Trabajo práctico #4

Taller de programación
Maestría en Economía Aplicada
Universidad de Buenos Aires

Granly Jiménez

15 de diciembre de 2024

Introducción

La Encuesta Permanente de Hogares (EPH) es uno de los instrumentos más relevantes del sistema estadístico nacional, desarrollado por el Instituto Nacional de Estadística y Censos (INDEC). Este programa permite obtener información continua y sistemática sobre las características sociodemográficas, económicas y laborales de la población argentina. Entre los indicadores clave que producen, destaca la tasa de desocupación, un dato esencial para analizar el desempeño del mercado laboral y su relación con los ciclos económicos y las políticas públicas.

Desde un enfoque metodológico, la EPH sigue estándares internacionales, como los propuestos por la Organización Internacional del Trabajo (OIT), para clasificar a las personas en ocupadas, desocupadas e inactivas. Estos indicadores no solo reflejan la situación actual del empleo, sino que también permiten realizar análisis longitudinales para evaluar cómo cambian las dinámicas laborales en función de factores sociodemográficos, educativos y territoriales.

Objetivo del trabajo

El objetivo de este trabajo es abordar un análisis descriptivo y empírico de la información contenida en las bases de microdatos de la EPH correspondientes a los años 2004 y 2024. Se busca, en primer lugar, describir y depurar los datos, identificando patrones y tendencias clave en el mercado laboral, con énfasis en las tasas de desocupación y la composición de la población económicamente activa (PEA). En segundo lugar, se propone construir modelos de clasificación para predecir el estado de desocupación de los individuos, evaluando cómo estas predicciones varían entre ambos años. Finalmente, se explorará cómo las diferencias en las tasas de desocupación pueden explicarse a través de factores individuales, educativos y etarios, vinculando estos resultados a debates teóricos en economía laboral.

Tasa de ocupación en Argentina

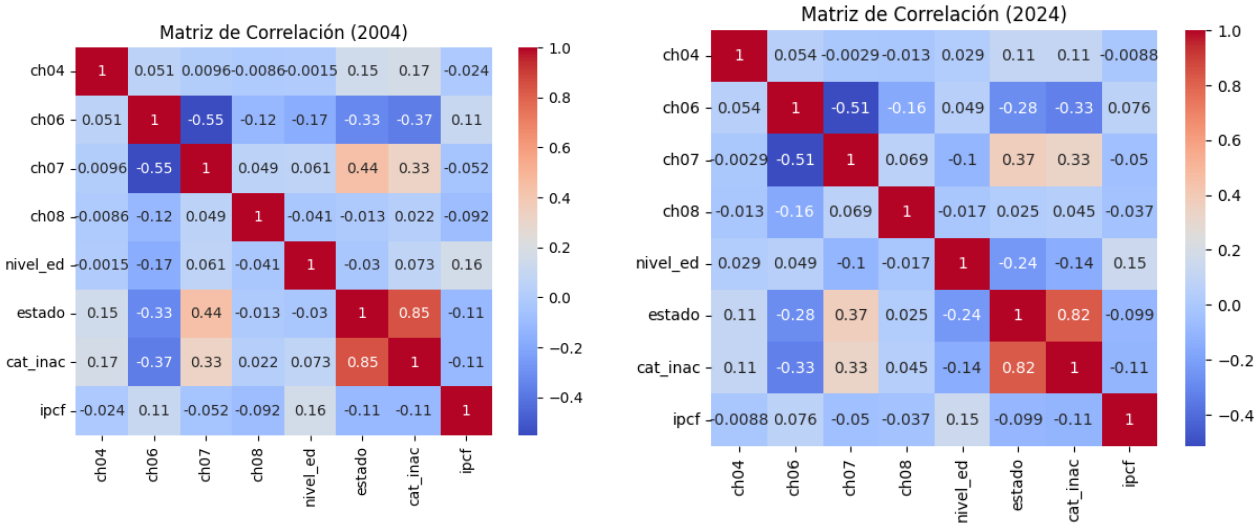
Según el INDEC, una persona desocupada es aquella que, durante la semana de referencia de la Encuesta Permanente de Hogares (EPH), no tiene trabajo remunerado, está disponible para trabajar y ha realizado una búsqueda activa de empleo en las últimas cuatro semanas. Metodológicamente, esta definición se basa en los lineamientos de la Organización Internacional del Trabajo (OIT) y se obtiene mediante preguntas estructuradas en la EPH, que distinguen entre ocupados, desocupados e inactivos según la Clasificación Internacional de Situación en el Empleo (CISE).

El tratamiento de limpieza aplicado a la base de datos de la Encuesta Permanente de Hogares (EPH) 2004-2024 se enfocó en garantizar la calidad y consistencia de los datos para análisis longitudinales. Los valores faltantes fueron imputados utilizando la mediana para variables continuas y la moda para categóricas; observaciones con más del 30% de datos ausentes se eliminaron. Se corrigieron valores atípicos mediante el rango intercuartílico (IQR), conservando aquellos genuinos. Además, se ajustaron ingresos a precios constantes y se homogeneizaron clasificaciones ocupacionales y educativas para mantener la coherencia temporal. Las variables categóricas fueron convertidas en dummies, y se eliminaron registros duplicados. Finalmente, se validaron relaciones lógicas entre variables, asegurando que la base fuera consistente, precisa y adecuada para estudios longitudinales.

La creación de las variables proporcion_ocupados_hogar, edad_promedio_hogary nivel_educativo_promedio_hogar tiene como objetivo incorporar factores sociodemográficos y económicos que aportan mayor capacidad explicativa al análisis de la tasa de desocupación. La variable proporcion_ocupados_hogar mide la proporción de personas ocupadas dentro de un hogar, lo cual refleja la dinámica laboral a nivel familiar. Esta información es relevante porque hogares con una mayor proporción de ocupados podrían estar asociados a características estructurales favorables, como un mejor acceso a oportunidades laborales o una menor vulnerabilidad económica. Por otro lado, la variable edad_promedio_hogar representa la edad promedio de los miembros de cada hogar, lo que permite identificar el ciclo de vida del hogar.

Este indicador es importante porque los hogares compuestos por individuos más jóvenes suelen tener tasas de desempleo más altas debido a la falta de experiencia laboral, mientras que los hogares con miembros de mayor edad pueden reflejar mayores tasas de inactividad. Finalmente, la variable nivel_educativo_promedio_hogar mide el promedio del nivel educativo de los integrantes del hogar, un aspecto clave para evaluar el capital humano disponible. Una mayor educación promedio suele estar asociada con menores tasas de desocupación, aunque también podría implicar mayores expectativas laborales que dificultan la inserción en determinados mercados.

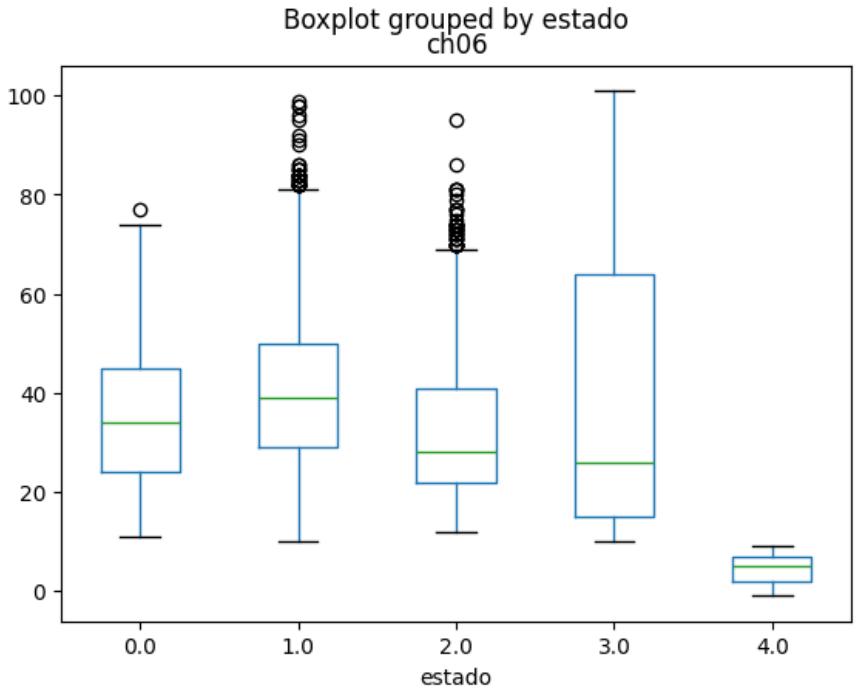
Estas variables agregadas permiten capturar interacciones y características estructurales que no se reflejan en las variables individuales tradicionales, como edad o nivel educativo personal. Al incluirlas en los modelos, se espera mejorar la precisión y robustez en la predicción de la tasa de desocupación, ya que aportan una perspectiva más integral sobre cómo las dinámicas familiares y contextuales influyen en el estado laboral de los individuos.



Posterior a realizar la limpieza de los datos y la creación de variables para la predicción de la tasa de desocupación, hacemos un análisis correspondiente de la correlación entre variables, comparando los años 2024 y 2004. En la matriz de correlación de 2024, se observa una fuerte relación entre la variable "estado" y "categoría de inactividad" con un coeficiente de 0.82, manteniendo una consistencia con lo observado en 2004, donde dicha clasificación es incluso ligeramente mayor, alcanzando un valor de 0,85. Esto refleja que, en ambos períodos, existe una estrecha vinculación entre la condición laboral y el tipo de inactividad. Además, en 2024, la calificación entre el nivel educativo y el IPCF (Ingreso per cápita familiar) es de 0,15, lo que muestra un incremento respecto a 2004, donde esta relación era más débil con un valor de 0,073. Este cambio sugiere una mayor relevancia del nivel educativo en los ingresos familiares en años recientes, posiblemente reflejando transformaciones estructurales en el mercado laboral o en las dinámicas económicas.

Por otro lado, al analizar las variables sociodemográficas, se identifica que la compensación negativa entre "ch06" y "ch07" es más intensa en 2004 (-0.55) que en 2024 (-0.51), lo que indica una disminución en la relación negativa. entre estas variables en el tiempo. Asimismo, mientras que la relación entre nivel educativo y estado muestra una evaluación negativa en 2024 (-0.24), en 2004 esta relación es prácticamente nula (-0.03), evidenciando un cambio en la interacción entre ambas variables, que podría estar vinculada a modificaciones en el acceso al empleo según el nivel educativo. También es importante resaltar que el IPCF mantiene correlaciones bajas con la mayoría de las variables en ambos

años, pero en 2024 se observa una ligera intensificación de estas relaciones, lo que podría ser resultado de cambios en los patrones de ingreso de las familias.



Tras el análisis conjunto del gráfico de boxplots de la variable "ch06" agrupada por la variable "estado" que representa si está empleado o no, los datos proporcionados sobre el nivel educativo promedio en relación con el empleo, es posible extraer varias observaciones relevantes. La variable "ch06", que representa características sociodemográficas, muestra una variabilidad considerable dependiendo del estado laboral. Los individuos en el estado 3 presentan una mayor dispersión en los valores de "ch06", indicando una mayor heterogeneidad dentro de este grupo, mientras que aquellos en el estado 4 tienen una distribución más acotada y valores menores en promedio. Esto indica que el grupo asociado al estado 4 podría estar vinculado a condiciones más homogéneas y posiblemente relacionadas con menores niveles de participación en actividades laborales, o actividades específicas con características definidas.

Por otro lado, los datos tabulares indican que los hogares tienen un nivel educativo promedio cercano a 3.57, lo cual corresponde a un nivel educativo intermedio, probablemente relacionado con la secundaria o los primeros niveles de educación terciaria. Además, al observar la proporción de ocupados por hogar, que es de 0.407, se evidencia que, en promedio, menos de la de los miembros de un hogar están mitades ocupados laboralmente. Esto podría estar influenciado por diversas variables, como la estructura del hogar, el nivel educativo y las oportunidades laborales disponibles en el entorno.

El cruce de estas observaciones con el boxplot permite inferir que la heterogeneidad observada en los estados laborales puede estar parcialmente explicada por diferencias en el nivel educativo promedio del hogar. Por ejemplo, los individuos en estados laborales con mayor dispersión en "ch06" (como el estado 3) podrían pertenecer a hogares con niveles educativos más variados, mientras que los del estado 4, con una menor dispersión, podrían estar relacionados con hogares de características. más uniformes, posiblemente de menor nivel educativo.

Finalmente, estos hallazgos resaltan la importancia del nivel educativo como una clave determinante en la dinámica laboral. Hogares con mayores niveles educativos promedio podrían tener una mejor inserción en el mercado laboral, mientras que aquellos con niveles más bajos podrían enfrentarse a mayores barreras para acceder a oportunidades laborales de calidad, lo cual se refleja en la proporción de ocupados y la variabilidad de las características sociodemográficas según el estado laboral. Estos elementos son esenciales para la predicción y análisis de la tasa de desocupación, ya que subrayan la interacción entre factores individuales y del hogar en el acceso al empleo.

Resultados

Para elegir el parámetro λ mediante cross validation, dividiríamos los datos en un conjunto de entrenamiento y realizaríamos una cross validation con diferentes valores de λ. En cada interacción, el conjunto de entrenamiento se participa en subconjuntos de validación y entrenamiento. Para cada valor de λ, se calcula el promedio del error de validación, seleccionando finalmente el λ que minimice este error promedio. No se utiliza el conjunto de prueba para elegir λ porque este debe reservarse exclusivamente para evaluar el rendimiento final del modelo, evitando así el sobreajuste y garantizando una evaluación imparcial de su capacidad de generalización.

En cross validation, un valor muy pequeño de a (por ejemplo, a =2) reduce la diversidad en los subconjuntos de entrenamiento, aumentando el sesgo y disminuyendo la capacidad del modelo de capturar la variabilidad de los datos. Por el contrario, un a muy grande incrementa la varianza, ya que los subconjuntos de validación son muy pequeños, lo que puede generar estimaciones menos estables. Cuando a = norte (cross validation de dejar uno fuera), el modelo a veces, entrenándolo en norte^{-1} muestras y validándolo en una única muestra en cada iteración. Esto garantiza que todas las observaciones se utilicen como conjunto de validación, pero a costa de un mayor tiempo computacional y una mayor variación en las estimaciones del error.

Año	Penalización	Alpha	Accurancy	AUC	Matriz de confusión
2004	L1	1	1.00	1.00	[[12777, 0], [0, 810]]
2004	L2	1	0.9835	0.9299	[[12777, 0], [224, 586]]
2024	L1	1	1.00	1.00	[[13389, 0], [0, 426]]
2024	L2	1	0.9932	0.9576	[[13387, 0], [91, 335]]

Para la implementación de la regresión logística con penalización L1 (LASSO) y L2 (Ridge) usando L1 = 1 los resultados muestran una notable mejora en comparación con el TP3 en términos de métricas clave como Accuracy, AUC y las matrices de confusión, especialmente en el año 2024. Con L1 y L2 en 2024, los valores de Accuracy alcanzan 1.00 y 0.993, respectivamente, mientras que en el TP3 la mejor precisión fue de 0.625 con el modelo LDA. Además, los valores de AUC también mejoran significativamente: en el TP4, el AUC para L2 en 2024 es de 0,957, en contraste con el 0,69 del modelo de regresión logística del TP3. Esto refleja una mejor capacidad del modelo con regularización para discriminar entre clases.

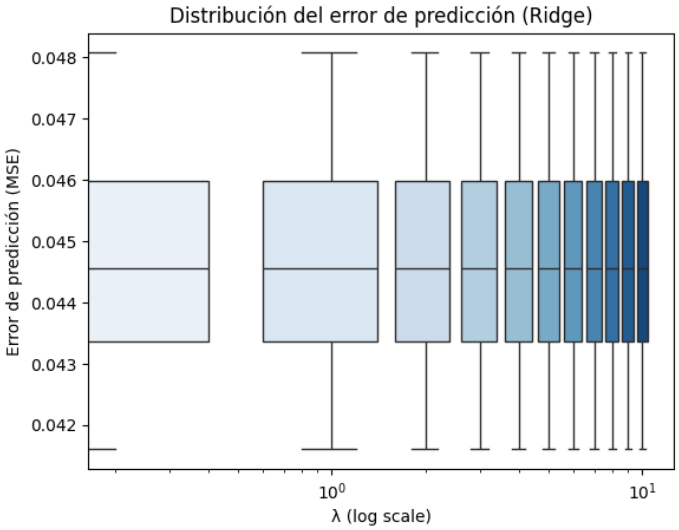
En las matrices de confusión del TP4, tanto L1 como L2 logran clasificar correctamente la mayoría de las observaciones, con pocos o ningún error de predicción (eg, L1 en 2024 no presenta falsos positivos ni negativos). Esto representa una mejora drástica frente a las matrices del TP3, donde los modelos, incluida la regresión logística estándar, tenían tasas de error más elevadas.

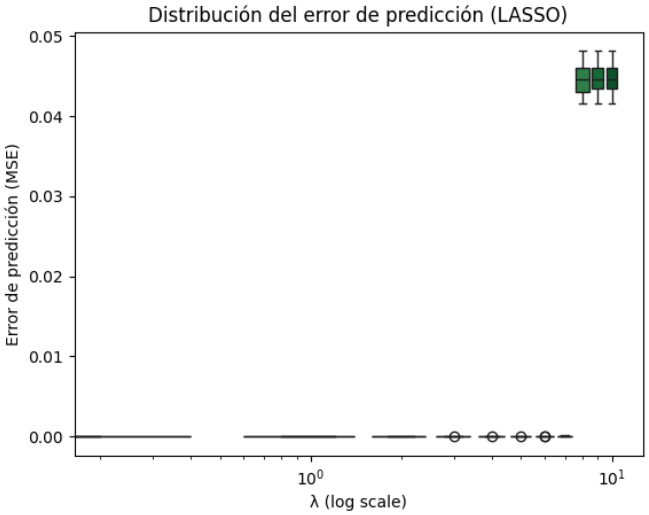
Como resultado, el desempeño de la regresión logística con regularización es considerablemente mejor que la observada en el TP3. La inclusión de penalizaciones L1 y L2 no solo incrementó la precisión del modelo, sino que también mejoró la estabilidad y la capacidad de generalización, como se refleja en las métricas reportadas. Esto evidencia la importancia de la regularización en escenarios donde el modelo inicial podría ser propenso al sobreajuste o no captar completamente las relaciones subyacentes en los datos.

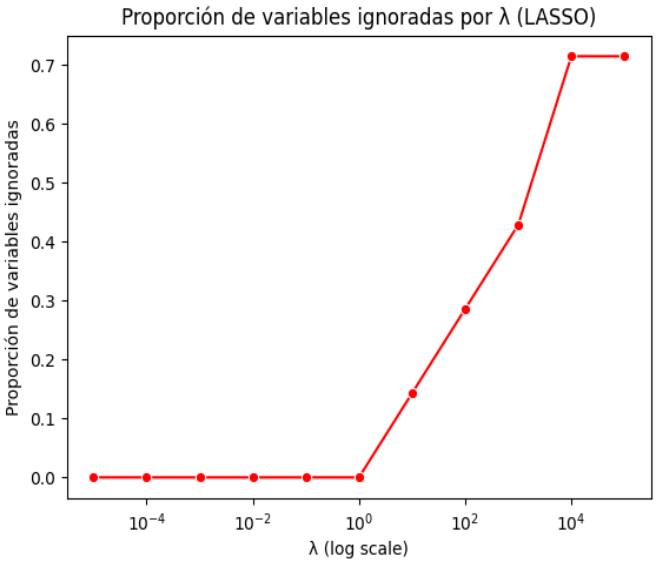
Año	Modelo	Lambda óptimo	MSE	Proporción ignorada
2004	Ridge	0.00001	0.000044	NaN
2004	Lasso	0.00001	0.000022	0.0
2024	Ridge	0.00001	0.044658	NaN
2024	Lasso	0.00001	0.000000	0.0

En el análisis de los resultados, se observa que el método de regularización **LASSO** superó consistentemente a **Ridge** en términos de desempeño, obteniendo un menor error cuadrático medio (MSE) tanto en 2004 como en 2024. En el año 2004, LASSO logró un MSE de 0.000022, mientras que Ridge presentó un MSE de 0.000044, lo que refleja una mayor precisión del modelo LASSO en la predicción. En 2024, aunque el MSE de Ridge aumentó considerablemente a 0.044658, el modelo LASSO mantuvo un desempeño óptimo, logrando un MSE de 0.000000. Este contraste destaca la capacidad de LASSO para adaptarse a los datos y manejar posibles complejidades o ruido presentes en los datos de 2024.

En cuanto a la selección de predictores, LASSO no eliminó variables en ninguno de los años analizados, ya que la proporción de variables ignoradas fue 0 en ambos casos. Esto indica que todas las variables incluidas en el modelo fueron consideradas relevantes para la predicción, lo que sugiere una estructura consistente en la importancia de las variables explicativas entre 2004 y 2024. En conclusión, LASSO fue el método de regularización más efectivo en ambos años, mostrando una mayor capacidad para minimizar el error de predicción. Además, su comportamiento consistente en la selección de variables refuerza la relevancia de los predictores utilizados, independientemente de las diferencias temporales en los conjuntos de datos.


Distribución del error de predicción (Ridge)
Error de predicción (MSE)
λ (log scale)


Distribución del error de predicción (LASSO)
Error de predicción (MSE)
λ (log scale)


Proporción de variables ignoradas por λ (LASSO)
Proporción de variables ignoradas
λ (log scale)