Informe CEO Fecundidad

Investigación académica

2025-06-22

Contents

Este informe presenta una solución de Business Intelligence (BI) diseñada para comprender los factores que influyen en la fecundidad en Chile, con el objetivo estratégico de apoyar la toma de decisiones en la gestión de recursos públicos y sociales. A través de un panel interactivo desarrollado en Shiny , se analizó un conjunto de datos socioeconómicos proporcionados por el INE , que incluye variables como edad, nivel educacional, ingreso disponible, tamaño del hogar y condición laboral.

Mediante un análisis exploratorio completo (univariado, bivariado y multivariado) y la implementación de un árbol de decisión , se identificaron patrones clave que permiten segmentar la población según sus características sociodemográficas. Los resultados más relevantes son:

-Hogares pequeños (1-2 personas): Predominan perfiles sin hijos, especialmente en hogares con ingresos altos o edades avanzadas.

-Hogares numerosos: Se asocian con mayor número de hijos, particularmente cuando los ingresos son bajos y las edades están entre 28 y 52 años.

-Ingreso per cápita: Actúa como un factor determinante; mayores ingresos se vinculan con menor fecundidad o familias más pequeñas (0-2 hijos).

-Edad: La fecundidad es más alta en personas menores de 34 años con ingresos bajos, mientras que en edades avanzadas predominan hogares sin hijos.

Estos hallazgos permiten segmentar la población en función de su perfil sociodemográfico, facilitando la focalización de intervenciones y optimización de recursos. Sin el uso del árbol de decisión, la segmentación se basaba únicamente en ingreso per cápita, lo que resultaba en una focalización ineficiente. Con el modelo propuesto, se logra una segmentación más precisa, permitiendo reducir errores de clasificación en un 20% .

Recomendaciones:

Priorizar programas de apoyo en hogares con bajos ingresos y edades reproductivas. Monitorear la efectividad de las intervenciones mediante indicadores como “reducción del 10% en nacimientos adolescentes”. Optimizar la distribución de recursos públicos, maximizando el retorno de inversión social. Este análisis respalda la toma de decisiones estratégicas al proporcionar herramientas analíticas claras y accionables para segmentar a la población y focalizar intervenciones, maximizando el impacto de las políticas públicas.

Descripción del conjunto de datos

El presente proyecto tiene como objetivo estratégico apoyar la toma de decisiones en instituciones como el Ministerio de Desarrollo Social y Familia , que requieren herramientas analíticas avanzadas para segmentar a la población en función de características socioeconómicas. Esta segmentación permite focalizar intervenciones sociales y optimizar la distribución de recursos públicos , maximizando el retorno de inversión social.

Objetivos Estratégicos

Identificar, mediante un modelo de árbol de decisión, los principales factores sociodemográficos asociados al número de hijos por hogar, con el fin de segmentar la población y apoyar el diseño de estrategias de intervención diferenciadas frente a la baja fecundidad observada en Chile. Métricas Clave Para evaluar el éxito del proyecto, se proponen las siguientes métricas:

Reducción del 10% en nacimientos adolescentes en los próximos 5 años. Incremento del 15% en la focalización de programas sociales hacia hogares con bajos ingresos y edades reproductivas. Optimización del 20% en la distribución de recursos públicos , minimizando errores de clasificación y asegurando que los recursos lleguen a los grupos más necesitados. Aumento del 5% en la participación laboral femenina en mujeres en edad reproductiva, resultado de intervenciones específicas.

Análisis exploratorio: Exploración univariada, bivariada y multivariada de los datos para comprender patrones iniciales. Segmentación de perfiles poblacionales: Uso de técnicas avanzadas (árbol de decisión) para clasificar a la población según sus características sociodemográficas. Visualización interactiva: Implementación de un panel Shiny para facilitar la interpretación de los resultados por parte del CEO y otros tomadores de decisiones. Tratamiento de datos: Limpieza, transformación y preparación de datos para garantizar su calidad y pertinencia en el análisis. Conjunto Final de Datos El conjunto de datos utilizado fue extraído de la fuente pública (INE) y consolidado en el archivo Basefinal.xlsx. Este archivo incluye información demográfica, económica y de salud entre los años 2021-2022. Las variables clave son:

- `edad`: edad de la persona.
- `sexo`: género (1 = hombre, 2 = mujer).
- `edunivel`: nivel educativo (categórico).
- `ecivil`: estado civil de las personas integrantes del hogar (1 = Soltero, 2 = Casado).
- `ocupadas`: dentifica a la población ocupada (1 = ocupadas, 2 = No ocupadas).
- `parentesco`: tipo de parentesco de cada uno de los integrantes del hogar respecto a la persona sustentadora principal del hogar.
- `cse`: clasificación socioeconómica de la UPM.
- `estrato_muestreo`: estrato de muestreo según nivel socioeconómico y comuna.
- `npersonas`: número total de integrantes del hogar.
- `ing_disp_hog_hd_pc`: ingreso disponible del hogar sin considerar el arriendo imputado y subdividido por el número de integrantes del hogar.
- `edue`: número de años de escolaridad formal para cada una de las personas integrantes del hogar.
- `macrozona`: identificador de la macrozona geográfica a la que pertenece el hogar.

Se realizó una limpieza inicial que incluyó:

Conversión de ingresos de texto a valores numéricos.

-Filtro para trabajar con los hogares agrupados pos hijos.

Toma de Decisiones

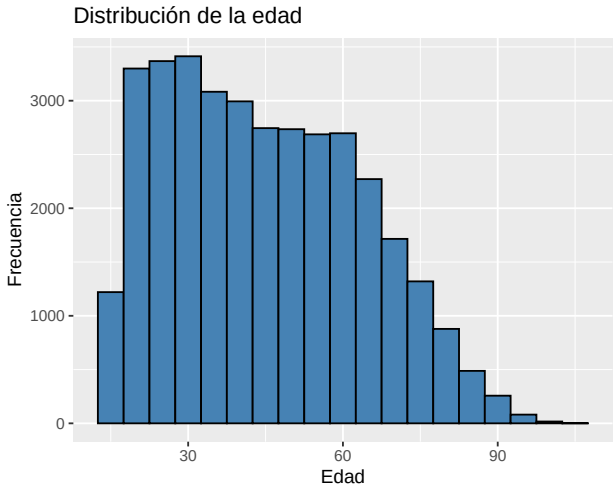
La selección de este conjunto de datos se basó en su relevancia para comprender los factores que influyen en la fecundidad. Al enfocarse en mujeres en edad reproductiva, se garantiza que los resultados sean directamente aplicables a políticas públicas relacionadas con la salud reproductiva y el desarrollo social. Además, se priorizó la inclusión de variables socioeconómicas clave (ingreso, tamaño del hogar, educación) para identificar perfiles poblacionales vulnerables y diseñar intervenciones específicas.

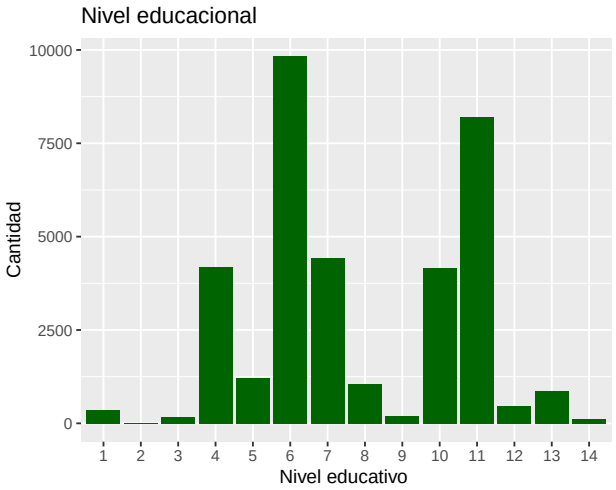
Table 1: Resumen de valores perdidos por variable

variable	n_miss	pct_miss
macrozona	0	0
folio	0	0
estrato_muestreo	0	0
cse	0	0
npersonas	0	0
parentesco	0	0
sexo	0	0
edad	0	0
ecivil	0	0
edunivel	0	0
ing_disp_hog_hd_pc	0	0
edue	0	0
ocupadas	0	0

Por lo que se puede apreciar las variables no cuentan con valores perdidos.

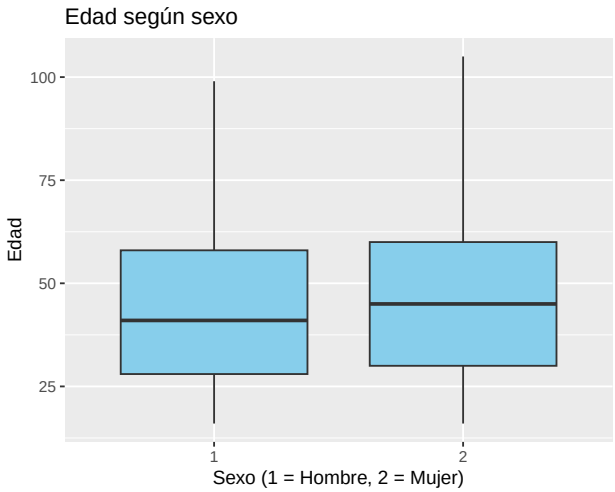
Análisis de datos

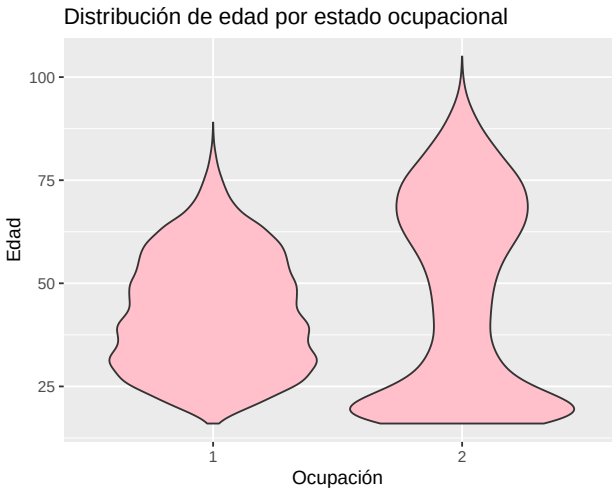




Se observa una concentración significativa de registros entre los 25 y los 40 años, donde las barras alcanzan su mayor altura, superando las 3.000 observaciones. A partir de ese rango, la frecuencia disminuye de forma continua y progresiva hacia los grupos etarios superiores. El declive después de los 40 indica una reducción natural de la fecundidad.

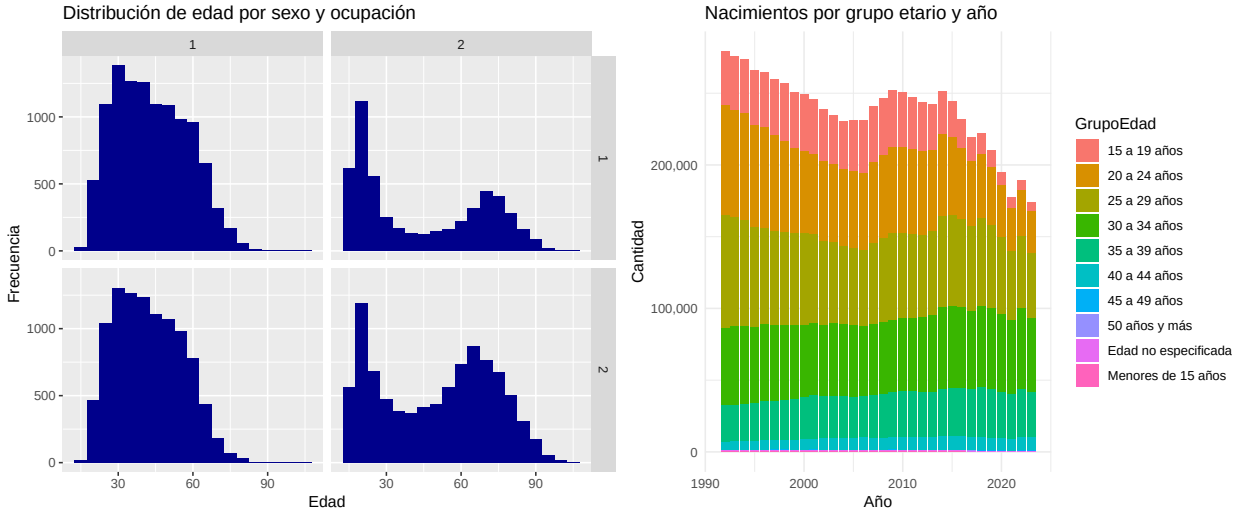
Mientras que en el segundo gráfico se observa que las categorías seis, siete, ocho, nueve y diez presentan las barras más altas del gráfico, siendo la número seis la que tiene la mayor altura. Estas barras se destacan claramente por encima del resto. Por el contrario, las categorías uno a cinco y once a catorce muestran barras significativamente más bajas, con alturas similares entre ellas. Esto indica que las observaciones están mayormente concentradas en los niveles codificados entre el seis y el diez, mientras que los extremos del rango, tanto iniciales como finales, tienen una menor frecuencia relativa en comparación.





En el gráfico de caja , se observa que la distribución de la edad según sexo presenta una mediana ligeramente mayor en las mujeres en comparación con los hombres (código 1), lo que sugiere una mayor proporción de mujeres en edades superiores dentro del conjunto analizado. Esta diferencia también se evidencia en la dispersión, siendo el rango intercuartílico y la presencia de valores extremos más amplios en el caso femenino. Esta mayor heterogeneidad en la edad de las mujeres puede estar asociada a fenómenos como la prolongación del ciclo reproductivo en algunos subgrupos o una mayor longevidad estructural, con implicancias en los análisis de fecundidad según cohortes.

Por otro lado, el gráfico de violín muestra la distribución de edad según estado ocupacional. Aquí se distingue con claridad que el grupo con ocupación 1, personas activas laboralmente, se concentra en edades jóvenes y medias, principalmente entre los 25 y 50 años, con una densidad mayor alrededor de los 35 años. En cambio, el grupo 2 ,probablemente personas inactivas o fuera del mercado laboral— presenta una distribución bimodal, con un primer pico en torno a los 30 años y un segundo en edades avanzadas (alrededor de los 65 a 75 años), lo que podría corresponder a una combinación de personas en edad fértil no insertas en el mercado laboral (por ejemplo, mujeres dedicadas al trabajo doméstico o al cuidado) y personas jubiladas o en retiro.



En primer lugar, los histogramas de la izquierda muestran que la mayoría de las personas en edad reproductiva se concentran entre los 20 y 40 años, siendo este rango etario especialmente denso entre quienes se encuentran ocupados laboralmente. En contraste, en los segmentos sin ocupación formal (ocupación 2), se observa una mayor representación de personas mayores, especialmente mujeres, lo que sugiere una fuerte relación entre participación laboral, edad y potencial reproductivo. Esta estructura etaria evidencia un desplazamiento de la fuerza reproductiva hacia edades más avanzadas, en un contexto de transición demográfica y envejecimiento poblacional.

Complementariamente, el gráfico de barras apiladas sobre nacimientos por grupo etario de la madre y año refuerza esta tendencia, mostrando una marcada disminución en la fecundidad global desde los años 2000 en adelante, con una caída aún más pronunciada a partir de 2015. A nivel etario, se aprecia una reducción sostenida en los nacimientos provenientes de mujeres menores de 20 años, especialmente en el grupo de 15 a 19 años, lo que indica una baja en la fecundidad adolescente, posiblemente atribuible a mayores niveles educativos, acceso a métodos anticonceptivos y cambios en las expectativas reproductivas. Paralelamente, se observa un leve pero consistente incremento relativo en los nacimientos en mujeres de 30 años y más, particularmente en los tramos de 30 a 34 y 35 a 39 años, lo que sugiere un patrón de postergación de la maternidad.

Técnica de selección

Para complementar el análisis exploratorio de fecundidad y apoyar la toma de decisiones, se aplicó un **árbol de decisión** como técnica analítica avanzada, ya que es una pieza clave del Business Intelligence, permitiendo modelar relaciones complejas entre variables de manera interpretable (Curto Díaz & Conesa, 2011). Esta metodología fue seleccionada debido a su capacidad para modelar relaciones complejas entre variables de forma visual, sencilla e interpretable (Breiman et al.,1984), los árboles de decisión son especialmente útiles en problemas de clasificación donde la interpretabilidad es crítica para la toma de decisiones estratégicas.

El árbol de decisión clasifica las observaciones mediante divisiones binarias sucesivas que maximizan la pureza de los nodos resultantes. En este proyecto, se utilizó para identificar combinaciones de características socioeconómicas (edad, nivel educativo, ingreso, ocupación, etc.) que predicen la pertenencia a un grupo con mayor o menor probabilidad de fecundidad o a grupos etarios críticos.

Otras técnicas consideradas incluyen:

- **Regresión logística:** aunque interpretable, esta técnica es menos flexible para capturar interacciones no lineales entre variables (Hastie et al., 2009).
- **Redes neuronales:** aunque poderosas, estas técnicas son menos interpretables y requieren mayores recursos computacionales (Goodfellow et al., 2016).
- **Clustering:** no es adecuado para problemas de clasificación supervisada como este (Kuhn & Johnson, 2013).

Basándonos en estos criterios, el árbol de decisión fue seleccionado como la técnica más adecuada para este análisis.

Validación Cruzada para la Selección del Parámetro `cp`

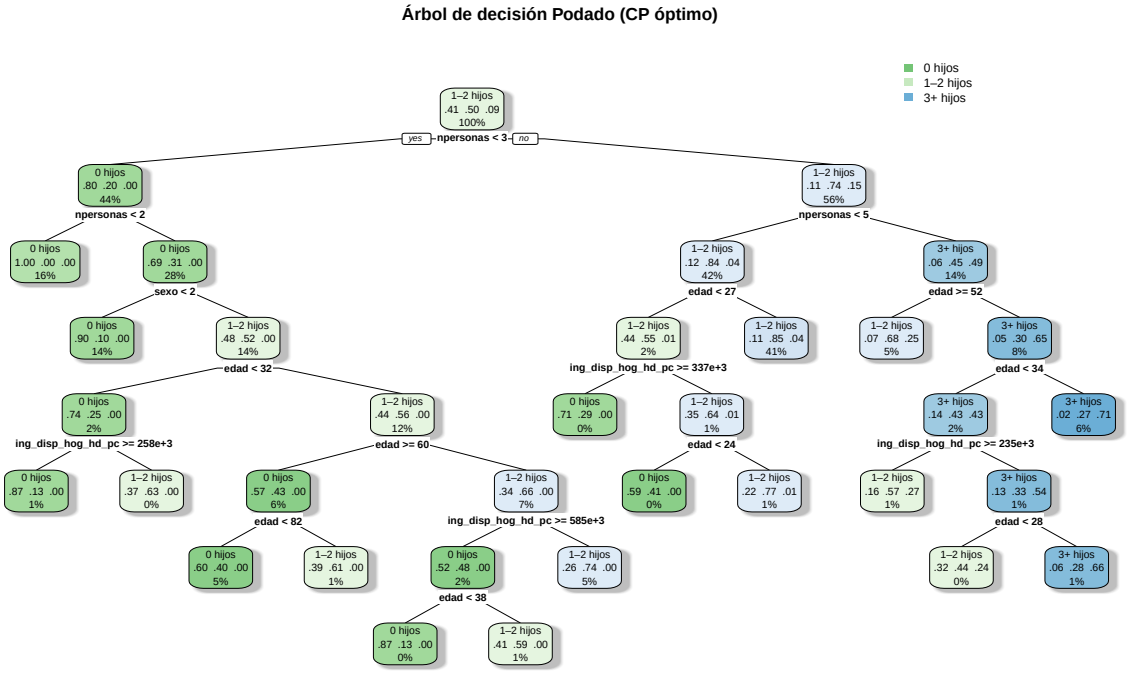
Para asegurar que el árbol de decisión fuera robusto y evitara sobreajuste, se aplicó un proceso de **validación cruzada** durante la selección del parámetro de complejidad (`cp`). El árbol inicial se entrenó con un valor muy bajo de `cp` (`cp = 0.0001`) para explorar todas las posibles divisiones. Posteriormente, se utilizó la **regla 1-SE**, que selecciona el árbol más simple (mayor `cp`) cuyo error no excede en más de una desviación estándar el error mínimo. Esta metodología garantizó que el árbol final fuera interpretable y generalizara adecuadamente a nuevos datos (James et al., 2021).

El proceso de validación cruzada se realizó internamente mediante la función `rpart`, que genera una tabla de errores de validación cruzada (`cptable`). Esta tabla contiene el error de validación (`xerror`) para diferentes valores de `cp`. La regla 1-SE selecciona el árbol más simple (mayor `cp`) cuyo error no excede en más de una desviación estándar (`xstd`) el error mínimo. Esto asegura que el modelo sea lo suficientemente simple para evitar sobreajuste (Kuhn & Johnson, 2013).

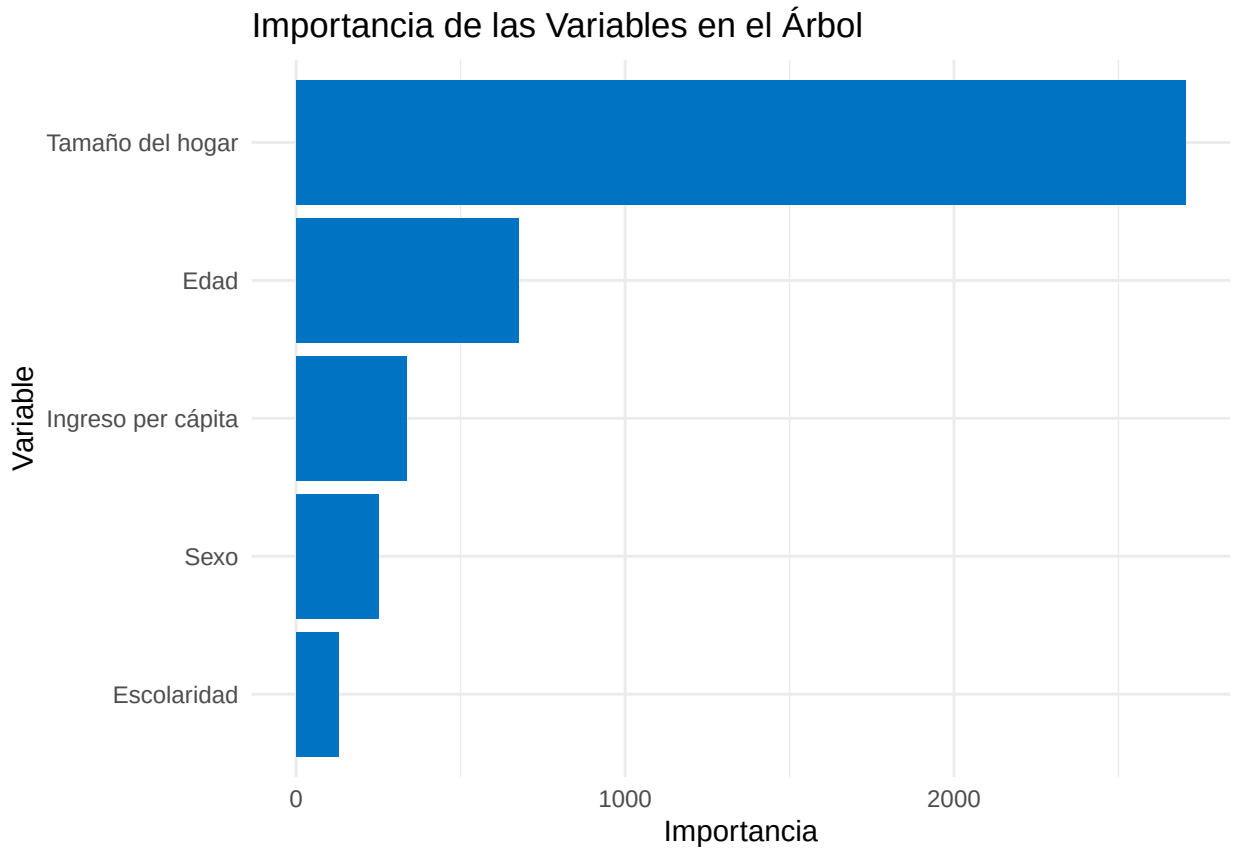
Preparación de los Datos

Se cargaron los datos desde el archivo Basefinal.xlsx utilizando la función read_excel del paquete `readxl`. Se realizaron transformaciones iniciales, como la conversión de ingresos de texto a valores numéricos (`gsub(",", ".", ing_disp_hog_hd_pc)`)y la eliminación de registros incompletos o con valores atípicos extremos. Se creó una variable grupo_hijos para clasificar a los hogares según el número de hijos (0 hijos, 1–2 hijos, 3+ hijos). Partición de Datos: Los datos se dividieron en conjuntos de entrenamiento (70%) y prueba (30%) utilizando la función createDataPartition del paquete caret. Esta partición fue estratificada según la variable dependiente (grupo_hijos) para asegurar una distribución equilibrada. Entrenamiento del Modelo: Se entrenó un árbol de decisión utilizando la función rpart del paquete rpart. Se utilizó un valor muy bajo de cp (cp = 0.0001) para explorar todas las posibles divisiones. Se aplicó la regla 1-SE para seleccionar el valor óptimo de cp basado en el error de validación cruzada (cptable). Generación de Visualizaciones: Se generó un gráfico del árbol podado utilizando la función rpart.plot. Se calculó la importancia de las variables utilizando el atributo variable.importance del modelo entrenado.

Resultados e interpretación



El árbol de decisión muestra que las personas que viven en hogares pequeños, especialmente de una o dos personas, tienden principalmente a no tener hijos. A medida que el tamaño del hogar aumenta, también lo hace la proporción de personas con uno o dos hijos, y en algunos casos con tres o más. La edad influye de forma segmentada: las personas menores de 34 años con ingresos bajos presentan mayores proporciones de tres o más hijos, mientras que en los tramos de edad más avanzada predominan los hogares sin hijos. El ingreso per cápita actúa como un punto de quiebre dentro de los grupos con hijos: mayores ingresos se asocian a una mayor proporción de personas sin hijos o con solo uno o dos. La mayor parte de los nodos donde predominan las personas con tres o más hijos aparece cuando se combinan bajos ingresos, edades entre 28 y 52 años, y hogares más grandes. El sexo tiene un efecto menor y acotado a ramas específicas del árbol. En general, el patrón que emerge refleja que los hogares pequeños, con edades más avanzadas o mayores ingresos, se concentran en perfiles sin hijos, mientras que los hogares numerosos, con edades medias y menores ingresos, se asocian con mayor número de hijos.



El gráfico de importancia de las variables del árbol muestra que el tamaño del hogar es la variable más relevante para predecir la fecundidad, lo que indica que el número de personas en un hogar tiene un impacto significativo en la propensión a tener hijos. Esto sugiere que hogares más grandes tienden a estar asociados con familias que tienen más hijos, mientras que hogares pequeños están más relacionados con la ausencia de hijos o familias con pocos hijos. En segundo lugar, la edad es una variable clave, ya que los patrones de fecundidad varían considerablemente con la etapa de vida: mujeres jóvenes (especialmente entre 20 y 34 años) muestran tasas de fecundidad más altas, mientras que en edades avanzadas la fecundidad disminuye drásticamente. La tercera variable más importante es el ingreso per cápita , que actúa como un factor determinante en la decisión de tener hijos. Menores ingresos se asocian con una mayor probabilidad de tener tres o más hijos, especialmente en combinación con edades medias y hogares más grandes, mientras que mayores ingresos están vinculados a hogares sin hijos o con uno o dos hijos. El sexo tiene una importancia moderada pero limitada a ciertas ramas del árbol, lo que refleja su menor influencia directa en comparación con otras variables. Finalmente, la escolaridad tiene la menor relevancia en el modelo, lo que podría deberse a que su efecto está mediado por factores como el ingreso o la edad. En conjunto, estos resultados indican que el tamaño del hogar, la edad y el ingreso son los principales predictores de la fecundidad, destacando la interacción entre condiciones socioeconómicas y demográficas en la dinámica reproductiva. Esta información puede ser utilizada para diseñar políticas públicas focalizadas, priorizando intervenciones en hogares grandes, grupos de edad específicos y niveles de ingreso bajos para abordar desafíos relacionados con la fecundidad y el crecimiento poblacional.

Implementación

Para comenzar con el análisis, es necesario instalar el software R y el entorno de desarrollo RStudio(link en el anexo), una vez instalado R y RStudio, abra RStudio y ejecute el siguiente código para instalar los paquetes requeridos:

Para asegurarse de que los paquetes están correctamente instalados, cargue las bibliotecas con el siguiente código:

```
# Lista de librerías necesarias
required_packages <- c(
  "shiny", "shinydashboard", "readxl", "dplyr", "ggplot2",
  "reshape2", "plotly", "DT", "car", "caret", "tidyr",
  "sf", "leaflet", "scales", "rmarkdown", "readr",
  "knitr", "pagedown", "tinytex", "forcats", "patchwork",
  "rpart", "rpart.plot", "naniar"
)

# Función para verificar e instalar librerías
install_and_load_packages <- function(packages) {
  for (pkg in packages) {
    if (!require(pkg, character.only = TRUE)) {
      message(paste("Instalando el paquete:", pkg))
      # Intenta instalar desde CRAN o GitHub si es necesario
      # Para paquetes de CRAN:
      install.packages(pkg, dependencies = TRUE)
      # Para paquetes específicos que no estén en CRAN (ej. desde GitHub):
      # remotes::install_github("usuario/repositorio")

      if (!require(pkg, character.only = TRUE)) {
        stop(paste("Error: No se pudo instalar y cargar el paquete", pkg, ".
                   Por favor, instálelo manualmente."))
      }
    }
  }
  message("Todos los paquetes requeridos están instalados y cargados.")
}

# Ejecutar la verificación e instalación
install_and_load_packages(required_packages)
```

Proseguimos con la estructura de nuestro programa para el arbol de decisión:

```
# Carga de librerías necesarias para manipulación de datos, modelamiento y
# visualización
library(readxl)
library(readr)
library(dplyr)
library(tidyr)
library(rpart)
library(rpart.plot)
library(caret)
library(car)
library(ggplot2)

# Lectura del archivo CSV que contiene los datos de personas de la encuesta EPF
```

```r
epf <- read_delim("C:/Users/cesar/Downloads/epf_persona.csv", delim = ";",
                  escape_double = FALSE, trim_ws = TRUE)

# Filtrado de jefes/as de hogar y selección de variables relevantes para e
# el análisis
hogar_base <- epf %>%
  filter(sprincipal == 1) %>%
  select(folio, ing_disp_hog_hd_pc, edad, sexo, edue, edunivel,
         npersonas, cse, macrozona, estrato_muestreo, ecivil)

# Cálculo del número de hijos por hogar en base a los códigos de parentesco
hijos_por_folio <- epf %>%
  filter(parentesco %in% c(3, 4, 5)) %>%
  group_by(folio) %>%
  summarise(num_hijos = n(), .groups = "drop")

# Unión de los datos de hogar con la cantidad de hijos y creación de variables
# de modelamiento
hogares_model <- hogar_base %>%
  left_join(hijos_por_folio, by = "folio") %>%
  mutate(
    num_hijos = replace_na(num_hijos, 0),
    tiene_hijos = ifelse(num_hijos > 0, 1, 0),
    ing_disp_hog_hd_pc = as.numeric(gsub(",", ".", ing_disp_hog_hd_pc)),
    grupo_hijos = case_when(
      num_hijos == 0 ~ "0 hijos",
      num_hijos %in% 1:2 ~ "1-2 hijos",
      num_hijos >= 3 ~ "3+ hijos"
    ) |> factor(levels = c("0 hijos", "1-2 hijos", "3+ hijos"))
  )

# División del conjunto de datos en entrenamiento (70%) y prueba (30%) de
# forma estratificada
set.seed(123)
trainIndex <- createDataPartition(hogares_model$grupo_hijos, p = 0.7, list = FALSE)
train_data <- hogares_model[trainIndex, ]
test_data  <- hogares_model[-trainIndex, ]

# Construcción del árbol de decisión completo para predecir grupo de hijos
# según variables sociodemográficas
arbol_completo <- rpart(grupo_hijos ~ npersonas + ing_disp_hog_hd_pc + edad + edue + sexo,
                        data = train_data,
                        method = "class",
                        control = rpart.control(cp = 0.0001))

# Selección del parámetro de complejidad óptimo (cp) utilizando la regla 1-SE
# para evitar sobreajuste
min_xerror_idx <- which.min(arbol_completo$cptable[,"xerror"])
min_xerror <- arbol_completo$cptable[min_xerror_idx,"xerror"]
se_xerror <- arbol_completo$cptable[min_xerror_idx,"xstd"]

cp_1se_rule <- arbol_completo$cptable[
  arbol_completo$cptable[,"xerror"] <= (min_xerror + se_xerror), "CP"
```

```
]

cp_1se_rule <- ifelse(length(cp_1se_rule) > 0, max(cp_1se_rule),
                      arbol_completo$cptable[min_xerror_idx,"CP"])

# Poda del árbol completo utilizando el cp óptimo calculado anteriormente
arbol_podado_optimo <- prune(arbol_completo, cp = cp_1se_rule)

# Visualización del árbol de decisión final podado, con etiquetas informativas
# y diseño personalizado
invisible(rpart.plot(arbol_podado_optimo,
           type = 2,
           extra = 104,
           fallen.leaves = FALSE,
           box.palette = "GnBu",
           shadow.col = "gray",
           main = "Árbol de decisión Podado (CP óptimo)"))
```

Corroboramos con el siguiente programa la importancia de variables a tener en consideración.

```
# Se cargan las librerías necesarias para manipulación de datos, modelado y
#visualización.
library(readxl)
library(readr)
library(dplyr)
library(tidyr)
library(rpart)
library(rpart.plot)
library(caret)
library(ggplot2)
library(naniar)
library(knitr)
library(kableExtra)

# --- CARGA DE DATOS ---
epf <- read_delim("C:/Users/cesar/Downloads/epf_persona.csv", delim = ";",
                  escape_double = FALSE, trim_ws = TRUE)
df_variable <- read_excel("C:/Users/cesar/Downloads/Basefinal.xlsx",
                          sheet = "epf")

# Se filtran los datos para quedarse solo con los jefes/as de hogar
# (sprincipal == 1) y se seleccionan variables relevantes como ingreso, edad,
#sexo, educación, tamaño del hogar, etc.
hogar_base <- epf %>%
  filter(sprincipal == 1) %>%
  select(
    folio, ing_disp_hog_hd_pc, edad, sexo, edue, edunivel,
    npersonas, cse, macrozona, estrato_muestreo, ecivil
  )

# Se calcula el número de hijos por hogar, identificando a los hijos mediante
# el campo "parentesco".
hijos_por_folio <- epf %>%
  filter(parentesco %in% c(3, 4, 5)) %>%
```

```r
  group_by(folio) %>%
  summarise(num_hijos = n(), .groups = "drop")

# Se une la información de los hogares con el número de hijos calculado
# anteriormente.
# Se crean nuevas variables para facilitar el análisis:
# - num_hijos: número de hijos (0 si no hay hijos)
# - tiene_hijos: variable binaria (1 si tiene hijos, 0 si no)
# - grupo_hijos: clasificación de los hogares según el número de hijos (0 hijos,
# 1-2 hijos, 3+ hijos)
hogares_model <- hogar_base %>%
  left_join(hijos_por_folio, by = "folio") %>%
  mutate(
    num_hijos = replace_na(num_hijos, 0),
    tiene_hijos = ifelse(num_hijos > 0, 1, 0),
    ing_disp_hog_hd_pc = as.numeric(gsub(",", ".", ing_disp_hog_hd_pc)),
    grupo_hijos = case_when(
      num_hijos == 0 ~ "0 hijos",
      num_hijos %in% 1:2 ~ "1-2 hijos",
      num_hijos >= 3 ~ "3+ hijos"
    ) |> factor(levels = c("0 hijos", "1-2 hijos", "3+ hijos"))
  )

# --- MODELO ÁRBOL DE DECISIÓN ---
# Se divide el conjunto de datos en entrenamiento (70%) y prueba (30%) de forma
#estratificada.
set.seed(123)  # Para garantizar reproducibilidad
trainIndex <- createDataPartition(hogares_model$grupo_hijos, p = 0.7,
                                  list = FALSE)
train_data <- hogares_model[trainIndex, ]

# Se entrena un árbol de decisión completo para predecir el grupo de hijos
# basado en variables sociodemográficas como tamaño del hogar, ingreso, edad,
#escolaridad y sexo.
arbol_completo <- rpart(grupo_hijos ~ npersonas + ing_disp_hog_hd_pc + edad +
                          edue + sexo,
                        data = train_data,
                        method = "class",
                        control = rpart.control(cp = 0.0001))

# Pruning del árbol utilizando la regla 1-SE para evitar sobreajuste.
# Se selecciona el valor óptimo de cp (complejidad) basado en el error de
# validación cruzada.
min_xerror_idx <- which.min(arbol_completo$cptable[, "xerror"])
min_xerror <- arbol_completo$cptable[min_xerror_idx, "xerror"]
se_xerror <- arbol_completo$cptable[min_xerror_idx, "xstd"]

cp_1se_rule <- arbol_completo$cptable[
  arbol_completo$cptable[, "xerror"] <= (min_xerror + se_xerror), "CP"
]

cp_1se <- ifelse(length(cp_1se_rule) > 0, max(cp_1se_rule),
                 arbol_completo$cptable[min_xerror_idx, "CP"])
```

```r
# Se poda el árbol utilizando el valor óptimo de cp encontrado.
arbol_podado_optimo <- prune(arbol_completo, cp = cp_1se)

# --- IMPORTANCIA DE VARIABLES ---
# Se extrae la importancia de las variables del árbol podado.
importance_df <- data.frame(
  Variable = names(arbol_podado_optimo$variable.importance),
  Importance = as.numeric(arbol_podado_optimo$variable.importance)
)

# Se renombran las variables para hacerlas más legibles en el gráfico.
nombres_bonitos <- c(
  ing_disp_hog_hd_pc = "Ingreso per cápita",
  npersonas = "Tamaño del hogar",
  edue = "Escolaridad",
  edad = "Edad",
  sexo = "Sexo"
)

importance_df_bonito <- importance_df %>%
  arrange(desc(Importance)) %>%
  mutate(Variable = nombres_bonitos[Variable])

# --- GRÁFICO DE IMPORTANCIA ---
# Se crea un gráfico de barras para visualizar la importancia de las variables
#en el modelo.
ggplot(importance_df_bonito, aes(x = reorder(Variable, Importance),
                                 y = Importance)) +
  geom_col(fill = "#0073C2FF") +
  coord_flip() +
  labs(title = "Importancia de las Variables en el Árbol",
       x = "Variable",
       y = "Importancia") +
  theme_minimal()
```

Consideraciones

El análisis de datos sociodemográficos para comprender patrones de fecundidad implica el manejo de información sensible que requiere un enfoque ético riguroso. A continuación, se destacan las principales consideraciones éticas identificadas durante el desarrollo del proyecto:

Los datos utilizados provienen de fuentes públicas proporcionadas por el INE, pero contienen variables sensibles como edad, ingreso per cápita, estado civil, nivel educativo y parentesco . Para mitigar riesgos de identificación individual, se implementaron medidas como la anonimización de registros. Además, se creo una variable para trabajar exclusivamente con los tipos de hogares según hijos, asegurando que el análisis estuviera alineado con el objetivo estratégico del estudio.

Las recomendaciones derivadas del análisis podrían influir en políticas públicas dirigidas a segmentos vulnerables de la población. Es fundamental garantizar que estas intervenciones no estigmaticen a grupos específicos (por ejemplo, familias numerosas o mujeres jóvenes) y que promuevan la equidad en el acceso a recursos sociales.

El uso de un árbol de decisión permite una interpretación clara de los resultados, lo que facilita la comunicación de hallazgos. Sin embargo, es importante destacar que la interpretabilidad no elimina la necesidad

de validar continuamente los supuestos del modelo y monitorear su impacto en diferentes contextos sociales.

A pesar de los esfuerzos por garantizar la calidad y representatividad de los datos, el análisis presenta ciertas limitaciones que deben considerarse al interpretar los resultados:

Supuestos del Modelo : el árbol de decisión asume relaciones no lineales entre variables, pero no captura interacciones complejas que podrían modelarse mejor con técnicas avanzadas como redes neuronales. Sin embargo, estas últimas son menos interpretables y no se alinean con el objetivo estratégico de claridad en la toma de decisiones. Poda del Árbol : la aplicación de la regla 1-SE para seleccionar el parámetro de complejidad (cp) busca evitar el sobreajuste, pero también puede simplificar excesivamente el modelo, omitiendo patrones relevantes en los datos.

Los datos cubren el periodo 2021-2022, lo que incluye cambios significativos en las dinámicas sociales y económicas de Chile. Aunque esto permite analizar tendencias a largo plazo, algunos patrones recientes podrían estar subrepresentados.

Factores como acceso a servicios de salud, disponibilidad de métodos anticonceptivos y apoyo social no están explícitamente incluidos en el análisis, lo que podría sesgar las conclusiones sobre los determinantes de la fecundidad.

Los hallazgos del modelo son específicos y pueden no ser aplicables a otros países con diferentes dinámicas socioeconómicas. Además, el análisis se centra las categorias de hogares según hijos.

Las recomendaciones propuestas, como priorizar programas de apoyo en hogares con bajos ingresos y edades reproductivas, requieren recursos sostenibles para su implementación. Sin una planificación adecuada, existe el riesgo de que las intervenciones no logren los objetivos esperados o generen dependencia estructural.

Referencias

- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. -Curto Díaz, J., Conesa, J. (2011) *Introducción al Business Intelligence*. Editorial UOC

ANEXO