Ciencia de Datos

Trabajo Práctico № 4

Ignacio Spiousas

Noelia Romero

Ignacia Auguste

Florencia Reverendo

Trinidad Morán

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

1. Variables que pueden ser predictivas de la desocupación
2. Eliminen las observaciones que no corresponden a los aglomerados de CABA y GBA y unan ambos trimestres en una sola base.

En este trabajo se utilizaron las bases de datos de la Encuesta Permanente de Hogares correspondientes al primer trimestre de 2004 y 2024. Cada año dispone de dos conjuntos de datos: uno sobre hogares y otro sobre personas. Para la elaboración de este análisis, estas bases se integraron utilizando los códigos de identificación de viviendas (CODUSU) y hogares (NRO\_HOGAR), los cuales permitieron vincular los registros de hogares con las personas correspondientes. A partir de esta integración, se generó una base de datos conjunta que incluye únicamente información de la Ciudad Autónoma de Buenos Aires y el Gran Buenos Aires, resultando en un total de 14.698 registros.

1. Limpieza de la base

Sobre esta base consolidada, se procedió a codificar las columnas CH04 (sexo), CH07 (estado civil), CH08 (cobertura médica), NIVEL\_ED (nivel educativo), ESTADO (condición de actividad) y CAT\_INAC (categoría de inactividad) como variables categóricas numéricas, asignando un valor numérico específico a cada categoría.

Con el objetivo de reducir posibles errores en los datos, se eliminaron aquellos registros con valores que carecían de sentido lógico. En particular, se descartaron los valores negativos en las variables CH06 (edad), PP03D (cantidad de ocupaciones), PP08D1 (ingresos mensuales), P21 (ingreso total habitual de la ocupación), IPCF (ingreso per cápita familiar), IX\_TOT (cantidad de personas en el hogar) y CAT\_INAC (categoría de inactividad), ya que estos valores deberían ser positivos para mantener la coherencia de los datos. Luego de la limpieza, la base de datos quedó conformada por 13.701 registros.

Además, se realizó un análisis de los datos faltantes en ciertas columnas de interés. Se identificaron los siguientes valores ausentes: 135 registros en la columna de edad (CH06), 13 registros en la columna de estado civil (CH07), 20 registros en la columna de cobertura médica (CH08), 1 registro en la columna de ingreso familiar per cápita (IPCF), 230 registros en la columna de cantidad de personas en el hogar (IX\_TOT).

Dado que el número de datos faltantes era relativamente bajo en comparación con el tamaño total de la base, se optó por eliminar los registros con valores faltantes, excepto en el caso de la columna de edad (CH06). Consideramos que la edad es un dato fundamental para el análisis, por lo que decidimos estimar los valores faltantes utilizando la mediana de la edad correspondiente a cada grupo de nivel educativo. Este enfoque permitió preservar la relevancia de la variable sin introducir sesgos significativos en los datos.

1. Construcción de 4 variables

PROP\_OCUPADOS: Esta variable nos permite conocer la proporción de las personas que trabajan en un hogar.

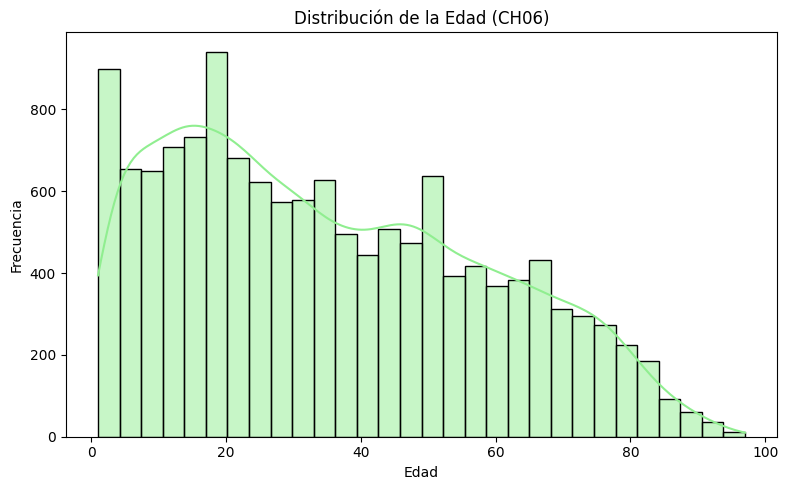
PROP\_MENORES: Nos muestra la proporción de menores de edad que habitan en el hogar.

DEP\_ECO: El fin de esta variable es mostrar la proporción de habitantes activos vs inactivos de un hogar para ver de cuántas personas depende el hogar.

EDUC\_PROM: Muestra el nivel educativo promedio de los habitantes del hogar.

1. Análisis descriptivo de variables

En este apartado, se presentan las estadísticas descriptivas de tres variables de la (EPH) que consideramos relevantes para analizar la desocupación: edad (CH06), nivel educativo (NIVEL\_ED) y categoría de inactividad (CAT\_INAC). A continuación, se exponen los análisis y comentarios correspondientes a cada variable.

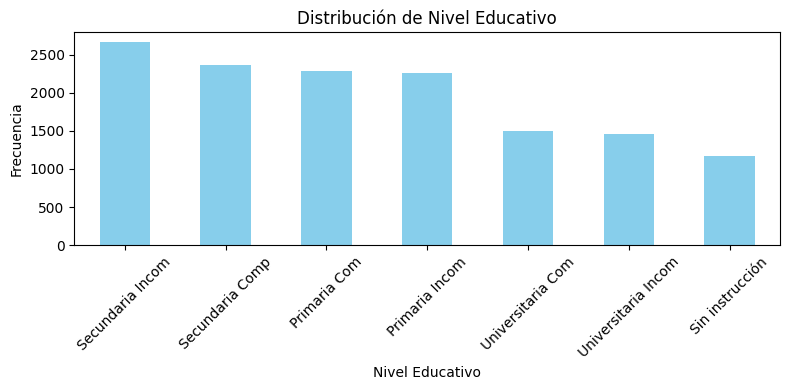
La edad es una variable clave para entender la desocupación, ya que los niveles de participación en el mercado laboral y las probabilidades de estar ocupado o desocupado varían significativamente según el grupo etario. Por ello, el análisis de esta variable resulta relevante para comprender la distribución de la desocupación en diferentes etapas de la vida.

En cuanto a las estadísticas descriptivas obtenidas, la edad promedio de la población analizada es de 35.27 años, lo que indica que la mayoría de las personas están dentro de la edad laboral activa. La desviación estándar, de 23.14 años, evidencia una amplia dispersión en los valores de la muestra, lo que sugiere que la población incluye tanto a menores de edad como a adultos mayores. Los valores extremos de 1 y 97 años confirman la presencia de personas fuera de la edad laboral activa, como niños pequeños y jubilados. Al analizar los percentiles, se observa que el 25% de la población tiene 16 años o menos, lo que representa un segmento con baja participación en el mercado laboral. Por otro lado, la mediana de 32 años refleja que la mitad de la población tiene esa edad o menos, mientras que el 75% alcanza hasta los 53 años, lo que abarca gran parte de la población económicamente activa.

El nivel educativo es un determinante clave en el análisis de la desocupación, ya que influye en la empleabilidad y en las oportunidades laborales de las personas. En general, niveles educativos más altos están asociados con una mayor empleabilidad, mientras que aquellos con menor educación enfrentan barreras significativas para acceder al mercado laboral. Por ello, esta variable resulta esencial para comprender las dinámicas de la desocupación.

En cuanto a la distribución de los niveles educativos, el grupo más frecuente es el de personas con educación secundaria incompleta, que representa el 19.42% de la población. Esto indica que una proporción considerable de individuos no alcanza a completar el nivel secundario, lo que puede limitar sus posibilidades de inserción laboral, especialmente en empleos que exigen una calificación mínima. Por su parte, el 17.29% de la población tiene secundaria completa, lo que señala que una cantidad significativa logra alcanzar este nivel, considerado un requisito básico en muchos ámbitos laborales.

El grupo de personas sin ningún tipo de instrucción constituye el 8.55% de la población, lo que sugiere que la mayoría cuenta con al menos algún grado de educación formal. Sin embargo, este segmento es particularmente vulnerable en el mercado laboral debido a la ausencia de calificaciones. En cuanto a los niveles superiores, el 10.95% de los individuos tiene educación universitaria incompleta, y el 10.66% ha completado este nivel, lo que en conjunto representa alrededor del 22% de la población. Esto evidencia que una parte importante accede a la educación superior, aunque no todos logran finalizarla.



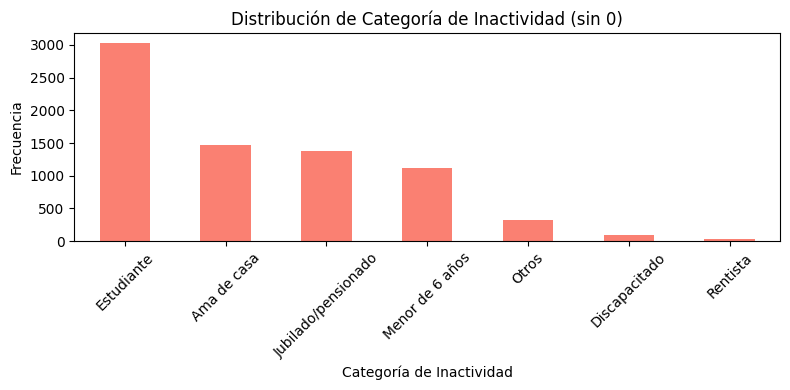
La variable CAT\_INAC es esencial para identificar los individuos que no participan activamente en el mercado laboral, tales como jubilados, estudiantes, amas de casa, menores de 6 años y personas con discapacidades. Esta categorización permite entender mejor los factores que influyen en la inactividad laboral, lo cual es crucial al estudiar la relación entre inactividad y desocupación dentro de los hogares.

En términos de las frecuencias y proporciones, una parte significativa de la muestra no tiene categoría asignada, con un 45.60% de los registros correspondientes al valor 0.0. Este alto porcentaje se corresponde con las personas que se encuentran en actividad, tienen asignado el valor 0 porque no les corresponde una clasificación de inactividad.

El grupo de estudiantes representa el mayor porcentaje dentro de las categorías válidas de inactividad, con un 22.14%. Este dato es esperado, dado que los estudiantes suelen estar fuera del mercado laboral debido a sus estudios. La categoría de amas de casa ocupa el tercer lugar con un 10.72%, reflejando a aquellos que no participan en el mercado laboral debido a sus responsabilidades domésticas. Su presencia en un hogar podría influir en las dinámicas económicas y en la proporción de personas activas que pueden estar buscando empleo.

El 10.11% de la población corresponde a jubilados y pensionados, lo que indica que una porción significativa de la muestra está compuesta por adultos mayores que han dejado el mercado laboral. En cuanto al grupo de menores de 6 años, que constituye el 8.13% de la muestra, aunque no participan en el mercado laboral, es relevante debido a la dependencia económica que generan en los hogares, lo cual puede influir en las decisiones económicas de las familias.

Por otro lado, las categorías de discapacitados (0.73%) y rentistas (0.23%) tienen proporciones pequeñas, pero no por ello menos importantes. Las personas con discapacidad pueden enfrentar barreras adicionales para participar en el mercado laboral, mientras que los rentistas podrían tener un ingreso pasivo que los mantiene fuera del ámbito laboral activo.



En conclusión, la variable categoría de inactividadofrece una visión detallada de los grupos inactivos dentro de los hogares y proporciona contexto adicional para comprender cómo estos grupos impactan en la economía familiar y en las tasas de desocupación.

2. Elegimos el parámetro de regularización a través de validación cruzada dividiendo el conjunto de entrenamiento en k-folds para optimizar el hiper parámetro. Se entrena el modelo en k -1 de los subconjuntos y validamos en el restante. Utilizaremos k = 10. Esto se repite rotando los conjuntos y calculamos el promedio del ECM como métrica de desempeño para elegir el λ que lo minimiza. No se utiliza el conjunto de prueba hasta la evaluación de rendimiento final para evitar caer en *overfitting.*

3. No tengo idea

con k = n el modelo se entrena en la cantidad de veces que tiene datos?