Un análisis exploratorio de la elección del tipo de gestión educativa de nivel secundario en Argentina con *machine learnig*[1]

Donadoni, Ana Clara y Harriague, María Marcela

Departamento de Economía - Centro de Investigación y Análisis en Economía (CIANECO)

Universidad Nacional de Río Cuarto

adonadoni@fce.unrc.edu.ar; mharriague@fce.unrc.edu.ar

Resumen

El trabajo aborda la elección entre educación de gestión pública y privada en en nivel secundario de Argentina, considerando la creciente participación del sector privado en la matrícula educativa. Se propone un análisis exploratorio mediante la técnica de "Classification and Regression Trees" (CART), para prever la elección de gestión educativa basándose en variables socioeconómicas y de gastos familiares. La metodología se aplica a datos de la Encuesta Nacional de Gastos de Hogares 2017-2018. La revisión de la literatura destaca estudios previos sobre segregación escolar y elección de gestión, principalmente basados en ingresos y nivel educativo del jefe de hogar. El trabajo argumenta la necesidad de incluir un conjunto más amplio de características. Se utilizan árboles de clasificación para el total del país y para el total urbano de cada región estadística, identificando variables clave como el clima educativo del hogar, el decil de ingreso per cápita provincial, condicional laboral del padre y la madre, cobertura de salud confort de la vivienda y junto a otras variables vinculadas al gasto educativo. Se concluye que estos modelos ofrecen una comprensión inicial, pero se reconocen limitaciones. El estudio sugiere vías para futuras investigaciones y plantea la importancia de analizar la no asistencia.

Palabras Clave

EDUCACIÓN SECUNDARIA - GESTIÓN EDUCATIVA PÚBLICA Y PRIVADA - MACHINE LEARNING

1. Introducción

El sistema educativo argentino tiene una larga historia de expansión, con logros significativos de cobertura de manera temprana respecto a la región (Llach, 2006) y con

[1] Se agradece al Dr. Marcos Herrera Gómez por los comentarios realizados y al Dr. Manuel Maffini (ISTE UNRC-CONICET) por la elaboración del mapa utilizado en este artículo. Se valoran también los comentarios vertidos por los miembros del CIANECO en seminarios de discusión internos.

una clara preeminencia del sector público en la prestación del servicio, especialmente en educación básica (niveles inicial, primario y secundario). A partir de la década de 1950, el subsidio público a la oferta se extiende al sector privado, especialmente a establecimientos religiosos, generando un sistema con dos áreas de gestión, establecimientos de gestión pública y de gestión privada, reconocidos por ley.

Si bien, históricamente, el sistema de gestión pública presenta una matrícula predominante en todos los niveles educativos, la participación del sector privado en la matrícula total de alumnos ha mostrado una tendencia creciente (Gasparini, Jaume, Serio y Vázquez, 2011), especialmente en los últimos años.

Asimismo, al ser Argentina un país federal, presentó durante muchos años concurrencia en la prestación de los servicios educativos entre el nivel nacional, los gobiernos provinciales y, en menor medida, los gobiernos municipales. Desde fines de la década del 60 y hasta 1978, se produce un proceso de descentralización desde el gobierno nacional hacia los gobiernos provinciales de la educación primaria (Ministerio de Educación, Ciencia y Tecnología, 2003) y, finalmente, en 1992 se transfieren a los gobiernos provinciales las escuelas normales, establecimientos de educación técnica e institutos terciarios que hasta ese momento seguían dependiendo del Ministerio de Cultura y Educación. De este modo, la responsabilidad de los sistemas educativos del nivel básico, tanto públicos como privados, pasa a los gobiernos provinciales (incluyendo a la Ciudad de Buenos Aires), así como también del subsidio a la gestión privada.

En 1993 se sanciona la Ley Federal de Educación (ley 25195), impulsada por el poder ejecutivo nacional. Esta ley, modifica la tradicional separación de los niveles educativos en inicial, primaria y media, proponiendo un sistema que establece 2 años de Educación Inicial, 3 ciclos de Educación General Básica (EGB) de 3 años cada uno y 3 años de Educación Polimodal. Asimismo, extiende la obligatoriedad en la sala de 5 años del nivel inicial y 2 años en el EGB3. Sin embargo, en el contexto del sistema descentralizado las provincias gradualmente optaron por distintas opciones, especialmente en el nuevo nivel entre el anterior primario y secundario ahora denominado EGB3. Algunas, como Córdoba, “secundarizaron” el séptimo grado de primaria y otras, como Buenos Aires, “primarizaron” el EGB3. Superado el crítico período de la salida de la convertibilidad, en el año 2005 se aprueba la Ley de Financiamiento Educativo y, en 2006, la Ley de Educación Nacional anuló la ley federal. A partir de ello, se volvió a la estructura de Inicial, Primaria y Media, pero

permitiendo una duración de 6 o 7 años para nivel primario y de 5 o 6 años para el secundario, estructurado en dos ciclos uno básico y otro orientado[2] (Palamidesi y Gorostiaga, 2022). De este modo, algunas provincias adoptaron la modalidad de primaria de siete años seguida de una secundaria de cinco años, mientras que otras implementaron el sistema de Educación General Básica (EGB) y Polimodal, donde los primeros seis años de EGB se consideran nivel primario y los últimos tres años de EGB se clasifican como nivel secundario. Esto hace que la estructura académica del nivel secundario difiere entre las distintas provincias, complejizando especialmente la interacción con los relevamientos de encuestas a hogares y la situación específica del sistema educativo en cada provincia.

Las tasas brutas de escolarización de la población en edad de acceder al nivel secundario, medidas a partir de los censos poblacionales, muestran un aumento muy relevante entre los censos del año 2010 y el del año 2022 (Cuadro Nº 1).

Cuadro N° 1. Población en viviendas particulares de entre 12 y 17 años que asiste a un establecimiento educativo, por grupo de edad. Argentina, total del país. Años 2001, 2010, 2022

Tramos de edad/ Año	2001	2010	2022
	% asistencia		
12-14 años	95,1	96,5	95,8
15-17 años	79,4	81,6	90,2

Fuente: Indec (2023). Censo Nacional de Población Hogares y Viviendas - Resultados Definitivos - Educación

Los datos de evolución de la matrícula del Ministerio de Educación entre 2010 y 2018 reflejan un aumento de la matrícula en los establecimientos de Gestión Privada de este nivel de más de un punto porcentual. Al agrupar los datos por las regiones estadísticas establecidas por INDEC se aprecian que los aumentos más relevantes se presentan en las regiones Metropolitana y Noreste y en menor medida en Cuyo y Pampeana (Cuadro Nº 2).

Cuando se analiza, a partir de la Encuesta Nacional de Gastos de Hogares 2017-2018 la asistencia a cada tipo de gestión se identifica por un lado una fuerte concentración de la asistencia en la gestión pública en en el quintil de menores ingresos. El 46% de los asistentes a este tipo de gestión pertenecen a este quintil en tanto que la asistencia al sector privado está menos concentrada. Sin embargo, se debe considerar

[2] La educación técnica y artística tienen un año adicional.

también que la participación de la matrícula privada es cercana al 25% en la encuesta y del 29,6% según los datos de matrícula[3]. Cuando se considera la participación de asistentes por tipo de gestión en cada quintil se destaca que aproximadamente el 90% de los asistentes del primer quintil corresponden a la gestión pública, en tanto que en el quintil de ingresos más altos el 74% de asisten a establecimientos de gestión privada.

Cuadro N° 2. Evolución de la participación porcentual de la oferta de gestión privada según región estadística.

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
Total Nacional	**28.4**	**28.4**	**28.8**	**29.0**	**29.0**	**29.4**	**29.5**	**29.7**	**29.6**
Metropolitana	36.7	36.9	37.6	37.8	38.2	39.0	38.9	39.2	38.9
Pampeana	30.8	31.2	31.3	31.5	31.5	31.6	31.6	31.6	31.5
Noroeste	21.6	20.9	21.6	21.6	21.4	21.3	21.6	21.7	21.7
Noreste	15.6	15.6	15.9	16.0	15.8	16.6	16.8	17.0	17.5
Cuyo	20.0	19.8	19.9	20.3	20.4	20.6	20.8	21.1	21.0
Patagonia	16.1	16.2	16.2	16.0	16.2	16.4	16.6	16.5	16.6

Fuente: Elaboración propia a partir del Anuario Estadístico Educativo - 2018. Ministerio de Educación

Cuadro N° 3. Asistentes al nivel secundario por quintiles de ingreso per cápita familiar provincial.

	Asistentes al secundario por quintiles					
	1	2	3	4	5	Total
Privado	139295	197887	203348	154921	177820	873271
Público	1220115	695822	442530	213167	61914	2633548
Total	1359410	893709	645878	368088	239734	3506819
	Participación por quintiles					
	1	2	3	4	5	Total
Privado	16.0	22.7	23.3	17.7	20.4	100
Público	46.3	26.4	16.8	8.1	2.4	100
Total	38.8	25.5	18.4	10.5	6.8	100
	Participación dentro de cada quintil					
	1	2	3	4	5	Total
Privado	10.2	22.1	31.5	42.1	74.2	24.9
Público	89.8	77.9	68.5	57.9	25.8	75.1
Total	100	100	100	100	100	100

Fuente: Elaboración propia a partir del ENGHo 2017-2018

[3] Cabe destacar que la encuesta tiene dominio urbano.

Esto no significa que sectores medios no opten por establecimiento públicos, en los deciles 3 y 4 casi el 69% y el 58% asisten al sector público respectivamente. Esto puede originarse tanto en aspecto de oferta educativa como en decisiones de los hogares. Al respecto, Narodowski y Gottau (2017), mediante entrevistas semi-estructuradas abiertas a un conjunto de 30 padres de clase media alta de la Ciudad de Buenos Aires analizan los motivos de la elección pública entre hogares de clase media (excluyendo a los establecimientos secundarios dependientes del sistema universitario caracterizado por su alta calidad) e identifican que a pesar de tener los recursos para optar por la privada en cierta medida rechazan el modelo de privatización de la educación y apoyan el sistema tradicional de monopolio estatal. Sin embargo, esto no necesariamente se traduce en una acción política colectiva efectiva para defender la educación pública. Si bien estas identidades se centran en lo estatal y colectivo, no necesariamente tienen articulación política. Por otro lado, a menudo recurren a sus redes personales para obtener ventajas (por ejemplo vacantes), lo que contradice el ideal de igualdad de oportunidades que dicen defender. Por lo que lo consideran más una afirmación individual que una lucha colectiva organizada.

Establecer los determinantes de la elección entre educación de gestión pública o privada es el primer paso para estudios más profundos orientados a la segregación escolar y los factores que inciden sobre ella.

El acceso a la educación de tipo pública o privada es un fenómeno multifactorial en el cual inciden aspectos socioeconómicos del hogar. Existe un enorme potencial para la utilización de técnicas de aprendizaje automático en el sector de la educación debido a la disponibilidad creciente de datos, con la posibilidad de ayudar en el diseño de políticas públicas tendientes a garantizar el derecho a la educación.

En este trabajo se propone un análisis exploratorio acerca de las principales variables que inciden sobre la predicción de la asistencia de los menores de entre 11 y 18 años a establecimientos educativos de nivel secundario de gestión pública o privada, con un enfoque de *machine learning*. Se utilizarán variables que representan determinadas características socioeconómicas y perfil de gastos del hogar al que pertenecen, en el marco de los datos provistos por la Encuesta Nacional de Gasto de Hogares de Argentina considerando, en esta primera etapa, el relevamiento 2017-2018. Adicionalmente, se estiman modelos por región, en el intento de explorar si existe

heterogeneidad en torno a las variables que resultan decisivas a nivel regional en contraposición al agregado nacional.

2. Estado actual de la literatura

Numerosos autores han estudiado los fenómenos de segregación escolar y los factores que inciden sobre la elección del tipo de gestión educativa en la Argentina.

Segnana y Adrogué (2021) analizan la incidencia de los factores socioeconómicos del hogar (nivel educativo del jefe de hogar, condicional en la posición del hogar en la distribución del ingreso) sobre la probabilidad de asistencia a la escuela pública o privada en primaria o secundaria, a partir de un modelo Probit utilizando datos trimestrales de la Encuesta Permanente de Hogares correspondientes al período 2016-2019. Los resultados que obtienen van en línea con lo que se observa en la literatura: la probabilidad de asistir a una institución de gestión pública disminuye a medida que aumenta el nivel educativo del jefe de hogar y la posición del hogar en la distribución del ingreso. Los resultados varían cuando se compara nivel primario y secundario, en tanto para el secundario hay un cambio en las decisiones de las familias de clase media en favor de la escuela pública.

Respecto a los determinantes de la asistencia escolar, cabe destacar el trabajo de Paz y Cid (2012), que analizan la asistencia escolar de los jóvenes de entre 15 y 18 años en las EPH puntual y contínua en el período 1997- 2009. Con un Modelo Logit Multinomial (MLMN) encuentran que la edad, el sexo y los salarios que perciben los jóvenes en el mercado de trabajo impactan en la decisión de asistencia de manera completa. Asimismo, el género y la condición socio-ocupacional de los jefes de hogar informales o desocupados impactan negativamente, y en mayor medida en los varones y la cantidad de niños en el hogar afecta en el mismo sentido. Por otro lado, el clima educativo del hogar se relaciona positivamente con la posibilidad de asistir. Estiman brechas de asistencia para hogares favorecidos, aquellos que vive en un hogar con jefe varón, ocupado en el sector formal, con clima educativo alto y pocos menores en el hogar y desfavorecidos a aquellos con jefa mujer, ocupada en el sector informal, con clima educativo bajo y muchos menores en el hogar. Simulan estas situaciones teniendo en cuenta el signo y la significancia estadística de los parámetros y encuentran que “Un joven que reside en un hogar con condiciones sociales favorables registra a los 17 años una tasa de asistencia del 93%, mientras que un joven de la misma edad, igual en todo al

anterior pero que reside en un hogar con condiciones sociales adversas alcanza una tasa de asistencia del 8%."

Utilizando también datos de la Encuesta Permanente de Hogares, Jaume (2011) estudia la segregación escolar por estrato socioeconómico para el período 1992-2010. Para ello, plantea un modelo de elección escolar, sobre la base de un modelo de respuesta multinomial, en función a las características del estudiante y el hogar: edad y género del estudiante, y logaritmo del ingreso per cápita familiar, tamaño de la familia y nivel educativo de los padres. Dicho modelo consiste en una maximización de utilidad considerando las siguientes respuestas mutuamente excluyentes: A) no asistir al colegio; B) asistir a un colegio público y C) asistir a un colegio privado. Encuentra que la segregación escolar se profundiza a lo largo del período considerado y que los resultados varían entre niveles educativos.

Por otra parte, existen también trabajos que utilizan datos de las encuestas de gastos de los hogares. Gasparini, Jaume, Serio y Vázquez (2011) estudian, a partir de datos de distintas encuestas a hogares realizadas en Argentina, el fenómeno de segregación escolar. En particular, con los datos de las ENGHo 1985/1986, 1996/97 y 2004/05 para la ciudad de Gran Buenos Aires, estiman la probabilidad condicionada de acceder a la escuela pública en función al nivel socioeconómico de las familias, medido a partir del percentil de ingreso per cápita familiar de cada estudiante. Los resultados sugieren una relación negativa entre el percentil de ingreso per cápita familiar y la probabilidad de asistencia a un establecimiento público para la ciudad de Buenos Aires, así como la profundización de la segregación escolar a lo largo del período considerado en el estudio.

Trabajos actuales como el de Sasserra (2022) abordan la segregación escolar desde una perspectiva territorial para la provincia de Chaco, considerando como fuentes de información el Relevamiento Anual y los resultados de las pruebas Aprender correspondientes al año 2019, que les permiten abordar a nivel departamental la concentración del Nivel Socio Económico, las disparidades de acceso y permanencia y los recursos materiales de los establecimiento. Utilizan técnicas de análisis socio-espacial y de análisis de cluster de K-medias e identifican circuitos diferenciados relacionados con las condiciones económicas y sociales de los departamentos y las condiciones escolares.

Recientemente, (CAF 2022), estudió la movilidad intergeneracional absoluta y relativa en América Latina con una diversidad de enfoques así como el estudio de la

movilidad intergeneracional del capital humano y de la igualdad de oportunidades. Por un lado, a partir de datos censales, identifican que en Argentina en los nacidos entre 1980 y 1990, se detectan una relación que ajusta bien a la línea recta para la movilidad relativa y no encuentran las no linealidades en el extremo superior de la distribución que se encuentran en países como Bolivia, Brasil, Costa Rica, Jamaica, Perú, República Dominicana, Trinidad y Tobago y Uruguay. A partir de una encuesta realizada en las principales capitales de la región (ECAF 2021), se observa que la educación de los padres está relacionada positivamente con las probabilidades de mejora en educación, salud y calidad de vivienda, aunque no de forma uniforme. Los hijos de padres que completaron la educación primaria tienen más probabilidades de lograr éxito en educación y calidad de vivienda en comparación con aquellos cuyos padres tienen educación primaria incompleta. Sin embargo, no se encuentran diferencias significativas en otras dimensiones del bienestar. Por otro lado, los hijos de padres con educación secundaria completa o superior tienen más probabilidades de mejorar en salud y calidad de vivienda que aquellos cuyos padres sólo completaron la primaria incompleta. La tenencia de vivienda u otras propiedades por parte de los padres también se asocia con mayores oportunidades de ascenso educativo, aunque no en otras áreas. En comparación con los hijos de padres desempleados a los 14 años, aquellos con padres empleados no muestran un aumento significativo en sus probabilidades de ascenso, aunque los hijos de padres con empleos de alta complejidad sí presentan mayores oportunidades en educación y salud.

El diseño de la encuesta permite evaluar además los efectos en generaciones no adyacentes. Los datos indican que los hijos de padres universitarios tienen un 29% más de probabilidad de completar la educación superior que aquellos cuyos padres no alcanzaron este nivel. Esta diferencia persiste en nietos y bisnietos, con brechas del 15% y 12%, respectivamente. Este hallazgo contrasta con las predicciones de un modelo de dos generaciones, donde la brecha habría desaparecido en la cuarta generación. Asimismo analizan la segregación escolar entre los subsectores público y privado utilizando la encuesta ECAF 2021 y encuentran que la probabilidad de asistir a una escuela privada aumenta 22 puntos porcentuales (controlando por características) para los hijos de padres que también asistieron a establecimientos privados, ubicándose Buenos Aires en niveles muy cercanos a este promedio (21 puntos porcentuales).

De la revisión de la literatura se encuentra que los trabajos existentes, en general, estudian la asistencia a escuelas públicas o privadas principalmente a partir de

la distribución de ingresos del hogar y, en particular, de una tipificación de clases sociales a partir de dicha distribución. Esto puede implicar cierta arbitrariedad en la selección de las variables que inciden sobre la asistencia a establecimientos educativos de un tipo de gestión u otra. Adicionalmente, limitar el número de variables consideradas para el análisis puede llevar a la omisión de características relevantes y, por otro lado, la tipificación de las predictoras puede influir sobre los resultados, reduciendo su variabilidad (Brunori et al, 2021).

Son numerosas las razones que fundamentan, entonces, la necesidad de incorporar un conjunto adicional de características representativas de los hogares y de los estudiantes, que permita enriquecer los análisis ya existentes circunscritos a la incidencia de la distribución del ingreso y el nivel educativo del/la jefe/a de hogar.

Sosa Escudero (2023) releva el uso de técnicas de machine learning para el estudio de la pobreza y el desarrollo, destacándose la importancia de estas técnicas para el la dimensionalidad de las caracterizaciones tanto en estudios de pobreza como en estudios sobre clases sociales específicas como el estudio de Edo, M.; Sosa Escudero, W. y, Svarc, M. (2022) que desarrollan un método moderno de selección de variables para reducir la dimensión del bienestar, y lo utilizan para detectar y medir la clase media argentina de manera que la clase media se halle en el centro de una distribución multivariada y clasificando todas las observaciones dentro de cuatro clases (pobres, clase media vulnerable, clase media segura y clase alta). Esto se asegura tanto mediante la generación de un índice de bienestar unidimensional como por la identificación de los diferentes grupos (pobres, clase media, clase alta) a través de cuantiles multivariados. Para reducir la dimensión del bienestar se utilizan técnicas de “machine learning” para identificación del subconjunto más pequeño de variables del espacio original de bienestar minimizando la pérdida respecto al subconjunto original.

Asimismo, en el contexto de imputación de datos faltantes de encuestas a hogares Indec (2020) realiza árboles de regresión con el fin de identificar las clases de imputación a partir de los nodos finales del subárbol óptimo, y posteriormente aplica hot deck utilizando “k-vecino más cercano”.

3. Aspectos metodológicos

En este trabajo se utilizó un enfoque de *machine learning* para predecir la asistencia de un menor a la escuela secundaria de gestión pública o privada, en función

a la información proporcionada por la Encuesta Nacional de Gastos de los Hogares 2017-2018.

3.1. Datos y Análisis Descriptivo

En esta sección se describen los datos, se observan relaciones y se presentan ejemplos vinculados a los mismos.

Los datos de la matrícula secundaria provistos por el Ministerio de Educación muestran que, para el año 2018, había 3.832.054 alumnos en el nivel secundario distribuidos entre las diferentes modalidades provinciales de implementación de la Ley Nacional de Educación. Esta es la población objetivo del presente estudio.

Cuadro N° 4. Matrícula total del Nivel Secundario y participación de la oferta de gestión privada según modalidad de implementación de la Ley Nacional de Educación. Argentina, 2018

5 años de nivel Secundario			**6 años de nivel Secundario**		
Jurisdicción	**Alumnos Nivel Medio**	**% Gestión privada**	**Jurisdicción**	**Alumnos Nivel Medio**	**% Gestión privada**
Ciudad de Buenos Aires	196,308	50.5	Buenos Aires	1,559,488	33.5
Chaco	101,995	17.0	Catamarca	43,634	18.6
Jujuy	67,607	16.1	Córdoba	329,729	40.0
La Rioja	31,735	14.4	Corrientes	106,242	17.5
Mendoza	130,644	23.0	Chubut	56,331	13.0
Misiones	100,341	22.8	Entre Ríos	127,908	24.6
Neuquén	50,231	13.8	Formosa	63,709	10.3
Río Negro	59,309	19.6	La Pampa	33,556	25.2
Salta	125,937	19.1	San Juan	75,559	21.8
Santa Cruz	28,733	15.5	San Luis	48,191	14.5
Santa Fe	245,267	31.8	Tierra del Fuego	155,158	27.0
Santiago del Estero	77,346	25.1	Tucumán	17,096	27.5
Subtotal	1,215,453	31.7	Subtotal	2,616,601	68.3
Total de jurisdicciones	3,832,054				
Gestión Privada	29.6%				

Fuente: Anuario Estadístico Educativo - 2018. Ministerio de Educación

En el Cuadro Nº 1 se destaca que, a nivel nacional, la proporción de matrícula del nivel secundario de establecimientos de gestión privada era del 29.6%, mientras que el 70.4% correspondía a escuelas de gestión pública. Se evidencian, no obstante, marcadas diferencias en la participación de la oferta privada entre las distintas jurisdicciones del país. Mientras que en la Ciudad de Buenos Aires el 50.5% de los alumnos asistía a escuelas privadas, en otras provincias como Formosa y Chubut este porcentaje era mucho menor, con solo el 10.3% y 13% respectivamente. Adicionalmente, se encuentran diferencias en la participación de la oferta privada según la duración del nivel secundario: en las jurisdicciones con nivel de 5 años de duración, la participación de la oferta privada era en promedio del 31.7%, mientras que en las de 6 años alcanzaba el 68.3%.

Esta información resulta relevante para analizar las diferencias en la preferencia por escuelas públicas o privadas a nivel nacional, considerando en el análisis la no uniformidad de la participación de la oferta privada entre provincias, así como para explorar los factores que podrían explicar estas disparidades.

3.2. Fuentes de datos

Para llevar adelante este estudio, se construyó una base de datos compuesta por los resultados de la Encuesta Nacional de Gastos de los Hogares (ENGHo) 2017-2018.

La ENGHo es una encuesta nacional, con cobertura en todas las provincias a partir del relevamiento de localidades de más de 2.000 o más habitantes, que tiene como objetivo proporcionar información sobre las condiciones de vida de la población en general y de distintos grupos de hogares en particular, desde el punto de vista de su participación en la distribución del ingreso y en el acceso a los bienes y servicios que constituyen el consumo de los hogares.

Se trabajó a partir de las bases usuario publicadas por el INDEC -base de hogares, de personas y de gastos-, y se seleccionaron las observaciones correspondientes a miembros del hogar de entre 11 y 18 años que asisten a un establecimiento educativo de nivel secundario, asignándoles las características socioeconómicas y perfil de gastos del hogar al que pertenecen, a partir de la clave de identificación de registro (id). Se identificó un conjunto de variables que caracterizan tanto al *estudiante* como a su grupo familiar. Muchas de estas variables son el resultado de la agregación de distintas variables disponibles en la encuesta.

Una fuente adicional de datos fue el Anuario Estadístico Educativo 2000, publicado por el Ministerio de Educación. Es importante destacar que se trata del único censo de datos educativos en el país y que provee una fuente de datos valiosa y de cobertura total para los establecimientos educativos, abarcando todas las modalidades, desde el nivel inicial hasta el superior no universitario.

3.3. Variables de estudio

La *variable dependiente* es el tipo de gestión del establecimiento educativo, una variable de tipo binaria que asume valor 1 si el estudiante asiste a un establecimiento de gestión pública y 0 si es de gestión privada (Tabla Nº 1).

Para seleccionar a los asistentes se procedió, en primer lugar, a hacer un filtrado por edad según la jurisdicción (entre 11 y 18 años, variable cp03 de la base personas de la ENGHo) y por asistencia a establecimiento público o privado (variable cp20 de la base personas de la ENGHo). En una segunda instancia, se seleccionó de acuerdo al nivel más alto que cursa (EGB, secundario y polimodal), eliminando aquellos casos que habían finalizado el nivel respectivo. Por último, para el nivel EGB se asignó a secundario a aquellos individuos cuyo último grado aprobado era 6, 7 u 8, consistente con las equivalencias entre niveles de acuerdo a la estructura del sistema educativo de cada provincia. De este modo, se definió una base de 6.142 observaciones del total muestreado por la ENGHo.

La *muestra* bajo estudio se restringió a los menores de entre 11 y 18 años que asisten a un establecimiento educativo de nivel secundario, con información disponible sobre el conjunto de variables socioeconómicas y perfil de gastos del hogar que se incluyen en el presente estudio. Se utilizó el factor pondera, proporcionado por la ENGHo, para expandir los 6.142 casos del muestreo en un total de 3.506.819 observaciones. De ese total, los asistentes a establecimientos de gestión pública son 2.633.548 y los de gestión privada 873.271.

Tabla 1. Variable dependiente.

	Variable	**Tipo**	**Definición**	**Concepto**
variable DEPENDIEN-TE	**asistencia**	Dummy	1 = público	Asiste a un establecimiento educativo
			0 = privado	

Las *predictoras* se seleccionaron entre variables que se obtuvieron y construyeron a partir de la ENGHo 2017-2018 y que, teóricamente, pueden ser factores

de incidencia. Esto es así ya que, si bien se debería incluir entre las predictoras a todas las variables que puedan afectar la elección del establecimiento educativo, al trabajar con los datos de la encuesta de gastos, sólo se dispone de un subconjunto de estos determinantes. Se identificó un conjunto de variables que caracterizan tanto al *estudiante* como al *hogar* al que pertenece. Muchas de estas variables son el resultado de la agregación de distintas variables disponibles en la encuesta.

La siguiente tabla resume las variables *explicativas* utilizadas, incluyendo el tipo de variable y la definición empleada en las estimaciones.

Tabla 2. Variables explicativas

	Variable	Tipo	Definición	Concepto
Variables asociadas al ESTUDIANTE	**trabaja**	Dummy	1 = si	¿La semana pasada trabajó al menos una hora? (cp26)
			2 = no	
	AUH	Dummy	1 = si	Percepción en los úlimos seis meses de Asignación Universal por Hijo (AUH) cp62
			2 = no	
	Progresar	Dummy	1 = si	Percepción en los úlimos seis meses de Progresar(cp64)
			2 = no	
	viajahab	Dummy	1 = Sí	Viaja al menos 1 vez por semana en trasporte público.
			2 = No	
Variables asociadas al HOGAR	**tipoviv**	Categórica	0 = rancho, casilla, pieza o local no construido para habitación	
			1 = departamento	
			2 = casa	
	confort	Categórica	0 = ninguno	No tiene cochera, jardín, pileta ni área deportiva.
			1 = bajo	Sólo tiene cochera o jardín.
			2 = confort medio	Tiene cochera y otra amenidad.
			3 = alto	Tiene 3 o más entre cochera, jardín, pileta y área deportiva.
	hacinamiento	Categórica	0 = sin hacinamiento	
			1 = moderado	
			2 = crítico	
	servicios	Categórica	0 = ninguno	Suma el acceso a servicios públicos de agua de red, cloacas y/o gas de red.
			1 = 1 servicio	
			2 = 2 servicios	
			3 = agua de red, cloacas y gas de red	
	tipohog	Categórica	1 = Unipersonal	Tipo de hogar
			2 = Nuclear sin hijos	
			3 =Nuclear con hijos	
			4 = Extendido	
	climaed	Categórica	1 Muy bajo	Clima educativo del hogar
			2 Bajo	
			3 Medio	

			4 Alto	
			5 Muy alto	
			99 Ns/nr	
	menhog	Numérica		Cantidad de miembros menores de 18 años en el hogar
	cobsalud	Categórica	0 = Sin cobertura	Indica el tipo de cobertura de salud, cuando hay más de una cobertura y está presente una prepaga se asigna a prepaga (directa o por obra social).
			1 = Obra social	
			2 = Prepaga	
	dipch.p	Categórica	valores de 1 a 10 (1 = menores ingresos y 10 = mayores ingresos)	Decil de ingreso per cápita del hogar provincial (histórico)
	condlab.madre	Categórica	0 = Directivo	Combina jerarquía ocupacional (jer ocup) con condición de actividad (estado), desagregando para el caso de asalariado con o sin aportes agrega los estados desocupados e inactivos de condición de actividad. Se considera asalariada sin aportes a los casos en donde el aporte quedó en blanco.
			1 = Cuenta propia	
			2 = Jefe	
			3 = Asalariado con aportes	
			4 = Asalariado sin aportes (incluye los casos en que la madre es asalariada pero dejó en blanco la pregunta sobre aportes)	
			5 = Desocupada	
			6 = Inactiva	
			7 = madre no está identificada o no vive en el hogar	
			99 = Indeterminado	
	condlab.padre	Categórica	0 = Directivo	
			1 = Cuenta propia	
			2 = Jefe	
			3 = Asalariado con aportes	
			4 = Asalariado sin aportes (incluye los casos en que el padre es asalariada pero dejó en blanco la pregunta sobre aportes)	
			5 = Desocupado	
			6 = Inactivo	
			7 = padre no está identificado o no vive en el hogar	
			99 = Indeterminado	
variables asociadas al GASTO del hogar	**gEnoFpm**	Numérica		gasto del hogar en educación no formal como % del gasto de consumo total bruto, por menor en edad en edad escolar
	gMyUpm	Numérica		gasto en mochila y uniforme como % del gasto de consumo total bruto, por menor en edad en edad escolar
	gLpm	Numérica		gasto del hogar en libros como % del gasto de consumo total bruto, por menor en edad en edad escolar
	gMatEdpm	Numérica		gasto del hogar en materiales educativos como % del gasto de consumo total bruto, por menor en edad en edad escolar

variables asociadas a la OFERTA educativa	**pMP2016.S**	Numérica		Porcentaje de matrícula educativa privada en el año 2016 para el nivel secundario

Respecto al *estudiante*, se incluyeron variables como el nivel educativo que cursa (realizando los ajustes necesarios debido a la estructura de los sistemas educativos de cada provincia), la cobertura de salud, la condición de percepción de transferencias monetarias (Progresar y AUH), si trabaja o no, y si utiliza al menos una vez por semana el transporte público.

Para el caso de la variable "cobertura de salud", fue necesario considerar las distintas posibilidades de múltiples coberturas y agrega tres alternativas considerando: "Sin cobertura" (0) a aquellas personas que disponen sólo de servicios de emergencia, Profe o planes públicos de salud; "Obra Social" (1) a las personas que declaran tener la cobertura de una obra social o Pami e incluye aquellos que combinan estas alternativas con sistema de emergencias; "Prepaga" (2) a aquellas personas que aportan a una prepaga o disponen de prepaga por aporte a la obra social y también incluye las distintas combinaciones siempre que esté presente la prepaga.

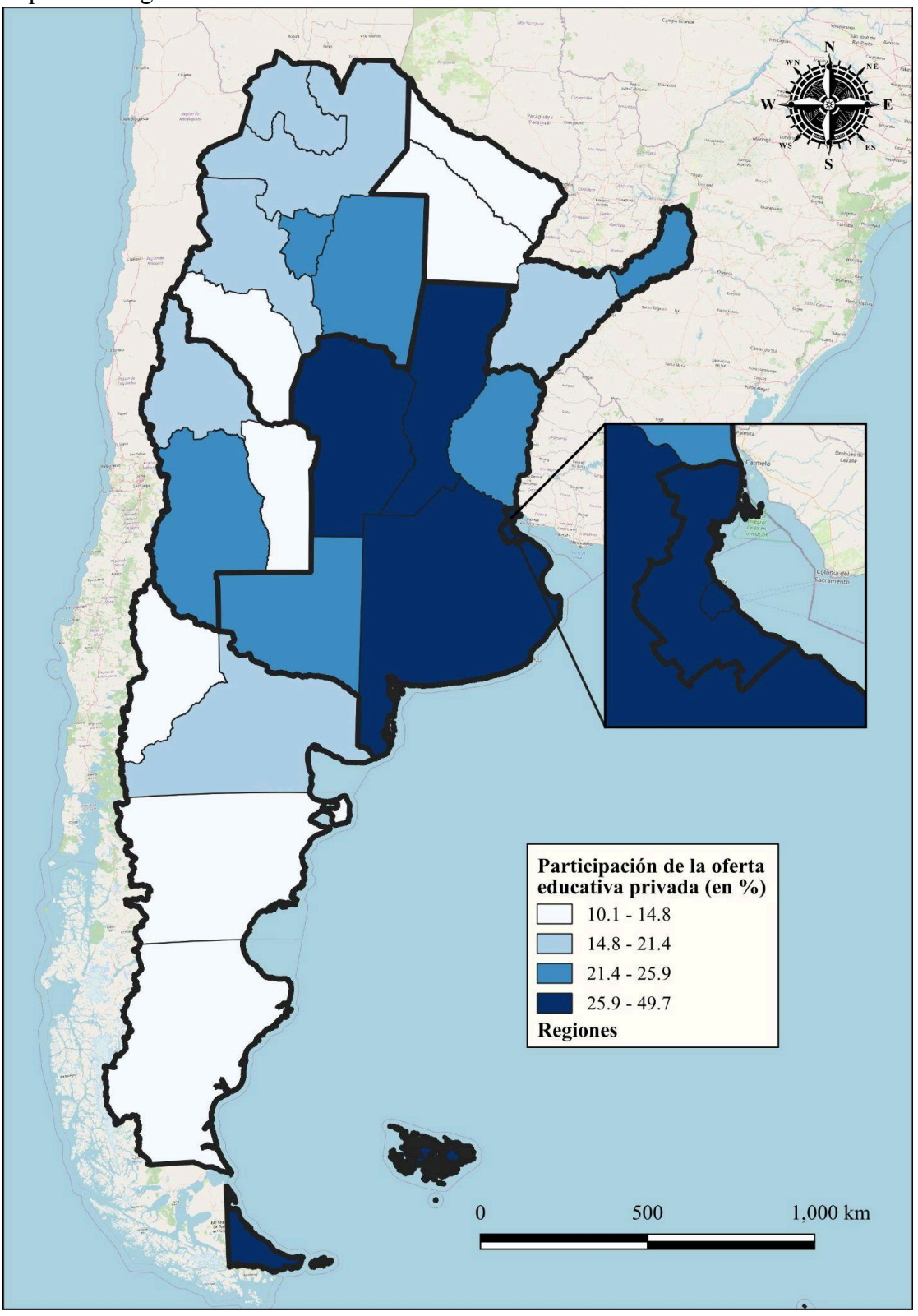
Se construyeron, adicionalmente, variables ocupacionales tanto para la madre como para el padre. La variable condición laboral de la madre (o padre) identifica primero al progenitor correspondiente de niños de entre 3 y 18 años y posteriormente integra las variables condición de actividad y jerarquía ocupacional, desagregando a los asalariados en registrados o no registrados.

Entre las variables que caracterizan al *grupo familiar* y reflejan las condiciones del hogar, se identificaron el decil de ingreso per cápita provincial, las características de la vivienda (hacinamiento, tipo de vivienda, acceso a servicios públicos y confort), el clima educativo del hogar, la condición laboral de la madre y el padre, el gasto de consumo per cápita y el tipo de hogar. Asimismo, se construyeron variables a partir de la base de gastos de la encuesta, para reflejar el gasto en educación no formal, libros, mochila y uniforme, como proporción del gasto total del hogar. Por otro lado, para eliminar el sesgo en los valores nominales debido a la presencia de inflación, las variables de gastos se tratan como proporciones respecto al total y, para considerar las diferencias por niveles de ingreso entre provincias, se considera el decil de ingreso o gastos per cápita provincial del hogar.

Respecto de las variables que caracterizan al hogar, se construyó una variable que tipifica los hogares en función de la disponibilidad de cochera, jardín, pileta y

espacios de uso deportivo, diferenciando entre quienes no tienen ninguna de estas características, quienes solo cuentan con algunas y quienes acceden a todas.

Figura Nº 1. Mapa de participación de la oferta educativa privada. Nivel secundario. República Argentina. Año 2016.



Fuente: Elaboración en base al Anuario Estadístico Educativo - 2018. Ministerio de Educación

Paralelamente se construyó una variable que refleja la importancia de la oferta educativa privada secundaria a nivel provincial en los distintos niveles educativos a partir del Relevamiento Anual del Sistema Educativo que realiza la Dirección de Información y Evaluación Educativa del Ministerio de Educación. Se calculó la proporción de matrícula privada en el año 2016 para cada provincia, excepto en el caso de la provincia de Buenos Aires que se calculó por separado para el Área Metropolitana de Buenos Aires (región metropolitana) y para el resto de la provincia (región pampeana). En la Figura Nº 1 se presenta la distribución de la participación porcentual de la oferta privada por provincias, en escala de color graduada por cuartiles.

3.4. Algoritmo CART

Se realizó un análisis exploratorio utilizando técnicas de *machine learning* para indagar acerca de las principales variables que inciden sobre la predicción de la asistencia de los menores a establecimientos educativos de nivel secundario de gestión pública o privada, a partir de variables que representan características socioeconómicas y perfil de gastos del hogar al que pertenecen. Para ello, se recurrió al algoritmo "Classification and Regression Trees" (CART), introducido originalmente por Morgan y Sonquist (1963) y popularizado por Breiman, Friedman, Olshen y Stone (1984), que se utiliza para construir árboles de decisión que permiten tanto la clasificación como la regresión. En el contexto de clasificación, el objetivo principal de CART es dividir un conjunto de datos en subgrupos homogéneos basados en las características de los datos, facilitando así la predicción de la clase de nuevas observaciones (Breiman et al., 1986).

Los árboles de clasificación particionan de manera iterativa y sin superposiciones el espacio de predicción, seleccionando un predictor que minimiza la tasa de error de clasificación en la muestra de entrenamiento en cada iteración. La técnica apunta a encontrar divisiones que maximicen la homogeneidad dentro de los nodos y minimicen la homogeneidad entre nodos, con el objetivo de evitar el sobreajuste del modelo y lograr una adecuada predicción por fuera de la muestra.

Una característica particular de esta técnica de clasificación es que permite trabajar tanto con variables categóricas como contínuas, es invariante a transformaciones monótonas así como incluir combinaciones lineales de las variables (Feldman y Gross, 2005).

3.4.1. Construcción y poda del Árbol

La construcción de un árbol de clasificación CART se lleva a cabo mediante un proceso recursivo que incluye la división de los nodos y la poda de árboles. Una de las características distintivas del algoritmo CART es su enfoque en la partición del espacio de atributos, dividiendo el espacio multidimensional de las características en regiones (subespacios) que corresponden a diferentes clases.

El espacio de atributos se define como el conjunto de todas las combinaciones posibles de características en el conjunto de datos. Cada punto en este espacio representa una instancia (observación) del conjunto de datos, con sus características como dimensiones.

En la etapa de división, CART realiza la partición del espacio de atributos de manera recursiva utilizando un enfoque basado en la minimización de la impureza de los nodos para decidir la mejor forma de dividir los datos en cada paso. La impureza se mide comúnmente utilizando el índice de Gini o la entropía.

El *Índice de Gini* se define como:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

donde p_i es la proporción de ejemplos de la clase i en el conjunto de datos D y C es el número total de clases. El objetivo es seleccionar la división que minimice el índice de Gini en los nodos resultantes (Breiman et al., 1986).

La *Entropía* se define como:

$$Entropy(D) = - \sum_{i=1}^{C} p_i \log_2(p_i)$$

De manera similar, se busca minimizar la entropía tras cada división (Shannon, 1948), indicando un mejor ajuste del modelo a los datos.

La división se realiza a través de un proceso iterativo que evalúa todas las posibles divisiones para cada característica X_j y cada punto de corte s, buscando el punto de corte óptimo para cada atributo. Para una variable continua, la división se realiza en un punto específico s que separa las instancias menores y mayores que s. Para variables categóricas, se evalúan todas las combinaciones posibles de categorías.

El mejor punto de corte s es aquel que minimiza la impureza ponderada de los nodos resultantes:

$$Impureza\ después\ de\ la\ división = \frac{N_{left}}{N} * Gini\left(D_{left}\right) + \frac{N_{right}}{N} * Gini\left(D_{right}\right)$$

donde: N es es el número total de instancias en el nodo original, y N_{left} y N_{right} son el número de instancias en los nodos izquierdo y derecho después de la división.

La división del nodo genera dos subespacios:

- Un subespacio que contiene todas las instancias que cumplen la condición (por ejemplo, $X_j \leq s$), y
- otro subespacio que incluye las instancias que no la cumplen (por ejemplo, $X_j > s$),

Este proceso se repite para cada subespacio creado, dividiendo aún más hasta que se cumplan los criterios de parada, como la pureza del nodo o el número mínimo de instancias en un nodo.

Una vez construido el árbol, se procede a la poda para evitar el sobreajuste. CART utiliza un enfoque de poda cost-complexity, que implica la selección de un parámetro de complejidad α que penaliza la complejidad del árbol:

$$R_\alpha(T) = R(T) + \alpha|T|$$

donde $R(T)$ es el error de clasificación del árbol T y $|T|$ es el número de nodos terminales. La poda busca minimizar esta función para encontrar un árbol que balancee precisión y simplicidad (Breiman et al., 1986).

Una vez que el árbol ha sido podado, la estimación de la clase para una nueva instancia se realiza siguiendo el camino desde la raíz hasta un nodo terminal, basado en las decisiones tomadas en cada nodo. La clase predicha corresponde a la clase más frecuente en el nodo terminal alcanzado (Loh, 2011).

El algoritmo CART de clasificación es una poderosa herramienta para la toma de decisiones, que permite la interpretación clara de los resultados a través de su estructura jerárquica. Su capacidad para manejar datos categóricos y continuos, junto con su enfoque en la minimización de la impureza, lo convierte en una opción popular en diversas aplicaciones de clasificación. Esta metodología proporciona un marco sólido para la construcción y optimización de modelos de clasificación, garantizando que los resultados sean tanto precisos como interpretables.

3.4.2. Implementación del Algoritmo CART con Validación Cruzada

Para llevar a cabo la clasificación utilizando el algoritmo CART, se empleó el paquete *rpart* en R, ampliamente utilizado para la construcción de árboles de decisión (Therneau y Atkinson, 2019). El proceso se estructuró en cuatro etapas clave,

comenzando con la validación cruzada, la selección del mejor modelo, la poda del árbol y, finalmente, la evaluación del modelo en un conjunto de testeo.

1. Validación Cruzada de 10-Folds.

Se implementó una validación cruzada de 10 pliegues (10-fold cross-validation) de manera manual para evaluar la robustez del modelo. Este método implica dividir el conjunto de datos en 10 subconjuntos, donde en cada iteración se utiliza uno de los subconjuntos como conjunto de prueba y los restantes como conjunto de entrenamiento. Este proceso se repite 10 veces, permitiendo que cada subconjunto sea utilizado como conjunto de prueba una vez. La precisión (accuracy) se calculó en cada iteración para medir el rendimiento del modelo.

La precisión se define como:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

donde TP, TN, FP y FN representan los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente (James et al., 2013). Se utilizó una semilla de aleatoriedad (123), para garantizar la replicabilidad de los resultados.

2. Selección del Mejor Modelo

Tras completar la validación cruzada, se seleccionó el modelo con la mayor precisión promedio. Esta elección se basa en la suposición de que un modelo con mejor rendimiento en datos de validación cruzada es más capaz de generalizar a datos no vistos (Kohavi, 1995; Efron, 1983).

3. Evaluación del Parámetro de Complejidad (cp)

Una vez identificado el mejor modelo, se procedió a evaluar el parámetro de complejidad (cp) para la poda del árbol. El cp controla la complejidad del árbol y ayuda a prevenir el sobreajuste. Se generó una tabla de complejidad utilizando la función `printcp()` del paquete `rpart`, que proporciona información sobre el error de clasificación y el tamaño del árbol para diferentes valores de cp.

La poda se llevó a cabo seleccionando un valor óptimo de cp que minimizara el error de clasificación en el conjunto de validación. Esto se realizó mediante la selección del cp que ofrecía el mejor equilibrio entre precisión y simplicidad del modelo (Breiman et al., 1986; Hastie et al., 2009).

4. Predicción con el Modelo Podado

Finalmente, se utilizó el modelo podado para realizar predicciones sobre el conjunto de prueba correspondiente al pliegue del mejor modelo. La predicción se llevó a cabo utilizando la función `predict()` del paquete `rpart`, que aplica el modelo a los nuevos datos para estimar las clases de las instancias en el conjunto de prueba. Allí se estimó la matriz de confusión y la tasa de error de clasificación.

Este enfoque estructurado permitió la implementación efectiva del algoritmo CART, maximizando la precisión del modelo y garantizando su capacidad de generalización. La combinación de validación cruzada, poda y evaluación robusta del modelo proporciona una metodología sólida para el análisis de clasificación en conjuntos de datos complejos.

4. Resultados

Los resultados de la estimación de los árboles de clasificación se presentan a continuación, en primer lugar para el total nacional y, posteriormente, se realiza un análisis de cada una de las regiones dada la disparidad de la oferta educativa. Los árboles regionales muestran distintas configuraciones respecto al modelo para el total del país, evidenciando particularidades regionales.

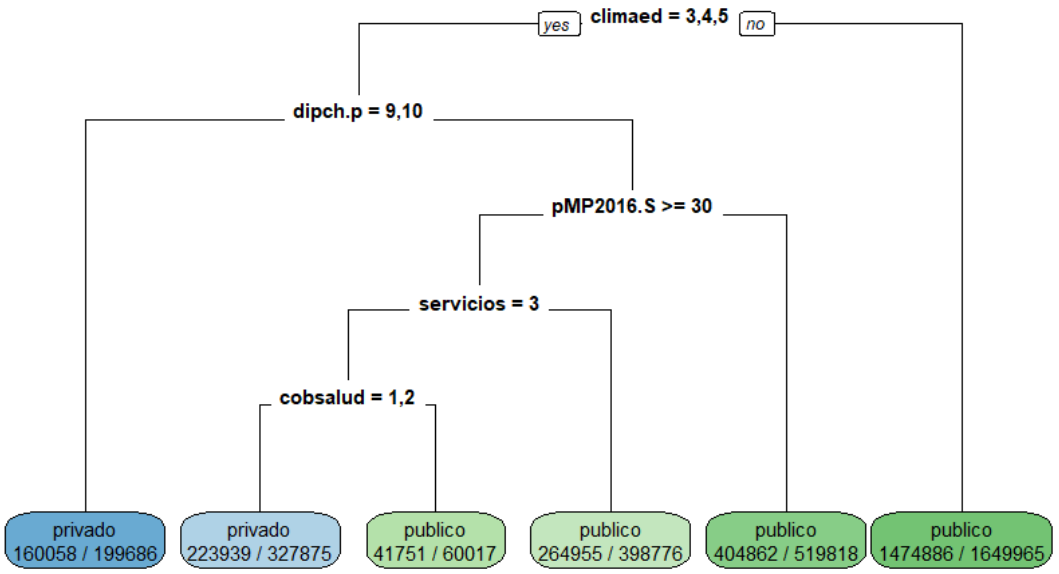
4.1. Nivel nacional

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 3.506.819 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 0.8025, para un subconjunto de entrenamiento de 3.156.137 observaciones. El mejor modelo estimado divide el espacio de predictoras en nueve nodos internos y diez nodos terminales, exhibiendo gran profundidad y complejidad. Esto puede significar que el árbol es muy específico a los datos de entrenamiento, lo que podría indicar sobreajuste (ver en anexo la representación gráfica del modelo). En consecuencia, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Se seleccionó un valor de cp = 0.02, ya que presentaba un buen balance entre bajo error de predicción y complejidad del modelo (en

el anexo se presenta el gráfico del parámetro de complejidad). El árbol resultante se muestra en la Figura 2.

Figura Nº 2. Árbol de clasificación para el nivel secundario - total Nacional. Aglomerados urbanos de Argentina 2017-2018.



El árbol correspondiente al nivel secundario en el total del país, particiona el espacio de atributos en cinco nodos internos y seis nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable clima educativo del hogar y predice que los hogares con clima educativo bajo tienden a ser clasificados como asistente a un establecimiento público. Los asistentes pertenecientes a hogares con clima educativo medio y alto se demarcan, en un segundo nodo, por el decil de ingreso per cápita del hogar provincial. Los asistentes pertenecientes a los deciles 9 y 10 de la distribución se clasifican al sector privado. Para los restantes deciles de la distribución, surge un nodo adicional que representa la oferta privada: cuando hay oferta educativa privada menor al 30%, los asistentes se clasifican al sector público. En cambio, si la oferta privada es mayor o igual al 30%, el árbol toma mayor profundidad. Para este caso, los hogares que disponen 1 o 2 servicios se clasifican en el sector público. En cambio, para los hogares que disponen de todos los servicios básicos (agua de red, cloaca y gas) surge un último nodo vinculado a la variable cobertura de salud. Si el hogar cuenta con cobertura de obra social o prepaga, asisten a establecimientos privados. Por otro lado, aquellos que aún disponiendo de 3 servicios en el hogar, no cuentan con cobertura de salud se clasifican en el sector público.

El modelo final, luego de la poda, se ajusta a los datos razonablemente bien, clasificando erróneamente alrededor del 13,35 % de las observaciones en el conjunto de testeo. Su precisión es de 0.8665 en el subconjunto de validación, lo que significa que el 86.65% de las predicciones realizadas por el modelo son correctas. Considerando que la proporción de asistentes a escuelas públicas para el total nacional es del 75 % en la muestra, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

4.2. Región metropolitana

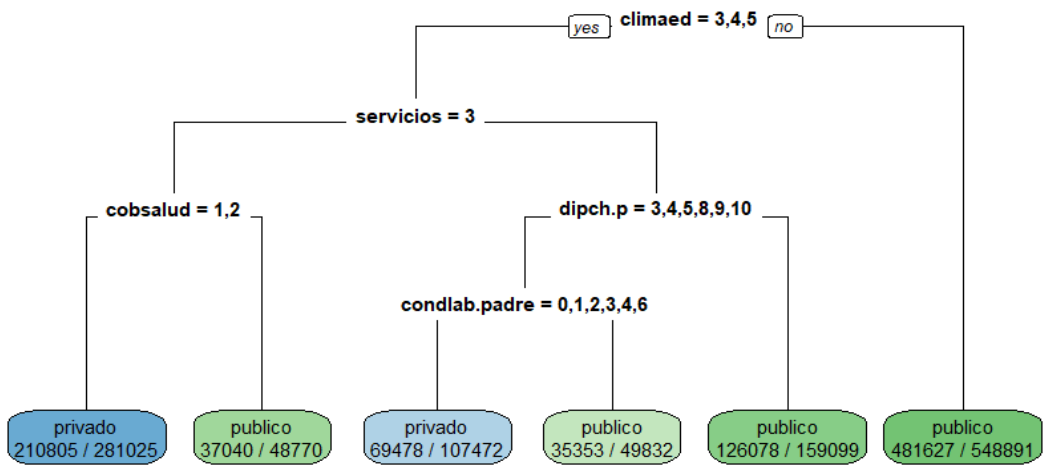
Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 1.300.396 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 77.87%, para un subconjunto de entrenamiento de 1.195.089 observaciones. El mejor modelo estimado divide el espacio de predictoras en once nodos internos y doce nodos terminales, exhibiendo gran profundidad y complejidad. Esto puede significar que el árbol es muy específico a los datos de entrenamiento, lo que podría indicar sobreajuste, al igual que en el modelo anterior (ver en anexo la representación gráfica del modelo). En consecuencia, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Se seleccionó un valor de cp = 0.02, ya que ofrecía un buen compromiso entre bajo error de predicción y complejidad del modelo. El árbol resultante se muestra en la Figura 3.

El árbol correspondiente al nivel secundario en la región Metropolitana, particiona el espacio de atributos en cinco nodos internos y seis nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable clima educativo del hogar y predice que los hogares con clima educativo bajo son clasificados como asistentes a establecimientos públicos. Los asistentes pertenecientes a hogares con clima educativo medio y alto se demarcan, en un segundo nodo, por la disponibilidad de servicios en el hogar. Para aquellos que disponen de los 3 servicios (agua de red, cloacas y gas de red) surge un nodo vinculado a la variable cobertura de salud. Si el hogar cuenta con cobertura de obra social o prepaga, se clasifican al sector privado. En

cambio si el hogar no cuenta con cobertura asisten a establecimientos públicos. Para aquellos que disponen sólo de 1 o 2 servicios se ramifica en un nodo adicional con el decil de ingreso per cápita del hogar. Aquellos pertenecientes a los deciles 1, 2, 6 y 7 optan por el sector público. En cambio, para aquellos cuyo hogar pertenece a los deciles 3, 4, 5, 8, 9, y 10 surge un nodo adicional relacionado con la variable condición laboral del padre. Aquellos cuyo padre tiene una condición laboral 5 "desocupado" o 7 "No identificado/no vive en el hogar" se clasifican en el sector público, todas las restantes condiciones ocupacionales se clasifican en el sector privado.

Figura Nº 3. Árbol de clasificación para el nivel secundario - región Metropolitana. Aglomerados urbanos de Argentina 2017-2018.



Cabe destacar que se trata de la región con mayor participación de oferta privada y la que registra el mayor crecimiento de la participación de la oferta privada entre 2010 y 2018. Por otro lado, la variable servicios podría estar más a la subregión conurbano.

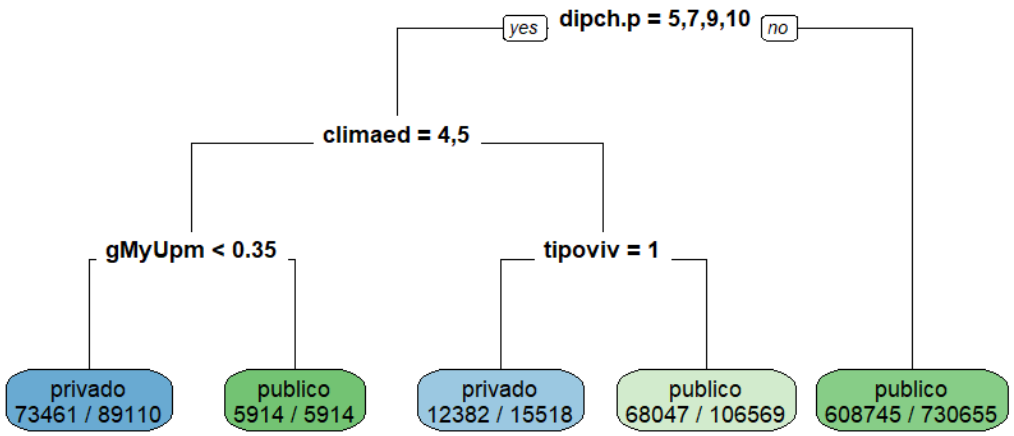
El modelo final, luego de la poda, se ajusta bien a los datos, clasificando erróneamente alrededor del 10,37 % de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 89,63%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 67,47% en la muestra, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

4.3 Región pampeana

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 1.053.073 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 74.62%, para un subconjunto de entrenamiento de 947.766 observaciones. El mejor modelo estimado divide el espacio de predictoras en trece nodos internos y catorce nodos terminales, exhibiendo gran profundidad y complejidad, incluso mayor a los modelos anteriores. Esto puede significar que el árbol es muy específico a los datos de entrenamiento, lo que podría indicar sobreajuste (ver en anexo la representación gráfica del modelo). En consecuencia, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Se seleccionó un valor de cp = 0.02, ya que ofrecía un buen compromiso entre bajo error de predicción y complejidad del modelo. El árbol resultante se muestra en la Figura 4.

Figura Nº 4. Árbol de clasificación para el nivel secundario - región Pampeana. Aglomerados urbanos de Argentina 2017-2018.



El árbol correspondiente al nivel secundario de la región Pampeana, particiona el espacio de atributos en cuatro nodos internos y cinco nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable decil de ingreso per cápita del

hogar provincial, los asistentes pertenecientes a hogares de los deciles de menores ingresos (1 a 4) y a los deciles 6 y 8 se clasifican en establecimientos de gestión pública. Para aquellos asistentes que pertenecen a hogares de los deciles 5, 7, 9 y 10 el árbol presenta un nodo con la variable clima educativo del hogar que muestra dos ramificaciones posteriores. Los asistentes pertenecientes a hogares con clima educativo medio y alto se demarcan, en un nuevo nodo delimitado por la variable que considera en gasto en mochila y uniformes como porcentaje del gasto bruto total por menor, aquellos con niveles de gasto inferior a 0.35 asisten a establecimientos privados como nodo final, en cambio aquellos con gasto superior a dicho porcentaje asisten a establecimientos del sector público. Finalmente, los asistentes que pertenecen a hogares con clima educativo bajo o medio (1, 2, 3) se dividen nuevamente por el tipo de vivienda, aquellos que viven en rancho, casilla, pieza o local no construido para habitación o en casas se clasifican a establecimientos públicos en cambio aquellos que viven en departamentos se clasifican en establecimientos de gestión privada.

El modelo final, luego de la poda, se ajusta razonablemente bien a los datos, clasificando erróneamente alrededor del 21,53 % de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 78,47%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 74,23% en la muestra, este porcentaje implica que el modelo aporta a la predicción bastante poco más que si se dejara la clasificación al azar. Estos resultados podrían indicar la necesidad de profundizar en la especificación del modelo para dicha región, en vistas a mejorar el desempeño por fuera de la muestra. En el anexo se presenta la matriz de confusión para el conjunto de validación.

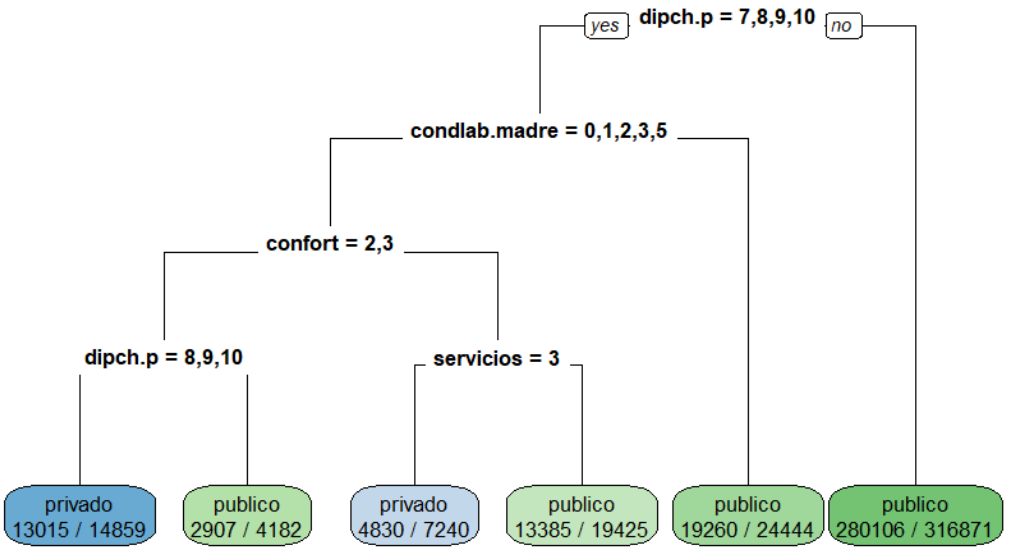
4.4. Región Noroeste

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 430.023 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 83.33%, para un subconjunto de entrenamiento de 387.021 observaciones. El mejor modelo estimado divide el espacio de predictoras en diez nodos internos y once nodos terminales, exhibiendo gran profundidad y complejidad, incluso mayor a los modelos anteriores. Esto puede significar que el árbol es muy específico a los datos de entrenamiento, lo que

podría indicar sobreajuste (ver en anexo la representación gráfica del modelo). En consecuencia, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Para seleccionar el cp se seleccionó aquel valor, dentro del rango entre 0.01 y 0.03 para el cual el modelo tiene el mejor rendimiento en términos de error, que ofrecía un mejor balance entre complejidad y rendimiento (cp = 0.03). El árbol resultante se muestra en la Figura 5.

Figura Nº 5. Árbol de clasificación para el nivel secundario - región Noroeste. Aglomerados urbanos de Argentina 2017-2018.



El árbol correspondiente al nivel secundario de la región Noroeste, particiona el espacio de atributos en cuatro nodos internos y seis nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable decil de ingreso per cápita del hogar provincial, los asistentes pertenecientes a hogares de los deciles de ingresos 1, 2, 3, 4, 5 y 6 se clasifican en establecimientos de gestión pública. Para aquellos asistentes que pertenecen a hogares de los deciles 7, 8, 9 y 10 el árbol presenta un nodo con la variable condición laboral de la madre. Aquellos asistentes que cuya madre tiene la condición laboral 4 asalariados sin aportes, 6 inactiva, 7 no está identificada o no vive en el hogar y 99 indeterminado se clasifican en establecimientos del sector público como nodo terminal. Para aquellos asistentes cuya madre tiene la condición laboral de 0 directiva, 1 cuenta propia, 2 jefe, 3 asalariada con aportes y 5 desocupada se clasifican según un

nodo adicional vinculado al confort de la vivienda. Para niveles de confort 3 alto (tiene tres o más vinculado a cochera, jardín, pileta y área deportiva) o 2 medio (cochera y otra amenidad) surge una nueva división en el árbol relacionada nuevamente a la variable ingreso per cápita del hogar provincial donde aquellos con deciles más altos son clasificados en establecimientos de gestión privada y aquellos pertenecientes al decil 7 en el establecimientos de gestión pública. Para los asistentes que viven en hogares de confort 0 nulo (no dispone de cochera, jardín, pileta o área deportiva) o 1 bajo (sólo dispone de cochera o jardín) se establece un nuevo nodo vinculado a los servicios de la vivienda, aquellos que cuentan con todos los servicios se clasifican en establecimientos de gestión privada en tanto que aquellos que no disponen de servicios o les falta 1 o 2 de los servicios de agua, cloacas y/o gas de red se clasifican en establecimientos públicos.

El modelo final, luego de la poda, se ajusta muy bien a los datos, clasificando erróneamente alrededor del 5.2 % de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 94.8%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 83.74% en la muestra, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

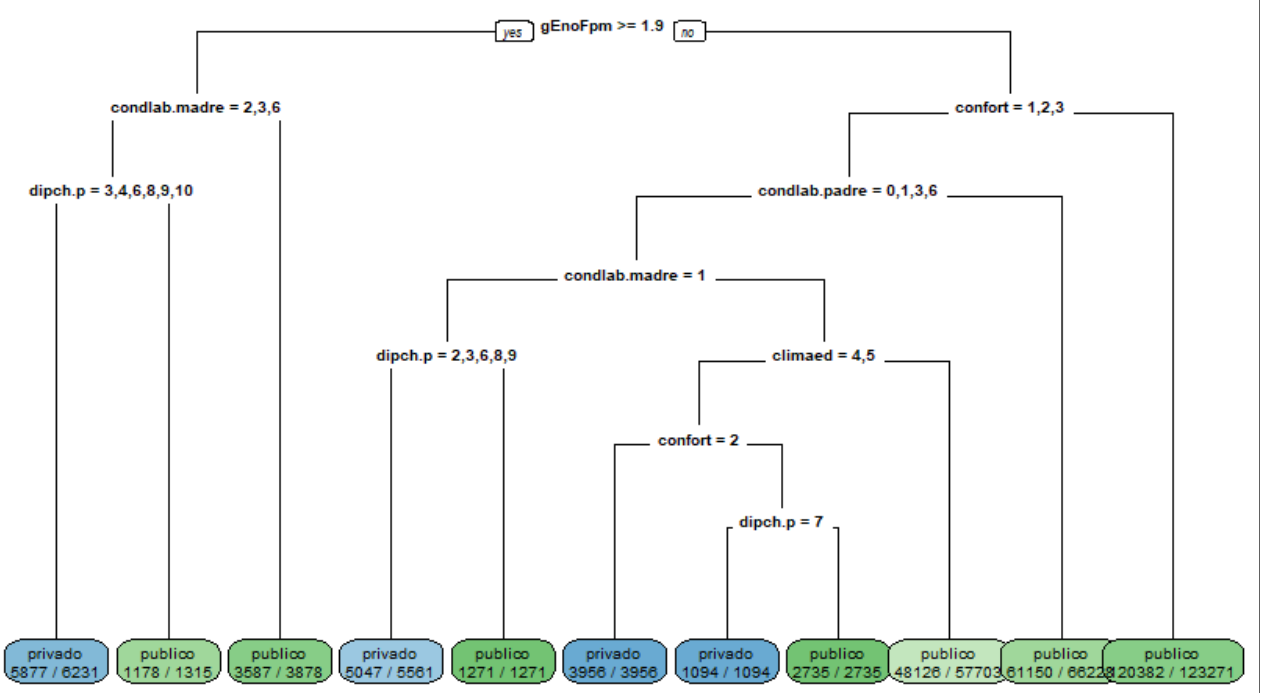
4.5. Región Noreste

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 303.603 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 85.59%, para un subconjunto de entrenamiento de 273.243 observaciones. El mejor modelo estimado divide el espacio de predictoras en veintidós nodos internos y veintitrés nodos terminales, exhibiendo una excesiva profundidad y complejidad (ver en anexo la representación gráfica del modelo). Para evitar el sobreajuste, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Para seleccionar el cp se optó por un valor, dentro del rango donde se estabiliza el error, para el cual el modelo

presenta mayor simplicidad, manteniendo un equilibrio entre complejidad y rendimiento (cp = 0.03). El árbol resultante se muestra en la Figura 6.

Figura Nº 6. Árbol de clasificación para el nivel secundario - región Noroeste. Aglomerados urbanos de Argentina 2017-2018.



Es una de las regiones donde ha aumentado la participación de la oferta privada en los últimos años, 1.99 puntos porcentuales entre 2010 y 2018. Es una de las regiones en donde la clasificación muestra mayor complejidad. Se presentan diez nodos y once nodos terminales u hojas.

El primer nodo está determinado por la variable gasto en educación no formal, aquellos con un gasto en educación no formal superior a 1,9% del gasto bruto del hogar se clasifican posteriormente por condición laboral de la madre, si esta tiene la condición laboral 2 jefa, 3 asalariada con aportes o 6 inactiva se dividen nuevamente por el decil de ingreso per cápita del hogar provincial, dentro de este grupo. aquellos de hogares con decil 3, 4, 6, 8, 9 y 10 se clasifican en establecimientos de gestión privada en tanto que los que pertenecen a los restantes deciles (1, 2, 5, 7) asisten a establecimientos públicos. Si la condición laboral de la madre es (0 directiva,1 jefa, 4 asalariada sin aportes, 5 desocupada, 7 no identificada o no vive en el hogar o 99 indeterminada) se clasifican en establecimientos públicos.

Para aquellos hogares con gasto en educación no formal menor a 1,9% del gasto bruto del hogar se clasifican posteriormente por la variable confort, si no dispone de valor 0, no tiene cochera ni jardín ni pileta ni área deportiva se clasifican en establecimientos públicos. En cambio, aquellos que tienen alguna o todas estas amenidades se clasifican nuevamente por la condición laboral del padre, si este es 2 jefe, 4 asalariado sin aportes, 5 desocupado, 7 no está identificado o no vive en el hogar o indeterminado asisten como nodo final a establecimientos públicos. En cambio si la condición laboral del padre es alguna de las restantes (0, 1, 3, 6) el nodo se divide nuevamente por la condición laboral de la madre. Si la madre es 1 cuenta propia el nodo se divide posteriormente por el decil de ingreso per cápita del hogar provincial, los deciles 2, 3, 6, 8 y 9 asiste como nodo final al sector privado en cambio si el hogar pertenece a los deciles 1, 4, 5, 7 o 10 asiste a establecimientos del sector público. Si la madre tiene una condición laboral distinta de cuenta propia el nodo se divide por el clima educativo del hogar, en donde los asistentes pertenecientes a hogares de clima educativo bajo a medio (1, 2 y 3) tienen como nodo final el sector público. En cambio si el clima educativo es alto o muy alto (4 y 5) el nodo se divide por la variable confort, aquellos con confort medio (2) como nodo final al sector privado; si el hogar tiene niveles de confort 1 o 3 se vuelve a dividir por el decil de ingreso per cápita familiar, si tiene ingreso per cápita del hogar del decil 7 asiste al sector privado caso contrario al sector público.

El modelo final, luego de la poda, se ajusta muy bien a los datos, clasificando erróneamente alrededor del 3.15% de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 96.85%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 88.17% en la muestra, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

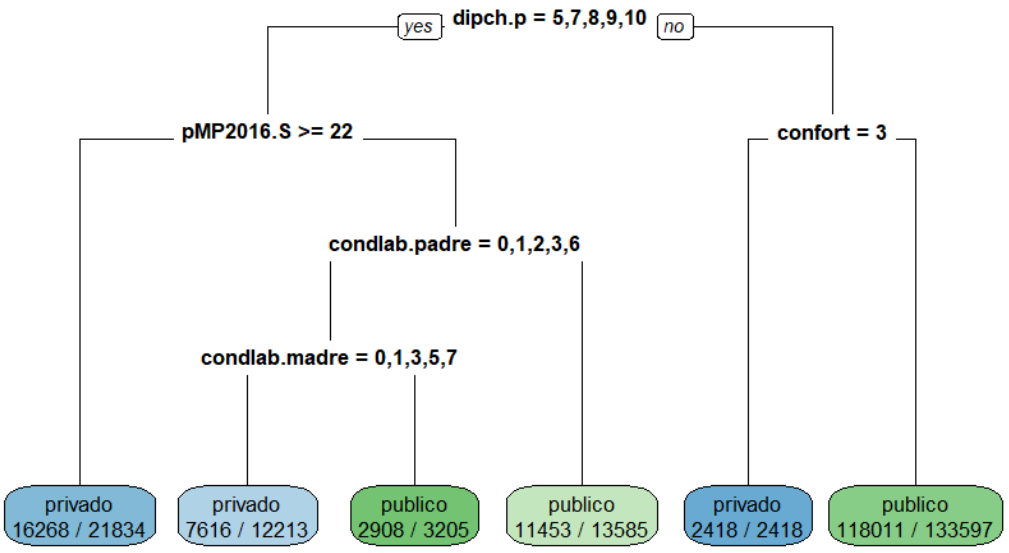
4.6. Región Cuyo

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 207.614 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 77.43%, para un

subconjunto de entrenamiento de 186.852 observaciones. El mejor modelo estimado divide el espacio de predictoras en veinte nodos internos y veintiún nodos terminales, exhibiendo una excesiva profundidad y complejidad (ver en anexo la representación gráfica del modelo). Para evitar el sobreajuste, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Para seleccionar el cp se optó por un valor, dentro del rango donde se estabiliza el error, para el cual el modelo presenta mayor simplicidad, manteniendo un equilibrio entre complejidad y rendimiento (cp = 0.03). El árbol resultante se muestra en la Figura 7.

Figura Nº 7. Árbol de clasificación para el nivel secundario - región Cuyo. Aglomerados urbanos de Argentina 2017-2018.



El árbol correspondiente al nivel secundario de la región Cuyo, particiona el espacio de atributos en cinco nodos internos y seis nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable decil de ingreso per cápita del hogar provincial, los asistentes pertenecientes a hogares de los deciles de menores ingresos (1 a 4) y a los deciles 6 se divide posteriormente por la variable confort en donde aquellos con niveles bajos o medio de confort se clasifican en el sector público en tanto que los que tienen confort alto se clasifican en el sector privado.

Para aquellos asistentes que pertenecen a hogares de los deciles 5, 7, 8, 9 y 10 el árbol presenta un nodo con la variable oferta educativa provincial es mayor o igual a 22

(Mendoza) el nodo final los asigna al sector privado. En cambio, si la oferta provincial es menor a dicho valor el nodo se divide primero por la condición laboral del padre, si el padre es asalariado sin aportes, desocupado, no está identificado o no vive en el hogar o indeterminado el nodo finaliza clasificando en establecimientos de gestión pública. Por otro lado, para las condiciones laborales del padre 0 directivo, 1 cuenta propia, 2 jefe, 3 asalariado con aportes o 6 inactivo el nodo se divide por condición ocupacional de la madre. Para las condiciones laborales del madre 0 directivo, 1 cuenta propia, 3 asalariada con aportes, 5 desocupada o 7 no identificado o no vive en el hogar se clasifica en el sector privado en tanto que las restantes condiciones laborales de la madre se clasifican en el sector público.

El modelo final, luego de la poda, se ajusta razonablemente bien a los datos, clasificando erróneamente alrededor del 22.62% de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 77.38%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 90.11% en la muestra, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

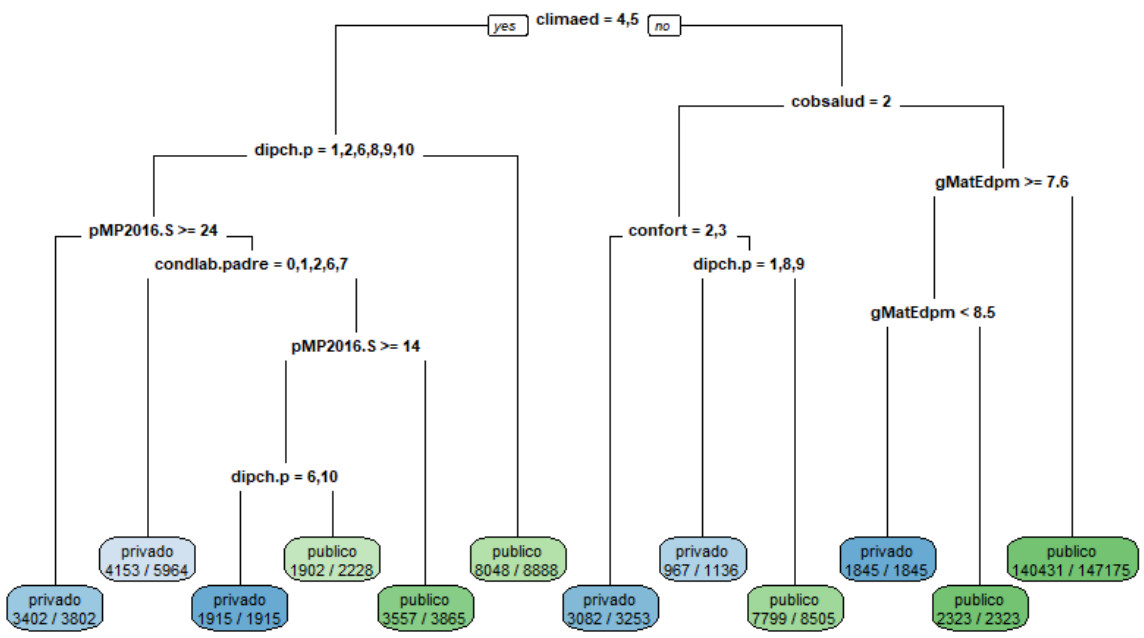
4.7. Región Patagonia

Se estimó un modelo de árbol de clasificación para predecir la asistencia a establecimientos de gestión pública o privada al nivel secundario, sobre una base de 212.110 individuos, a partir de un conjunto de 20 predictoras.

A partir de la estimación del modelo por validación cruzada de 10 pliegues (*10-fold cross validation*) se obtuvo una precisión media de 83.77%, para un subconjunto de entrenamiento de 190.899 observaciones. El mejor modelo estimado divide el espacio de predictoras en veinte nodos internos y veintiún nodos terminales, exhibiendo una excesiva profundidad y complejidad (ver en anexo la representación gráfica del modelo). Para evitar el sobreajuste, se hicieron pruebas adicionales para evaluar el tamaño óptimo del árbol; en particular, se estimó el parámetro de costo de complejidad (*cost complexity prune*), y se optó por podar el árbol para mejorar el ajuste del modelo por fuera de la muestra. Para seleccionar el cp se optó por un valor, dentro del rango donde se estabiliza el error, para el cual el modelo presenta mayor simplicidad, manteniendo un equilibrio entre complejidad y rendimiento (cp = 0.03) (en

el anexo se presentan la tabla y el gráfico del parámetro de complejidad). El árbol resultante se muestra en la Figura 8.

Figura Nº 8. Árbol de clasificación para el nivel secundario - región Patagonia. Aglomerados urbanos de Argentina 2017-2018.



El árbol correspondiente al nivel secundario en la región Patagonia, particiona el espacio de atributos en once nodos internos y doce nodos terminales u hojas. El primer nodo de clasificación corresponde a la variable clima educativo del hogar y predice que los hogares con clima educativo bajo o medio (1, 2 o 3) son clasificado posteriormente por la variable cobertura de salud, si la cobertura de salud es 0 o 1 (no tiene u obra social) el nodo se divide por el gasto del hogar en materiales educativos como proporción del gasto bruto, si esta proporción es menor a 7,6 se clasifican como asistentes a establecimientos públicos como nodo final. En cambio si es mayor a 7,6 el nodo se vuelve a dividir por la misma variable, y en este caso en los que es menor a 8,5 se clasifican asistiendo al sector privado pero cuando es mayor se clasifican en el sector público. Los asistentes que pertenecen a hogares de bajo o medio clima educativo pero que tienen como cobertura de salud a la prepaga se dividen posteriormente por la variable confort, aquellos con niveles de confort nulo o bajo se dividen posteriormente por el decil de ingreso per cápita provincial, aquellos que pertenecen a los deciles 1, 8 y

9 se clasifican al privado y en los restantes deciles al público. Aquellos que tienen confort medio y alto (2 y 3) se clasifican en establecimientos del sector privado.

Los asistentes que integran hogares de clima educativo alto o muy alto (4 y 5) se dividen posteriormente por la variable decil de ingreso per cápita del hogar provincial, si pertenecen a hogares con deciles 3, 4, 5 y 7 se clasifican como nodo final en establecimientos públicos. Aquellos que pertenecen a hogares con deciles 1, 2, 6, 8, 9 y 10 se vuelven a particionar por la oferta educativa y si la oferta educativa es mayor a 24% se clasifican como nodo final en establecimientos del sector privado. Si la oferta educativa es menor a 24% el nodo se divide por condición laboral del padre, si esta variable asume los valores 0, 1, 2, 6 o 7 se asignan al sector privado. Si asume los valores de 3, 4, 5 o 99 vuelve a dividirse por la oferta privada provincial, y si esta es inferior al 14% se clasifican en establecimientos públicos en tanto que en las jurisdicciones de oferta provincial superior al 14% (pero menor al 24%) se vuelve a dividir considerando el ingreso per cápita familiar provincial, si corresponden a hogares con decil 6 o 10 se clasifican a establecimientos de gestión privada y si no pertenecen a estos deciles a establecimientos de gestión pública. En esta región la participación de la oferta privada provincial es muy heterogénea, por eso aparece como relevante en las particiones del árbol.

El modelo final, luego de la poda, se ajusta razonablemente bien a los datos, clasificando erróneamente alrededor del 3.5% de las observaciones en el conjunto de testeo. Su precisión, es decir las predicciones correctamente realizadas por el modelo, es del 96.5%. Considerando que la proporción de asistentes a escuelas públicas para la región es del 90.11%, este porcentaje implica que el modelo aporta a la predicción más que si se dejara la clasificación al azar. En el anexo se presenta la matriz de confusión para el conjunto de validación.

5. Conclusiones

Este trabajo constituye una primera aproximación al estudio de los determinantes de la elección entre establecimientos de gestión pública o privada para el nivel secundario con datos de la ENGHo 2017-2018, a partir de un modelo de Árbol de Clasificación.

De la estimación de árboles para el total del país y para región, surgen resultados interesantes para analizar el proceso de decisión del tipo de gestión educativo que

complementan los trabajos existentes en el área, a la vez que introducen nuevas variables que hasta ahora no habían sido consideradas como relevantes para el análisis.

Cabe destacar que, la variable clima educativo del hogar es suministrada por INDEC en la base de datos e integra el nivel educativo de los mayores de 18 años que forman parte del hogar. Dicha variable, en sus niveles más bajos, aparece en el agregado y en algunas regiones asociada en los nodos de decisión del sector público así como las variables de ingreso per cápita provincial. Las nuevas variables asociadas a la asistencia a establecimientos públicos o privados que cobran relevancia en este trabajo pueden sintetizarse en gastos del hogar en bienes y servicios vinculados a la educación, la condición laboral tanto del padre como de la madre, la cobertura de salud y variables vinculadas al confort de la vivienda. La estrategia utilizada, a partir del empleo de las técnicas de *machine learning* descriptas previamente, permitió identificar los factores que mejor determinan la probabilidad de elección de un establecimiento público o privado para el nivel secundario en el total de país y por grupo de provincias, optimizando la predicción de la asistencia. Esto posibilita complementar los resultados obtenidos mediante metodologías tradicionales en estudios anteriores sobre la temática. Adicionalmente, el estudio muestra diferencias significativas en la participación de la oferta de gestión privada entre las distintas regiones estadísticas de Argentina, lo que sugiere que los factores que influyen en la elección del tipo de gestión educativa pueden variar según la región.

Este primer trabajo, al aplicar estas técnicas a una encuesta tan relevante como la ENGHo, abre un camino importante de trabajo. En primer lugar, es necesario evaluar más en profundidad estas relaciones a partir de la aplicación de técnicas de Random Forest y posteriormente profundizar el análisis el análisis con las técnicas de regresión logística y regularización. Por otro lado, debe considerarse la posibilidad de realizar estudios comparativos, especialmente con ENGHo 2004-2005. Asimismo, puede ser relevante complementar este análisis con el estudio de los determinantes de la no asistencia, ya que este nivel si bien ha aumentado la cobertura no han logrado la universalización.

Es necesario, además, considerar que el tipo de fuente de información que se utiliza impone una limitación a los resultados obtenidos ya que no es posible captar un conjunto de variables que pueden contribuir a explicar el fenómeno bajo estudio. Entre ellas, las diferencias que presentan ambos tipos de gestión educativa en términos de

localización precisa e infraestructura de los establecimientos educativos, planta docente y rendimiento académico medio (Gasparini et. al., 2011).

6. Bibliografía.

Breiman, L., Friedman, J., Olshen, R., y Stone, C. (1984). *Classification and Regression Trees.* Wadsworth and Brooks/Cole, Monterey.

Brunori, P., y Neidhöfer, G. (2021). The evolution of inequality of opportunity in Germany: A machine learning approach. *Review of Income and Wealth, 67*(4), 900-927.

CAF (2022). *Desigualdades heredadas: El rol de las habilidades, el empleo y la riqueza en las oportunidades de las nuevas generaciones*. Bogotá, Colombia – octubre.

Chan, F. y Mátyás, L. ed. (2022) *Econometrics with Machine Learning.* Advanced Studies in Theoretical and Applied Econometrics Vol. 53. Springer.

Edo, M.; Sosa Escudero, W. y, Svarc, M. (2022). A multidimensional approach to measuring the middle class. *The Journal of Economic Inequality* (2021) 19:139–162

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements to the cross-validation approach. *Journal of the American Statistical Association, 78*(382), 316-331.

Feldman, D. y Gross, S. (2005) Mortgage Default: Classification Trees Analysis. The Journal of Real Estate Finance and Economics, 30:4, 369–396, 2005.

Gasparini, L., Jaume, D., Serio, M., y Vázquez, E. (2011). La segregación entre escuelas públicas y privadas en Argentina. Reconstruyendo la evidencia. *Desarrollo Económico: Revista de Ciencias Sociales*, 189-219.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Herrera, M., Araoz, M. F., de la Fuente, G., María Lucrecia, D. J., Granado, J., Michel, A., y Paz, C. (2005). Técnicas para datos multinivel: aplicación a los determinantes del rendimiento educativo. Munich Personal RePEc Archive. Anales AAEP 2005. https://bd.aaep.org.ar/anales/works/works2005/herrera_otros.pdf

INDEC (2020). *Encuesta Nacional de Gastos de los Hogares 2017-2018: Manual de uso de la base de datos usuario*. 1a ed. Ciudad Autónoma de Buenos Aires.

INDEC (2020). *Encuesta Nacional de Gastos de los Hogares 2017-2018 Metodología de imputación Nota técnica n° 5* – Mayo de 2020

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2021). *An introduction to statistical learning.* Vol. 112, 2da edición. Springer.

Jaume, D. (2011). *Evolución de la segregación escolar en Argentina*. [Tesis de Maestría, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/18221?show=full

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1137-1145).

Llach, J.J (2006). *El desafío de la equidad educativa. Diagnóstico y propuestas.*.Ed. Granica

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14-23.

Ministerio de Educación de la Nación (2022). *Anuario Estadístico Educativo 2020* - 1a ed - Ciudad Autónoma de Buenos Aires.

Ministerio de Educación, Ciencia y Tecnología (2003). *Descentralización y estrategia en educación: caso Argentina*.

Lustig, N. (2017) The Impact of Taxes and Social Spending on Inequality and Poverty in Latin America: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica,

Ecuador, El Salvador, Guatemala, Honduras, Mexico, Peru, and Uruguay. *CGD Working Paper* 427, Washington, DC.

Morgan, J. y Sonquist, J. (1963). *Problems in the Analysis of Survey Data, and a Proposal*. Journal of the American Statistical Association 58(302), p. 415 – 34.

Narodowski, M., y Gottau, V. (2017) Clases medias y escuela pública. La elección escolar como resistencia. *Perfiles educativos*, vol. XXXIX, núm. 157, Universidad Nacional Autónoma de México.

Palamidessi, M. y Gorostiaga, J. (2022). Las políticas nacionales para la educación básica. En Gamallo, G (comp), "De Alfonsín a Macri. Democracia y Política Social en Argentina (1983-2019)" (p. 139-184). Eudeba.

Paz, J; Cid, J. C. (2012) Determinantes de la asistencia escolar de los jóvenes en la Argentina. *REDIE. Revista Electrónica de Investigación Educativa*, vol. 14, núm. 1, pp. 136-152.

Pincay-Ponce, J.; De Giusti, A.; Sánchez-Andrade, D. y Figueroa-Suárez, J. (2024). CatBoost: Aprendizaje automático de conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, (38),* (31-39). doi:10.24215/18509959.38.e3.

Segnana, J., y Adrogué, C. (2021). Factores socioeconómicos del hogar en la elección del tipo de gestión del establecimiento educativo en Argentina. *Páginas de Educación*, *14*(1), 112-126.

Sasserra, J (2022) Desigualdades educativas y socioespaciales en la educación secundaria en Argentina: aportes para la identificación de circuitos educativos en los departamentos de la provincia de Chaco. *Itinerarios educativos, núm. 16, e0025, 2022. Universidad Nacional del Litoral, Argentina*

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379-423.

Sosa Escudero, W. (2023) Big Data y Algoritmos para la Medición de la Pobreza y el Desarrollo, Cedlas, Documento de trabajo 319.

Sosa-Escudero, W., Anauati M. V.y Brau, W. Poverty, Inequality and Development Studies with Machine Learning en Chan, F. y Mátyás, L. ed. (2022) *Econometrics with Machine Learning.* Capítulo 9.

Therneau, T. M., & Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

7. Anexo

7.1. Modelo para el total nacional

Figura Nº 1. Árbol de clasificación para el nivel secundario - total del país. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).

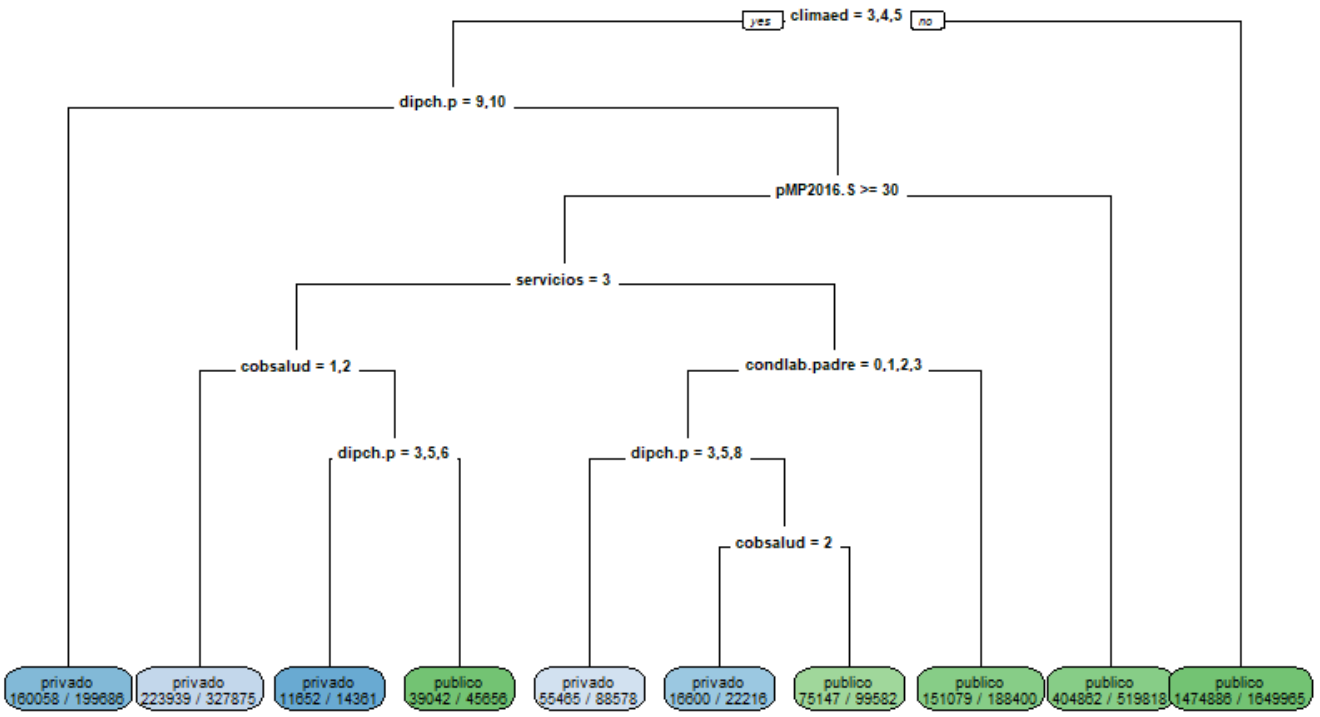
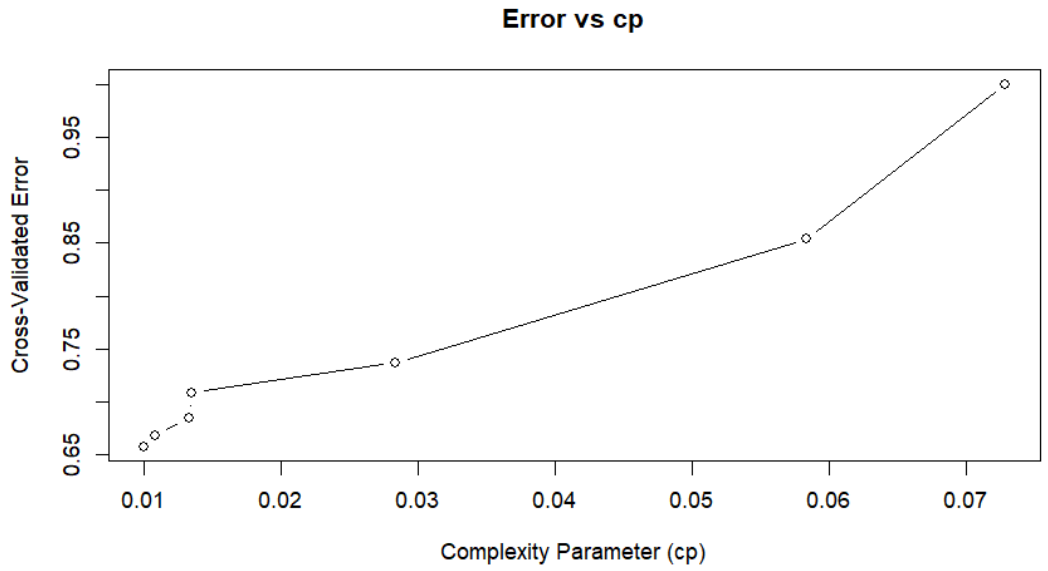
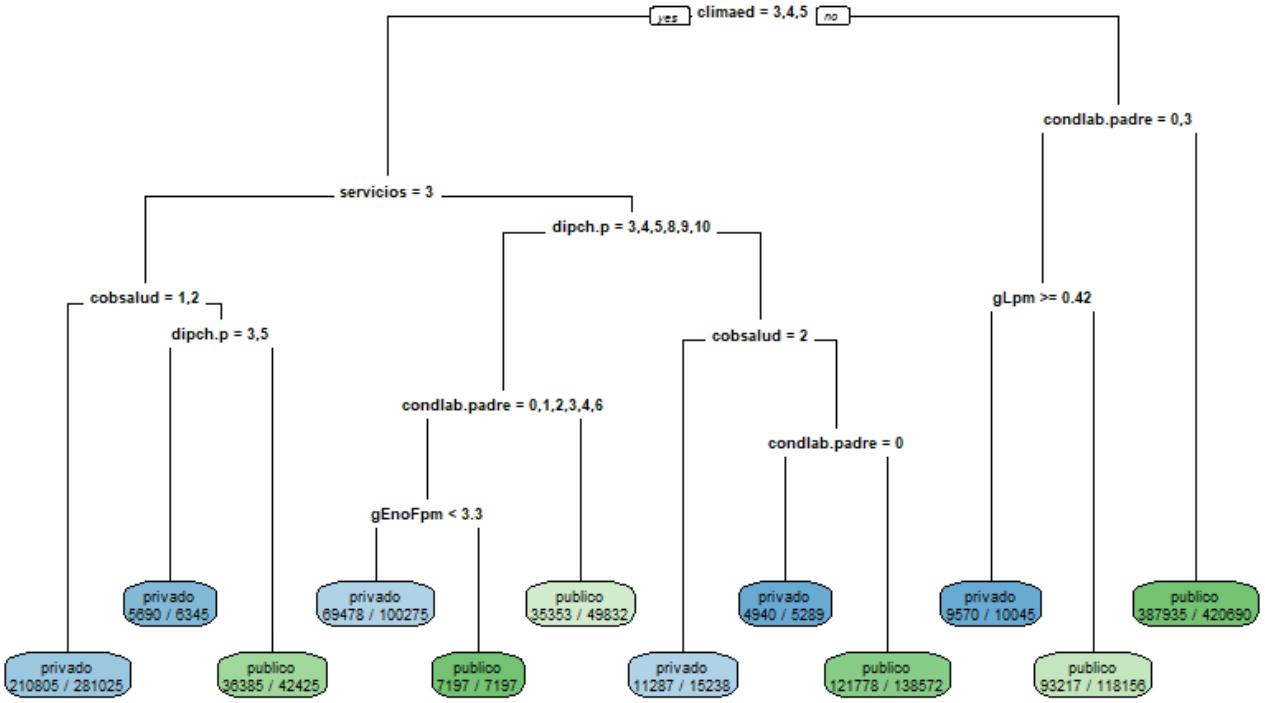


Figura Nº 2. Parámetro de complejidad (cost complexity prune) - total del país



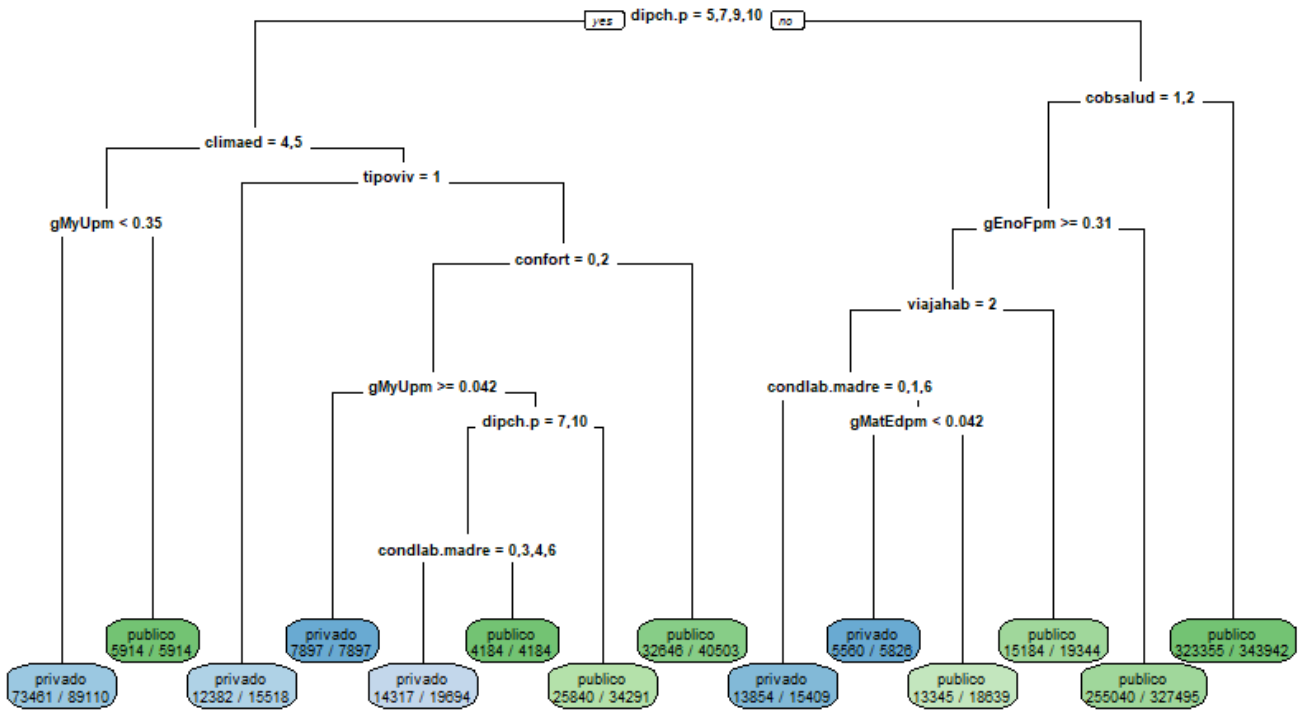
7.2. Modelo para la región Metropolitana

Figura Nº 3. Árbol de clasificación para el nivel secundario - región Metropolitana. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).



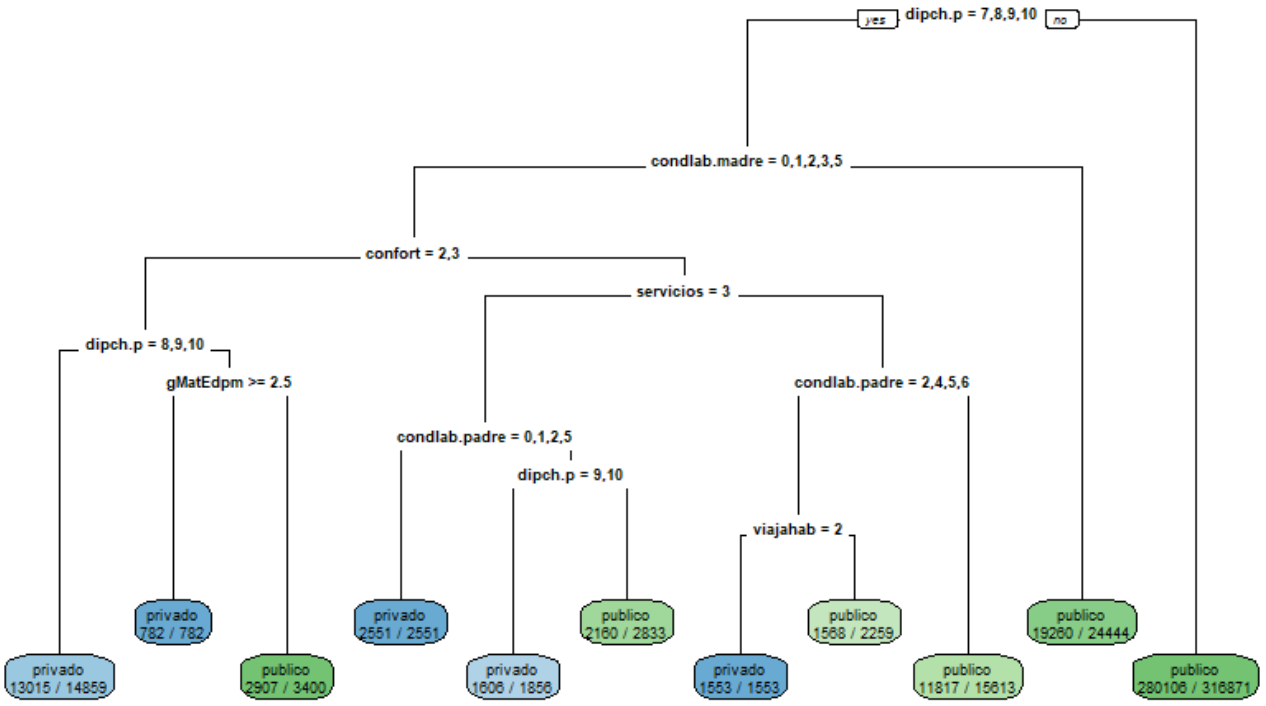
7.3. Modelo para la región Pampeana

Figura Nº 4. Árbol de clasificación para el nivel secundario - región Pampeana. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).



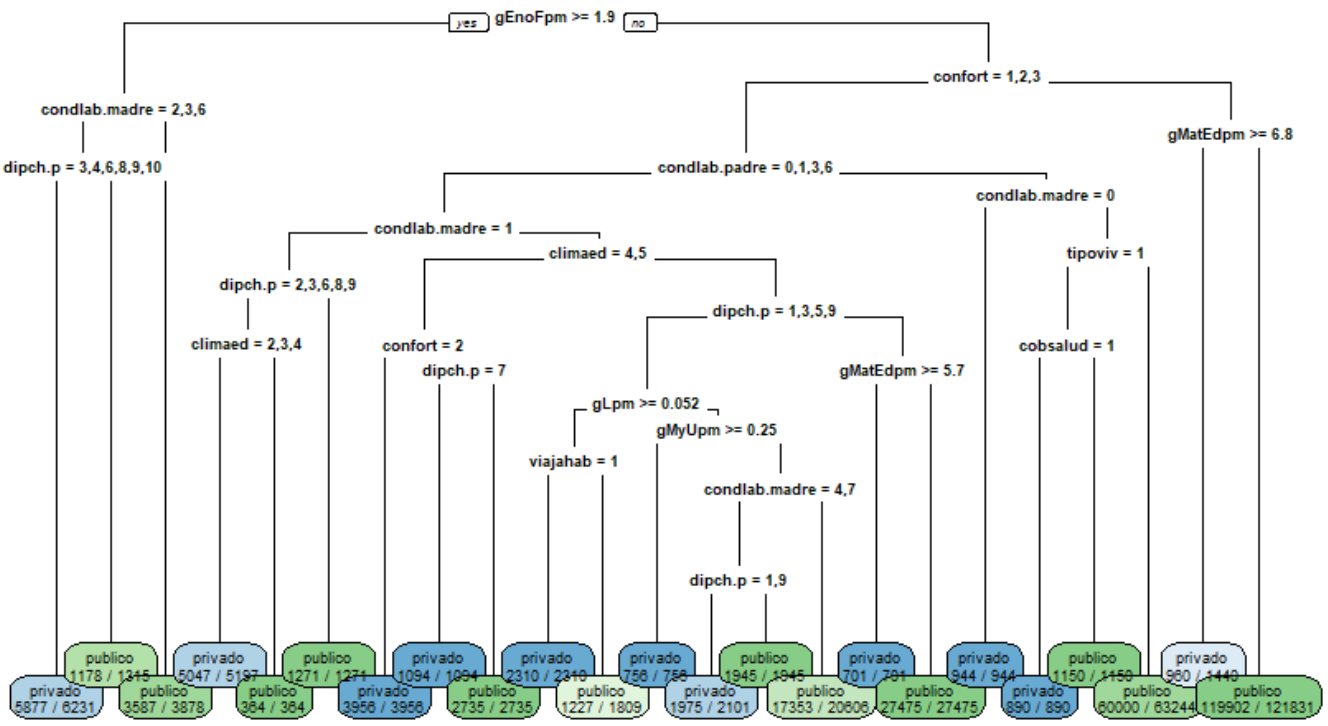
7.4. Modelo para la región Noroeste

Figura Nº 5. Árbol de clasificación para el nivel secundario - región Noroeste. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).



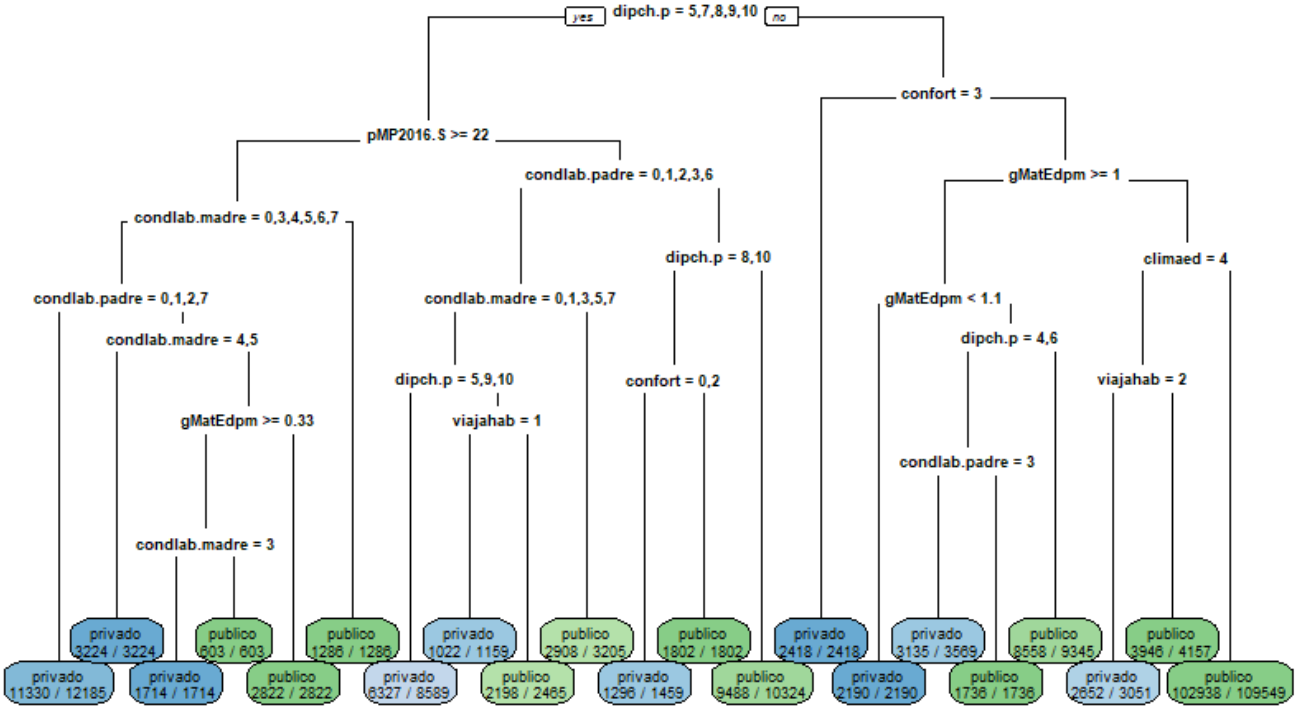
7.5. Modelo para la región Noreste

Figura Nº 6. Árbol de clasificación para el nivel secundario - región Noreste. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).



7.6. Modelo para la región Cuyo

Figura Nº 7. Árbol de clasificación para el nivel secundario - región Cuyo. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).



7.7. Modelo para la región Patagonia

Figura Nº 8. Árbol de clasificación para el nivel secundario - región Patagonia. Aglomerados urbanos de Argentina 2017-2018 (modelo completo).