Master in Management + Analytics

Deserción escolar en Argentina

Autoría: *Michel Torino, Tomás*
Año académico: *2023*

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

TESIS

DESERCIÓN ESCOLAR EN ARGENTINA.

MAYO 2023

Autor:

Tomás Michel Torino

Director:

Federico Favata

Fecha:

Mayo 2023

El presente trabajo es la tesis final requerida por la Universidad Torcuato Di Tella (UTDT) para la Maestría en Management + Analytics. La misma fue dirigida por Federico Favata a quien le agradezco encarecidamente por su labor académica, así como su aporte personal. Todos los errores son exclusiva responsabilidad del autor.

Resumen

A la luz de los esfuerzos realizados en América Latina y el Caribe en la implementación de Sistemas de Alerta Temprana para la prevención del abandono escolar, el presente estudio tiene como objetivo construir un modelo de predicción de deserción escolar que pueda ser aplicado en Argentina. Sin embargo, en nuestro país se evidencia una notable falta de herramientas y metodologías adecuadas para identificar a los jóvenes en riesgo de abandonar su educación debido a la falta de información requerida por un sistema de seguimiento educativo a nivel nacional.

Como una primera aproximación, se trabajó en la construcción del modelo propuesto, el cual fue entrenado, validado y probado utilizando datos socioeconómicos y habitacionales obtenidos de la Encuesta Permanente de Hogares. Es importante tener en cuenta las limitaciones de este enfoque, dado que se utilizó una muestra acotada que no fue específicamente diseñada con fines educativos. Con el objetivo de identificar y brindar apoyo a los jóvenes más vulnerables que no logran completar su educación obligatoria, se plantea una alternativa factible que permitiría al país obtener los principales indicadores de cada estudiante, generando así una base de datos suficiente para implementar este modelo de predicción a nivel federal.

Con la muestra utilizada se logró desarrollar un modelo de *Boosting* que ha demostrado un muy buen desempeño del 87.4% bajo la métrica AUC-ROC. Además, se lograron identificar los atributos con mayor influencia como factores desencadenantes del abandono escolar. Tras este ejercicio de identificación, se concluye el estudio con una propuesta para adaptar una política pública que facilite la provisión de un acompañamiento individualizado de manera eficiente en términos de costos, superando los desafíos específicos que cada provincia enfrenta en la gestión de datos.

Palabras clave: Argentina; deserción; educación; EPH; predicción; vulnerabilidad.

Abstract

In light of the efforts made in Latin America and the Caribbean regarding the implementation of Early Warning Systems for the prevention of school dropout, the aim of this study is to construct a predictive model for school dropout that can be applied in Argentina. However, our country exhibits a noticeable lack of appropriate tools and methodologies to identify youth at risk of dropping out of education due to the insufficient information required by a nationwide educational monitoring system.

As a first approach, we worked on the construction of the proposed model, which was trained, validated, and tested using socioeconomic and housing data obtained from the Permanent Household Survey. It is important to acknowledge the limitations of this approach, as it utilized a limited sample that was not specifically designed for educational purposes. To identify and provide support to the most vulnerable youth who are unable to complete their compulsory education, a feasible alternative is proposed, which would enable the country to obtain key indicators for each student, thus generating a sufficient database to implement this predictive model at a federal level.

Using the utilized sample, a Boosting model was developed, demonstrating a very good performance of 87.4% under the AUC-ROC metric. Furthermore, the attributes with the greatest influence as triggering factors for school dropout were identified. Following this identification exercise, the study concludes with a proposal to adapt a public policy that facilitates the provision of cost-effective individualized support, overcoming the specific challenges that each province faces in data management.

Keywords: Argentina; dropout; education; EPH; prediction; vulnerability.

Índice.

1. Introducción y objetivos 5
2. Revision de literatura 7
2.1. El impacto de la pandemia en la educacion 9
2.2. Factores desencadenantes de la desercion escolar 10
2.3. Sistemas de alerta temprana para la prevencion de la desercion escolar 12
2.3.1. SAT: Antecedentes en la region 13
2.4. Panorama actual y desafios para la Argentina 15
2.5. ¿Por qué usar machine learning? 17
3. Datos 18
3.1. Analisis descriptivo de los datos 19
3.2. Ingenieria de atributos 20
3.3. Limpieza de base 21
3.4. Tratamiento de clases desbalanceadas 23
4. Metodología 25
4.1. Conjuntos de entrenamiento, validacion y testeo 26
4.2. Modelos 28
4.2.1. Regresion Logistica 28
4.2.2. Ridge y LASSO 29
4.2.3. KNN 30
4.2.4. Arboles de Decision 31
4.2.5. Bagging 33
4.2.6. Random Forest 33
4.2.7. Boosting 34
4.2.8. Support Vector Machines 35
4.3. Evaluacion de Modelos 35
4.4. Optimizacion de Hiperparametros 38
5. Resultados 37
6. Discusion y propuesta para las politocas publicas 44
7. Conclusiones 47
8. Referencias 49
Apéndice 1: Datos 52
Apéndice 2: Análisis de significatividad de coeficientes 82

1. Introducción y objetivos

La deserción escolar en Argentina y en la región es una problemática alarmante. Actualmente en nuestro país, la tasa de abandono interanual afecta al 15,9%[1] de los jóvenes en el último año de la secundaria. Tan solo el 29%[2] de los estudiantes logra completar sus estudios secundarios dentro de los plazos establecidos en el cronograma educativo, mientras que la cifra llega al 54% al incluir a los repitentes, y al 71% considerando a los jóvenes-adultos que finalizan en institutos acelerados o programas de asistencias especiales.

Resulta preocupante constatar que únicamente 7 de cada 10 personas logran finalizar sus estudios obligatorios. Esta cifra, no obstante, oculta las profundas desigualdades existentes en el sistema educativo argentino. Mientras que en el segmento de mayores ingresos, 9 de cada 10 jóvenes logran culminar su educación obligatoria, en el grupo perteneciente al decil ingresos más bajos, la cifra se reduce a tan solo 3 de cada 10 jóvenes. Esta desigualdad excede el promedio de los países de América Latina y el Caribe, lo que nos enfrenta a un desafío aún mayor para garantizar la continuidad educativa de toda la población (CIPPEC, 2021).

La deserción escolar refiere a la interrupción o abandono de los estudios obligatorios por parte de los niños, niñas y adolescentes, y está relacionada con problemas sociales y económicos más amplios, como la pobreza, la falta de acceso a recursos educativos adecuados, la violencia en las escuelas y en los entornos cercanos a estas, la falta de motivación e interés tanto de los estudiantes como por parte de los padres o tutores legales, la necesidad de trabajar para contribuir al ingreso familiar, la falta de acceso a servicios básicos como agua potable y electricidad, así como la distancia a las escuelas, entre otros. Además, la pandemia de COVID-19 ha exacerbado este problema, ya que muchos estudiantes no tienen acceso a tecnologías adecuadas para seguir las clases virtualmente.

La educación no solo se enfoca en la adquisición de conocimientos, sino también en el desarrollo de habilidades, competencias y valores que promueven el crecimiento individual y colectivo, contribuyendo en el conjunto al progreso social y económico toda la nación. La problemática de la deserción tiene graves consecuencias tanto para los jóvenes en cuestión como para la sociedad en general. A largo plazo, el abandono escolar suele desencadenar, de forma redundante, en bajos niveles educativos, menores oportunidades de empleo y menores ingresos. La falta de recursos también suele asociarse a peores condiciones de salud, tanto física como mental, y con mayores tasas de criminalidad. Partiendo de mayores niveles de pobreza, la deserción no hace más que perpetuar esta desigualdad intergeneracionalmente, especialmente si consideramos que hay una brecha de 55 puntos en la concreción del nivel secundario entre las puntas de la distribución de los deciles de ingresos familiares (Templado *et al*. 2021).

Todos los niños tienen derecho a la educación, pero actualmente son muchos quienes están privados de ese derecho. Desde la perspectiva de Sen (1999), mientras que la educación es un

[1] Tasa de Abandono Interanual 2020/2021. Fuente: Ministerio de Educación de la Nación.
[2] Informe Nacional de Indicadores Educativos 2021. Fuente: Ministerio de Educación de la Nación.

medio para empoderar a las personas y promover la igualdad de oportunidades, el abandono escolar representa una barrera que perpetúa las desigualdades, limitando el desarrollo humano y restringiendo significativamente las libertades para alcanzar una mejor calidad de vida de estos miles de jóvenes afectados en nuestro país.

El acalorado debate educativo es un tema de constante disputa en el panorama político, pero el difícil escenario económico y social, que se agudizo tras el estallido de la pandemia, no ha logrado más que erosionar los niveles de asistencia escolar. Algunas de las más recientes propuestas para palear esta problemática consisten en flexibilizar drásticamente las normas de repitencia, argumentando que esta no es una verdadera "segunda oportunidad" sino más bien la antesala de la partida final. ¿pero acaso bajar las exigencias parece ser la forma de mejorar el sistema educativo? La calidad educativa no depende de la mayor o menor presión que se ponga para aprobar a los estudiantes, sino de los procesos de enseñanza y los mecanismos de acompañamiento que se den a quienes necesitan apoyo para fortalecer sus aprendizajes y continuar sus estudios (Kit *et al.* 2022).

Frente a un escenario de recesión económica, de la mano de una partida presupuestaria acotada y con recortes generalizados entre los cuales la educación asciende al podio con una reducción del 9% de la partida presupuestaria en términos reales[3], resulta de vital importancia comprender adecuadamente cuáles son los factores que tienen un mayor impacto en los niveles de deserción, para poder aplicar políticas focalizadas costo-efectivas que acompañen a los jóvenes con mayor riesgo de abandono.

Existe una gran falta de herramientas y metodologías adecuadas para identificar a los jóvenes desertores o en riesgo de deserción, para medir su situación, para evaluar las razones de su exclusión y consecuentemente para planificar políticas para su tratamiento. Es necesario adquirir una mejor visión general de los datos existentes, como los recopilados a través de registros administrativos y encuestas de hogares, para hacer un uso más eficaz de dicha información y poder aprovecharla. El primer paso para apoyar a estos jóvenes es comprender su situación. ¿Quiénes son los que no van a la escuela, dónde viven y por qué no van? Nuestro país tiene problemas para responder estos interrogantes debido a la falta de información confiable. Sin un sistema de seguimiento nacional eficaz, es imposible obtener cifras precisas. Además, con datos limitados y poco confiables, las posibilidades de analizar las causas de la exclusión son severamente limitadas, por lo que existe un gran riesgo de que las políticas y estrategias no estén fundamentadas en pruebas, o peor aún, se basen en pruebas poco fiables (*United Nations International Children's Emergency Fund*, 2016).

Considerando las implicancias de la deserción escolar, el objetivo de esta tesis consiste en predecir el abandono educativo mediante la aplicación de modelos de *machine learning*, con el fin de proporcionar nuevas herramientas que permitan aumentar la permanencia de los jóvenes en el sistema educativo y así lograr que cada vez más estudiantes finalicen sus estudios obligatorios en tiempo. A partir de esta información, es posible desarrollar intervenciones más

[3] Presupuesto educativo 2023 en comparación con el del 2022.

efectivas y enfocadas en aquellos estudiantes cuyas circunstancias personales los estén distanciando del sistema educativo, mejorando así las oportunidades académicas y profesionales de los jóvenes. A través este modelo, se busca identificar los grupos que requieren asistencia adicional mediante programas sociales como las tutorías o potenciales complementos a la Asignación Universal por Hijo (AUH), los cuales promueven la continuidad en el sistema escolar. Según un estudio realizado por Edo *et al.* (2015), la AUH incrementó la tasa de asistencia escolar en un 3,9% para los beneficiarios de 15 a 17 años.

Utilizando modelos de predicción de deserción, es posible identificar de manera temprana a los estudiantes con mayor riesgo de abandonar la escuela, y así tomar un enfoque más personalizado y efectivo, que tenga en cuenta los factores individuales y contextos específicos que pueden estar contribuyendo a esta situación. Además, un modelo de predicción de deserción escolar también puede ayudar a identificar las tendencias y patrones a nivel de grupo, lo que puede ser de utilidad para desarrollar políticas y programas de prevención de deserción escolar más efectivos, en los diferentes niveles de gobierno: municipal, provincial y/o nacional.

En el presente trabajo se desarrollará un modelo de predicción, elaborado con variables sencillas que la Encuesta Permanente de Hogares[4] (EPH) hoy en día ya recolecta, y que por ende también son variables que el gobierno, en gran medida, tiene en sus bases administrativas. La EPH cuenta con datos que tienen una estrecha relación con nuestro objeto de estudio, entre las cuales encontramos como las más explicativas de nuestro modelo al avance de la edad de los estudiantes, la temprana inserción en el mercado laboral o realización de tareas domésticas, condiciones de los padres como su nivel educativo y edad, así como las condiciones habitacionales y los niveles de ingresos monetarios familiares.

Dado que los factores más influyentes identificados por nuestro modelo son coincidentes con los de las poblaciones más vulnerables alcanzadas por asistencias sociales como la AUH o la Ayuda Escolar Anual, se propondrá aprovechar estos programas mediante la inclusión de un nuevo cuestionario en sus requisitos, con el cual sería posible obtener información sobre las condiciones habitacionales de las familias más vulnerables. Esta información, en conjunto con los datos socioeconómicos ya recopilados por la ANSES, permitiría generar una base de datos nacional comparativa a la EPH. Como medida final para abordar la problemática planteada, se analizará la viabilidad de implementar un sistema de tutorías a distancia dirigido a los jóvenes identificados como más propensos a la deserción escolar, buscando proporcionar un acompañamiento individualizado de manera eficiente en términos de costos, que pueda superar los desafíos específicos que cada provincia enfrenta en la gestión de datos y, de esta manera, ofrecer una política educativa de alcance nacional.

En esta línea, la presente investigación se desarrollará en 7 capítulos. En el próximo, se estudiará la situación actual de esta problemática, la cual tras el prolongado cierre de las escuelas a causa

[4] La EPH es una encuesta nacional a cargo del Instituto Nacional de Estadísticas y Censos (INDEC), que se publica trimestralmente y representa la población urbana del país en Argentina, recopilando de forma sistemática y permanente microdatos sobre las características demográficas, educativas, laborales y socioeconómicas de la población: *www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos*

de la pandemia ha generado la mayor erosión alguna vez registrada del sistema educativo para toda la región de América Latina y el Caribe, por tanto, será necesario comprender los factores que causan la deserción escolar y, en consecuencia, analizar los sistemas de alerta temprana que se están incorporando para atacar el problema. Adaptando esta situación al contexto argentino, continuaremos por desarrollar nuestro SAT prototipo con las bases de datos públicas que tenemos a disposición. El tercer capítulo estará dedicado a la descripción, limpieza y tratamiento de los datos que finalmente utilizaremos en el cuarto capitulo para entrenar todos los modelos que serán puestos a prueba a fin de lograr el mayor poder predictivo en datos desconocidos. Tras el análisis de los resultados en el quinto capitulo, en el sexto estaremos desarrollando nuestra propuesta para poder recabar la información necesaria para realizar este ejercicio de predicción de deserción a nivel federal, a fin de poder brindar un mayor soporte individualizado a los jóvenes identificados, mediante la adaptación de los sistemas de tutoría a distancia ya existentes. Finalmente, en el último capítulo se resumirán las conclusiones del trabajo.

2. Revisión de literatura

La deserción escolar tiene consecuencias importantes tanto a nivel social como individual. El analfabetismo es la manifestación más grave de este problema. Los costos sociales son difíciles de cuantificar, pero incluyen una fuerza laboral menos calificada y competente, mayor desempleo, mayor criminalidad, menor cohesión social, menores ingresos fiscales y mayores gastos de bienestar y salud pública, lo que en consecuencia genera impacto en el crecimiento económico. Además, se generan mayores gastos en programas sociales y transferencias para los sectores que no pueden generar recursos por sí mismos, lo que representa un costo social. Sin nombrar que hasta aquí no adentramos en la gravedad de esta decisión en todo el desarrollo futuro de la vida de un joven desertor. El abandono escolar también contribuye a la reproducción intergeneracional de las desigualdades sociales y la pobreza, lo que dificulta la integración social y la profundización de la democracia. (Muñoz, 2011).

En la literatura de educación existe gran cantidad de autores que encuadran a la deserción como el resultado final de un proceso de desvinculación de la escuela, en el cual esta decisión es impulsada por una gran variedad de factores, sobre los cuales es posible agrupar a los alumnos por estos tipos de condicionantes. En consecuencia, el tratamiento de esta problemática dependerá del contexto de cada estudiante, pudiendo encontrar soluciones efectivas mediante intervenciones particulares, como lo pueden ser las transferencias condicionadas de dinero, la provisión de información sobre los retornos futuros de la educación, o hasta el fortalecimiento de habilidades socioemocionales (UNICEF. 2016). A lo largo de este capítulo, profundizaremos en el panorama educativo de la región en el escenario postpandemia, los factores condicionantes de los logros educativos, y los sistemas de alertas temprana para la identificación y prevención de la deserción, finalizando en el análisis de la situación y las perspectivas de la Argentina.

2.1. El impacto de la pandemia en la educación

La pandemia COVID-19 generó un doble impacto, sanitario y económico, que afectó al sector educativo de América Latina y el Caribe de manera sin precedentes, llevando al cierre masivo de escuelas y afectando[5] a más de 170 millones de estudiantes en toda la región. A pesar de los esfuerzos realizados por los países para mitigar la falta de educación presencial mediante sistemas de educación a distancia, el impacto en la educación ha sido demasiado alto en toda la región. Se espera un aumento de más del 20% en la "pobreza de aprendizaje"[6] al final de la educación primaria y que más de dos tercios de los estudiantes de educación secundaria caigan por debajo de los niveles mínimos de rendimiento esperados, con mayores pérdidas de aprendizaje para los estudiantes más desfavorecidos. Estas pérdidas en el aprendizaje afectan principalmente al quintil inferior en la escala de ingresos, lo que puede exacerbar aún más la brecha socioeconómica en materia de resultados educativos. Además, La pandemia también ha afectado a la nutrición de los estudiantes, ya que muchos dependían de las comidas que se les servían en la escuela. Según las estimaciones realizadas por el Banco Mundial (2021), la enorme pérdida de educación, capital humano y productividad podría traducirse en una caída de ingresos agregados a nivel regional de 1.700 millones de dólares, equivalentes a aproximadamente el 10% de los ingresos totales.

Con la extensión del período de educación no presencial en 2020, que en América Latina y el Caribe fue el más prolongado del mundo, fue creciendo la preocupación sobre si la continuidad educativa sin presencialidad llevaría a un deterioro de los aprendizajes, una profundización de las desigualdades educativas y una mayor tasa de deserción escolar. La pandemia de COVID-19 ha generado una mayor demanda de información sobre el sector educativo y sus implicaciones sociales, lo que ha llevado a la necesidad de recopilar y analizar datos para diseñar políticas efectivas y mitigar los efectos negativos de la pandemia en la educación. Además de la preocupación por el impacto de la educación no presencial, también surgió la necesidad de información sobre la situación socioeconómica de las familias y su capacidad para acceder a los recursos tecnológicos necesarios para el aprendizaje en línea. Esto se debió a que la educación no presencial requería el uso de dispositivos electrónicos y conectividad a internet, lo que creaba barreras para el acceso a la educación para aquellos que no tenían disponibilidad de estos recursos[7]. Por lo tanto, la necesidad de información se extendió a la identificación de las

[5] Debido al cierre masivo de escuelas, a febrero de 2021, alrededor de 120 millones de niños en edad escolar habían perdido o corrían el riesgo de perder un año completo presencial del calendario escolar, con graves impactos educativos.

[6] La “pobreza de aprendizaje”, definida como el porcentaje de niños de 10 años incapaces de leer y comprender un relato simple, podría haber crecido de 51% a 62,5%. Esto podría equivaler a 7,6 millones adicionales de niños y niñas en educación primaria “pobres de aprendizaje” en la región de América Latina y el Caribe. (Banco Mundial. 2021)

[7] En América Latina y el Caribe tan solo el 16% de hogares más pobres cuenta con una computadora en casa y el 23% tiene acceso a internet. En comparación, el 68% de los hogares ricos cuenta con una computadora en casa y un 74% tiene acceso al internet. (BID, 2020)

desigualdades en el acceso a la tecnología y el impacto de estas desigualdades en los resultados educativos. (*United Nations Educational, Scientific and Cultural Organization*, 2021).

Los efectos de la pandemia trascienden el impacto directo sobre el aprendizaje o los años de escolaridad; la crisis ha aumentado los niveles de ansiedad y estrés entre los estudiantes y ha tenido un impacto en su bienestar emocional y psicológico que en muchos persistirá durante toda la vida. Muchos estudiantes, particularmente de los grupos de menores ingresos y los que ya estaban aprendiendo muy poco incluso antes de la crisis, ahora se encuentran en mayor riesgo de abandonar la escuela debido a la pandemia, situación que incluso se ve agravada en este contexto de dificultades económicas y recesión. Algunas simulaciones sugieren que la deserción escolar en América Latina y el Caribe pudo aumentar en un 15% debido a la pandemia. Es importante tener en cuenta todos estos factores al abordar la crisis educativa en la región y tomar medidas para abordar los efectos a largo plazo que la pandemia ha tenido en los estudiantes (Banco Mundial, 2021). De acuerdo con las estimaciones del BID (2020), en América Latina y el Caribe, aproximadamente 1,2 millones de niños, niñas y adolescentes se vieron obligados a abandonar sus estudios debido a las consecuencias directas de la pandemia. Esta situación representa un retroceso de casi una década en términos de inclusión educativa en la región.

En palabras de Bucciarelli *et al.* (2022), menores aprendizajes y un contexto socioeconómico deteriorado son factores que incrementan las probabilidades de que, aún con clases presenciales, muchos niños, niñas y adolescentes interrumpan su escolaridad. Por esto, es prioritario avanzar en políticas orientadas no solo a reincorporar a los estudiantes que hayan abandonado a causa de la pandemia, sino que se pueda identificar a quienes presentan mayor riesgo de abandonar para así brindarles el apoyo que garantice trayectorias escolares completas y de calidad en los años por venir.

2.2. Factores desencadenantes de la deserción escolar

La medición del abandono escolar y sus causas es un requisito en el objetivo de lograr la universalidad de la educación primaria y secundaria. La Iniciativa Mundial por los Niños y Niñas fuera de la Escuela liderada por la UNESCO y la UNICEF contribuyó a una mayor comprensión de las causas del abandono escolar y a la identificación de estudiantes en riesgo de abandonar sus estudios. Cada caso de abandono es único y puede estar influenciado por una infinidad de factores que a menudo se interconectan y refuerzan mutuamente, acrecentando la probabilidad de deserción mientras más indicadores se evidencien, como, a modo de ejemplo, lo pueden ser la pobreza, la falta de infraestructura escolar, los problemas familiares y del hogar, la discriminación, las barreras geográficas, entre otros.

La deserción escolar es el último eslabón en la cadena de fracasos escolares. Antes de desertar, lo mas probable es que un alumno haya repetido, estirando su trayectoria escolar, dañando su autoestima y perdiendo de vista su formación como un logro esperanzador. La deserción escolar

rara vez es un evento inesperado, sino que se presenta como una cadena de hechos que van elevando el riesgo de abandono, a medida que se avanza en edad y se experimentan crecientes dificultades de rendimiento y adaptación. En Argentina un repitente tiene alrededor de un 20% más de probabilidades de abandonar la escuela. (Muñoz, 2011).

El **cuadro 1** presenta un resumen de algunos de los factores asociados al riesgo del abandono escolar, aunque debe destacarse que las razones son infinitas, siendo imposible resumir todas las posibles causas.

Cuadro 1. Factores asociados al riesgo de abandono escolar.

Tipología	Descripción del problema
Desempeño académico	Dificultades en el desempeño académico.
Conducta	Baja vinculación con la escuela. Incluye bajos niveles de atención y concentración, no hace tareas, no participa en actividades extracurriculares, socialmente aislado.
	Problemas de conducta: suspensión por mala conducta grave, como comportamiento antisocial, intimidación, violencia, robo, uso de sustancias o problemas con la ley.
Ausentismo crónico	Ausentismo frecuente: 10-20 por ciento de días perdidos durante el año escolar actual.
	Ausentismo severo: más del 20 por ciento de los días perdidos durante el año escolar actual.
Discapacidad	Basado en el nivel de dificultad para participar y aprender en clase debido a la condición de discapacidad específica.
Acceso y progresión en el sistema educativo	Al menos dos años mayor que la edad esperada para su grado, pero no ha repetido un grado.
	Ha abandonado anteriormente, pero regresó a la escuela.
	Sin educación preprimaria (para estudiantes de primaria).
	Repitiendo o ha repetido un grado.
Responsabilidades del adulto temprano	Carga de trabajo del estudiante fuera de la escuela (remunerada o no) que dificulta las posibilidades de escolarización. Esto puede incluir el trabajo doméstico o el cuidado de miembros de la familia (por ejemplo, familiares con discapacidades, que padecen enfermedades crónicas o hermanos menores).
	Matrimonio.
	Embarazo o paternidad.
Circunstancias familiares y con pares	Dificultad para ir a la escuela (por ejemplo, vive lejos de la escuela y no puede pagar o no tiene acceso al transporte público).
	Vive en circunstancias difíciles que se definirán más específicamente de acuerdo con el contexto nacional, por ejemplo, pobreza, condiciones de vida afectadas por situaciones extremas como desastres naturales o

	eventos equivalentes, condición de refugiado o de personas desplazadas, o tiene uno o más miembros de la familia discapacitados.
	Familia numerosa.
	Parece tener problemas familiares en el hogar que afectan significativamente al estudiante, como conflictos, peleas, separación de los padres, pérdida de un ser querido, abuso de drogas o alcohol, abuso físico, sexual o emocional, etcétera.
	Tiene dificultades con el idioma que se usa en la escuela porque se habla un idioma diferente en el hogar.
	Bajo involucramiento familiar: por ejemplo, padres o tutores que no asisten a las reuniones escolares, se niegan a hablar sobre su hijo con los maestros, no expresan interés en la educación del niño, etcétera.
	Presión de pares: amigos o hermanos han abandonado la escuela.
	Víctima de acoso escolar.

Fuente: Adaptado por Perusia (2021) de UNICEF y UIS. (2016). Monitoring Education Participation: Framework for Monitoring Children and Adolescents who are Out of School or at Risk of Dropping Out. UNICEF Series on Education Participation and Dropout Prevention, Vol I. UNICEF Regional Office for Central and Eastern Europe and the Commonwealth of Independent States. Geneva, UNICEF.

2.3. Sistemas de alerta temprana (SAT)

Entender las razones por las cuales los estudiantes abandonan la escuela es crucial. Sin embargo, para poder aplicar políticas públicas adecuadas a esta problemática, es aún más importante identificar con precisión quiénes son los estudiantes con mayor riesgo de abandonar la escuela. En situaciones donde los recursos son limitados y hay múltiples prioridades, como lo es en la mayoría de los sistemas educativos del mundo, identificar precisamente a los estudiantes en riesgo es especialmente importante, ya que permitirá enfocar las intervenciones de manera efectiva donde más se necesitan.

Los sistemas de alerta temprana (SAT) son herramientas preventivas para identificar a los estudiantes en riesgo de abandonar la escuela. Se basan en señales específicas de problemas que pueden contribuir al abandono escolar. Los SAT buscan identificar estas señales de alerta a tiempo para que las escuelas y equipos competentes implementen el apoyo adecuado para contribuir a la continuidad educativa. En la práctica, la mayoría de la predicción del abandono escolar se basa en indicadores de la participación y el aprendizaje de los estudiantes, como la asistencia, las infracciones de comportamiento, las calificaciones de los cursos y el rendimiento en los exámenes (Adelma *et al.* 2018). Aunque estos indicadores son efectivos, un abordaje integral para la prevención de este fenómeno debería también prestar atención a los factores inherentes a los estudiantes, factores propios del sistema educativo y factores socioeconómicos. Los SAT pueden contribuir a identificar estos problemas, pero no son la mejor herramienta para profundizar en un diagnóstico sobre estas situaciones.

Tal como describe Perusia (2021) para lograr el buen funcionamiento de los SAT es importante establecer la metodología de recolección de datos y aprovechar fuentes de información ya existentes, como los sistemas de información para la gestión educativa (SIGED). Sin embargo, es importante considerar la oportunidad de la información, ya que no siempre estará actualizada para propiciar intervenciones oportunas. Los SIGED típicamente registran las condiciones finales de aprobado o desaprobado de cada estudiante, sin embargo, esta información ya es tardía a los fines del SAT, si con este se pretende evitar la repitencia de los alumnos. Es recomendable tener información a nivel individual para mayor focalización de las acciones, pero si no está disponible, se pueden identificar tendencias en el comportamiento agregando indicadores por escuelas o áreas territoriales.

Los sistemas de información y gestión educativa son fundamentales para planificar y desarrollar políticas educativas efectivas. Como se explicará en el apartado siguiente, casi todos los países de la región están trabajando arduamente para recopilar información sobre los estudiantes y otros aspectos del sistema educativo. Dado el gran volumen de datos generados diariamente por los sistemas educativos y su potencial de uso, los países deben aprovechar la tecnología para mejorar la calidad y la accesibilidad a esta información de manera oportuna. El desafío radica en desarrollar sistemas de datos eficaces que puedan utilizarse para mejorar continuamente el sistema y tomar decisiones informadas. Finalmente, un paso crucial es considerar cómo transmitir la información a las escuelas y cómo debería utilizarse allí para evitar la estigmatización de los estudiantes. Se recomienda que los sistemas de alerta temprana (SAT) se desarrollen como una herramienta de monitoreo de los estudiantes y que los reportes del sistema estén acompañados por una capacitación que cubra las áreas de protección y seguridad de los datos y manejo de información sensible.

2.3.1. SAT: Antecedentes de la región

El desarrollo de los SAT comenzó en Estados Unidos para combatir el abandono escolar en la educación secundaria y mejorar las tasas de graduación en este nivel educativo. Alrededor de un tercio de los estudiantes de instituciones educativas públicas abandonaban sus estudios a principios del siglo XXI, lo que impulsó la creación de estas herramientas que para el año escolar 2014/2015, ya estaban presentes en el 52 por ciento de las escuelas secundarias públicas en ese país. En Europa, para el año 2012 ya había 15 países que tenían herramientas similares en funcionamiento. En comparación, en la región de América Latina y el Caribe, la experiencia en la utilización de los SAT es limitada y reciente. Sin embargo, la pandemia generó una mayor preocupación por la exclusión escolar, lo que sirvió para dar el puntapié inicial e impulsar su desarrollo generalizado. En 2020, los ministros de Educación de los países miembros del Sistema de Integración Centroamericana (SICA) adoptaron el Plan de Contingencia en Educación, que recomendaba el desarrollo y/o fortalecimiento de mecanismos de alerta temprana para prevenir la exclusión escolar (UNESCO. 2021).

Si bien el desarrollo de los SAT en América Latina tiene una trayectoria acotada, algunos países de la región ya están acumulando experiencia en el desarrollo de estas herramientas para el combate contra la deserción escolar. Colombia fue el precursor de la región, quien desde el 2012 posee el Sistema de Información para el Monitoreo, Prevención y Análisis de la Deserción Escolar, SIMPADE. En 2017 Guatemala lanzó un programa piloto, que utiliza datos para identificar a los alumnos en riesgo de desertar, y también brinda pautas con estrategias a los directores, incluyendo acciones sencillas para alentar a los alumnos a seguir en el colegio[8]. Además, las escuelas con los mejores resultados en la prevención del abandono reciben certificados de reconocimiento formal. En el caso de Costa Rica, desde 2020 se encuentra operativo un sistema de alerta temprana que depende de la Unidad para la Permanencia, Reincorporación y Éxito Educativo (UPRE) del Ministerio de Educación Pública de ese país. En El Salvador se implementó ese mismo año el Mecanismo de Alerta Temprana, cuyo desarrollo se había iniciado en 2018. Perú inició en 2020 el sistema «Alerta Escuela», una herramienta dirigida a los directores y docentes de todas las instituciones educativas públicas y privadas de la educación básica, que emplea *machine learning* y datos a nivel de alumnos para identificar a aquellos en riesgo de desertar. La información se actualiza mensualmente y los directores y docentes reciben las estrategias pedagógicas y administrativas para apoyar a los estudiantes en riesgo. En Chile, el sistema de alerta temprana, lanzado inicialmente en 2019 solo en algunas regiones del país, se está ampliando a escala nacional para mitigar la deserción escolar como respuesta al cierre de escuelas. Por su lado, Honduras inicio en 2021 el Sistema de Alerta y Respuesta Temprana (SART). Finalmente, el Ministerio de Educación de Panamá constituyó la Red de Prevención y Retención Escolar, aunque el sistema aún está en desarrollo (UNESCO. 2021 ; World Bank. 2021).

El **cuadro 2** presenta los sistemas de alerta temprana identificados en países de América Latina y el Caribe, y específica sobre qué variables de riesgo recogen información y el año en que inician su funcionamiento.

Cuadro 2. Sistemas de alerta temprana en países de América Latina y el Caribe y variables de riesgo asociadas al abandono escolar

	Belice (2015)	Chile (2020)	Colombia (2012)	Costa Rica (2020)	Honduras (2021)	Panamá (2020)	Perú (2020)	El Salvador (2020)	Uruguay (2016)
Individuales									
Desempeño académico	x	x	x	x	x	x	x	x	x
Asistencia	x	x	x	x	x	x	x	x	x

[8] La evidencia procedente de Guatemala muestra que los datos administrativos se pueden utilizar para identificar correctamente el 80 por ciento de los estudiantes de sexto grado con riesgo de deserción. Según el estudio realizado por Haimovich et al. (2018), el programa redujo la probabilidad de abandono escolar en las escuelas bajo tratamiento en 1,3 puntos porcentuales y en 3,1 puntos porcentuales cuando se considera el impacto sobre quienes cumplieron con el programa de capacitación y de estímulos conductuales. Los resultados sugieren que el impacto es explicado principalmente por la intervención básica de proporcionar orientación y capacitación sobre la manera de prevenir la deserción.

Aspectos emocionales y de conducta	x		x	x				
Trayectoria educativa (interrupciones o repitencia)		x	x		x		x	x	x
Tiene un empleo formal o informal								x
Familiares								
Embarazo adolescente o tareas de cuidado			x	x			x	x
Matrimonio temprano							x	
Nivel socioeconómico		x		x		x		x
Entorno social y familiar (nivel educativo de sus padres, situación de desempleo, mudanza)	x	x	x	x	x	x		
Institucionales								
Clima escolar (acoso escolar, discriminación)			x	x			x	
Condiciones edilicias o hacinamiento			x					
De contexto								
Condiciones de vulnerabilidad (migración, explotación sexual, trata de personas, trabajo infantil y adolescente)			x	x	x		x	
Manifestaciones de violencia (delincuencia, sexual y de género, drogadicción)	x			x	x		x	x

Fuente: Adaptado de Perusia, J. y A. Cardini (2021). Sistemas de alerta temprana en la educación secundaria: prevenir el abandono escolar en la era del COVID-19. Documento de Política Pública 233. Buenos Aires: CIPPEC

Aunque los SAT han demostrado tener algunos resultados positivos, todavía hay poca evaluación sobre su eficacia. Los nuevos SAT creados en respuesta a la pandemia son muy recientes para ser evaluados y analizar sus resultados. Es por eso que esta herramienta de apoyo a la política y gestión escolar todavía se considera una innovación en la región. Se pueden tener expectativas positivas, pero al mismo tiempo se debe reflexionar sobre las mejores condiciones para su funcionamiento y sostenibilidad a largo plazo. La efectividad de los SAT debe evaluarse en función de su capacidad para reducir el abandono escolar, lo cual depende de la eficacia de las intervenciones para abordar el riesgo de desvinculación. La producción de los datos necesarios para identificar a los estudiantes en riesgo de abandonar es un componente fundamental de los SAT y también presenta desafíos en el contexto actual de la región. Por lo tanto, se considera recomendable la interconexión entre los SIGED y los SAT para aprovechar al máximo la información recopilada por los primeros. Sin embargo, el éxito de los SAT dependerá de la consolidación de los SIGED como sistemas de información con altos estándares de calidad, preferiblemente con registros individualizados y digitalizados que cubran ampliamente el sistema. (UNESCO. 2021).

2.4. Panorama actual y desafíos para la Argentina

En Argentina, la implementación de los SAT es limitada debido a que las iniciativas en esta área no han perdurado en el tiempo o son recientes. Existen programas como "Asistiré" y "SInIDE

Acompañar" a nivel nacional, y planes como "Vuelvo a Estudiar" en Santa Fe y la "Red de Apoyo a la Protección de Trayectorias Educativas" en Mendoza, que buscan recopilar información nominalizada sobre la trayectoria educativa de los estudiantes y brindar intervenciones a aquellos en situación de riesgo. También se han establecido marcos normativos, como la Ley Programa de Cédula Escolar en 2018 y el acuerdo del Consejo Federal de Educación en 2020, para avanzar en sistemas compatibles con los SAT. No obstante, la infraestructura digital en Argentina es deficitaria en gran parte del sistema educativo nacional, lo que se presenta como uno de los principales obstáculos en la implementación de estos sistemas. Para poner en marcha los SAT, es necesario acelerar la consolidación de los Sistemas de Información para la Gestión Educativa (SIGED) a nivel nacional y jurisdiccional, para poder recopilar datos de calidad de cada estudiante de manera individualizada y digitalizada de manera flexible y periódica. Además, se requiere el diseño de protocolos de intervención ligados a las principales alarmas y el establecimiento de un marco sólido para la protección de los datos personales. (Bucciarelli *et al.* 2022).

El principal desafío es asegurar que todos los estudiantes de los niveles obligatorios estén en la base de datos nominalizada. Sin embargo, hay una gran disparidad en los niveles de avance de los SIGED, ya que no todas las provincias tienen un sistema nominalizado que permita el seguimiento de la trayectoria de cada estudiante y su vinculación a un establecimiento educativo. En tal sentido, el mayor esfuerzo nacional está puesto en el Sistema Integral de Información Digital Educativa (SInIDE)[9], el cual, si bien se encuentra en constante crecimiento y ampliando el numero de provincias incluidas en el aplicativo[10], aún tiene una enorme trayectoria por recorrer en busca de su universalidad, considerando que tan solo un tercio de los estudiantes en niveles obligatorios están incluidos en el. Es crucial avanzar en la digitalización de los registros escolares para aumentar la base de datos para los Sistemas de Alerta Temprana y reducir la carga de trabajo sobre las instituciones educativas.

Para una implementación efectiva de los Sistemas de Alerta Temprana, es importante garantizar que los Sistemas de Información para la Gestión Educativa sean interoperables con otros sistemas que también cuentan con información relevante sobre los estudiantes. Acceder a variables importantes relacionadas con el abandono escolar contribuirá a una detección más precisa del riesgo de abandono y al desarrollo de intervenciones más completas. Es crucial asegurar la protección de los datos personales y la información utilizada por los SAT para generar las alertas. Para fortalecer los SIGED y desarrollar los SAT, es fundamental contar con recursos económicos, técnicos y humanos disponibles para su implementación efectiva. Muchos países de la región han recurrido a fondos multilaterales para financiar estos sistemas. Por ejemplo,

[9] sinide.educacion.gob.ar

[10] A principios del 2023, el SInIDE cuenta con 5 provincias que ya están utilizando el sistema (Tierra del Fuego, Misiones, La Rioja, Catamarca y Salta), 4 provincias que están en proceso de integración de sus propios sistemas al sistema nacional (Buenos Aires, Córdoba, Corrientes y Río Negro), otras 10 se adhirieron este año y están comenzando la carga (Chaco, Chubut, Entre Ríos, Formosa, Jujuy, La Pampa, Mendoza, Santa Cruz, Santiago del Estero y Tucumán), mientras que las 5 restantes no cargan ni transfieren información al SInIDE (Ciudad de Buenos Aires, Neuquén, San Luis, San Juan y Santa Fe)

Chile y Uruguay recibieron financiamiento del Banco Interamericano de Desarrollo (BID), y El Salvador y Honduras contaron con el apoyo de UNICEF y USAID, respectivamente.

2.5. ¿Por qué usar machine learning?

El desarrollo de los Sistemas de Alerta Temprana para la identificación de los estudiantes en riesgo de abandono requiere de la adopción de una metodología, la cual puede ser basada en indicadores o en análisis de datos utilizando técnicas de *machine learning*. Como su nombre lo indica, el primer modelo requiere de indicadores con los que se establecen umbrales para determinar qué estudiantes se encuentran en riesgo de desertar. Algunos de los indicadores ya mencionados son las inasistencias o las notas insuficientes. Por el otro lado, los modelos basados en *machine learning* se alimentan de grandes volúmenes de información que pueden incluir características de las trayectorias, las familias o las condiciones de vida de cada alumno, mediante los cuales se estima la probabilidad de abandono escolar de cada joven. La adopción de un modelo u otro dependerá de la disponibilidad de la información con la que se cuente y/o su factibilidad de obtención.

Adelman *et al.* (2018) realizaron un ejercicio de predicción de deserción escolar en Guatemala y Honduras con el que lograron identificar correctamente al 80 % de los estudiantes de sexto grado que abandonarán los estudios durante el próximo año, basándose en las variables disponibles en los datos existentes (características sociodemográficas básicas de los estudiantes, sus hogares y sus comunidades) y sin incluir otros factores que a menudo se usan en la predicción de la deserción escolar en otros países, como calificaciones académicas, registros de asistencia, e infracciones de conducta. Según su análisis, los modelos funcionaron mejor en Guatemala, donde las tasas de deserción escolar medidas por los datos administrativos corresponden muy de cerca a las estimaciones basadas en encuestas de hogares, logrando incluso mejor desempeño que otros enfoques de focalización comúnmente utilizados y también que los modelos utilizados en los EEUU.

Para el desarrollo de nuestro trabajo de predicción, optamos por evaluar diversos modelos basados en *machine learning* ya que consideramos son una potente herramienta capaz de identificar patrones y tendencias en los alumnos, con las que podremos predecir quienes son los de mayor riesgo de abandono. Siguiendo la idea del trabajo de Adelman, y considerando la disponibilidad de información que poseemos a nuestro alcance, decidimos afrontar este análisis utilizando la Encuesta Permanente de los Hogares del INDEC, la cual recolecta las principales características demográficas, educativas, laborales y socioeconómicas de la población. En el siguiente capítulo ahondaremos en el detalle de esta base de datos y en cómo fue adaptada a nuestro objeto de estudio.

3. Datos

El presente trabajo fue elaborado con los microdatos de la Encuesta Permanente de Hogares. La EPH es una encuesta continua que se realiza trimestralmente en Argentina con el objetivo de obtener información sobre la situación socioeconómica de los hogares y las personas que los componen. La encuesta se realiza mediante un cuestionario estructurado que se aplica a un grupo de hogares seleccionados de forma aleatoria a partir de una muestra representativa de la población. Este cuestionario abarca temas como el empleo, los ingresos, la educación, la vivienda, la salud y otros aspectos relevantes para la medición de la calidad de vida de la población.

Cada hogar es entrevistado durante cuatro trimestres, dos de ellos consecutivos, seguidos por dos trimestres sin entrevista, y luego se realizan otras dos entrevistas en los trimestres consecutivos siguientes, con el fin de permitir análisis longitudinales. Como resultado, la muestra utilizada en este trabajo no será tan grande como la suma de las cuatro muestras trimestrales originales utilizadas para entrenar los modelos, ya que se eliminaron las observaciones repetidas y se conservó sólo la encuesta más reciente de cada joven y su hogar. Los encuestados rotan cada trimestre para asegurar que la muestra sea representativa y no esté sesgada por eventos específicos que puedan afectar a ciertos hogares en un momento determinado.

Diversos autores han optado por utilizar esta encuesta para sus trabajos asociados a las problemáticas educativas[11], ya que la misma contiene una significativa cantidad de variables de los hogares y las personas que posibilitan el análisis de las principales características demográficas y socioeconómicas de la población. No obstante, resulta atinado mencionar que la EPH presenta limitaciones ya que únicamente tiene representatividad urbana, dejando fuera de análisis a los poblados rurales, y no fue específicamente diseñaba con fines educativos, por lo que no cuenta con cierta información que sería deseable para nuestro objeto de estudio, como podrían serlo las inasistencias, las calificaciones o las infracciones de conducta de los alumnos.

En Argentina no hay encuestas públicas disponibles que hayan sido elaboradas para estudiar la deserción escolar y/o la graduación de las escuelas. Existen evaluaciones estandarizadas que dan cuenta de los resultados académicos de los estudiantes[12], pero estas no capturan al alumno que abandonó la escuela y no brindan tanta información socioeconómica de los jóvenes y sus familias como la que proporciona la EPH. El sistema educativo argentino no posee un sistema centralizado de información, lo que genera una gran disparidad entre los datos que cada una de las escuelas recolecta, especialmente considerando que todavía hay provincias que no centralizan los registros de las escuelas. A nivel federal, la Administración Nacional de la Seguridad Social (ANSES) es el organismo que posee mayor disponibilidad de datos, registrando

[11] Sosa Escudero y Marchionni (1999); Giovagnoli (2007); Edo *et al. (*2015); Agrogue y Orlicki (2018), entre otros.

[12] Como las pruebas aprender o las evaluaciones PISA.

las composiciones de cada familia y las condiciones laborales de cada individuo. Como se desarrollará más adelante, a raíz de la información de la EPH, se pretende simular el mejor nivel de información que un organismo federal como la ANSES podría llegar a obtener.

Cada publicación trimestral de la EPH incluye dos bases de datos con registros de los hogares por un lado y de las personas por el otro. De la unión de ambas a través del código de vivienda y el número de hogar, nos fue posible obtener las características de cada individuo (género, edad, participación en el mercado laboral y tipo de escuela a la que asiste/asistió), las características del jefe del hogar (género, años de educación, situación laboral y si tiene cónyuge), y las características de la vivienda (ingreso per cápita, tamaño, número de habitantes, acceso a agua corriente, ubicación y materiales de la construcción), entre otros[13].

3.1. Análisis descriptivo de los datos

Para nuestro análisis y entrenamiento de modelos, decidimos utilizar las encuestas de los últimos dos trimestres del 2021 y los dos primeros del 2022 a fin de poder capturar los patrones del último año completo disponible, mientras que por separado utilizamos la encuesta del tercer trimestre del 2022 para realizar el testeo de nuestros modelos, siendo esta la última encuesta publicada al momento. Cada una de las encuestas cuenta con el relevamiento de aproximadamente 50 mil individuos pertenecientes a un promedio de 17 mil hogares de los cuales se obtienen 88 atributos de cada hogar y 177 de cada individuo.

Tras la unión de las bases de datos mencionadas, y habiendo descartado a los individuos y los atributos duplicados entre las diferentes encuestas, nos quedamos con una muestra total de 144 mil observaciones con 242 atributos de cada una de ellas. Nuestra muestra cuenta con un total de 38 mil jóvenes de hasta 18 años de edad, de los cuales 880 declaran haber abandonado sus estudios obligatorios.

De una primera lectura de los datos, es posible mencionar los siguientes aspectos de interés[14]:

- Mientras que la muestra total se encuentra equitativamente distribuida por sexos, la muestra de los desertores cuenta con una marcada preponderancia de varones quienes representan casi dos tercios del total, muy probablemente relacionado con el acceso al mercado laboral y la desigualdad de ingresos entre géneros.
- El 46,9% de los desertores viven en hogares donde el jefe de hogar está desocupado. Esto más que duplica a la proporción de jefes desocupados de la muestra total. Esta situación de vulnerabilidad de ingresos presiona a los jóvenes hacia una temprana incorporación al mercado laboral que se contrapone con la demanda de las escuelas. Al analizar la inserción laboral de los jóvenes, se observa que un tercio de los desertores

[13] Ver diseño de registro y estructura para las bases preliminares de Hogar y Personas: *www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_registro_3T2022.pdf*

[14] Ver Apéndice 2: Análisis de significatividad de coeficientes

se encuentra laboralmente activo, mientras que para los no desertores esa proporción es 6,2 veces menor. También se observa una gradual reducción de los niveles de deserción escolar inversamente relacionada con el crecimiento de los ingresos familiares.

- El nivel educativo del jefe de hogar evidencia una notable influencia en el máximo nivel educativo alcanzado por sus hijos. Mientras en de la muestra total el 56% de los jefes de hogares tienen estudios secundarios completos o superiores, por el otro lado, los jóvenes que desertaron provienen en su amplia mayoría de hogares con menor educación, donde tan solo el 26% de los jefes de hogar finalizaron la educación secundaria, y más del 50% ni siquiera llegaron a ingresar a dicho nivel.

3.2. Ingeniería de atributos

La ingeniería de atributos es el proceso de seleccionar, transformar y crear atributos a partir de los datos originales, permitiéndole al modelo extraer más y mejor información. Es una tarea crucial en el proceso de aprendizaje automático, ya que la calidad y relevancia de los atributos pueden influenciar drásticamente en el rendimiento del modelo. Gran parte de los esfuerzos en la implementación de los algoritmos de *Machine Learning*, se dedica en el preprocesamiento y transformación de los datos (Bengio *et al,* 2013).

Como fue mencionado anteriormente, la EPH no fue específicamente diseñada con fines educativos, por lo que hay ciertos aspectos en los que no será posible profundizar de la forma deseada. No obstante, es de destacarse el potencial que brinda esta base de acceso público por la gran cantidad hogares encuestados de una amplia cobertura geográfica, así como por la basta cantidad de información que les es solicitada a cada uno de ellos. Si bien la encuesta no deja de ser una herramienta que brinda datos puramente declarativos, la información recabada sistemáticamente está muy bien estructurada y limpia, permitiendo ampliar el análisis por sobre el cuestionario realizado.

Para llevar a cabo este trabajo fue necesario incorporar variables adicionales que no fueron incluidas en el cuestionario original, por lo que se decidió generarlas a partir de la información ya existente en la EPH. Lógicamente, la primera de ellas que inevitablemente se debió incorporar, fue la variable objetivo que se busca predecir. En tal sentido, se creó la variable 'DESERTO', la cual fue definida con valor 1 para todos los jóvenes de hasta 18 años de edad que declararon no haber terminado la escuela y no continuar en la misma, mientras que para el resto de los casos se le asignó el valor 0.

Además, se incorporaron 6 nuevos atributos a la base de datos en función de las observaciones realizadas en la sección anterior, ya que consideramos que pueden ser de gran utilidad para predecir la deserción escolar. Estos nuevos atributos indican el nivel educativo, la condición laboral, el sexo y la edad del jefe de hogar, así como la existencia y la condición de empleo de su

cónyuge. De la misma forma que para el atributo de deserción, estas nuevas variables fueron creadas a raíz de la información contenida en la EPH, con el objetivo de adaptar la base de datos a las necesidades del análisis.

Por otro lado, esta base de datos contiene una gran cantidad de atributos, entre los que se pueden encontrar tanto variables continuas como categóricas. Ahora bien, esto puede resultar un problema, ya que no todos los modelos de aprendizaje automático saben interpretar las variables cuyas categorías no tienen un orden inherente, a las que por lo tanto no se pueden asignar valores numéricos de manera natural. Por tal motivo, se aplicó la técnica de *One-Hot-Encoding* para convertir cada valor de las variables categóricas en nuevas variables binarias (0 o 1), buscando evitar que los modelos asuman relaciones de orden inexistentes entre estas variables.

Si bien la utilización de esta técnica puede mejorar considerablemente la precisión de las predicciones, también hay que tener en consideración que se estará generando una nueva variable binaria por cada uno de los atributos de las variables categóricas a las que se aplique (Gerón, 2022). En esta base de datos fue factible incorporarlo, ya que las variables a las que les fue aplicado contaban con una cantidad acotada de categorías, por lo que el ralentizamiento al momento de entrenar el modelo, continúo siendo aceptable. Para su viabilidad, también se destaca la basta cantidad de observaciones que superan ampliamente el numero de variables incluso después de su codificación.

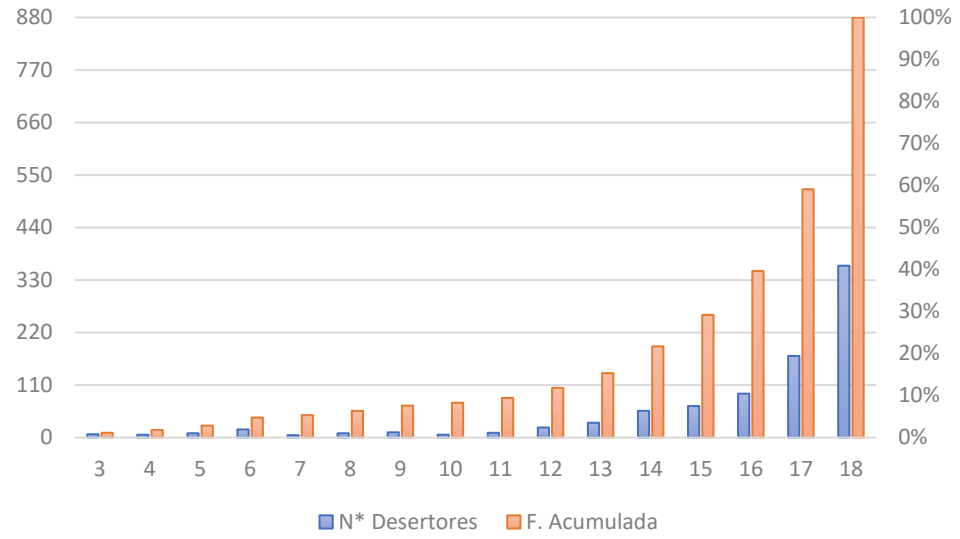
3.3. Limpieza de base

Tras unificar las bases de datos de los distintos trimestres y remover a todos los registros duplicados entre ellas, se obtuvo la muestra total de 144 mil individuos. Los registros de los mayores de edad fueron utilizados únicamente para crear las variables nuevas referidas a las características de los jefes de hogar y sus cónyuges, tras lo cual fueron removidos por no ser necesarios para nuestro estudio. De esta forma nos quedamos con una base de 38 mil jóvenes de hasta 18 años de edad, la cual cuenta con 880 desertores.

A fin de acotar la base de datos a nuestra población objetivo, fueron removidos adicionalmente todos los niños de hasta 10 años de edad, conservando así únicamente a los 18 mil jóvenes de 11 a 18 años inclusive. Si bien de esta forma se está aceptando la pérdida del 8% de los desertores, también se esta acotando el número de observaciones al rango etario donde el abandono tiene mayor frecuencia, equivalente al 4,3%[15] de nuestra muestra.

[15] Notar que este porcentaje corresponde al total de desertores en la muestra de jóvenes comprendidos entre las edades de 11 a 18 años, inclusive. Como se puede apreciar en el Gráfico 1, este porcentaje aumenta sustancialmente a medida que se incrementan las edades consideradas. Aunque se observa que la mayor proporción de jóvenes que abandonan sus estudios inconclusos se encuentra entre los 19 y 21 años, estas edades no fueron consideradas en el análisis debido a la naturaleza del modelo, que se enfoca en la predicción temprana de la deserción escolar, para el diseño de políticas públicas que puedan prevenir su ocurrencia.

Gráfico 1. Frecuencia acumulada de desertores por edad.



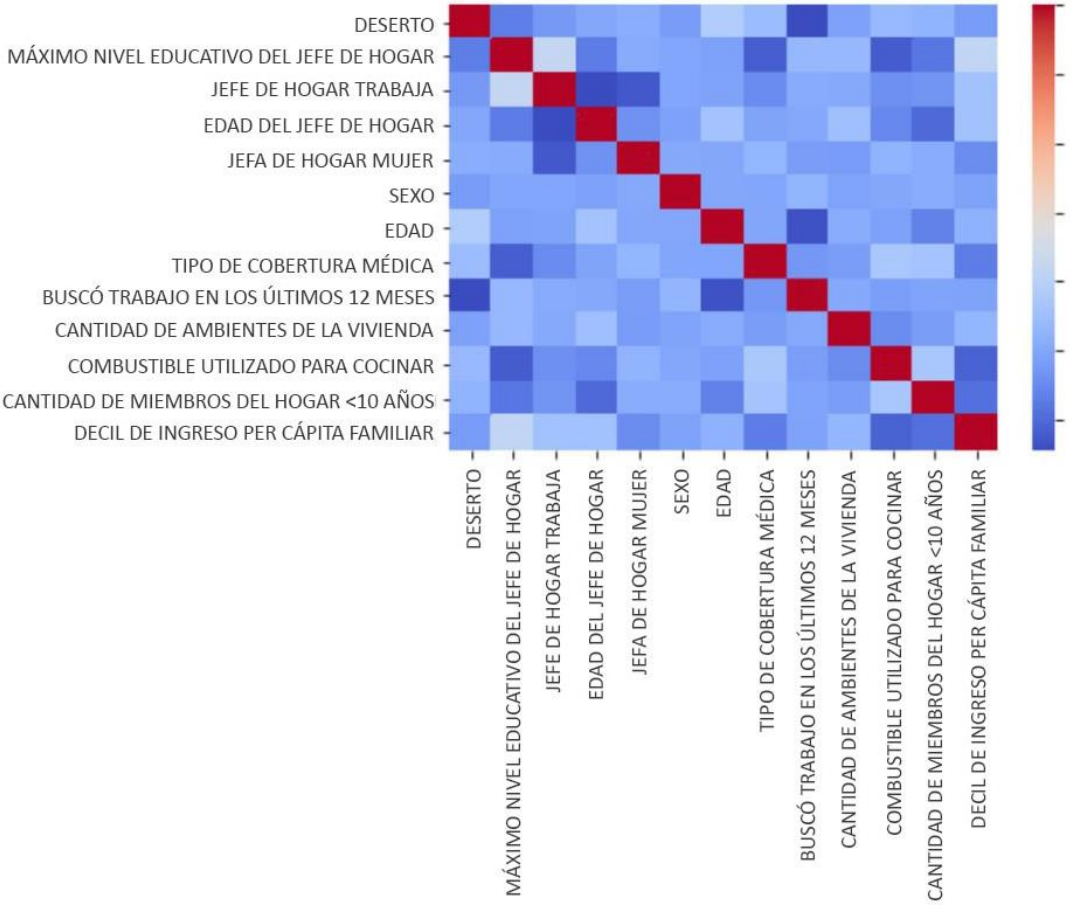
Además de focalizar la muestra, es importante realizar una buena limpieza de atributos para eliminar datos irrelevantes y mejorar la eficiencia de los modelos, así como también para mejorar la calidad y confiabilidad de sus predicciones al reducir los errores y el ruido en los datos. Para limpiar la base en primera instancia se optó por descartar todas las variables que presentan más de un 25% de observaciones vacías. Estas en su mayoría existen porque son preguntas dirigidas a subgrupos de encuestados, como por ejemplo a los desempleados o a los individuos de determinadas zonas geográficas. Para el grupo de variables que conservábamos, remplazamos los valores vacíos por el promedio de esa misma variable. Tras la realización de esta limpieza, fueron eliminados un total de 106 atributos.

En una segunda instancia de selección, se procedió a remover manualmente otras 80 variables según diversos criterios que se mencionan a continuación:

- Se removieron todas las variables referidas a la identificación de los hogares y/o los individuos, así como las fechas de nacimiento o los momentos en los que fueron encuestados, ya que estas variables no aportan valor a las predicciones.
- Se eliminaron todas las variables monetarias ya que el análisis incluye 5 trimestres diferentes en un contexto de alta inflación que dificulta la comparación entre periodos. Buena parte de estas variables también están expresadas en términos de deciles de ingresos, las cuales fueron conservadas.
- No se tomaron en consideración las variables que fueron utilizadas para generar las variables nuevas que incorporamos al modelo, de la misma forma que no se incluyeron las que se utilizaron para construir la variable objetivo (como nivel educativo, año aprobado o la subcategoría de condición de inactividad por ser estudiante).
- Se transformaron las variables categóricas no binarias mediante la aplicación de *One-Hot-Encodign* tras lo cual se quitaron las variables originales pre-codificacion.

- Por último, también fue analizada la correlación de las variables, eliminando aquellas que estaban muy estrechamente relacionadas. A modo esquemático, se presenta a continuación un *heatmap*, el cual refleja el bajo nivel de correlación presente entre algunas de las principales variables que se conservaron para el entrenamiento de los modelos.

Gráfico 2. *Heatmap* de correlaciones entre variables predictoras.



Fuente: Elaboración propia en base a los datos de las EPH 2021 y 2022.

De esta forma se logro reducir el set de datos original de 144mil registros con 242 atributos asociados, a su versión focalizada de 18 mil jóvenes de entre 11 y 18 años, conservando 63 atributos (previos a la aplicación de la técnica de *One-Hot-Encoding*), entre los cuales se incluyen las 7 variables generadas. A modo de sugerencia para el perfeccionamiento futuro, queda pendiente analizar el tratamiento de los valores atípicos, considerando que para ello se requiere un profundo conocimiento que permita tomar decisiones a detalle según cada una de las variables incluidas. Para más información, ver el apéndice Datos, el cual contiene la descripción de las variables generadas, junto al diseño de registro de la EPH, identificando las variables que fueron conservadas.

3.4. Tratamiento de clases desbalanceadas

Es importante destacar la limitación que existe en este estudio debido a que la variable objetivo presenta clases sumamente desbalanceadas en donde el número de desertores equivale a tan solo el 4,3% de la muestra de jóvenes, especialmente considerando que la mayoría de los algoritmos de aprendizaje automático no pueden manejar correctamente esta situación ya que asumen distribuciones equilibradas.

La presencia de bases de datos con clases desbalanceadas supone un problema para los modelos predictivos ya que estos tienden a centrar su atención sobre los casos de la clase mayoritaria obteniéndose resultados que aparentemente son buenos, pero que en definitiva pueden estar asignando todas las predicciones a la clase de mayor peso (Miravet. 2021). Puesto en otras palabras, si solo el 4,3% de los jóvenes abandona la escuela sin finalizar sus estudios obligatorios, un modelo que prediga que nadie va a abandonar tendría una precisión muy alta del 95,7%, aun cuando no este acertando en la predicción de ni un solo desertor. Esta situación se acentúa mientras mayor sea el desbalance de la clase, generando estimaciones que virtualmente tienen mayor precisión. Es por esto que el análisis de datos desbalanceados merece ser tratado distinto tanto a la hora de entrenar como a la hora de evaluar, en comparación con los datos de clases balanceadas.

Existen dos formas principales de abordar este problema. La primera es modificar los modelos para asignar una ponderación mayor a las observaciones en la categoría minoritaria (Desertores). La segunda forma es utilizar métodos de remuestreo aleatorio que buscan equilibrar la proporción de observaciones de las categorías de la variable de respuesta, como el *oversampling* que aumenta aleatoriamente el número de observaciones minoritarias para igualar ambas categorías, y el *undersampling* que subrepresenta la clase mayoritaria con el mismo propósito (Somasundaram y Reddy. 2016).

Un inconveniente de la técnica de remuestreo es que no aporta información nueva, sino que simplemente repite los datos existentes, lo cual puede llevar al sobreajuste del modelo afectando negativamente el desempeño en datos desconocidos[16]. Si bien la técnica de submuestreo tiene la contra de puede perder información importante de la clase mayoritaria, se optó por la misma para balancear las clases aprovechando que la muestra es lo suficientemente grande como para permitirlo. De esta forma, fueron seleccionados aleatoriamente 807 individuos no desertores para conformar la base de datos final con 1614 registros equitativamente distribuidos entre las clases.

Después de realizar los procesos mencionados, se logró unificar las cinco bases de datos previas que contenían 50 mil observaciones y 242 atributos cada una. Como resultado, se obtuvo finalmente la base de datos unificada, correctamente focalizada y equitativamente balanceada,

[16] Por esta misma razón, es pertinente aclarar que en el presente estudio no se llevó a cabo la ponderación de los datos, ya que esto no contribuiría con información adicional, sino que redundaría en la replicación de los datos utilizados.

compuesta por 1614 registros con 127 atributos (63 antes de aplicar *one-hot-encoding*). De estos atributos, 7 fueron generados a partir del procesamiento previo de la información. Además, se logró eliminar todos los valores nulos y reducir significativamente los niveles de correlación entre las variables. Con esta base final, se espera mejorar la precisión y eficacia de los modelos de aprendizaje automático que trabajaremos en el siguiente capítulo.

4. Metodología

Machine Learning es una subdisciplina de la Inteligencia Artificial que se enfoca en el desarrollo de algoritmos y modelos de aprendizaje automático que pueden aprender a partir de datos y realizar tareas complejas sin ser explícitamente programados. En *Machine Learning*, los algoritmos son entrenados con grandes volúmenes de datos y, a partir de estos, aprenden a realizar tareas específicas, como la clasificación, la regresión o la clusterización. Este tipo de modelos se basan en la identificación de patrones y relaciones en los datos que permitirán predecir con cierto grado de certeza si un estudiante abandonará sus estudios.

De acuerdo con Mitchell T. (2006), esta área de la inteligencia artificial se centra en el desarrollo de sistemas informáticos que puedan mejorar su desempeño mediante la experiencia. Hay dos tipos principales de aprendizaje estadístico: supervisado y no supervisado. En el aprendizaje supervisado, se dispone de una respuesta *Y* para las variables predictoras *X*, y el objetivo es encontrar un modelo que pueda predecir la respuesta para nuevas observaciones o entender la relación entre las variables predictoras y la respuesta. En el aprendizaje no supervisado, se dispone de valores para las variables, pero no se tiene una respuesta para predecir, y el objetivo es encontrar patrones en los datos, como agrupaciones de observaciones similares.

Para el ejercicio planteado en este trabajo, se utilizaron diversos modelos de aprendizaje supervisado ya que, si bien el data set original no contiene la variable de deserción, se pudo elaborar la misma e incorporarla al modelo como la 'Y' objetivo. A su vez, los ejercicios de aprendizaje supervisado pueden ser divididos entre problemas de clasificación o problemas de regresión, dependiendo de si la variable respuesta es categórica o continua. Para este trabajo se utilizaron herramientas de clasificación binaria, ya que la variable dependiente tomara el valor 1 en caso de predecir la deserción de un alumno y el valor 0 en caso contrario.

Mediante la evaluación de diversos modelos de predicción como Regresión Logística, KNN, Árboles de Decisión, *Support Vector Machine* o *Boosting* entre otros, se buscó identificar patrones que permitan elaborar herramientas para prevenir el abandono escolar. La pretensión de este trabajo es conectar este ejercicio de predicción con la viabilidad de su utilización para mejorar o complementar políticas públicas, buscando aumentar la permanencia de los jóvenes en el sistema educativo y así lograr que cada vez más estudiantes finalicen sus estudios obligatorios en tiempo.

4.1. Conjuntos de entrenamiento, validación y testeo

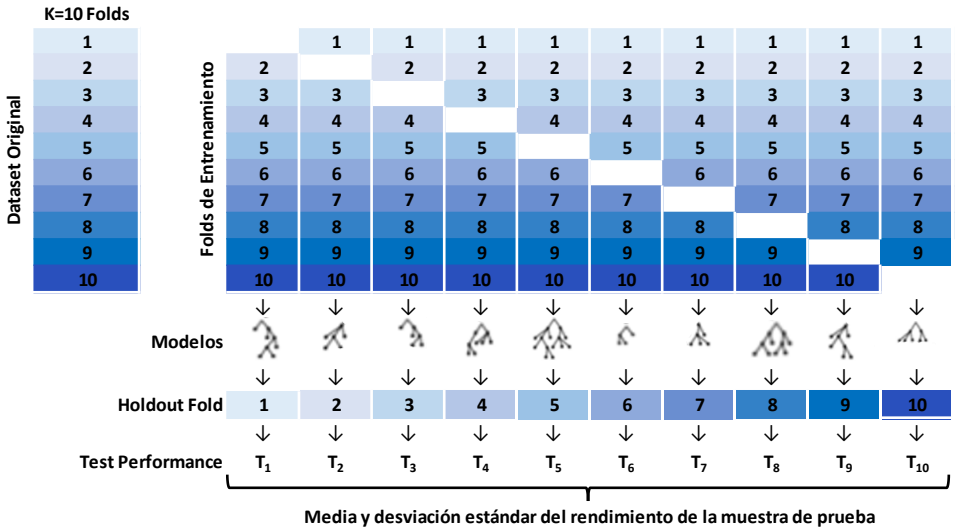
En primera instancia, antes de avanzar con la configuración de los modelos, se particionó la muestra en diferentes conjuntos de entrenamiento, validación y testeo. Estas particiones son herramientas fundamentales en el proceso de desarrollo para garantizar modelos precisos y generalizables. Resumidamente, la funcionalidad de cada uno de estos conjuntos es:

- Entrenamiento: es el conjunto de datos utilizado para entrenar los modelos, buscando obtener la configuración optima de los parámetros y así mejorar su precisión. El modelo utiliza los datos de entrenamiento para aprender las relaciones entre las características (variables independientes) y la variable objetivo (variable dependiente).
- Validación: es el conjunto de datos utilizado para ajustar probar diferentes valores de hiperparámetros y seleccionar los que consiguen el mejor rendimiento del modelo. Los hiperparámetros son parámetros que se establecen antes del entrenamiento del modelo y afectan su capacidad para generalizar a nuevos datos.
- Testeo: es el conjunto de datos utilizado para evaluar el rendimiento final del modelo en datos nuevos y no vistos durante el entrenamiento ni la validación.

La importancia de los conjuntos de entrenamiento, validación y testeo radica en que el uso inadecuado de los datos de entrenamiento puede llevar a generar modelos que se sobreajustan (*overfitting*) o subajustan (*underfitting*) a los datos con los que son entrenados, lo que resulta en un rendimiento deficiente al querer probarlos con datos desconocidos. La justificación subyacente a esta práctica radica en el hecho de que si empleamos un conjunto de datos para la selección de un modelo, el modelo elegido seguramente arrojará buenos resultados en dicho conjunto de datos utilizando la misma métrica de evaluación. Sin embargo, lo que es realmente relevante es la métrica de evaluación en datos desconocidos, ya que estos representan la prueba de la capacidad de generalización del modelo (James *et al.* 2013).

El objetivo es evaluar los modelos varias veces utilizando diferentes conjuntos de datos para seleccionar la configuración óptima basada en el promedio de los puntajes de cada evaluación. Dado que el conjunto de datos es limitado, se optó por la validación cruzada *K-folds*, que permite realizar remuestreos para simular el uso de diferentes conjuntos de datos y mejorar el modelo a través de la combinación de conjuntos de entrenamiento y validación.

Gráfico 3. Ilustración esquemática de *K-Fold Cross-Validation* con K=10



Fuente: Elaboración propia.

Aprovechando la secuencialidad temporal de las encuestas, se utilizaron las EPH del tercer y cuarto trimestre de 2021, junto con las del primer y segundo trimestre de 2022 para generar la base de datos con la que se entrenó y se realizó la validación cruzada con K=10 *folds*. Por el otro lado, se reservaron los datos de la EPH del tercer trimestre de 2022 para realizar el testeo final. Este enfoque ayuda a prevenir la posibilidad de realizar *data leakage*, ya que se están utilizando conjuntos de datos diferentes.

Gráfico 4. Ilustración esquemática de los conjuntos de entrenamiento, validación y testeo.

Fuente: Elaboración propia.

Para el conjunto de testeo, se aplicaron los mismos procedimientos de limpieza e ingeniería de atributos, a excepción del balanceo, eliminando todos los registros de los jóvenes que habían participado de encuestas anteriores. De esta forma se obtuvo la base de testeo con 3706 registros de los cuales 139 son desertores. Se utilizó esta base de datos que no había sido utilizada previamente para realizar la evaluación final del rendimiento con el fin de simular el desempeño del modelo en el caso de que fuera implementado. Los resultados obtenidos fueron similares a los obtenidos en la instancia de validación cruzada, lo que confirma que el modelo no presenta sobreajuste de datos y puede ser generalizado.

4.2. Modelos

En el presente trabajo, se procedió a entrenar 10 modelos de aprendizaje supervisado utilizando el conjunto de entrenamiento. Estos modelos incluyen: Regresión Lineal, Regresión Logística, Regresión Lineal Ridge, Regresión Lineal LASSO, KNN, Árbol de decisión (CART), *Bagging, Random Forest, Boosting* y SVM. Para llevar a cabo esta tarea, se utilizó el paquete SKLearn para entrenar, validar y evaluar cada uno de estos modelos, con el objetivo de identificar aquel que presente la mejor performance en los resultados y que por tanto tenga el mejor poder de predicción.

A continuación, se procederá a explicar detalladamente el funcionamiento de los 10 modelos identificados como los más relevantes para el caso de estudio y que por tanto serán comparados para este ejercicio de predicción.

4.2.1. Regresión Logística

La regresión logística es un modelo de aprendizaje supervisado utilizado para la clasificación binaria, es decir, la predicción de la probabilidad de que una observación pertenezca a una de dos categorías mutuamente excluyentes. La técnica se basa en la regresión lineal y utiliza la función logística para transformar la variable de respuesta a una escala de probabilidades entre 0 y 1.

El modelo *logit* es conocido como la técnica más tradicional y es utilizado para modelar la probabilidad de que una variable de resultado dicotómica (en nuestro caso: Desertó/No Desertó), y una variable predictiva continua que está relacionada con la probabilidad o "*odds*" de la variable de resultado. También se puede utilizar con predictores categóricos y con múltiples predictores. Si la probabilidad estimada es mayor al 50%, el modelo predecirá que la variable pertenece a la clase positiva (etiquetada como "1"), caso contrario el modelo predice que pertenece a la clase negativa (etiquetada como "0"). La probabilidad estimada se calcula a partir de la siguiente ecuación:

$$\hat{p} = \sigma(\mathrm{X}^T\theta)$$

$$\hat{y}\begin{cases} 0\ if\ \hat{p} < 0{,}5 \\ 1\ if\ \hat{p} \geq 0{,}5 \end{cases}$$

En donde $\hat{p}$ es la probabilidad estimada, σ es la función sigmoide, X es la matriz de características, θ es el vector de parámetros del modelo y $\hat{y}$ corresponde a la predicción. (Muñoz Jaramillo 2021). La función Logit toma valores continuos que van desde menos infinito hasta más infinito, lo que permite a los modelos de regresión lineal predecir la probabilidad de que ocurra un evento en función de las variables predictoras.

Luego de obtener cada predicción de probabilidades individuales, es decir para cada joven en particular, se crea un umbral de decisión σ para determinar si una observación pertenece a una

clase o a la otra. Es así que, si para cada joven en particular la probabilidad predicha es mayor a σ, entonces se clasificará como Desertor. En caso contrario, si es menor que σ, se clasificará como No Desertor.

4.2.2. Ridge y Lasso

Ridge regression es un método de análisis de regresión que utiliza la técnica de regularización para evitar el sobreajuste y reducir la varianza del modelo. En Ridge, se agrega un término de penalización a la función objetivo que minimiza el error cuadrático medio. Este término de penalización está controlado por un hiperparámetro λ, que controla el grado de regularización aplicado al modelo. Las estimaciones de los coeficientes de regresión ridge, $\hat{\beta}^R$, son los valores que minimizan

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2$$

Al igual que en los mínimos cuadrados, la regresión Ridge busca estimaciones de coeficientes que se ajusten bien a los datos, al hacer que la suma de los residuos cuadrados (RSS) sea pequeña. Sin embargo, el segundo término, $\lambda\sum_j\beta_j^2$, llamado *shrinkage penalty* (penalización de encogimiento), se hace pequeño cuando $\beta_1,\dots,\beta_p$ están cerca de cero, y tiene el efecto de encoger las estimaciones de β_j hacia cero. El parámetro de ajuste λ sirve para controlar el impacto relativo de estos dos términos en las estimaciones de los coeficientes de regresión. Cuando λ = 0, el término de penalización no tiene efecto y la regresión Ridge producirá las estimaciones de mínimos cuadrados. Sin embargo, a medida que $\lambda \rightarrow \infty$, el impacto de la penalización de encogimiento crece y las estimaciones de los coeficientes de regresión Ridge se acercarán a cero (James *et al.* 2013).

Esto implica que las variables menos importantes tendrán coeficientes más cercanos a cero, pero no exactamente cero. La ventaja de Ridge es que permite que todas las variables predictoras contribuyan al modelo, incluso las que pueden tener un impacto relativamente menor en la predicción. Esto puede ser beneficioso cuando se sospecha que todas las variables son relevantes en cierta medida y se desea evitar la exclusión prematura de variables útiles.

Por otro lado, Lasso (*Least Absolute Shrinkage and Selection Operator*), es otro método de regresión que también utiliza la regularización para evitar el sobreajuste. Al igual que Ridge, Lasso agrega un término de penalización a la función objetivo, pero en este caso, se utiliza la norma L1 para la penalización. Los coeficientes de Lasso, $\hat{\beta}_\lambda^L$, minimizan la cantidad

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

Lasso y regresión Ridge tienen formulaciones similares. La única diferencia es que el término β_j^2 en la penalización de regresión Ridge se ha reemplazado por $|\beta_j|$ en la penalización de Lasso. En terminología estadística, Lasso utiliza una penalización L1 en lugar de una penalización L2. La norma L1 de un vector de coeficientes β se define como $\|\beta\|_1 = \sum |\beta_j|$

La principal diferencia entre Lasso y *Ridge regression* radica en la forma en que se reduce el tamaño de los coeficientes. Mientras que Ridge reduce los coeficientes hacia cero pero no exactamente a cero, Lasso tiene la capacidad de forzar algunos coeficientes exactamente a cero cuando el parámetro de ajuste λ es suficientemente grande. Esto significa que Lasso realiza selección de variables al eliminar completamente las variables menos importantes del modelo. Como resultado, los modelos generados por Lasso suelen ser mucho más fáciles de interpretar que los producidos por la regresión Ridge. Decimos que Lasso produce modelos dispersos, es decir, modelos que involucran solo un subconjunto de las variables (James *et al.* 2013).

La capacidad de Lasso para realizar selección de variables es especialmente útil cuando se trabaja con conjuntos de datos con un gran número de variables predictoras, donde es deseable identificar las variables más relevantes y descartar las irrelevantes para la predicción del resultado. La elección entre Ridge y Lasso depende de los objetivos específicos del análisis y las características de los datos, considerando si se prefiere mantener todas las variables en el modelo (Ridge) o seleccionar solo las variables más relevantes (Lasso).

4.2.3. K-Nearest Neighbors

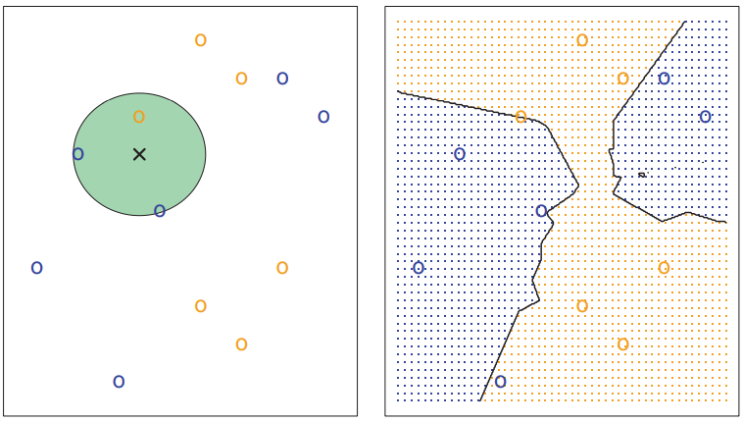
El modelo KNN *(K-Nearest Neighbors* o K-Vecinos más cercanos) es un algoritmo de aprendizaje supervisado utilizado tanto para la clasificación como para regresión. En el contexto de la clasificación binaria, el modelo KNN se utiliza para predecir la clase de una observación desconocida en función de las clases de las observaciones cercanas en un conjunto de datos de entrenamiento. Para el cálculo de la distancia entre las observaciones se suele utilizar la distancia Euclidiana, que es la separación entre dos puntos en un espacio n-dimensional.

Para una clasificación binaria, se determina la clase más común entre los K vecinos más cercanos, asignando la clasificación positiva cuando la mayoría de los vecinos más cercanos pertenecen a la clase positiva; de lo contrario, se clasifica como negativa.

$$\Pr(Y = \mathrm{j}|X = x_0) = \frac{1}{K}\sum_{\mathrm{i}\in N_o} \mathrm{I}(y_i = \mathrm{j})$$

Dada una observación a clasificar (X_0) y un valor de K, se identifican los K vecinos (N_0) más cercanos de X_0, se estima la probabilidad condicional de pertenecer a la clase j como la fracción de observaciones de N_0 que pertenecen a j y se asigna como clase de X_0 a aquella clase para la cual se obtuvo la mayor probabilidad condicional (James *et al.* 2013).

Gráfico 5. Ilustración del modelo KNN para clasificación con K=3.



Fuente: James et al. (2013). Cap 2. Pg 40. Figure 2.14.

Una ventaja del modelo KNN es que es no paramétrico, lo que significa que no hace suposiciones sobre la distribución de los datos. También es fácil de implementar y puede ser útil cuando se tienen datos no lineales. Sin embargo, puede ser computacionalmente costoso para conjuntos de datos grandes y puede no funcionar bien si los datos están muy sesgados o si las dimensiones son demasiado altas.

Si bien es un modelo relativamente simple, podría brindarnos mucha información acerca de las características de los jóvenes que deciden abandonar sus estudios, y cuáles de ellas son relevantes para poder realizar una predicción correcta. Aunque no esperamos obtener la mejor predicción de este modelo, en un análisis futuro se podría explotar la facilidad con la que este se extiende a múltiples categorías para poder tener en cuenta distintas características cualitativas o cuantitativas de los jóvenes (como lo pueden ser su género, edad, participación en el mercado laboral, tipo de escuela a la que asiste, condiciones de su vivienda, nivel educativo de sus padres o el nivel de ingresos monetarios de la familia, entre otros) y conseguir un modelo de propensión al abandono.

4.2.4. Arboles de decisión (CART)

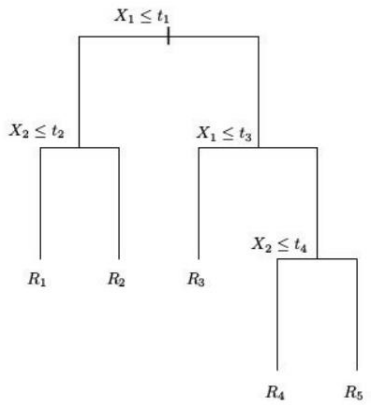
Los árboles de decisión son uno de los métodos más simples y de fácil interpretación utilizados para realizar predicciones tanto en problemas de clasificación, como de regresión. Fueron introducidos por primera vez en 1984 por Breiman *et al*, brindando la posibilidad de capturar y modelar automáticamente relaciones no lineales entre los predictores.

Para la explicación de este algoritmo utilizaremos la metodología de CART (*Classification and Regresion Trees*), introducida por Breiman et al. (1984), ya que es la base de posteriores algoritmos más sofisticados. CART se construye a partir de una serie de decisiones basadas en las variables predictoras, donde cada nodo representa una pregunta, y las ramas del nodo representan las posibles respuestas. La construcción del árbol continúa hasta que se llega a un nodo final o hoja, que proporciona una predicción para la clase de interés. Los árboles implementan una estrategia de “divide y conquistarás” en un formato no paramétrico, mediante

la construcción de una estructura de datos jerárquica, que le permite a los datos indicar la distribución, sin requerir estimar parámetros. Esto permite que los árboles pueden detectar relaciones complejas entre las variables explicativas o independientes "X" y las variables dependientes "Y". Este método funciona muy bien para estructuras no-lineales y también resulta ser uno de los mejores modelos a la hora de predecir con una base de datos desbalanceada.

El procedimiento es una partición recursiva binaria que busca encontrar el mejor ajuste global, es decir, la mejor variable predictora en cada nodo y el punto de partición óptimo se definen como el corte que proporciona la mayor ganancia de información o la mayor reducción en la entropía en la división de los datos en dos grupos. Una vez que se ha construido el árbol con la división del espacio de los X_1, X_2, ... , X_p predictores en j regiones distintas y no superpuestas, R_1, R_2, ... , R_j, se puede utilizar para hacer predicciones sobre nuevos datos de prueba. Siguiendo las ramas del árbol hasta llegar a una hoja, a cada observación que cae en la región R_j se la clasifica en la categoría k que resulta más frecuente en esa región. Cada variable y cada punto de partición se eligen de modo que se maximice la homogeneidad de la variable respuesta al interior de cada nodo. El inconveniente es que resulta ser un método poco robusto a los datos, ya que una pequeña alteración en estos puede dar como resultado un nuevo árbol completamente diferente. Es un método que presenta una varianza muy alta, debido a las características jerárquicas que poseen, donde un ligero cambio en los datos puede alterar todos los sucesivos cortes.

Gráfico 6. Ilustración de un árbol de partición binaria recursiva.



Fuente: James et al. (2013). Cap 8. Pg 308. Figure 8.3.

Por lo general, los árboles de decisión no son competitivos contra otros algoritmos de aprendizaje supervisado en términos de precisión de predicción. Sin embargo, de ellos se desprenden algoritmos mucho más potentes como *Random Forest, Bagging y Boosting*, los cuales trabajaremos en los apartados siguientes. Cada uno de estos enfoques implica la producción de múltiples árboles que luego se combinan para producir una sola predicción de consenso. La combinación de una gran cantidad de árboles a menudo puede resultar en mejoras drásticas en la precisión, a expensas de hacer más difícil la interpretación del modelo resultante. (James *et al.* 2013).

4.2.5. Bagging

Los algoritmos de árboles de decisión tienen bajo sesgo, aunque sufren de alta varianza, por lo que si dividiéramos aleatoriamente los datos de entrenamiento en dos partes y entrenáramos un árbol para cada mitad, los resultados obtenidos posiblemente serían muy diferentes. Para solucionar este problema, Breiman (1996) introduce los métodos de ensamble a los árboles de decisión mediante el modelo *bagging* (nombre derivado de *bootstrap aggregation*).

Este consiste en tomar muchas muestras aleatorias con reemplazo del conjunto de datos de entrenamiento. Estimar un árbol nuevo con cada una de estas muestras e ir guardando todas las predicciones individuales. Cada modelo en el conjunto produce una predicción, a partir de las cuales obtenemos la predicción final promediando las predicciones individuales, asignándole igual peso de voto a cada una.

El *Bagging* reduce efectivamente la varianza del modelo, aumenta la precisión de las predicciones y logra evitar el sobreajuste (*overfitting*), ya que los modelos en el conjunto tienen menos probabilidad de memorizar los datos de entrenamiento debido a que cada uno tiene una muestra aleatoria diferente. La intuición detrás se fundamenta en la noción de que la varianza del promedio es menor que la varianza de un único árbol: Dado n muestras de observaciones independientes Z_1, Z_2…, Z_n, cada una con varianza σ^2, la varianza de la media de las observaciones es σ^2/n. Es por esto que, la predicción conjunta de todos los árboles, es más robusta que la individual.

Si bien el método de *bagging* introduce la aleatoriedad en la muestra generando que todos los árboles sean distintos, en la práctica estos pueden resultar sumamente similares. El ensamble de *bagging* suele generar arboles con ramas o nodos finales diferenciados, pero que con frecuencia tienen estructuras muy similares en la parte alta de los árboles, debido a la presencia de predictores de mayor relevancia que terminan siendo incluidos en el primer corte en la mayoría de los árboles.

4.2.6. Random Forest

También introducido por Breiman (2001), el algoritmo de *Random Forest* (o bosques aleatorios) es una variante del método *bagging* diseñado para trabajar específicamente con árboles de decisión. *Random Forest* introduce una aleatoriedad adicional en la construcción de cada uno de los árboles que finalmente constituirán el bosque. En lugar de buscar el mejor atributo al dividir un nodo, de todas las ρ variables predictoras, se seleccionan al azar $m < \rho$ predictores que van a ser los candidatos para el corte. Esta aleatoriedad adicional se introduce antes de hacer cada uno de los cortes, generando que la selección del mejor candidato para la división sea extraído de este subconjunto reducido de atributos.

Si bien suele ser necesaria una mayor cantidad de árboles que en *bagging, Random Forest* es computacionalmente más eficientes ya que al evaluar sólo unos pocos predictores en cada

corte, la construcción de cada árbol logra ser más rápida. Este algoritmo genera arboles más diferenciados, rompiendo con la correlación descripta en *bagging*, logrando así reducir la varianza a costa de aceptar un sesgo ligeramente superior. Otro aspecto a destacar de este modelo es que logra un muy buen poder de predicción, pero evitando caer en el sobreajuste (Gerón A. 2019).

4.2.7. Boosting

De la misma forma que *Bagging, Boosting* otro modelo de aprendizaje automático que utiliza los métodos de ensamble para reducir la varianza y así mejorar el rendimiento predictivo mediante la combinación de múltiples modelos simples. Si bien ambos buscan construir un modelo final que supere el rendimiento de cualquier modelo individual, la principal diferencia radica en como lo hacen. Mientras que *Bagging* busca entrenar modelos independientes, *Boosting* entrena modelos secuenciales que se ajustan a los datos de entrenamiento de manera iterativa, con cada modelo subsecuente corrigiendo los errores del modelo anterior. *Boosting* realiza la construcción de un modelo fuerte a partir de un conjunto de modelos débiles (*weak learners*). El modelo fuerte se construye a través de una combinación ponderada de los modelos débiles, con las ponderaciones asignadas en función de la precisión del modelo débil en los datos de entrenamiento.

Boosting aprende lentamente al ir construyendo pequeños árboles secuenciales cuyo objetivo es predecir los residuos del árbol anterior, en lugar del resultado "Y". Es decir, partiendo del árbol inicial, se generan pequeños nuevos árboles a partir de los residuos del anterior. Luego agregamos este nuevo árbol de decisión en la función ajustada para actualizar los residuos. Al ajustar pequeños árboles a los residuos, lentamente vamos mejorando el árbol inicial en las áreas donde no se desempeñaba adecuadamente. A diferencia de lo que sucede con *Bagging*, la generación de cada árbol depende en gran medida de los árboles generados anteriormente. El parámetro λ es el encargado de ralentizar aún más el proceso generando arboles chicos y diferenciados, para evitar que se generen arboles muy extensos que provoquen *overfitting*. Los modelos de *boosting* han demostrado una gran efectividad tanto para problemas de clasificación como regresión al lograr mejorar la precisión reduciendo tanto la varianza como el sesgo (*bias*) del modelo. (James *et al.* 2013).

Existen diferentes algoritmos de boosting actualmente muy difundidos como: AdaBoost (Adaptive Boosting), Gradient Boosting, XGBoost, LightGBM, entre otros. Cada uno de estos algoritmos utiliza una estrategia diferente para construir el modelo final. En este trabajo se utilizó el *Gradient Boosting*, que es una técnica iterativa que ajusta un conjunto de modelos de árbol de decisión para minimizar una función de pérdida en cada iteración. En cada iteración, el modelo se ajusta para corregir los errores del modelo anterior.

4.2.8. Support Vector Machines

El modelo de SVM (*Support Vector Machines* o Máquina de Vector Soporte), es un método de aprendizaje supervisado originalmente desarrollado para la clasificación binaria, aunque actualmente existen extensiones dentro de esta metodología para clasificación con más categorías, regresión y detección de datos atípicos. El modelo SVM encuentra el límite de clasificación no lineal mediante la construcción de límites lineales de decisión, específicamente, hiperplanos separadores, en un espacio transformado del conjunto de los predictores originales (James *et al.* 2013).

El modelo SVM funciona de la siguiente manera: Primero, se representan las observaciones en un espacio de características, que es una representación matemática de los datos. Si los datos originales son de alta dimensión, se puede utilizar una técnica de reducción de dimensionalidad para transformar los datos a un espacio de características de menor dimensión. A continuación, de gran infinidad de hiperplanos posibles, se encuentra el que mejor separa las dos clases. Este hiperplano es el límite de decisión que maximiza la distancia entre las observaciones más cercanas de cada clase, conocidas como vectores de soporte. Una vez que se encuentra el hiperplano, se utiliza para clasificar nuevas observaciones. Si la nueva observación está en un lado del hiperplano, se clasifica como perteneciente a una clase; de lo contrario, se clasifica como perteneciente a la otra clase

El modelo SVM utiliza una función de pérdida para minimizar el error de clasificación y encontrar el hiperplano que mejor separa las clases. No obstante, este modelo suele ser poco robusto y sensible a los datos, generando con facilidad cambios en el hiperplano de óptima separación. La función de pérdida también incluye un término de regularización que controla la complejidad del modelo, siendo crucial para evitar generar un sobreajuste y, por consiguiente, una mala predicción por fuera de la muestra. Otro problema recurrente es que las observaciones de la muestra no siempre son perfectamente separables de modo lineal. Los modelos SVM pueden ser lineales o no lineales, lo que significa que pueden encontrar hiperplanos lineales o curvos que separan las clases. Una ventaja del modelo SVM es que puede funcionar bien en conjuntos de datos de alta dimensionalidad y es resistente al sobreajuste. Sin embargo, puede ser computacionalmente costoso para grandes conjuntos de datos y puede ser sensible a la selección de parámetros, como la función de pérdida y los parámetros de regularización.

4.3. Evaluación de Modelos

Contando con diversidad de modelos, es necesario definir una métrica relevante para poder compararlos y determinar cuál de ellos nos brinda las mejores predicciones al problema planteado dada la base de datos que disponemos. Para entender que tan cerca están las respuestas verdaderas de las predichas, debemos definir la métrica de error a utilizar, cuya elección debe basarse en la comprensión del problema y los objetivos del modelo. Si bien el

entrenamiento y comparación de los modelos estará guiado por una única métrica, en los resultados se expondrán la precisión (ACC), el área bajo la curva ROC (AUC), el error cuadrático medio (ECM), verdaderos positivos (TP), falsos negativos (FN), verdaderos negativos (TN) y falsos positivos (FP) ya que esto permitirá tener una evaluación más completa de los resultados.

Una de las métricas más ampliamente utilizadas para comprender el rendimiento de los clasificadores es el *Accuracy* o Precision, y corresponde a la proporción de observaciones clasificadas correctamente por el modelo, la cual se calcula como la suma de los verdaderos positivos y los verdaderos negativos dividida por el total de observaciones. No obstante, esta métrica puede volverse engañosa y dejar de ser adecuada para evaluar modelos derivados de conjuntos de datos desequilibrados como lo es el con la base de alumnos desertores. En la práctica, considerando un caso hipotético en el que solo el 4% de los alumnos abandonen prematuramente sus estudios, un modelo que predice que la totalidad de los alumnos va a finalizar sus estudios, tendría una precisión del 96% aun cuando este no haya podido identificar ni a un solo individuo de nuestra población objetivo.

Es importante tener en cuenta que las métricas de evaluación adecuadas dependen del problema de predicción específico que se esté abordando. Dado que la medida de precisión trata a cada clase como igualmente importante, puede no ser adecuada para analizar conjuntos de datos desequilibrados, donde la clase rara es considerada la importante (Tan *et al.* 2006). A continuación, se muestra la matriz de confusión que resume el número de instancias predichas correcta o incorrectamente por un modelo de clasificación.

		Clase Predicha		Instancias Totales
		+	-	
Clase Real	+	TP	FN	P
	-	FP	TN	N

True Positive (TP): es el número de observaciones clasificadas correctamente como positivas.

False Positive (FP): es el número de observaciones clasificadas incorrectamente como positivas.

True Negative (TN): es el número de observaciones clasificadas correctamente como negativas.

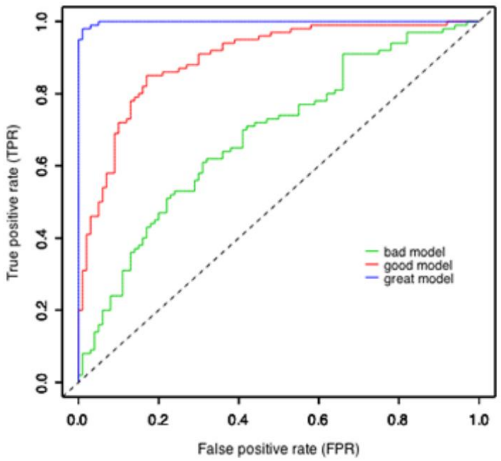
False Negative (FN): es el número de observaciones clasificadas incorrectamente como negativas.

True Positive (TP), *False Positive* (FP), *True Negative* (TN) y *False Negative* (FN) son métricas útiles para evaluar la calidad de la clasificación binaria cuando una clase del problema es mucho más importante que la otra. Al estar trabajando sobre políticas públicas focalizadas, resulta prioritario buscar alcanzar a la totalidad de los jóvenes que sufren de mayor vulnerabilidad, por lo que para la elección del modelo, se tendrá en consideración aquel que maximice el *True Positive Rate* (TPR):

$$TPR = \frac{TP}{TP + FN}$$

La elección final del mejor modelo de predicción estará definida por la maximización del área bajo la curva ROC (*Receiver Operating Characteristic Curve*). Esta métrica es la más acertada para el ejercicio propuesto ya que es buena para evaluar modelos de clasificación binaria cuando las clases no están balanceadas. La misma tiene en cuenta tanto la sensibilidad $(Recall = \frac{TP}{TP+FN})$ como la especificidad $(specificity = \frac{TN}{TN+FP})$. La curva ROC permite obtener un enfoque grafico para mostrar el *trade off* entre el *True Positive Rate* (TPR) y el *False Positive Rate* (FPR) de un clasificador a lo largo de diferentes umbrales de clasificación. La tasa TPR se representa a lo largo del eje Y, mientas que la tasa FPR (1- especificidad) se muestra en el eje X. Cada punto a lo largo de la curva corresponde a uno de los modelos inducidos por el clasificador.

Gráfico 7. Ilustración de la curva ROC según el desempeño de los modelos.



Fuente: James et al. (2013). Cap 4. Pg 148. Figure 4.8.

El AUC-ROC (*Area under the ROC Curve*) es una métrica que mide la capacidad del modelo para distinguir entre las dos clases. Un buen modelo de clasificación dará una curva que se ubicara en el grafico tan cerca como sea posible del límite superior izquierdo ubicándose en [$TPR=1$ $FPR=0$] , mientras que un modelo aleatorio se situara sobre la diagonal puntea que conecta los puntos [$TPR=0, FPR=0$] y [$TPR=1, FPR=1$]. Por lo tanto, cuanto mayor sea el área bajo la curva AUC-ROC, mejor se estará distinguiendo entre las clases, lo que permitirá diseñar políticas focalizadas que alcancen correctamente al mayor número posible de desertores, pero que a su vez logren excluir correctamente a quienes no son nuestra población objetivo, a fin de optimizar la utilización de los recursos públicos.

4.4. Optimización de hiperparámetros

Los hiperparámetros son parámetros de un modelo de aprendizaje automático que no se aprenden directamente de los datos, sino que se establecen antes de entrenar el modelo. Son valores que controlan el comportamiento del modelo y afectan drásticamente su capacidad para aprender patrones en los datos. Algunos ejemplos comunes de hiperparámetros son la tasa de

aprendizaje, el número de capas en una red neuronal, la profundidad del árbol de decisión, la cantidad de vecinos en K-NN, entre otros.

La elección de los hiperparámetros adecuados es una parte crucial del proceso de construcción de modelos de aprendizaje automático y puede mejorar significativamente su rendimiento. Si los hiperparámetros no están bien ajustados, el modelo puede sobreajustarse o subajustarse, lo que puede resultar en un bajo rendimiento en la predicción. Por ello, es importante explorar diferentes combinaciones de los mismos y evaluar su rendimiento en un conjunto de validación antes de seleccionar los mejores valores. (Bergstra & Bengio, 2012).

La optimización es el proceso de seleccionar los mejores valores de los hiperparámetros de un modelo de aprendizaje automático para maximizar su rendimiento en un conjunto de datos de prueba. El proceso de optimización de hiperparámetros debe ser iterativo y se debe repetir varias veces para encontrar la mejor combinación de hiperparámetros. Es importante tener en cuenta que la mejor configuración puede variar dependiendo del conjunto de datos y esta no necesariamente garantiza un buen rendimiento del modelo en datos nuevos.

Asignando un amplio rango de valores posibles para cada uno de los hiperparámetros, se entrenaron los modelos probando a fuerza bruta todas las combinaciones posibles de valores, seleccionando el conjunto de hiperparámetros que produjo el mejor rendimiento en el conjunto de validación. Si bien la búsqueda de cuadrícula es computacionalmente costosa, la misma permitió realizar la evaluación de un amplio rango de combinaciones posibles desempeñándose en tiempos adecuados.

5. Resultados.

En este capítulo se presentarán los resultados más relevantes obtenidos al evaluar la capacidad predictiva de los 10 modelos propuestos. A continuación, se muestran dos tablas que resumen los valores de las medidas de rendimiento predictivo alcanzados por cada modelo, utilizando la configuración óptima de hiperparámetros. Se incluyen 7 métricas de error diferentes para comparar los errores en las etapas de entrenamiento y de testeo. Es importante destacar que durante las instancias de entrenamiento y validación se trabajó con muestras balanceadas, donde la proporción de desertores y no desertores fue equilibrada. Sin embargo, en la instancia de prueba se utilizó la muestra sin balancear, a fin de replicar la situación real donde se busca identificar a los jóvenes[17] de entre 11 y 18 años que abandonan la escuela prematuramente.

[17] En la muestra de testeo, el porcentaje de desertores es de solo el 4,3% ya que, por un lado se analizó una franja etaria de 8 años en donde la proporción de desertores es mucho menor mientras menos años tienen. Por otro lado, también se menciona que se utilizó la EPH sin expandir.

Tabla de Resultados en Validacion:

Modelo & Parametros	ACC	AUC	ECM	TP	FN	TN	FP
LDA tolerance = 0.001	0,8136068	0,8680741	0,1863932	496	172	591	77
Regresion Logística	0,7649422	0,8373365	0,2350578	**517**	**151**	505	163
Regularizador Lasso con λ= 0.00001	0,8151049	0,8696192	0,1848951	498	170	591	77
Regularizador Ridge con λ= 0.01	0,8158512	0,8689686	0,1841488	498	170	592	76
SVC con C = 1.0	0,5194423	0,5195519	0,4805577	58	610	**636**	**32**
Regression Tree with maxdepth = 10	0,7559645	0,7497794	0,2440355	473	195	537	131
BAG n_estimator = 85	0,8091348	0,8655722	0,1908652	516	152	565	103
RFC n_estimator = 40	0,8031590	0,8623439	0,1968410	506	162	567	101
BOOST n_estimator = 75	**0,8173437**	**0,8732546**	**0,1826563**	516	152	576	92
KNN n_neighbors = 7	0,7282853	0,7731935	0,2717147	456	212	517	151

Tabla de Resultados en Test:

Modelo & Parametros	ACC	AUC	ECM	TP	FN	TN	FP
LDA tolerance = 0.001	0,8078791	0,8647555	0,1921209	**107**	**32**	2887	680
Regresion Logística	0,7579601	0,8317531	0,2420399	106	33	2703	864
Regularizador Lasso con λ= 0.00001	0,8138154	0,8657014	0,1601443	**107**	**32**	2909	658
Regularizador Ridge con λ= 0.01	0,8167836	0,8651367	0,1590578	**107**	**32**	2920	647
SVC con C = 1.0	**0,9028602**	0,5243328	**0,0971398**	16	123	**3330**	**237**
Regression Tree with maxdepth = 10	0,7865623	0,7259138	0,2134377	88	51	2827	740
BAG n_estimator = 85	0,8240691	0,8586231	0,1759309	106	33	2948	619
RFC n_estimator = 40	0,8486239	0,8494079	0,1513761	100	39	3045	522
BOOST n_estimator = 75	0,8472747	**0,8741764**	0,1527253	104	35	3036	531
KNN n_neighbors = 7	0,7412304	0,7701472	0,2587696	95	44	2652	915

Tanto en el conjunto de validación como en el conjunto de prueba, se lograron obtener niveles de *accuracy* o precisión (ACC) superiores al 80% para este desafio de prediccion de desercion escolar en Argentina. Estos resultados son especialmente buenos considerando que para el entrenamiento de estos modelos, se utilizaron encuestas habitacionales que no fueron específicamente diseñadas para abordar cuestiones relacionadas a la educación. El ejercicio realizado buscó predecir un fenómeno social complejo y multicausal, lo cual añade dificultad adicional a la tarea de predicción realizada.

Los resultados obtenidos son consistentes, ya que las predicciones realizadas sobre la muestra desconocida, correspondiente a un periodo temporal posterior y compuesta por jóvenes cuyos datos no fueron utilizados para entrenar el modelo, logran valores de AUC (área bajo la curva ROC) prácticamente idénticos a los de entrenamiento. Las ligeras discrepancias en los rendimientos observados reflejan un buen poder predictivo de los modelos, sin evidencia de sobreajuste a los datos de entrenamiento. Por otro lado, eran de esperarse las variaciones de mayor magnitud observadas en términos de precisión (ACC), considerando que esta métrica es susceptible a las bases de datos desbalanceadas.

En términos generales, se puede afirmar que todos los modelos logran predicciones considerablemente buenas, a excepción del *Support Vector Machine* (SVC), el cual exhibe

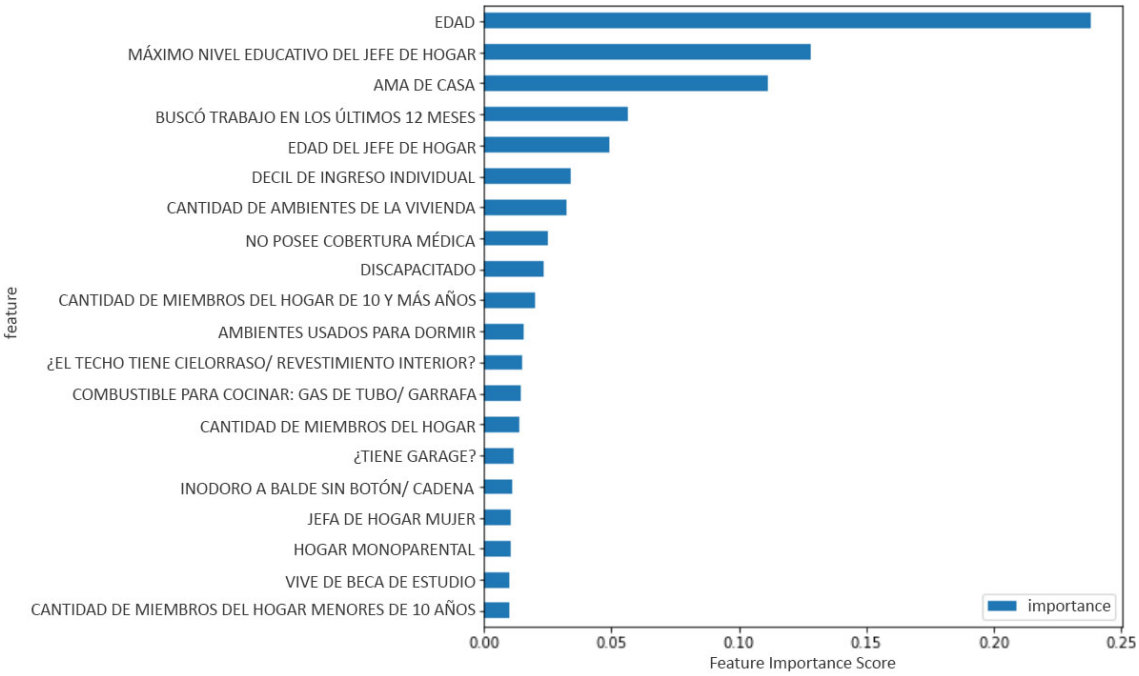
predicciones casi aleatorias que no alcanza ni el 52% de aciertos en la etapa de validación, pero que sorprendentemente supera el 90% al ser evaluado con los datos de testeo. Incluso más, este modelo presenta el mejor desempeño en 4 de las 7 métricas evaluadas: precisión global (ACC), error cuadrático medio (ECM), verdaderos negativos (TN) y falsos positivos(FP). Sin embargo, este resultado debe ser interpretado con cautela debido al desbalance de la muestra utilizada. Como se explicó anteriormente, en este caso particular, el SVC simplemente asigna la amplia mayoría de las clasificaciones a la clase mayoritaria (no desertores), lo que resulta en una precisión virtual sobresaliente de más del 90%, pero que apenas logra identificar a 16 individuos de nuestra población objetivo (desertores). Este modelo carece de valor para nuestros propósitos considerando que buscamos alcanzar a la mayor cantidad de niños vulerables, por lo que esta cantida de aciertos positivos es inaceptable.

Como se ha mencionado a lo largo de este estudio, el objetivo del trabajo es seleccionar el mejor modelo para predecir la deserción escolar en Argentina. Tras realizar un exhaustivo análisis, se concluye que el modelo de *boosting* es el más adecuado para este propósito, ya que maximiza el área bajo la curva ROC (AUC), obteniendo un valor de 0.8741. Este resultado indica que el modelo tiene una muy buena capacidad para distinguir entre las clases positivas y negativas, lo que nos permite identificar a un mayor número de jóvenes en riesgo de deserción, al tiempo que logra discernir correctamente aquellos que no son parte de nuestra población objetivo, lo que nos brinda la posibilidad de enfocar las políticas en los jóvenes más vulnerables sin desperdiciar recursos en aquellos que no lo necesitan. Si bien en la tabla de resultados el modelo de boosting no es el mejor en el resto de las métricas, se caracteriza por su buena performance general. Logra posicionarse dentro de los tres mejores clasificadores o muy cerca del mejor modelo en en cada una de las métricas. De esta manera, el modelo de boosting alcanza una precisión del 84.73% y un error cuadrático medio de 0,1527. En cuanto a las métricas de clasificación, logra predecir correctamente 104 verdaderos positivos (TP) y 3.036 verdaderos negativos (TN), cometiendo solo 35 falsos negativos (FN) y 531 falsos positivos (FP).

Por otro lado, aunque el modelo de árbol de decisión muestra un rendimiento resagado en comparación con los modelos de ensambles, es importante destacar que, para alcanzar el consenso en políticas públicas, es necesario brindar reglas claras, transparentes y fácilmente comprensibles para la población. En este sentido, los árboles de decisión son particularmente deseables, ya que nos permiten comprender de forma sumamente clara los criterios utilizados para identificar a la población objetivo. Además, la estructura del árbol nos brinda información sobre los factores que tienen un mayor impacto como desencadenantes de la deserción escolar.

Al analizar la relevancia de atributos, el árbol de decision arroja el siguiente listado de variables, las cuales expresan en orden decreciente el aporte de cada una al poder preductivo del modelo:

Gráfico 8. Aporte de cada variable al poder predictivo del modelo.



Fuente: Elaboración propia en base a los datos de las EPH 2021 y 2022.

Los resultados obtenidos resultan sumamente coherentes y son consistentes con estudios previos de los determinantes de la desercion escolar en Argentina. A continuacion se explica el significado de las principales variables identificadas y su relación con el abandono. El detalle de cada variable se encuentra en el Apéndice de Datos.

- **EDAD: ¿Cuántos años cumplidos tiene?**
  A medida que los jóvenes avanzan en edad, se observa un aumento en la probabilidad de que abandonen la escuela. Esta tendencia, tal como fue evidenciada en investigaciones anteriores (Paz y Cid, 2012; Alderete *et al*., 2017; Ibañez *et al.,* 2020), es comprensible debido al incremento en el costo de oportunidad que implica optar por la participación en el mercado laboral en detrimento de la asistencia escolar. Este fenómeno se atribuye a la presión creciente con los años que enfrentan los adolescentes para contribuir a la economía familiar.

- **Maximo nivel educativo alcanzado por el jefe de hogar.**
  Existe una relación positiva entre el clima educativo del hogar y la probabilidad de asistencia escolar. Esta relación se basa en la importancia que se le otorga a la inversión en educación, la cual está estrechamente relacionada con el nivel educativo de los padres. Por lo tanto, es probable que los padres con mayor nivel educativo realicen mayores esfuerzos para asegurar la asistencia regular de sus hijos a la escuela. Además, se argumenta que los padres con mayor educación poseen recursos y habilidades que les permiten apoyar de manera más efectiva el proceso educativo de sus hijos. Estos hallazgos concuerdan con las investigaciones de Sosa y Marchionni (1999), Bertranou (2002), Binstok y Cerrutti (2005), Paz y Cid (2012), Alderete et al. (2017).

- **Categoría de inactividad = Ama de Casa**

  Se observa un efecto negativo en la asistencia escolar de los jóvenes que tienen responsabilidades en las tareas domésticas y el cuidado del hogar. Este fenómeno es coherente con la necesidad de destinar tiempo equivalente, e incluso mayor, al trabajo que implica el cuidado del hogar, de las personas mayores, los niños o de cualquier persona que dependa de una asistencia constante. Estos hallazgos coinciden con los resultados obtenidos por Miranda (2010) e Ibañez *et al.* (2020), quienes indican que en Argentina es común que las personas que realizan tareas domésticas de forma regular tengan una menor participación en la educación secundaria. Además, concluyen que este fenómeno ocurre con mayor frecuencia en familias con bajos niveles de ingresos y especialmente en el caso de las mujeres.

- **Buscó trabajo momento de los ultimos 12 meses**

  Al igual que con el cuidado del hogar, se observa un efecto negativo en la asistencia escolar de los jóvenes que están involucrados en actividades laborales. Como se mencionó anteriormente, el trabajo y la educación compiten por el tiempo de los individuos, convirtiéndose en opciones excluyentes para los jovenes que tienen una mayor dependencia de su propia obtencion de ingresos. Binstock y Cerruti (2005) asi como Groisman (2012) coinciden en que la inserción en el mercado laboral es un factor determinante del abandono y retraso escolar.

- **Edad del Jefe de hogar**

  La edad del jefe de hogar genera un efectecto negativo en la asistencia escolar cuanto mas a los extremos se ubican sus edades en la curba de distribucion. Curiosamente esta variable es la que genera mas ramificaciones en el arbol obtenido, denotando que los jefes de hogar mas jovenes o mas a ancianos afectan negativamente los logros escolares de los jovenes bajo su cuidado. Si bien los motivos subyacentes trascienden los límites de nuestro estudio, se puede asumir que los padres de edad adulta poseen mayor madurez, experiencia y estabilidad económica, lo que, como mencionamos previamente, favorece el entorno educativo. Por otro lado, la presencia de jefes de hogar muy jóvenes o muy ancianos puede indicar que los jóvenes no están bajo el cuidado directo de sus padres, sino de hermanos mayores, tíos, abuelos u otros responsables debido a la falta de presencia de sus progenitores.

- **Cantidad de ambientes de la vivienda**

  Cantidad de miembros del hogar de 10 y más años

  Diversos autores encuentran a las viviendas en condiciones de hacinamiento entre uno de los mayores determinantes negativos de los logros educativos. Estas se caracterizan por tener pocos ambientes de espacios reducidos con alta densidad de personas. El hacinamiento es un indicador de condiciones habitacionales deficientes que afectan tanto la salud como las oportunidades de estudio de los individuos. Además, la escasez

de espacio físico usualmente es acompañada por la escasez de materiales y recursos didácticos, como libros, escritorios o computadoras. (Herrero, 2005; Groisman y Calero, 2010; Paz y Cid, 2012; Alderete et al., 2017)

- **Numero de decil de ingreso total individual del total EPH**
  No posee cobertura medica
  La falta de cobertura médica y la ubicación en los deciles inferiores de ingresos indican que el hogar se encuentra excluido del mercado laboral, o al menos del mercado laboral formal. La ausencia de ingresos y de aportes de cobertura médica impulsan a los adolescentes a una temprana inserción en la población laboralmente activa y los expone a mayores riesgos y dificultades en caso de enfrentar problemas de salud (Binstock y Cerruti, 2005; Groisman y Calero, 2010; Paz y Cid, 2012, Ibañez *et al.,* 2020).

- **Categoría de inactividad = Discapacitado**
  Los niños con discapacidad poseen una mayor propensión a no asistir a la escuela, ncrementándose de manera significativa en los casos que presentan múltiples discapacidades o en aquellas de mayor severidad. En el nivel secundario, el 35% de los jovenes que poseen más de una dificultad funcional no asisten a la escuela[18]. Debido a las caracteristicas propias de esta poblacion, su analisis y tratamiento debe ser realizado de forma particular, trascendiendo el alcance de la presente investigación.

A raíz de los resultados obtenidos, se puede afirmar que el riesgo de abandono escolar entre los jóvenes de 11 a 18 años en Argentina está estrechamente vinculado con diversos factores. Entre los mas relevantes, se destaca la influencia directa del avance en la edad, donde a medida que los adolescentes se acercan a la etapa adulta, aumenta la probabilidad de deserción. Además, la participación activa del adolescente en el mercado laboral o bien la responsabilidad de las tareas domésticas también se relacionan con un mayor riesgo de abandono escolar. Asimismo, vivir en condiciones de hacinamiento o presentar deficiencias habitacionales impacta negativamente en la continuidad educativa. Por otro lado, se observa una relación inversa entre el abandono escolar y el clima educativo del hogar, el acceso a cobertura de salud y el nivel de ingresos del grupo familiar. Es decir, un entorno familiar propicio para el aprendizaje, el acceso a servicios de salud y una situación económica estable actúan como factores protectores frente a la deserción.

Considerando las condiciones de vulnerabilidad mencionadas, propias de la población identificada como desertores, en el siguiente capítulo se presentará una propuesta para la implementación a nivel federal de forma rápida y económica del modelo de predicción de abandono desarrollado. En este sentido, se incluirá una propuesta de adaptación de una política pública existente con el objetivo de abordar y mitigar esta problemática de manera efectiva.

[18] Fact Sheet: Children with Disabilities. UNICEF. August 2022 https://www.unicef.org/media/128976/file/UNICEF%20Fact%20Sheet%20:%20Children%20with%20Disabilities.pdf

6. Discusión y propuesta para las políticas públicas

En la mayoría de los países de la región, existe un consenso generalizado sobre la necesidad de focalizar los esfuerzos en la prevención de la deserción escolar, por lo que se ha apostado por el desarrollo de los Sistemas de Alerta Temprana. Estas herramientas permiten detectar y abordar tempranamente las situaciones de vulnerabilidad que afectan a los jóvenes buscando reducir la tasa de abandono escolar. Para la implementación efectiva de estos sistemas es fundamental contar con Sistemas de Información para la Gestión Educativa (SIGED) sólidos y actualizados a nivel nacional y jurisdiccional. Estos sistemas permiten recopilar y monitorear periódicamente datos de calidad de cada estudiante.

En el caso de Argentina, se ha iniciado el desarrollo del Sistema Integral de Información Digital Educativa (SInIDE), sin embargo, la infraestructura digital del sistema educativo nacional aún presenta importantes deficiencias. Actualmente, solo un tercio de los estudiantes en niveles obligatorios se encuentran registrados en el registro nacional. Esta situación representa uno de los principales obstáculos en la implementación de los sistemas de alerta temprana en nuestro país. Además, la disparidad en los niveles de avance de los SIGED es evidente, ya que no todas las provincias cuentan con un sistema nominalizado que permita el seguimiento de la trayectoria de cada estudiante y su vinculación a un establecimiento educativo.

La falta de un sistema centralizado de información genera una gran disparidad entre los datos que recopila cada una de las escuelas, en especial porque todavía hay provincias que no centralizan los datos escolares. Este problema dificulta el desarrollo de políticas educativas federales en un sistema educativo que se sustenta y organiza de forma descentralizada entre las diferentes provincias e incluso dentro de sus municipios. A nivel federal, la ANSES es el organismo que posee mayor disponibilidad de información de cada habitante del suelo argentino. Registra las composiciones de cada hogar y las condiciones de empleo, así como los niveles de ingreso de cada familia. Sin embargo, esto no es suficiente para cubrir todas las necesidades de información del sistema educativo. El desarrollo de un sistema de alerta temprana de alcance nacional requeriría del esfuerzo conjunto y coordinado de todos los actores involucrados en el sistema educativo.

En este estudio se emplearon herramientas de *machine learning* para entrenar un modelo de predicción de deserción escolar, que logro alcanzar un AUC-ROC del 87% en datos desconocidos, utilizando datos obtenidos de la encuesta permanente de hogares. Aunque esta información solo representa una muestra de la población urbana del país, una parte de ella ya es recolectada regularmente por la ANSES a nivel nacional. Por otro lado, como fue descripto en el capítulo anterior, es importante destacar que la mayoría de los jóvenes que abandonan sus estudios presentan condiciones de vulnerabilidad social similares a las de los beneficiarios de la Asignación Universal por Hijo o de la Ayuda Escolar Anual, especialmente considerando que el 80% de los desertores identificados, provienen de los 3 deciles inferiores de ingresos familiares, por lo que muy probablemente sean beneficiarios de estos programas. Debido a la similitud en

cuanto a la población objetivo de bajos recursos, estos programas nacionales en gran medida alcanzan a los mismos jóvenes que no logran finalizar su educación obligatoria.

Para poder implementar un programa efectivo que ayude a reducir las brechas educativas y de desigualdad en todo el país, se sugiere la incorporación de un cuestionario obligatorio para todos los padres que reciben la AUH y/o la Ayuda Escolar Anual. La AUH es el programa de transferencias condicionadas más grande de Argentina, que tiene como objetivo garantizar un ingreso mínimo para los niños del país y fomentar el desarrollo de habilidades a través del cumplimiento de condicionalidades. Resumidamente este consiste en una prestación mensual que se paga por cada hijo menor de 18 años -con un máximo de 5 hijos por familia y sin límite de edad cuando se trate de un hijo discapacitado-, a uno solo de los padres -priorizando a la madre- siempre y cuando ambos se encuentren desempleados o trabajando informalmente. Tiene como condicionalidades el cumplimiento de chequeos anuales de salud, el calendario de vacunación y la asistencia escolar. Por su lado, la Ayuda Escolar Anual es un apoyo económico que reciben una vez al año los titulares de asignación familiar y asignación universal por cada hija o hijo en edad escolar, para lo cual, también es necesaria realizar la acreditación de escolaridad.

Sobre la base de estos programas nacionales ya existentes, se propone la incorporación de un cuestionario adicional que los padres deberán presentar anualmente junto con las acreditaciones de asistencia escolar. Este formulario debe ser fácil de entender y que permita recopilar información de las principales variables requeridas para el ejercicio de predicción, contando con preguntas referidas a las condiciones habitacionales, el nivel educativo de los padres, la condición de repitencia de los jóvenes o situaciones de embarazo adolescente, entre otros. Al combinar esta información con los datos socioeconómicos recopilados por la ANSES, se podría crear de forma sencilla y económica, una base de datos federal que contenga los principales indicadores necesarios para entrenar un modelo de predicción similar al utilizado en este estudio.

La ANSES cuenta con un alcance universal en todo el territorio argentino, así como con una capacidad de gestión de información que la convierte en un organismo público idóneo para recopilar los datos solicitados en estos cuestionarios, los cuales serían incorporados como un requisito para la continuidad de las asignaciones que la ANSES misma provee. En estas condiciones, la generación de esta base de datos implicaría un costo marginal para la escala de esta política pública. De esta manera, sería posible llevar a escala nacional el modelo de predicción planteado para los jóvenes en situación de mayor vulnerabilidad.

Los jóvenes identificados con mayor probabilidad de deserción serían asignados automáticamente a un sistema centralizado de tutorías telefónicas. Este sistema permitiría profundizar en los riesgos que afronta cada alumno en particular y trabajar en herramientas para fortalecer su inserción en el sistema escolar. Un aspecto central a destacar aquí, es que la implementación de este programa nacional evitaría los problemas de federalismo en la negociación con todas las provincias en temas de educación y permitiría abordar de manera más efectiva la problemática de la desigualdad educativa y la deserción escolar en Argentina.

Como fue explicado anteriormente, la deserción es el resultado final de un proceso de desvinculación de la escuela, en el cual esta decisión es impulsada por una gran variedad de factores, sobre los cuales es posible agrupar a los alumnos por estos tipos de condicionantes. La atención y seguimiento personalizado por parte de los tutores, posibilitaría identificar mejor cuales son los retos particulares que enfrenta cada alumno, con lo cual se le podrían ofrecer intervenciones personalizadas que ataquen de forma efectiva cada problemática. Algunos ejemplos de asistencias que se podrían brindar incluyen el seguimiento y asistencia para tratar los desafíos académicos, el acompañamiento psicológico para fortalecer las habilidades socioemocionales, y programas de refuerzos económicos que podrían ser incluidos como anexos de la AUH, la Ayuda Escolar o incluso de PROGRESAR, facilitando un monto de asignación superior para las familias más carenciadas.

Según la investigación llevada a cabo por Adrogué *et al.* (2022), actualmente en Argentina, los ingresos acumulados a lo largo de la vida laboral son un 43.7% mayores para aquellos individuos que han completado la educación secundaria en comparación con aquellos que solo finalizaron el nivel educativo primario. Además, los autores sostienen que la educación en Argentina conlleva retornos positivos en términos de ingresos laborales, con un incremento promedio del 10% por cada año adicional de estudio. Desde esta perspectiva, se argumenta que la inversión pública destinada a financiar estas asistencias educativas sería fácilmente justificable debido a los retornos futuros esperados para sus beneficiarios, los cuales incluso repagarían los costos de dichas asistencias mediante el aumento de los ingresos tributarios generados.

El Banco Interamericano de Desarrollo (BID) está actualmente brindando apoyo para implementar un sistema de tutorías telefónicas en Argentina y otros países de la región[19]. Tras realizar un diagnóstico inicial, se diseña un recorrido específico para cada estudiante, considerando sus necesidades individuales. Bajo el sistema ya implementado, a cada niño se le asigna un tutor, quien envía ejercicios y material didáctico por WhatsApp a los padres, y realiza llamadas semanales para su explicación y seguimiento. La atención personalizada y el vínculo establecido con el tutor permiten que los estudiantes reconozcan sus dificultades de aprendizaje específicas. A los tutores, esta metodología les permite motivar y brindar apoyo individualizado y herramientas necesarias para que el estudiante resuelva las tareas y continúe aprendiendo en la escuela (Bergamaschi *et al.* 2022).

Este sistema ha demostrado resultados satisfactorios desde su reciente implementación. Según los estudios realizados por el Centro para la Evaluación de Políticas basadas en Evidencia (CEPE) de la Universidad Torcuato Di Tella, tras la implementación de tutorías virtuales en las ciudades de Buenos Aires y Mendoza, los estudiantes con dificultades de concentración han experimentado una mejora del 13% en su desempeño en matemáticas gracias a la asistencia brindada. La sistematización desarrollada por los alumnos durante los procesos de tutorías, reflejan dejar mayores motivaciones en los niños para continuar realizando sus tareas escolares en un 40% de los casos. Los tutores encargados de impartir las tutorías han afirmado que estas

[19] El sistema ya está en ejecución en Argentina, México, Guatemala y El Salvador y preparando sus próximos lanzamientos en Brasil, Ecuador, Paraguay, Perú, República Dominicana y Uruguay.

brindan un apoyo socioemocional significativo y aceleran el proceso de aprendizaje, lo cual ha sido corroborado por los padres de los estudiantes tratados, quienes han manifestado la observación de mejoras en los hábitos de estudio, el orden y la disciplina con la tarea, así como una mayor motivación y curiosidad por la escuela. A su vez, la evidencia empírica proveniente de otros países respalda estos resultados, entre los que se destacan un incremento del 11% en la dedicación al estudio de los jóvenes y una disminución de síntomas de depresión de 0.16 desviaciones estándar en Italia, una reducción del 8,9% en la repitencia en España y una reducción del 4% del miedo antes de asistir a la escuela en El Salvador (Dorna *et al. 2022*).

La generalización del uso de la tecnología en respuesta al cierre de escuelas debido a la pandemia ha permitido la implementación de esta estrategia de tutorías telefónicas que resulta altamente eficiente, individualizada, sencilla y de rápida implementación. Esta práctica profesional se caracteriza por un bajo costo, ya que puede ser aplicado de forma centralizada a nivel nacional. En contraste con otros países, en Argentina, las tutorías fueron realizadas por estudiantes de formación docente, quienes han manifestado de manera prácticamente unánime su gratitud por el proceso de desarrollo personal y profesional generado por su participación en esta iniciativa. Ante la acogida positiva de esta práctica, cabría la posibilidad de que en un futuro se extienda como parte de un programa de prácticas profesionales para estudiantes de pedagogía, psicología y profesorados, ofreciéndose a todo el país mediante convenios con universidades docentes en un número limitado de provincias.

7. Conclusiones

La deserción escolar constituye un problema de magnitud alarmante que acarrea graves consecuencias a lo largo de todo el desarrollo de la vida del individuo, además de tener un impacto negativo en el ámbito social. En Argentina, existe una notable carencia de herramientas y metodologías idóneas para identificar a los jóvenes desertores o en riesgo de abandonar su educación, así como para evaluar su situación, analizar las causas de su exclusión y, en consecuencia, diseñar políticas eficaces para abordar este fenómeno. El primer paso fundamental para brindar apoyo a estos estudiantes en riesgo consiste en comprender su situación; no obstante, dada la inexistencia de un sistema de seguimiento nacional, resulta imposible obtener datos precisos.

Utilizando los datos socioeconómicos y habitacionales provenientes de la Encuesta Permanente de Hogares, se logró desarrollar un modelo de *boosting* que ha demostrado un muy buen desempeño del 87.4% bajo la métrica AUC-ROC. Además, se lograron identificar los atributos con mayor influencia como factores desencadenantes del abandono escolar. Sin embargo, aunque estos datos son extrapolables a nivel nacional, al proceder de una muestra acotada, no permiten una identificación real de las personas que requieren asistencia para fortalecer su inserción en el sistema educativo.

Dado que los factores más influyentes identificados por nuestro modelo son coincidentes con los de las poblaciones más vulnerables alcanzadas por asistencias sociales como la Asignación Universal por Hijo o la Ayuda Escolar Anual, proponemos apalancarnos de estos programas en colaboración con la Administración Nacional de la Seguridad Social. A través de la inclusión de un nuevo cuestionario, el cual se requeriría junto con la presentación de la certificación de escolaridad para la continuidad de estos programas, sería posible obtener información sobre las condiciones habitacionales de las familias más vulnerables. Esta información, en conjunto con los datos socioeconómicos ya recopilados por la ANSES, nos permitiría generar una base de datos comparativa a la Encuesta Permanente de Hogares, pero de alcance nacional.

Como medida final para abordar la problemática planteada, proponemos incluir a todos los jóvenes identificados con mayor riesgo de deserción en un sistema de tutorías a distancia. A modo de sugerencia para futuras investigaciones, se podría analizar la viabilidad de integrar estas tutorías como parte de un programa de prácticas profesionales dirigido a estudiantes de pedagogía, psicología y profesorados. Sobre la viabilidad de este programa, también debe plantearse que es necesario garantizar el compromiso de los tutores, debiendo mantener continuidad con cada alumno y un adecuado seguimiento de las necesidades que presenten. Aun más, en su desarrollo deberán asentarse las bases para lograr garantizar la continuidad del programa más allá de la cara política que asuma el gobierno, especialmente al considerar que las inversiones en educación dan frutos a largo plazo, pero implican mayoritariamente costos en el corto, lo cual en muchas ocasiones genera desincentivos a su inversión. De lograr su exitosa incorporación al programa de prácticas profesionales, permitiría proporcionar un acompañamiento individualizado de manera eficiente en términos de costos, superando los desafíos específicos que cada provincia enfrenta en la gestión de datos y, de esta manera, ofrecer una política educativa de alcance nacional.

En el futuro, cuando el Sistema Integral de Información Digital Educativa (SInIDE) esté completamente operativo, sería ideal vincularlo con la base de datos propuesta. Esto permitiría generar una base de datos nacional exhaustiva que posibilitaría un abordaje integral de la trayectoria educativa de los estudiantes, abarcando tanto su rendimiento académico como sus condiciones habitacionales, así como también los factores socioeconómicos que los afectan. El propósito fundamental de esta integración es comprender de manera detallada la situación individual de cada alumno, con el objetivo de mejorar la capacidad predictiva del modelo. Asimismo, sería deseable poder actualizar los datos con la mayor frecuencia posible, con el objetivo de identificar a los jóvenes en mayor riesgo de deserción con la suficiente antelación antes de que abandonen el sistema educativo.

8. Referencias

- Adelman, M., Haimovich, F., Ham, A. y Vázquez, E. (2018). Predicting school dropout with adminis trative data: New evidence from Guatemala and Honduras. Education Economics, 26:4, 356-372.
- Adrogué, C., Catri, G., Nistal, M. y Volman, V. (2022). Retornos de la educación. ¿Vale la pena estudiar? Observatorio de Argentinos por la Educación.
- Adrogué, C., Orlicki, M. E. (2018) Estudiantes en riesgo: un análisis de los factores asociados al abandono de la escuela secundaria en la Argentina desde 2003; Universidad Nacional de Comahue; Pilquen; 15; 1; 7-2018; 21-32
- Alderete, M. V., Formichella, M. M. y Krüger, N. (2017). Acceso y uso de TIC en barrios vulnerables de Bahía Blanca: su efecto sobre los resultados educativos. Asociación Argentina de Economía Política.
- Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432.
- Bergamaschi, A., Dellepiane, P., Hevia, P., Signorini M., P., Vinacur, T., y Zoido, P. (21 de junio de 2022). *Compromiso, creatividad, acción: apoyar estudiantes mediante tutorías remotas*. BID. https://blogs.iadb.org/educacion/es/tutorias-remotas-2
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.
- Bertranou, E., (2002). Determinantes del avance en los niveles de educación en Argentina. Análisis empírico basado en un modelo probabilístico secuencial. IIE, Working Papers 038, IIE, Universidad Nacional de La Plata.
- BID. (2020). Hablemos de política educativa. ¿Una década perdida? Los costos educativos de la crisis sanitaria en América Latina y el Caribe. División de Educación - Sector Social.
- Binstock, G. y Cerutti, M. (2005). Carreras Truncadas. El abandono escolar en el nivel medio en la Argentina. Buenos Aires: UNICEF.
- Breiman, L., Friedman, J. H., Stone, C. J., y Olshen, R. A. (1984). Classification and Regression Trees. Taylor; Francis.
- Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123-140.
- Bucciarelli, M. E., Paparella, C. y Perusia, J. C. (2022). ¿Cómo implementar un sistema de alerta temprana (SAT) para prevenir el abandono escolar? Buenos Aires: CIPPEC.
- Cerrutti, M. y Binstock, G. (2004). Camino a la exclusión: determinantes del abandono escolar en el nivel medio en la Argentina, Congresso da Associação Latino Americana de População, Brasil.

- Cetrángolo, O. y Curcio, J., (2018). Análisis y propuestas de mejoras para ampliar la Asignación Universal por Hijo. UNICEF.
- CIPPEC. (2021). El impacto de la pandemia en la educación secundaria en Argentina y América Latina
- Dorna, G., Gertner, G., Mateo, M., Regalía, F., Bergamaschi, A., Vinacur T., Levy Yeyati, E. y Cruces, J., J., (18 de octubre de 2022). *Tutorías Remotas en Argentina 2022* [Jornada de cierre]. Universidad Torcuato Di Tella, Argentina. https://www.youtube.com/watch?v=IM6ssAc301w
- Edo, M., Marchionni, M., & Garganta (2015). Conditional cash transfer programs and enforcement of compulsory education laws. The case of Asignación Universal por Hijo in Argentina. CEDLAS Working Paper N. 190.
- Geron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.
- Giovagnoli, P. I. (2007). Failures in school progression. CEDLAS
- Groisman, F. (2012). Determinantes de la escolarización y participación económica de los adolescentes en Argentina. Frontera Norte, vol. 24, núm. 48, México, Colegio de la Frontera Norte.
- Groisman, F. y Calero, A. V. (2010). Educación y participación económica de los jóvenes en Argentina. Un análisis de sus determinantes (2004-2009). Asociación Argentina de Economía Política.
- Herrero, V. (2005). Determinantes de situaciones de riesgo educativo en la población en edad escolar en Argentina. VIII Jornadas Argentinas de Estudios de Población de la Asociación de Estudios de la Población de Argentina.
- Ibáñez, M. M., Formichella M. M., y Costabel L. E. (2020). Exclusión social: explorando la dimensión educativa en Argentina. Revista Latinoamericana de Economía.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer Science & Business Media.
- Kit, I., España, S., Catri, G., Nistal, M. y Volman, V. (2022). Desgranamiento y aprendizajes desiguales: las dos caras de la misma moneda. Observatorio de Argentinos por la Educación.
- Miranda, A. (2010). Educación secundaria, desigualdad y género en Argentina. Revista Mexicana de Investigación Educativa -RMIE, núm. 45.
- Miravet, B. A. (2021). Mejora de las predicciones en muestras desbalanceadas, Universidad Autónoma de Madrid.
- Mitchell, T. M. (2006). The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department.

- Muñoz Jaramillo, V. D. (2021). Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín, Universidad Nacional de Colombia.
- Muñoz, R. R. (2011). Deserción escolar en la Argentina.
- Paz, J. A. y Cid, J. C. (2012). Determinantes de la asistencia escolar de los jóvenes en la Argentina. Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California.
- Perusia, J. C. (2021). Los sistemas de alerta temprana para prevenir el abandono escolar en América Latina y el Caribe. UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean
- Perusia, J. C. y Cardini A. (2021). Sistemas de alerta temprana en la educación secundaria: prevenir el abandono escolar en la era del COVID-19. Documento de Política Pública 233. Buenos Aires: CIPPEC.
- Sen, A., (1999). Development as freedom. New York: Oxford University Press. ISBN 9780198297581.
- Somasundaram, A., y Reddy, U. S. (2016). Data imbalance: effects and solutions for classification of large and highly imbalanced data. ICRECT.
- Sosa Escudero, W. y Marchionni, M. (1999). Household structure, gender, and the economic determinants of school attendance in Argentina. Documentos del Banco Mundial, núm. 37583.
- Tan, P.M., Steibach M., Kumar, V. (2006). Introduction to Data Mining.
- Templado, I., Catri, G., Nistal, M. y Volman, V. (2021) Evidencia sobre desigualdad educative en la Argentina. Observatorio de Argentinos por la Educación.
- UNICEF & UIS. (2016). Monitoring Education Participation: Framework for Monitoring Children and Adolescents who are Out of School or at Risk of Dropping Out. UNICEF Series on Education Participation and Dropout Prevention, Vol I. Geneva, UNICEF
- World Bank. (2021). Actuemos ya para Proteger el Capital Humano de Nuestros Niños; Los Costos y la Respuesta ante el Impacto de la Pandemia de COVID-19 en el Sector Educativo de América Latina y el Caribe. World Bank.

Apéndice 1: Datos.

A continuación, se expone el diseño de registro de la encuesta utilizada en el presente trabajo.

Para el entrenamiento, validación y testeo de los modelos, fueron utilizados únicamente **los atributos resaltados en negrita**.

Las variables no incluidas en el ejercicio de predicción fueron removidas según se explica en la sección 3.3. Limpieza de base.

Los atributos subrayados fueron transformados a *dummies*.

Variables Generadas

TIPO DE CAMPO: N = numérico ; C = Character

CAMPO	TIPO (longitud)	DESCRIPCIÓN
DESERTO	**N (1)**	**Jóvenes de hasta 18 años de edad cumplida que declaran no haber terminado la escuela y no continuar en la misma.** **0 = No desertó** **1 = Desertó** Criterios de elaboración: 'CH06' (¿Cuántos años cumplidos tiene? < 19) + 'NIVEL_ED' (¿Cuál es el máximo nivel educativo alcanzado? = Primario incompleto o Primario completo o Secundario incompleto) + 'CH10' (¿Asiste o asistió a algún establecimiento educativo? = No asiste, pero asistió).
JEFE_NIVEL_ED	**N (1)**	**Máximo nivel educativo alcanzado por el jefe de hogar:** **0 = Sin instrucción** **1 = Primario incompleto** **2 = primario completo** **3 = Secundario incompleto** **4 = secundario completo** **5 = Superior universitario incompleto** **6 = Superior universitario completo** ~~7 = Sin instrucción~~ Criterios de elaboración: 'CH03' (Relación de parentesco = Jefe de Hogar.) + 'NIVEL_ED' (Nivel educativo). Nota: se reordeno la categoría 7 (sin instrucción) pasando su valor al 0 para poder considerar el nivel educativo como una variable continua, bajo el supuesto de que cada nivel implica en promedio 3 años más de instrucción formal completa.

JEFE_TRABAJA	**N (1)**	**0 = hogares cuyo jefe no trabaja.** **1 = hogares cuyo jefe trabaja.** Criterios de elaboración: 'CH03' (Relación de parentesco = Jefe de hogar.) + 'ESTADO' (Condición de actividad = Ocupado).
JEFE_EDAD	**N (2)**	**¿Cuántos años cumplidos tiene el jefe de hogar?** Criterios de elaboración: 'CH03' (Relación de parentesco = Jefe de hogar) + 'CH06' (¿Cuántos años cumplidos tiene?).
JEFA_MUJER	**N (1)**	**Sexo del jefe de hogar** **0 = jefe de hogar varón.** **1 = jefa de hogar mujer.** Criterios de elaboración: 'CH03' (Relación de parentesco = Jefe) + 'CH04' (Sexo = Mujer).
CONYUGE_EXISTE (Hogar Monoparental)	**N (1)**	**Existencia de cónyuge del jefe de hogar.** **0 = Jefe de hogar sin cónyuge (monoparental)** **1 = Jefe de hogar con cónyuge** Criterios de elaboración: 'CH03' (Relación de parentesco = Cónyuge).
CONYUGE_TRABAJA	**N (1)**	**Condición de actividad del cónyuge del jefe de hogar** **0 = Cónyuge inactivo/desocupado.** **1 = Cónyuge ocupado**. Criterios de elaboración 'CH03' (Relación de parentesco = Cónyuge) + 'ESTADO' (Condición de actividad = Ocupado).

IDENTIFICACIÓN		
CAMPO	**TIPO (longitud)**	**DESCRIPCIÓN**
CODUSU	C (29)	Código para distinguir viviendas, permite aparearlas con Hogares y Personas. Además, permite hacer el seguimiento a través de los trimestres
NRO_HOGAR	N (1)	Código para distinguir Hogares, permite aparearlos con Personas.
REALIZADA	N (1)	Entrevista realizada = Sí = No (hogar no respuesta)
ANO4	N (4)	Año de relevamiento (4 dígitos)
Trimestre	N (1)	Ventana de observación = 1° trimestre = 2° trimestre = 3° trimestre = 4° trimestre
REGION	N (2)	Código de región 01 = Gran Buenos Aires 40 = NOA 41 = NEA 42 = Cuyo 43 = Pampeana 44 = Patagonia
MAS_500	C (1)	Aglomerados según tamaño N = Conjunto de aglomerados de menos de 500.000 habitantes S = Conjunto de aglomerados de 500.000 y más habitantes

AGLOMERADO	N (2)	Código de Aglomerado 02 = Gran La Plata 03 = Bahía Blanca - Cerri 04 = Gran Rosario 05 = Gran Santa Fé 06 = Gran Paraná 07 = Posadas 08 = Gran Resistencia 09 = Comodoro Rivadavia - Rada Tilly 10 = Gran Mendoza 12 = Corrientes 13 = Gran Córdoba 14 = Concordia 15 = Formosa 17 = Neuquén – Plottier 18 = Santiago del Estero - La Banda 19 = Jujuy-Palpalá 20 = Río Gallegos 22 = Gran Catamarca 23 = Gran Salta 25 = La Rioja 26 = Gran San Luis 27 = Gran San Juan 29 = Gran Tucumán - Tafí Viejo 30 = Santa Rosa – Toay 31 = Ushuaia - Río Grande 32 = Ciudad Autónoma de Buenos Aires 33 = Partidos del GBA 34 = Mar del Plata 36 = Río Cuarto 38 = San Nicolás – Villa Constitución 91 = Rawson – Trelew 93 = Viedma – Carmen de Patagones
PONDERA	N (6)	Ponderación

CARACTERÍSTICAS DE LA VIVIENDA		
CAMPO	**TIPO (longitud)**	**DESCRIPCIÓN**
IV1	**N (1)**	**Tipo de vivienda (por observación)** 1. **casa** 2. **departamento** 3. **pieza de inquilinato** 4. **pieza en hotel / pensión** 5. **local no construido para habitación**
IV1_Esp	C (45)	6. otros, especificar:
IV2	**N (2)**	**¿Cuántos ambientes/habitaciones tiene la vivienda en total? (sin contar baño/s, cocina, pasillo/s, lavadero, garage)**

IV3	**N (1)**	**Los pisos interiores son principalmente de...** 1. **mosaico / baldosa / madera /cerámica / alfombra** 2. **cemento / ladrillo fijo** 3. **ladrillo suelto / tierra**
IV3_Esp	C (45)	4. otros, especificar:
IV4	**N (2)**	**La cubierta exterior del techo es de...** 1. **membrana / cubierta asfáltica** 2. **baldosa / losa sin cubierta** 3. **pizarra / teja** 4. **chapa de metal sin cubierta** 5. **chapa de fibrocemento / plástico** 6. **chapa de cartón** 7. **saña / tabla / paja con barro / paja sola** **9. N/S. Departamento en propiedad horizontal**
IV5	**N (1)**	**¿El techo tiene cielorraso / revestimiento interior?** 1 **= Sí** 2 **= No**
IV6	**N (1)**	**Tiene agua...** 1. **por cañería dentro de la vivienda** 2. **fuera de la vivienda, pero dentro del terreno** 3. **fuera del terreno**
IV7	**N (1)**	**El agua es de...** 1. **red pública (agua corriente)** 2. **perforación con bomba a motor** 3. **perforación con bomba manual**
IV7_Esp	C (45)	4. otra fuente, especificar:
IV8	N (1)	¿Tiene baño / letrina? 1 = Sí 2 = No
IV9	**N (1)**	**El baño o letrina está...** 1. **dentro de la vivienda** 2. **fuera de la vivienda, pero dentro del terreno** 3. **fuera del terreno**
IV10	**N (1)**	**El baño tiene...** **1. inodoro con botón / mochila / cadena y arrastre de agua** 2. **inodoro sin botón / cadena y con arrastre de agua (a balde)** 3. **letrina (sin arrastre de agua)**
IV11	**N (1)**	**El desagüe del baño es...** 1. **a red pública (cloaca)** 2. **a cámara séptica y pozo ciego** 3. **solo a pozo ciego** 4. **a hoyo/excavación en la tierra**

IV12_1	**N (1)**	**La vivienda está ubicada cerca de basural/ es (3 cuadras o menos)** 1 **= Sí** 2 **= No**
IV12_2	**N (1)**	**La vivienda está ubicada en zona inundable (en los últimos 12 meses)** 1 **= Sí** 2 **= No**
IV12_3	**N (1)**	**La vivienda está ubicada en villa de emergencia (por observación)** 1 **= Sí** 2 **= No**

CARACTERÍSTICAS HABITACIONALES DEL HOGAR

CAMPO	TIPO (longitud)	DESCRIPCIÓN
II1	**N (2)**	**¿Cuántos ambientes / habitaciones tiene este hogar para su uso exclusivo?**
II2	**N (2)**	**De esos, ¿cuántos usan habitualmente para dormir?**
II3	**N (1)**	**Utiliza alguno exclusivamente como lugar de trabajo (para consultorio, estudio, taller, negocio, etc.)** 1 **= Sí** 2 **= No**
II3_1	N (1)	Si utiliza alguno exclusivamente como lugar de trabajo, **¿cuántos?**
II4	N (1)	Tiene además...
II4_1	**N (1)**	**cuarto de cocina** 1 **= Sí** 2 **= No**
II4_2	**N (1)**	**lavadero** **1 = Sí** **2 = No**
II4_3	**N (1)**	**garage** **1 = Sí** **2 = No**
II5	N (1)	De esos ... (los sí de la pregunta 4) ¿usan alguno para dormir? 1 = Sí 2 = No
II5_1	N (2)	Si utiliza alguno para dormir, **¿cuántos?**

II6	N (1)	De esos... (los sí de la pregunta 4) utiliza alguno de estos exclusivamente como lugar de trabajo (consultorio, estudio, taller, negocio, etc.) 1 = Sí 2 = No
II7	**N (2)**	**Régimen de tenencia** **01 = Propietario de la vivienda y el terreno** **02 = Propietario de la vivienda solamente** **03 = Inquilino / arrendatario de la vivienda** **04 = Ocupante por pago de impuestos / expensas** **05 = Ocupante en relación de dependencia** **06 = Ocupante gratuito (con permiso)** **07 = Ocupante de hecho (sin permiso)** **08 = Está en sucesión**
II7_Esp	C (45)	09 = Otra situación (especificar)
II8	**N (1)**	**Combustible utilizado para cocinar** **01 = Gas de red** **02 = Gas de tubo / garrafa** **03 = Kerosene / leña / carbón**
II8_Esp	C (45)	04 = Otro (especificar)
II9	**N (1)**	**Baño (tenencia y uso)** 01**= Uso exclusivo del hogar** 02**= Compartido con otro/s hogar/es de la misma vivienda** 03**= Compartido con otra/s vivienda/s** 04**= No tiene baño**

ESTRATEGIAS DEL HOGAR

CAMPO	TIPO (longitud)	DESCRIPCIÓN
• **¿En los últimos tres meses, las personas de este hogar han vivido...**		
V1	**N (1)**	**...de lo que ganan en el trabajo?** 1 **= Sí** 2 **= No**
V2	**N (1)**	**...de alguna jubilación o pensión?** 1 **= Sí** 2 **= No**
V21	**N (1)**	**...aguinaldo de alguna jubilación o pensión cobrada el mes anterior?** 1 **= Sí** 2 **= No**

V22	**N (1)**	**...retroactivo de alguna jubilación o cobró el mes anterior?** 1 **= Sí** 2 **= No**
V3	**N (2)**	**...de indemnización por despido?** 1 **= Sí** 2 **= No**
V4	N (1)	...de seguro de desempleo? 1 = Sí 2 = No
V5	**N (1)**	**...de subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.?** 1 **= Sí** 2 **= No**
V6	**N (1)**	**...con mercaderías, ropa, alimentos gobierno, iglesias, escuelas, etc.?** 1 **= Sí** 2 **= No**
V7	**N (1)**	**...con mercaderías, ropa, alimentos de familiares, vecinos u otras personas que no viven en este hogar?** 1 **= Sí** 2 **= No**
V8	**N (1)**	**...algún alquiler (por una vivienda, terreno, oficina, etc.) de su propiedad?** 1 **= Sí** 2 **= No**
V9	N (1)	...ganancias de algún negocio en el que no trabajan? 1 = Sí 2 = No
V10	N (1)	...intereses o rentas por plazos fijos / inversiones? 1 = Sí 2 = No
V11	**N (1)**	**...una beca de estudio?** 1 **= Sí** 2 **= No**
V12	**N (1)**	**...cuotas de alimentos o ayuda en dinero de personas que no viven en el hogar?** 1 **= Sí** 2 **= No**
V13	**N (1)**	**...gastar lo que tenían ahorrado?** 1 **= Sí** 2 **= No**
V14	**N (1)**	**...pedir préstamos a familiares / amigos** 1 **= Sí** 2 **= No**

V15	N (1)	...pedir préstamos a bancos, financieras, etc.? 1 = Sí 2 = No
V16	N (1)	¿Compran en cuotas o al fiado con tarjeta de crédito o libreta? 1 = Sí 2 = No
V17	N (1)	¿Han tenido que vender alguna de sus pertenencias? 1 = Sí 2 = No
V18	N (1)	Tuvieron otros ingresos en efectivo (limosnas, juegos de azar, etc.) 1 = Sí 2 = No
V19_A	N (1)	menores de 10 años ayudan con algún dinero trabajando? 1 = Sí 2 = No
V19_B	N (1)	menores de 10 años ayudan con algún menores pidiendo? 1 = Sí 2 = No

• Resumen del hogar

IX_Tot	**N (2)**	**Cantidad de miembros del hogar**
IX_Men10	**N (2)**	**Cantidad de miembros del hogar menores de 10 años**
IX_Mayeq10	**N (2)**	**Cantidad de miembros del hogar de 10 y más años**

• Ingreso total familiar

ITF	N (12)	Monto de ingreso total familiar
DECIFR	C (2)	Nº de decil del ingreso total del hogar del total EPH
IDECIFR	C (2)	Nº de decil del ingreso total del hogar del interior
RDECIFR	C (2)	Nº de decil de ingreso total del hogar de la región
GDECIFR	C (2)	Nº de decil de ingreso total del hogar del Conjunto de aglomerados de 500.000 y más habitantes

PDECIFR	C (2)	Nº de decil de ingreso total del hogar del conjunto de aglomerados de menos de 500.000 habitantes
ADECIFR	C (2)	Nº de decil de ingreso total del hogar del aglomerado .

• Ingreso per cápita familiar		
IPCF	N (12)	Monto de ingreso per cápita familiar
DECCFR	**C (2)**	**Nº de decil del ingreso per cápita familiar del total EPH**
IDECCFR	C (2)	Nº de decil del ingreso per cápita familiar del interior
RDECCFR	C (2)	Nº de decil de ingreso per cápita familiar de la región
GDECCFR	C (2)	Nº de decil de ingreso per cápita familiar del conjunto de aglomerados de 500.000 y más habitantes
PDECCFR	C (2)	Nº de decil de ingreso per cápita familiar del conjunto de aglomerados de menos de 500.000 habitantes (**Anexo I**)
ADECCFR	C (2)	Nº de decil de ingreso per cápita familiar del aglomerado
PONDIH	C (6)	Ponderador del ingreso total familiar y del ingreso per cápita familiar
• Organización del hogar		
VII 1_1	N (2)	Realización de las tareas de la casa. Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar
VII 1_2	N (2)	Realización de las tareas de la casa Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar
VII 2_1	N (2)	Otras personas que ayudan en las tareas de la casa. Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar 98: Ninguna
VII 2_2	N (2)	Otras personas que ayudan en las tareas de la casa. Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar 98: Ninguna

VII 2_3	N (2)	Otras personas que ayudan en las tareas de la casa. Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar 98: Ninguna
VII 2_4	N (2)	Otras personas que ayudan en las tareas de la casa. Número de componente del hogar 96: Servicio doméstico 97: Otra persona que no vive en el hogar 98: Ninguna

Diseño de registros de la base Personas

TIPO DE CAMPO

N = numérico	C = character

IDENTIFICACIÓN

CAMPO	TIPO (longitud)	DESCRIPCIÓN
CODUSU	C (29)	Código para distinguir VIVIENDAS, permite aparearlas con Hogares y Personas. Además, permite hacer el seguimiento a través de los trimestres.
NRO_HOGAR	N (2)	Código para distinguir hogares 51 = Serv. doméstico en hogares 71 = Pensionistas en hogares
COMPONENTE	N (2)	Número de componente: no de orden que se asigna a las personas que conforman cada hogar de la vivienda. Casos especiales: 51 = Servicio doméstico en hogares 71 = Pensionistas en hogares
H15	N (1)	Entrevista individual realizada 1 = Sí 2 = No
ANO4	N (4)	Año de relevamiento (4 dígitos)
TRIMESTRE	N (1)	Ventana de observación 1 = 1° trimestre 2 = 2° trimestre3 = 3° trimestre 4 = 4° trimestre
REGION	N (2)	Código de región 01 = Gran Buenos Aires 40 = Noroeste 41 = Nordeste 42 = Cuyo 43 = Pampeana 44 = Patagonia
MAS_500	C (1)	Aglomerados según tamaño N = Conjunto de aglomerados de menos de 500.000 habitantes S = Conjunto de aglomerados de 500.000 y más habitantes

AGLOMERADO	02 N (2)	03 Código de aglomerado 04 = Gran La Plata 05 = Bahía Blanca - Cerri 06 = Gran Rosario 07 = Gran Santa Fé 08 = Gran Paraná 09 = Posadas 10 = Gran Resistencia 11 = Comodoro Rivadavia - Rada Tilly 12 = Gran Mendoza 13 = Corrientes 14 = Gran Córdoba 15 = Concordia 16 = Formosa 17 = Neuquén - Plottier 18 = Santiago del Estero - La Banda 19 = Jujuy - Palpalá 20 = Río Gallegos 21 = Gran Catamarca 22 = Gran Salta 23 = La Rioja 24 = Gran San Luis 25 = Gran San Juan 26 = Gran Tucumán - Tafí Viejo 27 = Santa Rosa - Toay 28 = Ushuaia - Río Grande 29 = Ciudad Autonoma de Buenos Aires 30 = Partidos del GBA 31 = Mar del Plata 32 = Río Cuarto 33 = San Nicolás - Villa Constitución 34 = Rawson - Trelew 35 = Viedma - Carmen de Patagones
Pondera	N (6)	Ponderación

CARACTERÍSTICAS DE LOS MIEMBROS DEL HOGAR

CH03	**N (2)**	**Relación de parentesco** 01 **= Jefe/a** 02 **= Cónyuge/pareja** 03 **= Hijo / hijastro/a** 04 **= Yerno / nuera** 05 **= Nieto/a** 06 **= Madre / padre** 07 **= Suegro/a** 08 **= Hermano/a** 09 **= Otros familiares** 10 **= No familiares**

CH04	**N (1)**	**Sexo** 1 **= varón** 2 **= mujer**
CH 05	date	Fecha de nacimiento (día, mes y año)
CH06	**N (2)**	**¿Cuántos años cumplidos tiene?**
CH07	**N (1)**	**¿Actualmente está...** 1 **= ... unido?** 2 **= ... casado?** 3 **= ... separado/a o divorciado/a?** 4 **= ... viudo/a?** 5 **= ... soltero/a?**
CH08	**N (3)**	**¿Tiene algún tipo de cobertura médica por la que paga o le descuentan?** **1 = Obra social (incluye PAMI)** **2 = Mutual / prepaga / servicio de emergencia** **3 = Planes y seguros públicos** **4 = No paga ni le descuentan** **9 = Ns/Nr** **12 = Obra social y mutual / prepaga / servicio de emergencia** **13 = Obra social y planes y seguros públicos** **23 = Mutual / prepaga / servicio de emergencia/ Planes y seguros públicos** **123 = obra social, mutual/prepaga/ servicio de emergencia y planes y seguros públicos**
CH09	**N (1)**	**¿Sabe leer y escribir?** 1 **= Sí** 2 **= No** 3 **=** Menor de 2 años
CH10	N (1)	¿Asiste o asistió a algún establecimiento educativo? (colegio, escuela, universidad) 1 = Sí, asiste 2 = No asiste, pero asistió 3 = Nunca asistió
CH11	N (1)	Ese establecimiento es... 1 = ... público 2 = ... privado 9 = ... Ns/Nr

CH12	N (2)	¿Cuál es el nivel más alto que cursa o cursó? 1 = Jardín/preescolar 2 = Primario 3 = EGB 4 = Secundario 5 = Polimodal 6 = Terciario 7 = Universitario 8 = Posgrado universitario 9 = Educación especial (discapacitado)
CH13	N (1)	¿Finalizó 1 ese nivel? 1 = Sí 2 = No 3 = Ns/Nr
CH14	N (2)	¿Cuál fue el último año que aprobó? 00 = Ninguno 01 = Primero 02 = Segundo 03 = Tercero 04 = Cuarto 05 = Quinto 06 = Sexto 07 = Séptimo 08 = Octavo 09 = Noveno 98 = Educación especial 99 = Ns/Nr

CH15	**N (1)**	**¿Dónde nació?** **1. En esta localidad** **2. En otra localidad de esta provincia** **3. En otra provincia (especificar)** **4. En un país limítrofe (especificar: Brasil, Bolivia, Chile, Paraguay, Uruguay)** **5. En otro país (especificar)** **9. Ns/Nr**
CH15_Cod	C (3)	Especificar: contiene el código que corresponde a: 3. En otra provincia 4. En un país limítrofe 5. En otro país
CH16	**N (1)**	**¿Dónde vivía hace 5 años?** **1. En esta localidad** **2. En otra localidad de esta provincia** **3. En otra provincia (especificar)** **4. En un país limítrofe (especificar: Brasil, Bolivia, Chile, Paraguay, Uruguay).** **5. En otro país (especificar)** **6. No había nacido** **9. Ns/Nr**
CH16_Cod	C (3)	Especificar: contiene el código que corresponde a: 3. en otra provincia 4. en un país limítrofe 5. en otro país
NIVEL_ED	N (1)	Nivel educativo 1 = Primario incompleto (incluye educación especial) 2 = primario completo 3 = Secundario incompleto 4 = secundario completo 5 = Superior universitario incompleto 6 = Superior universitario completo 7 = Sin instrucción 9 = Ns/ Nr
ESTADO	**N (1)**	**Condición de actividad** **0 = Entrevista individual no realizada (no respuesta al cuestionario individual)** **1 = Ocupado** 2 **= Desocupado** 3 **= Inactivo** 4 = Menor de 10 años

CAT_OCUP	**N (1)**	**Categoría ocupacional** **(Para ocupados y desocupados con ocupación anterior)** **1 = Patrón** 2 **= Cuenta propia** 3 **= Obrero o empleado** 4 **= Trabajador familiar sin remuneración** 9 = Ns/Nr
CAT_INAC	**N (1)**	**Categoría de inactividad** 1 **= Jubilado / Pensionado** 2 **= Rentista** 3 = Estudiante 4 **= Ama de casa** 5 **= Menor de 6 años** 6 **= Discapacitado** 7 = Otros
IMPUTA	N (1)	Indica los casos que han sido imputados ¿De qué manera estuvo buscando trabajo?
PP02C1	N (1)	Hizo contactos, entrevistas
PP02C2	N (1)	Mandó currículum, puso, contestó avisos (diarios, internet)
PP02C3	N (1)	Se presentó en establecimientos
PP02C4	N (1)	Hizo algo para ponerse por su cuenta
PP02C5	N (1)	Puso carteles en negocios, preguntó en el barrio
PP02C6	N (1)	Consultó a parientes, amigos
PP02C7	N (1)	Se anotó en bolsas, listas, planes de empleo, agencias, contratistas, o alguien le está buscando trabajo
PP02C8	N (1)	De otra forma activa
PP02E	**N (1)**	**Durante esos 30 días, no buscó trabajo porque...** **1= está suspendido** **2= ya tiene trabajo asegurado** **3= se cansó de buscar trabajo** **4= hay poco trabajo en esta época del año** **5= por otras razones**
PP02H	**N (1)**	**En los últimos 12 meses ¿buscó trabajo en algún momento?** 1 **= Sí** 2 **= No**
PP02I	N (1)	En los últimos 12 meses ¿trabajó en algún momento? 1 = Sí 2 = No

• Ocupados que trabajaron en la semana de referencia

PP03C	N (1)	La semana pasada, ¿tenía... 1 = ...un solo empleo / ocupación / actividad? 2 = ...más de un empleo / ocupación / actividad?
PP03D	N (1)	Cantidad de ocupaciones
PP3E_TOT	N (5,1)	Total de horas que trabajó en la semana en la ocupación principal
PP3F_TOT	N (5,1)	Total de horas que trabajó en la semana en otras ocupaciones
PP03G	N (1)	La semana pasada, ¿quería trabajar más horas? 1 = Sí 2 = No
PP03H	N (1)	¿Si hubiera conseguido más horas... 1 = ...podía trabajarlas esa semana? 2 = ...podía empezar a trabajarlas en dos semanas a más tardar? 3 = ...no podía trabajar más horas? 9 = ...Ns/Nr
• Para todos los ocupados		
PP03I trabajar	N (1)	En los últimos treinta días, ¿buscó más horas? 1 = Sí 2 = No 9 = Ns/Nr
PP03J	N (1)	Aparte de este/os trabajo/s, ¿estuvo buscando algún empleo / ocupación / actividad? 1 = Sí 2 = No 9 = Ns/Nr
INTENSI	N (1)	1 = Subocupado por insuficiencia horaria 2 = Ocupado pleno 3 = Sobreocupado 4 = Ocupado que no trabajó en la semana 9 = Ns/Nr
• Ocupación principal		
PP04A	N (1)	¿El negocio / empresa / institución / actividad en la que trabaja es... (se refiere al que trabaja más horas semanales) 1 = ...estatal? 2 =... privada? 3 =... de otro tipo? (especificar)

PP04B_COD	N (5)	¿A qué se dedica o produce el negocio / empresa / institución? (Ver Clasificador de Actividades Económicas para Encuestas Sociodemográficas del Mercosur, CAES-Mercosur)
PP04B1	N (1)	Si presta servicio doméstico en hogares particulares 1 = casa de familia
PP04B2	N (1)	¿En cuántas casas trabaja? (cantidad)
• ¿Cuánto tiempo hace que trabaja allí? (en la casa que tiene más horas)		
PP04B3_MES	N (2)	mes
PP04B3_ANO	N (2)	año
PP04B3_DIA	N (2)	día
PP04C	N (2)	¿Cuántas personas, incluido ... trabajan allí en total? 1 = 1 persona 2 = 2 personas 3 = 3 personas 4 = 4 personas 5 = 5 personas 6 = de 6 a 10 personas 7 = de 11 a 25personas 8 = de 26 a 40personas 9 = de 41 a 100 personas 10 = de 101 a 200 personas 11 = de 201 a 500 personas 12 = más de 500 personas 99 = Ns/Nr
PP04C99	N (1)	1 = hasta 5 2 = de 6 a 40 3 = más de 40 9 = Ns/Nr
PP04D COD	C (5)	Código de ocupación (Ver Clasificador Nacional de Ocupaciones – CNO - versión 2001)
PP04G	N (2)	¿Dónde realiza principalmente sus tareas? 1 = En un local / oficina / establecimiento / negocio / taller/ chacra / finca 2 = En puesto o kiosco fijo callejero 3 = En vehículos: bicicleta, moto, auto, barco, bote (no incluye servicio de transporte). 4 = En vehículo para transporte de personas y mercaderías – aéreos, marítimo, terrestre – incluye taxis, colectivos, camiones, furgones, transporte de combustible, mudanzas, etc.) 5 = En obras en construcción, de infraestructura, minería o similares 6 = En esta vivienda. (sin lugar exclusivo).

		7 = En la vivienda del socio o del patrón 8 = En el domicilio/local de los clientes 9 = En la calle, espacios públicos, ambulante, de casa en casa puesto callejero, móvil. 10 = En otro lugar (especificar)

OCUPACIÓN PRINCIPAL DE LOS TRABAJADORES INDEPENDIENTES		
• ¿Cuánto tiempo hace que trabaja en ese empleo en forma continua...?		
PP05B2_MES	N (2)	meses
PP05B2_ANO	N (2)	años
PP05B2_DIA	N (2)	días
• ¿En ese negocio/empresa/actividad, tiene...?		
PP05C_1	N (1)	...maquinarias / equipos? 1 = propio (del negocio) 2 = prestado / alquilado 3 = no tiene
PP05C_2	N (1)	...local (incluye kiosco / puesto fijo) 1 = propio (del negocio) 2 = prestado / alquilado 3 = no tiene
PP05C_3	N (1)	...vehículo? 1 = propio (del negocio) 2 = prestado/alquilado 3 = no tiene
PP05E	N (1)	¿Para la actividad del negocio, en los últimos 3 meses, tuvo que gastar en la compra de materias primas, productos, pagar servicios u otros gastos 1 = Sí 2 = No
PP05F	N (1)	¿Ese negocio/empresas/actividad, trabaja habitualmente para... 6 = un solo cliente? (persona, empresa) 7 = distintos clientes? (incluye público en general)
PP05H	N (1)	¿Durante cuánto tiempo ha estado trabajando er ese empleo en forma continua? (con interrupciones laborales no mayores a 15 días) 1 = menos de 1 mes 2 = de 1 a 3 meses 3 = más de 3 a 6 meses 4 = más de 6 meses a 1 año 5 = más de 1 a 5 años 6 = más de 5 años 9 = Ns/Nr
• Ingresos de la ocupación principal de los trabajadores independientes		

PP06A	N (1)	En ese negocio / empresa / actividad tiene socios o familiares asociados? 1 = Sí 2 = No
PP06C	N (10)	Monto de ingreso de patrones y cuenta propia sin socios
PP06D	N (10)	Monto de ingreso de patrones y cuenta propia con socios
PP06E	N (1)	Ese negocio / empresa / actividad... 1 = ...es una sociedad jurídicamente constituida? (SA, SRL, Comandita por Acciones, etc.) 2 = ...es una sociedad de otra forma legal? 3 = ...o es una sociedad convenida de palabra?
PP06H	N (10)	¿Es una actividad familiar? (solo para 2 y 3 de PP06E) 1 = Sí 2 = No
• Ocupación principal de los asalariados (excepto servicio doméstico)		
PP07A	N (1)	¿Cuánto tiempo hace que está trabajando en ese empleo en forma continua? (sin interrupciones de la relación laboral en la misma empresa / negocio / institución) 1 = menos de 1 mes 2 = 1 a 3 meses 3 = más de 3 a 6 meses 4 = más de 6 a 12 meses 5 = más de 1 a 5 años 6 = más de 5 años 9 = Ns/Nr
PP07C	N (1)	¿Ese empleo tiene tiempo de finalización? 1 = Sí (incluye changa, trabajo transitorio, por tarea u suplencia, etc.) 2 = No (incluye permanente, fijo, estable, de planta) 9 = Ns/Nr
PP07D	N(1)	¿Por cuánto tiempo es ese trabajo? (Para los que tienen en PP07C= 1) 1 = solo fue esa vez / solo cuando lo llaman 2 = hasta 3 meses 3 = más de 3 a 6 meses 4 = más de 6 a 12 meses 5 = más de 1 año 9 = Ns/Nr

PP07E	N (1)	¿Ese trabajo es... 1 = ...un plan de empleo? 2 = ...un período de prueba? 3 = ...una beca / pasantía / aprendizaje 4 = ninguno de estos
• Ocupación principal de los asalariados (incluido servicio doméstico)		
PP07F1	N (1)	¿En este trabajo le dan... (no excluyentes) ...de comer gratis en el lugar de trabajo? 1 = Sí 2 = No
PP07F2	N (1)	...vivienda? 1 = Sí 2 = No
PP07F3	N (1)	… algún producto o mercadería?
PP07F4	N (1)	...algún otro beneficio? (automóvil, teléfono celular, pasajes, etc.) 1 = Si 2 = No
PP07F5	N (1)	¿No recibe ninguno? 1 = Sí ¿En este trabajo tiene... (no excluyentes)
PP07G1	N (1)	...vacaciones pagas? 1 = Si 2 = No
PP07G2	N (1)	…aguinaldo? 1 = Sí 2 = No
PP07G3	N (1)	...días pagos por enfermedad? 1 = Sí 2 = No
PP07G4	N (1)	...obra social? 1 = Sí 2 = No
PP07G_59	N (1)	no tiene ninguno 5 = Sí
PP07H	N (1)	¿Por ese trabajo tiene descuento jubilatorio? 1 = Sí 2 = No
PP07I	N (1)	¿Aporta por sí mismo a algún sistema jubilatorio? 1 = Sí 2 = No

PP07J	N (1)	¿El turno habitual de trabajo es... 1 = ...de día? (mañana / tarde) 2 = ...de noche? 3 = ...de otro tipo? (rotativo, día y noche, guardias con franco)
PP07K	N (1)	¿Cuando cobra... 1 = ...le dan recibo con sello / membrete / firma del empleador? 2 = ...le dan un papel / recibo sin nada? 3 = ...entrega una factura? 4 = ...no le dan ni entrega nada? 5 = no cobra, es trabajador sin pago / ad honorem

• Ingresos de la ocupación principal de los asalariados

PP08D1	N (10)	¿Cuánto cobró por ese mes por estos conceptos? **Monto total de sueldos / jornales, salario familiar, horas extras, otras bonificaciones habituales y tickets, vales o similares**
PP08D4	N (10)	Por el mes de ... (mes) ... ¿cobró ... **monto percibido en tickets?**
PP08F1	N (10)	¿Cuánto cobró por ese mes de ... (mes) ... **monto en pesos cobrado por comisión por venta / producción?**
PP08F2	N (10)	¿Cuánto cobró por ese mes de ... (mes) ... **monto en pesos cobrado por propinas?**
PP08J1	N (6)	¿Cuánto cobró por ese mes de ... (mes) ... **monto aguinaldo?**
PP08J2	N (6)	¿Cuánto cobró por ese mes de ... (mes) ... **monto otras bonificaciones no habituales?**
PP08J3	N (6)	¿Cuánto cobró por ese mes de ... (mes) ... **monto retroactivo?**

• Movimientos interurbanos (sólo para ocupados)

PP09A	N (1)	Solo ocupados de: Ciudad Autónoma de Buenos Aires y partidos del GBA. En su ocupación (o en la de más horas), ¿trabaja .... en... 1 = Ciudad de Buenos Aires? 2 = partidos del GBA? 3 = ambos? 4 = otro lugar?

PP09A_ESP	C (90)	Especificar: contiene la descripción de otro lugar Solo ocupados de: Posadas, Formosa, Corrientes, Resistencia, Santa Fe, Paraná y Neuquén.
PP09B	N (1)	En su ocupación (o en la de más horas), ¿trabaja en esta ciudad? 1 = Si 2 = No
PP09C	N (1)	¿Dónde trabaja .....? 1 = en otro lugar de esta provincia 2 = en otra provincia 3 = en otro país
PP09C_ESP	C (90)	Descripción de otro lugar (según pregunta PP09C)
• Desocupado		
PP10A	N (1)	¿Cuánto hace que está buscando trabajo? 1 = ...menos de 1 mes 2 = ...de 1 a 3 meses 3 = ...más de 3 a 6 meses 4 = ...más de 6 a 12 meses 5 = ...más de 1 año
PP10C	N (1)	¿Durante ese tiempo hizo algún trabajo/changa? 1= Sí 2= No
PP10D	N (1)	¿Ha trabajado alguna vez? 1= Sí 2= No
PP10E	N (1)	¿Cuánto tiempo hace que terminó su último trabajo/changa...? 1 = ...menos de 1 mes? 2 = ...de 1 a 3 meses? 3 = ...más de 3 a 6 meses? 4 = ...más de 6 a 12 meses? 5 = ...más de 1 a 3 años? 6 = ...más de 3 años?
• Desocupados con empleo anterior: última ocupación / changa (finalizada hace 3 años o menos)		
PP11A	N (1)	¿El negocio / empresa / institución / actividad en la que trabajaba era 1 = ...estatal? 2 = ...privada? 3 = ...de otro tipo?
PP11B_COD	N (5)	A qué se dedicaba o qué producía ese negocio / empresa / institución?

		(Ver Clasificador de Rama de Actividad, CAES-Mercosur)
PP11B1	N (1)	Si prestaba servicios domésticos en hogares particulares 1 = casa de familia
• ¿Cuánto tiempo trabajó allí?		
PP11B2_MES	N (2)	meses
PP11B2_ANO	N (2)	años
PP11B2_DIA	N (2)	días
PP11C	N (2)	¿Cuántas personas, incluido... trabajaban allí en total? 1 = 1 persona 2 = 2 personas 3 = 3 personas 4 = 4 personas 5 = 5 personas 6 = 6 a 10 personas 7 = 11 a 25 personas 8 = 26 a 40 personas 9 = 41 a 100 personas 10 = 101 a 200 personas 11 = 201 a 500 personas 12 = más de 500 personas 99 = Ns/Nr
PP11C99	N (1)	Para los que responden PP11c=99 1 = hasta 5 2 = de 6 a 40 3 = más de 40 9 = Ns/Nr
PP11D_COD	(5)	¿Cómo se llamaba la ocupación que tenía? (Ver Clasificador Nacional de Ocupaciones, CNO, versión 2001)
• ¿Cuánto tiempo seguido estuvo trabajando en ese lugar...?		
PP11G_ANO	N (2)	años
PP11G_MES	N (2)	meses
PP11G_DIA	N (2)	días

PP11L	N (1)	¿Cuál fue la razón principal por la que dejó esa actividad? 1= falta de clientes / clientes que no pagan 2= falta de capital / equipamiento 3= trabajo estacional 4= tenía gastos demasiado altos 5= otras causas laborales (especificar) 6= jubilación / retiro 7= causas personales (matrimonio, embarazo, cuidado de hijos o familiar, estudio, enfermedad).
PP11L1	N (1)	¿Ese trabajo era... 1 = ...una changa, trabajo transitorio, por tarea u obra, suplencia, etc.? 2 = ...un trabajo permanente, fijo, estable, de planta, etc.? 3 = Ns/Nr
PP11M	N (1)	¿Ese trabajo era... 1 = ...un plan de empleo? 2 = ...un período de prueba? 3 = ...otro tipo de trabajo?
PP11N	N (1)	¿En ese trabajo le hacían descuento jubilatorio? 1 = Sí 2 = No 9 = Ns/Nr
PP11O	N (2)	¿Cuál fue la razón principal por la que dejó ese trabajo? 1 = despido / cierre (quiebra / venta / traslado de la empresa, reestructuración o recorte de personal/ falta de ventas o clientes) 2 = retiro voluntario del sector público 3 = jubilación 4 = fin de trabajo temporario / estacional 5 = le pagaban poco / no le pagaban 6 = malas relaciones laborales / malas condiciones de trabajo (insalubre, cambios de horarios, etc.) 7 = renuncia obligada/pactada 8= otras causas laborales (especificar) 9= razones personales (matrimonio, embarazo, cuidado de hijos o familia, estudio, enfermedad).
PP11P	N (1)	¿Cerró la empresa? (Para quienes responden en pp11o=1) 1 = Sí 2 = No 9 = Ns/Nr

PP11Q	N (1)	¿Fue la única persona que quedó sin trabajo? 1 = Sí 2 = No 9 = Ns/Nr
PP11R	N (1)	¿Le enviaron telegrama? 1 = Sí 2 = No
PP11S	N (1)	¿Le pagaron indemnización? 1 = Sí 2 = No
PP11T	N (1)	¿Está cobrando seguro de desempleo? 1 = Sí 2 = No
• Ingresos de la ocupación principal		
P21	N (10)	Monto de ingreso de la ocupación principal
DECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal del total EPH
IDECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal del interior EPH
RDECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal de la región
GDECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal del conjunto de aglomerados de 500.000 y más habitantes
PDECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal del conjunto de aglomerados de menos de 500.000 habitantes
ADECOCUR	C (2)	Nº de decil de ingreso de la ocupación principal del aglomerado
PONDIIO	N (6)	Ponderador del ingreso de la ocupación principal
• Ingreso de otras ocupaciones		
Tot_p12	N (12)	Monto de ingreso de otras ocupaciones (incluye ocupación secundaria, ocupación previa a la semana de referencia, deudas/ retroactivos por ocupaciones anteriores al mes de referencia, etc.)
• Ingreso total individual		
p47T	N (10)	Monto de ingreso total individual (sumatoria ingresos laborales y no laborales - ver **Anexo I**)
DECINDR	**C (2)**	**Nº de decil de ingreso total individual del total EPH**

IDECINDR	C (2)	Nº de decil de ingreso total individual del interior EPH
RDECINDR	C (2)	Nº de decil de ingreso total individual de la región
GDECINDR	C (2)	Nº de decil de ingreso total individual del conjunto de aglomerados de 500.000 y más habitantes
PDECINDR	C (2)	Nºde decil de ingreso total individual del conjunto de aglomerados de menos de 500.000 habitantes
ADECINDR	C (2)	Nº de decil de ingreso total individual del aglomerado
PONDII	N (6)	Ponderador para ingreso total individual
• Ingresos no laborales		
V2_M	N (6)	Monto del ingreso por jubilación o pensión
V3_M	N (6)	Monto del ingreso por indemnización por despido
V4_M	N (6)	Monto del ingreso por seguro de desempleo
V5_M	N (6)	Monto del ingreso por subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.
V8_M	N (6)	Monto del ingreso por alquiler (vivienda, terreno, oficina, etc.) de su propiedad
V9_M	N (6)	Monto del ingreso por ganancias de algún negocio en el que trabajó
V10_M	N (6)	Monto del ingreso por intereses o rentas por plazos fijos / inversiones
V11_M	N (6)	Monto del ingreso por beca de estudio
V12_M	N (6)	Monto del ingreso por cuotas de alimentos o ayuda en dinero de personas que no viven en el hogar
V18_M	N (6)	Monto del ingreso por otros ingresos en efectivo (limosnas, juegos de azar, etc.)
V19_AM	N (6)	Monto del ingreso por trabajo de menores de 10 años
V21_M	N (6)	Monto del ingreso por aguinaldo
T_Vi	**N (12,4)**	**Monto total de ingresos no laborales**
• Ingreso total familiar		
ITF	N (12,2)	Monto del ingreso total familiar
DECIFR	C (2)	Nº de decil de ingreso total familiar del total EPH

IDECIFR	C (2)	Nº de decil de ingreso total familiar del interior EPH
RDECIFR	C (2)	Nº de decil de ingreso total familiar de la región
GDECIFR	C (2)	Nº de decil de ingreso total familiar del conjunto de aglomerados de 500.000 habitantes
PDECIFR	C (2)	Nº de decil de ingreso total familiar del conjunto de aglomerados de menos de 500.000 habitantes
ADECIFR	C (2)	Nº de decil de ingreso total familiar del aglomerado
• Ingreso per cápita familiar		
IPCF	N (12,2)	Monto del ingreso per cápita familiar
DECCFR	C (2)	Nº de decil de ingreso per cápita familiar del Total EPH
IDECCFR	C (2)	Nº de decil de ingreso per cápita familiar del interior EPH
RDECCFR	C (2)	Nº de decil de ingreso per cápita familiar de la región
GDECCFR	C (2)	Nº de decil de ingreso per cápita familiar del Conjunto de aglomerados de 500.000 y más habitantes
PDECCFR	C (2)	Nº de decil de ingreso per cápita familiar del Conjunto de aglomerados de 500.000 habitantes
ADECCFR	C (2)	Nº de decil de ingreso per cápita familiar del aglomerado
PONDIH	N (6)	Ponderador del ingreso total familiar y del ingreso per cápita familiar, para hogares

Apéndice 2: Análisis de significatividad de coeficientes.

OLS Regression Results

Dep. Variable:	DESERTO	R-squared:	0.093
Model:	OLS	Adj. R-squared:	0.093
Method:	Least Squares	F-statistic:	272.6
No. Observations:	18594	Prob (F-statistic):	0.00
Df Residuals:	18586	Log-Likelihood:	4104.7
Df Model:	7	AIC:	-8193.
Covariance Type:	nonrobust	BIC:	-8131.

	coef	std err	t	P>|t|	[0.025	0.975]
constante	0.2257	0.052	4.319	0.000	0.123	0.328
SEXO_VARON	0.0097	0.003	3.401	0.001	0.004	0.015
ESTADO_OCUPADO	0.1363	0.052	2.601	0.009	0.034	0.239
ESTADO_DESOCUPADO	0.0636	0.054	1.183	0.237	-0.042	0.169
ESTADO_INACTIVO	-0.1343	0.052	-2.586	0.010	-0.236	-0.033
JEFE_TRABAJA	-0.0089	0.004	-2.543	0.010	-0.016	-0.002
JEFE_NIVEL_ED	-0.0138	0.001	-13.949	0.000	-0.016	-0.012
DECIL_INGRESO_FAM	-0.0013	0.000	-3.119	0.002	-0.002	-0.000

A excepción del estado de empleo “Desocupado”, se considera que los coeficientes son estadísticamente significativos.