Trabajo Práctico 4

Otoño 2024

Big Data

Integrantes: Martina Murga, Valentina Benitez, Benicio García Oliver

<u>Parte I:</u> Análisis de la base de hogares y cálculo de pobreza

1)Las variables en la base de hogares de la EPH incluyen información sobre características del hogar y sus miembros. Algunas variables que están presentes y que podrían ser muy útiles para predecir la pobreza incluyen:

- Variables relacionadas al ingreso del hogar: ITF, DECIFR, IPCF, etc. Estás porque el nivel de ingresos es un indicador directo del bienestar económico de un hogar. Los hogares con ingresos más bajos son más propensos a estar en situación de pobreza.
- Variables relacionadas al número de miembros por hogar: IX_Tot, IX_Men10, IX_Mayeq10, etc. Esto porque un mayor número de miembros puede indicar una mayor carga económica para el hogar, especialmente si hay muchos dependientes.
- La variable asignada al nivel educativo del jefe del hogar o de los miembros del hogar:Esto dado que la educación está fuertemente correlacionada con las oportunidades de empleo y los niveles de ingreso.
- Variables acerca de las condiciones en la que se encuentra la vivienda: Las condiciones de vivienda son un indicador de la calidad de vida y pueden reflejar la capacidad económica del hogar. Y esto puede afectar a varios factores influyentes sobre la pobreza, por ejemplo la educación, ingreso, salud, etc.

3) La limpieza de la base de datos mostrada en el código se justifica por las siguientes razones:

1. Filtrado de variables binarias para valores válidos (1 o 2):
    - Justificación: Este paso asegura que las variables binarias (como 'V1', 'V2', 'V3', etc.) solo contengan valores válidos, es decir, 1 o 2. Esto es crucial para evitar errores en el análisis posterior debido a valores fuera de rango o erróneos.
2. Eliminación de columnas que no aportan valor:
    - Justificación: Al eliminar columnas que no aportan valor, como 'IV1_ESP', 'IV3_ESP', 'IV7_ESP', etc., se simplifica el conjunto de datos y se reduce la carga computacional. Estas columnas pueden contener información irrelevante.
3. Filtrado de ingresos, edad, horas trabajadas, cantidad de habitantes en el hogar y habitaciones:
    - Justificación: Este paso asegura que las variables clave (como 'IPCF', 'CH06', 'PP3E_TOT', etc.) solo contengan valores dentro de un rango lógico. Por ejemplo, asegurar que los ingresos sean mayores o iguales a cero, que las horas trabajadas no excedan un límite razonable, y que la cantidad de habitantes en el hogar sea coherente con el número de habitaciones. Este tipo de filtrado es esencial para prevenir valores atípicos o errores de entrada que puedan sesgar el análisis y conducir a conclusiones incorrectas.

5) Estadísticas descriptivas de cinco variables relevantes para predecir pobreza de la encuesta de hogar:

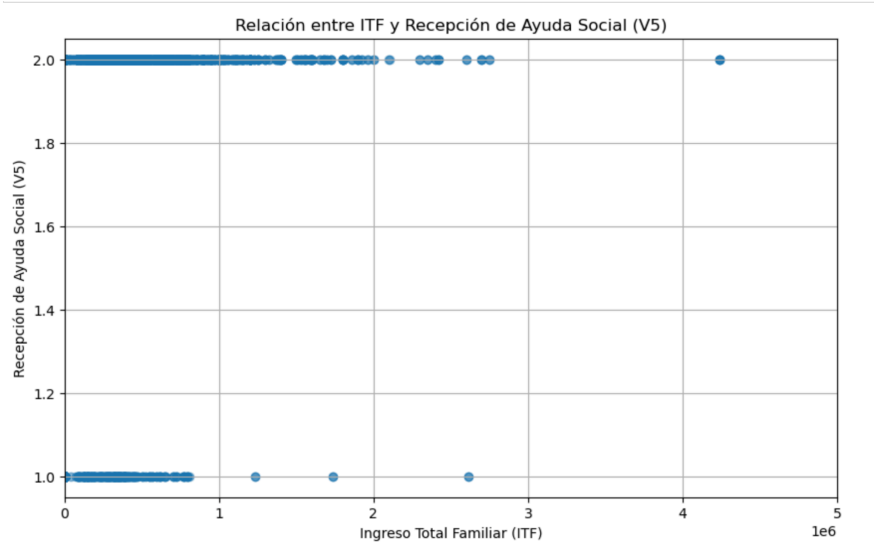
```
Estadísticas Descriptivas:
       habitantes_por_cuarto  ingreso_por_adulto          IPCF     PP3E_TOT  \
count            3443.000000        3.443000e+03  3.443000e+03  3443.000000
mean                     inf        1.453690e+05  1.314657e+05    36.579437
std                      NaN        2.101405e+05  1.960944e+05    17.548398
min                 0.333333        0.000000e+00  0.000000e+00     0.000000
25%                 1.333333        0.000000e+00  0.000000e+00    24.000000
50%                 1.500000        8.666667e+04  7.333333e+04    40.000000
75%                 2.000000        2.150667e+05  2.000000e+05    48.000000
max                      inf        2.100000e+06  2.100000e+06   133.000000

       asistencia_familiar
count          3443.000000
mean           4008.039791
std           15738.100568
min              -9.000000
25%               0.000000
50%               0.000000
75%               0.000000
max          166000.000000
```

- habitantes_por_cuarto: Tiene 3443 observaciones. La media es infinita y la desviación estándar no está disponible, indicando posibles problemas en los datos. El valor mínimo es 0.33, el percentil 25 es 1.33, la mediana es 1.5, el percentil 75 es 2 y el máximo es infinito.
- ingreso_por_adulto: Con 3443 observaciones, la media es 145,369 con una desviación estándar de 210,140. El valor mínimo es 0, el percentil 25 es 0, la mediana es 86,667, el percentil 75 es 215,067 y el máximo es 2,100,000.
- IPCF: También con 3443 observaciones, tiene una media de 131,465 y una desviación estándar de 196,094. El valor mínimo es 0, el percentil 25 es 0, la mediana es 73,333, el percentil 75 es 200,000 y el máximo es 2,100,000.
- PP3E_TOT: Tiene una media de 36.58 y una desviación estándar de 17.55. El valor mínimo es 0, el percentil 25 es 24, la mediana es 40, el percentil 75 es 48 y el máximo es 133.
- asistencia_familiar: Cuenta con 3443 observaciones, una media de 4008.04 y una desviación estándar de 15,738.1. El valor mínimo es -9, el percentil 25 es 0, la mediana es 0, el percentil 75 es 0 y el máximo es 166,000.

En resumen hay posibles problemas de datos en "habitantes_por_cuarto" y una gran variabilidad en "ingreso_por_adulto" y "asistencia_familiar".

6)



El gráfico ilustra la relación entre el Ingreso Total Familiar (ITF) y la Recepción de Ayuda Social (V5). El eje horizontal muestra el ITF, mientras que el eje vertical indica la recepción de ayuda social, con valores de 1 (no recibe ayuda) y 2 (recibe ayuda).

La mayoría de los puntos están en el nivel de ingreso más bajo, lo que sugiere que las familias con menores ingresos son las que más reciben ayuda social (valor 2 en el eje vertical). A medida que los ingresos aumentan, la cantidad de familias que reciben ayuda social disminuye notablemente. La concentración de puntos en la parte inferior izquierda indica una fuerte correlación negativa: a menores ingresos, mayor es la probabilidad de recibir ayuda social, y viceversa. Hay pocos puntos dispersos en los niveles de ingreso altos, lo que indica que muy pocas familias con ingresos elevados reciben ayuda social.

Parte III: Clasificación y Regularización

3) Para elegir λ por validación cruzada, seguiríamos estos pasos:

1. División de datos: Dividimos el conjunto de datos en k partes iguales (e.g., 5 o 10) para realizar k-fold cross-validation.
2. Entrenamiento y validación: Para cada valor de λ y cada una de las kk particiones:
    - Entrenamos el modelo con k−1 particiones.
    - Evaluamos el modelo en la partición restante.
3. Promedio de errores: Calculamos el error promedio (e.g., precisión, AUC) en las k particiones para cada λ.
4. Selección de λ: Elegimos el valor de λ que minimiza el error promedio.

No usaríamos el conjunto de prueba para la elección de λ porque el conjunto de prueba debe usarse exclusivamente para evaluar el rendimiento final del modelo, no para su ajuste o selección de hiper parámetros. Además, usar el conjunto de prueba para elegir λ introduciría sesgo y llevaría a una sobreestimación del rendimiento del modelo en datos no vistos.

4) En validación cruzada, usar un k pequeño tiene muchas implicancias, tales como: varianza alta: Los modelos entrenados pueden tener una alta variabilidad entre las diferentes particiones de los datos, lo que puede llevar a una estimación inestable del rendimiento. Luego, hay sesgo bajo: Cada modelo se entrena con una gran cantidad de datos, lo que generalmente produce modelos menos sesgados.

Usar un k grande tiene implicancias. La primera es que hay una varianza baja: Las particiones son más pequeñas, lo que puede llevar a una estimación más estable del rendimiento. La segunda es que hay un sesgo alto: Cada modelo se entrena con menos datos, lo que puede producir modelos más sesgados.

Cuando k=n (Leave-One-Out Cross-Validation, LOOCV) tenemos implicancias. En primer lugar, hay una varianza muy baja: Cada partición consiste en un solo punto de prueba y el resto de los puntos de datos como entrenamiento, lo que produce una estimación muy estable. En segundo lugar, hay un sesgo muy alto: Cada modelo se entrena con casi todos los datos, pero se prueba en un solo punto, lo que puede hacer que el modelo sea muy específico para el conjunto de datos de entrenamiento. En tercer lugar, es computacionalmente costoso: Se estima el modelo nn veces, donde nn es el número de muestras, lo que puede ser muy costoso computacionalmente para conjuntos de datos grandes.

En resumen, un k pequeño puede dar lugar a estimaciones con alta varianza y bajo sesgo, mientras que un k grande tiende a dar estimaciones con baja varianza y alto sesgo. Cuando k=n, el modelo se estima n veces, una por cada muestra del conjunto de datos.

5) En este ejercicio tuvimos problemas haciendo que nuestras computadoras corran el código dada la cantidad de datos, nuestras computadoras tardaron mucho tiempo y no llegamos a ver los resultados. Igualmente continuamos haciendo el código de los ejercicios 6 y 7, y en el ejercicio 9 elegimos utilizar lasso para predecir qué personas son pobres. Entendemos que el código de los incisos 6, 7 y 9 están bien pero no podemos corroborarlo dado que el código del ejercicio 5 tarda demasiado en correr y no sabemos cómo solucionarlo.