

# Análisis del 10% de mayores ingresos en Chile

Marcos González, Agustín Rabie

19 January, 2025

## Contents

Abstract . . . . .	1
## Objetivos . . . . .	1
Preparación y descripción de los datos . . . . .	1
Modelo Entidad-Relación . . . . .	3
Análisis descriptivo inicial . . . . .	5
Visualización de la distribución por deciles . . . . .	7
Caracterización del 10% superior vs resto de la población . . . . .	8
Análisis territorial . . . . .	8
Sexo jefe/a de hogar . . . . .	10
Edad . . . . .	11
Nivel educacional . . . . .	11
Resumen variables potencialmente predictoras . . . . .	12
Variables potenciales . . . . .	12
Variables de bienestar de EBS . . . . .	14
Propuesta de modelo de Machine Learning . . . . .	15
Bibliografía: . . . . .	15

## Abstract

Este estudio analiza los determinantes de la pertenencia al 10% superior de ingresos en Chile utilizando técnicas de machine learning. Mediante el uso combinado de las encuestas CASEN 2020 y EBS 2021, desarrollamos un modelo predictivo que identifica las características socioeconómicas, demográficas y de bienestar subjetivo asociadas a la pertenencia a este grupo. El análisis contribuye a la literatura sobre élites económicas en Chile y América Latina, aportando evidencia sobre los mecanismos de reproducción de la desigualdad económica.

La literatura internacional muestra que el 10% de mayores ingresos presenta características sociodemográficas distintivas que tienen profundas implicaciones para la cohesión social y las políticas públicas: Influencia Política (Gilens, 2012); impacto en Desigualdad y Bienestar (Wilkinson & Pickett, 2010; 2019); Segregación Social (Méndez & Gayo, 2024; Hernando & Mitchell, 2023). En ese contexto, un modelo predictivo de Machine Learning permitirá: - Validar si los patrones internacionales se replican en Chile - Identificar variables no teorizadas previamente

## ## Objetivos

### Objetivo General

Desarrollar y evaluar un modelo de machine learning para predecir la pertenencia al 10% superior de ingresos en Chile utilizando datos socioeconómicos y de bienestar.

### Objetivos Específicos

- Identificar un conjunto de variables capaces de predecir la pertenencia al 10% superior de ingresos
- Examinar la distribución espacial y las características socioeconómicas del 10% superior de ingresos
- Evaluar el desempeño predictivo de diferentes algoritmos de machine learning
- Analizar la importancia relativa de las variables en la predicción

## Preparación y descripción de los datos

Para comenzar el análisis, se cargan las bibliotecas necesarias y se establecen los parámetros de configuración. Se utiliza una combinación de paquetes para manipulación de datos (haven, dplyr, tidyr, purrr, magrittr, stringr), visualización (ggplot2, scales, viridis, DiagrammeR), manejo de datos espaciales (sf, geodata, chilemapas, rmapshaper) y presentación de resultados (kableExtra), presentación (kableExtra, webshot2, naniar) y utilidades (devtools, rlang).

##2. Carga y preparación inicial de datos El análisis utiliza dos fuentes principales de datos:

CASEN 2020 (versión reducida, en formato rds para poder subirse a GitHub): Proporciona información socioeconómica detallada EBS 2021: Complementa con información adicional (especialmente de bienestar) y factores de expansión actualizados

En el proceso de preparación, se realizan los siguientes pasos: a. Cálculo de deciles de ingreso para identificar el 10% superior según la muestra total de la encuesta CASEN, no de la submuestra de la EBS. b. Creación de variable binaria para el grupo objetivo. c. Merge de ambas bases de datos manteniendo la estructura de la CASEN.

La tabla de resumen del proceso de merge que aparece en el documento muestra que:

- CASEN original tiene 185,437 observaciones: Presenta sólo 98 NAs en la variable ytotcorh de ingreso, lo cual es marginal. También debe considerarse que estamos tomando al hogar como unidad y no a individuos (y los individuos que responden la encuesta pueden no corresponder a los/las jefes/as de hogar).
- EBS y la base final fusionada tienen 10,921 observaciones (debe considerarse también que quienes responden no necesariamente con la misma persona que en el caso de CASEN, a pesar de ser el mismo hogar).

Esto implica una retención del 5.9% de la muestra original de CASEN, lo cual es esperable dado que la EBS es una submuestra de CASEN. Lo crucial aquí es verificar que esta submuestra mantiene representatividad, especialmente para el grupo de interés (10% superior).

### Representatividad y factores de expansión:

Del análisis de los factores de expansión se observa:

- El 10% superior representa 1,859,535 personas en la población.
- El resto representa 13,304,855 personas.

- La razón de expansión promedio es mayor para el grupo elite (1,909 vs 1,337).

Table 1: Resumen del proceso de merge CASEN-EBS

Etapa	N observaciones	Pérdida de registros	% retención
CASEN original	185437	NA	NA
EBS original	10921	NA	NA
Después del merge	10921	174516	5.9

Table 2: Retención de observaciones por decil después del merge

Decil	N en CASEN	N después del merge	% retención
1	18534	1117	6.0
2	18534	1204	6.5
3	18534	1099	5.9
4	18534	1093	5.9
5	18534	1096	5.9
6	18534	1082	5.8
7	18534	1100	5.9
8	18534	1104	6.0
9	18534	1052	5.7
10	18533	974	5.3
NA	98	NA	NA

Los 974 casos que pertenecen al 10% superior representan aproximadamente el 9% de nuestra muestra, lo cual es consistente con la distribución poblacional esperada. Esto nos proporciona una base sólida para el modelamiento:

- La proporción 9%-91% (vs un 10%-90% ideal) representa un desbalance moderado.
- Es manejable con ponderación de clases en Random Forest.
- No requiere submuestreo de la clase mayoritaria.

Rothman y Greenland (2018) argumentan que es más apropiado planificar el tamaño del estudio basándose en la precisión deseada rather que en el poder estadístico tradicional. En nuestro caso, con 974 casos de 10,921 observaciones, alcanzamos una precisión adecuada.

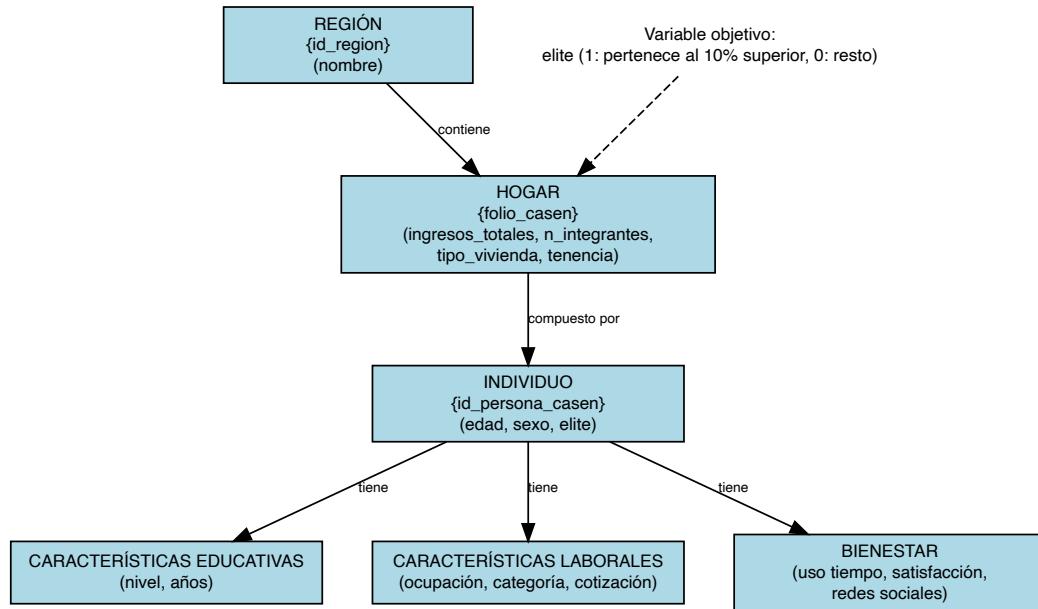
En términos del tamaño muestral requerido para el análisis predictivo, considerando aproximadamente 45 predictores candidatos (6-1[dummy] en bienestar, 9-1 en tenencia, 5-1 en nivel educacional, 12-1 intervalos de edad, 16-1 en Región, más zona y sexo), los 974 eventos nos proporcionan 21.6 eventos por variable, superando los umbrales mínimos estándar en la literatura de predicción. Si bien este criterio es simplificador (van Smeden et al., 2019), proporciona una base razonable para proceder con el análisis predictivo propuesto.

## Modelo Entidad-Relación

El diagrama presentado muestra la estructura de los datos, que se organiza en cuatro entidades principales:

- REGIÓN {id\_region}: Permite el análisis territorial de la distribución del ingreso.
- HOGAR {folio\_casen} Unidad fundamental de análisis económico.

- INDIVIDUO {id\_persona\_casen} Unidad básica de observación.
- CARACTERÍSTICAS: Atributos del individuo (sea o no jefe/a de hogar).



## Análisis descriptivo inicial

Examinación de valores faltantes El gráfico de patrones de valores faltantes nos muestra una estructura importante:

- Variables sin datos faltantes (0% missing): Variables sociodemográficas básicas (edad, sexo, zona, región) Variables económicas (ytotcorh, elite, fexp) Variables educativas (e6a) Variables de vivienda (v13)
- Variables con alta tasa de missing (~42%): Variables de bienestar laboral (j3a\_1, j3a\_4, j3a\_5): 4,606 casos missing Variables de uso del tiempo (a3\_4, c3\_2, c6): 4,606 casos missing
- Variables con tasa media de missing (~35%): Tenencia de vivienda (v13\_propia): 3,805 casos missing

Esto sugiere: Las variables socioeconómicas básicas están completas. Hay un patrón sistemático de no respuesta en variables de bienestar. La pérdida de información en variables de bienestar afecta aproximadamente al 42% de la muestra.

### Patrón de valores faltantes en variables principales

Se muestran todas las variables relevantes para el modelo

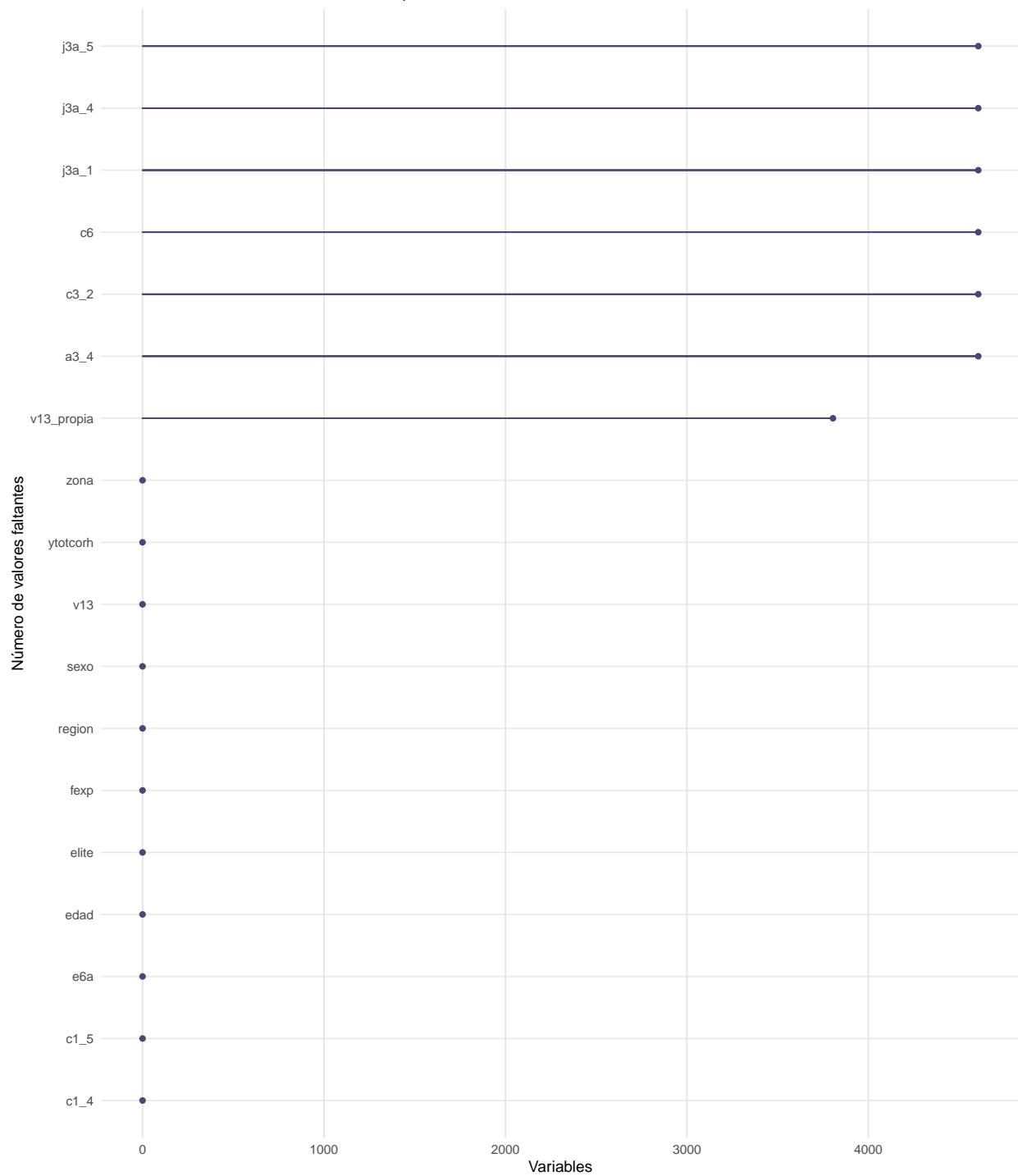


Table 3: Resumen detallado de valores faltantes por variable

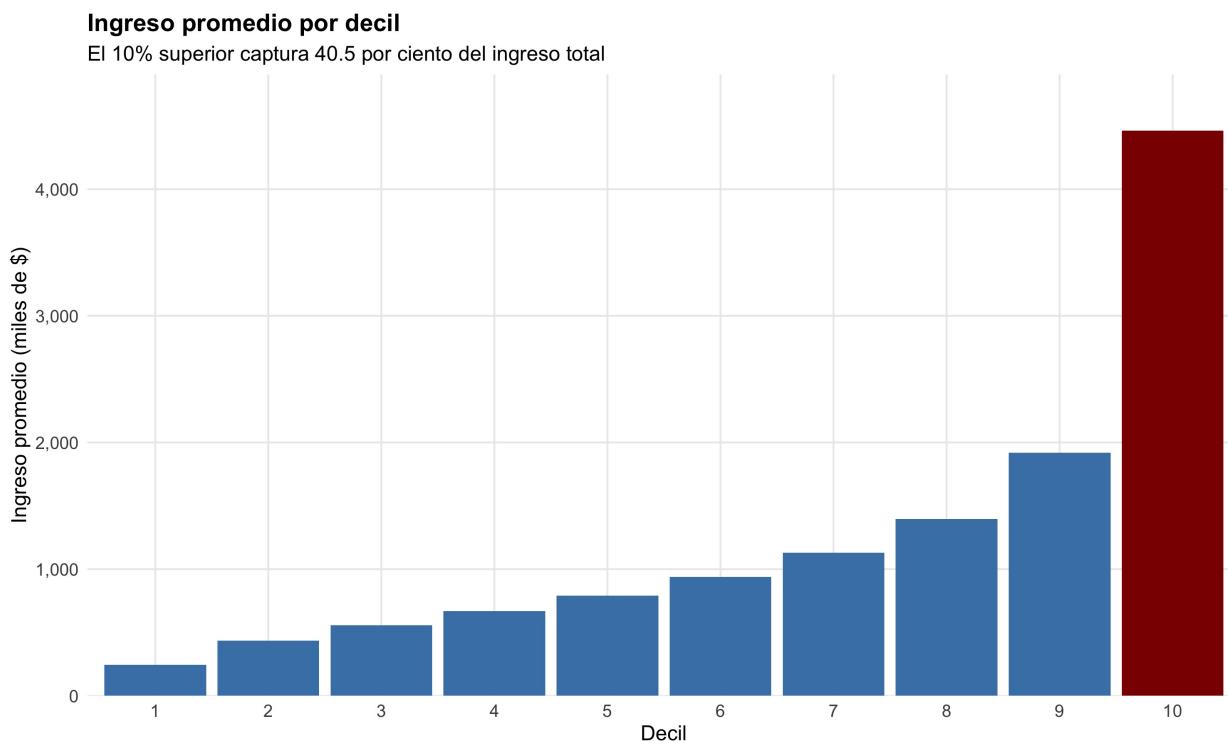
Variable	N Missing	% Missing	Total Obs	Obs Disponibles
j3a_1	4606	42.18	10921	6315
j3a_4	4606	42.18	10921	6315

a3_4	4606	42.18	10921	6315
j3a_5	4606	42.18	10921	6315
c3_2	4606	42.18	10921	6315
c6	4606	42.18	10921	6315
v13_propia	3805	34.84	10921	7116
edad	0	0.00	10921	10921
sexo	0	0.00	10921	10921
zona	0	0.00	10921	10921
region	0	0.00	10921	10921
e6a	0	0.00	10921	10921
v13	0	0.00	10921	10921
ytotcorh	0	0.00	10921	10921
elite	0	0.00	10921	10921
fexp	0	0.00	10921	10921
c1_4	0	0.00	10921	10921
c1_5	0	0.00	10921	10921

## Visualización de la distribución por deciles

El gráfico de ingresos promedio por decil muestra una marcada desigualdad. El décimo decil (en rojo) tiene un ingreso promedio de más de 4 millones de pesos. La diferencia con el noveno decil es especialmente notoria, pues el 10% superior captura 40.5% del ingreso total.

```
## pdf
## 2
```



## Caracterización del 10% superior vs resto de la población

La tabla siguiente compara características clave entre este grupo y el resto de la población. Se incluyen variables territoriales (región, urbano/rural), demográficas (edad, sexo), socioeconómicas (educación) y de bienestar. Los valores están ponderados usando los factores de expansión provistos por la EBS: seleccionamos éstos en vez de los de CASEN por tratarse de una encuesta bifásica.

\begin{table} \caption{Características del 10% superior vs resto}

elite_label	n	n_expandido	ingreso_promedio	desv_est	edad_promedio	prop_hombres	prop_urbano
10% superior	974	1,859,535	4,460,648	2,993,049	41.27	0.56	0.95
Resto	9,947	13,304,855	917,774	500,956	44.87	0.48	0.87

\end{table}

## Análisis territorial

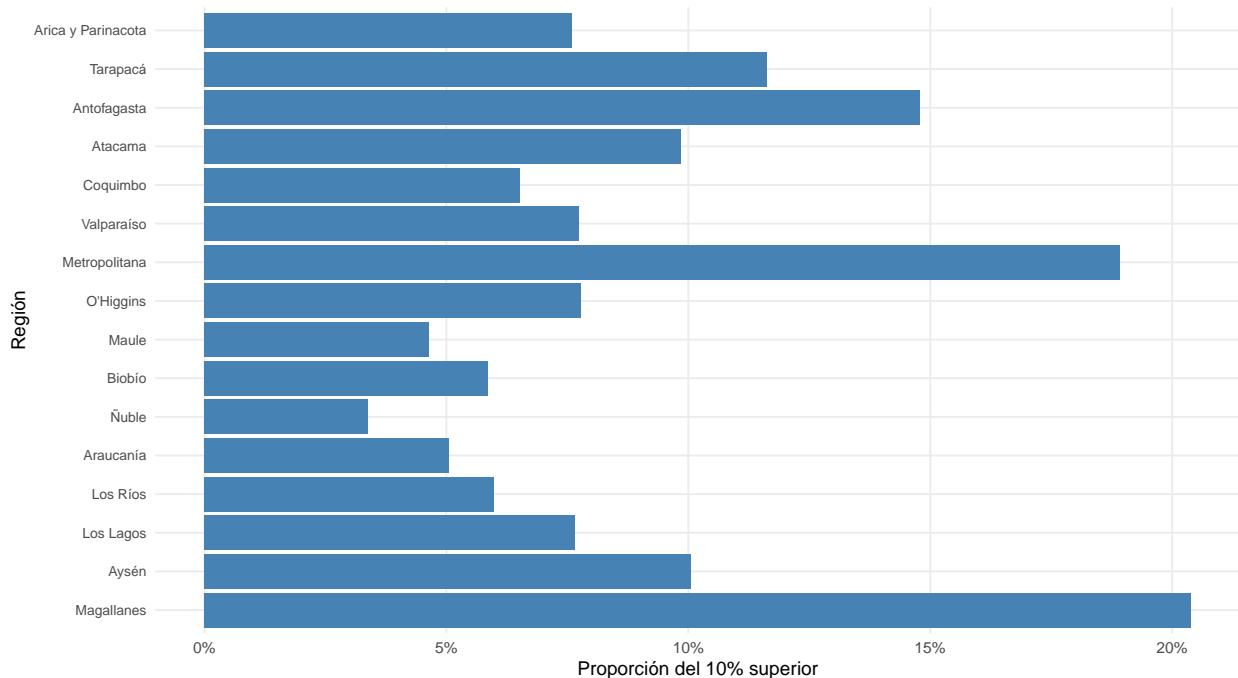
El análisis territorial se desarrolla en dos dimensiones complementarias:

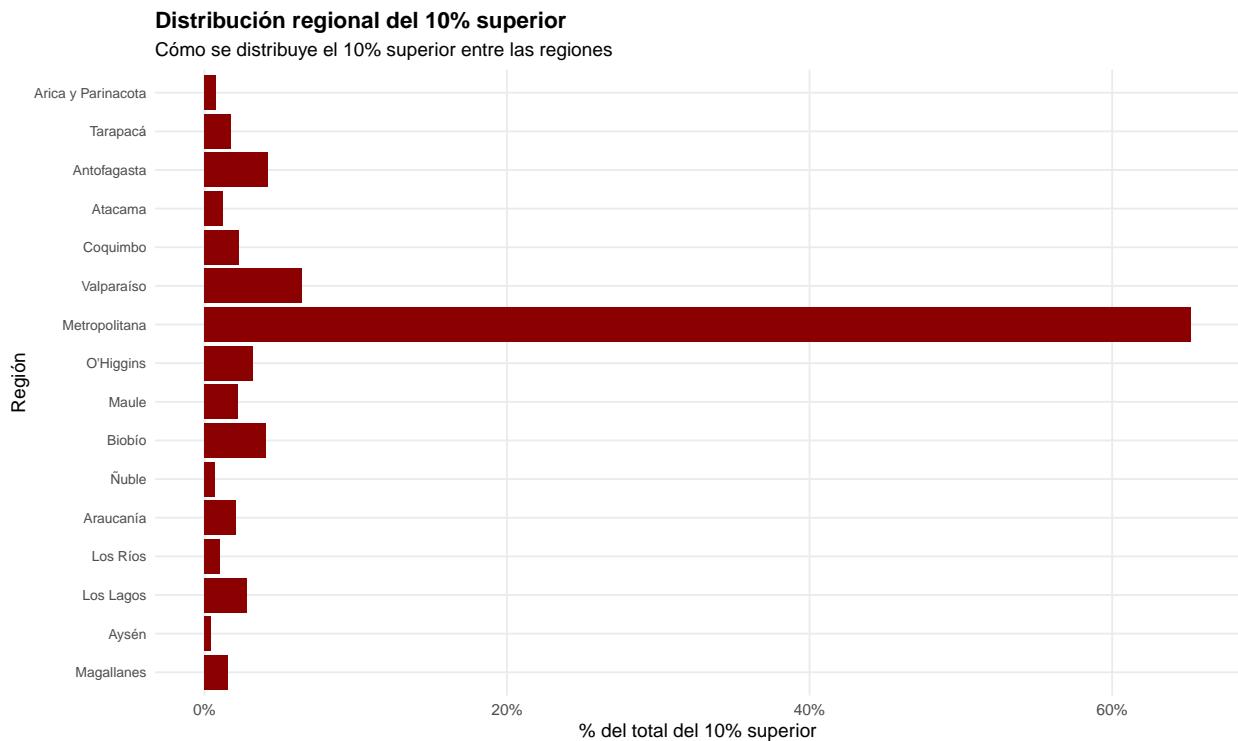
Proporción interna: Qué porcentaje de la población de cada región pertenece al 10% superior Distribución nacional: Cómo se distribuye el total del 10% superior entre las regiones

Para facilitar la interpretación, se ordenan las regiones de norte a sur.

### Proporción del 10% superior por región

% de habitantes de cada región que pertenece al 10% superior





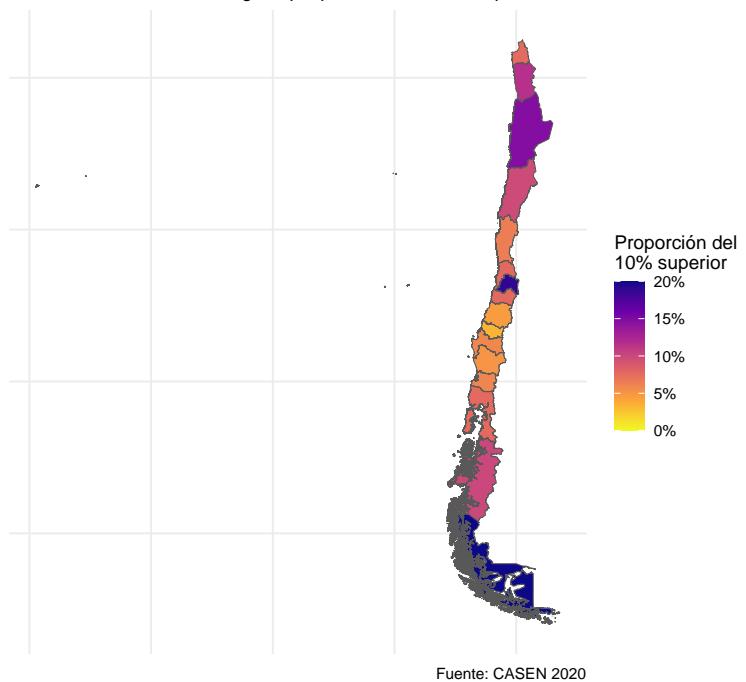
Los gráficos resultantes revelan patrones interesantes. La proporción de elite dentro de cada región (gráfico azul) muestra una concentración en ciertas regiones (Magallanes, RM, Antofagasta,) La distribución del total de la elite (gráfico rojo) evidencia una fuerte centralización en RM.

##8. Visualización espacial Aquí generamos mapas que permiten una visualización más intuitiva de los patrones espaciales. Se utilizan dos mapas que corresponden a las mismas dimensiones analizadas en los gráficos de barras:

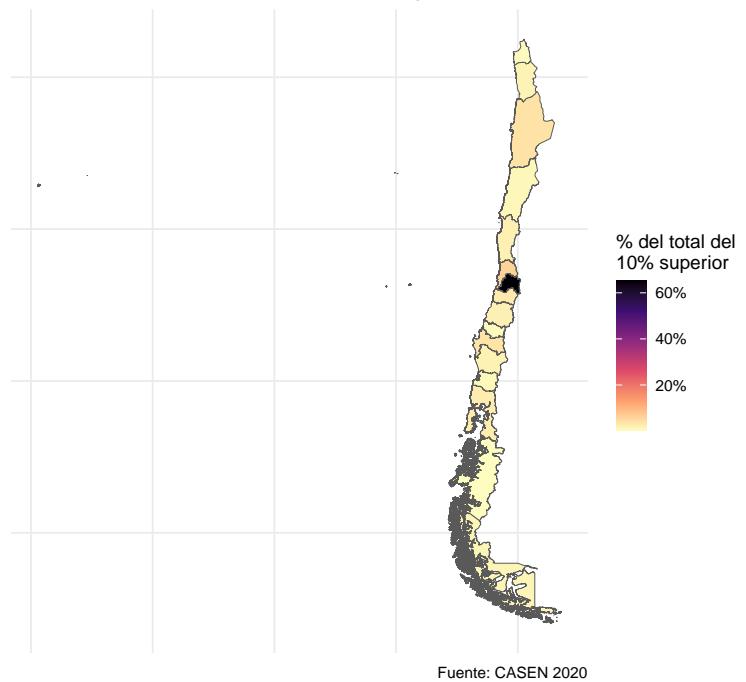
El primer mapa muestra la proporción de habitantes de cada región que pertenece al 10% superior. Este mapa ayuda a identificar dónde es más probable encontrar miembros de la elite económica. El segundo mapa (en tonos magma) visualiza cómo se distribuye el total de ese 10% entre las regiones.

Para la construcción de estos mapas, se enfrentaron varios desafíos técnicos: La necesidad de compatibilizar diferentes codificaciones de regiones: Magallanes tenía problemas de visualización que lo convertía persistentemente en NA.

**Proporción del 10% superior por región**  
% de habitantes de cada región que pertenece al 10% superior



**Distribución regional del 10% superior**  
Cómo se distribuye el 10% superior entre las regiones



### Sexo jefe/a de hogar

Recodificación para identificar sexo de jefe de hogar. No es posible saberlo para el 100% de los casos, pero con variables de sexo y de relación de quien responde con jefe/a de hogar, se puede reducir para cerca del 70%.

Table 4: Distribución de parentesco

Parentesco	Frecuencia
Jefe(a) de Hogar	5164
Esposo(a) o pareja de distinto sexo	2442
Esposo(a) o pareja de igual sexo	18
Hijo(a) de ambos	1078
Hijo(a) sólo del jefe(a)	1331
Hijo(a) sólo del esposo(a)/pareja	65
Padre o madre	123
Suegro(a)	34
Yerno o nuera	119
Nieto(a)	234
Hermano(a)	138
Cuñado(a)	20
Otro Familiar	98
No familiar	57

Table 5: Distribución y proporción de elite por sexo del jefe de hogar

Sexo	N	N expandido	Proporción elite	Proporción muestra	Proporción población
Mujer	2710	3129383	6.2%	0.525	0.454
Hombre	2454	3757075	14.1%	0.475	0.546

```
## [1] "\nDistribución del sexo del jefe de hogar (incluyendo inferidos):"
##
##      1    2 <NA>
## 4357 3267 3297
```

Observaciones importantes: - Los hogares con jefatura masculina tienen más del doble de probabilidad de pertenecer al 10% superior. - Hay una discrepancia entre proporción muestral y poblacional que sugiere sobremuestreo de - hogares con jefatura femenina. - La proporción de elite es notablemente menor en hogares con jefatura femenina.

## Edad

Se utilizan intervalos quinquenales desde los 30 años e intervalos más amplios en edades extremas.  
 Concentración en edades medias: El peak de proporción elite está en 30-34 años (19.5%) y 35-39 años (17.3%). Declina consistentemente después de los 40.

Tiene muy baja representación en extremos: jóvenes (2.2% en 18-29) y casi nula en mayores de 80 (0.3%).  
 Declina fuertemente después de los 70 años.

Diferencias muestra-población: - Ligera subrepresentación de grupos 30-39 años. - Sobrerepresentación de grupos mayores. - Factor de expansión utilizados.

## Nivel educacional

Decisiones clave aquí: - Agrupación en 5 categorías principales - Distinción entre educación técnica y universitaria - Postgrado como categoría separada, a pesar del bajo N, por la sobrerepresentación del 10% superior. - Manejo explícito de valores NA

Table 6: Distribución y proporción de elite por grupo de edad

Grupo de edad	N	N expandido	Proporción elite (ponderada)	Proporción muestra	Proporción población
18-29	293	435288	2.2%	0.057	0.063
30-34	397	670537	19.5%	0.077	0.097
35-39	433	694170	17.3%	0.084	0.101
40-44	486	770243	13.2%	0.094	0.112
45-49	564	674818	14.3%	0.109	0.098
50-54	584	717096	9.3%	0.113	0.104
55-59	593	782963	8.9%	0.115	0.114
60-64	563	625209	8.7%	0.109	0.091
65-69	466	552611	6.5%	0.090	0.080
70-74	361	436181	5.5%	0.070	0.063
75-79	248	282900	5.4%	0.048	0.041
80 o más	176	244442	0.3%	0.034	0.035

Table 7: Distribución y proporción de elite por nivel educacional

Nivel educacional	N	N expandido	Proporción elite (ponderada)	Proporción muestra	Proporción población
Hasta básica	2453	2879582	1.3%	0.225	0.190
Hasta media	4613	6227468	4.3%	0.422	0.411
Postgrado	188	334451	58.7%	0.017	0.022
Técnica superior	1302	1879090	9.9%	0.119	0.124
Universitaria	2365	3843799	30.5%	0.217	0.253

Gradiente educacional marcado: - Postgrado: 58.7% pertenece al 10% superior; Universitaria: 30.5%; Técnica superior: 9.9%; Media o menos: < 5%.

Marcada diferencia con distribución poblacional: 41.1% tiene hasta educación media; 25.3% tiene educación universitaria; Solo 2.2% tiene postgrado.

### Resumen variables potencialmente predictoras

Para preparar la fase de modelamiento, se realiza un análisis exploratorio de las variables que podrían predecir la pertenencia al 10% superior. El análisis de estas variables se realiza considerando: - Su distribución diferenciada entre elite y no elite - La presencia de valores faltantes que podrían afectar el modelamiento - La necesidad de transformaciones o recodificaciones para su uso en modelos predictivos

Decisiones relevantes: - Combinación de tipo (v13) y propiedad (v13\_propia) - Distinción entre vivienda pagada y en proceso de pago - Categorización explícita de situaciones irregulares

Patrones destacados: Alta proporción elite en vivienda propia pagándose (36.1%), mucho más que en vivienda propia pagada (9.7%). Respecto a la muestra total, hay un predominio de vivienda propia pagada (55.1%) e importante sector de arriendo (19.2%).

Casos especiales: Alto porcentaje elite en propiedad compartida pagándose (88.4%), pero muy pocos casos (n=3).

### Variables potenciales

La tabla resultante proporciona una primera aproximación a la capacidad predictiva de cada variable, mostrando diferencias significativas en varias dimensiones entre el grupo elite y el resto de la población.

Table 8: Distribución y proporción de elite por tenencia de vivienda

Tenencia	N	N expandido	Proporción elite (ponderada)	Proporción muestra	Proporción
Propia pagada	6022	7424366	9.7%	0.551	
Propia pagándose	1052	2100421	36.1%	0.096	
Propia compartida (pagada)	39	65280	12.9%	0.004	
Propia compartida (pagándose)	3	6793	88.4%	0.000	
Arrendada	2097	3307550	8.4%	0.192	
Cedida	1305	1707647	3.9%	0.119	
Usufructo	301	406743	5.1%	0.028	
Ocupación irregular	74	111732	0.2%	0.007	
Poseedor irregular	28	33858	0.0%	0.003	

Table 9: Estadísticas de edad por grupo

elite_label	N	Media	DE	NA's (%)
10% superior	974	41.27	16.05	0%
Resto	9947	44.87	17.84	0%

Table 10: Distribución por sexo

sexo	N sin ponderar	N ponderado	N elite	% del total	% Elite (ponderado)	NA's (%)
Mujer	6308	7753344	477	57.8%	10.6%	0%
Hombre	4613	7411046	497	42.2%	14.0%	0%

Table 11: Distribución por zona

zona	N sin ponderar	N ponderado	N elite	% del total	% Elite (ponderado)	NA's (%)
Urbana	9307	13405487	916	85.2%	13.2%	0%
Rural	1614	1758903	58	14.8%	5.3%	0%

Table 12: Distribución por región

region_nombre	N sin ponderar	N ponderado	N elite	% del total	% Elite (ponderado)	NA's (%)
Metropolitana	1138	6412096	205	10.4%	18.9%	0%
Valparaíso	851	1541851	66	7.8%	7.7%	0%
Biobío	761	1289286	39	7.0%	5.9%	0%
Araucanía	664	776319	31	6.1%	5.1%	0%
Ñuble	645	399028	23	5.9%	3.4%	0%
O'Higgins	644	766289	40	5.9%	7.8%	0%
Los Lagos	633	687519	48	5.8%	7.7%	0%
Los Ríos	631	315343	34	5.8%	6.0%	0%
Coquimbo	623	641782	33	5.7%	6.5%	0%
Antofagasta	623	527209	85	5.7%	14.8%	0%
Tarapacá	623	285450	68	5.7%	11.6%	0%
Atacama	621	233615	57	5.7%	9.9%	0%
Maule	620	876150	26	5.7%	4.6%	0%
Aysén	618	80017	67	5.7%	10.1%	0%
Arica y Parinacota	617	191704	47	5.6%	7.6%	0%
Magallanes	609	140732	105	5.6%	20.4%	0%

Table 13: Distribución por nivel educacional

educ_rec	N sin ponderar	N ponderado	N elite	% del total	% Elite (ponderado)	NA's (%)
Hasta media	4613	6227468	184	42.2%	4.3%	0%
Hasta básica	2453	2879582	34	22.5%	1.3%	0%
Universitaria	2365	3843799	535	21.7%	30.5%	0%
Técnica superior	1302	1879090	125	11.9%	9.9%	0%
Postgrado	188	334451	96	1.7%	58.7%	0%

Table 14: Distribución por tenencia de vivienda

tenencia_vivienda	N sin ponderar	N ponderado	N elite	% del total	% Elite (ponderado)	NA's (%)
Propia pagada	6022	7424366	446	55.1%	9.7%	0%
Arrendada	2097	3307550	131	19.2%	8.4%	0%
Cedida	1305	1707647	51	11.9%	3.9%	0%
Propia pagándose	1052	2100421	329	9.6%	36.1%	0%
Usufructo	301	406743	13	2.8%	5.1%	0%
Ocupación irregular	74	111732	1	0.7%	0.2%	0%
Propria compartida (pagada)	39	65280	2	0.4%	12.9%	0%
Poseedor irregular	28	33858	0	0.3%	0.0%	0%
Propria compartida (pagándose)	3	6793	1	0.0%	88.4%	0%

Por ejemplo, el 10% es más joven en promedio (41,27 vs 44,87) y tiene mayor proporción de hombres (14% vs 10,6%) y hay fuerte concentración en urbes (13,2% vs 5,3%).

Sin embargo, hay importantes diferencias en la calidad de los datos. Las variables básicas tienen pocos valores faltantes (0% NA's) y hay buena representatividad en categorías principales.

## Variables de bienestar de EBS

Decisiones importantes aquí: - Uso de scale() para normalizar variables - Restricción a valores 1-5 en la escala original - Codificación explícita de NA para otros valores

Patrones clave identificados - Diferencias positivas para el 10% superior: Mayor logro de metas, mayor apoyo empleabilidad; mejor balance trabajo-vida. - Diferencias negativas/neutrales: Menor interferencia doméstica; Diferencias mínimas en satisfacción con tiempo y flexibilidad de ausencias.

Implicancias para el análisis: - Las variables de bienestar muestran diferencias sistemáticas. - Potencial predictivo moderado pero consistente. - Sin embargo, alto número de NAs. - Se podría considerar un índice compuesto de bienestar.

Table 15: Comparación de variables de bienestar entre grupos (valores estandarizados)

Variable	Media no elite	DE no elite	Media elite	DE elite
Apoyo a empleabilidad	-0.03	1.00	0.20	0.95
Balance trabajo-vida	-0.02	1.00	0.14	0.96
Satisfacción con tiempo	-0.01	1.00	0.05	1.02
Logro de metas	-0.05	1.01	0.40	0.84
Interferencia doméstica	0.01	1.01	-0.09	0.95
Flexibilidad ausencias	0.78	0.42	0.83	0.38

## Propuesta de modelo de Machine Learning

A partir de lo anterior, se sugiere un Random Forest para predecir la pertenencia al 10% superior de ingresos. Esta elección se fundamenta en varias características específicas de nuestros datos y objetivos analíticos. En primer lugar, hemos identificado relaciones no lineales importantes, particularmente en variables como edad y educación. Random Forest, al construir múltiples árboles de decisión, puede capturar estas no linealidades sin necesidad de especificación explícita.

Además, nuestros datos presentan una estructura jerárquica territorial, con importantes efectos de interacción entre variables socioeconómicas y ubicación geográfica. Por ejemplo, el significado de un determinado nivel de ingreso en variables de bienestar podría variar según la región (Mac-Clure et al., 2014). Random Forest maneja naturalmente estas interacciones a través de su proceso de construcción de árboles y la selección aleatoria de variables en cada división.

El desbalance moderado en nuestra variable objetivo (9% vs 91%) también se alinea con las fortalezas de Random Forest, especialmente cuando se implementa con pesos de clase ajustados. Esto es particularmente relevante dado que nuestro interés principal es identificar correctamente a los miembros de un grupo minoritario.

Para la implementación específica, proponemos utilizar la librería ranger en R, que permite un manejo eficiente de datos de gran escala y ofrece opciones para importancia de variables permutacional. Sugerimos una configuración inicial con 1000 árboles, mtry igual a la raíz cuadrada del número de predictores, y un tamaño mínimo de nodo de 5 observaciones. La validación cruzada se realizará mediante una estratificación espacial por región para mantener la representatividad geográfica.

Las variables de bienestar subjetivo, que presentan patrones significativos de valores faltantes (42%), serán manejadas mediante imputación múltiple antes del modelamiento, creando cinco conjuntos de datos completos. También estamos considerando construir un índice de bienestar.

## Bibliografía:

- Gilens, M. (2012). *Affluence and influence: Economic inequality and political power in America*. Princeton, NJ: Princeton University Press.
- Hernando, M. & Mitchell, G. (2023). *Uncomfortably off: Why the Top 10% of Earners Should Care about Inequality*. Bristol: Policy Press.
- Mac-Clure, O., Barozet, E., & Maturana, V. (2014). Desigualdad, clase media y territorio en Chile: ¿clase media global o múltiples mesocracias según territorios?. *Revista EURE - Revista De Estudios Urbanos Regionales*, 40(121).
- Méndez, M.L., & Gayo, M. (2024). *The Politics of the Elite: Ideological Orientations, Mothering, and Social Mobilities in Neoliberal Chile*. London: Routledge.
- Rothman, K. J., & Greenland, S. (2018). Planning Study Size Based on Precision Rather Than Power. *Epidemiology*, 29(5), 599-603.
- van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8), 2455-2474.
- Wilkinson, R. & Pickett, K. (2010). *The spirit level: Why equality is better for everyone*. London: Penguin.
- Wilkinson, R. & Pickett, K. (2019). *The inner level: How more equal societies reduce stress, rest*