Trabajo Práctico 3

Otoño 2024

Big Data

Integrantes: Martina Murga, Valentina Benitez, Benicio García Oliver

Parte I: Analizando la base

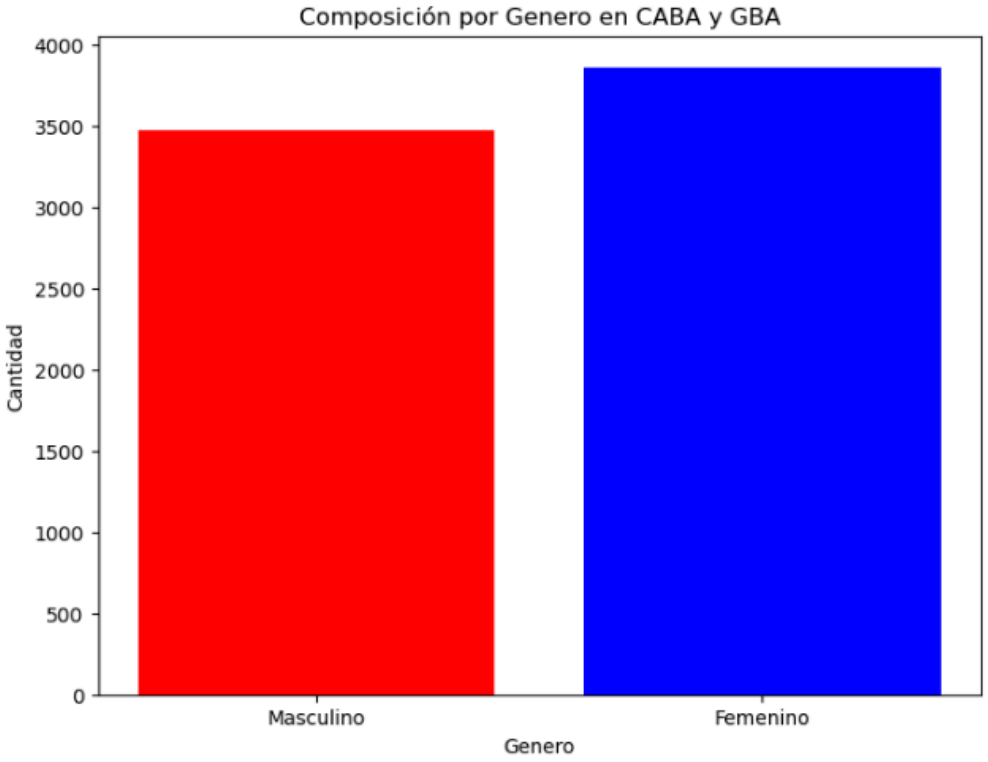
1.

La noción de pobreza como entiende el indec, sigue el método de medición indirecta, denominado también como "línea". La "Línea de Pobreza"[1] Incluye y toma en cuenta no solo los consumos alimentarios mínimos sino también otros consumos básicos no alimentarios que son considerados necesarios. La suma de estos consumos básicos conforma la Canasta Básica Total (CBT) esta canasta es también contrastada con los datos encontrados en la EPH. Entonces, para lograr calcular la línea de pobreza se necesita la canasta básica alimentaria(CBA) y añadir los bienes y servicios no alimentarios (transporte, educación, salud, etc) para llegar a la canasta básica total(CBT). Luego la "Línea de Pobreza" se construye para cada hogar dependiendo de su tamaño y composición y se contrasta con el ingreso del hogar. Si el ingreso del hogar es menor a la línea de pobreza construida para este, entonces estos estarán debajo de la línea de pobreza y se considerará personas pobres.

2.

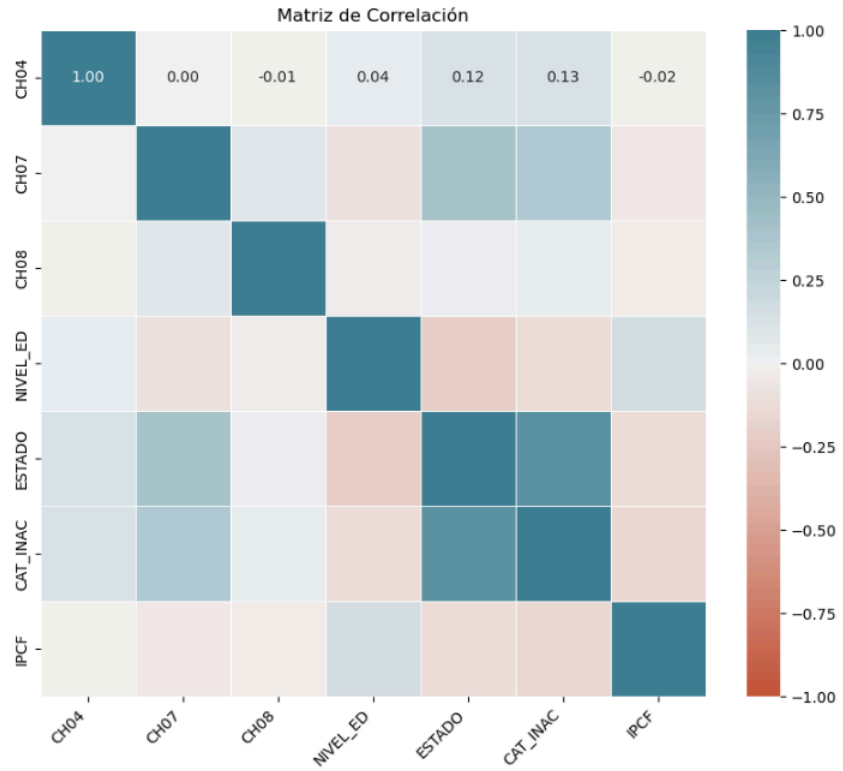
a. En Notebook
b. El enunciado pide eliminar las observaciones con valores que no tienen sentido y que son absurdas, dado esto eliminamos las observaciones relacionadas a ingreso y edad que son negativas. También se eliminan valores negativos relacionados con observaciones provenientes del capítulo: ¿Cuánto tiempo hace que trabaja allí? (en la casa que tiene más horas) que también sería absurdo encontrarse un número negativo. O observaciones como la de ¿En cuántas casas trabaja? (cantidad)(PP04B2) que tampoco tendría sentido encontrarse un número negativo. Eliminar estas observaciones absurdas es importante dado que nos brindará una mayor calidad y robustez al modelo predictor que armaremos después. Esto porque datos inválidos o absurdos pueden sesgar los resultados del modelo y así afectar la precisión de la predicción. Además eliminando estas observaciones corregimos la base de datos que pudo haber tenido errores de tipeo o relacionados a la recolección de los datos. Nos servirá solucionar estos problemas para continuar con los incisos que siguen.
c.

[1] https://www.indec.gob.ar/indec/web/Nivel3-Tema-4-46



Luego de haber realizado la limpieza de la base y habiéndonos quedado con las observaciones útiles de los aglomerados de Ciudad Autónoma de Buenos Aires y del Gran Buenos Aires, procedimos a hacer un grafico de barras mostrando la composicion por sexo de la base de datos. Lo que encontramos en el gráfico es que hay un mayor número de observaciones de género Femenino que del masculino, del género Femenino se encontraron 3860 observaciones y del Masculino 3470. Esto encaja con la proporción de hombres y mujeres que viven en el país y en el mundo, que según el indec en el total del país hay 94,8 hombres por cada 100 mujeres. Entonces habiendo hecho los cambios de incisos anteriores a la base de datos original, no se daño esta proporción. Esto brinda mayor precisión al predecir valores de ingreso o pobreza en incisos a continuación.

d.



	CH04[2]	CH07	CH08	NIVEL_ED	ESTADO	CAT_INAC	IPCF
CH04	1.000000	0.000501	-0.009665	0.037974	0.120873	0.128221	-0.022446
CH07	0.000501	1.000000	0.072854	-0.101498	0.399829	0.342766	-0.049593
CH08	-0.009665	0.072854	1.000000	-0.031640	0.026661	0.055910	-0.042539
NIVEL_ED	0.037974	-0.101498	-0.031640	1.000000	-0.214697	-0.118471	0.151536
ESTADO	0.120873	0.399829	0.026661	-0.214697	1.000000	0.827529	-0.138020
CAT_INAC	0.128221	0.342766	0.055910	-0.118471	0.827529	1.000000	-0.151449
IPCF	-0.022446	-0.049593	-0.042539	0.151536	-0.138020	-0.151449	1.000000

Como podemos ver en la matriz de correlación, la mayoría de las variables tenidas en cuenta no están correlacionadas entre ellas, a excepción de las variables de ESTADO y CAT_INAC que tienen una correlación alta de 0.827. La variable estado muestra el estado laboral en el que se encuentra el encuestado, toma valor 1 si está ocupado, valor 2 si está desocupado, valor 3 si está inactivo o valor 4 si es menor a 10 años. La variable CAT_INAC muestra la categoría de inactividad, toma valor 1 si es

[2] https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_4T2023.pdf

jubilado, 2 si es rentista, 3 si es estudiante, 4 si es ama de casa, etc. Podemos entender como claramente estas variables están correlacionadas y una de ellas no aporta al modelo, podríamos decir que es mejor eliminar la variable CAT_INAC. ¿Qué problemas pueden traer para el modelo que las variables estén correlacionadas entre sí? Un problema de multicolinealidad, si sus variables están muy correlacionadas como en este caso con las dos variables explicadas anteriormente, puede resultar en un problema de multicolinealidad que dañara la precisión del modelo dado que puede aumentar la varianza y hacer difícil el estudio de la interpretación de los coeficientes. Por suerte no hay una gran cantidad de variables correlacionadas haciendo que el modelo no sufra tanto este problema.

e.

Número de ocupados: 3490
Número de desocupados: 240
Número de inactivos: 2765

Media de IPCF según estado:	
Ocupados	132041.486499
Desocupados	58012.198417
Inactivos	84993.676882

5.

Luego de realizar la comparación del ITF y del ingreso que necesita esa familia para no caer debajo de la línea de la pobreza, llegamos al resultado de que hay 2196 personas/hogares que se encuentran debajo de la línea de la pobreza y son considerados pobres.

Parte II: Clasificación

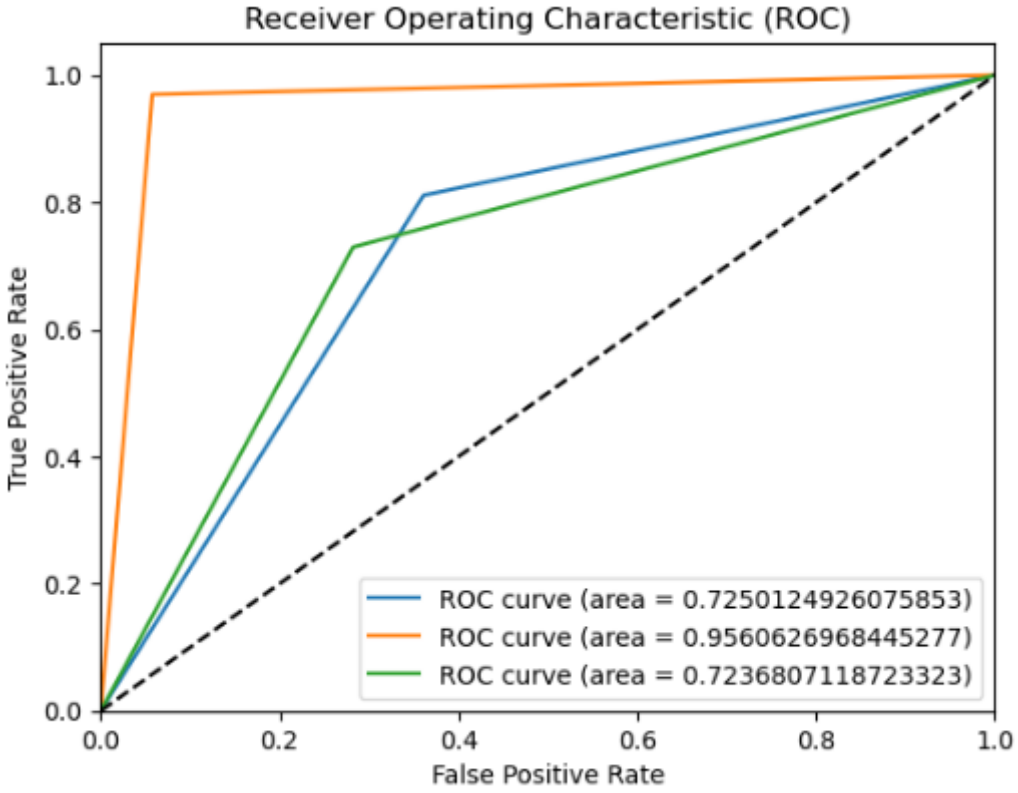
1. En notebook

2. En notebook

3.

Resultados:

	Regresión Logística	Análisis de discriminante lineal	KNN (k=3)
Matriz de Confusión	TP = 537, FP = 238, TN = 421, FN = 125	TP = 642, FP = 38, TN = 621, FN = 20	TP = 483, FP = 186, TN = 473, FN = 179
AUC	0.7250124926075853	0.9560626968445277	0.7236807118723323
Accuracy	0.7252081756245269	0.9560938682816048	0.7236941710825132



4.

Las medidas de precisión que conocemos son, la matriz de confusión, la curva ROC y los valores de AUC y de Accuracy de cada uno de los métodos que implementamos. La matriz de confusión nos brinda datos sobre los resultados de la clasificación en términos de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Como buscamos minimizar los errores de clasificación, buscamos modelos con menos FP y FN. En el caso de el AUC, que es el área bajo la curva ROC, buscamos que este sea lo más cercano posible a 1. En el caso del Accuracy, este mide la proporción de predicciones correctas realizadas por el modelo.

Teniendo en cuenta lo explicado y los datos obtenidos en el inciso anterior, podemos concluir que el mejor modelo en este caso es el Análisis Discriminante

Lineal, seguido por la Regresión Logística(logit) y por último el KNN(k = 3). Esto debido a que el Análisis Discriminante Lineal presenta el mayor nivel de accuracy y AUC ambos cercanos a 0.96 comparado a los demas metodos, y tambien ,como podemos notar en su matriz de confusión, tiene menos cantidad de falsos positivos y falsos negativos en comparación con los otros métodos. Esto nos ayuda a concluir que es el mejor método a implementar en este caso.

5.

Los resultados nos dicen que la proporción de las personas que no respondieron y pudieron ser identificadas como pobres es de 0,26; es decir un 26%.

6.

Para correr los tres métodos se utilizaron todas las variables disponibles como predictores, esto puede ser ventajoso cómo puede perjudicar al modelo y su predicción. En primer lugar, utilizar todas variables disponibles nos brinda la ventaja de no tener problemas de omisión de variables relevantes y tener la maxima informacion posible para generar la predicción, ademas no eliminar variables nos brinda simplicidad al crear el modelo y nos ahorra tiempo de filtrar todas las variables para usar sólo las relevantes.

Pero por otro lado usar todas las variables de la base de datos no nos parece correcto, dado distintos problemas claros que puede traernos esto. Uno de ellos es problemas de multicolinealidad, esto significa que al usar tantas variables relacionadas a temas similares muchas de estas pueden estar correlacionadas y traernos problemas al implementar el modelo. También podría traernos problemas de overfitting, esto significa que incluir todas las variables nos puede hacer llegar a un modelo que se ajuste demasiado bien a los datos de entrenamiento, haciendo que éste capture ruido en vez del patrón real de interés. Esto obviamente disminuye la capacidad del modelo a generalizar sus resultados en datos nuevos. Además, si se incluyen variables irrelevantes o muy redundantes y similares a otras, esto puede causar el empeoramiento del modelo ya que añadirá ruido en vez de información útil.

Dado esto, selecciona variables de interés que no se correlacionan mucho entre ellas, como la edad, el sexo, variables de empleo(pero no todas, solo las mas relevantes). Intentaría incluir las variables no correlacionadas, que entienda brinden información útil y no mayor ruido al modelo.

Nuestro código arroja los siguientes resultados:

- Precisión del modelo con todas las variables: 0.88
- AUC del modelo con todas las variables: 0.92
- Precisión del modelo con variables seleccionadas: 0.88
- AUC del modelo con variables seleccionadas: 0.92
- Diferencia en precisión: 0.01

    - Diferencia en AUC: 0.00

- Entrenamiento con todas las variables: Entrenamos un modelo de regresión logística utilizando todas las variables disponibles y calculamos las medidas de precisión (accuracy y AUC).
- Entrenamiento con variables seleccionadas: Entrenamos un segundo modelo utilizando solo un subconjunto de variables seleccionadas y calculamos las mismas medidas de precisión.
- Comparación de resultados: Comparamos las medidas de precisión de ambos modelos.
- Predicción en *norespondieron*: Utilizamos el modelo con las variables seleccionadas para hacer predicciones en la base *norespondieron.*