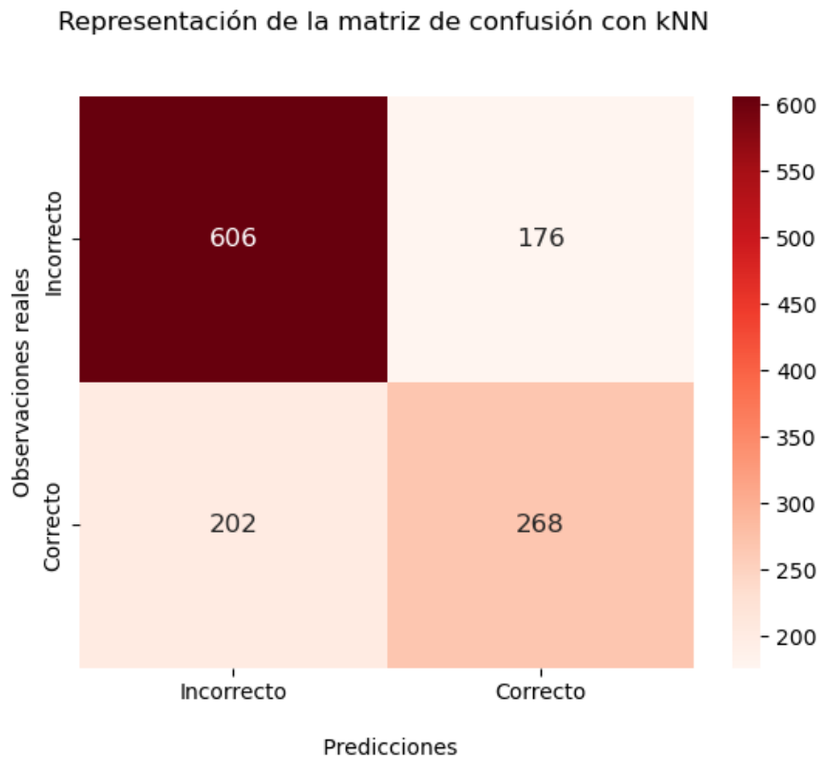
Ejercicios parte II

- 4) A continuación los resultados de la aplicación de los tres modelos de clasificación para predecir la pobreza sin usar variables de ingresos. El desafío de este enfoque radica en la exclusión de variables de ingresos, que suelen ser determinantes en este tipo de predicciones.

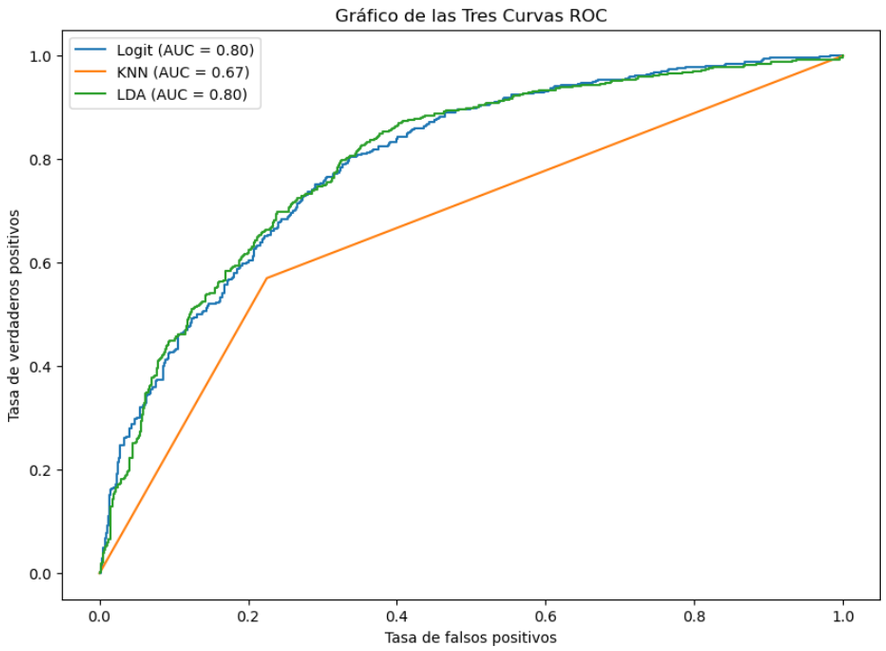


Para el modelo de **Regresión Logística (Logit)**, el área bajo la curva (AUC) fue de 0.702 y la precisión (Accuracy) alcanzó un valor de 0.729233. En cuanto al **Análisis Discriminante Lineal (LDA)**, se obtuvo un AUC de 0.702 y una precisión de 0.734026. Finalmente, para el modelo **kNN (k Nearest Neighbors)**, el AUC resultó ser 0.67 con una precisión de 0.698.





Comparando los tres modelos, se puede observar que todos tienen un desempeño similar en términos del AUC, siendo los modelos Logit y LDA ligeramente superiores con un AUC de 0.702, mientras que el kNN tiene un AUC de 0.67. En cuanto a la precisión, el modelo LDA tiene la precisión más alta con 0.734026, seguido de cerca por la regresión logística con 0.729233, y finalmente el kNN con 0.698.



Al examinar el gráfico que contiene las tres curvas ROC, es evidente que los modelos de **Regresión Logística** y **Análisis Discriminante Lineal** muestran un desempeño muy similar entre sí, evidenciando una capacidad predictiva comparable. Sin embargo, el modelo **kNN** claramente se desempeña de manera inferior, lo que indica una menor capacidad para discriminar entre las clases.

La gestión de valores faltantes es una tarea esencial en el análisis de datos. Existen diversas estrategias para tratar con ellos, como la eliminación completa, imputación con valores fijos, técnicas estadísticas, métodos de aprendizaje automático, entre otros. Cada enfoque tiene sus ventajas y desventajas, y la elección adecuada depende del contexto, tipo de datos y el objetivo del análisis. Después de una consideración, en este trabajo, se eligió imputar los valores faltantes con ceros. Esta decisión se basó en su simplicidad, facilidad de implementación

y en la interpretación particular que el valor cero puede tener en el contexto específico de nuestro estudio. Es crucial ser consciente de sus limitaciones y validar los resultados para asegurarse de que la imputación no haya introducido sesgos en el análisis.

Es crucial señalar que estos resultados son altamente sensibles a las decisiones adoptadas durante la fase de preprocesamiento de datos. Específicamente, la elección subjetiva de tratar los valores faltantes (NAs) reemplazándolos con ceros y la decisión de utilizar variables numéricas pueden influir significativamente en la performance de los modelos. Cualquier variación en estas decisiones puede alterar los resultados, lo que subraya la importancia de considerar diferentes estrategias y validar la robustez de los hallazgos.

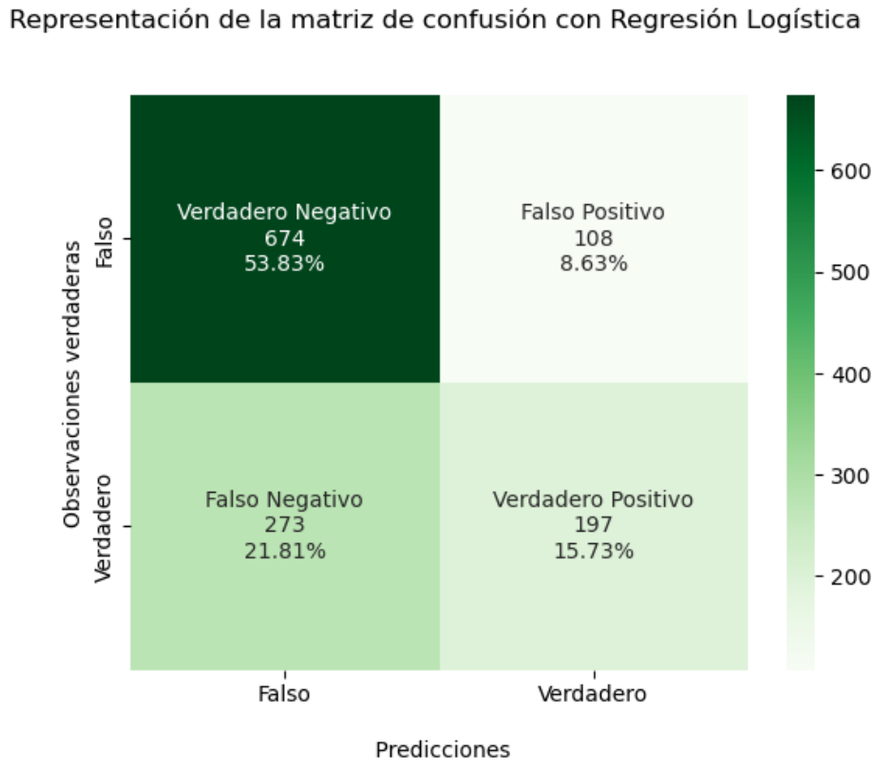
- 5) En este trabajo , empleamos el método de Regresión Logística para identificar la proporción de individuos pobres en la base *no respondieron*. Al computar la media de quienes caían por debajo de la línea de la pobreza sobre el conjunto *respondieron*, identificamos que el 37.17% de la muestra fue clasificado como pobre. Al aplicar el modelo logit al conjunto *no respondieron*, el modelo predijo que el 39.76% de esta muestra sería pobre.

  Es importante destacar que, dada la ausencia de la variable objetivo en el conjunto *no respondieron*, no es posible realizar una validación post hoc de las predicciones. Ni tampoco una evaluación del método en este subconnjunto. Sin embargo, la proporción predicha es comparable a la observada en la muestra *respondieron*, lo que sugiere una consistencia en las características observadas entre ambas muestras y un rendimiento consitente del método validado previamente.

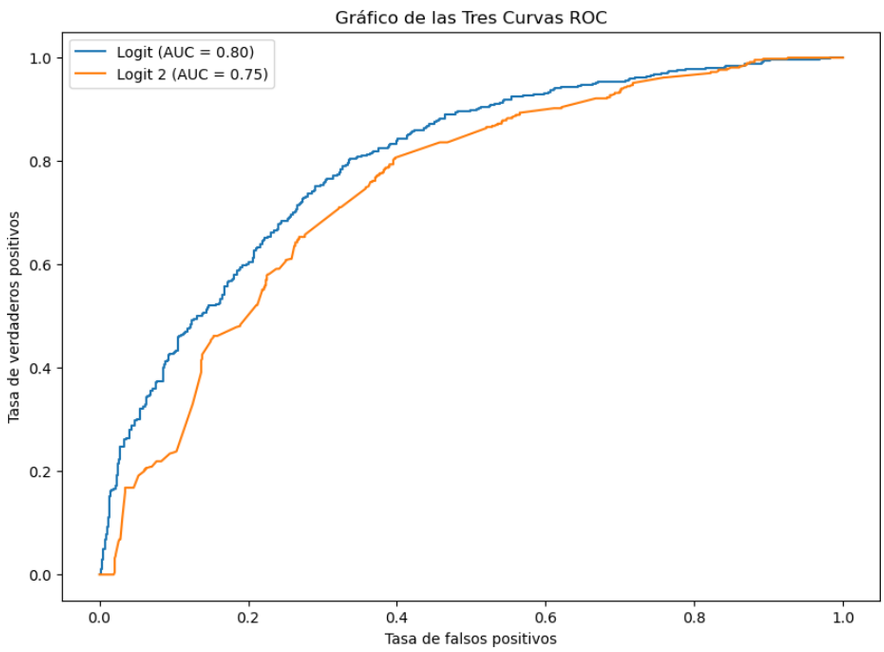
- 6) En el modelo previo, se utilizaron todas las variables disponibles, excluyendo aquellas relacionadas directamente con el ingreso, como predictores. Si bien este enfoque puede parecer exhaustivo, conlleva el riesgo de saturar el modelo. En términos de aprendizaje automático, la inclusión indiscriminada de todas las variables disponibles puede llevar a un sobreajuste (overfitting), donde el modelo se ajusta demasiado bien a los datos de entrenamiento pero pierde poder predictivo en datos nuevos o no vistos.

  Para el nuevo modelo, se ha optado por una selección más acotada y deliberada de variables, utilizando como predictores el sexo, la cobertura médica, la alfabetización, el nivel educativo, el monto total de ingresos no laborales y el monto de ingresos por intereses o rentas por plazos fijos o inversiones. Esta selección tiene el propósito de mejorar la generalización del modelo y evitar los problemas anteriormente mencionados.

  La matriz de confusión para el nuevo logit nos indica que el modelo ha clasificado correctamente a 674 individuos como no pobres, mientras que ha identificado erróneamente a 108 individuos como pobres. Por otro lado, ha clasificado correctamente a 197 individuos como pobres, pero ha fallado en identificar a 273 verdaderos pobres. Estas cifras nos proporcionan un valor de precisión de 0.696, lo que implica que el modelo es capaz de clasificar correctamente al 69.6% de los individuos en la muestra de prueba.



Al observar la curva ROC, es evidente que el modelo logístico previamente implementado tiene un mejor rendimiento en términos de tasa de verdaderos positivos en comparación con el nuevo modelo. Específicamente, la curva del modelo anterior está consistentemente por encima de la curva del nuevo modelo, lo que indica que tiene una mayor capacidad para identificar correctamente a los verdaderos positivos, manteniendo una tasa de falsos positivos similar.



En resumen, mientras que el nuevo modelo presenta una precisión aceptable, el modelo logístico anterior demostró ser más efectivo al predecir los verdaderos positivos. Estos hallazgos enfatizan la importancia de seleccionar cuidadosamente las variables y el modelo al abordar problemas de clasificación en contextos económicos.