Informe TP4

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

1. Luego de explorar el diseño de registro de la base de hogar, consideramos las siguientes variables como predictivas de la desocupación:

Características de la Vivienda:

- ➔ **IV1 (Tipo de Vivienda)**: El tipo de vivienda puede reflejar el nivel socioeconómico del hogar y, por ende, influir en la probabilidad de desocupación.
- ➔ **IV2 (Cantidad de Ambientes):** Una vivienda con pocos ambientes puede ser indicativo de hacinamiento y condiciones de vida precarias, factores que podrían estar asociados a la desocupación.
- ➔ **IV6, IV7 (Acceso y Tipo de Agua):** La falta de acceso a agua potable o la dependencia de fuentes no seguras podrían estar relacionadas con la pobreza y la desocupación.
- ➔ **IV8, IV9, IV10 (Saneamiento):** Las condiciones de saneamiento, como la presencia o ausencia de baño y su tipo, pueden ser indicadores del nivel socioeconómico y, por ende, de la vulnerabilidad a la desocupación.
- ➔ **IV12_1, IV12_2, IV12_3 (Ubicación de la Vivienda):** La proximidad a basurales, la ubicación en zonas inundables, o la residencia en villas de emergencia pueden ser factores de riesgo para la desocupación.

Características del Hogar:

- **ITF (Ingreso Total Familiar):** Un ITF bajo puede estar correlacionado con mayores tasas de desocupación.
- **DECIFR, IDECIFR, RDECIFR, GDECIFR, PDECIFR, ADECIFR (Deciles de Ingreso Total Familiar)**: La ubicación del hogar en los deciles de ingreso puede ser un predictor significativo de la desocupación. Los hogares en deciles más bajos podrían enfrentar mayores tasas de desocupación.
- **II7 (Régimen de Tenencia)**: El régimen de tenencia de la vivienda podría estar relacionado con la desocupación. Por ejemplo, los ocupantes de hecho podrían mostrar una mayor propensión a la desocupación.

Estrategias del Hogar:

- **V1 a V18 (Estrategias del Hogar):** Las estrategias que los hogares utilizan para afrontar sus necesidades económicas pueden ser indicativas de su

situación laboral. Por ejemplo, la dependencia de subsidios o ayudas sociales (V5), el uso de ahorros (V13), o la venta de pertenencias (V17) podrían ser señales de riesgo de desocupación.

Resumen del hogar

- **IX_Tot, IX_Men10, IX_Mayeq10**: La cantidad de miembros del hogar, especialmente la presencia de menores, puede influir en la necesidad de generar ingresos y, por ende, en la probabilidad de desocupación.

Ingreso Total Familiar (ITF)

- ITF: se calcula como la suma de los ingresos individuales totales de todos los miembros del hogar.
- Además del monto, la base Hogar incluye variables que ubican al hogar en diferentes escalas decílicas de ingreso, como **DECIFR** (decil del ingreso total del hogar del total EPH), **IDECIFR** (decil del interior), **RDECIFR** (decil de la región), **GDECIFR** (decil del conjunto de aglomerados de 500.000 y más habitantes), **PDECIFR** (decil del conjunto de aglomerados de menos de 500.000 habitantes) y **ADECIFR** (decil del aglomerado).
- Estas variables decílicas se calculan utilizando el ponderador PONDIH, que corrige por la no respuesta de ingresos4.

Organización del Hogar:

- ➔ **IX_Tot:** Cantidad de miembros del hogar.
- ➔ **IX_Men10:** Cantidad de miembros menores de 10 años.
- ➔ **IX_Mayeq10**: Cantidad de miembros de 10 años y más
- ➔ **VII1_1 y VII1_2:** Número de componente del hogar que realiza las tareas de la casa (se pueden registrar hasta dos personas)
- ➔ **VII2_1 a VII2_4:** Número de componente del hogar que ayuda en las tareas de la casa (se pueden registrar hasta cuatro personas)

La limpieza de datos

El primer paso consistió en reducir el conjunto de datos a las variables que son relevantes para nuestro análisis (ver inciso 1) y además, agregamos variables como “ano4” y “codusu”, “nro_hogar”. A partir del conjunto original de datos, seleccionamos solo las columnas que nos proporcionaron información importante para predecir la desocupación, según lo identificado previamente en el proyecto.

Para manejar los valores faltantes en el conjunto de datos, identificamos inicialmente qué columnas contienen valores NaN. Detectamos solamente que “idecifry” y “pdecifr” estaban completamente llenas de NaN . Estas columnas no contienen información útil para el análisis, por lo que decidimos eliminarlas del DataFrame.

Para identificar los valores atípicos en las variables numéricas relevantes, seguimos estos pasos:

1. **Visualización mediante Boxplots** : Utilizamos gráficos de boxplot para observar visualmente las distribuciones de las variables numéricas principales. Este método nos permitió identificar los valores extremos que se encuentran fuera de los "bigotes" del diagrama, los cuales representan el rango de datos esperado.

2. **Criterio del Rango Intercuartílico (IQR)** : Aplicamos el criterio del IQR para definir los límites superior e inferior de las variables:

Luego de visualizar los boxplot identificamos dos variables que podrían tener outliers: Se aplicó este procedimiento a las siguientes variables relevantes: Ingreso Total Familiar (ITF) e Ingreso Per Cápita Familiar (IPCF).

El código transformó variables categóricas y numéricas para preparar el conjunto de datos para análisis y modelado. Se identificaron variables categóricas relevantes, como tipo de vivienda, acceso al agua, tipo de baño y nivel educativo, y se reemplazaron sus valores textuales por números, respetando su lógica y orden (por ejemplo, "Casa" como **1** y "Departamento" como **2**). Variables dicotómicas como "Sí" y "No" se codificaron como **1** y **2**, mientras que respuestas como "No sabe/No responde" se asignaron a **0**. Además, se aseguró que todas las variables relevantes fueron convertidas explícitamente a tipo numérico para facilitar su uso en análisis estadísticos y modelos de aprendizaje automático.

Se identificaron las variables categóricas presentes en el conjunto de datos y se generaron nuevas columnas representativas utilizando la función pd.get_dummies(), manteniendo todas las categorías originales. Las variables ficticias creadas se revisaron y, en caso de estar en formato booleano, se convirtieron en valores enteros (1 y 0) para asegurar su compatibilidad con técnicas estadísticas y modelos predictivos.

Construcción de variables

En el análisis, se crearon tres nuevas variables relevantes que no estaban inicialmente en la base de datos, pero que son significativas para predecir la desocupación de los individuos.

En primer lugar, se construyó la variable **Proporción de Personas Ocupadas en el Hogar** (“proporcion_ocupados”). Esta variable representa el porcentaje de miembros del hogar que se encuentran ocupados, calculada como la relación entre el número de personas ocupadas y el total de miembros del hogar. Posteriormente, se redondeó a dos decimales para mejorar su legibilidad y precisión en el análisis.

En segundo lugar, se creó Proporción de mujeres en el hogar (proporcion_mujeres): Se calculó como la proporción de mujeres en relación con el total de miembros del hogar. Esta variable puede ser relevante para analizar cómo la composición de género influye en la ocupación, considerando posibles diferencias en las oportunidades laborales.

1. **Proporción de Adultos Mayores en el Hogar** (proporcion_adultos_mayores): Representa el porcentaje de miembros del hogar mayores de 65 años respecto al total. Esta variable es útil para evaluar si la presencia de adultos mayores impacta en las dinámicas laborales de los demás miembros, como el tiempo dedicado al cuidado.

Estadísticas descriptivas

Estadísticas	Tamaño del hogar (ix_tot)	Monto de Ingreso Total Familiar (itf)	Proporción de Ocupados (proporcion_ocupados)
count	13.402	13.402	13.402
mean	3.69	139.073,56	43%
std	1.57	286.629,35	29%
min.	1	0	0
25%	2	210	25%
50% (mediana)	4	950	40%
75%	5	105000	60%
max	7	1412500	100%

Tabla 1 - Estadística descriptiva

En el análisis de las tres variables seleccionadas: Tamaño del hogar (ix_tot) , Monto de Ingreso Total Familiar (itf) y Proporción de Ocupados (proporcion_ocupados) —se observan patrones significativos que pueden ayudar a predecir la desocupación. El tamaño del hogar muestra que los hogares más grandes tienden a tener mayores presiones económicas, lo cual podría estar vinculado con un mayor riesgo de desocupación. Los hogares con más miembros pueden enfrentar dificultades para encontrar empleo para todos, lo que aumenta la probabilidad de que algunos miembros permanezcan desocupados.

Por otro lado, como se puede observar en la Tabla 1, el ingreso total familiar tiene una alta dispersión, con una media de 139.073,56 , pero una desviación estándar de

286.629,35 , lo que indica que existen hogares con ingresos muy bajos y otros con ingresos significativamente altos. Esto sugiere que los hogares con menores ingresos tienen una mayor probabilidad de desocupación, debido a la falta de acceso a oportunidades laborales y recursos. Finalmente, la proporción de ocupados dentro del hogar, con un promedio de 43% , refleja cómo la empleabilidad de los miembros de un hogar impacta en la tasa de desocupación. Los hogares con menos miembros ocupados tienen más riesgos de enfrentar situaciones de desempleo. En resumen, estas variables ofrecen una visión clara de los factores socioeconómicos que influyen en la desocupación y son cruciales para un análisis predictivo más preciso.

Parte II: Clasificación y regularización

2. El procedimiento para elegir el parámetro λ implicó primero definir un rango de valores posibles, semanalmente en una escala logarítmica. A continuación, se entrenó y evaluó el modelo utilizando validación cruzada, que generalmente se realiza con k-fold (usualmente con k=5 o k=10), y se registró la métrica de evaluación correspondiente, como el error cuadrático medio en problemas de regresión del AUC en clasificación. Finalmente, se seleccionó el valor de λ que minimiza el error en los datos de validación cruzada, asegurando así la mejor performance del modelo.

El conjunto de prueba no se debe usar ya que se reserva exclusivamente para evaluar el rendimiento final del modelo. Usar el conjunto de prueba para elegir λ introduce fuga de datos (data leakage), lo que significa que el modelo "ve" información del conjunto de prueba durante su ajuste, lo que lleva a sobreestimación del rendimiento en el cual el modelo podría parecer más preciso de lo que realmente es en datos nuevos y falta de generalización donde el modelo se ajusta no solo a los datos de entrenamiento, sino también a las características específicas del conjunto de prueba, comprometiendo su capacidad para funcionar en datos completamente nuevos.

Por estas razones, se utiliza validación cruzada dentro del conjunto de entrenamiento para seleccionar λ, mientras que el conjunto de prueba se usa únicamente para medir el rendimiento real del modelo.

3. La validación cruzada (cross-validation) divide los datos en k subconjuntos (folds) para entrenar y evaluar un modelo en diferentes particiones de los datos. El valor de k tiene un impacto significativo en el balance entre la precisión de la estimación y el costo computacional.

Cuando k es muy pequeño (por ejemplo, k=2), las implicancias pueden ser que la evaluación del modelo puede ser más sensible al azar, ya que se entrena en grandes porciones de los datos y se prueba en un conjunto muy pequeño. La estimación del rendimiento puede no ser representativa, ya que cada prueba se

realiza sobre un subconjunto muy limitado de los datos. Menos particiones implican que el modelo se entrena y evalúa pocas veces, lo que reduce el tiempo de cálculo. Típicamente, se usa en situaciones donde los datos son muy grandes, y el costo computacional es una prioridad.

Cuando k es muy grande (por ejemplo, cercano a n), las implicancias pueden ser que cada partición de prueba es muy pequeña, por lo que el modelo se entrena casi con todos los datos en cada iteración, lo que resulta en una estimación más precisa del rendimiento. Aumenta el número de veces que el modelo debe ser entrenado, ya que se entrena k veces. El modelo aprovecha casi todo el conjunto de datos para entrenarse, pero la evaluación puede ser más sensible a pequeñas variaciones en los datos de prueba. Sin embargo, es útil cuando se busca máxima precisión en la evaluación, especialmente con conjuntos de datos pequeños.

k	Sesgo	Varianza	Costo computacional	Estabilidad
Pequeño (k=2,3)	Alto	Bajo	Bajo	Baja
Grande (k≈n)	Bajo	Alto	Alto	Muy alta

Tabla 2

4. Los modelos con penalidad L1 para los años 2004 y 2024 alcanzan un AUC perfecto (1.00), lo que sugiere un desempeño ideal en la separación de clases, aunque esto podría indicar sobreajuste. En contraste, los modelos con penalidad L2 presentan rendimientos más modestos, con AUC de 0.64 para 2004 y 0.56 para 2024, siendo apenas mejor que el azar para este último. Esto resalta que la penalidad L1 parece ser más efectiva, pero requiere análisis adicionales para garantizar generalización.

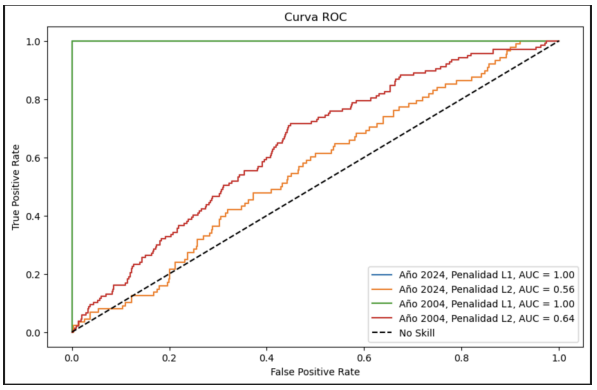

Figura 1 - Curva ROC

5. El boxplot muestra cómo el "Mean Squared Error" (MSE) varía con diferentes valores de λ (escala logarítmica) en Ridge Regression. Aunque la mediana del MSE se mantiene estable, los valores bajos de λ presentan mayor dispersión, indicando que menor regularización introduce más

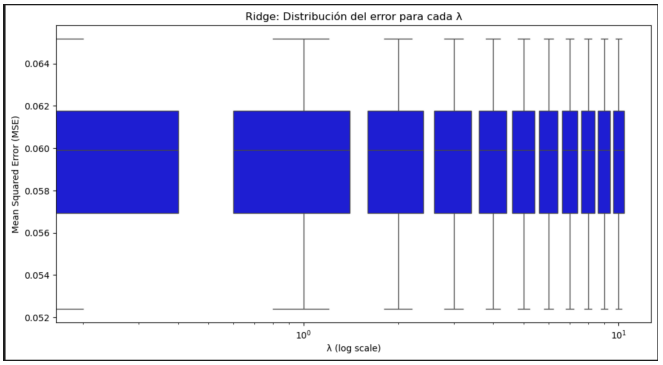

Figura 3 - Boxplot de la distribución de MSE en Ridge

variabilidad en los errores. A medida que λ aumenta, la dispersión se reduce, mostrando un modelo más consistente pero sin un impacto significativo en el error promedio. Esto sugiere que valores altos de λ aportan estabilidad sin sacrificar precisión.

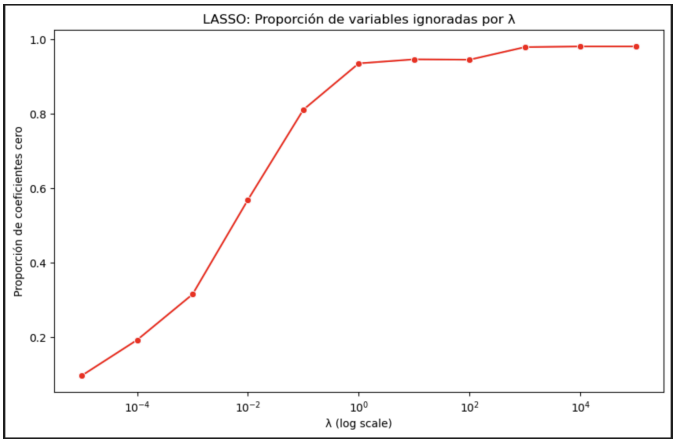

Figura 3 - LASSO y la proporción de coeficientes 0

La Figura 3 muestra cómo LASSO elimina progresivamente variables al aumentar el valor de λ (escala logarítmica). Para valores bajos de λ, menos del 20% de las variables son eliminadas, mientras que con λ más alto, la proporción de coeficientes cero crece rápidamente, alcanzando el 100% cuando λ es muy grande. Esto indica que LASSO favorece modelos más parsimoniosos al incrementar λ, reduciendo la complejidad del modelo y mitigando el riesgo de sobreajuste, aunque valores excesivos pueden llevar a un subajuste al eliminar demasiadas variables.

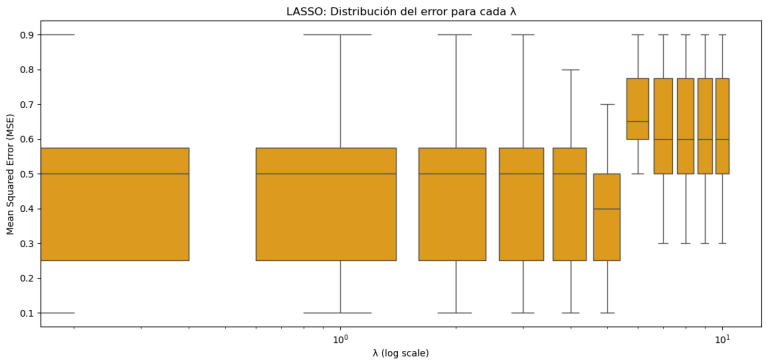

Figura 4 - LASSO y MSE

La Figura 4 ilustra la distribución del error cuadrático medio (MSE) en función de diferentes valores del parámetro de regularización λ en un modelo de regresión Lasso. En el eje horizontal se representan los valores de λ en una escala

logarítmica, mientras que el eje vertical muestra el MSE. Cada caja refleja la distribución del MSE para un valor específico de λ, con la línea central indicando la mediana, los bordes representando los cuartiles 25% y 75%, y los bigotes extendiéndose hasta el rango intercuartílico. Se observa una tendencia general donde el MSE disminuye a medida que aumenta λ, lo que sugiere que la regularización Lasso ayuda a reducir el sobreajuste del modelo y mejora la precisión de las predicciones.

6. El modelo LASSO con λ=0.0001 seleccionó la mayoría de las variables como relevantes para predecir la desocupación, descartando únicamente 31 de un total de 214. Este comportamiento es consistente con nuestras expectativas, dado que un valor de λ tan chico ejerce una penalización leve, permitiendo que el modelo retenga coeficientes pequeños en lugar de eliminarlos completamente. Las variables descartadas incluyen principalmente categorías específicas o marginales, como *estado_Entrevista individual no realizada* y otras con sufijos *_0* y *_9*, que suelen ser categorías poco informativas o menos relevantes.

En relación con lo que se respondió en el inciso 1 de la Parte I, este resultado reafirma que las variables seleccionadas por LASSO son coherentes con las características clave previamente identificadas, como el nivel educativo (*nivel_ed)*, el ingreso per cápita familiar (*ipcf*) y la proporción de adultos mayores en el hogar (*proporcion_adultos_mayores*). Estas variables reflejan factores críticos asociados a la desocupación, destacando su importancia en el modelo predictivo.

Por lo tanto, el modelo no sólo confirmó nuestras hipótesis iniciales sobre las variables más relevantes, sino que también permitió simplificar el conjunto de datos al descartar variables menos útiles, mejorando la interpretabilidad sin perder precisión.

7. En el análisis de regularización con Ridge y LASSO para los años 2004 y 2024, LASSO mostró un mejor desempeño al lograr un MSE=0.0 en ambos años, superando a Ridge (MSE=0.0126 en 2004 y MSE=0.0025 en 2024). Además, LASSO seleccionó distintas variables relevantes en cada año, descartando 44 variables en 2004 y 62 en 2024. Esto indica que la relación entre las variables predictoras y la desocupación cambió entre los dos períodos, y refleja posibles cambios estructurales en el mercado laboral. Estos resultados destacan la capacidad de LASSO para realizar predicciones precisas y, al mismo tiempo, simplificar el modelo eliminando predictores irrelevantes.