Master in Management + Analytics

Repitencia en el primer año de secundarias públicas de la Ciudad de Buenos Aires – Un análisis de Machine Learning

Autoría: Mayorca, Guillermina
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

REPITENCIA EN EL PRIMER AÑO DE SECUNDARIAS PÚBLICAS DE LA CIUDAD DE BUENOS AIRES – UN ANÁLISIS DESDE MACHINE LEARNING

TESIS

Guillermina Mayorca

Junio 2025

Tutor: Ramiro Gálvez

Resumen

El abandono escolar en el nivel secundario es un desafío estructural del sistema educativo argentino, con solo el 73% de los estudiantes logrando completar este nivel. La repitencia se ha identificado como un factor crítico asociado a la deserción, por lo que su detección temprana puede contribuir a mejorar las trayectorias educativas. Este estudio propone el desarrollo de un modelo de aprendizaje automático para predecir la repitencia en el primer año del nivel secundario, utilizando datos académicos, conductuales y contextuales de los estudiantes. En particular, se evalúa el impacto del Informe Puente Primario Secundario (PPS) en la predicción de trayectorias escolares.

Los resultados indican que el modelo logra identificar patrones asociados a la repitencia con precisión, incluso en ausencia de variables socioeconómicas detalladas. Se destacan dos aplicaciones principales: la planificación pedagógica al inicio del ciclo lectivo y la intervención temprana a lo largo del año escolar. Asimismo, se identifican oportunidades de mejora mediante la incorporación de más cohortes y la estandarización de datos provenientes de instituciones privadas.

Este estudio demuestra el potencial del uso de datos educativos y modelos predictivos en la toma de decisiones escolares. Futuras investigaciones podrían optimizar los modelos mediante nuevas variables y metodologías, así como evaluar el impacto real de su implementación en la dinámica escolar.

Abstract

School dropout at the secondary level is a structural challenge in the Argentine education system, with only 73% of students completing this stage. Grade repetition has been identified as a critical factor associated with dropout, making its early detection crucial for improving educational trajectories. This study proposes the development of a machine learning model to predict grade repetition in the first year of secondary education, using students' academic, behavioral, and contextual data. In particular, it evaluates the impact of the Puente Primario Secundario (PPS) report on predicting school trajectories.

The results indicate that the model identifies patterns associated with grade repetition, even in absence of detailed socioeconomic variables. Two key applications stand out: pedagogical planning at the beginning of the school year and early intervention throughout the academic cycle. Additionally, the study identifies opportunities for improvement, such as incorporating data from more student cohorts and standardizing information from private institutions.

This study highlights the potential of using educational data and predictive models in school decision-making. Future research could refine the models by incorporating new variables and methodologies, as well as assessing the real impact of its implementation on school dynamics.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

El nivel secundario es el último tramo de la educación obligatoria en Argentina desde la sanción de la Ley de Educación Nacional N° 26.206 en 2006. A pesar de que, según datos de la Encuesta Permanente de Hogares (EPH), la totalidad de los adolescentes accede a este nivel, solo el 73% logra finalizarlo (Torre, Paparella, & D'Alessandre, 2023). Esta problemática refleja una de las principales debilidades del sistema educativo: la dificultad para garantizar trayectorias escolares completas y exitosas para todos los estudiantes.

El rendimiento académico es un factor determinante en la permanencia dentro del sistema educativo. Tal como resaltan Torres, Acevedo, & Gallo, (2015):

> *"La razón por la que el rendimiento académico es un factor de alerta en la deserción es porque puede ser una forma de identificar la población más vulnerable frente al abandono de estudios, puesto que es probable que un estudiante que posea bajas calificaciones, pobre desempeño en las clases y en general en el desarrollo de sus actividades académicas, posea menos elementos a los cuales aferrarse ante un obstáculo y decida desertar".*

En este sentido, la repitencia es identificada como una variable altamente relacionada con un posterior abandono del sistema educativo (Gamboa-Cruzado, et al., 2023). La repitencia y el posterior abandono escolar representan desafíos estructurales dentro del sistema educativo argentino, afectando no solo la formación académica de los estudiantes, sino también sus oportunidades futuras de desarrollo personal y profesional (Avalis & Caro, 2024). En el caso uruguayo, Salas, Vigorito, & Failache (2015) identifican que 53% de los estudiantes desertores habían repetido previamente, y 17% de los repitentes termina abandonando el sistema educativo.

La importancia de poder prevenir la repitencia, y como consecuencia intentar reducir la deserción escolar, radica en el hecho de que la finalización de la trayectoria educativa obligatoria tiene un gran impacto en el desarrollo personal, profesional y laboral de las personas. Tal como se expone en un informe de CIPPEC publicado en agosto de 2023, la posesión de un título de nivel secundario permite el acceso a un salario un 37% mayor para los jóvenes que el que obtiene una persona de su misma edad pero que no completó los estudios obligatorios.

Es por esto que una detección temprana de perfiles propensos a la repitencia puede generar impacto en tanto reduzca los rezagos en la trayectoria educativa y así desincentive el abandono escolar. Alineados con este análisis, diversos proyectos buscan la implementación de sistemas

que permitan detectar desvíos en la trayectoria escolar. Los mismos toman como variables principales la sobreedad, el rendimiento, el comportamiento, y el nivel socioeconómico de la persona y su familia.

En este contexto, los Sistemas de Alerta Temprana (SAT) emergen como una herramienta clave para identificar a tiempo a los estudiantes en riesgo y diseñar intervenciones personalizadas que permitan reducir los índices de repitencia y deserción. Estas herramientas buscan proveer a los docentes de un análisis predictivo que permita el monitoreo y acción sobre estudiantes en distintas situaciones, siendo la prevención de deserción el uso más común. Sin embargo, su eficacia depende de la capacidad de las instituciones educativas para integrar estos sistemas dentro de un enfoque más amplio de mejora educativa, que incluya políticas de apoyo pedagógico, acompañamiento socioemocional y estrategias de inclusión escolar.

Actualmente, los sistemas de gestión de las escuelas secundarias públicas de la Ciudad de Buenos Aires recopilan una considerable cantidad de información sobre la trayectoria académica de los estudiantes. Estos datos cumplen un rol fundamental en la gestión administrativa, facilitando la emisión de títulos, boletines e informes. Sin embargo, su uso para generar intercambios y retroalimentación acerca de la trayectoria del estudiante dentro del sistema educativo sigue siendo limitado. Aunque docentes y alumnos aportan información de manera constante, no existen hasta el momento herramientas que les devuelvan análisis o *insights* que contribuyan a mejorar los procesos de enseñanza y aprendizaje. Esto representa una oportunidad para fortalecer el uso de los datos en beneficio de la comunidad educativa.

Dentro de la literatura, la elaboración acerca de los Sistemas de Alerta Temprana se alimenta de datos vinculados a lo socioeconómico y al rendimiento escolar. Sin embargo, no incorporan información sobre habilidades no cognitivas, a pesar de que estas han demostrado ser determinantes clave en la trayectoria educativa de los estudiantes (Hsin & Xie, 2011).

En este sentido, la Ciudad de Buenos Aires cuenta con una herramienta potencialmente valiosa: el informe Puente Primario Secundario (PPS). El PPS es un informe que se realiza para los estudiantes de séptimo grado de instituciones públicas donde se recopila información acerca de la manera que tienen los estudiantes de relacionarse entre sí, con sus docentes, la actitud que adoptan frente a las instancias de aprendizaje y la manera en que sus familias se vinculan con el colegio.

En este contexto, el PPS representa una oportunidad para entender de manera tangible el impacto de dimensiones como la actitud hacia el aprendizaje, la capacidad de vinculación y la organización en las tareas en la trayectoria escolar de los estudiantes de la Ciudad.

El objetivo de este proyecto de investigación es desarrollar, en base a información académica, conductual y de contexto de los estudiantes, un modelo basado en aprendizaje automático que identifique patrones comunes entre los estudiantes que repiten el primer año de educación secundaria.

Se focaliza el análisis en la Ciudad Autónoma de Buenos Aires dado que cuenta con información sistematizada sobre aspectos no cognitivos de los estudiantes. Esta disponibilidad de datos específicos permite explorar con mayor profundidad el fenómeno de la repitencia y su vinculación con variables como la actitud frente al aprendizaje, el vínculo con docentes y compañeros, y el acompañamiento familiar.

Este modelo busca evaluar el poder predictivo del informe Puente Primario Secundario elaborado para los estudiantes de séptimo grado del sistema educativo público de la Ciudad. Se espera que la información que el mismo provee acerca de la actitud y el comportamiento de los estudiantes sea influyente en la capacidad de los modelos predictivos para entender si un estudiante repetirá o no. A fin de poder incorporar esta información es que el modelo se centrará en analizar a la cohorte 2022 del nivel primario público de la Ciudad Autónoma de Buenos Aires, la primera en tener la información de PPS sistematizada.

Previo a continuar con el desarrollo cabe realizar una mención acerca de los debates alrededor de la repitencia. Si bien durante el periodo analizado y en la actualidad la repitencia existe en el nivel secundario de la Ciudad de Buenos Aires, hay otros distritos del país en donde esta figura de la trayectoria escolar se ha eliminado o está al menos siendo cuestionado.

Las modificaciones adoptadas son diversas, pero en general tienden a reemplazar la repitencia del año completo por esquemas similares a los del nivel universitario, donde el estudiante debe recursar únicamente las materias desaprobadas. Estos cambios se llevan a cabo como parte de una estrategia de flexibilización que busca aumentar la capacidad de la escuela secundaria para lograr la permanencia en la escuela de la diversidad de perfiles que se encuentran hoy en día en este nivel (Steinberg, Tiramonti, & Ziegler, 2018).

Un ejemplo concreto es el de la Nueva Escuela Secundaria Rionegrina (ESRN), implementada en escuelas públicas de Río Negro desde 2017. Este modelo propone una reorganización del

régimen de promoción y una reestructuración curricular basada ya no en materias sino en áreas de conocimiento como ciencias sociales y humanidades, educación científica y tecnológica, entre otras. Además establece que salvo en primer año, donde no existe la posibilidad de repetir, es un comité académico —compuesto por docentes, directivos, estudiantes y familias— el encargado de decidir si un estudiante repite o no. [1]

Estas experiencias reflejan que, en distintas jurisdicciones del país, se están explorando alternativas al régimen tradicional de repitencia, lo que permite contextualizar este trabajo en un escenario educativo en movimiento, atravesado por debates y transformaciones recientes.

El presente trabajo se divide en 5 secciones. La sección 2 consta de una explicación de los datos utilizados y las transformaciones aplicadas para poder emplearlos dentro de los modelos. En la sección 3 se presenta un análisis exploratorio de los mismos y una primera aproximación los comportamientos esperados por la bibliografía. La sección 4 corresponde al desarrollo acerca del modelo elegido, sus particularidades y cómo fue implementado en este caso. En la sección 5 se encuentran los resultados de dicha experimentación y la interpretación de los mismos. Finalmente en la sección 6 se presentan conclusiones de la investigación y cuáles son las posibilidades de mejora del modelo.

2. Datos

Esta tesis se basa en datos proporcionados por el Ministerio de Educación de la Ciudad de Buenos Aires, correspondientes a 13.737 estudiantes que completaron sus estudios de nivel primario en el sistema público de la Ciudad durante el año 2022 y continuaron su educación en el nivel secundario público durante los años 2023 y 2024. Este grupo de estudiantes es especialmente relevante, ya que reúne a todos aquellos que cuentan con el informe PPS. Dicho informe, elaborado para los estudiantes de séptimo grado de las instituciones públicas, se utiliza para identificar el tránsito entre la educación primaria y secundaria, y se considera un insumo fundamental para el análisis de las trayectorias educativas.

La decisión de no incorporar instituciones de educación privada en el modelo se debe principalmente a la falta de uniformidad y periodicidad en la información que las escuelas privadas reportan al Ministerio de Educación. Esta disparidad en el reporte de datos dificulta la comparación y la integración de estos en el análisis. Además, las escuelas primarias privadas aún

[1] https://educacion.rionegro.gov.ar/seccion/24

no han implementado la figura del Puente Primario Secundario de la misma manera que las instituciones públicas, lo que limita aún más la disponibilidad de información clave para el modelo. En consecuencia, una de las variables más relevantes del estudio no se encuentra disponible para las instituciones privadas, lo que ha generado la exclusión de este sector educativo en el análisis de las trayectorias.

En relación con los datos sobre el contexto socioeconómico de los estudiantes y sus familias, si bien se reconoce su relevancia, se optó por no incorporarlos al modelo debido a limitaciones en su calidad y disponibilidad. Se consultaron diversos repositorios que reúnen información sobre ingresos y condiciones de vivienda, construidos a partir de fuentes como la plataforma de gestión educativa, reportes de becas y programas de formación del Gobierno de la Ciudad. Estos datos incluyen el salario declarado del grupo familiar, subsidios recibidos y tipo de vivienda. Sin embargo, presentan un alto nivel de nulidad y no están disponibles para la totalidad de los estudiantes, ya que no provienen directamente del sistema de matrícula. Ante la elevada proporción de datos faltantes, se consideró que cualquier estrategia de imputación podría distorsionar la realidad de los casos, por lo que se decidió excluir estas variables del modelo.

2.1. Fuentes de datos

Los datos que se utilizan en este trabajo provienen del Ministerio de Educación de la Ciudad de Buenos Aires, como se mencionó con anterioridad. Los datasets que se presentarán a continuación no tienen todos el mismo origen. Algunos de ellos corresponden a datos de los estudiantes que se crean a medida que avanzan en su recorrido educativo y otros son datos que se solicitan en las inscripciones. La multiplicidad de orígenes implica que, para cada uno de estos conjuntos, puede haber un distinto nivel de completitud de datos para cada uno de los estudiantes a evaluar en este trabajo.

Los datos de matrícula, pases de institución y calificaciones provienen de AprendeBA (ex MiEscuela)[2], sistema de gestión escolar que utilizan los colegios para poder registrar la información de sus estudiantes. Ese mismo origen tienen los datos del Puente Primario Secundario, módulo que tuvo su primer ciclo lectivo de uso en 2022. Los datos de responsables provienen del sistema “Inscripciones en Línea”, y los de servicios de la nómina de cada uno de esos sistemas de becas y subsidios.

[2] https://buenosaires.gob.ar/educacion/aprendeba

2.1.1. Matrícula

El conjunto de datos de matrícula posee información para los ciclos lectivos 2022, 2023 y 2024 del sistema de educación público de la Ciudad. Esta información refleja la última escuela en la que se encontraron los alumnos en cada uno de estos periodos, y contiene datos únicamente para aquellos estudiantes que completan el calendario académico. Es decir, si un estudiante abandona sus estudios durante el año, el mismo no se encontrará presente en este dataset.

Cada uno de los registros representa a un estudiante para un ciclo lectivo en específico, con las siguientes columnas:

Tabla 1 - Variables de la matrícula

Variable	Tipo	Observaciones
ciclo_lectivo	Numérica	Indica el año escolar de la información.
id_miescuela	Numérica	Identificador único del estudiante.
documento	Caracter	Número de identificación de la persona, puede ser aquel otorgado por RENAPER u otras entidades.
genero	Caracter	Género reportado en la matrícula escolar
anio	Caracter	Grado que cursa el estudiante.
turno	Caracter	Turno al que asiste el estudiante.
jornada	Caracter	Tipo de jornada del curso al que asiste el estudiante.
capacidad_maxina	Numérica	Cantidad máxima de estudiantes permitidos en el curso.
cueanexo	Numérica	Identificador único del establecimiento al que asiste el estudiante.
dependencia_funcional	Caracter	Tipo de institución a la que asiste el estudiante. Ej: técnica, media, artística, etc.
distrito_escolar	Numérica	Distrito escolar en el cual se encuentra la institución a la que asiste el estudiante. Determina el supervisor que tendrá el establecimiento.
comuna	Numérica	Comuna de la ciudad en que se encuentra el establecimiento.

barrio	Caracter	Barrio de la ciudad en que se encuentra el establecimiento.
repite	Numérica	Indica si en el ciclo lectivo del registro el estudiante realiza el mismo curso que realizó en el ciclo lectivo inmediatamente anterior.
sobreedad	Numérica	Indica si el estudiante tiene, al 30 de junio, mayor edad que la esperada para el grado que cursa.

Sobre las columnas repite y sobreedad cabe realizar una aclaración. En tanto el estudiante repita curso, su valor de sobreedad será 1 desde ese momento en adelante entendiendo que de allí en más siempre estará por encima de la edad esperada para el curso al día 30 de junio. La columna repite valdrá 1 solo en el ciclo lectivo donde el estudiante realice el mismo año que en el ciclo lectivo inmediatamente anterior.

La elección de estas variables deriva de una investigación previa y revisión bibliográfica que las identifica como relevantes a la hora de estudiar las trayectorias educativas con rezago.

En el caso de *género*, la misma se incorpora debido a que en la literatura se identifica una mayor tendencia a la repitencia entre varones debido a factores socioemocionales y de conducta. Además, particularmente en contextos económicos adversos, la inversión de tiempo y recursos en la educación puede ser vista como un costo frente al beneficio inmediato del trabajo. Este efecto es mayor en el caso de los hombres, tanto por las expectativas sociales de la división del trabajo así como por la mayor oportunidad que tiene este grupo de acceder a oportunidades laborales informales relacionadas al trabajo de fuerza (Méndez-Errico & Ramos, 2015 ; Llambí, Messina, & Perera, 2009).

La *sobreedad*, otra característica personal del estudiante, fue incorporada al ser variable clásica de los sistemas de alertas tempranas. Este es un predictor esencial en tanto permite dar cuenta de ciertas dificultades de aprendizaje previas y suele estar relacionada con mayores probabilidades de repitencia (Boado & Fernández, 2010).

En lo que hace a las características de la escuela a la que asiste el estudiante, las variables de *turno* y *jornada* son importantes en tanto se postula que aquellas escuelas con una jornada que requiera que el estudiante pase más tiempo en las aulas tienen menores tasas de repitencia (Amarante, Colafranchesi, & Vigorito, 2011).

En esta misma línea, el *cueanexo* y la *dependencia funcional* son importantes en tanto resaltan diferencias entre colegios y formas de aprendizaje que tienen un correlato en la vinculación de los estudiantes con las instituciones (Dari, Cervini, & Quiroz, 2021).

La variable de *capacidad máxima* fue incluida para comprender si, tal como resaltan Vries, León, Romero, & Hernández (2011) para el caso de programas universitarios, un mayor número de estudiantes por profesor tiene un impacto en el rezago y/o abandono de la trayectoria educativa.

En última instancia, las variables *comuna*, *distrito escolar* y *barrio* se incluyen con el fin de aproximar el contexto socioeducativo y geográfico de los estudiantes. En concordancia con el argumento presentado acerca del género de los estudiantes, se espera que áreas con mayores niveles de pobreza posean una mayor concentración de casos de repitencia (Duncan, et al., 2007; Torres, Acevedo, & Gallo, 2015).

Dentro de los modelos se utilizan las variables correspondientes al ciclo lectivo 2022 y 2023. Del ciclo lectivo 2024 se toma únicamente la variable 24_repite dado que es la variable de interés a predecir en este trabajo.

Sobre las variables que se utilizaron en este trabajo se aplicó una estrategia de One Hot Encoding para poder hacer a las mismas compatibles con el modelo que se usará luego. Este procedimiento es necesario en tanto la predicción se realiza mediante un modelo de boosting que admite únicamente variables numéricas. Las variables que fueron sometidas a este proceso fueron: *genero, turno, jornada, cueanexo, dependencia_funcional, distrito_escolar, comuna y barrio.*

Así, la estructura de la tabla final utilizada en base a esta información tiene las siguientes columnas para representar esa información:

Tabla 2 - Variables OHE de matrícula

Variable	Tipo	Observaciones
genero_<genero>	Numérica	Vale 1 para el género para el estudiante en la matrícula
<cl>_turno_<tipo_turno>	Numérica	Vale 1 para el turno que le corresponda al estudiante en ese ciclo lectivo y 0 en caso contrario.

<cl>_jornada_<tipo_jornada>	Numérica	Vale 1 para el tipo de jornada que le corresponda al estudiante en ese ciclo lectivo y 0 en caso contrario.
<cl>_dependencia_funcional_<tipo_df>	Numérica	Vale 1 para la dependencia funcional del colegio al que asistió el estudiante en ese ciclo lectivo y 0 en caso contrario.
<cl>_distrito_escolar_<numero_de >	Numérica	Vale 1 para el distrito escolar del establecimiento y 0 en caso contrario.
<cl>_comuna_<numero_comuna >	Numérica	Vale 1 para la comuna del establecimiento y 0 en caso contrario.
<cl>_barrio_<nombre_barrio >	Numérica	Vale 1 para el barrio del establecimiento y 0 en caso contrario.

2.1.2. Puente Primaria Secundaria

El conjunto de datos de Puente Primario Secundario nuclea información del módulo PPS de la plataforma de gestión educativa MiEscuela correspondiente al ciclo lectivo 2022. El mismo fue completado por las escuelas de gestión estatal de la Ciudad. Tal como indica la resolución N° 3930/MEGC/17[3] del Boletín Oficial de la Ciudad de Buenos Aires: "*El Informe correspondiente al Puente Primaria - Secundaria está compuesto por distintas dimensiones, las cuales serán completadas por el docente a cargo del grado del estudiante tomando en consideración el aporte de los docentes curriculares, el coordinador de ciclo, el equipo directivo y la información relevante brindada por los adultos responsables.*"

El informe se divide en los siguientes ejes:

- Actitud: incluye información acerca del comportamiento y compromiso del estudiante con las actividades propuestas por la escuela.
- Convivencia: incluye información acerca del respeto del estudiante a sus pares, su capacidad de resolver conflictos, la manera en que se vincula con adultos y cómo participa en situaciones recreativas.
- Trayectoria: incluye información acerca de la existencia de apoyos a la trayectoria escolar, áreas en las que el estudiante destaca y ajustes que haya necesitado en los contenidos del programa.

[3] https://documentosboletinoficial.buenosaires.gob.ar/publico/PE-RES-MEGC-MEGC-3930-17-ANX.pdf

- Vínculo: incluye información acerca de el o los adultos responsables del estudiante frente a la escuela, la frecuencia con la que los mismos participan ante lo requerido por el establecimiento y el involucramiento en la formación de los alumnos.
- Antecedentes: incluye información acerca de los antecedentes de salud del estudiante.
- Intervenciones: incluye información de las intervenciones del Equipo de Orientación Escolar, en caso de que el estudiante hubiese recibido apoyo de este.
- Jornada: incluye información acerca de las jornadas extracurriculares en las que el estudiante haya participado.

Este tipo de datos son esenciales para evaluar factores contextuales que inciden en el éxito o fracaso escolar. En este sentido, varios estudios demuestran la influencia de factores actitudinales y socioemocionales en el desempeño educativo de los estudiantes. Cardozo, Silveira, & Fonseca (2022) muestran cómo las características personales y la actitud hacia el aprendizaje observable desde la primera infancia impactan en el desempeño académico, lo cual puede contribuir a la predicción de trayectorias de rezago escolar. En el caso de la educación de nivel medio, estas habilidades no académicas son importantes en tanto reflejan la capacidad del estudiante de amoldarse a las demandas cognitivas y de comportamiento de las instituciones de educación formal.

Además, Torres, Acevedo, & Gallo (2015) destacan como la interrelación entre los docentes y alumnos, así como la relación entre alumnos, es un factor de relevancia para entender la repetición escolar. Estos elementos, capturados a través del eje de convivencia, ofrecen una visión más completa del estudiante.

Según Hsin & Xie (2011), las habilidades no cognitivas —como la actitud hacia el aprendizaje, la capacidad para generar vínculos y la regulación emocional— tienen un impacto significativo en las trayectorias escolares de los estudiantes. A diferencia de las habilidades cognitivas, que están más relacionadas con el acceso a la educación y los recursos económicos del hogar, las habilidades no cognitivas son moldeadas por la estructura social y la forma en que los estudiantes interactúan con su entorno pudiendo trascender la barrera económica.

Dado que el PPS capta información sobre estos aspectos actitudinales y relacionales, se espera que su incorporación en el modelo de predicción permita detectar patrones de riesgo de repitencia más allá de los factores tradicionales, como el rendimiento académico o la situación socioeconómica.

Dentro de la información recolectada por el PPS se combinan campos de respuesta cerrada —con valores limitados a un set de posibles respuestas predefinido— y campos de texto libre, que recogen respuestas abiertas sin estandarización previa. A diferencia de las respuestas cerradas, las respuestas de texto libre dependen de la iniciativa de los docentes para ser completadas, lo que introduce un alto grado de variabilidad en cuanto al nivel de detalle, redacción y contenido. Esta falta de uniformidad complica su procesamiento automático y limita la posibilidad de realizar comparaciones consistentes entre estudiantes. La única excepción a esto es el caso del vínculo, donde si bien el docente tiene la oportunidad de completar con texto libre, es más sencillo normalizar las respuestas para definir categorías como madre, padre, abuelo, abuela, tío, tía, hermano, hermana, etc.

La información del PPS que se incluye en el estudio para cada uno de los alumnos tiene las siguientes columnas:

Tabla 3 - Variables procesadas de Puente Primaria Secundaria para el modelo

Variable	Tipo	Observaciones
actitud_como	Caracter	Modalidad de participación (remota, presencial o bimodal).
actitud_logra	Caracter	Indica con qué frecuencia el estudiante logra comprometerse con su aprendizaje reconociendo logros y dificultades.
actitud_cumple	Caracter	Indica con qué frecuencia el estudiante cumple con las tareas/actividades/producciones en los tiempos establecidos.
actitud_trabaja	Caracter	Indica con qué frecuencia el estudiante trabaja de manera colaborativa.
actitud_autonomo	Caracter	Indica con qué frecuencia el estudiante logra ser autónomo en su trabajo en términos de organización y resolución de las actividades/propuestas.
actitud_consulta	Caracter	Indica con qué frecuencia el estudiante consulta, revisa y corrige su trabajo teniendo en cuenta las intervenciones y sugerencias de sus docentes.

actitud_demuestra	Caracter	Indica con qué frecuencia el estudiante demuestra motivación e interés sobre las actividades propuestas y cómo sus aportes evidencian compromiso frente a las tareas.
actitud_participa	Caracter	Indica con qué frecuencia el estudiante participa activamente de las propuestas de actividades de los docentes.
actitud_pedagógica	Caracter	Indica si el estudiante mantiene o no la continuidad pedagógica.
actitud_manifiesta	Caracter	Indica con qué frecuencia el estudiante manifiesta predisposición ante las actividades planteadas.
actitud_puedeorganizarse	Caracter	Indica con qué frecuencia el estudiante puede organizarse y cumplir con las actividades propuestas a través de los diversos formatos.
trayectoria_requirio	Caracter	Indica si el estudiante requirió o no apoyos en su trayectoria.
trayectoria_cuales	Caracter	Indica, en caso de haber necesitado configuraciones de apoyo, cuáles fueron.
trayectoria_interrumpida	Caracter	Indica si el estudiante interrumpió su trayectoria escolar.
trayectoria_requiriopedagogico	Caracter	Indica si el estudiante requirió un plan de aprendizaje diferente respecto al de los demás estudiantes de su curso.
convivencia_acude	Caracter	Indica con qué frecuencia el estudiante acude a un adulto frente a situaciones complejas.
convivencia_respeta	Caracter	Indica con qué frecuencia el estudiante respeta los acuerdos de convivencia en las diferentes instancias.

convivencia_vincula	Caracter	Indica con qué frecuencia el estudiante se vincula de manera respetuosa con los adultos.
convivencia_resuelve	Caracter	Indica con qué frecuencia el estudiante resuelve sus conflictos mediante el diálogo.
convivencia_vinculapares	Caracter	Indica con qué frecuencia el estudiante se vincula con sus pares en los espacios propuestos por el colegio.
convivencia_mantiene	Caracter	Indica con qué frecuencia el estudiante mantiene una comunicación sostenida y sistemática con sus docentes.
vinculo_adulto	Carácter	Indica quién es el adulto responsable del estudiante frente al colegio.
vinculo_acompaña	Carácter	Indica con qué frecuencia el adulto acompaña al estudiante en las instancias requeridas por el colegio.
vinculo_participa	Carácter	Indica con qué frecuencia el adulto participa de las instancias en las que el colegio lo requiere.

Para obtener las columnas mencionadas se trabajó en la apertura de las respuestas del informe que se encontraban en una única columna para cada uno de los ejes en formato JSON. En el caso de las columnas de frecuencia, se optó por traducir las palabras a una escala numérica mediante *ordinal encoding* para su inclusión en el modelo mediante la siguiente regla:

Tabla 4 - Transformación de valores de frecuencia en Puente Primario Secundario

Valor original	Valor resultante
Con poca frecuencia	0
Frecuentemente	1
Siempre	2

En el caso de la variable vinculo_adulto, el procesamiento fue diferente dado que la pregunta en el sistema admitía respuestas de texto libre. En primer lugar, se realizó una reducción de cardinalidad, agrupando las distintas formas en que los profesores identificaron a los

responsables, y luego con esos grupos se procedió a realizar One Hot Encoding obteniendo como resultado las siguientes columnas:

1) vinculo_nadie (no se indica vinculo adulto o se manifiesta que el docente no conoce a sus padres)
2) vinculo_madre
3) vinculo_padre
4) vinculo_hermana
5) vinculo_hermano
6) vinculo_cuñados
7) vinculo_abuela
8) vinculo_abuelo
9) vinculo_tios
10) vinculo_padrino_madrina
11) vinculo_pareja_progenitores
12) vinculo_docentes
13) vinculo_adulto_hogar (en referencia a estudiantes que viven en un hogar de niños y sus referentes son los tutores responsables de dichas instituciones).

2.1.3. Pases de institución

El conjunto de datos de pases concentra información acerca de los eventos en los cuales los estudiantes se cambian de institución educativa durante los ciclos lectivos 2022 y 2023. Esta información resulta importante puesto que los pases de institución indican cierto nivel de discontinuidad en la trayectoria escolar, derivados del proceso necesario de adaptación del estudiante al nuevo entorno, sus compañeros y métodos de enseñanza (Heckman, 2008).

Siendo el foco de este estudio el primer año del nivel secundario, donde los estudiantes en su mayoría tienen una edad de 12 o 13 años, la inestabilidad a nivel del entorno social es particularmente importante. En esta instancia, la interrupción en el proceso de formación de vínculos de amistad o la discontinuación de vínculos ya existentes pueden generar cierto sentimiento de aislamiento y soledad que tenga impacto en la vida del estudiante, y como consecuencia de ello en su rendimiento académico.

El dataset de pases cuenta con información acerca de los estudiantes que piden pases, desde qué establecimiento lo hacen y hacia cual se transfieren. De esta información se conservó, a fines del modelo, un único registro por estudiante para cada ciclo lectivo que indica cuántos pases aprobados obtuvo en el año. Las transformaciones se realizaron mediante un proceso de conteo de cantidad de pases únicos solicitados por los estudiantes y aprobados por sus establecimientos para cada uno de los ciclos lectivos.

Esto permite entender cuántas veces el estudiante efectivamente se transfirió entre escuelas, siendo que los pases pueden también ser rechazados. Es importante destacar que en caso de que el estudiante cambie de curso dentro de la misma institución no se registra como pase, por lo que esa información no es capturada en las columnas 22_cant_pases y 23_cant_pases. Estas variables tampoco contabilizan los cambios de institución que se realicen de un ciclo lectivo al siguiente, es decir que si un estudiante finaliza el ciclo lectivo en una escuela determinada y el siguiente año continua su trayectoria en otra eso no aumenta el valor de las columnas que contabilizan pases.

Tabla 5 - Variables procesadas de pases para el modelo

Variable	Tipo	Observaciones
documento	Caracter	Número de identificación de la persona, puede ser aquel otorgado por RENAPER u otras entidades.
22_cant_pases	Numérica	Cantidad de pases aprobados para el estudiante en el ciclo lectivo 2022.
23_cant_pases	Numérica	Cantidad de pases aprobados para el estudiante en el ciclo lectivo 2023.

2.1.4. Calificaciones

El conjunto de datos de calificaciones reúne información acerca de las notas de los estudiantes en las materias lengua y matemática. La decisión de tomar las calificaciones de estos dos espacios curriculares deriva de dos razones principales. En primer lugar, al tener los secundarios distintas modalidades y planes de estudio, se encuentra que matemática y lengua son materias transversales a estas especificidades. En segundo lugar, diversos estudios coinciden en que los conocimientos impartidos y evaluados en estos espacios curriculares correlacionan fuertemente con trayectorias educativas con rezago e incluso deserción escolar (Duncan, et al., 2007; Boado & Fernández, 2010).

En el caso del ciclo lectivo 2022 se cuenta con las notas de los 4 bimestres, mientras que en el caso del ciclo lectivo 2023 se utilizan únicamente las notas de los bimestres 1 y 2 con el fin de posicionarse a mitad del año para comprender si el estudiante repetirá de curso o no de cara al ciclo lectivo 2024.

Cada registro de la tabla de calificaciones cuenta con 10 columnas: *ciclo_lectivo, id_alumno, n1_mate, n2_mate, n3_mate, n4_mte, n1_lengua, n2_lengua, n3_lengua y n4_lengua*. Cada una de las columnas que inicia con n1, n2, n3 y n4 contiene la calificación del estudiante para la materia indicada en el bimestre señalado por el número, es decir la columna n1_lengua indica la calificación del estudiante para el primer bimestre en la materia lengua, para el ciclo lectivo al que corresponde el registro.

Las calificaciones no siempre son numéricas, de modo que para poder darles un sentido lógico que el modelo pueda interpretar, se realizaron normalizaciones de acuerdo al significado de los valores a nivel pedagógico:

Tabla 6 - Conversión de valores para calificaciones

Valor original	Valor resultante
Sobresaliente (S)	10
Avanzado	9.5
Muy Bueno (MB)	9
Bueno (B)	8
Suficiente	7.5
Regular (R)	7
Insuficiente (I)	5
Promoción Acompañada	5
En Proceso	5

Los estudiantes para los cuales se recolectan calificaciones se dividen en dos grupos: aquellos para los cuales se expidió un informe PPS en el ciclo lectivo 2022, y aquellos que asisten a clase con este primer grupo durante los ciclos lectivos 2022 y 2023.

Para aquellos alumnos que no poseen un PPS, la incorporación de sus notas es temporal y estrictamente enfocada en la imputación de valores faltantes para los alumnos con PPS, y la elaboración de un ranking de desempeño de alumnos dentro de su curso. En ambos casos se conserva el criterio de utilizar únicamente las calificaciones de lengua y matemática del ciclo lectivo 2022 y los primeros dos bimestres de 2023. Estas dos estrategias se abordaron de la siguiente manera:

- Imputación de valores faltantes

Si bien XGBoost puede manejar valores nulos, en este caso se optó por imputar las calificaciones faltantes por la baja cantidad de nulos y porque era necesario tener un promedio de desempeño de todos los estudiantes para poder elaborar el ranking de desempeño relativo dentro de cada sección. El porcentaje de valores nulos en las columnas de calificaciones es menor al 1% de los estudiantes en todos los casos, por lo que la imputación no implica un riesgo significativo de distorsión de los datos.

Para la imputación de nulos en cada una de las columnas de calificación se adoptó una estrategia de tres pasos. En primer lugar, se añadieron las filas faltantes al dataset. Estas "filas faltantes" corresponden a casos de estudiantes para los que no había calificaciones en todo el año, por lo que la tabla de origen no tenía una fila para ese evento. Luego de ello se buscan en el dataset de matrícula dos datos: la sección a la que concurrió el estudiante y el distrito escolar del establecimiento al que asistió, identificadas por las columnas *id_seccion_miescuela* y *distrito_escolar* respectivamente. Una vez teniendo esas filas se procedió a imputar nulos con dos enfoques secuenciales:

Imputación de la media de la sección: se busca, para el estudiante con calificaciones nulas, la sección a la que pertenece (identificada por la columna *id_seccion_miescuela*). En caso de que para ese evento, la columna correspondiente al bimestre y la materia donde el estudiante no tiene una calificación, la sección tuviese a al menos un 25% de sus estudiantes calificados se imputa para la nota faltante la nota media de la sección. El umbral del 25% se establece dado que imputar una media basada en pocas calificaciones puede sesgar el valor tanto positiva como negativamente.

Imputación de la media del distrito escolar: en caso de que al estudiante no se le imputen las calificaciones mediante la estrategia por sección, ya sea porque nadie de la sección fue calificado para ese bimestre en esa materia o porque el porcentaje de estudiantes calificados es menor al 25%, se busca hacerlo mediante la asignación del promedio del distrito escolar al que pertenece el establecimiento. De esta manera se asigna el promedio de calificaciones obtenido por los primeros años o séptimos grados del distrito escolar en ese bimestre, para esa materia y en ese mismo ciclo lectivo.

- Creación de las columnas promedio:

Una vez que todos los estudiantes cuentan con valores de calificación para todos los eventos seleccionados, se procede a calcular el *promedio_mate* y *promedio_lengua* para cada uno en cada ciclo lectivo correspondientes al promedio por estudiante de las calificaciones de cada

materia en el ciclo lectivo. En el caso del ciclo lectivo 2022 con las calificaciones de los 4 bimestres mientras que en el caso de 2023 con las correspondientes a la primera mitad del año.

2.1.4.1 Calificaciones relativas

Una vez completa la imputación de valores faltantes y el cálculo de los promedios se decidió incorporar dos variables extra. Las mismas refieren al desempeño relativo del estudiante en su curso y son de importancia dado que permiten evaluar si el estudiante está teniendo resultados parecidos, superiores o inferiores a los de los demás estudiantes de su curso.

Asimismo, sirve para poder captar si un estudiante con una calificación más cercana a la nota mínima de aprobación se encuentra allí por un mal desempeño personal, porque no hay comprensión generalizada de los temas en el aula (todos sus compañeros de clase tienen desempeños parecidos) o si incluso estando cercano a la nota mínima de aprobación está por encima del resultado promedio de sus compañeros.

$$rk\_mate_{e,s} = rank_{dense,\downarrow}(promedio\_mate_{e,s})$$

$$rk\_lengua_{e,s} = rank_{dense,\downarrow}(promedio\_lengua_{e,s})$$

Estos dos campos indican la posición relativa del estudiante *e* respecto a sus compañeros de clase en estas dos materias. El mismo se calcula ordenando los estudiantes de la sección *s* de mayor a menor según el promedio de cada materia e indicando su posición en base al promedio de calificaciones de dicho espacio curricular. Para hacer que esta columna fuese de utilidad para el modelo a fines de comparar entre secciones, se añade un paso de cálculo en el cual se normalizan los valores entre 0 y 1 dividiendo la posición del estudiante en el ranking entre la cantidad total de estudiantes de su misma sección:

$$N_s = Cantidad\ de\ estudiantes\ en\ la\ sección\ s$$

$$rk\_mate_{e,s}^{norm} = \frac{rk\_mate_{e,s}}{N_s}$$

$$rk\_lengua\ _{e,s}^{norm} = \frac{rk\_lengua_{e,s}}{N_s}$$

Así, un estudiante con un valor de *rk_mate* o *rk_lengua* más cercano a 0 indicara un mejor desempeño dentro de la sección que un estudiante con un valor de ranking más cercano a 1.

Lo que se esperaría según lo expuesto por Dari, Cervini, & Quiroz (2018) es que cursos con mayor heterogeneidad favorezcan la repitencia en tanto el estudiante con menor rendimiento quedaría atrasado en su aprendizaje dado que el ritmo de aprendizaje del curso sería diferente al de su velocidad de aprendizaje.

2.1.5. Servicios

El conjunto de datos de servicios provee datos acerca de las becas y beneficios económicos que reciben los estudiantes en los ciclos lectivos 2022 y 2023. La información refiere a los programas: Becas Alimentar, Becas Media y Boleto estudiantil.

Las Becas Alimentarias[4] corresponden a la cobertura por parte del Gobierno de la Ciudad del valor del servicio de comedor para los estudiantes. Las mismas podrán ser del 50 o del 100 por ciento dependiendo de la situación económica de la familia del solicitante.

El programa Becas Media[5] corresponde a un incentivo otorgado por el Gobierno de la Ciudad a distintos grupos de personas que asistan a las escuelas secundarias públicas de la Ciudad. Sus requisitos son más flexibles que aquellos de las Becas Alimentarias por lo que cuenta con más beneficiarios.

El objetivo de incluir esta información radica en que estos programas requieren el cumplimiento de ciertas condiciones de carácter económico para ser otorgadas. Esto permite aproximar cierto nivel socioeconómico para los estudiantes ante la falta de completitud de sus datos socioeconómicos personales en las plataformas de gestión educativa dentro del Ministerio de Educación.

Las condiciones de cada uno de los programas se listan aquí debajo:

- Becas Alimentarias:
    - Se otorgarán becas totales cuando la suma de los ingresos mensuales del grupo familiar al que pertenecen no supere el equivalente al sueldo mínimo fijado en el convenio para empleados de comercio, multiplicado por 2,5.
    - Se otorgarán medias becas cuando los referidos ingresos superen el tope establecido para las becas totales sin llegar a exceder el equivalente al sueldo mínimo fijado en el convenio para empleados de comercio, multiplicado por 3,5.
    - Cuando el grupo familiar esté integrado por más de un niño que concurra a la escuela pública, el tope de ingresos fijado para determinar la modalidad de la

[4] https://buenosaires.gob.ar/becas-alimentarias
[5] https://boletinoficialpdf.buenosaires.gob.ar/util/imagen.php?idn=125647&idf=1

beca a otorgarse de conformidad a lo dispuesto en el presente artículo, se incrementará en un quince por ciento (15%) por cada niño que se sume.

- Se concederán también becas totales, con prescindencia de los ingresos del grupo familiar, cuando se trate de hijos de personal docente, en el caso en que dicho personal se encuentre a cargo de los alumnos en turno de comedor que almuercen dentro del ámbito de éste.

En caso de que un integrante del grupo familiar se encuentre afectado por una enfermedad crónica que requiera tratamiento continuo o por períodos prolongados, los gastos derivados de dicho tratamiento podrán deducirse del monto del ingreso mensual del grupo familiar que se toma como base para el otorgamiento de becas totales y medias becas. Estas circunstancias deberán ser debidamente acreditadas.

- Becas Media: se otorgan a quienes cumplan al menos uno de los siguientes requisitos:
  - Estudiantes que integren un hogar cuyo ingreso mensual resulte igual o menor a 1, 5 del salario mínimo, vital y móvil.
  - Concurrir a escuelas de reingreso (aquellas apuntadas a jóvenes y adultos que no están en edad de concurrir a escuelas de educación media de modalidad común con menores en edad de escolaridad).
  - Estudiantes que se encuentren en situación de calle.
  - Estudiantes que se encuentren alejados/as de sus familias en hogares convivenciales, o en otros tipos de programas de asistencia.
  - Estudiantes que integren un hogar con Necesidades Básicas Insatisfechas (NBI).
  - Estudiantes que residan en Núcleos Habitacionales Precarios o Transitorios.
  - Estudiantes que sean padres o madres adolescentes, o adolescentes embarazadas.
  - Estudiantes que tengan necesidades especiales o enfermedad graves.
  - Estudiantes que integren un hogar con más de tres miembros en edad escolar, siempre que los ingresos del mismo resulten hasta un 50% por encima de lo establecido en el punto a).
  - Estudiantes que tengan a su padre y/o madre privado de la libertad en el sistema penal, siempre que los ingresos de su hogar resulten hasta un 50 % por encima de lo establecido en el primer punto).
  - Estudiantes que presenten otros indicadores de vulnerabilidad socioeconómica que, según la autoridad de aplicación, resulten suficientes para acreditar tal condición de vulnerabilidad.

Alumnos que reciban otras becas no están excluidos de recibir dinero mediante este programa pero el monto será reducido respecto a aquellos que únicamente perciben este beneficio. En el caso de estudiantes que concurran a escuelas de modalidad técnica o artística pueden percibir, además de este beneficio, otros que se orienten exclusivamente a cubrir costos asociados a los materiales necesarios por la modalidad de educación a la que asisten.

- Boleto estudiantil: es un beneficio que alcanza a los estudiantes de gestión estatal de la Ciudad y a aquellos estudiantes que concurran a escuelas de gestión privada con 100% de subsidio o modalidad de cuota 0. Los mismos deben ser residentes de la Ciudad de Buenos Aires. Lo que implica este programa es el 100% de cobertura en 50 viajes mensuales en el trayecto comprendido entre el domicilio del estudiante y el establecimiento donde asiste a clases. Para obtener este beneficio los estudiantes deben realizar un trámite donde presenten constancia de sus estudios y luego deben obtener una tarjeta SUBE asociada a su persona donde se les aplicará el descuento.

Para mantener estos beneficios, los estudiantes deben ser alumnos regulares de las instituciones a las que asisten. La irregularidad como condición de asistencia implica la quita del subsidio/beneficio.

2.1.6. Responsables

El conjunto de datos referente a responsables concentra información de los adultos que se hacen responsables por los estudiantes incluidos en el modelo. Se estima que el nivel educativo de los progenitores es un buen aproximador del valor que la familia le da a la educación de sus hijos, la habilidad de estos adultos para asistirlos en sus tareas y también, en el caso de las mujeres puntualmente, la actitud frente a su educación Salas, Vigorito, & Failache (2015). Por otro lado Hsin & Xie (2011) resaltan la importancia de los antecedentes educativos familiares en el desarrollo de habilidades cognitivas de los estudiantes que influyen en su desempeño académico.

Estos datos se obtuvieron de lo reportado en el sistema "Inscripción En Línea"[6], herramienta por la cual debe registrarse a todo estudiante que desee ingresar a una escuela pública de nivel medio de la Ciudad. Dentro de este dataset contamos originalmente con las siguientes columnas: *id_miescuela* (identificador del alumno), *doc_alu* (documento del alumno), *doc_resp* (documento del responsable), *nac_resp* (nacionalidad del responsable), *vinculo* (vinculo del

[6] https://buenosaires.gob.ar/educacion/estudiantes/inscripcionescolar

responsable con el estudiante) y *nivel_educativo* (máximo nivel de estudios alcanzado por el responsable). A fines de transformar los datos para volverlos utilizables por el modelo se realizaron las siguientes operaciones:

- Mapeo de nivel educativo: el nivel educativo de los responsables de los estudiantes es señalado como una variable de importancia en la predicción de desempeño escolar. A nivel sistema, la información del nivel educativo se recibe en texto. Para convertirla a números y poder incluirla en el modelo se realizó el siguiente *ordinal encoding*:

Tabla 7 - Transformación de niveles educativos de los responsables

Valor original	Valor resultante
Sin estudios	0
Primario incompleto	1
Primario completo	2
Secundario incompleto	3
Secundario completo	4
Terciario incompleto	5
Terciario completo	6
Universitario incompleto	7
Universitario completo	8
Posgrado	9

- Mapeo de nacionalidades: la nacionalidad del responsable de los estudiantes provee información acerca de la cercanía cultural del entorno del alumno con los códigos sociales de la escuela. Pero fundamentalmente, lo que permite la información acerca de la nacionalidad del responsable es comprender si, ante dudas o consultas al momento de realizar las tareas en el hogar, su responsable podrá comprender las consignas planteadas y ayudar en la resolución. En este sentido se encontró cierto nivel de diversidad, por lo que se redujo la dimensionalidad de la variable utilizando dos grupos: hispanoparlantes y no hispanoparlantes. Estas categorías se capturaron mediante la columna nacionalidad_hispana que toma valor 1 en caso de que el responsable provenga de un país cuya primera lengua es el español y 0 en caso contrario.
- Mapeo de vínculos: al igual que en el caso de los vínculos en el Puente Primario Secundario, se cuenta con distintos tipos de vínculos señalados al momento de la inscripción. En este caso la diferencia radica en que este sistema provee a quien completa el formulario de ciertas opciones para hacerlo. En base a eso se crearon

columnas que toman valor 1 si el vínculo corresponde a esa categoría y 0 en caso contrario. Las mismas son: *resp_nadie, resp_padres, resp_hermanos, resp_abuelos, resp_tios, resp_padrastros, resp_tutores, resp_primos*.

3. Análisis exploratorio de los datos

Esta sección presenta un análisis exploratorio de los conjuntos de datos introducidos en la sección 2, con el objetivo de identificar patrones, relaciones y particularidades relevantes de cara al modelado.

3.1. Repitencias

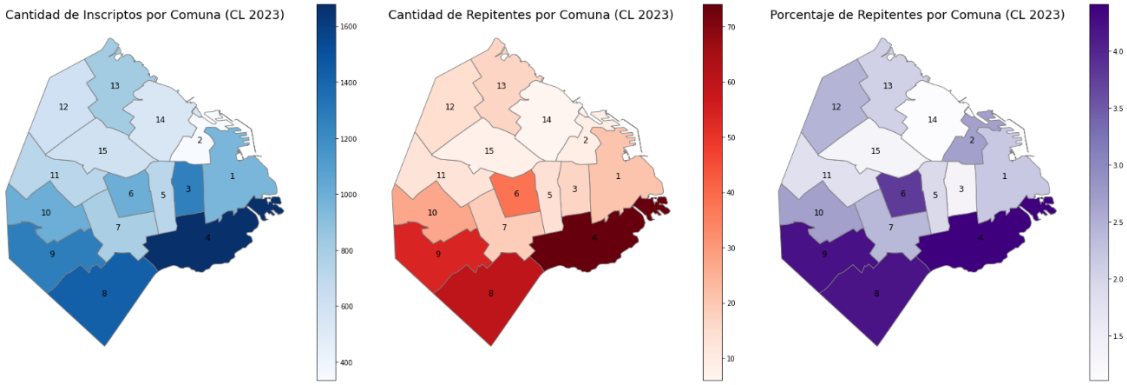
Antes de avanzar con el análisis especifico de patrones en los datos, es pertinente realizar un comentario acerca de lo que se observa para la variable dependiente de este estudio, 24_repite.

Esta variable se construye observando si el estudiante se encuentra matriculado para el ciclo_lectivo 2024 en el mismo curso que lo estuvo para el ciclo_lectivo 2023. 24_repite será 1 para aquellos estudiantes que en 2024 hayan realizado el mismo curso que habían realizado en 2023. Dentro del conjunto de datos usado en este trabajo 24_repite vale 1 para 392 estudiantes, siendo esto un 2,85% del total de la matrícula disponible (13.737).

3.2. Geografía

Se realizó un análisis exploratorio en términos de la distribución de los datos y la información en el territorio de la Ciudad. Se observa una concentración de la cantidad de matriculados en las comunas de la zona sur (a la izquierda), asimismo al realizar un análisis acerca de la distribución de los casos de repitencia se replica este comportamiento (gráfico del centro). Para entender si esta concentración de repitentes deriva de lo observado en primera instancia, más casos de repitencia por una concentración de matriculados en esa zona, o de un patrón de la variable *repite,* se realizó un gráfico que muestra el porcentaje de la matrícula de primer año que repite en 2024 siendo esto lo que los modelos buscaran predecir (a la derecha).

Figura 1 - Distribución de los casos en las comunas de la Ciudad de Buenos Aires donde los estudiantes realizan su primer año de secundaria (2023)



Lo que puede verse es que la concentración de la cantidad de repitentes en la zona sur no deriva del número absoluto de estudiantes y casos de repitencia sino que a nivel porcentual de la propia matrícula, hay una mayor proporción de los estudiantes que repiten en estas comunas. Como puede observarse, el porcentaje de repitentes de las comunas 4, 8 y 9 es de aproximadamente 4,5% mientras que si se observa a nivel ciudad, el porcentaje de la matrícula que repite curso en 2024 respecto a 2023 es del 2,85%.

Como se resaltó anteriormente, la variable a predecir está altamente influenciada por la estructura económica del hogar del estudiante. Si bien no se tienen variables directamente vinculadas a ello, puede pensarse que la manera en que se concentran los casos de repitencia en el espacio geográfico de la Ciudad guarda un vínculo cercano con este punto.

Según los datos declarados por el Instituto de Estadística y Censos de la Ciudad Autónoma de Buenos Aires[7], la relación entre el ingreso promedio per cápita del año 2022 a nivel Ciudad versus las comunas con más casos de repitencia fue la siguiente:

- Comuna 4: 39,84% menor que el promedio de la Ciudad.
    - Comprende los barrios de La Boca, Barracas, Parque Patricios y Nueva Pompeya.
- Comuna 6: 17,5% mayor que el promedio de la Ciudad.
    - Comprende al barrio de Caballito.
- Comuna 8: 58,35% menor que el promedio de la Ciudad.
    - Comprende a los barrios Villa Soldati, Villa Riachuelo y Villa Lugano.
- Comuna 9: 29,28% menor que el promedio de la Ciudad.
    - Comprende los barrios de Mataderos, Liniers y Parque Avellaneda.
- Comuna 10: 18,77% menor que el promedio de la Ciudad.

[7] https://www.estadisticaciudad.gob.ar/eyc/?p=82456

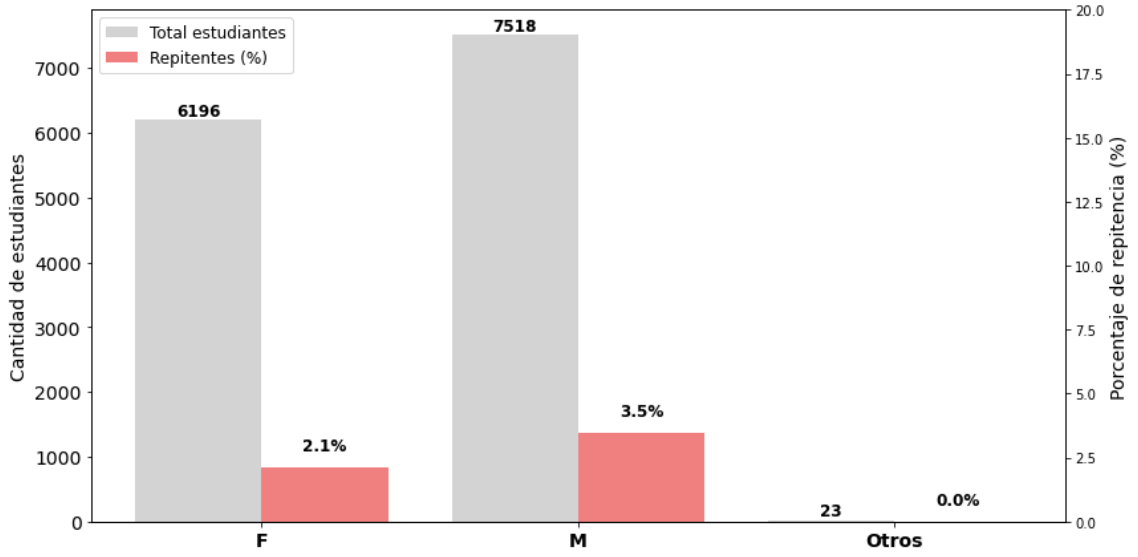
- Comprende los barrios de Villa Real, Monte Castro, Versalles, Floresta, Vélez Sarsfield y Villa Luro.

En este sentido, escapa a esta relación negativa entre condiciones socioeconómicas y repitencia la Comuna 6 donde el ingreso está por encima del promedio de la ciudad.

3.3. Género

Existe en la literatura cierto consenso acerca de la influencia del género en las trayectorias educativas. Lo que se propone allí es que en nivel medio, los varones exhiben mayores tasas de repitencia que sus pares de otros géneros. En el caso uruguayo, por ejemplo, se observa que los estudiantes de género masculino muestran mayores tasas de retraso en sus trayectorias educativas, estando estas altamente asociadas a conflictos en su conducta (Salas, Vigorito, & Failache, 2015).

Figura 2 - Distribución de la matrícula y los casos de repitencia por género



En la *Figura 2* el eje X representa los distintos géneros declarados por los estudiantes. La barra gris, graficada sobre el eje Y de la izquierda, representa la cantidad de estudiantes que reportan ese género. La barra color rojo, graficada en el eje Y de la derecha, representa el porcentaje de los alumnos que declaran ese género para los cuales 24_repite toma valor 1.

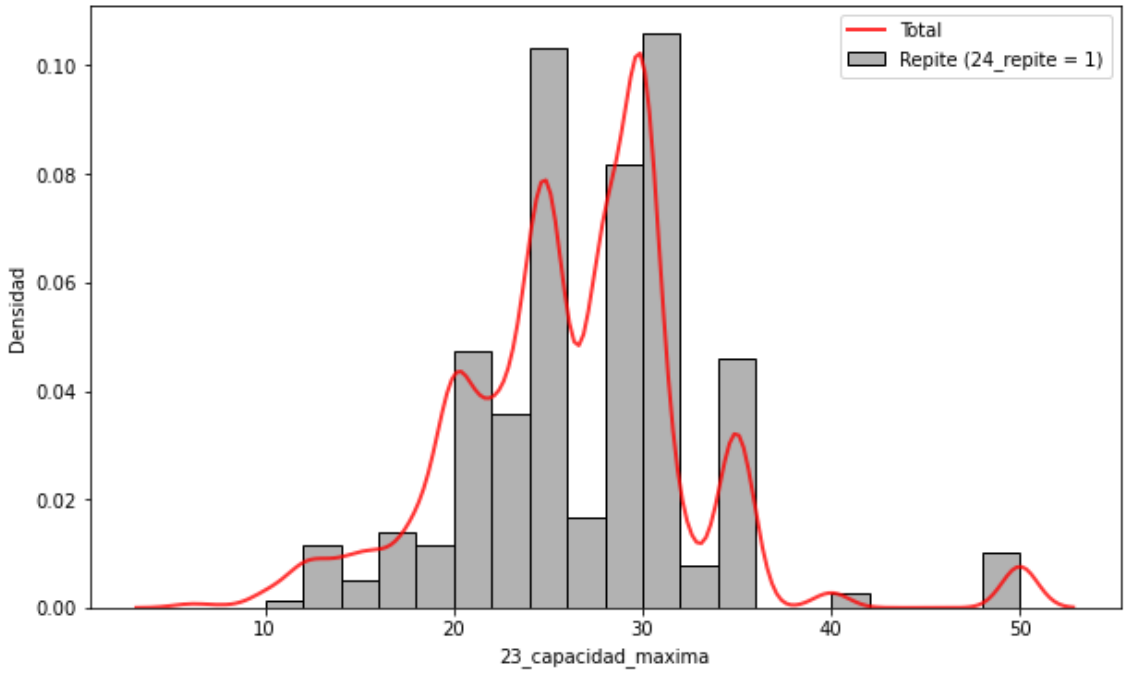
Al analizar la muestra a trabajar, se encuentra que de los 13.737 estudiantes, 7.518 (54,72%) declaran género masculino (M), 6.196 (45,10%) declaran género femenino (F) y 23 declaran como género otros (0,16%).

A nivel de los casos de repitencia, lo que se observa preliminarmente es que parecería existir una predominancia del género masculino dentro de los casos de repitencia por sobre los demás, teniendo una diferencia de 1.1 puntos porcentuales en el porcentaje de repitencia observado respecto a las mujeres de la muestra.

3.4. Cantidad de estudiantes por curso

En la sección 2.1.1, donde se explicaron aquellos datos que se obtendrían de la matrícula para el presente trabajo, se referenció al estudio de programas universitarios de Vries, León, Romero, & Hernández (2011) quienes presentan la hipótesis de que un mayor radio de estudiantes por profesor incrementaría la probabilidad de rezago en la trayectoria escolar.

Figura 3 - Distribución de los casos según capacidad máxima de las secciones



En la Figura 3 se representa la distribución de estudiantes según la capacidad máxima de las aulas en las que toman clases en el año 2023. En este gráfico, la línea roja representa al total de los estudiantes, y las columnas grises a los estudiantes para los cuales 24_repite toma valor 1. El eje X da cuenta de la capacidad máxima de las aulas de los estudiantes, mientras que el eje Y la densidad de la probabilidad – es decir, que tan concentrados están los estudiantes de los grupos entre las aulas de distintas capacidades máximas. El área tanto de las barras grises como de la línea roja se encuentran normalizadas, de modo que la suma da 1 para cada uno de los casos. Que tanto la línea roja como las barras grises se comporten de manera similar en tanto

tamaño y fluctuación nos indicaría que la capacidad máxima del curso no ejerce mayor influencia sobre la repitencia.

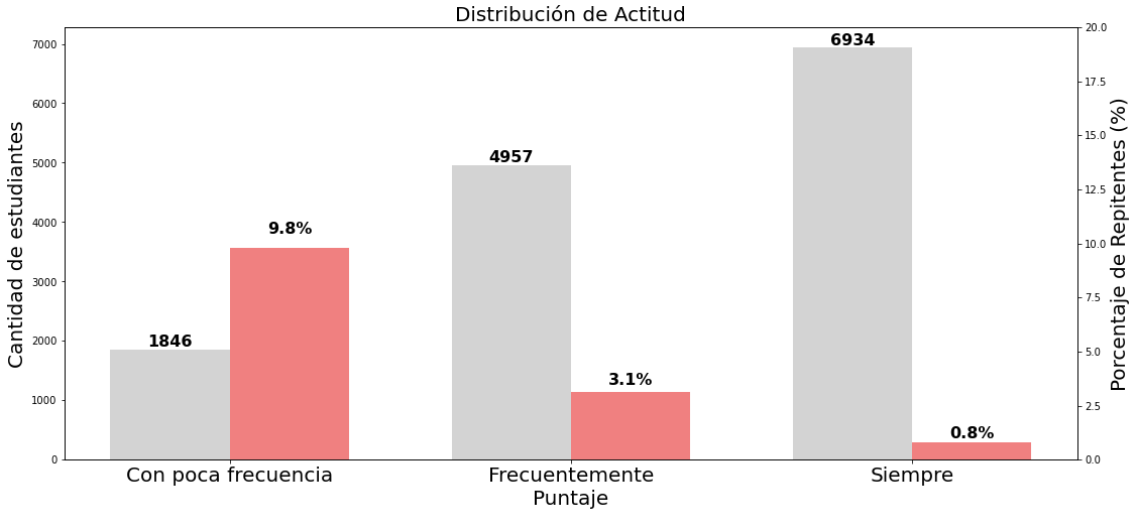
Lo que se observa es que la capacidad máxima del curso tendría un impacto en aulas de entre 25 y 35 estudiantes. Dentro de este rango se observa que la columna gris está por encima de la línea roja, indicando una mayor concentración de casos de repitencia que de matriculados en aulas de estas características.

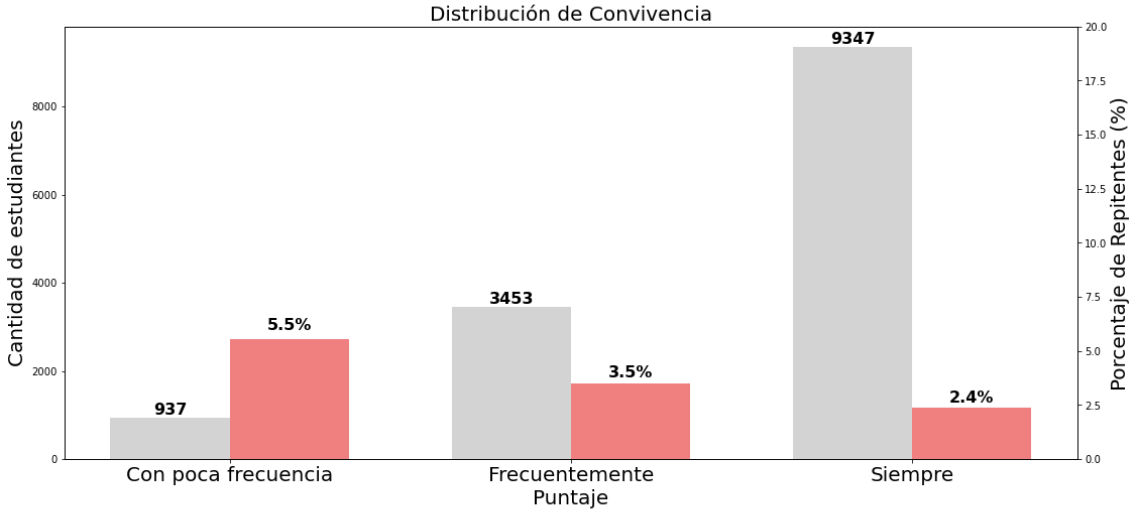
3.5. Puente Primario Secundario

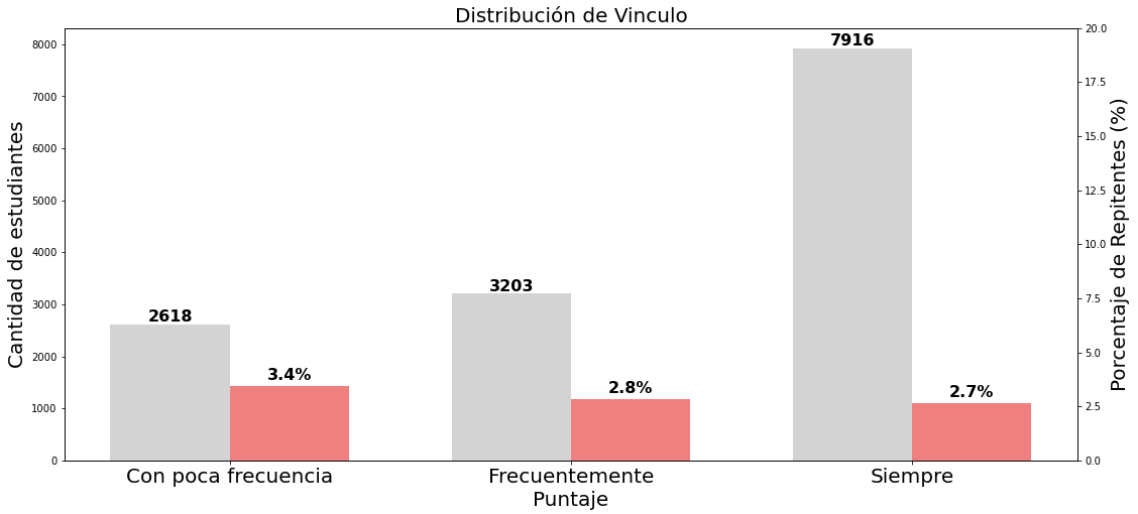
Según lo expuesto en la bibliografía, se espera que estudiantes con mejores habilidades en su actitud para el aprendizaje, la convivencia en el aula y otras variables no estrictamente relacionadas con el desempeño académico tengan mayor tasa de éxito en sus trayectorias educativas. Se estima que la capacidad de vincularse con otros guarda estrecha relación con la actitud frente al aprendizaje, estando esta última fuertemente asociada al éxito académico (Hsin & Xie, 2011).

En esta línea se realizó la Figura 4, donde se observa la distribución de los estudiantes según el promedio de su puntaje en los ejes actitud, convivencia y vínculo del PPS, y que tanto porcentaje de esos grupos repetía. Al igual que en la Figura 2, correspondiente al análisis por género, las barras grises muestran el total de estudiantes y las barras rojas el porcentaje de esos estudiantes que repiten. El eje Y primario representa la cantidad de estudiantes de las barras grises, y el secundario el porcentaje a representar en las barras rojas. Ejemplo: existen 6.934 estudiantes para los que al promediar su puntaje del eje actitud y redondearlo obtienen un valor de 2 correspondiente a la categoría “Siempre”, de ese grupo solo el 0.8% repite curso el curso realizado en 2023.

Figura 4 - Distribución de la matrícula y los casos de repitencia según desempeño en PPS







Lo que se puede observar es que en el eje de actitud, que captura información acerca del compromiso y comportamiento del estudiante frente al aprendizaje, los casos de repitencia se concentran entre aquellos estudiantes de menor puntaje. Con menor contundencia, se observa algo parecido para el caso de convivencia, indicando que ambas variables tienen un comportamiento alineado con aquello indicado por la bibliografía: mejores puntajes, menor probabilidad de repitencia.

En lo que respecta al eje de vinculo, se promediaron las columnas *vinculo_acompaña* y *vinculo_participa*, dado que las demás son columnas binarias que se crean producto del agrupamiento de los vínculos señalados por el docente. Lo que vemos en este caso es que la participación/acompañamiento de los adultos a las actividades del estudiante no guardan una relación fácilmente observable con la variable dependiente.

3.6. Calificaciones

Como se mencionó en secciones anteriores, el presente trabajo incorpora las calificaciones de lengua y matemática como indicadores de desempeño académico de los estudiantes. Las mismas fueron seleccionadas dada su relevancia y por ser espacios curriculares transversales a todas las dependencias funcionales.

De todas formas, según lo expuesto por Duncan, et al. (2007) las calificaciones de matemática guardarían una relación más estrecha con los resultados académicos de mediano plazo de los estudiantes. En su estudio acerca de la relación entre las habilidades de los estudiantes al momento de entrar al sistema educativo y sus logros académicos de largo plazo, encuentran que los resultados obtenidos en matemática tienen el doble de poder predictor que aquellos obtenidos en lengua.

Las figuras 5 y 6 representan la distribución de los estudiantes en general, y los repitentes en términos de su promedio en las materias de lengua y matemática durante séptimo grado. La línea roja representa al total de la matrícula mientras que las barras grises a aquellos estudiantes que repiten el grado cursado en 2023. Al igual que en el caso de la Figura 3, las áreas bajo la línea roja y de las columnas grises se encuentran normalizadas haciendo a ambas cosas comparables.

Figura 5 - Distribución de promedios de matemática del ciclo lectivo 2022

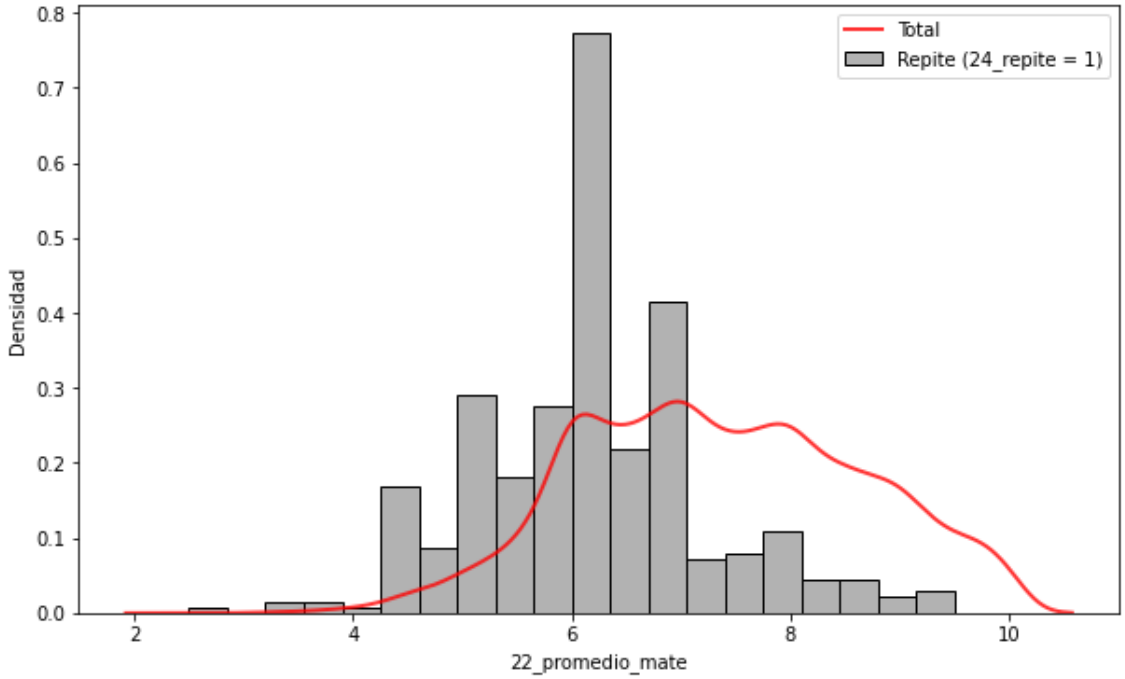
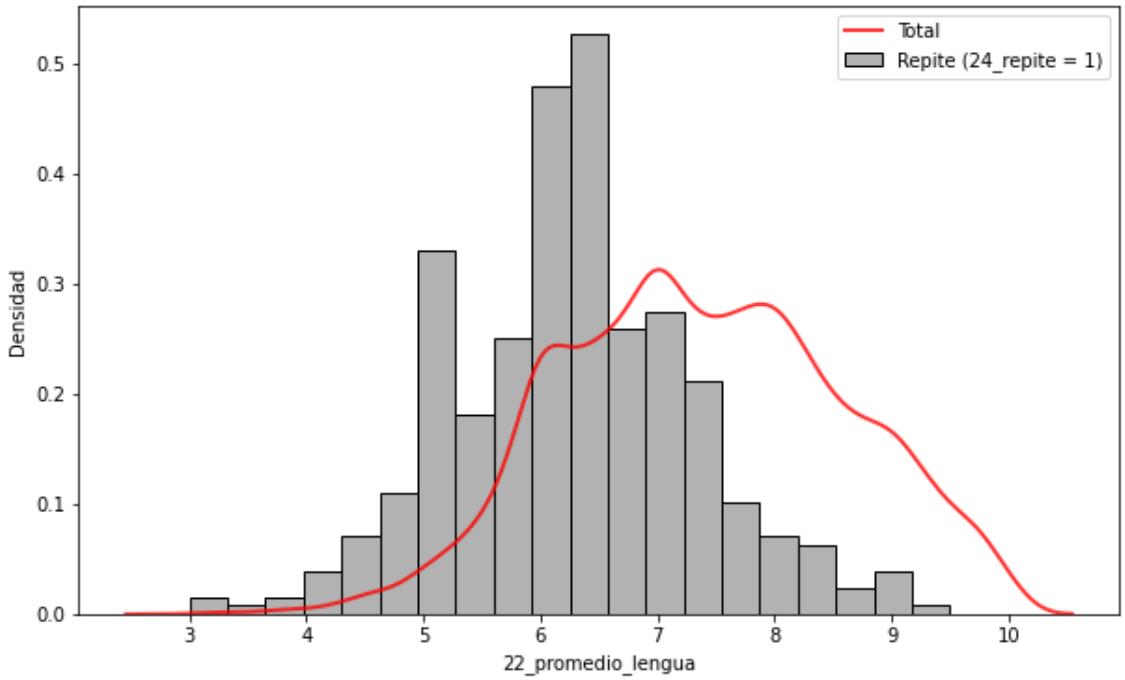


Figura 6 - Distribución de promedios de lengua del ciclo lectivo 2022



Observando el comportamiento de los promedios del ciclo lectivo 2022 completo, podemos ver que en el caso de matemática parece haber una mayor diferencia entre quieres repiten y aquellos que no lo hacen. Si se observa el rango de promedios entre 5 y 7, lo que se ve es que la densidad de probabilidad del total de los estudiantes se encuentra visiblemente por debajo que la de aquellos estudiantes que terminan por repetir. De la misma forma la línea roja sugiere que

el total de estudiantes se concentra en promedios entre el 6 y el 10 mientras que los promedios de quienes repiten se nuclean entre el 4 y el 8 aproximadamente.

En el caso de lengua el movimiento de la línea del total comparado con las columnas de los repitentes no muestra mayores diferencias en su comportamiento. Ambos siguen una distribución similar, aunque en el caso de los repitentes el promedio de notas quede situado a la izquierda, indicando que quienes repiten tendrían notas menores que quienes no lo hacen.

Es decir que, observando el ciclo lectivo previo, aquellos chicos que acaban por repetir tienen un desempeño relativo de notas más bajas que el total de los estudiantes en el caso de matemática. En el caso de lengua, si bien las notas de quienes repiten son menores, la distribución se mantiene. Esta diferencia en patrones en el caso de matemática indicaría que tal como menciona la bibliografía, el desempeño en ese área permite identificar tempranamente a aquellos estudiantes que podrían tener cierto rezago en su trayectoria.

3.7. Servicios

Como se mencionó en el apartado 2.1.6, este trabajo incorpora datos vinculados a beneficios económicos condicionados a la asistencia escolar de los estudiantes. Estos programas exigen como requisito de permanencia la regularidad en la asistencia a clases. En Argentina, la Asignación Universal por Hijo (AUH) —programa de alcance nacional— establece condiciones similares de escolaridad regular a las de los beneficios analizados en este estudio.

En su investigación sobre la AUH, Edo, Marchionni y Garganta (2017) evidencian que esta política de transferencias condicionadas incrementa en 3,9 puntos porcentuales la probabilidad de asistencia a clases en el nivel secundario. Asimismo, el efecto positivo resulta más significativo entre estudiantes varones y entre aquellos que provienen de familias numerosas con jefes de hogar con menor nivel educativo.

Si bien el estudio no analiza de manera directa los efectos sobre la repitencia escolar, el aumento en la asistencia regular podría, en ciertos contextos, contribuir a una mayor permanencia y continuidad educativa, lo cual podría estar vinculado —aunque no de manera automática— con menores tasas de repitencia. No obstante, cabe señalar que en algunos casos estos programas podrían contribuir a evitar el abandono escolar sin necesariamente mejorar los niveles de promoción, dado que los estudiantes permanecen en el sistema pero no alcanzan los requisitos académicos necesarios.

En términos del presente trabajo, lo elaborado por los autores permite formular la hipótesis de que estudiantes que acceden a dichos beneficios podrían presentar, en promedio, una menor tasa de repitencia que aquellos cuyas familias no los perciben.

Figura 7 - Distribución de casos según si el estudiante es beneficiario de Becas Media

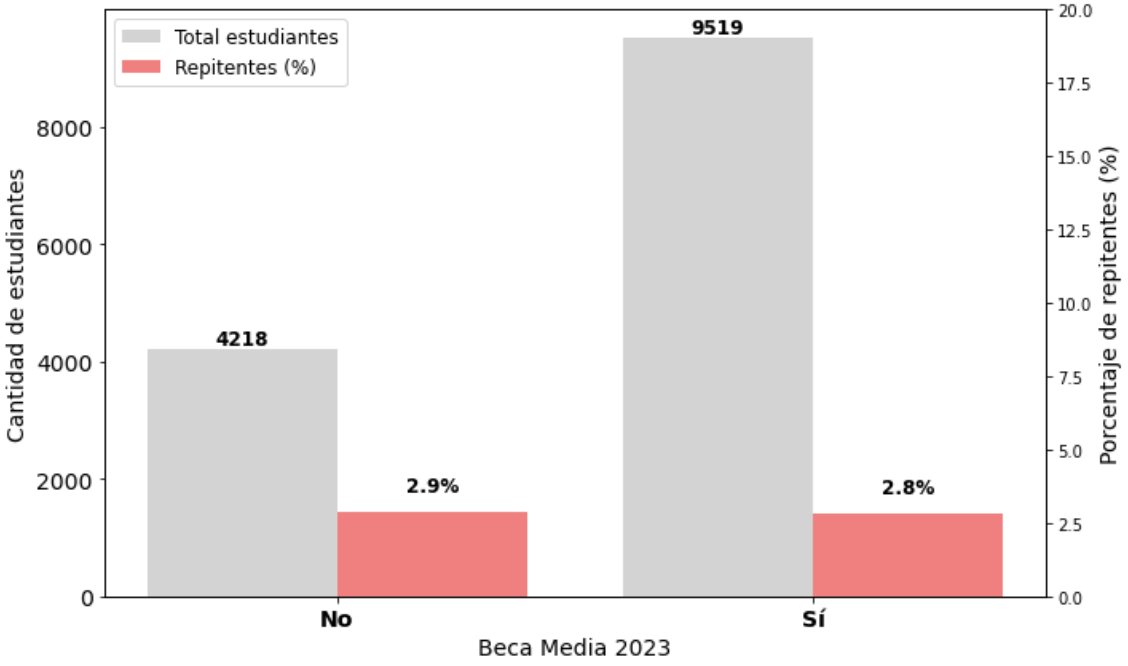
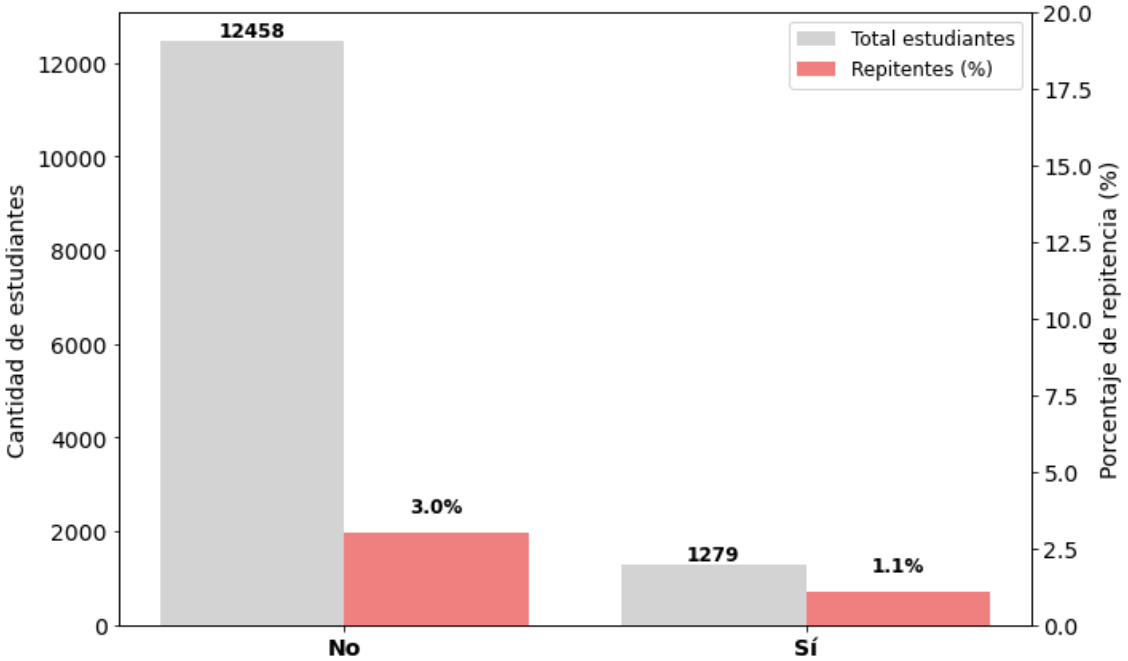


Figura 8 - Distribución de los casos según si el estudiante es beneficiario de Becas Alimentarias



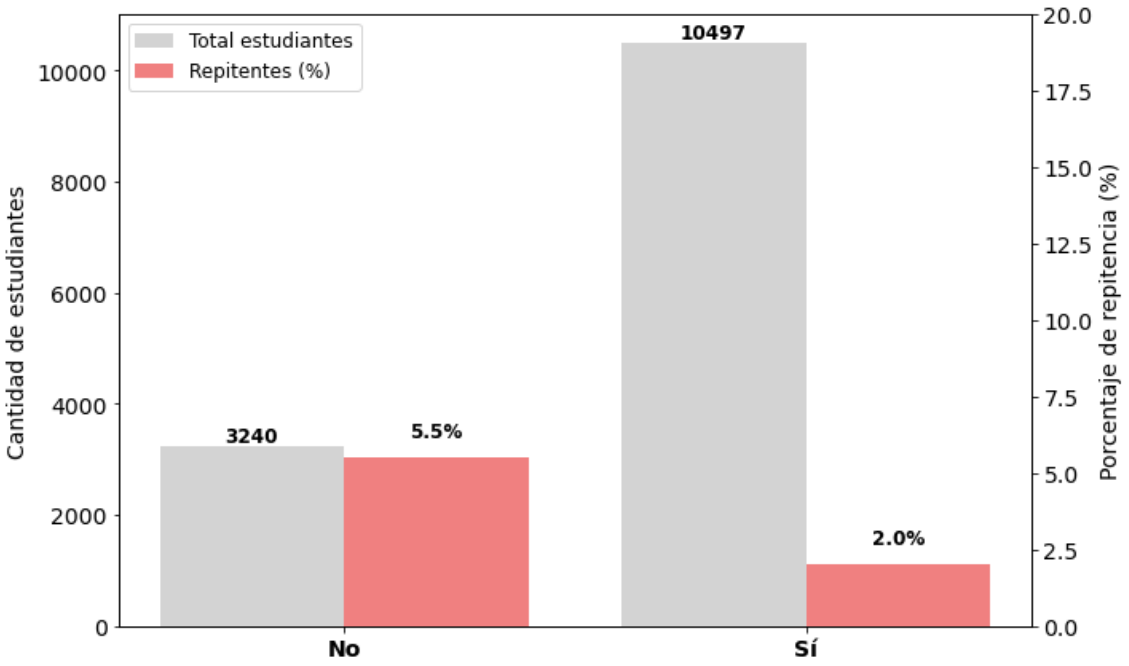
En el caso de Becas Media, representado en la Figura 6, lo que se observa es que no habría un impacto del programa en la repitencia del alumno. Tanto en el caso de los beneficiarios como los no beneficiarios del programa, el porcentaje de repitentes es prácticamente el mismo.

En el caso de las Becas Alimentarias sí se evidencia una diferencia entre quienes son o no beneficiarios del programa. La cantidad de casos de repitencia entre aquellos que obtuvieron el beneficio es casi dos veces menor que entre aquellos que no lo reciben.

Lo que se observa también es que el programa de Becas Alimentarias parece ser más restrictivo en cuanto a las barreras de entrada, resultando en una identificación tal vez más precisa de casos de vulnerabilidad económica. En este sentido resulta coherente ver que el ser beneficiario o no de este programa guarde mayor relación con una reducción en la repitencia que en el caso del programa Becas Media.

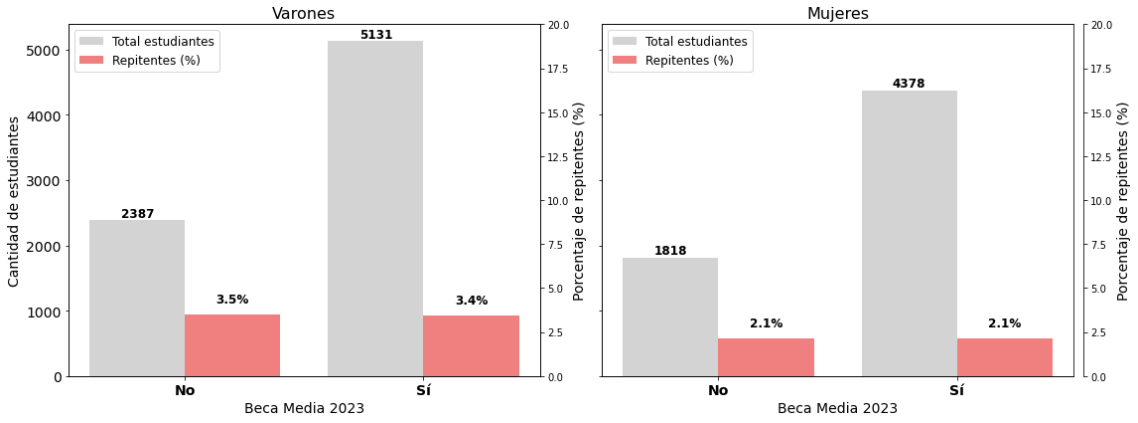
Se realizó el mismo análisis pero sobre los beneficiarios de boleto estudiantil, encontrando que entre aquellos que no poseen este beneficio las tasas de repitencia son más altas que las de aquellos que si lo perciben. De todas maneras al ser este beneficio adquirible solo por la porción de los estudiantes que residen en la Ciudad, se entiende que la distribución de los casos no se puede tomar como algo exclusivamente relacionado al beneficio en sí mismo. Lo que se sospecha es que esta diferencia podría estar relacionada al domicilio de los estudiantes, dato no disponible para ser incluido en el modelo, que implicaría una mayor distancia con el establecimiento educativo aumentando así el tiempo de viaje y reduciendo las posibilidades e incentivos de los estudiantes para tener un mejor desempeño educativo.

Figura 9 - Distribución de los cases según si el estudiante es beneficiario de Boleto Estudiantil



Por otro lado, se evaluó la distribución de casos por género en los programas Becas Media y Becas Alimentarias. Lo que se observa en las Figura 10, correspondiente al programa Becas Media es que para ambos géneros, el ser beneficiario de la beca no reporta un impacto perceptible en la repitencia.

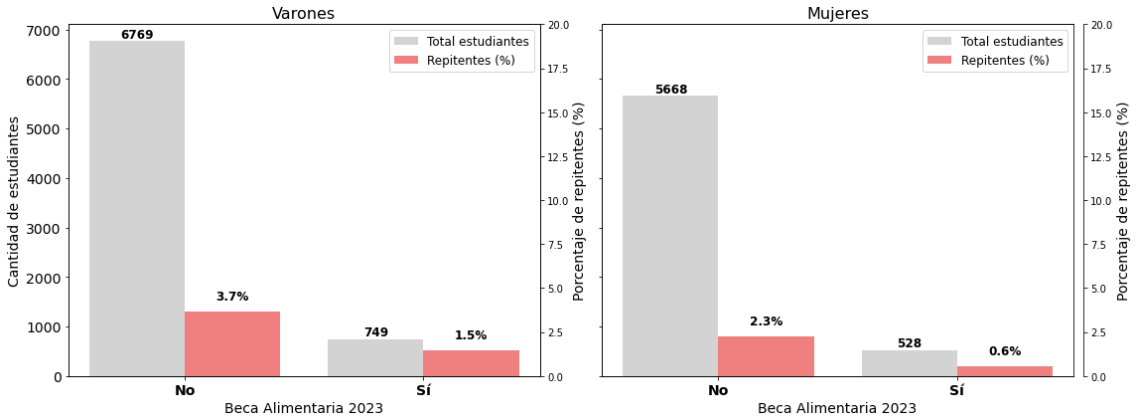
Figura 10 - Distribución de casos según si el género del estudiante y si este es beneficiario de Becas Media



Por el contrario, en la Figura 11 puede verse que el efecto de ser o no beneficiario de Becas Alimentarias parecería tener un impacto distinto según el género del estudiante. En el caso de los varones, la tasa de repitencia entre aquellos que son beneficiarios del programa es del 1.5% mientras que en el caso de las mujeres es del 0.6%. Si se observa esto comparado con los porcentajes de repitencia de los no beneficiarios de cada género, se obtiene que en el caso de los estudiantes varones la repitencia entre los beneficiarios es un 59,45% menor que entre los

no beneficiarios. Sin embargo, entre las mujeres hay una caída del 73,91% entre los valores de las no beneficiarias y quienes si reciben la beca. Esto indicaría que, en la muestra usada para este trabajo, la diferencia por géneros existe pero en el sentido inverso al indicado en el estudio de Edo, Marchionni, & Garganta (2017) siendo las mujeres y no los varones el grupo con mayor impacto positivo de la obtención de un subsidio condicional.

Figura 11 - Distribución de casos según si el género del estudiante y si este es beneficiario de Becas Alimentarias

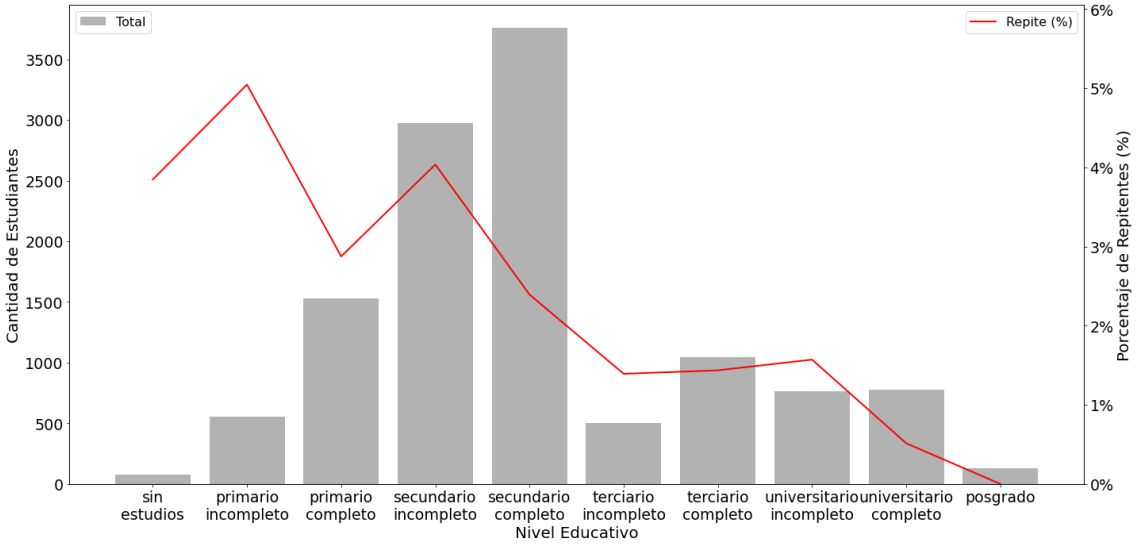


3.8. Responsables

En la sección 2.6 se desarrolló acerca de la importancia del nivel educativo del adulto que acompaña al estudiante en su desempeño académico. Esta variable es importante no solo como aproximación del nivel socioeconómico del hogar del alumno, en tanto un aumento en el nivel educativo está vinculado a un incremento en los ingresos (Avalis & Caro, 2024), sino porque el nivel educativo de los responsables aproxima la capacidad que los mismos tendrán para ayudar a los estudiantes que tengan dudas o dificultades para realizar sus tareas (Salas, Vigorito, & Failache, 2015).

En este sentido, lo que se espera es que aquellos estudiantes cuyos responsables tengan un mayor nivel educativo, tengan una menor probabilidad de repetir de curso. La Figura 12 refleja el comportamiento del total de los estudiantes y de aquellos que repiten según el nivel de estudios alcanzado por sus responsables. El eje X indica qué casos de nivel de estudios se reportan en la muestra. El eje Y principal se asocia a las barras grises y representa la cantidad de estudiantes para los cuales los responsables reportan tener ese nivel de estudios alcanzado, el eje Y secundario por su parte se asocia a la línea roja y representa el % de los estudiantes de la barra gris que repiten el curso realizado en 2023 – es decir, que tienen 24_repite=1.

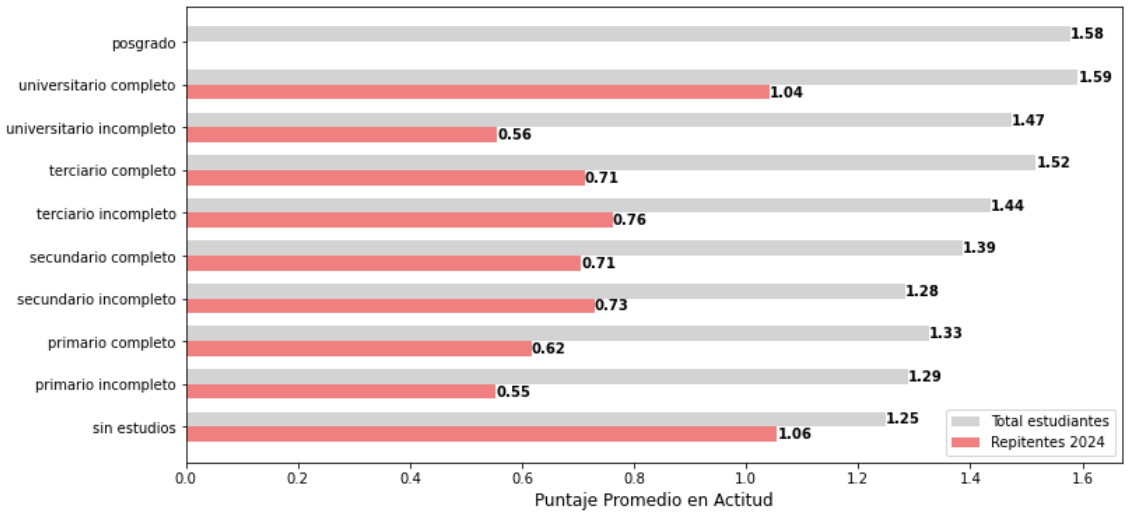
Figura 12 - Distribución de los casos según el nivel educativo de los responsables



Como se observa en la figura, la distribución de casos de repitencia en el conjunto de datos utilizado en este trabajo corresponde a lo elaborado por la bibliografía. Los casos de estudiantes que repiten de curso se concentran allí donde los responsables poseen un menor nivel educativo.

La Figura 13 representa el puntaje promedio de los estudiantes en el apartado de *actitud* del PPS según el nivel de estudios reportado por sus responsables, ordenado desde el mayor nivel de estudios alcanzado al menor. Las barras grises representan el promedio del puntaje del total de los estudiantes de cada grupo mientras que las barras rojas representan el promedio de aquellos estudiantes que repitieron en 2023.

Figura 13 - Puntajes actitudinales según nivel educativo del responsable



Lo que se observa es que en lo que hace a la relación entre el nivel educativo de los responsables y el eje actitudinal de los estudiantes, lo que se encuentra es que no hay una relación tan clara entre cómo se comporta la misma entre los casos de repitentes y no repitentes. Entre los estudiantes cuyos responsables reportan no tener estudios, el puntaje en el eje actitudinal es igual de alto que el de aquellos repitentes con responsables universitarios. En los niveles educativos intermedios, tanto el promedio general en el eje de actitud como el promedio entre repitentes no reporta diferencias con una tendencia identificable.

4. Metodología

La presente sección metodológica tiene como objetivo describir el enfoque utilizado para abordar el problema de predicción planteado en este estudio. En primer lugar, en la Sección 4.1 se define la variable dependiente del estudio. A continuación, en la Sección 4.2, se presenta el modelo de predicción, seguido de una descripción de los espacios de búsqueda de hiperparámetros en la Sección 4.2.1. Posteriormente, en la Sección 4.3, se explica la métrica de evaluación utilizada para medir el rendimiento del modelo. En la Sección 4.4, se describen los conjuntos de datos de entrenamiento y testeo empleados para validar la efectividad del enfoque propuesto. En la Sección 4.5, se detallan las estrategias para utilizar los distintos conjuntos de datos en los modelos de predicción. Finalmente, en la Sección 4.6, se presentan los modelos de benchmark utilizados para comparar los resultados obtenidos.

4.1. Definición de la variable dependiente

La repitencia escolar se define como la reinscripción de un estudiante en el mismo grado o nivel educativo en el que estuvo matriculado previamente debido a la no adquisición de los conocimientos, competencias o criterios de evaluación requeridos para la promoción al siguiente nivel (UNESCO, 2012).

Este fenómeno se captura en el presente trabajo a través de la variable *24_repite*. Esta variable se construye observando si el estudiante se encuentra matriculado para el *ciclo_lectivo* 2024 en el mismo curso que lo estuvo para el *ciclo_lectivo* 2023[8]. Específicamente para el caso de esta tesis, al estar trabajando con estudiantes que en el año 2023 realizaron 1er año del nivel secundario, el valor de *24_repite* será 1 para aquellos estudiantes que en 2024 hayan realizado

[8] En caso de que el estudiante se cambie de colegio la variable repite tomará valor 1 dado que lo que observa la misma es el curso al que asiste el estudiante independientemente del colegio en que lo haga.

1er año de nivel secundario por segunda vez. La misma será la variable dependiente en los modelos que se desarrollen.

Dentro del conjunto de datos, *24_repite* toma valor 1 en 392 ocasiones representando esto un 2,85% de los datos. Como consecuencia de ello, se adoptan en el presente trabajo estrategias de clases desbalanceadas.

4.2. Definición del modelo

El presente trabajo trata con un problema de predicción de una variable binaria, *24_repite*, con un desbalance de clases teniendo aproximadamente 1 caso positivo por cada 34 negativos dentro de la muestra. En este contexto, se eligió Extreme Gradient Boosting (XGBoost) como el modelo a utilizar para generar esta predicción. XGBoost es una implementación de la técnica de boosting, que en el contexto de la clasificación se basa en una secuencia de entrenamientos y reclasificaciones aditivas sobre los casos que no fueron correctamente clasificados en iteraciones anteriores. En este tipo de modelos, las predicciones se construyen de manera secuencial, donde cada nuevo árbol se entrena utilizando la información de los árboles previos.

La elección de este modelo proviene de su capacidad para construir predictores robustos a partir de modelos débiles. XGBoost mejora el modelo aprendiendo iterativamente a partir de una aproximación de los residuos de la predicción anterior. Cada nuevo árbol se entrena para corregir los errores residuales del conjunto anterior (Chen & Guestrin, 2016). Específicamente en cuanto al desbalance de clases, los algoritmos de boosting son una buena alternativa en tanto permiten ajustar la penalización de los errores de cada clase (Witten, Hastie, Gareth, & Tibshirani, 2017).

Dentro de este tipo de modelos se pueden realizar pruebas de hiperparámetros, entre los cuales se encuentran los siguientes:

- n_estimators: Número de árboles del modelo. Admite valores entre 0 e infinito.
    - Valores más bajos dan modelos más simples, valores altos implican mayor complejidad pero con riesgo de overfitting.
- learning_rate: Tasa de aprendizaje del modelo. Admite valores entre 0 y 1.
    - Valores bajos implican un aprendizaje más lento, pero con mejor generalización. Con valores altos el modelo aprende rápido pero con mayor tendencia al overfitting.
- max_depth: Máxima profundidad de los árboles del modelo. Admite valores entre 0 e infinito.

    - Valores más bajos dan modelos más simples, valores altos implican mayor complejidad pero con riesgo de overfitting.
- colsample_bytree: Fracción de las variables utilizadas para entrenar. Admite valores entre 0 y 1.
    - Valores bajos incorporan menos variables al modelo, valores más altos le dan mayor precisión.
- subsample: Fracción de los registros usados para entrenar. Admite valores entre 0 y 1.
    - Valores más bajos reducen el sobreajuste pero aumentan la varianza. Valores más altos son más precisos pero con riesgo de overfitting.
- gamma: Regularización para evitar sobreajuste, determina la mínima ganancia de información requerida para la división de un nodo. Admite valores entre 0 e infinito.
    - Valores bajos aumentan la cantidad de divisiones volviendo al modelo más flexible, valores altos reducen las divisiones simplificando el modelo.
- max_delta_step: Control de estabilidad, restringe cuánto puede cambiar la estimación d la clase en cada iteración.
    - Valores bajos son menos restrictivos generando mayor inestabilidad en problemas con desbalance de clases, valores altos permiten cambios pequeños en los pesos volviendo más lento el aprendizaje del modelo.
- min_child_weight: Cantidad mínima de observaciones que debe tener una hoja del árbol para que se admita su subdivisión. Admite valores entre 0 e infinito.
    - Valores bajos permiten más divisiones dando como resultado arboles más profundos pero con mayor riesgo de overfitting, valores más altos restringen la división que en caso de ser extrema lleva a underfitting.

Cada uno de los parámetros mencionados tiene un impacto directo en la capacidad del modelo para generalizar y evitar problemas como el overfitting o el underfitting. Dada la importancia de estos parámetros, se debe realizar una búsqueda cuidadosa de sus valores óptimos para encontrar el balance adecuado que maximice el rendimiento del modelo.

4.2.1. Búsqueda de hiperparámetros

El modelo elegido permite la búsqueda de hiperparámetros, pudiendo esta realizarse mediante *grid search* o *random search*. El método grid define un conjunto finito de valores a evaluar para cada uno de los parámetros, evalúa todas las combinaciones posibles entre los valores definidos y luego hace una validación cruzada para seleccionar la mejor. A nivel computacional resulta

costoso dado que si se busca probar combinaciones entre varias alternativas para cada parámetro el tiempo y costo de cómputo puede resultar muy alto (Chen & Guestrin, 2016).

El método de búsqueda aleatoria, o *random search*, elige aleatoriamente algunas combinaciones dentro de las posibles por los valores especificados. De este modo, no todo el conjunto de posibilidades es evaluado reduciendo el tiempo y el costo de cómputo sobre todo en datasets de tamaño considerable. La filosofía detrás de este método es que no todos los parámetros tienen el mismo impacto en el modelo, por lo que explorar aleatoriamente es suficiente.

Para el presente trabajo se adoptó una estrategia de *random search* sobre los siguientes parámetros: *n_estimators, learning_rate, max_depth, colsample_bytree, subsample, gamma, max_delta_step,* y *min_child_weight*. Tanto para estos modelos como para los que se utilizarán de benchmark (ver Sección 4.6), el número de iteraciones definido (*n_iter*) es 100. Esto implica que el algoritmo seleccionará 100 combinaciones aleatorias distintas entre los valores definidos para los hiperparámetros y evaluará cada una de ellas mediante validación cruzada. De este modo, se garantiza una exploración amplia del espacio de búsqueda sin necesidad de probar todas las combinaciones posibles, lo que reduce significativamente el tiempo de cómputo manteniendo una buena probabilidad de encontrar configuraciones cercanas al óptimo.

En una primera ronda de entrenamiento, el conjunto de parámetros aleatorios tenía como espacio de trabajo para *colsample_bytree*, *subsample* y *min_child_weight* los siguientes conjuntos:

- *colsample_bytree* : [0.7, 0.8, 0.9]
- *subsample*: [0.7, 0.8, 0.9]
- *min_child_weight*: [1, 5, 10]

Lo que se observó al entrenar los modelos con el conjunto de parámetros seteado de esta forma fue que, en todos los casos, estos tres hiperparámetros se concentraban en los extremos. En el caso de *colsample_bytree* y *subsample* siempre se adoptaba el menor valor, mientras que *min_child_weigt* siempre se situaba en el mayor valor permitido. En el caso de *colsample_bytree* y *subsample* todos los modelos con mejor performance en AUC-ROC promedio de los folds tomaban valor 0.7, exceptuando únicamente el último caso que tomó valor 0.9. En lo que respecta a *min_child_weight* el comportamiento fue el contrario, tomando este siempre el valor más grande permitido por el conjunto de hiperparámetros dados: 10.

Como consecuencia de esto, los modelos fueron nuevamente entrenados pero ampliando los valores de estos tres parámetros obteniendo el siguiente espacio de búsqueda:

Tabla 8 - Segunda configuración del espacio de búsqueda aleatoria de parámetros en XGBoost

Parámetro	Definición	Valores indicados
n_estimators	Número de árboles del modelo	[100, 300, 500, 700]
learning_rate	Tasa de aprendizaje del modelo	[0.01, 0.05, 0.1, 0.2]
max_depth	Máxima profundidad de los árboles del modelo	[4, 6, 8, 10]
colsample_bytree	Fracción de las variables utilizadas para entrenar	[0.5,0.6,0.7,0.8,0.9]
subsample	Fracción de los registros usados para entrenar	[0.5,0.6,0.7,0.8,0.9]
gamma	Regularización para evitar sobreajuste	[0, 1, 3, 5]
max_delta_step	Control de estabilidad	[0, 1, 5]
min_child_weight	Peso mínimo de la hoja	[1, 5, 10, 12, 14]

La manera en que se configuró el nuevo espacio de búsqueda responde a lo observado en el primer caso. La manera en que se definían *colsample_bytree*, *subsample* y *min_child_weight* sugería que el modelo estaba eligiendo una configuración conservadora, por lo que se les amplió el espacio de búsqueda en ese sentido. El uso de menos columnas y registros, y la creación de hojas con pesos mínimos más altos favorece a la creación de modelos más conservadores.

4.3. Métrica de evaluación del modelo

Los modelos de XGBoost usados en este trabajo serán optimizados en base a la métrica AUC-ROC. El área bajo la curva ROC mide la capacidad de un modelo para diferenciar entre clases positivas y negativas en problemas de clasificación binaria. A diferencia de la métrica de Accuracy, el AUC-ROC ofrece una evaluación más informativa del desempeño del modelo al considerar tanto la tasa de verdaderos positivos como la de falsos positivos.

Mientras que *Accuracy* se fija en el porcentaje de observaciones bien clasificadas, pudiendo ser estas únicamente las correspondientes a la clase mayoritaria, AUC-ROC se centra en entender la capacidad del modelo de asignar las observaciones a sus correspondientes categorías independientemente de la proporción de cada una de ellas dentro del conjunto de datos.

AUC-ROC se basa en las fórmulas de tasa de verdaderos positivos TPR y la tasa de falsos positivos FPR. En la teoría se define como la integral de la curva TPR en función de FPR.

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$
$$AUC = \int_0^1 \frac{TP}{TP + FN}\, d\left(\frac{FP}{FP + TN}\right) \quad (1)$$

Siendo que las iteraciones son eventos discretos y no continuos como deberían serlo para integrar, el área bajo la curva se aproxima mediante la sumatoria de las áreas bajo la curva que surgen de la evaluación de un modelo en un numero finito de umbrales, que oscilan entre 0 y 1.

$$AUC = \sum_{i=1}^{n} (FPR_i - FPR_{i-1}) \left(\frac{TPR_i + TPR_{i-1}}{2}\right) \quad (2)$$
$$AUC = \sum_{i=1}^{n} \left(\frac{FP_i}{FP_i + TN_i} - \frac{FP_{i-1}}{FP_{i-1} + TN_{i-1}}\right) \left(\frac{\frac{TP_i}{TP_i + FN_i} + \frac{TP_{i-1}}{TP_{i-1} + FN_{i-1}}}{2}\right)$$

Los valores resultantes de esta operación se encuentran dentro del rango [0;1] donde valores más altos indican mejor performance del modelo en separación de clases. Un valor de 0,5 indica una performance del modelo comparable con el azar y un valor AUC-ROC < 0,5 indica que el modelo performa incluso peor que un clasificador que clasifica de manera totalmente azarosa.

4.4. Definición de conjuntos de entrenamiento y testeo

Como se mencionó con anterioridad, el presente trabajo utiliza datos de la cohorte 2022 de nivel primario. Dicha limitación proviene de la inclusión en el modelo de información del módulo PPS cuyas respuestas sistematizadas no están disponibles para años previos.

Al evaluar la opción de incluir información acerca de la cohorte 2023, que realizó su primer año de secundaria en 2024, se encontró que al no haber las escuelas efectivizado a sus alumnos hasta el momento de elaboración de esta tesis no se poseía información fiable acerca de en qué curso se encontrarían los estudiantes durante el ciclo lectivo 2025[9]. De esta forma la única información disponible tanto para entrenamiento como para evaluación y el testeo es aquella correspondiente a quienes terminaron el primario en el 2022.

El desbalance de clases dentro de los datos, siendo los positivos un 2,85% del total de los casos, lleva a que las estrategias de train, validation y test split comúnmente utilizadas no sean

[9] La efectivización es un proceso administrativo realizado por las escuelas públicas de la Ciudad dentro de la plataforma AprendeBA donde las mismas informan la distribución de los estudiantes que asistirán al establecimiento ese año entre los distintos cursos. De allí es que puede inferirse quienes de ellos repitieron el curso realizado en el ciclo lectivo previo y quienes no.

aplicables de manera directa. Para dividir entre los conjuntos de entramiento, testeo y validación se debe tener en cuenta el desbalance de clases, y para ello se utiliza una estrategia en dos pasos.

En primer lugar se separan los datos entre un 80% a utilizar en el entrenamiento y la validación, y un 20% a reservar para testear el modelo. La división entre estos dos grupos se realiza de manera estratificada utilizando el parámetro *stratify* provisto por el módulo *train_test_split* de sklearn[10]. De esta separación se reservan 2748 casos para testeo, siendo 78 de estos casos positivos. Para el entrenamiento y validación se conservan 10989 casos, con 314 casos positivos.

El objetivo de reservar un 20% de datos ajenos al entrenamiento es poder contar con un conjunto que no haya sido considerada por el modelo en ninguna de las validaciones internas, para luego predecir esos datos con el mejor modelo resultante del entrenamiento y poder aproximar cómo se comportaría el modelo ante datos desconocidos.

Dentro del 80% de los datos reservado para entrenamiento y validación, es necesario aplicar una estrategia que permita generar estos dos subgrupos de forma robusta. Para ello, se empleó el método *RepeatedStratifiedKFold* provisto por *scikit-learn*[11]. Al igual que *RepeatedKFold*, este método se utiliza como esquema de validación cruzada (*cross-validation*), dividiendo los datos en *k* particiones (*folds*) que se rotan de manera tal que cada *fold* actúe como conjunto de validación una vez, mientras los demás se usan para entrenar el modelo. En el contexto de entrenamiento de modelos con XGBoost, este esquema permite evaluar el desempeño del modelo en múltiples combinaciones de entrenamiento/validación.

La diferencia que existe entre *RepeatedKFold* y *RepeatedStratifiedKFold* es que en el primer caso los grupos se dividen de manera completamente aleatoria, lo cual no sirve en un contexto de clases desbalanceadas donde la utilidad del modelo debe validarse con cierta consideración del desbalance. Lo que permite *RepeatedStratifiedKFold* es que cada uno de los folds creados para la validación mantengan la proporción de casos positivos y negativos del dataset original.

Los parámetros que deben definirse para utilizar esta estrategia son: *n_splits*, *n_repeats* y *random_state* (en caso de querer dar replicabilidad a la generación de los sets). En este trabajo, se definió n_splits=3 y n_repeats=5. Lo que implica que para cada uno de los modelos evaluados los datos se dividen en tres grupos distintos, usando dos para entrenar y uno para validar, y todo

[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[11] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html

este proceso se repite cinco veces, lo cual permite reducir la varianza de las métricas obtenidas y mejora la robustez de las estimaciones de desempeño.

4.5. Definición de los datos a utilizar en cada modelo

Teniendo en cuenta la multiplicidad de conjuntos de información disponibles, se evaluó incluir distintos conjuntos de variables predictoras en cada modelo. La definición de qué incluir en cada uno de ellos fue la siguiente:

Tabla 9 - Diagramación de los modelos

Conjunto de datos	M1	M2	M3	M4	M4B	M4C	M5	M6	M6B**
Matrícula	X	X	X	X	X	X	X	X	X
Puente Primario Secundario		X	X	X	X	X	X	X	X
Pases de institución			X	X	X	X	X	X	X
Calificaciones				X	X		X	X	X
Desempeño relativo*					X	X	X	X	X
Promedios*					X	X	X	X	X
Servicios							X	X	X
Responsables								X	X

*Conjuntos de variables creadas a partir de datos de calificaciones

** Modelo idéntico a M6 pero solo con información del ciclo lectivo 2022, con el fin de entender la capacidad de predecir repitencia al momento en que el estudiante ingresa al nivel secundario.

Sobre el modelo M4C vale una aclaración. El hecho de que no contenga información de calificaciones significa que no incluye los registros bimestrales de las calificaciones del estudiante. En su lugar, solo utiliza su promedio anual de 2022 y semestral de 2023, junto con las columnas que reflejan su desempeño relativo dentro de la sección a la que pertenece.

4.6. Modelos de benchmark

Para poder evaluar el desempeño de los modelos con búsqueda aleatoria de hiperparámetros, primero se entrenaron modelos básicos de XGBoost para clasificación binaria, midiendo su efectividad mediante AUC-ROC al igual que los modelos más complejos que se utilizarán posteriormente.

Para mantener la coherencia con los modelos a evaluar en este trabajo, el clasificador binario simple a usar como benchmark se entrenó utilizando *Repeated Stratified K-Fold* (RSKF) con

n_splits=3 y *n_repeats*=5. Siguiendo el mismo criterio que se aplicará a los modelos complejos, el total de los datos se dividió en dos conjuntos que comprendieron el 80% y el 20% de los datos respectivamente. El conjunto con el 80% de los datos se utilizó para entrenamiento y validación cruzada, y el restante 20% desconocido por el modelo como conjunto de testeo.

El objetivo de este análisis es determinar si, con los mismos datos, un modelo más complejo tiene mayor poder predictivo, justificando así su uso. En la Tabla 10, donde se presentan los resultados de los modelos de benchmark, mean AUC RSKF representa el promedio del AUC-ROC los entrenamientos del mejor de los 100 modelos entrenados para cada caso.

Tabla 10 - Modelos de Benchmark

Modelo	Mean AUC RSKF*	AUC test
M1	50,25%	57,16%
M2	67,31%	67,01%
M3	67,32%	67,07%
M4	78,71%	79,52%
M4B	75,11%	81,72%
M4C	74,93%	79,86%
M5	74,92%	81,60%
M6	75,10%	81,72%
M6B	70,35%	68,80%

* RSKF – Repeated Stratified K Fold

Lo que puede verse en los modelos de benchmark es una considerable inestabilidad entre los resultados en train y en test. La brecha de AUC-ROC entre validación y test alcanza valores cercanos a los 7 puntos porcentuales. Lo que nos indican estos resultados es que el modelo básico sin tuneo de hiperparámetros no está teniendo la capacidad de generalizar sobre datos desconocidos. Los únicos dos casos donde se muestra cierta estabilidad de parte del modelo son los modelos M4, aquel que incorpora la información de las calificaciones de los estudiantes, y M6B, donde se realiza una predicción con todos los datos del estudiante correspondientes al ciclo lectivo 2022.

5. Resultados

5.1. Experimentación con modelos y análisis de resultados

En esta sección se presentan los resultados de los modelos explicados en la sección 4.3. En la Tabla 11 se observan los resultados obtenidos de la primera configuración de hiperparámetros (HP1) comentada en la sección 4.2.1. La Tabla 12 por su parte muestra aquellos resultados obtenidos de la segunda configuración de hiperparámetros presentada en ese mismo apartado (HP2).

Como se comentó en la sección 4.2.1, la búsqueda de hiperparámetros se realizó con un *n_iter*=100. Es por esto por lo que en tanto en la Tabla 11 como en la Tabla 12 *Mean Auc CV* representa el promedio de los AUC-ROC obtenidos en las rondas de entrenamiento y validación presentadas en la sección 4.4 para el mejor de esos 100 modelos.

Tabla 11 - Resumen de los parámetros por modelo usando la primera configuración de hiperparámetros

Modelo	n_estimators	learning_rate	max_depth	colsample_bytree	subsample	gamma	max_delta_step	min_child_weight	Mean AUC RSKF	AUC test
M1	100	0,05	10	0,7	0,7	5	1	10	66,70%	69,06%
M2	300	0,01	4	0,7	0,7	1	1	10	78,79%	79,14%
M3	300	0,01	4	0,7	0,7	1	1	10	78,83%	79,04%
M4	300	0,01	8	0,7	0,8	3	0	10	89,32%	90,28%
M4B	700	0,01	8	0,7	0,7	3	0	10	88,77%	90,08%
M4C	100	0.05	10	0,7	0,7	5	1	10	84,89%	88,57%
M5	700	0,01	8	0,7	0,7	3	0	10	88,94%	90,54%
M6	700	0,01	8	0,7	0,7	3	0	10	89,14%	90,45%
M6B	500	0,01	4	0,9	0,7	0	1	5	82,64%	81,62%

Lo que se observa en la Tabla 11, correspondiente a la primera configuración de hiperparámetros elegida, es que los parámetros *min_child_weight*, *subsample* y *colsample_bytree* se sitúan en valores extremos indicando que tal vez el modelo está buscando explorar más allá de lo permitido. Sin embargo, en comparación con los modelos de benchmark presentados en la sección 4.6 se observa un aumento considerable de la estabilidad en el poder predictivo del modelo entre la validación y el entrenamiento.

De una brecha máxima de casi 7 puntos porcentuales entre el promedio del AUC-ROC de validación y testeo en los modelos de benchmark, esta se reduce a un valor levemente menor que 3.7 puntos porcentuales usando HP1. El modelo que mantiene mayor brecha entre los conjuntos en este caso es el modelo M4C, correspondiente a aquel que trabaja solo con las notas promedios de los estudiantes, con un AUC-ROC promedio de entrenamiento de 84.89% y un AUC-ROC en test de 88.57%.

Tabla 12 - Resumen de los parámetros por modelo usando la segunda configuración de hiperparámetros

Modelo	n_estimators	learning_rate	max_depth	colsample_bytree	subsample	gamma	max_delta_step	min_child_weight	Mean AUC RSKF	AUC test
M1	300	0,01	10	0,5	0,9	0	1	12	66,89%	69,51%
M2	500	0,01	4	0,5	0,7	3	0	14	78,85%	79,72%
M3	500	0,01	4	0,5	0,7	3	0	14	78,88%	79,94%
M4	500	0,01	4	0,5	0,7	3	0	14	89,41%	90,15%
M4B	300	0,01	10	0,5	O.9	0	1	12	88,85%	89,25%
M4C	300	0,01	10	0,5	0,9	0	1	12	88,62%	88,46%
M5	700	0,01	6	0,7	0,5	5	0	14	89,10%	90,45%
M6	700	0,01	6	0,7	0,5	5	0	14	89,21%	90,09%
M6B	500	0,01	4	0,5	0,7	3	0	14	82,72%	81,76%

Lo que se observa es que los modelos de la Tabla 12, correspondientes a HP2, eligen efectivamente valores más bajos de colsample_bytree y más altos de *min_child_weght* cuando se les da la oportunidad. En el caso de *subsample,* los valores adoptan un comportamiento menos uniforme obteniendo tanto valores más altos como más bajos que en el caso de la Tabla 11.

En lo que hace al análisis de la estabilidad de los modelos entre el AUC-ROC promedio que resulta de la estrategia de RSKF y el de testeo, hay una reducción incluso más pronunciada de la que se observaba en la Tabla 11 respecto de los modelos de benchmark expuestos en la Tabla 10. Con esta nueva configuración de parámetros la diferencia máxima entre el entrenamiento y el testeo de los modelos es de 2.62 puntos porcentuales - casi 5 puntos porcentuales menor a la de los modelos de benchmark y 1.1 puntos porcentuales menor a la que se obtuvo de la aplicación de HP1.

Resumiendo lo obtenido tanto de los modelos con las dos configuraciones de hiperparámetros como de los modelos de benchmark, lo que se observa es lo siguiente:

Tabla 13 - Resumen de los resultados de los diferentes modelos

Modelo	HP1 Mean AUC RSKF	HP1 AUC Test	HP2 Mean AUC RSKF	HP2 AUC Test	Benchmark AUC Test
M1	66,70%	69,06%	66,89%	69,51%	57,16%
M2	78,79%	79,14%	78,85%	79,72%	67,01%
M3	78,83%	79,04%	78,88%	79,94%	67,07%
M4	89,32%	90,28%	89,41%	90,15%	79,52%
M4B	88,77%	90,08%	88,85%	89,25%	81,72%
M4C	84,89%	88,57%	88,62%	88,46%	79,86%
M5	88,94%	90,54%	89,10%	90,45%	81,60%
M6	89,14%	90,45%	89,21%	90,09%	81,72%
M6B	82,64%	81,62%	82,72%	81,76%	68,80%

HP1 refiere a la primera configuración de hiperparámetros empleada
HP2 refiere a la segunda configuración de hiperparámetros empleada

Los modelos entrenados en este trabajo con el objetivo de predecir repitencia escolar identifican como variables principales aquellas propuestas por la bibliografía citada. Tanto en el caso de HP1 como HP2, en términos de AUC-ROC el modelo M5 se observa como el mejor en términos de capacidad de predicción de casos de repitencia, seguido muy de cerca por el modelo M4 y el M6. Lo que incluyó cada uno fue presentado en la Tabla 10.

En ambas configuraciones de hiperparámetros, así como en el modelo de benchmark, los modelos tienen una apreciable mejora en tanto se añaden los datos del Puente Primario Secundario (M2) que termina por consolidarse cuando se adjunta información de las calificaciones (M4).

El modelo que incorpora promedios y desempeños relativos (M4B) no mejora los resultados, sino que los mantiene levemente por debajo de aquel que trabaja con calificaciones bimestrales (M4). El modelo que solamente emplea notas agregadas y rankings (M4C) tiene un desempeño levemente peor que su predecesor (M4B).

El modelo que añade datos acerca de pases de institución (M3) genera una leve mejora en el caso de la segunda configuración de hiperparámetros y una sutil desmejora en el caso de la primera configuración.

El modelo que incorpora información acerca de los responsables (M6) reduce levemente el rendimiento del modelo, haciéndolo menos preciso que el modelo previo (M5). En última

instancia su par M6B, correspondiente a la misma información pero situándose a comienzos de 2023 sin datos del año académico que el estudiante acabará por repetir, presenta un valor razonable de AUC-ROC.

Transversal a los 9 modelos puede verse que la ampliación del rango en los parámetros de *subsample*, *colsample_bytree* y *min_child_weight* provista por el conjunto de hiperparámetros número 2 aumenta la estabilidad de los modelos. Tomando el promedio de la diferencia entre el AUC-ROC obtenido en validación y test, HP1 presenta un 1.47 puntos porcentuales de diferencia promedio entre los conjuntos mientras que HP2 reduce esta diferencia al 0.76 puntos porcentuales. Lo que esto sugiere es que el segundo conjunto permite una mejor generalización dando mayor estabilidad a la predicción, y mayor consistencia en la performance ante nuevas observaciones.

Comparados con el benchmark, todos los modelos pertenecientes a HP1 y HP2 presentan mayor rendimiento. Esto corresponde a que el modelo base de XGBoost puede no ser el mejor para trabajar en contexto de clases desbalanceadas.

El *learning_rate* por default de XGBoost es mucho más alto (0.3) que aquel empleado por los modelos cuando se les permite explorar otros espacios (ubicándose en general en 0.01). Este parámetro controla el sobreajuste por lo que un valor más bajo hace que el modelo tenga mayor capacidad de generalizar. Lo mismo sucede con *subsample* y *colsample_bytree*, ambos de valor 1 en el modelo base, que toman valores más conservadores que previenen el overfitting en los modelos con hiperparámetros.

En el caso de *min_child_weight* se da la mayor diferencia en tanto es uno de los parámetros que toma el valor máximo permitido en HP1, 10, y luego aumenta su valor a 12 o 14 cuando se le permite hacerlo. El valor por defecto de este parámetro en XGBoost es 1.

Lo que se observa a nivel general es que para poder capturar información relevante teniendo en cuenta el desbalance de clases, el modelo necesita de una configuración de parámetros que le permita ser más conservador de lo que es por defecto.

5.2. Importancia de variables

El usar modelos de boosting como XGBoost eleva la capacidad de predicción pero implica a la vez que al interactuar unos árboles con otros no sea tan sencillo presentar la importancia de cada una de las variables dado que la misma no es constante en todos los árboles del modelo,

como si lo es en las regresiones u otros modelos lineales. Por otro lado, al tomar cada árbol del modelo las variables en distinta proporción y orden puede que la importancia de la variable no sea estática en ese sentido tampoco.

Ante esto se recurre a los Shapley Additive Explanations, más conocidos como SHAP Values, para poder volver este tipo de datos comprensibles en el contexto. Lo que plantea este enfoque es entender a cada predicción como la suma de la contribución atribuida a cada variable usada por el modelo:

$$g(z') = \emptyset_0 \sum_{i=1}^{M} \emptyset_i z'_i \qquad (3)$$

donde $\emptyset_0$ es la probabilidad de base, M es el conjunto de variables utilizadas por el modelo y $\emptyset_i$ es la contribución de la característica z'_i a la predicción $g(z')$ entregada por el modelo (Lundberg & Lee, 2017).

Este método para la medición de importancia de las variables implica que si una de las variables adquiere mayor importancia en la predicción su valor SHAP debe aumentar. Valores SHAP mayores a cero indican un aumento en la probabilidad de que el modelo prediga 1 mientras que valores menores a cero indican una reducción de esa probabilidad. Valores cercanos a cero indican que la variable en cuestión no es importante a la hora de la predicción.

En este trabajo los valores SHAP se calcularon utilizando, para cada uno de los nueve modelos, la configuración de XGBoost aplicando HP2 con mejor Mean AUC RSKF - siendo estos los presentados en la Tabla 12. Se realizó un gráfico donde puede verse la importancia media de cada una de las variables, para poder ver cuáles de ellas ejercen más peso, y un gráfico de tipo swarm donde puede verse cuál es el impacto de las variables según el valor que toman en cada instancia y en qué sentido cada uno de esos valores influye en la predicción final.

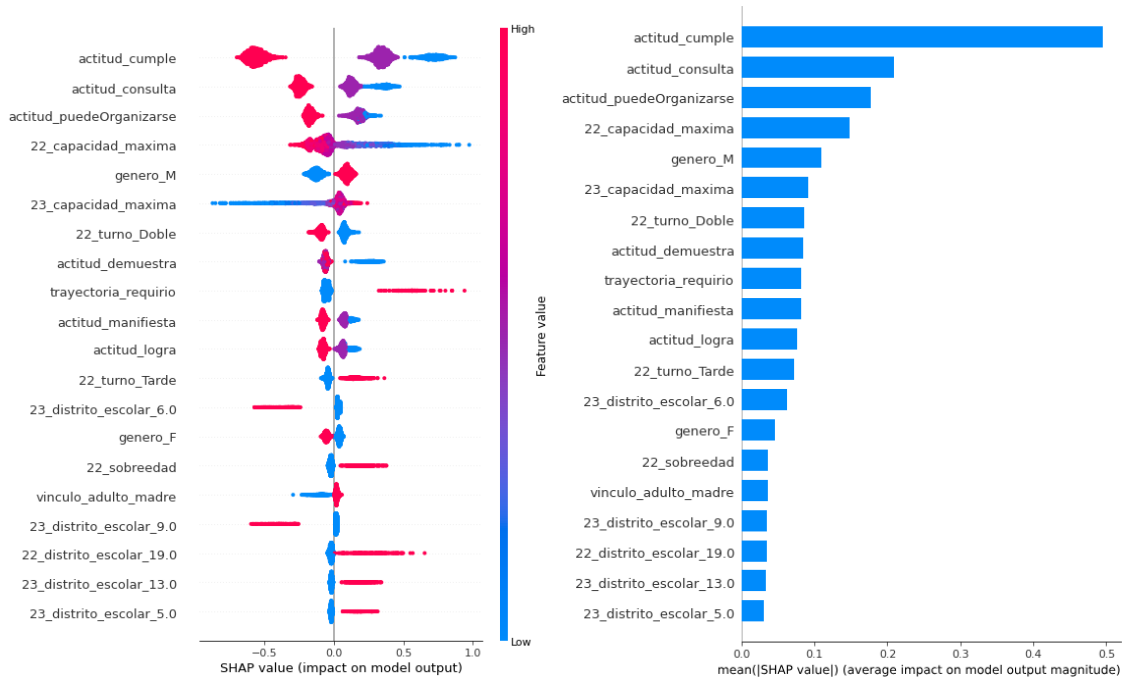
A continuación se presentan los gráficos de los modelos más relevantes dentro de la experimentación con el segundo set de hiperparámetros (HP2). El conjunto total de estos gráficos puede encontrarse en el Apéndice B del presente trabajo. Los gráficos correspondientes a los valores SHAP obtenidos en la utilización de HP1 se presentan en el Apéndice A.

El **Modelo 1**, donde sólo se incluyen datos de la matrícula, devuelve como su variable con mayor importancia el género masculino (*genero_M*). Lo que propone este modelo es que los varones tienen mayor probabilidad de repetir que los demás géneros y que además, es esta variable la más importante a la hora de la predicción. Se observa además que aquellos que asisten a una

escuela sin ninguna orientación tendrían una mayor probabilidad de repetir frente a sus pares que asistan a una escuela con una orientación en su plan de estudios. De manera coincidente con lo planteado en la bibliografía, asistir a una escuela de doble jornada, incluso en el año previo a la medición (2022), está asociado a una menor probabilidad de repitencia.

Al incorporar datos del PPS en el **Modelo 2**, el género del estudiante queda relegado en términos de importancia versus las variables de actitud que provee el informe. El top 3 de variables con mayor importancia es ocupado por las variables *actitud_cumple*, *actitud_consulta* y *actitud_puedeOrganizarse*.

Figura 14 - SHAP Values del modelo M2 con el set de hiperparámetros número 2



Como puede verse en la Figura 14, los valores obtenidos en *actitud* muestran una separación considerablemente precisa de los casos. Aquellos estudiantes que mayor cumplimiento en las tareas (*actitud_cumple)* tienen una probabilidad mucho menor de repetir que aquellos que no la demuestran. Algo similar sucede con las preguntas acerca de la frecuencia en que los estudiantes consultan a los profesores y como logran organizarse en sus tareas (*actitud_consulta)*. Los datos provistos acerca de la capacidad de convivencia con sus pares y el vínculo de sus familias con la institución no mostraron ser relevantes para la predicción.

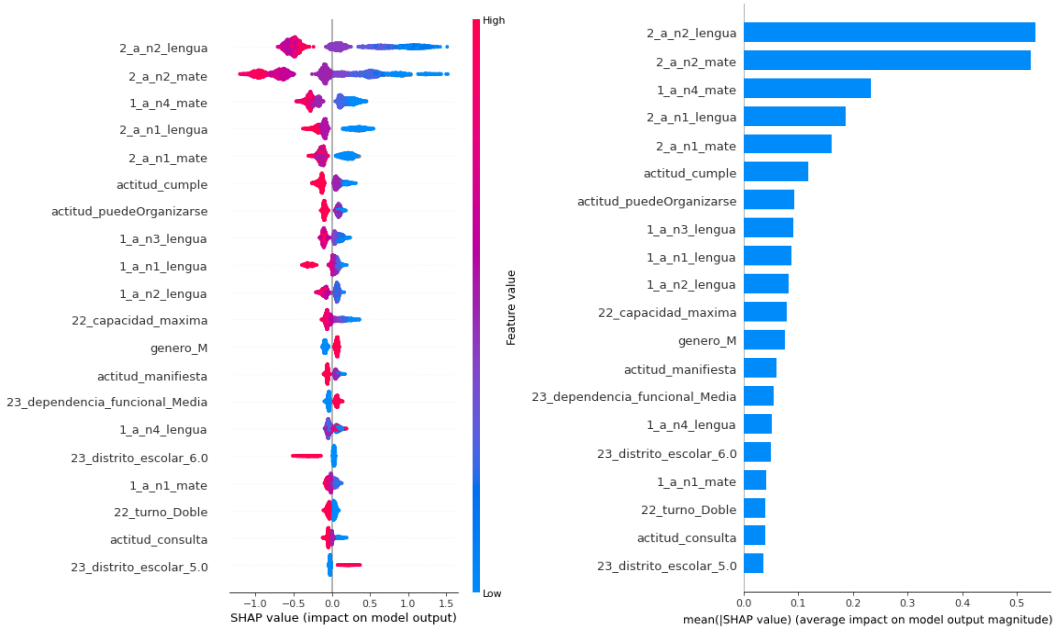
Más allá de lo particular de las variables del PPS en sí, resulta interesante como la inclusión de las mismas mejora la capacidad del modelo de separar más consistentemente el impacto del género y el turno, variables clave del primer modelo.

Esto indicaría que existe una relación entre las columnas género y turno y aquellas del PPS, y que al incluirse el segundo grupo en el modelo el algoritmo detecta de mejor manera los casos teniendo esto un correlato en los valores de AUC-ROC obtenidos.

En términos del **Modelo 3**, que incluye información de pases de institución a los ya utilizados en el Modelo 2, no se observa un gran impacto en el AUC-ROC ni se encuentran cambios en las variables más importantes obtenidas de la predicción sobre el conjunto de testeo.

Distinto es el caso del **Modelo 4**. Al incorporar información de las calificaciones de los estudiantes no solo se observa un aumento significativo en los valores de AUC-ROC, sino que también se produce una alteración de las variables de mayor importancia en el modelo.

Figura 15 - SHAP Values del modelo M4 con el set de hiperparámetros número 2



Como se observa en la Figura 15, las notas referentes al ciclo lectivo 2023 (aquellas que llevan como prefijo 2_) son las que tienen mayor importancia y el modelo parece estar pudiendo separar de manera considerable entre grupos denotando que mayores calificaciones implican una menor probabilidad de repetir.

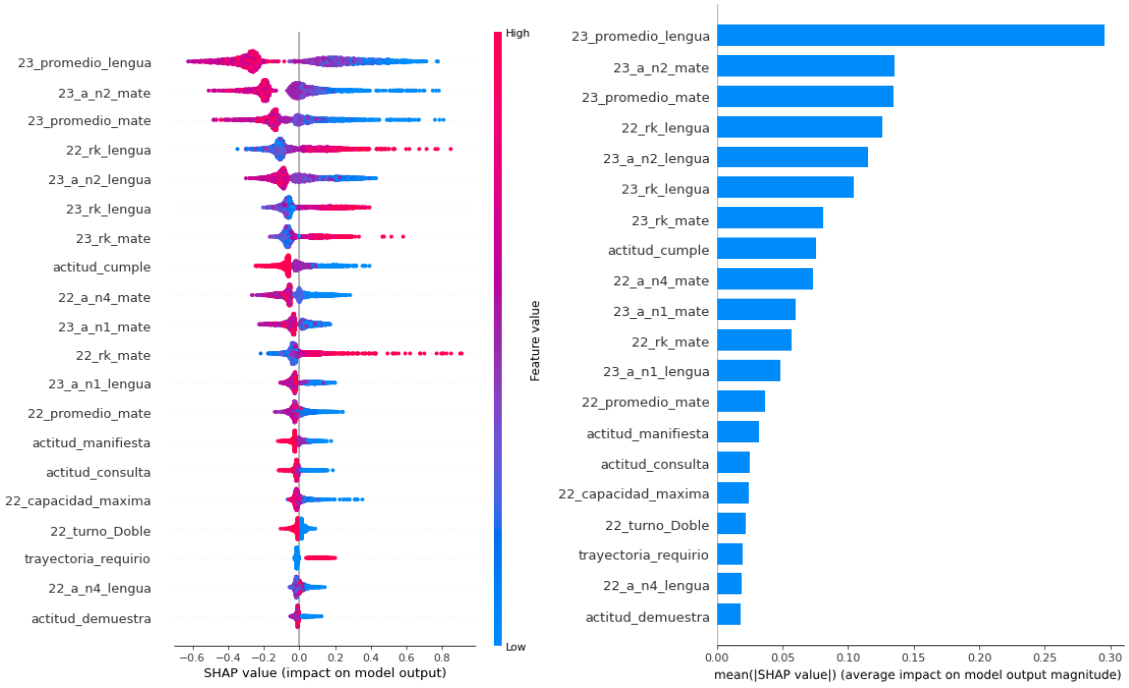
En términos de las materias, es importante notar que las notas de lengua parecen ser las más importantes del primer año de los estudiantes, tanto en el caso del primer bimestre como en el del segundo (*2_a_n1_lengua* y *2_a_n2_lengua*, respectivamente).

Lo que es particularmente interesante de este modelo es que la inclusión de las calificaciones tiene un efecto también en la importancia de las variables principales de los modelos 1, 2 y 3. Si

bien estas se mantienen en el top 20, su SHAP value queda mucho más cerca del 0 indicando que su importancia relativa versus las calificaciones es muy baja.

Cuando el **Modelo 4B** incorpora información de calificaciones relativas, usando promedios y rankings, son estas variables las que se posicionan como las más importantes a la hora de la predicción. Como se observa en la Figura 16, el incrementar la cantidad de variables relacionadas a las notas desplaza a casi todas las variables que tenían protagonismo en el Modelo 3, quedando dentro de las 20 más relevantes únicamente variables actitudinales del PPS.

Figura 16 - SHAP Values del modelo M4B con el set de hiperparámetros número 2



En general, alineado con lo que se espera en la literatura y lo que deriva del propio concepto de repitencia, aquellos estudiantes con menores calificaciones tienen una mayor probabilidad de repitencia.

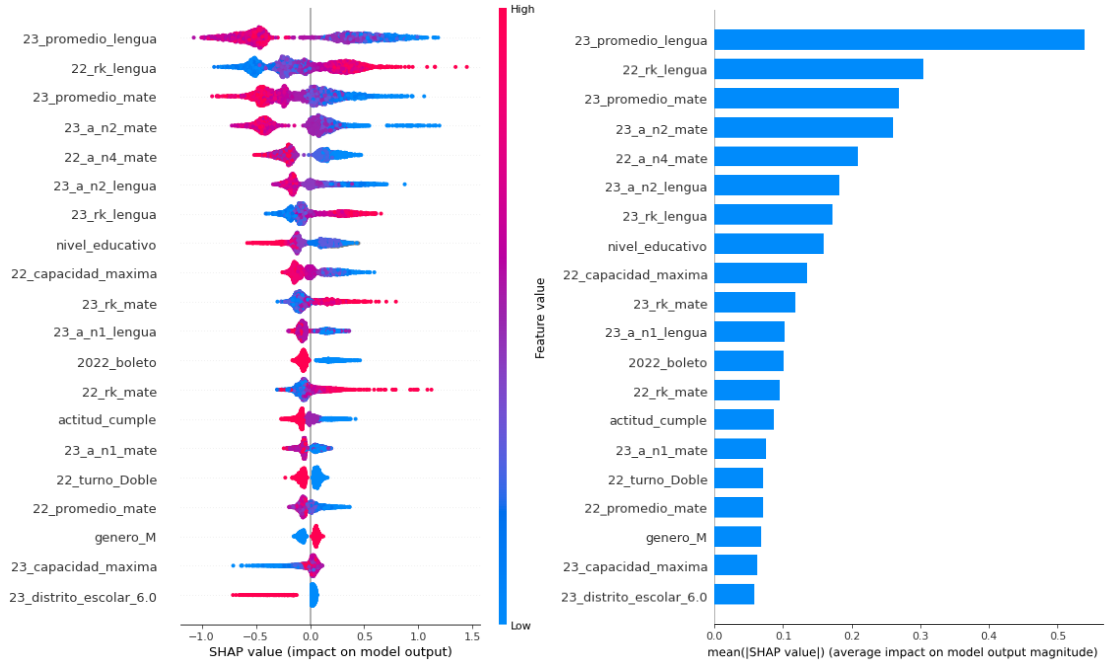
En el **Modelo 4C**, donde se usan datos de calificaciones pero solo en su formato promedio y ranking y no en su desagregación bimestral, las variables de mayor importancia son nuevamente los promedios y los rankings.

La incorporación de datos relativos a servicios en el **Modelo 5** da como resultado el modelo con mejor AUC-ROC de todos los entrenados. A nivel de importancia de variables sin embargo no se presentan grandes cambios versus el Modelo 4B. De las nuevas variables incluidas solo 2022_boleto – que indica si el estudiante fue beneficiario del boleto estudiantil en ese año –

ingresa al conjunto de las 20 variables más importantes mostrando que el tener este beneficio reduce la probabilidad de repitencia aunque con un valor de SHAP de magnitud reducida.

En última instancia, el **Modelo 6** incorpora información de los responsables de los estudiantes. Este modelo tiene un AUC-ROC en el test levemente menor al de M4 (-0.04pp).

Figura 17 - SHAP Values del modelo M6 con el set de hiperparámetros número 2



Lo que se observa en la Figura 17 es que, teniendo en este punto todas las variables presentadas para el trabajo, las calificaciones siguen siendo el conjunto de variables más importantes. Sin embargo, nuevamente las correspondientes a lengua se sitúan por encima de las correspondientes a matemática, contradiciendo la hipótesis de que las calificaciones de matemática poseían el doble de poder predictor que las de lengua (Duncan, et al., 2007).

Mas allá de esto, hay ciertas variables de otros datasets que la bibliografía postulaba como relevantes y exhiben serlo incluso con el protagonismo de las calificaciones. El nivel educativo de los responsables, que se incorpora en este modelo, muestra un comportamiento considerablemente definido identificando que un mayor nivel de estudios alcanzado por los adultos reduce la probabilidad de repitencia de los estudiantes.

Asimismo, la capacidad máxima del curso al que asisten los estudiantes en su último año de primaria (*22_capacidad_maxima*) tiene considerable importancia, aunque en el sentido contrario de lo esperado por Vries, León, Romero, & Hernández (2011). En su trabajo acerca de los cursos universitarios, los autores postulaban que cursos más grandes tienen un impacto negativo en la trayectoria educativa de los alumnos. Por lo que se observa en la Figura 17,

pertenecer a un curso más numeroso en séptimo grado reduce las probabilidades de repitencia en primer año. En el caso de la capacidad máxima de 2023, correspondiente al primer año de la educación de nivel secundario, una mayor cantidad de estudiantes por curso parecería estar relacionada con una mayor probabilidad de repitencia.

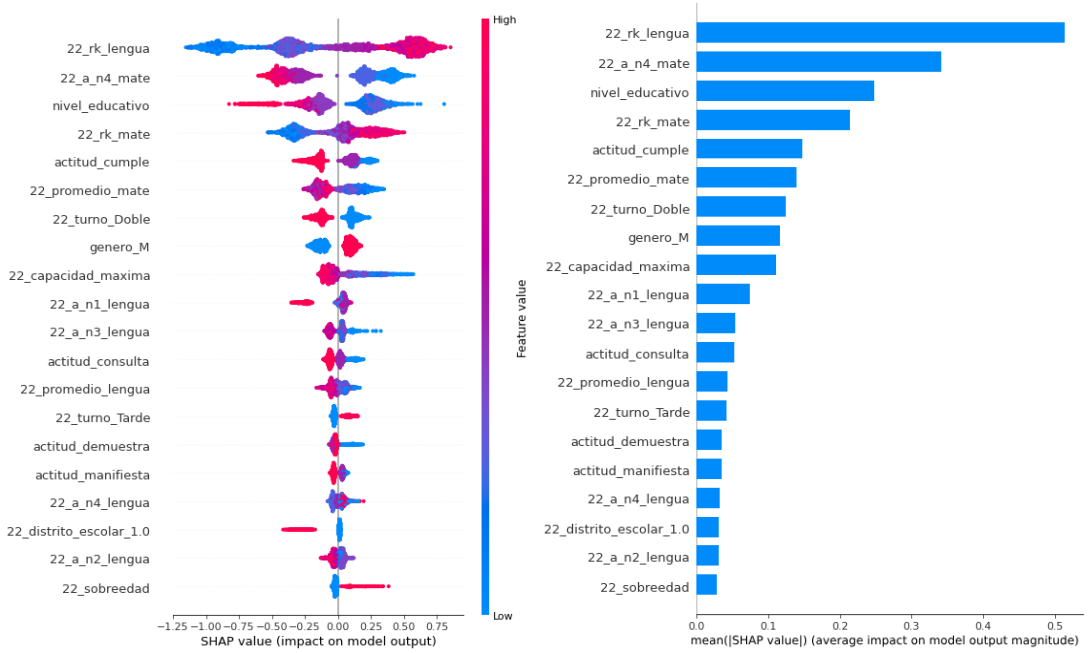
Comparando el tamaño de las clases de 2022, de nivel primario, y 2023, de nivel secundario, lo que se encuentra es que en promedio los cursos cuentan con una cantidad similar de estudiantes: 24.7 para el nivel primario y 26.3 para el nivel secundario. En este sentido, dada la similitud en el tamaño promedio de las clases se descartaría que la diferencia en el impacto de estas dos variables derive de cuestiones estructurales de los cursos de ambos niveles.

Partiendo de esto cabría estudiar si el espíritu de la educación primaria, donde el foco está puesto más fuertemente en el componente social de la educación, podría verse favorecido por la convivencia con más estudiantes. Mientras tanto en la secundaria, donde la educación está más enfocada a resultados, cursos menos numerosos podrían reducir el riesgo de repitencia en tanto permiten mayor interacción de los alumnos con los profesores para reforzar conocimientos.

En lo que hace a la ubicación geográfica aparecen dentro del conjunto de variables de mayor importancia en algunos de los modelos los distritos escolares 5 y 6. En el caso distrito 6, asistir a una escuela en ese espacio geográfico está asociado a una disminución de la probabilidad de repitencia, mientras que en el caso del distrito 5 el efecto es el contrario. Siendo que el distrito 5 refiere al área geográfica de la Comuna 4, la relación resulta esperable en tanto había allí una mayor concentración de casos de repitentes. En lo que respecta al distrito 6, correspondiente a los barrios de Boedo, San Cristóbal y sectores de Almagro y Balvanera, se estima que esta relación podría venir de las bajas tasas de repitencia de la Comuna 3 donde se ubican San Cristóbal y Balvanera.

En último lugar se analiza la información correspondiente al **Modelo M6B**, representada en la Figura 18.

Figura 18 - SHAP Values del modelo M6B con el set de hiperparámetros número 2



En este caso la ausencia de las variables relacionadas al ciclo lectivo 2023 hace que dentro del top 20 se encuentre mayor protagonismo de datos que hacen a las notas de séptimo grado del alumno, información acerca de su persona y su entorno. Datos como el nivel educativo de los padres, algunas de las preguntas de actitud y el género así como datos de los programas de los que los estudiantes son beneficiarios cobran mayor protagonismo.

Se observa que la capacidad del modelo para encontrar patrones entre los valores de las variables es aceptable. En el gráfico de la izquierda puede verse una separación clara de los colores rojo y azul indicando que las variables son buenas en la tarea de dividir a los casos de quienes repiten de aquellos que no lo hacen.

Siendo que el mismo cuenta solo con información del nivel primario, un AUC-ROC en test del 81,76% parece ser razonable como base para tener una aproximación de los alumnos a dar seguimiento desde el principio del ciclo lectivo.

6. Conclusiones

Se propuso como objetivo de este trabajo desarrollar un modelo basado en aprendizaje automático que permitiese identificar patrones comunes entre los estudiantes que repiten su primer año de educación secundaria en el sistema de educación público de la Ciudad de Buenos Aires. Adicionalmente, se buscó comprender si el Puente Primario Secundario provee de información actitudinal y conductual relevante para la predicción de este fenómeno.

Lo que se encontró en la experimentación es que los dos conjuntos de datos con mayor impacto en la precisión de la predicción fueron los correspondientes al PPS y a las calificaciones. La inclusión de cada uno de estos conjuntos produjo un aumento en AUC-ROC de casi 10 puntos porcentuales versus el modelo anterior a su incorporación.

En el caso puntual del PPS el conjunto de variables con mayor impacto en la predicción fueron aquellas relacionadas a la actitud del estudiante, indicando que estudiantes con mayor valoración en cuanto a su capacidad de organizarse, su actitud frente al aprendizaje y el nivel de consultas a los docentes son aquellos con menores probabilidades de repetir.

De todos los modelos evaluados, aquel con mayor AUC-ROC en test fue el Modelo 5 que contuvo información acerca de: matrícula, PPS, pases de institución, calificaciones bimestrales, promedios de calificaciones, ranking de desempeño relativo y datos de servicios. El AUC-ROC sobre test de este modelo usando la segunda configuración de hiperparámetros fue de 90,45% mientras que su valor en entrenamiento con la estrategia de validación de RSKF fue de 89,10%.

6.1. Limitaciones y posibles mejoras

Uno de los grandes desafíos para la construcción del presente trabajo fue la obtención de los datos. Si bien se obtuvo acceso a las bases de la plataforma de gestión escolar, lo que no fue posible obtener fueron datos estandarizados de domicilio y de ingresos, variables centrales para el modelo. En versiones futuras de la investigación, se esperaría que incorporar esas variables permita obtener mejores resultados.

Las mismas no pudieron ser tomadas de datos agregados dado que el trabajo trata con información a nivel alumno, identificable únicamente a través de su documento de cara a datos públicos y con su id_alumno a nivel interno. En este sentido, la información de acceso público, con la finalidad de resguardar la privacidad de los ciudadanos, no provee el nivel de detalle necesario para el trabajo.

En este mismo sentido se espera que el trabajar con mayor cantidad de cohortes, algo que se dará en tanto finalicen nuevos ciclos lectivos, permita crear modelos que tengan una mayor capacidad de generalización.

Por otro lado, sería de gran utilidad la incorporación al modelo de información acerca de instituciones privadas. Actualmente, su exclusión se debe a la falta de uniformidad y periodicidad en la información que reportan al Ministerio de Educación, que dificulta la comparación e integración de los datos. Además, la ausencia de una implementación

homogénea del Puente Primario Secundario en las escuelas privadas limita la disponibilidad de información clave para el modelo.

Transversal a escuelas públicas y privadas surge también la necesidad de comenzar a evaluar estrategias alternativas de análisis de trayectorias en tanto el concepto de repitencia se encuentra siendo desafiado por nuevas maneras de diseñar los caminos de aprendizaje de los estudiantes.

Del análisis de los modelos utilizados en esta tesis se deriva una potencial nueva línea de investigación vinculada a la variable capacidad máxima. Llama particularmente la atención la relación inversa entre capacidad máxima del curso al que asiste el estudiante y la probabilidad de repitencia que se asocia a ello dependiendo de si se trata de nivel primerio o secundario cuando la teoría en general espera que convivir con más estudiantes aumente la probabilidad de repitencia. Como se comentó en el análisis de los resultados en la sección 5.2 se observa que cursos más pequeños están asociados con un valor SHAP positivo en el nivel primario mientras exhiben una tendencia inversa en el nivel secundario.

Una primera aproximación realizada en torno a esto es que podría darse que el efecto del tamaño del curso no sea independiente del nivel y que esto se vincule con la diferencia en el rol que ocupa la sociabilidad en cada caso. La hipótesis es que una reducción en la probabilidad de repitencia en cursos más grandes del nivel primario podría tener que ver con un mayor protagonismo de la sociabilidad en ese contexto, favorecida por el volumen de gente con la que se convive e interactúa. Mientras tanto, un valor SHAP negativo asociado a cursos más numerosos en nivel secundario podrían tener que ver con una mirada más orientada a los resultados donde una menor cantidad de compañeros de curso puede ser beneficioso en tanto permite mayor acceso a los docentes y, consecuentemente, una mayor posibilidad de comprender los contenidos.

6.2. Aplicaciones

El objetivo principal de este trabajo fue desarrollar una herramienta que brinde previsibilidad a docentes y directivos sobre la probabilidad de que los estudiantes de primer año, a quienes aún no conocen en profundidad, repitan el curso.

Además, este estudio buscó evaluar la relevancia de las variables capturadas por el Informe PPS en la predicción del rendimiento académico. Los resultados sugieren que esta información es valiosa para anticipar la trayectoria escolar de los estudiantes. En este sentido, el modelo también podría utilizarse como una herramienta para analizar la utilidad de las preguntas

incluidas en el informe, evaluar posibles mejoras en su contenido y en la forma en que se recopilan los datos.

En cuanto a sus aplicaciones concretas, se identifican dos usos principales para el modelo:

1. Planificación pedagógica al inicio del ciclo lectivo

El modelo M6B ofrece información clave para la organización de las clases desde el comienzo del año escolar. El mismo permitiría a los docentes y directivos tomar decisiones estratégicas, como la disposición del aula para mejorar la concentración de los estudiantes con mayor riesgo de repitencia.

También podría contribuir al diseño de estrategias de apoyo en materias troncales, con el fin de reforzar el aprendizaje de los alumnos más vulnerables y reducir su riesgo de repetir.

2. Intervenciones durante el año escolar

Los demás modelos desarrollados en este trabajo pueden utilizarse para realizar predicciones a mitad de año. Compartir esta información con docentes y directivos permitiría tomar medidas en el corto y mediano plazo, activando estrategias de intervención temprana. Esto no solo ayudaría a evaluar la evolución de los estudiantes que, por sus antecedentes en primaria, ya presentaban un riesgo de repitencia, sino también identificar a aquellos que, sin un historial con tendencia a la repitencia, empiezan a mostrar dificultades en el transcurso del año.

De cara al docente, la predicción de mitad de año es crucial en tanto le permite entender a nivel general si el curso en que imparte clases se encuentra bien encaminado o si merece una revisión de los métodos y materiales de enseñanza ante una concentración de casos con altas probabilidades de repitencia.

La combinación de ambos modelos resulta fundamental: mientras que el primero permite establecer un diagnóstico inicial y delinear estrategias generales para todo el año, el segundo posibilita monitorear la evolución de los estudiantes y ajustar las intervenciones según su desempeño real.

Por otro lado, la incorporación de este modelo como herramienta de trabajo para las escuelas permitiría generar en estas una voluntad de reportar más y mejor información de sus estudiantes en tanto queda en evidencia que la misma se está utilizando para darles información útil para el planeamiento pedagógico y el trabajo en las aulas.

6.3. Conclusión

Este trabajo ha demostrado la viabilidad de utilizar modelos de aprendizaje automático para predecir la repitencia escolar en el primer año del nivel secundario, ofreciendo una herramienta con potencial para mejorar la planificación pedagógica y la intervención temprana en las trayectorias educativas.

A partir de los resultados obtenidos, y teniendo en cuenta la consistencia de las predicciones y las variables utilizadas en todos ellos, se concluye que la experimentación ha resultado favorable. Incluso ante la falta de datos importantes como el contexto socioeconómico, los modelos han podido predecir correctamente los casos de repitencia. El AUC-ROC obtenido en la utilización de HP2 indica que el modelo es estable y puede generalizar a datos desconocidos manteniendo la performance.

Más allá de sus resultados específicos, este estudio pone de manifiesto el valor del uso de datos educativos en la toma de decisiones dentro del sistema escolar. La disponibilidad de información en los sistemas de gestión educativa representa una oportunidad para fortalecer el seguimiento de los estudiantes y optimizar la asignación de recursos pedagógicos, siempre enmarcado en un enfoque que priorice la equidad y la inclusión.

Futuras investigaciones podrían profundizar en la optimización de los modelos predictivos, explorando nuevas variables y metodologías, así como evaluar el impacto real de la implementación de este tipo de herramientas en la dinámica escolar. La posibilidad de transformar los datos en conocimiento accionable para la mejora de las trayectorias educativas representa un camino prometedor hacia un sistema educativo más inclusivo y eficiente.

Referencias

Witten, D., Hastie, T., Gareth, J., & Tibshirani, R. (2017). *An Introduction to Statistical Learning.* Springer.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining.* Pearson.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree. *Proceedings of the 22nd ACM SIGKDD International Conference, KDD '16* (págs. 785-794). New York, NY: ACM.

Torres, J., Acevedo, D., & Gallo, L. (2015). Causas y consecuencias de la deserción y repitencia escolar: una visión general en el contexto Latinoamericano. *Cultura Educación y Sociedad*, 157-187.

Gamboa-Cruzado, J., Alvarez-Cuellar, C., Martinez-Medina, S., Turpo Chaparro, J., Sifuentes, D., & Rodríguez Kong, M. (2023). Predicción de repitencias a nivel escolar usando Machine Learning: Una revisión sistemática. 45-65.

Salas, G., Vigorito, A., & Failache, E. (2015). Desempeños en salud y desarrollo en la infancia y trayectorias educativas de los adolescentes en Uruguay: Un estudio en base a datos de panel. *Instituto de Economía*.

Hsin, A., & Xie, Y. (2011). Social Determinants and Consequences of Children's Non-Cognitive Skills: An Exploratory Analysis. *Spring Meeting of the ISA RC28.* UK: University of Essex.

Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A., & Klebanov, P. (2007). School readiness and later achievement. *Development Psychology*, 1428-1446.

Boado, M., & Fernández, T. (2010). Trayectorias académicas y experiencias laborales de los jóvenes uruguayos. El panel PISA 2003-2007. *Facultad de Ciencias Sociales, Montevideo*.

Méndez-Errico, L., & Ramos, X. (2015). Selección y logros educativos: ¿Por qué algunos niños se quedan atrás? Evidencia para un país de ingresos medios. *Instituto de Economía, Facultad de Ciencias Económicas y de Administración. Universidad de la Reepública*.

Llambí, C., Messina, P., & Perera, M. (2009). Desigualdad de oportunidades y el rol del sistema educativo n los logros de los jóvenes uruguayos. *Documentos de Trabajo CINVE* .

Amarante, V., Colafranchesi, V., & Vigorito, A. (2011). La desigualdad en el acceso y logro educativo en Uruguay: ¿qué nos dicen los datos? *Documento de trabajo, Universidad de la República.*

Dari, N., Cervini, R., & Quiroz, S. (2021). Repitencia escolar y desempeño en ciencias en Argentina: Estudio multinivel con base en datos de PISA 2015. . *Revista de Educación*, 45-62.

Vries, W., León, P., Romero, J., & Hernández, I. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de educación superior*, 29-49.

Cardozo, S., Silveira, A., & Fonseca, B. (2022). Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de machine learning. *RLEE Nueva Época México*.

Hunt, J. (2008). Education and Economic Growth: Why Not? *Economic Policy Institute*.

Avalis, F., & Caro, N. (2024). El efecto de la educación en los ingresos en Argentina a través de un estudio econométrico. *Revista Educación*.

Edo, M., Marchionni, M., & Garganta, S. (2017). Compulsory education laws or incentives from conditional cash transfer programs? Explaining the rise in secondary school attendance rate in Argentina. *Education Policy Analysis Archives*.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *NeurIPS.*

Heckman, J. (2008). Schools, skills, and synapses. *Economic Inquiry*, 289-324.

Dari, N., Cervini, R., & Quiroz, S. (2018). Repitencia y rendimiento escolar en la educación primaria de América Latina. Los datos del TERCE. En R. Cervini, *El fracaso escolar: diferentes perspectivas disciplinarias* (págs. 197-215).

UNESCO. (2012). Repetition and drop-out in secondary education: A global perspective. PARIS: UNESCO.

Torre, E., Paparella, C., & D'Alessandre, V. (2023). *Sistema de alerta temprana para prevenir la exclusión escolar.* CIPPEC.

Steinberg, C., Tiramonti, G., & Ziegler, S. (2018). *Políticas Educativas para transformar la Educación Secundaria. Estudio de casos a nivel provincial.* Buenos Aires: UNICEF-FLACSO.

Apéndice A. Valores SHAP de la experimentación con el primer set de hiperparámetros

Figura A1 - SHAP Values del modelo M1 con el set de hiperparámetros número 1

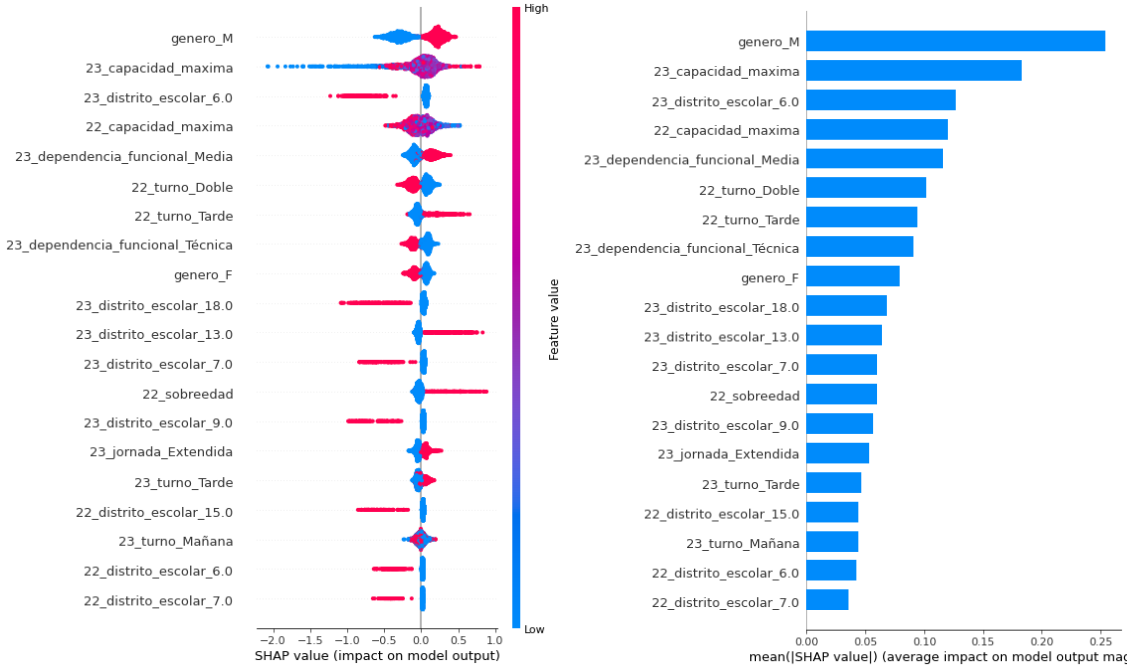


Figura A2 - SHAP Values del modelo M2 con el set de hiperparámetros número 1

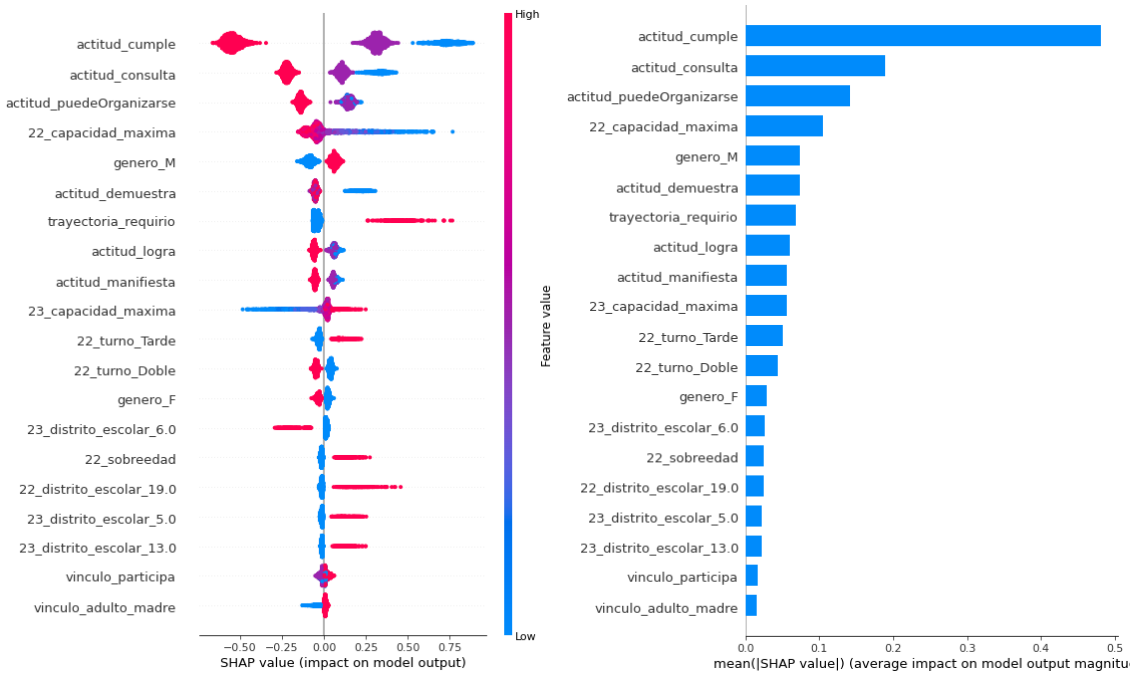


Figura A3 – SHAP Values del modelo M3 con el set de hiperparámetros número 1

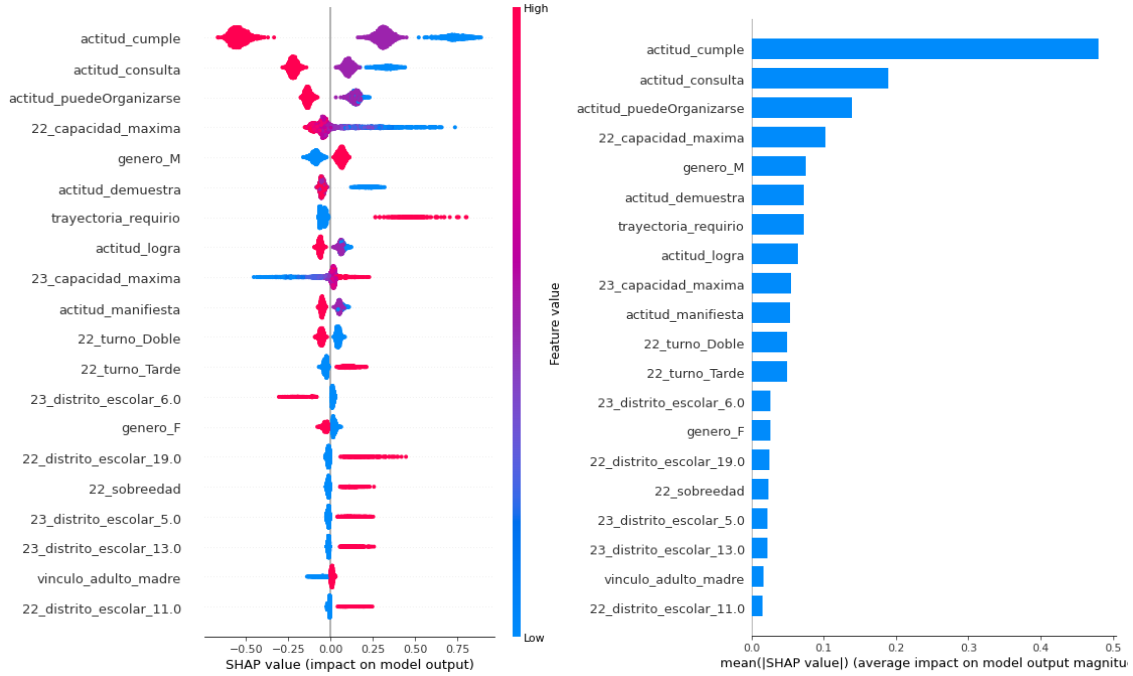


Figura A4 – SHAP Values del modelo M4 con el set de hiperparámetros número 1

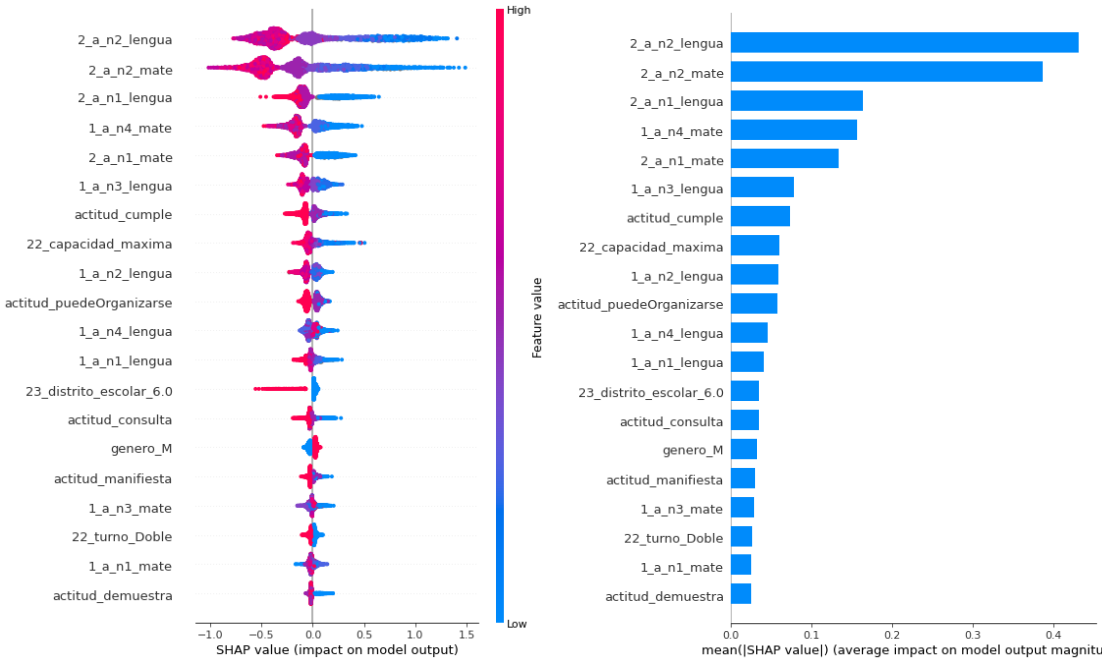


Figura A5 – SHAP Values del modelo M4B con el set de hiperparámetros número 1

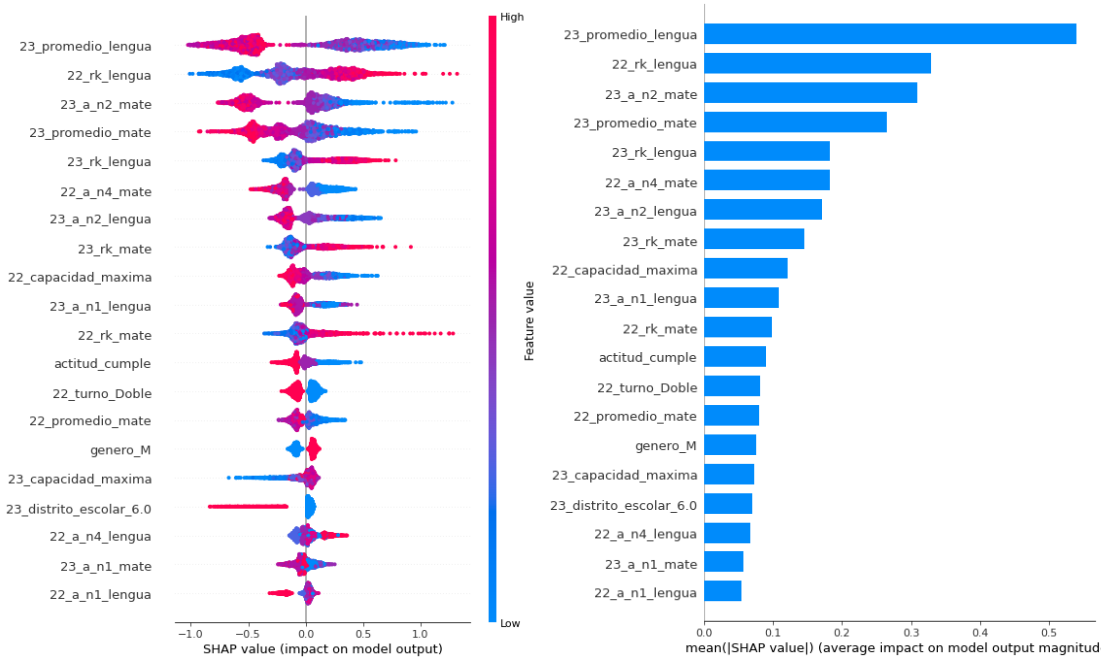


Figura A6 – SHAP Values del modelo M4C con el set de hiperparámetros número 1

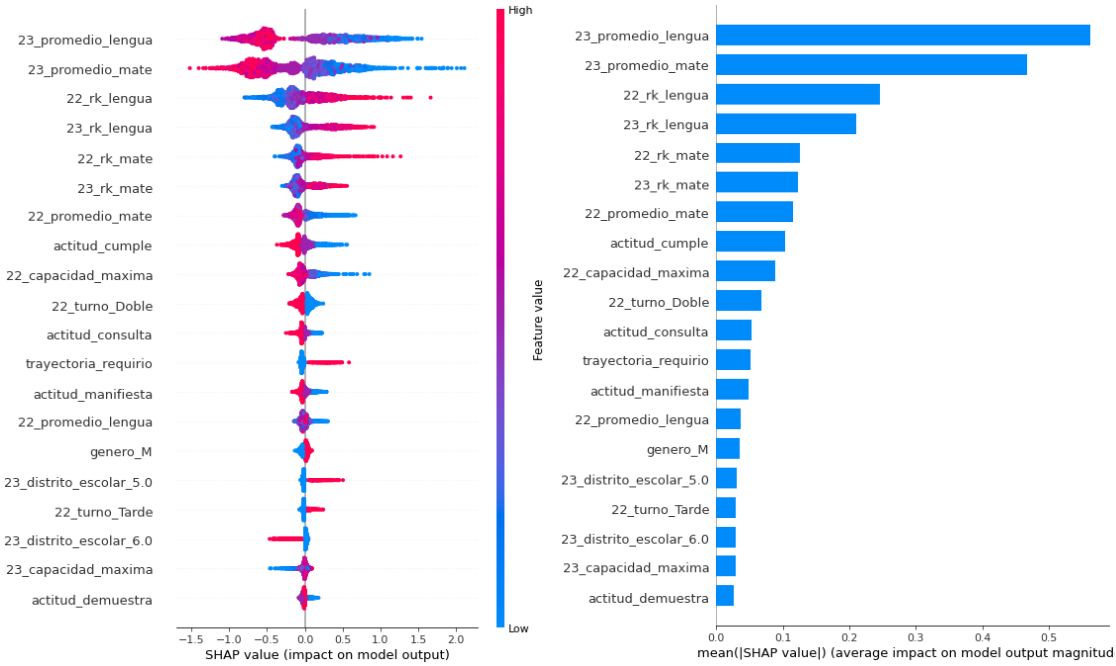


Figura A7 – SHAP Values del modelo M5 con el set de hiperparámetros número 1

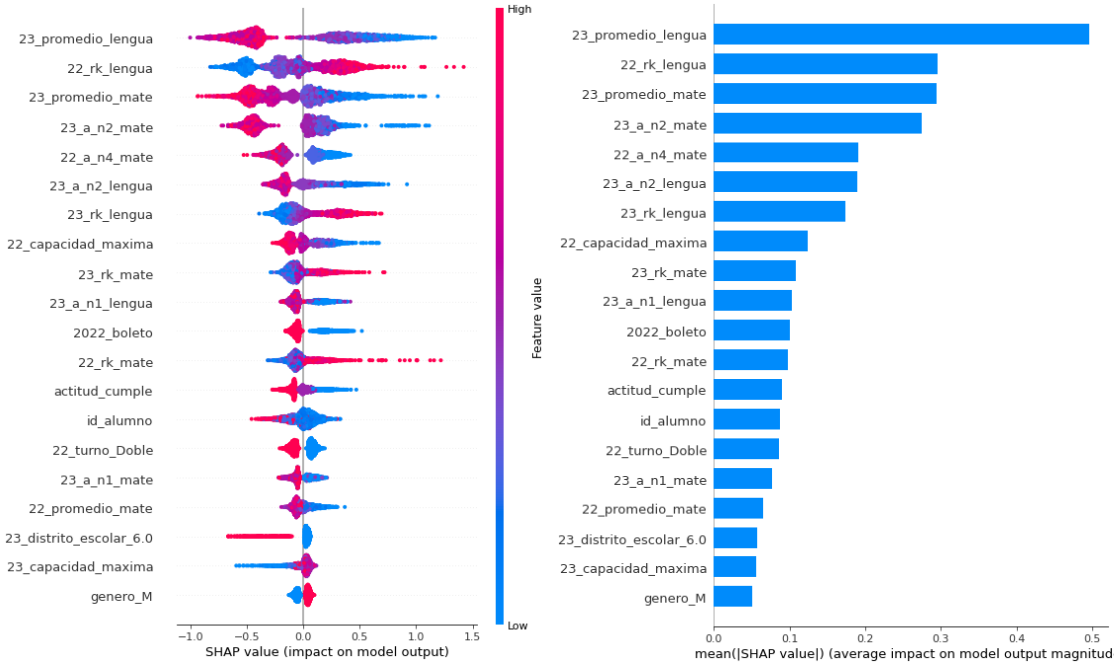


Figura A8 – SHAP Values del modelo M6 con el set de hiperparámetros número 1

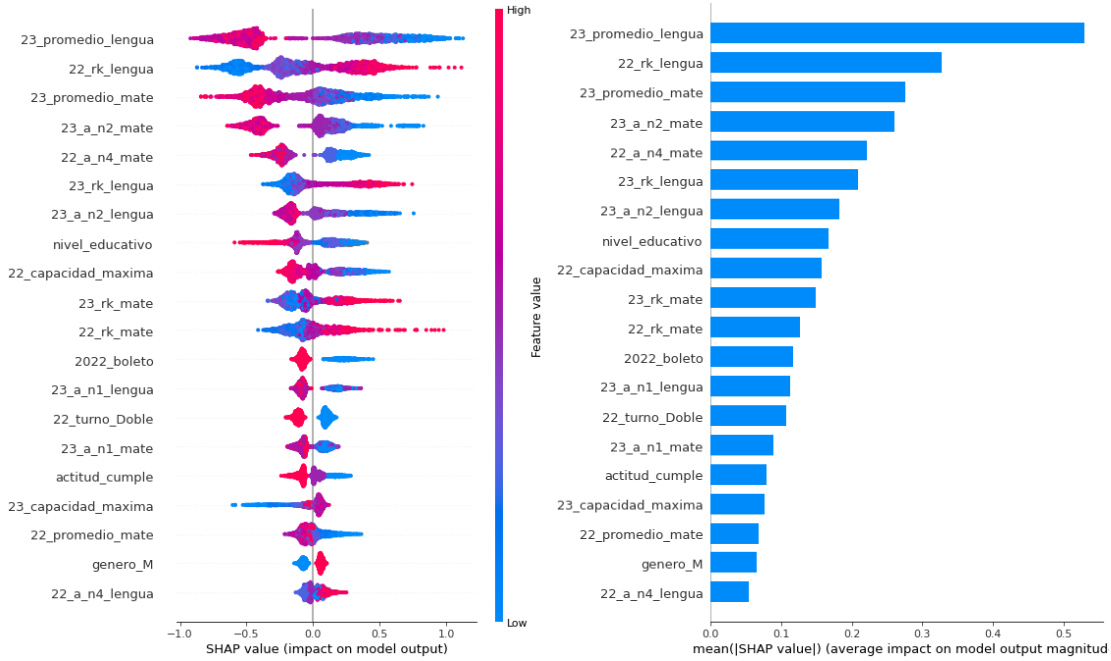
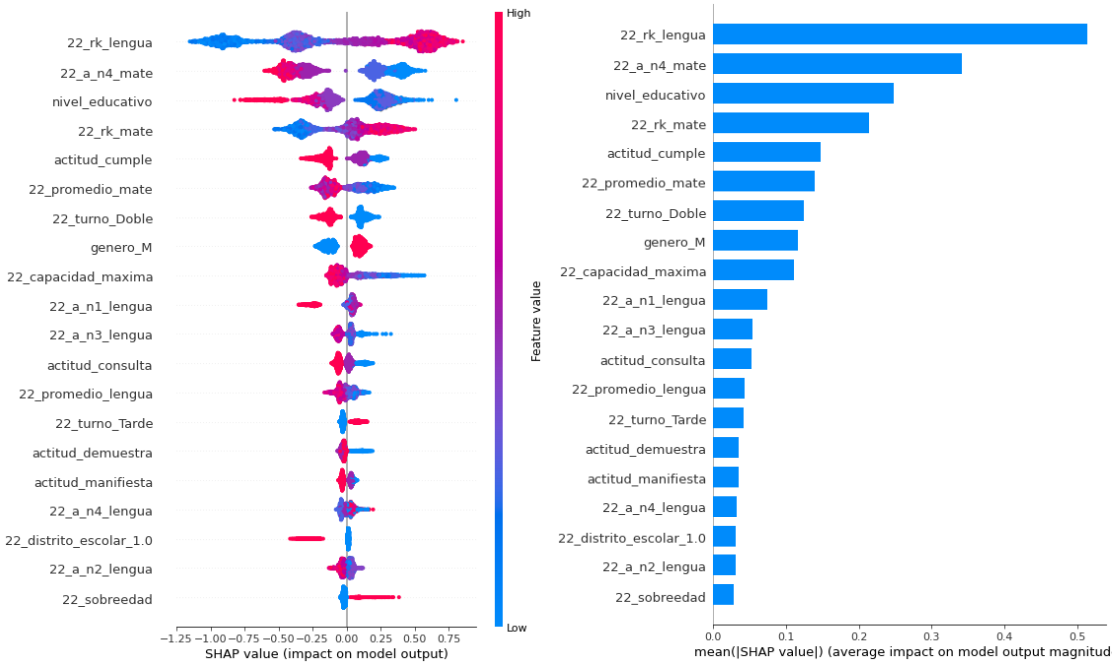


Figura A9 – SHAP Values del modelo M6B con el set de hiperparámetros número 1



Apéndice B. Valores SHAP de la experimentación con el segundo set de hiperparámetros

Figura B1 - SHAP Values del modelo M1 con el set de hiperparámetros número 2

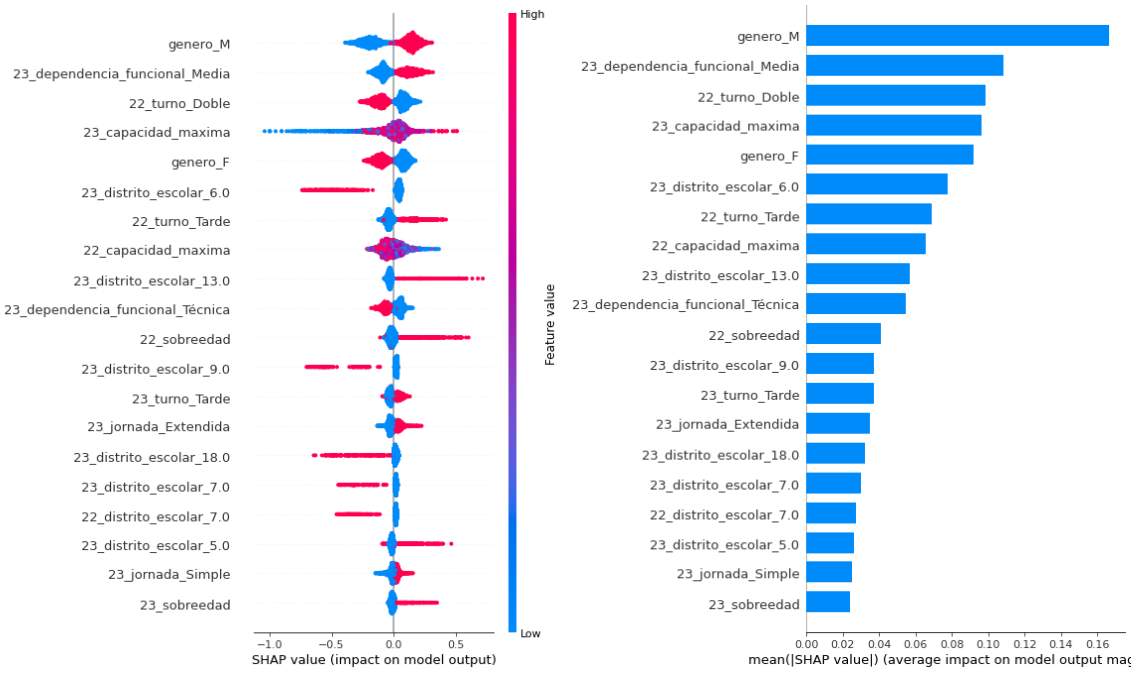


Figura B2 - SHAP Values del modelo M2 con el set de hiperparámetros número 2

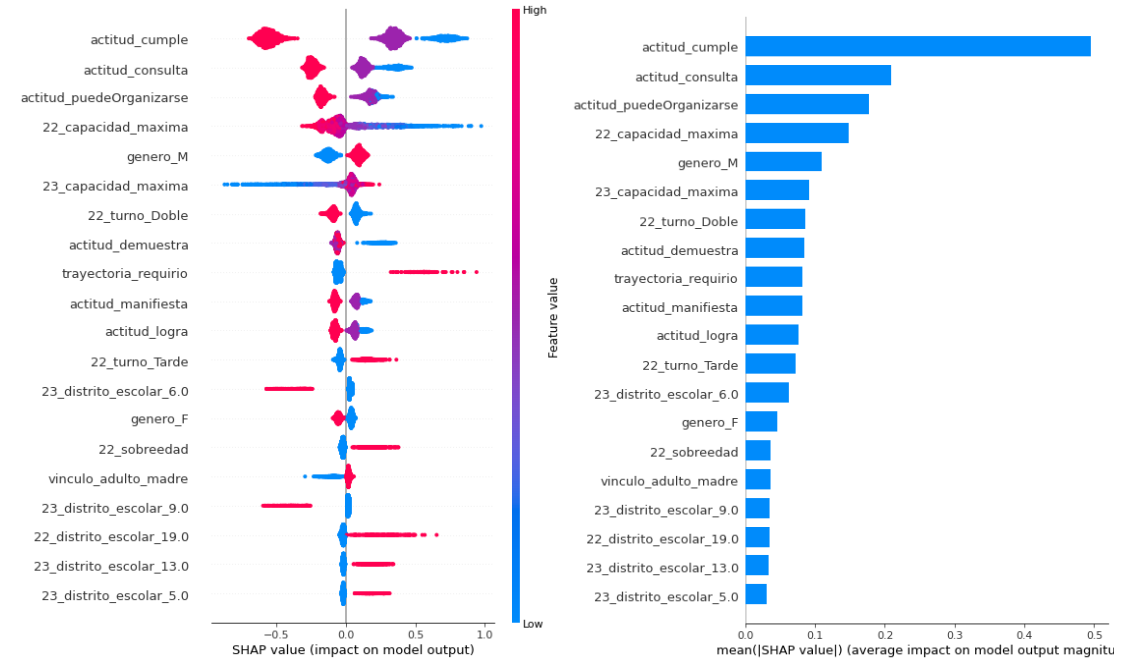


Figura B3 - SHAP Values del modelo M3 con el set de hiperparámetros número 2

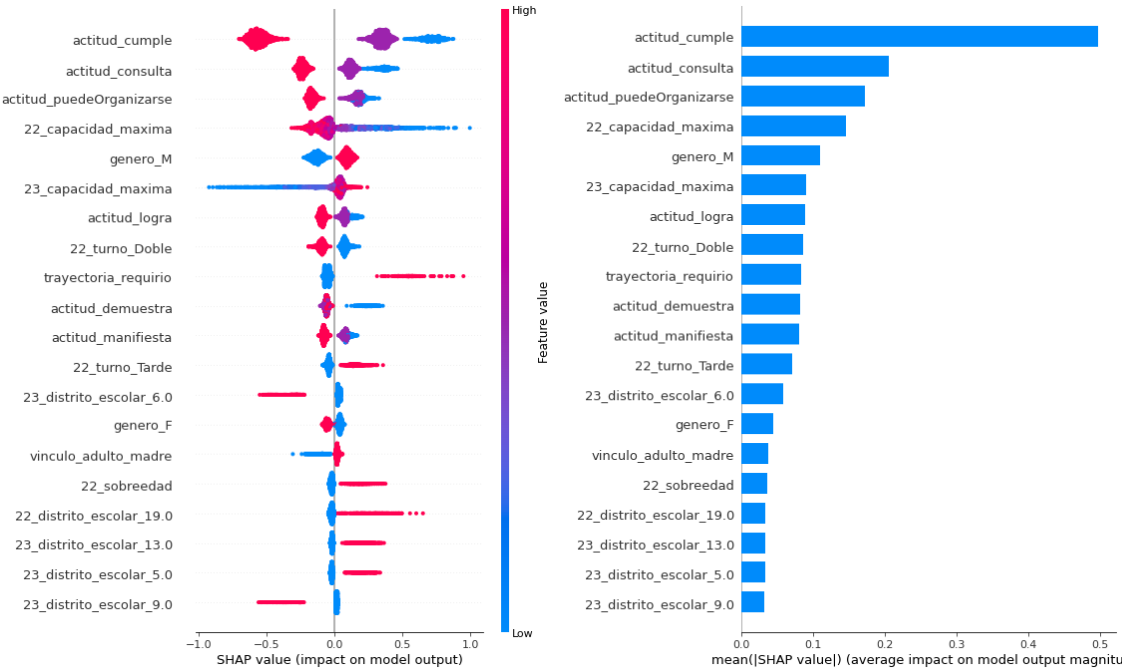


Figura B4 - SHAP Values del modelo M4 con el set de hiperparámetros número 2

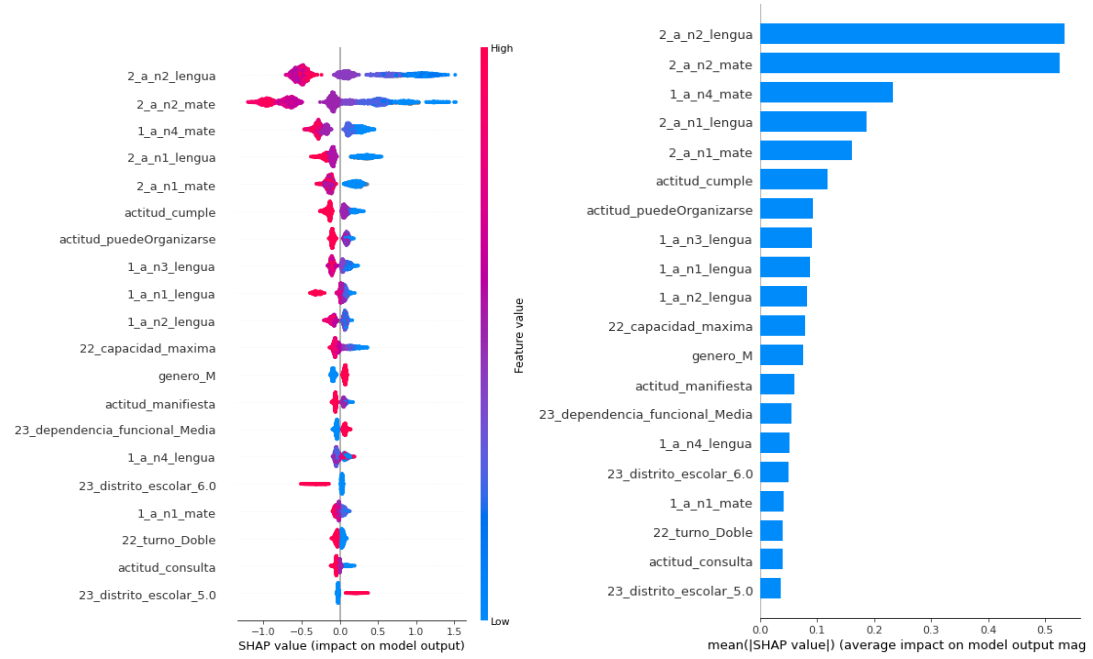


Figura B5 - SHAP Values del modelo M4B con el set de hiperparámetros número 2

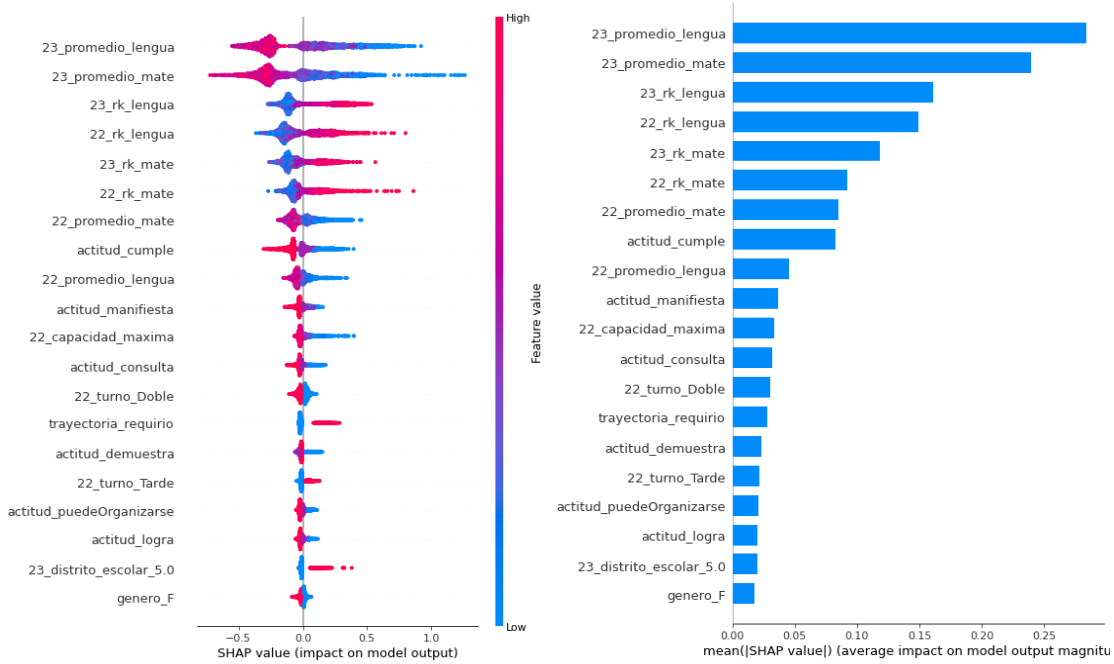


Figura B6 - SHAP Values del modelo M4C con el set de hiperparámetros número 2

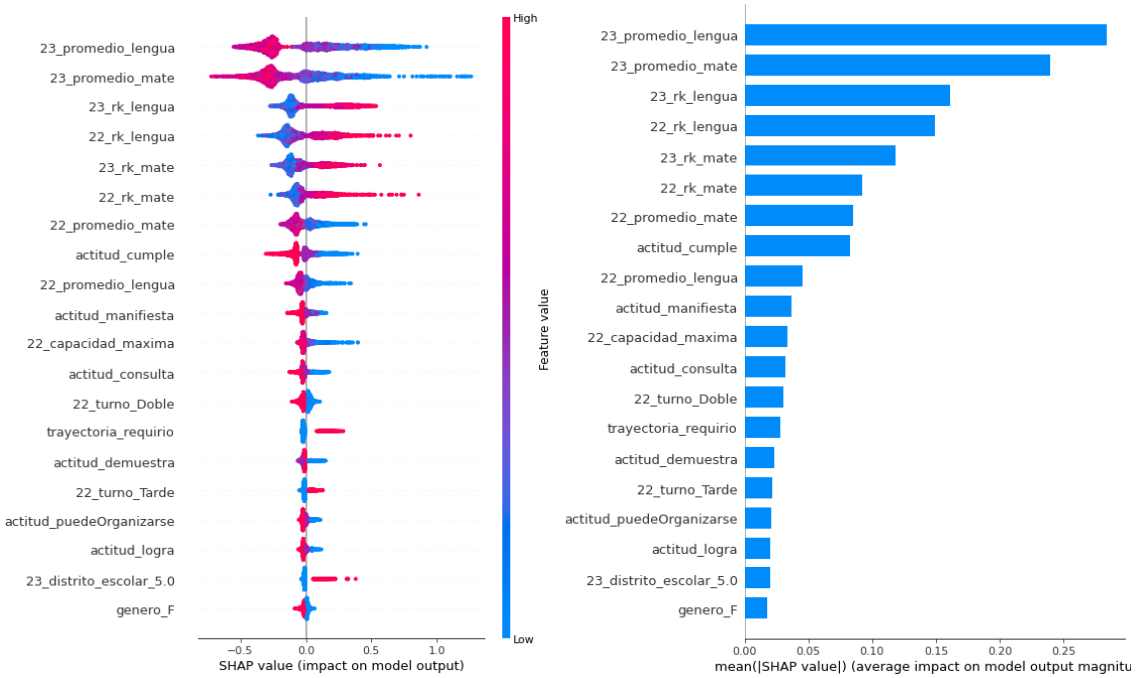


Figura B7 - SHAP Values del modelo M5 con el set de hiperparámetros número 2

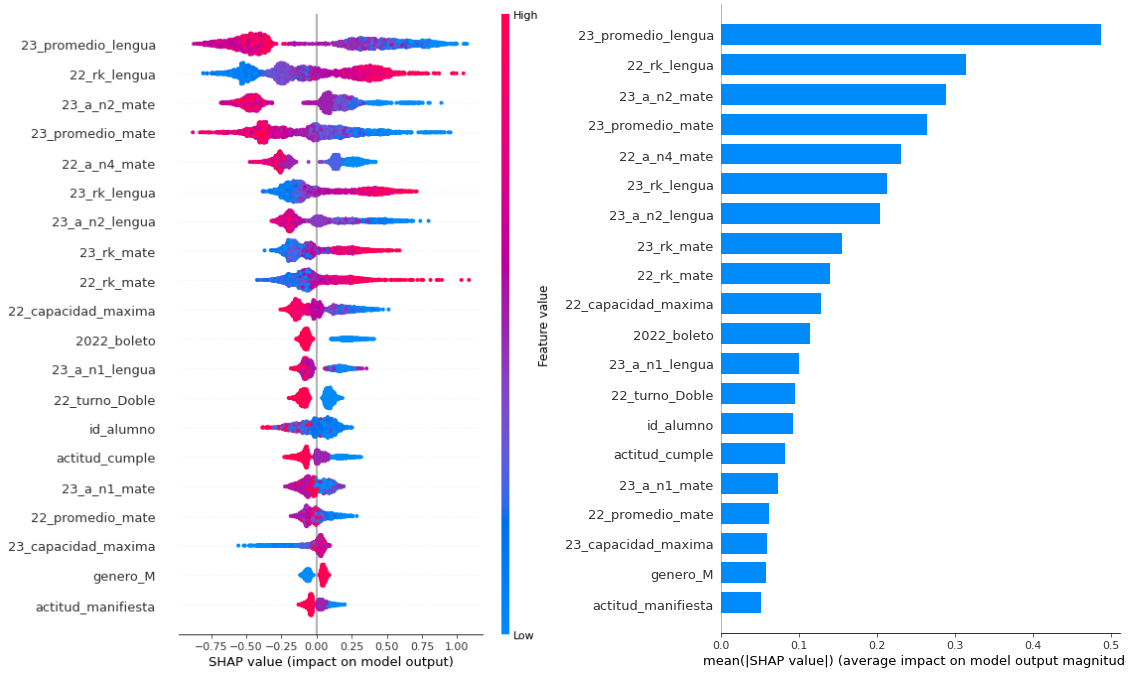


Figura B8 - SHAP Values del modelo M6 con el set de hiperparámetros número 2

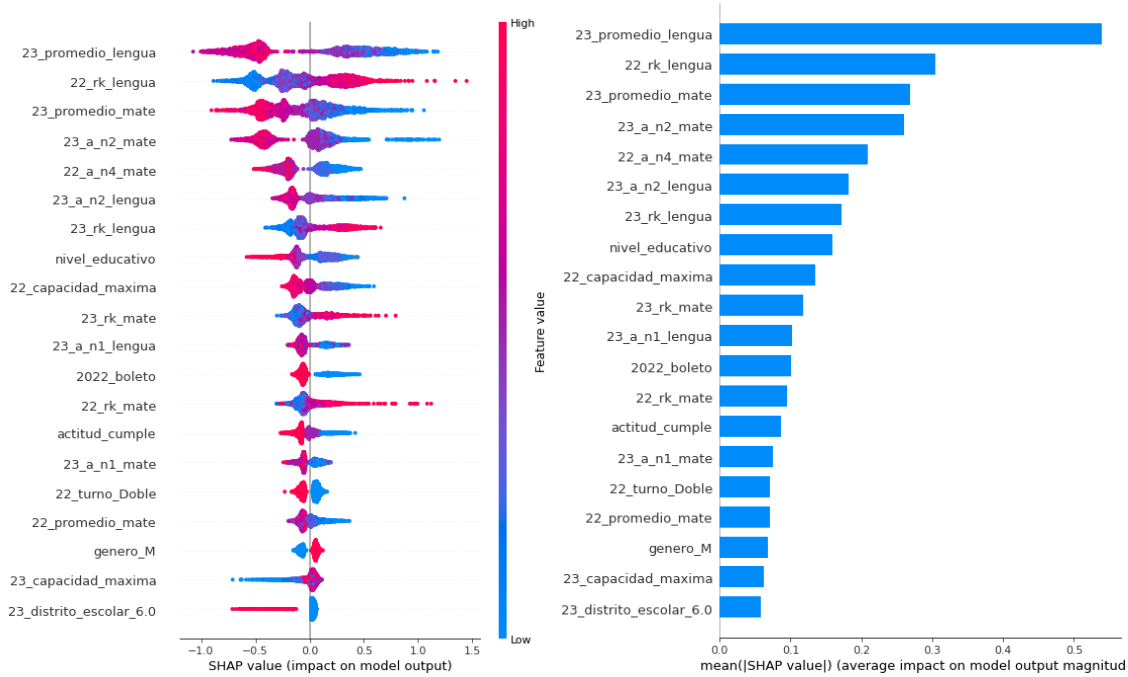
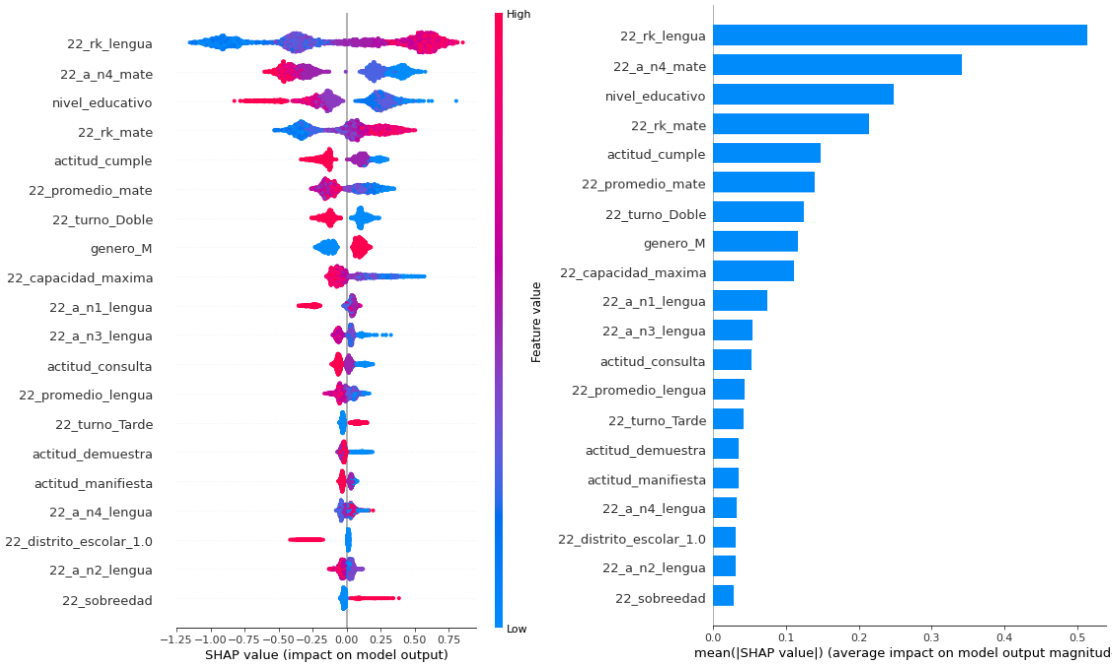
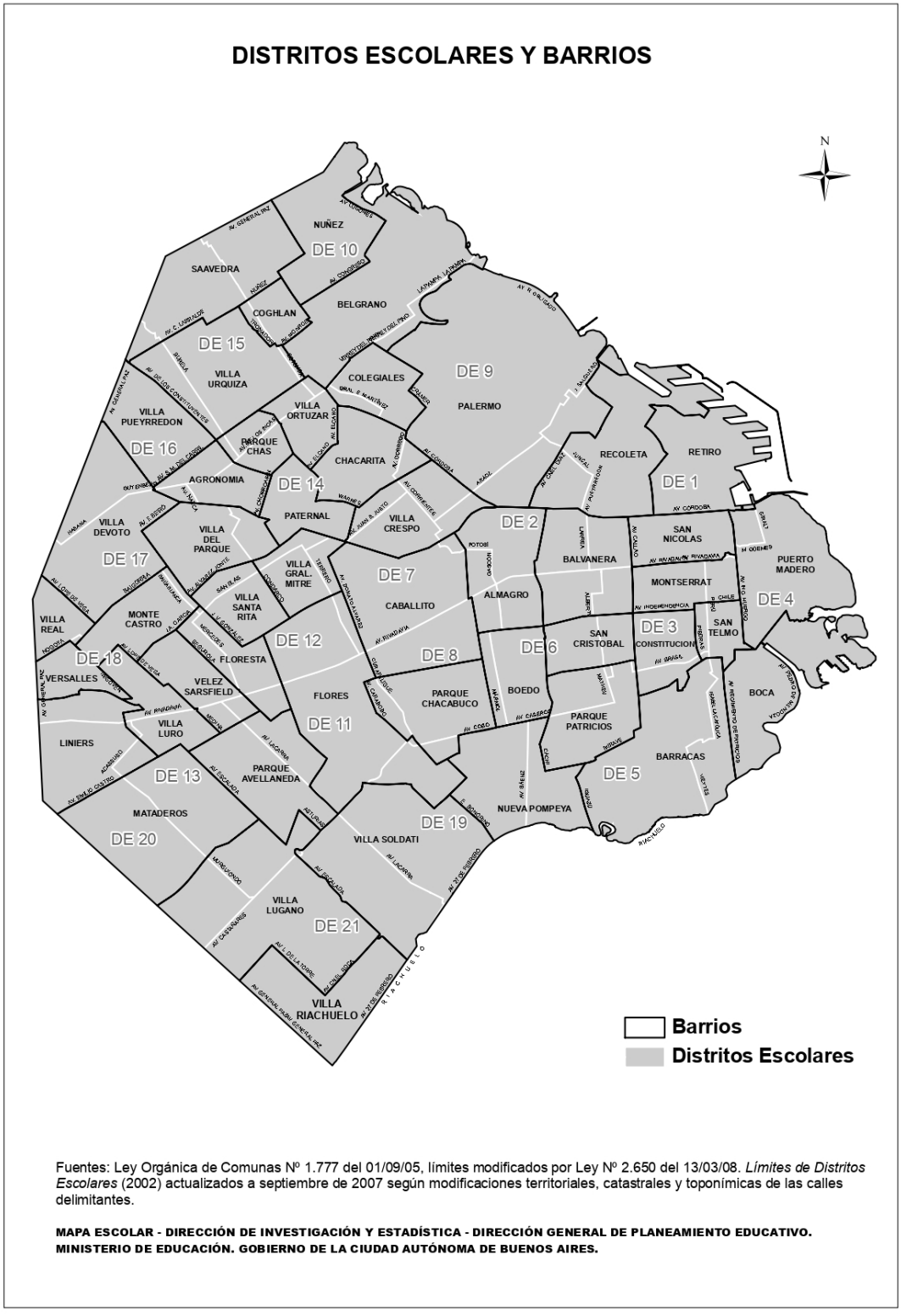


Figura B9 - SHAP Values del modelo M6B con el set de hiperparámetros número 2



Apéndice C. Mapa de distritos escolares y barrios de la Ciudad de Buenos Aires



Fuentes: Ley Orgánica de Comunas Nº 1.777 del 01/09/05, límites modificados por Ley Nº 2.650 del 13/03/08. *Límites de Distritos Escolares* (2002) actualizados a septiembre de 2007 según modificaciones territoriales, catastrales y toponímicas de las calles delimitantes.

MAPA ESCOLAR - DIRECCIÓN DE INVESTIGACIÓN Y ESTADÍSTICA - DIRECCIÓN GENERAL DE PLANEAMIENTO EDUCATIVO. MINISTERIO DE EDUCACIÓN. GOBIERNO DE LA CIUDAD AUTÓNOMA DE BUENOS AIRES.