Materia: Taller de Programación
Grupo 2: Jimena Teran, Juan Lynch, Renzo Falciglia
Fecha: 16/12/2024

Trabajo Práctico 4

Parte I: Análisis de la base de hogares y tipo de ocupación

Punto 1

Para perfeccionar el ejercicio del TP3, sería útil incluir variables que puedan predecir la desocupación. Algunas de estas variables se relacionan con la composición familiar y las características del hogar, y se pueden encontrar en la base de Hogares de la EPH. Las que seleccionamos para este trabajo son las siguientes:

- **ITF (Ingreso total familiar):** esta variable representa el monto total del ingreso familiar. Es lógico pensar que hogares con un ITF bajo tienen mayor probabilidad de estar en situación de pobreza.
- **IPCF (Ingreso per cápita familiar):** esta variable representa el ingreso promedio por miembro del hogar. Un IPCF bajo puede ser un indicador de desempleo, especialmente en hogares con muchos miembros.
- **DECIFR (Decil del ingreso total familiar):** esta variable ubica al hogar en una escala de 10 grupos, donde el decil 1 representa a los hogares con menores ingresos y el decil 10 a los de mayores ingresos. Los hogares en los deciles más bajos tienen mayor probabilidad de ser pobres y sus habitantes de estar desempleados.
- **DECCFR (Decil del ingreso per cápita familiar):** similar a DECIFR, pero utilizando el ingreso per cápita. También, los hogares en los deciles inferiores de esta variable tienen más probabilidad de tener habitantes en situación de desempleo.
- **Características de la vivienda:** Variables como el tipo de vivienda (IV1) y la cantidad de ambientes (IV2) sugieren la situación económica y por lo tanto laboral de sus ocupantes.
- **Características habitacionales del hogar:** variables como la cantidad de ambientes para uso exclusivo del hogar (II1) y la cantidad de ambientes utilizados para dormir (II2) también son útiles ya que dan información sobre la situación económica de sus ocupantes.
- **Estrategias del hogar:** las variables V1 a V19 reflejan las estrategias que los hogares utilizan para afrontar sus necesidades económicas., como por ejemplo la necesidad de recurrir a ayudas sociales (V5).
- **Organización del hogar:** variables como la cantidad de miembros del hogar (IX_Tot) también son útiles para analizar la situación económica.

Punto 2 y 3

En estos incisos se procesaron las bases de hogar de 2004 y 2024 y se las unió al df de las bases individuales del TP3. Se utilizaron las variables de CODUSU y NRO_HOGAR para realizar el merge. Luego, se realizaron las siguientes transformaciones:

A. Transformación de la columna 'aglomerado':

- Se eliminó la columna llamada aglomerado.y.
- La columna aglomerado.x fue renombrada como “aglomerado”.

B. Transformación de la variable 'IV1' (tipo de vivienda):

- Se asignaron valores numéricos a la columna IV1 de acuerdo con las categorías, para homogeneizar entre las dos bases:
    - Departamento → 2
    - Casa → 1
    - Pieza de inquilinato → 3
    - Pieza en hotel/pensión → 4
    - Otro → 5

C. Transformación de la variable 'IV5' (recibe algún tipo de subsidio):

Se asignaron valores numéricos a la columna IV5 de acuerdo con las categorías:

- Si → 1
- No → 2
- Ns./Nr. → 9

D. Transformación de las columnas 'DECIFR' (decil del ingreso total del hogar del total EPH) y 'DECCFR' (decil del ingreso per cápita familiar del total EPH):

- Eliminación de ceros iniciales: se quitaron los ceros iniciales de los valores en las columnas DECIFR y DECCFR.
- Conversión a formato numérico: las columnas DECIFR y DECCFR fueron convertidas a tipo *float64* después de eliminar los ceros iniciales.
- Cualquier valor vacío o no numérico en las columnas DECIFR y DECCFR fue reemplazado por NaN, y luego se convirtió en 0 si era necesario.

Punto 4

Se crearon 3 variables a partir de la información de la encuesta individual y la combinación con la encuesta de hogares. Dichas variables ayudan a predecir si una persona está desempleada o no. Las variables creadas son:

- **mayor_garage:** si la persona es mayor a 50 años y vive en una casa con garage se asigna un 1, sino un 0. Esta variable fue elegida ya que es probable que las personas mayores que viven en una casa con garage sean propietarias y tengan un auto, lo que implica que tienen más probabilidades de tener un empleo.

- **mujer_emergencia:** si la persona es mujer y vive en una villa de emergencia esta variable toma un 1, sino un 0. Esta variable fue elegida porque se espera que las mujeres tengan mayores dificultades para estar empleadas, y mayores dificultades aún si su lugar de residencia es una villa de emergencia.

- **mujer_hogar:** si la persona es mujer y vive en un hogar con más de 3 personas, esta variable toma un 1, sino un 0. Esta variable fue elegida porque suelen recaer sobre las mujeres las tareas de cuidado de otros miembros del hogar, por lo tanto, las mujeres que viven en hogares con más miembros tienen menores probabilidades de tener un empleo, porque se dedican a las tareas de cuidado.

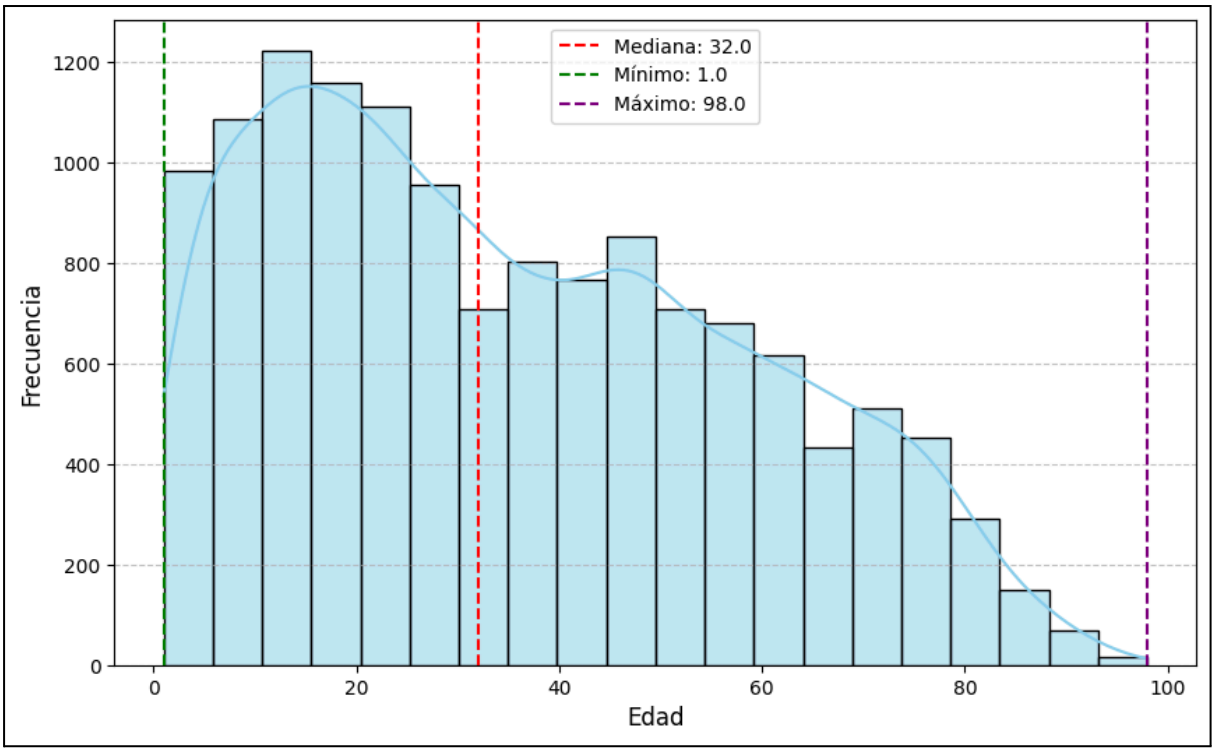
Punto 5

En este ejercicio se presentan diferentes estadísticas descriptivas de tres variables: edad, nivel educativo y tipo de hogar.

Edad

La variable CH06, que representa la edad, muestra una distribución con un rango entre 1 año (mínimo) y 98 años (máximo). La mediana de 32 años sugiere que la población está equilibrada alrededor de esta edad. La desviación estándar indica una dispersión moderada en los datos. Los gráficos confirman que la mayoría de las personas están concentradas entre ciertos rangos de edad, y se identifican posibles valores atípicos en los extremos. A continuación se presenta un histograma con la distribución de la variable:

Gráfico 1 - Histograma variable edad

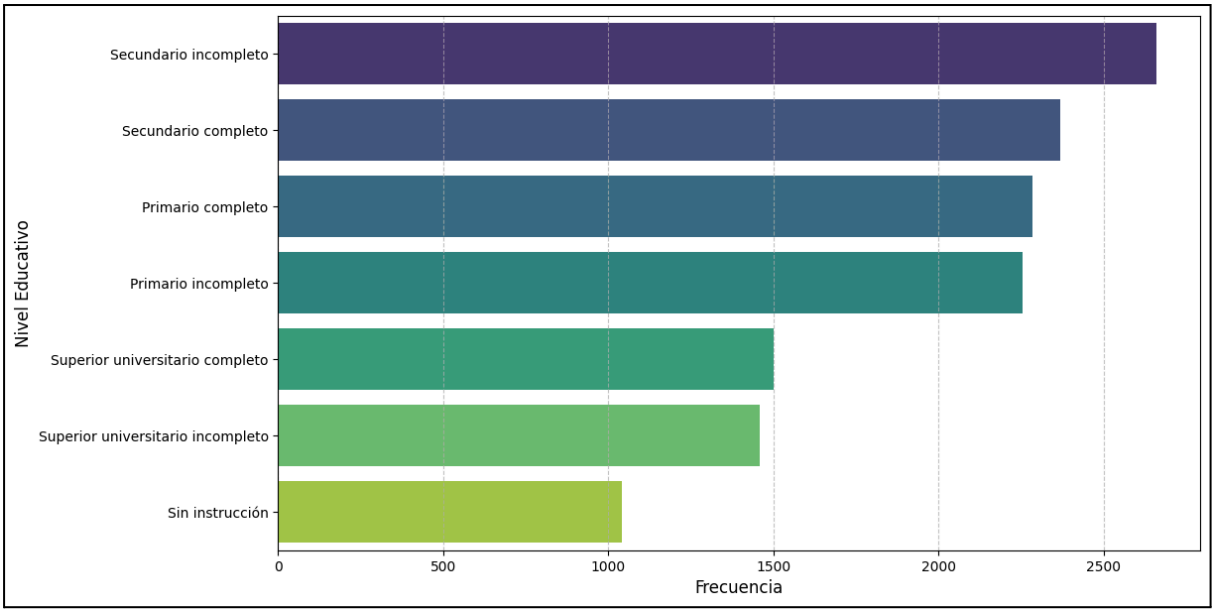


Nivel educativo

La variable NIVEL_ED, que representa el nivel educativo, revela una distribución que destaca las siguientes observaciones principales:

- **Categorías predominantes:** las categorías con mayor representación son la de "secundario incompleto" y la de "secundario completo", indicando que una proporción significativa de la población ha alcanzado estos niveles educativos.
- **Categorías menos frecuentes:** las categorías como "sin instrucción" o "primaria incompleta" tienen menor representación. En conjunto estas dos categorías representan casi el 25% de la muestra, lo que indica que el 75% de la población alcanzó un nivel educativo al menos de primaria completa. La categoría "sin instrucción" representa el 7% de la muestra.

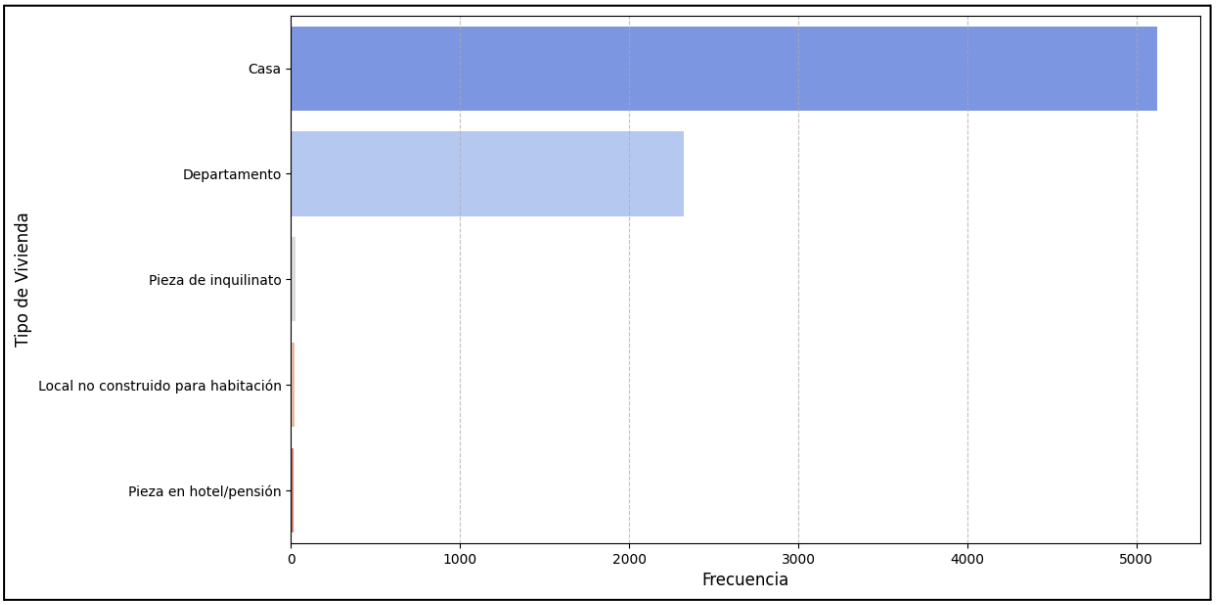
Gráfico 2 - Frecuencia de Nivel Educativo



Tipo de vivienda

Del análisis de los tipos de vivienda se desprende que la gran mayoría de las personas viven en una casa o un departamento, lo cuál es lógico por el tipo de país que es Argentina. En particular, el 68% de la muestra vive en casa, mientras que el 30% vive en departamento. En conjunto, el 98% de la muestra vive en alguno de estos dos tipos de vivienda.

Gráfico 3 - Frecuencia de Tipo de Vivienda



Parte II: Clasificación y regularización

Punto 1

Para la implementación del modelo de clasificación se procedió a dividir las bases de datos de los años 2004 y 2024 en conjuntos de entrenamiento y prueba. Se utilizó una proporción de 70% para entrenamiento y 30% para prueba, estableciendo random_state=101 para asegurar reproducibilidad. La variable dependiente 'desocupado' se construyó como una variable binaria que toma valor 1 cuando ESTADO es igual a 2 y 0 en caso contrario. Las variables independientes seleccionadas incluyen las características tanto individuales como del hogar, incorporando una columna de unos según lo solicitado para el término constante en la estimación.

Punto 2 y 3

Para la selección del parámetro λ por validación cruzada se procedió a dividir los datos de entrenamiento en k partes. Este método consiste en estimar el modelo k veces, utilizando en cada iteración k-1 partes para entrenar y la restante para validar, rotando sistemáticamente qué parte se utiliza para validación. Para cada valor candidato de λ se calcula el error promedio de validación y se selecciona el valor que minimice dicho error. La decisión de no utilizar el conjunto de prueba para esta elección se fundamenta en la necesidad de mantener estos datos completamente separados para la evaluación final del modelo, evitando así obtener una estimación sesgada del verdadero error de predicción.

La elección del número de partes k en la validación cruzada presenta diferentes consideraciones prácticas. Un k pequeño permite maximizar los datos disponibles para estimación en cada iteración, pero hace que los resultados sean más sensibles a qué observaciones quedaron en cada grupo. Por otro lado, un k grande brinda una evaluación más robusta al probar el modelo en más grupos diferentes, pero cada estimación utiliza menos datos para entrenar. En el caso particular donde k es igual al número de observaciones (leave-one-out), se debe estimar el modelo n veces, usando n-1 observaciones para entrenar y una para validar en cada iteración, lo cual resulta computacionalmente costoso dadas las dimensiones de nuestra base de datos.

Punto 4

Se implementó regresión logística con dos tipos de regularización (L1-LASSO y L2-Ridge) utilizando λ=1. A diferencia del TP3, donde se utilizaron principalmente variables individuales (AGLOMERADO, CH04, CH06, CH07, CH08, NIVEL_ED, CAT_INAC, IPCF, P21), en este trabajo se incorporaron variables del hogar como DECIFR, DECCFR, IV1, II1, II2, V5, IX_TOT, IV12_3, II4_3 y V1.

Para 2004, ambos métodos de regularización mostraron resultados similares entre sí pero inferiores al TP3. El modelo LASSO obtuvo un AUC de 0,897 y Ridge de 0,895, con una precisión de 70,6% y 68,8% respectivamente. Sin embargo, la sensibilidad fue notablemente baja (15,7% LASSO, 14,4% Ridge), indicando una dificultad significativa para identificar desocupados. En comparación, el modelo sin regularización del TP3 logró identificar mejor

los casos de desempleo, con 149 verdaderos positivos frente a solo 24 y 22 en los modelos regularizados.

Para 2024, aun con las nuevas variables del hogar, la performance de los modelos regularizados empeoró significativamente. Tanto LASSO como Ridge mostraron una precisión reducida (40% y 33,3% respectivamente) y una sensibilidad extremadamente baja (2,17% en ambos casos), aunque mantuvieron un AUC alto de 0,890. El modelo sin regularización del TP3 nuevamente mostró mejor capacidad para identificar desocupados, con 83 verdaderos positivos comparados con solo 2 en los modelos regularizados.

La incorporación de regularización y nuevas variables del hogar resultó en modelos más conservadores en la clasificación de desocupados, produciendo más falsos negativos. Esta performance inferior sugiere que, para este caso particular, ni la regularización ni las variables adicionales del hogar mejoraron la capacidad predictiva del modelo original del TP3.

Punto 5

Se realizó un barrido de valores para el parámetro λ utilizando **10-fold cross-validation** en ambos modelos de regresión logística regularizada: **L1-LASSO** y **L2-Ridge**. Para cada valor de λ, se calculó el error medio de validación (MSE) y se seleccionó el λ que minimizó este error.

Resultados

1. **LASSO:**
    - El λ óptimo encontrado fue **log10(λ) ≈ -5**, lo que indica una regularización débil.
    - Con este λ, LASSO no descartó ninguna variable en ninguno de los años (2004 y 2024), ya que todas las variables predictoras fueron consideradas relevantes por el modelo.
    - La distribución de errores mostró que a medida que λ aumentaba, el error de predicción crecía significativamente debido a una mayor penalización.

2. **Ridge:**
    - También mostró un λ óptimo de **log10(λλ) ≈ -5**.
    - En este rango, Ridge presentó un menor MSE en 2024 (0.052) en comparación con 2004 (0.063), lo que sugiere que los datos más recientes tienen patrones más claros para la predicción.

3. **Relación entre λ y el modelo:**
    - En ambos casos, un λ mayor llevó a modelos más simples pero con peor desempeño, reflejado en un aumento del MSE. Esto se evidenció en la gráfica de distribución de errores para diferentes valores de λ.

Punto 6

En este punto, se buscó identificar las variables descartadas por el modelo de **regresión logística LASSO (penalización L1)** utilizando el valor óptimo de λ obtenido en el barrido realizado en el Punto 5. Sin embargo el resultado fue que LASSO **no descartó ninguna variable** en 2004 ni en 2024. Esto indica que todas las variables predictoras se consideran relevantes para el modelo.

Esto es consistente con la expectativa de que las variables seleccionadas en el inciso 1 del Punto I (como **DECIFR, DECCFR, IPCF, IV1, II1, V5**, entre otras) aportan información útil para predecir la desocupación.

Punto 7

En este punto, se compararon los resultados de los modelos de **regresión logística regularizada** entre los años 2004 y 2024, evaluando el rendimiento y las diferencias en la selección de predictores con los distintos métodos de regularización (**Ridge y LASSO**). También se analizó cuál de los métodos de regularización funcionó mejor.

1. **Rendimiento:**
    - **Ridge:** Mejor ajuste en 2024 (MSE: **0.052**) que en 2004 (MSE: **0.063**).
    - **LASSO:** MSE de **0.063** en 2004 y sin mejora significativa en 2024.
2. **Selección de variables:**
    - LASSO no descartó ninguna variable en ninguno de los años, indicando que todas eran relevantes para el modelo.
3. **Mejor método:**
    - Ridge tuvo un ajuste ligeramente mejor en 2024, pero las diferencias no fueron significativas.

Conclusión: Ambos métodos tuvieron un desempeño similar, pero ninguno mejoró sustancialmente al modelo sin regularización del TP3.