UNIVERSIDAD DE LA LAGUNA

Análisis Predictivo del Rendimiento Académico en Estudiantes Canarios de Primaria usando Modelos Lasso

Autor: Sebastián Barbuzano Narváez

Tutores: Carlos Gabriel Bethencourt Marrero y Ángel Simón Marrero Llinares

2023/2024

Escuela de Doctorado y Estudios de Posgrado

Máster Universitario en Análisis Aplicado para las Ciencias Sociales

RESUMEN

La educación es crucial para el desarrollo individual y social, promoviendo el crecimiento económico, progreso cultural y mejora de la calidad de vida. Predecir el rendimiento académico es esencial para identificar a estudiantes en riesgo de fracaso escolar y tomar medidas preventivas. Este trabajo tiene como objetivo construir modelos predictivos del rendimiento académico de estudiantes de primaria utilizando datos de la Agencia Canaria de Calidad Universitaria y Evaluación Educativa (ACCUEE) de estudiantes en tercer grado en 2015/16 y en sexto grado en 2018/19. La metodología emplea técnicas de *machine learning*, específicamente modelos Lasso. Se desarrollaron tres modelos para predecir calificaciones en matemáticas, lengua y rendimiento académico general (agregado). Los principales resultados son: el modelo agregado tiene la mejor capacidad predictiva debido a su completitud, a pesar de tener la muestra más pequeña; los signos y magnitudes de las variables relevantes son consistentes entre modelos y con la literatura, indicando una explicación uniforme para matemáticas, lengua y el agregado; el efecto inercia es significativo en todos los modelos, capturando la mayor parte de la variabilidad; las variables explicativas son más relevantes en conjunto que por separado; la coincidencia de los signos en las predicciones cruzadas muestra una fuerte interseccionalidad entre variables; y las principales diferencias entre modelos están en las magnitudes, no en los signos.

Palabras clave: Lasso; rendimiento académico; machine learning; políticas públicas.

ABSTRACT

Education is crucial for individual and social development, promoting economic growth, cultural progress, and improved quality of life. Predicting academic performance is essential for identifying students at risk of academic failure and taking preventive measures. This work aims to build predictive models of primary school students' academic performance, using data from the Canary Agency for University Quality and Educational Evaluation (ACCUEE) of students in the third grade in 2015/16 and in the sixth grade in 2018/19. The methodology employs machine learning techniques, specifically Lasso models. Three models were developed to predict grades in mathematics, language, and overall academic performance (aggregate). The main results are: the aggregate model has the best predictive capacity due to its completeness, despite having the smallest sample; the signs and magnitudes of the relevant variables are consistent across models and with the literature, indicating a uniform explanation for mathematics, language, and the aggregate; the inertia effect is significant in all models, capturing most of the variability; the explanatory variables are more relevant together than separately; the matching of signs in cross-predictions shows a strong intersectionality between variables; and the main differences between models are in magnitudes, not in signs.

Keywords: Lasso; academic performance; machine learning; public policies.

ÍNDICE

1. INTRODUCCIÓN

La educación se constituye como uno de los pilares fundamentales para el desarrollo individual y social, impulsando dimensiones como el crecimiento económico, el progreso cultural y la mejora de la calidad de vida. Su relevancia radica en su capacidad para reducir las desigualdades existentes y generar oportunidades. Por todo ello, predecir el rendimiento académico de los estudiantes es crucial, especialmente para identificar a aquellos en riesgo de fracaso escolar y actuar de manera preventiva. Contar con información precisa sobre los factores que influyen en el rendimiento académico puede ayudar en la puesta en marcha de políticas públicas eficaces que disminuyan las consecuencias negativas del fracaso escolar y contribuyan a la construcción de un futuro más próspero para los individuos en particular, y para la sociedad en general.

El objetivo del presente trabajo, por tanto, es la construcción de un modelo predictivo del rendimiento académico futuro de los estudiantes de educación primaria.

Con la predicción del rendimiento académico de los estudiantes, este trabajo pretende conocer cuál es la incidencia de distintos factores que influyen en el entorno educativo de los estudiantes para establecer un punto de partida en la toma de decisiones encaminadas a su corrección y/o mejora, con el fin de contribuir al progreso del sistema educativo.

En este trabajo se han utilizado datos longitudinales provenientes de las evaluaciones de diagnóstico llevadas a cabo por la Agencia Canaria de Calidad Universitaria y Evaluación Educativa (ACCUEE). Particularmente, se han utilizado datos censales de un mismo conjunto de estudiantes encuestados en 3º de primaria en el curso 2015/16 y, 3 años después, en 6º de primaria en el curso 2018/19. El caso de Canarias es especialmente llamativo ya que, aparte de ser una de las Comunidades Autónomas con menor nivel de desarrollo económico (Chinea, 2018), también es una de las que alcanzan una menor puntuación en las evaluaciones internacionales sobre educación, tales como PISA (Hernández, 2018).

La metodología que se ha utilizado se basa en las técnicas de *machine learning*. En concreto, se han estimado un conjunto de modelos predictivos *Lasso* (*least absolute shrinkage and selection operator*, en inglés) en los que la variable a predecir es el rendimiento educativo del estudiante en 6º de primaria y las variables explicativas son distintos factores del mismo estudiante en 3º de primaria.

En primer lugar, se especificaron tres modelos de predicción distintos: uno para prever las calificaciones en la asignatura de matemáticas, otro para prever las calificaciones en la asignatura de lengua, y un tercero destinado a predecir el rendimiento académico general, calculado como la media de las calificaciones en matemáticas, lengua e inglés. Cada modelo se sometió a una fase inicial de filtrado de variables mediante la metodología de selección *Lasso Cross Validation*.

En segundo lugar, los modelos de matemáticas y lengua fueron evaluados en una etapa de cruce de modelos, generando así dos modelos adicionales: el modelo de matemáticas se utilizó para predecir las calificaciones en lengua y el modelo de lengua se empleó para predecir las calificaciones en matemáticas. El objetivo de este enfoque fue el de minimizar el margen de error en los cálculos y permitir una comparación más exhaustiva entre modelos. Este ejercicio permitió identificar aquellas variables más

relevantes en las diferentes asignaturas analizadas, así como las principales diferencias entre ellas.

Los principales resultados obtenidos son: primero, el modelo agregado demuestra la mejor capacidad predictiva debido a su mayor nivel de completitud, tanto en variables dependientes como independientes, a pesar de tener la muestra más pequeña para su entrenamiento. Segundo, los signos y magnitudes de las variables son robustos independientemente del modelo utilizado y consistentes con la literatura, indicando una uniformidad en la explicación aportada por cada variable, ya sea en matemáticas, lengua o el agregado, sin inconsistencias significativas entre asignaturas. Tercero, el efecto inercia descrito en la literatura aparece en todos los modelos con un alto nivel de significación, capturando la mayor parte de la variabilidad. Cuarto, la relevancia individual de cada variable explicativa es menor cuando se estudian por separado que cuando se analizan en conjunto. Por ello, los modelos para matemáticas y lengua mantienen sus capacidades predictivas incluso al alterar las variables dependientes para las que fueron diseñados. Quinto, el alto índice de coincidencia en los signos de las variables independientes en las predicciones cruzadas indica una fuerte interseccionalidad entre las diferentes variables. Sexto, las principales diferencias entre los modelos no están en los signos, sino en las magnitudes, lo que muestra que, en términos generales, las variables impactan en la misma dirección, independientemente de la asignatura estudiada.

Este trabajo se estructura de la siguiente manera: en la segunda sección se realiza una revisión sobre los principales trabajos de la literatura sobre predicción de rendimiento académico. En la tercera sección se describen los datos utilizados para el estudio. La cuarta sección detalla la metodología de *machine learning* utilizada para la predicción del rendimiento académico. La quinta sección presenta los principales resultados obtenidos. Por último, en la sexta sección se muestran las conclusiones.

2. REVISIÓN DE LA LITERATURA

En el ámbito académico, a lo largo de los años, ha existido un considerable interés por comprender las causas del rendimiento escolar, lo que ha generado una extensa investigación centrada en explorar y estudiar las variables que afectan al rendimiento de los estudiantes. La literatura ha abordado una amplia gama de factores con el objetivo de profundizar en los determinantes subyacentes del desempeño estudiantil. En este sentido, se han identificado una serie de variables que influyen significativamente en el rendimiento académico. Estos factores clave han sido abordados tanto desde el punto de vista teórico como empírico. Se pueden clasificar en factores: socioeconómicos y demográficos; de motivación y autoestima; ambiente familiar y escolar; habilidades cognitivas y socioemocionales; tecnológicos; culturales y contextuales; y programas individualizados o especializados de aprendizaje.

Respecto a los factores Socioeconómicos y Demográficos, la literatura ha evidenciado la existencia de una correlación significativa entre el rendimiento académico de los estudiantes y variables como el nivel educativo de los padres, el ingreso familiar, el estatus socioeconómico y la etnia. Por ejemplo, algunos estudios recientes (Muñoz et al., 2016; Bethencourt y Marrero, 2022) indican que los niños provenientes de familias con mayores recursos económicos tienden a mostrar un mejor rendimiento académico, habitualmente debido a su incidencia en el acceso a oportunidades educativas de calidad, tales como escuelas con mejores recursos, clases extracurriculares y tutores particulares (Fonseca et al., 2014). Además, el nivel educativo de los padres se asocia frecuentemente con un entorno doméstico más propicio para el aprendizaje, incluyendo la disponibilidad de libros y materiales educativos, así como una mayor capacidad para apoyar con las tareas escolares.

En cuanto a la motivación y la autoestima, está demostrado que los estudiantes que tienen una fuerte creencia en sus propias habilidades tienden a tener un mejor desempeño académico (Chilca, 2017). Los estudios más recientes, como Pendones (2021), encuentran que la autoestima es un aliciente en la motivación, pues permite que los estudiantes aborden las tareas con más confianza y resiliencia, mientras que aquellos con baja autoestima pueden experimentar ansiedad y evitación, afectando negativamente a su desempeño. Estos resultados han motivado el surgimiento de intervenciones del tipo de programas de mentoría o talleres de habilidades socioemocionales, con el objetivo de fomentar la motivación y autoestima, teniendo repercusión así de manera indirecta sobre el rendimiento académico (Ferrei et al., 2014). La teoría de la autodeterminación y la teoría de la expectativa-valor son marcos teóricos comúnmente utilizados para entender este aspecto (Usán y Salavera, 2018; Lamas, 2008; Broc, 2006), en la que la primera hace referencia a la capacidad del individuo a afrontar retos de manera motivada, con autonomía y competencia, mientras que la segunda expresa cómo las personas se encuentran más motivadas cuando creen que pueden tener éxito (como en el ambiente que plantea el entorno educativo), y cuando valoran la tarea (por lo que supone el éxito escolar, en futuros ambientes como el laboral).

El ambiente familiar, incluyendo el apoyo parental y el acceso a recursos educativos en el hogar y la calidad de la relación entre padres e hijos, influyen de manera sustancial en el rendimiento académico de los estudiantes (Apaza, 2015). Además, factores del entorno escolar como el tamaño de la clase, la calidad del profesorado y

el clima escolar también son determinantes en el rendimiento estudiantil (López, 2015; Morales y Zafra, 2013).

Los diversos estudios encuentran que las habilidades cognitivas como la inteligencia, la memoria y la capacidad de razonamiento son predictores importantes del rendimiento académico (Guzmán et al., 2017). Además, las habilidades socioemocionales como la inteligencia emocional, la resiliencia y la capacidad para manejar el estrés también desempeñan un papel crucial. (Berger et al., 2014; Muñoz et al., 2016; Russo, 2019; Mena et al., 2008).

Investigaciones actuales han explorado el impacto de la tecnología en el aprendizaje, incluyendo el uso de dispositivos móviles, plataformas en línea y herramientas de enseñanza digital (Javier et al., 2023; Means et al., 2009; Guamán et al., 2018). Estos estudios encuentran que la tecnología, pese a tener un enorme potencial como herramienta en materia educativa, también funciona como vía de acceso a otros entornos que entorpecen el desarrollo escolar (adicciones, deterioro de relaciones, florecimiento de intereses inapropiados…) (Cabrera et al., 2015). Además, internet y el resto de tecnologías son especialmente relevantes por ser generadores de desigualdades de oportunidades, a través de la brecha digital ocasionada por las barreras de entrada (tanto económicas como de conocimiento) que tienen algunas de las herramientas digitales a disposición del usuario hoy en día. Olivar et al. (2007), encuentran que la incidencia de la brecha tecnológica afecta en mayor medida por no poder acceder a ellas (barrera económica), que por la etapa de aprendizaje que suponen (barrera de conocimiento), dando ventaja por tanto a aquellos hogares que cuentan con estos recursos, sobre los que no. Esta brecha recibió gran atención por parte de los investigadores durante los años comprendidos entre 2019 y 2021, época en la que las dificultades educativas provocadas por la COVID-19 pusieron de manifiesto las necesidades de adaptación del alumnado a las tecnologías actuales, debido a que, durante este período, la digitalización no se comportó como un complementario a la educación tradicional, sino como un sustituto (ver Fernández et al., 2020).

Las diferencias culturales en las prácticas de crianza, las expectativas educativas y las normas sociales afectan también al desempeño estudiantil (Martínez et al., 2020; Hopson et al., 2014). Tal y como describen Ramón y Sánchez (2009), un ambiente familiar puede, en un extremo, fomentar la motivación, brindar apoyo, modelar relaciones saludables, y dar lugar al desarrollo de habilidades socio-cognitivas, mientras que en el otro extremo, el entorno familiar también puede desarrollar una adaptación incorrecta al sistema educativo, a través de prácticas controladoras, involucramiento ausente, conflictos y falta de apoyo emocional. Por otro lado, los programas de tutoría, así como las intervenciones de enriquecimiento académico, técnicas de estudio efectivas y programas de desarrollo de habilidades tienen efectos significativos en el rendimiento académico (Cavero, 2000; Rodríguez, 2020), así como también el apoyo de los compañeros de clase y el buen ambiente en el aula, favorecen el alza de las calificaciones y rendimiento general de los estudiantes (Morente et al., 2017).

Todos estos factores son relevantes en la explicación del rendimiento académico de los estudiantes. Sin embargo, la importancia relativa de los mismos, así como las posibles correlaciones que existen entre ellos es diferente y cambia a lo largo del ciclo escolar. Investigaciones recientes, como Vázquez et al. (2012), han puesto de manifiesto tales diferencias, evidenciando que el estudiante se pueda ver mayor o

menormente afectado por cada uno de estos factores, dependiendo del contexto preciso en el cual está desarrollando su actividad académica.

Estos resultados motivan el surgimiento de una nueva literatura en el ámbito de la educación que trata de predecir el rendimiento académico, explotando las posibles correlaciones que puedan existir entre los potenciales factores explicativos, y evaluar cómo la evolución de la importancia relativa de los mismos deriva de la propia dinámica del ciclo escolar. Estos estudios implican el tratamiento de un gran volumen de datos y, por tanto, su desarrollo ha ido paralelo al desarrollo de nuevas tecnologías que permiten dicho tratamiento. En concreto, las investigaciones más recientes recurren al uso de técnicas de *machine learning* (ver Castrillón et al., 2020), las cuales incluyen desde árboles de decisión, regresiones logísticas y lineales, modelos de optimización, hasta redes neuronales (Russo et al., 2016).

Algunos de los principales trabajos en esta nueva línea de métodos de *machine learning* aplicados al estudio del rendimiento escolar son los de Hoffait y Schyns (2017), y Waheed et al. (2020). Los primeros tienen como objetivo predecir las calificaciones de los estudiantes utilizando para ello el análisis del rendimiento académico en años anteriores. Este análisis les permite, además, medir el efecto inercia, el cual hace referencia a la tendencia que tienen las calificaciones obtenidas por el estudiante a permanecer constantes, o a cambiar de manera predecible a lo largo del tiempo, estando las puntuaciones futuras relacionadas con las pasadas, planteamiento que confirma el estudio en cuestión. Los segundos se centran en predecir el rendimiento de los estudiantes a través de variables demográficas y geográficas, encontrando que el ambiente (tanto del hogar, del lugar de residencia y del colegio) que rodea al estudiante, afecta significativamente a su rendimiento. Otro trabajo relevante es Xu et al. (2019). Estos autores tienen como objetivo evaluar el uso de internet por parte de los estudiantes como determinante de su rendimiento académico, y encuentran que se trata de un factor clave para predecir dicho rendimiento. Por su parte, Babić (2017) construye un modelo de redes neuronales capaces de estimar el rendimiento académico, utilizando como fuente de información predictiva toda aquella variable referida a la motivación del estudiante. Ella demuestra no sólo la efectividad de su modelo (con una precisión del 76.92%), sino también la importancia que tiene la información referida a valores tales como la autodeterminación y la resiliencia, entre otros, como variables predictoras. Adicionalmente, Costa-Mendes et al. (2020), se centran en evaluar las variables socio-demográficas que rodean al estudiante como factores claves de predicción del rendimiento académico de los mismos. Tales variables incluyen la edad, nivel cultural, lugar de residencia, empleo, ingresos, etc. Ellos encuentran la existencia de diferencias entre grupos de análisis; por ejemplo, el tamaño de la clase influye en materias de índole cualitativa, pero no en las de naturaleza cuantitativa. En esencia, se trata de un trabajo que transmite la importancia de conocer bien las variables objetivo, puesto que la gran interrelación que presentan las variables explicativas en materia de lo social puede generar que éstas alcancen impactos sustancialmente diferentes.

El trabajo más reciente es Iddrisu et al. (2023). Este artículo analiza la validez de los resultados de 84 estudios que han aplicado técnicas de *machine learning* para identificar los factores determinantes del rendimiento académico. En su análisis de tales estudios, los autores encuentran múltiples errores estadísticos en las muestras seleccionadas que impiden utilizar sus resultados para hacer inferencia. Insisten en la

importancia de emplear bases de datos adecuadas y perfectamente tratadas. Además, señalan las dificultades que presentan las regresiones logísticas, los árboles de decisión y los *random forests* cuando se consideran las variables por separado para predecir el rendimiento académico de los estudiantes. Tales problemas hacen que estos modelos sean muy poco adecuados para los estudios de ciencias sociales en general y se recomienda el uso de un tratamiento multidimensional de la información. Estos resultados han motivado que en el presente trabajo de investigación se empleen modelos derivados de las regresiones lineales (como primera aproximación), a efectos de no incurrir en las deficiencias estadísticas antes descritas.

3. DATOS

En este trabajo se utilizará una base de datos elaborada por la Agencia Canaria de Calidad Universitaria y Evaluación Educativa (ACCUEE). La ACCUEE es un organismo público fundado con el propósito de trabajar en la mejora continua de la enseñanza, tanto a nivel universitario como en el resto de etapas educativas. Adscrita a la consejería que gobierne en cada momento en materia de educación, centra su campo de actuación en el territorio canario.

Entre sus actividades, la ACCUEE se encarga de realizar evaluaciones de diagnóstico periódicas en educación primaria y educación secundaria obligatoria en Canarias. En este estudio utilizamos la información de los cursos académicos 2015/16 y 2018/19. Tanto en los cursos 2015/16 como en 2018/19, las EEDD se aplicaron a todo el alumnado de 6º curso de educación primaria, por lo que tuvieron carácter censal. Asimismo, en el curso 2015/2016, también se aplicaron, con carácter censal, las EEDD a 3º de educación primaria.

Existe armonía entre los intereses de la ACCUEE y los objetivos de este trabajo, ya que las evaluaciones de diagnóstico pretenden analizar no sólo el rendimiento académico de los estudiantes de manera aislada, sino también generar conocimiento acerca de los agentes que rodean a los propios resultados educativos, contextualizando así cada una de las variables estudiadas por los cuestionarios.

Asimismo, entre la información censal recogida por la ACCUEE en los cursos académicos mencionados, se dispone de datos longitudinales, es decir, información para un mismo estudiante en dos momentos del tiempo (específicamente, 3º y 6º de primaria). Este tipo de información, que es la utilizada en este trabajo, permite evaluar el progreso y desarrollo de los individuos a lo largo del tiempo y corregir ciertos problemas estadísticos, tales como los intrínsecos a las coyunturas temporales, al mismo tiempo que enriquece la calidad de los datos, pudiendo realizar estimaciones y comparaciones del individuo consigo mismo, siendo éste su mejor sustituto posible.

La base de datos proporcionada por la ACCUEE consta de toda la información censal para los cursos mencionados (un estudiante para cada fila), entre la que se encuentra la información longitudinal. Además, cada estudiante abarca un total de 561 columnas que representan las variables estudiadas en los cuestionarios de contexto. Como primer paso, por tanto, se llevó a cabo una limpieza o reorganización de la base de datos. Las variables (columnas) asociadas a cada estudiante se dividen en 7 bloques temáticos:

- Bloque 1: Variables ID. Bloque compuesto por variables que identifican al individuo a través de diferentes enfoques (organizacionales, centro educativo, curso académico, ID de la encuesta…).
- Bloque 2: Variables Informativas. Comprenden códigos que identifican si los individuos encuestados respondieron a los diferentes bloques de preguntas.
- Bloque 3: Calificaciones obtenidas. Compuesto por las calificaciones obtenidas por los estudiantes en las asignaturas de referencia (matemáticas, lengua e inglés), utilizando una clasificación numérica continua.
- Bloque 4: Cuestionario del alumno. Bloque de preguntas dirigidas a conocer el grado de acuerdo (categóricas), o situación (codificación/continuas) del alumno encuestado.

- Bloque 5: Cuestionario del director. Bloque de preguntas dirigidas a conocer el grado de acuerdo (categóricas), o situación (codificación/continuas) del director del centro del alumno encuestado. Se presupone coherencia para los mismos centros educativos.
- Bloque 6: Cuestionario de familias. Bloque de preguntas dirigidas a conocer el grado de acuerdo (categóricas), o situación (codificación/continuas) de la familia del alumno encuestado. Se presupone coherencia para las mismas unidades familiares.
- Bloque 7: Cuestionario de profesor: Bloque de preguntas dirigidas a conocer el grado de acuerdo (categóricas), o situación (codificación/continuas) del tutor del alumno encuestado. Se presupone coherencia para los estudiantes de un mismo tutor/a.

Entre los datos existe un alto porcentaje de valores faltantes o *missing values*. Nótese que esta información faltante puede generar sesgos cuando se utilizan los datos para realizar predicciones. Por ejemplo, podrían existir correlaciones entre no responder el cuestionario familiar o el cuestionario del estudiante, y la variable objetivo (rendimiento académico). No obstante, el tratamiento de estos *missing values* no es objeto de este trabajo, pero se presenta como una futura línea de investigación.

3.1. ELABORACIÓN DE LA MUESTRA

Uno de los puntos de mayor importancia en el análisis realizado por este trabajo es la construcción de la muestra de individuos con la que se trabajará, puesto que la información proporcionada por la ACCUEE no se presenta de manera consolidada. Como se mencionó en párrafos anteriores, la encuesta recopila información seccionada en varios bloques. Por lo tanto, para obtener la muestra completa de estudiantes con todas sus características es necesario extraer y combinar la información de los siete bloques. La matriz resultante constituirá la muestra definitiva de individuos, donde cada individuo estará representado por una fila y cada columna se corresponderá con una variable o atributo relevante.

El primer desafío se centra en la estandarización de los diferentes valores de identificación (IDs) existentes en la base de datos, puesto que la ACCUEE no ha sincronizado adecuadamente las distintas recopilaciones de datos, lo que ha dado lugar a la duplicación de valores de identificación para diferentes estudiantes, así como a la presencia de diversos formatos de numeración. Por consiguiente, para facilitar el indexado, es necesario armonizar la identificación de los individuos. Una vez finalizado el proceso de unificación de las variables IDs, la base de datos resultante consta de un panel de datos de 16.759 individuos de los que se cuenta con información tanto en 3º como en 6º curso de educación primaria.

A continuación, es preciso abordar el problema de los *missing values*. El objetivo es impedir que el porcentaje de *missing values* sea demasiado elevado. Con ello se pretende evitar, en la medida de los posible, que los resultados del estudio se vean afectados negativamente, cuestionando por tanto la validez estadística del mismo. De esta manera, estableciendo una relación entre el número de observaciones existentes y los posibles *missing values* de cada variable explicativa, se ha propuesto un límite máximo aceptable del 40% de *missing values* para cada variable explicativa analizada en el cuestionario, puesto que más allá de este porcentaje, los métodos de imputación de valores o corrección de datos faltantes comienzan a ser imprecisos. (Rubin y

Schenker, 1986). Esto implica que cada variable ha de tener al menos 10.000 valores; de lo contrario, la variable es eliminada del estudio.

Como resultado, se obtiene un panel de datos compuesto por 16.759 individuos, y 213 variables explicativas (ya clasificadas en continuas y categóricas).

3.2. VARIABLES UTILIZADAS

Con el objetivo de facilitar los estudios y análisis derivados de la utilización de la base de datos elaborada por la ACCUEE, Marrero et al. (2024) realizaron un diccionario de variables donde se encuentra disponible la descripción de cada una de ellas y las respuestas posibles.

No obstante, para este trabajo en particular, resulta necesario acudir a un glosario de variables que también precisa de cierta elaboración para ser práctico y operativo. Dada su extensión, y considerando la naturaleza de la metodología propuesta, la extensión completa del glosario se encontrará junto a la base de datos adjunta a este documento, al igual que las variables que se empleen a lo largo del trabajo.

Esta investigación, como se ha detallado, emplea un panel de datos. Esta característica implica que cada variable explicativa estudiada posee[1] tantos valores observados como momentos del tiempo son estudiados (tercero y sexto grado). Asimismo, por cuestiones metodológicas, resulta necesario separar, al menos, las variables que recogen las calificaciones de los alumnos en matemáticas y lengua (score_MAT y score_LEN, respectivamente), en un par de variables que sólo recojan las calificaciones obtenidas en tercer grado (score_MAT3 y score_LEN3) y que funcionarán como variables independientes, y en otro par que sólo recoja las calificaciones obtenidas en sexto grado (score_MAT6 y score_LEN6) que actuarán como variables dependientes a estimar.

El conjunto que ha superado la prueba crítica de *missing values*, y que por tanto será utilizado a lo largo de este trabajo se compone de 213 variables, cuyo detalle también podrá consultarse en la base de datos utilizada.

3.3. ANÁLISIS DESCRIPTIVO

A lo largo de esta sección se presenta un primer estudio exploratorio de los datos. Se realiza un análisis estadístico descriptivo de las variables dependientes a predecir, así como también de aquellas posibles variables explicativas que la literatura ha identificado como relevantes.

Dada la intención de este estudio de contribuir a la literatura en materia a través de nuevas hipótesis y planteamientos, el trabajo plantea analizar el rendimiento académico (y evolución) de los individuos, a través de:

- En primer lugar, un desagregado por asignaturas, teniendo como objetivo Matemáticas y Lengua (no incluyendo inglés por su aparición única en las observaciones de sexto).

[1] Siempre y cuando no se consideren los *missing values*.

- En segundo lugar, un agregado del rendimiento, compuesto por las medias aritméticas[2] de las calificaciones obtenidas en las asignaturas de matemáticas, lengua e inglés en sexto.

De esta manera, el trabajo posibilita el estudio de las diferencias entre variables explicativas según la asignatura objetivo, al mismo tiempo que ofrece una visión general y más próxima y cercana a la literatura al considerar el agregado del rendimiento académico. Finalmente, permite su validación a través del cruce de modelos, y comparación con el agregado.

En la literatura parece existir un consenso acerca de la fuerte inercia que experimentan los estudiantes en sus calificaciones a lo largo del ciclo escolar. Por ello, -y como se demostrará más adelante-, parece interesante analizar el comportamiento de las calificaciones que los alumnos obtienen en tercer grado, para realizar una comparativa posterior con sus puntajes en sexto grado y demostrar esta relación.

A través de la Figura 1 mostrada a continuación, los puntajes obtenidos en cada una de las asignaturas analizadas por este trabajo quedan representados a través de histogramas separados. Se trata de una indagación preliminar que muestra el comportamiento de las calificaciones en las que este trabajo pretende ahondar a través del desagregado de las variables explicativas que lo componen.

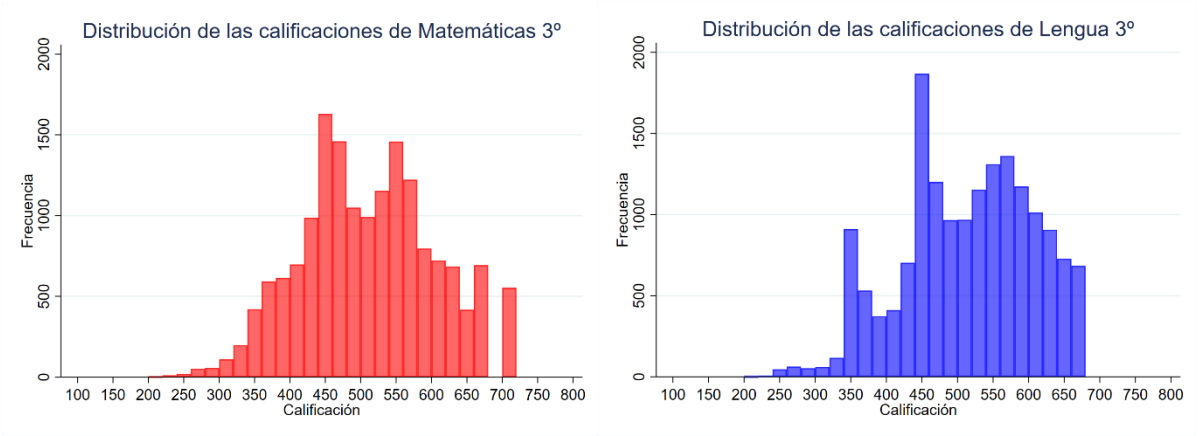

Figura 1. Distribución de las calificaciones de matemáticas y lengua en tercero. Elaboración propia (2024).

En lo que respecta a la distribución de cada uno de los histogramas, cabe mencionar el parecido que se observa en las calificaciones de matemáticas con respecto a la normalidad, que parece estar más difuso en las calificaciones correspondientes a lengua. Asimismo, en ambos casos, es notoria la ausencia de colas a la derecha de las distribuciones, pero sí a la izquierda, lo que demuestra que la escasez de valores anómalos altos (cercanos a 750 para el caso de matemáticas, y de 700 para el caso de lengua) no se deben a acotados estrictos en las escalas de medición, sino a la mayor presencia de individuos con calificaciones muy bajas, en comparación a la de individuos con calificaciones muy altas.

[2] Se emplean las medias aritméticas en lugar de las medias geométricas debido a la existencia de variables que pueden tomar el valor 0, en detrimento de la mayor robustez de las geométricas.

4. METODOLOGÍA LASSO

4.1. MODELOS LASSO

La metodología propuesta para predecir el rendimiento académico de los estudiantes de la muestra descrita parte de los modelos de regresión Lasso (*Least Absolute Shrinkage and Selection Operator).* Los modelos Lasso se insertan dentro de los métodos del *machine learning*, permitiendo que sea el propio procedimiento de estimación del modelo el que dictamine las variables explicativas estadísticamente significativas a la hora de predecir la variable objetivo (rendimiento académico) de una manera óptima, incorporando así aquellas variables que tienen un nivel alto de significatividad, y eliminando aquellas otras que puedan ser causantes de sobreestimación (*over-fitting*), o que solo estén incorporando ruido en el modelo y en las predicciones.

Para lograr esto, la regresión Lasso parte de principios semejantes a los de las regresiones lineales de Mínimos Cuadrados Ordinarios (MCO) (*Ordinary Least Squares*, OLS en inglés), donde se asocia un parámetro a cada variable explicativa, el cual, una vez estimado el modelo, expresa el efecto sobre la variable dependiente del incremento en una unidad[3] en dicha variable explicativa.

En concreto el modelo que se propone es:

$$\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 {x_2}^2 + \cdots + \beta_j x_j + \varepsilon + \lambda \sum |\beta_j|$$

En el que los primeros términos, -hasta incluir el término de error (ε)-, describen un modelo OLS estándar en el que se incorporan tantas variables explicativas (x) como variables observadas existan, y en el que habrá tantos factores de impactos relativos (β) como variables se incorporen. Además, incluye un parámetro β_0 que indica el punto de partida de la regresión (corte con el eje de ordenadas para el valor 0 en el eje de abscisas), y un parámetro ε que indica el error o variabilidad no explicada por el modelo. En su caso, cabría plantearse también recoger comportamientos no lineales del impacto de ciertas variables explicativas sobre la dependiente (γ), añadiendo términos cuadráticos de las independientes (x), que incorporan a su vez sus propios valores beta (β), tal y como se plantea en el supuesto de la variable independiente x_2.

Finalmente, en el caso de que existan variables categóricas (cuyos valores son mejor representados por cambios discretos que por cambios unitarios), es necesario transformarlas en variables *dummies*[4], en donde se asignan tantas *dummies* (con sus respectivos factores de impacto (β)) como categorías tenga la variable categórica.

Considerando lo descrito en los párrafos anteriores, se pone de manifiesto una de las principales dificultades estadísticas; la incorporación de un conjunto excesivo de variables, lo que a su vez, lleva a que el modelo incurra en el conocido *over-fitting*.

[3] Se emplea el cambio discreto para las variables categóricas.

[4] Las variables *dummies* son el resultado de desagregar una variable categórica en sus diferentes valores posibles. Tomará valor 1 cuando el valor de la categoría sea igual al de la *dummy*, y 0 en caso contrario.

Para corregir este inconveniente, se incorpora la segunda parte de la ecuación $\lambda \sum |\beta_j|$, denominada penalización L1[5] (valor absoluto) aplicada a los coeficientes de regresión. Este término de penalización puede ser ajustado por la variable Lambda (λ), y se encarga de excluir del modelo aquellas variables cuyo coeficiente posee una magnitud (en términos absolutos) menor que el término de penalización Lambda (λ). Esta penalización restringe la magnitud de los coeficientes y puede llevar algunos de ellos a cero, lo que, en la práctica, supone una multiplicación de las variables independientes no significativas por cero, dando lugar así a una simplificación del modelo.

Mencionar que un λ = 0 anularía la penalización propuesta y, por tanto, se estaría procediendo de la misma forma que a través de las regresiones OLS; mientras que en el sentido opuesto, un λ = ∞ incorporaría una penalización tan alta que ningún parámetro poseería la suficiente magnitud como para incorporarse a la regresión.

Es por ello que resulta necesario adoptar mecanismos de estimación del λ óptimo para el modelo propuesto de cara a equilibrar la balanza entre el *over-fitting,* y la estimación de un modelo incompleto, en términos de variables explicativas.

Adicionalmente, el Lambda (λ) que se estime ha de ser capaz también de minimizar los errores al cuadrado medios de la predicción, de manera que se constituya no sólo como una herramienta para seleccionar el número óptimo de variables significativas, sino para también minimizar los errores de predicción.

En el caso particular que atañe a este trabajo, se utilizará el modelo Lasso para predecir las calificaciones de los estudiantes en sexto curso de educación primaria, utilizando para ello la información explicativa recogida por los diferentes valores de las variables independientes de ellos mismos, en el tercer curso de educación primaria. Esto permitiría, como se explicó anteriormente, anticipar el futuro rendimiento académico de los estudiantes y, más concretamente, anticipar posibles fracasos, lo que, en suma, conduce a una información valiosa para la toma de decisiones en materia educativa a fin de poder elaborar políticas públicas encaminadas a actuar y corregir aquellos factores que influyen sobre el rendimiento académico de los estudiantes, prestando especial atención a aquellos que resulten más significativos.

El análisis propuesto pretende estimar el rendimiento académico en: el agregado (calculado como el promedio para cada estudiante de las asignaturas de matemáticas, lengua e inglés en sexto de primaria), en la asignatura de matemáticas de manera aislada y, por último, en la asignatura de lengua de manera aislada. Para ello se proponen tres modelos Lasso (uno para cada variable objetivo).

4.2. LASSO CROSS-VALIDATION

La optimización del modelo Lasso -y consigo, obtención del mínimo error cuadrático medio de predicción- depende intrínsecamente de la correcta selección del valor Lambda. Como se describió anteriormente, un Lambda pequeño genera un error de predicción semejante al de utilizar un Lambda mayor, por lo que no cabe lógica estadística en utilizar un Lambda que no sea el que minimice los errores al cuadrado.

[5] La penalización L1, también conocida como norma L1 o regularización L1 proviene del ámbito de la optimización y se utiliza para describir la técnica de regularización que implica la suma de los valores absolutos de los coeficientes en la función de pérdida del modelo establecido por el valor Lambda.

Por tanto, la verdadera cuestión del problema que plantea el modelaje Lasso es la de emplear una técnica adecuada de estimación del Lambda oportuno para cada caso.

Este trabajo ha decidido emplear la técnica de Cross-Validation como algoritmo para seleccionar el Lambda óptimo para cada una de las variables objetivo definidas (rendimiento académico agregado y en las asignaturas de matemáticas y lengua en sexto de primaria).

Se trata de una técnica ampliamente extendida y de creciente relevancia en el machine learning, principalmente por la eficiencia que muestra, siendo de aplicación sencilla al mismo tiempo que consigue una aproximación al mejor Lambda posible (Usai et al, 2009).

En general, el procedimiento es el siguiente:

1. En primer lugar, se toman n-1 secciones de la muestra (habitualmente la muestra se secciona en 5 o 10 partes, en función del número de observaciones disponibles) con las que realizar los análisis estadísticos pertinentes para dar lugar a la creación de un modelo predictivo.
2. Emplear el modelo predictivo creado en todas las secciones -exceptuando la que ha quedado fuera del análisis (n-1)-, teniendo además disponibilidad para observar la variable objetivo en cada sección, a fin de comprobar la precisión lograda.
3. Una vez se ha ajustado el modelo predictivo a través de su utilización y comprobación en las n-1 secciones (de manera que las predicciones resultantes sean precisas), se emplea el modelo para estimar la variable objetivo en la sección que se había dejado fuera del análisis durante la primera etapa; esta vez, sin conocer los resultados.
4. En caso de que el modelo no sea apropiado para predecir los valores de la sección enésima, es necesario reajustar. En caso de que sea apropiado (muestre un grado de error reducido), realizar la misma operativa, dejando fuera otra de las secciones de la base de datos.
5. Así, se repite el proceso hasta que todas las secciones de la base de datos hayan sido utilizadas (individualmente, como la sección enésima) para entrenar los modelos propuestos.
6. Entre todos los modelos se realiza una comparación general y ponderación de la aproximación de cada uno de ellos, de manera que entre todos resulte un único modelo que sea apropiado para realizar estimaciones para todas las secciones de la base de datos.

De esta manera, con una única base de datos, se puede predecir, entrenar y validar el modelo.

En particular, en los modelos de regresión Lasso, la validación cruzada se centra en el empleo de diferentes valores Lambda para la estimación de la variable objetivo en cada una de las secciones de la base de datos, obteniendo así tantos lambda óptimos de predicción como secciones se hayan propuesto. En concreto, los programas estadísticos actuales generan los modelos Lasso resultantes de emplear un valor lambda cada vez mayor (con saltos unitarios, decimales, centenales…), y predecir con cada uno de ellos, esperando reducir cada vez más el error cuadrático de la estimación, hasta encontrarse con el punto de inflexión en el que la falta de variables comienza a empeorar las predicciones, ajustando y precisando el Lambda alrededor

de esa franja. Así es como se alcanza un valor Lambda óptimo (que minimiza los errores al cuadrado medios) para cada una de las secciones. Este ejercicio, además, demuestra la capacidad de validación del propio modelo sin la necesidad de recurrir a fuentes de datos externas, puesto que se entrena con secciones de sí misma y cuenta con una fase final tras el entrenamiento, de validación con la sección que ha quedado fuera.

Finalmente, fruto del entrenamiento de múltiples modelos con una misma variable objetivo, los resultados -y consigo, el Lambda seleccionado- obtienen robustez por haber sido sometido a tantas pruebas de entrenamiento y validación como secciones se planteen sobre la muestra. Con el Lambda óptimo validado, resta incorporarlo en el modelo original y permitir que sea el propio factor de penalización el que seleccione las variables cuyo impacto es lo suficientemente significativo para mantenerse en el modelo como variables explicativas, y las que no.

Además, este primer análisis ya permite -de manera aproximada-, cuantificar las diferencias que existen entre los factores de impacto (β) de unas variables y otras, lo que, a su vez, facilita que el tomador de decisiones identifique si las variables están teniendo un impacto positivo o negativo en el rendimiento educativo, así como también comprobar cuáles tienen un impacto mayor.

5. RESULTADOS

A lo largo de este capítulo se presentarán los resultados obtenidos por la aplicación de la metodología antes descrita. En primer lugar, se presentan los resultados de la estimación de tres modelos de predicción diferentes: uno para estimar la calificación agregada del estudiante, otro para estimar la calificación del estudiante en matemáticas y otro para estimar la calificación del estudiante en lengua. Estos modelos son los Modelos 1, 2 y 4 en la Tabla 1, cuyas variables dependientes son el promedio de las puntuaciones en matemáticas, lengua e inglés para el Modelo 1 y las puntuaciones en matemáticas y lengua para el Modelo 2 y 4, respectivamente.

Este primer ejercicio permite identificar qué factores son relevantes para predecir la puntuación final de las asignaturas y, por tanto, puedan ser tenidos en cuenta para el posible diseño de políticas de refuerzo o excelencia educativa.

De la misma forma, a pesar de que la magnitud de los parámetros no es estrictamente comparable entre los diferentes modelos predictivos ya que la especificación de los mismos es diferente; este primer análisis sí permite comparar la significatividad y signo de las variables en los diferentes modelos predictivos.

En segundo lugar, se propone un ejercicio de validación que consiste en intercambiar las variables objetivo de los Modelos 2 y 4. De esta forma, se estiman dos nuevos modelos: el Modelo 3 para predecir las calificaciones de lengua utilizando para ello la especificación del modelo de matemáticas (Modelo 2), y, el Modelo 5, para predecir las calificaciones de matemáticas, utilizando para ello la especificación del modelo de lengua (Modelo 4).

En la literatura, esta práctica recibe el nombre de validación cruzada, y se utiliza generalmente para comparar la capacidad predictiva y generalización de los modelos en cuestión. Por tanto, si un modelo diseñado para un área específica (p. ej., matemáticas) es capaz de predecir razonablemente bien en otra área (p. ej., lengua), esto sugeriría que el modelo tiene un buen nivel de generalización, y que las variables explicativas tienen un nivel considerable de relevancia en ambos dominios. De la misma forma, unos resultados de predicción pobres o con márgenes de error muy grandes podrían estar indicando la existencia de variables contextuales que afectan de manera diferente a cada materia, o que cobra importancia para unas, pero no para otras, lo que abriría nuevas vías de investigación sobre la existencia de factores con incidencia distinta.

Además, este ejercicio permite ahondar aún más en la comparación de variables relevantes comunes. Al contrario que en los Modelos 1, 2 y 4 (utilizados para predecir las puntuaciones agregadas, las de matemáticas, y las de lengua), los Modelos 2 y 3, y los Modelos 4 y 5 están especificados de la misma forma, permitiendo la comparabilidad directa de los parámetros. Es decir, los resultados del Modelo 2 para matemáticas, y sus resultados cuando se utilizan para predecir lengua (y los resultados del Modelo 4 de Lengua para predecir matemáticas), son directamente comparables, tanto en nivel de significación como en magnitud de los coeficientes. Esta propiedad ayuda a entender la interrelación existente entre asignaturas, y permite

detectar aquellas variables que son significativas en ambos modelos, siendo así un factor clave en el diseño de políticas educativas que tengan como objetivo una mejora del rendimiento en varias asignaturas al mismo tiempo.

Tabla 1. Resultados de la estimación de los modelos Lasso

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Variable	**GEN**	**MAT**	**MAT-LEN**	**LEN**	**LEN-MAT**
Puntuaciones en 3º de primaria					
Puntuaciones en Matemáticas	0.271***	0.365***	0.176***	0.178***	0.396***
	(21.833)	(24.078)	(11.081)	(12.661)	(29.5)
Puntuaciones en Lengua	0.22***	0.204***	0.272***	0.268***	0.19***
	(16.199)	(12.224)	(15.536)	(17.748)	(13.206)
Sexo					
Mujer (Ref.: Hombre)	5.659***	-10.358***	24.105***	22.712***	-6.769***
	(2.928)	(-4.408)	(9.778)	(10.45)	(-3.259)
Repetidor					
Repetidor una o más veces (Ref.: No repetidor)	-24.049***				
	(-2.987)				
Frecuencia de faltas sin excusa a clase					
Nunca (Ref.: Una vez a la semana, una vez cada dos semanas, una vez al mes)	6.672**	8.452**	5.257		
	(2.13)	(2.263)	(1.342)		
Frecuencia de uso de ordenadores en casa.					
Frecuentemente (Ref.: Nunca, a veces)		-7.089**	-1.285		
		(-2.171)	(-0.375)		
Siempre (Ref.: Nunca, a veces)		-4.463	-0.613		
		(-1.089)	(-0.142)		
Siempre (Ref.: Nunca, a veces, frecuentemente)	0.411				
	(0.119)				
Frecuencia de uso de ordenadores en otros lugares.					
A veces (Ref.: Nunca)	-4.251*				
	(-1.867)				
Frecuentemente (Ref.: Nunca)	-8.301**				
	(-2.253)				
Siempre (Ref.: Nunca)	-7.899				
	(-1.337)				
A veces (Ref.: Nunca, frecuentemente)		-2.559	-2.377		
		(-0.925)	(-0.819)		
Siempre (Ref.: Nunca, frecuentemente)		-10.992	1.794		
		(-1.565)	(0.244)		
Frecuencia de uso de internet: buscar información sobre deportes					
Cada día (Ref.: Nunca, una o dos veces al mes, una o dos veces a la semana)	-4.454				
	(-1.327)				
Frecuencia de uso de internet: buscar información de otros temas					
Una o dos veces a la semana (Ref.: Nunca, una o dos veces al mes, cada día)	5.063**				
	(2.288)				
Trabajo con los profesores: ellos explican durante la mayor parte de la clase					
A veces (Ref.: Nunca, frecuentemente, siempre)	4.315**				
	(2.007)				
Siempre (Ref.: Nunca, a veces, frecuentemente)		-6.549***	-1.535		
		(-2.649)	(-0.591)		
Trabajo con los profesores: mientras explican, los estudiantes hacen preguntas.					
A veces		-5.054	1.744		
		(-1.582)	(0.52)		
Siempre		3.107	-2.749		
		(1.168)	(-0.985)		
Siempre (Ref.: Nunca, a veces, la mayoría de las veces)	1.308				
	(0.659)				
Trabajo con los profesores: se hacen debates en clase					
Siempre (Ref.: Nunca, a veces, frecuentemente)	2.664				
	(0.958)				
Material de clase: libros de texto					
En la mayoría de las clases (Ref.: Ninguno, en algunas clases, en todas las clases)	4.253**	1	4.019		
	(1.983)	(0.38)	(1.455)		
Material en clase: material preparado por el profesor					
En todas las clases (Ref.: Ninguno, en algunas clases, en la mayoría de las clases)	-7.273***	-5.875**	-5.909*	-6.815***	-6.333**
	(-3.046)	(-2.026)	(-1.941)	(-2.625)	(-2.554)
Material en clase: pizarra digital					
En todas las clases (Ref.: Ninguno, en algunas clases, en la mayoría de las clases)	-3.762*	-9.093***	1.478		
	(-1.846)	(-3.614)	(0.56)		

Evaluación: los profesores hacen preguntas antes de comenzar un tema					
Frecuentemente (Ref.: Nunca, a veces, siempre)		5.882**	-3.416		
		(2.201)	(-1.219)		
Siempre (Ref.: Nunca, a veces, frecuentemente)	-4.351**				
	(-1.991)				
Evaluación: los profesores evalúan las actividades que realizan los estudiantes en clase					
A veces (Ref.: Nunca, frecuentemente, siempre)	-6.127*				
	(-1.653)				
Evaluación: los profesores valoran el interés y la participación en clase					
A veces (Ref.: Nunca, frecuentemente)	-6.477*	-5.17	-8.534*		
	(-1.724)	(-1.115)	(-1.749)		
Siempre (Ref.: Nunca, frecuentemente)	-0.609	1.52	0.748		
	(-0.254)	(0.526)	(0.247)		
Grado de acuerdo: los estudiantes se quedan a solas con los profesores					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	-0.943				
	(-0.164)				
Grado de acuerdo: los profesores escuchan a los estudiantes					
De acuerdo (Ref.: Muy en desacuerdo, desacuerdo, muy de acuerdo)	7.588***				
	(3.359)				
Grado de acuerdo: el estudiante se siente solo					
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)	-8.932*			-5.681	-3.684
	(-1.754)			(-1.038)	(-0.704)
Grado de acuerdo: los compañeros del estudiante lo dejan fuera de los juegos					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	4.035*				
	(1.706)				
Grado de acuerdo: el estudiante está asustado de sus compañeros					
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)	-5.39	0.095	-18.618***	-13.417***	-0.743
	(-1.157)	(0.018)	(-3.379)	(-2.704)	(-0.157)
Grado de acuerdo: los compañeros del estudiante lo ayudan en clase					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	8.923***				
	(2.796)				
Grado de acuerdo: el estudiante está interesado en lo que el profesor dice					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)		-9.947	1.093		
		(-1.131)	(0.117)		
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)	4.98*				
	(1.864)				
Grado de acuerdo: los profesores proponen cosas interesantes a los estudiantes					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	-8.181*				
	(-1.81)				
Grado de acuerdo: los profesores responden claramente las preguntas de los estudiantes					
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)	3.839				
	(1.627)				
Grado de acuerdo: los profesores hacen cosas diferentes para ayudar a los estudiantes					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	-4.283	-11.541	6.552		
	(-0.715)	(-1.607)	(0.864)		
Grado de acuerdo: al estudiante realmente le gusta trabajar en clase					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	15.475***				
	(2.832)				
Grado de acuerdo: el estudiante querría cambiar de colegio					
Desacuerdo (Ref.: Muy en desacuerdo, de acuerdo, muy de acuerdo)	8.961*				
	(1.918)				
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)				-11.745**	-1.031
				(-2.336)	(-0.214)
Grado de acuerdo: el estudiante aprende mucho en el colegio					
Muy de acuerdo (Ref.: Muy en desacuerdo, desacuerdo, de acuerdo)	4.681	2.72	6.712*		
	(1.578)	(0.821)	(1.928)		
Edad del padre					
Edad del padre (variable continua)	0.191			0.043	0.157

	(1.216)			(0.248)	(0.944)
Educación de la madre					
Educación terciaria (Ref.: sin educación / educación primaria, educación secundaria baja, educación secundaria alta)		2.214	5.999**		
		(0.843)	(2.172)		
Educación del padre					
Educación terciaria (Ref.: Sin educación / educación primaria, educación secundaria baja, educación secundaria alta)	7.596***	9.559***	5.613*	9.784***	14.886***
	(3.542)	(3.387)	(1.893)	(4.12)	(6.557)
Edad de escolarización					
Entre tres y cuatro años (Ref.: Menos de tres años, cuatro o más años)	-1.177	-5.802**	2.374		
	(-0.569)	(-2.278)	(0.888)		
Con qué frecuencia el estudiante usa en casa: libros					
Una o dos veces al mes (Ref.: Nunca, una o dos veces a la semana, siempre)	-6.548*	-2.25	-6.645	-12.294***	-2.228
	(-1.954)	(-0.548)	(-1.541)	(-3.269)	(-0.62)
Con qué frecuencia el estudiante usa en casa: enciclopedias					
Siempre (Ref.: Nunca, una o dos veces al mes, una o dos veces a la semana)	4.725**	5.519*	2.195		
	(2.017)	(1.915)	(0.725)		
Número de libros en el hogar (Categorías: Menos de 11 libros, de 11 a 50 libros, de 51 a 100 libros, más de 100 libros).					
De 11 a 50 libros	-0.573			-3.306	-0.744
	(-0.265)			(-1.484)	(-0.35)
Más de 100 libros	5.912**	5.741**	5.157*		
	(2.343)	(1.993)	(1.705)		
Con qué frecuencia: los padres leen historias al estudiante					
Una vez al mes (Ref.: Nunca, a veces, una vez a la semana, diariamente)		6.735	6.583		
		(1.223)	(1.145)		
Diariamente (Ref.: Nunca, a veces, una vez al mes, una vez a la semana)	-3.477			-6.522**	-0.659
	(-1.526)			(-2.512)	(-0.266)
Con qué frecuencia: la madre asiste al colegio para comprobar el rendimiento que está obteniendo el estudiante					
A veces (Ref.: Nunca, una vez al mes, una vez a la semana)	3.198				
	(0.619)				
En una entrevista con el profesor, los padres se interesan en: dificultades en el proceso de aprendizaje					
Siempre (Ref.: Nunca, a veces, frecuentemente)	-7.921***	-10.53**	-6.151		
	(-2.797)	(-2.483)	(-1.382)		
En una entrevista con el profesor, los padres se interesan en: rendimiento del grupo					
A veces (Ref.: Nunca, frecuentemente, siempre)	0.435				
	(0.147)				
Con qué frecuencia: los padres ayudan al estudiante a hacer su tarea					
Siempre (Ref.: Nunca, a veces, frecuentemente)	-7.758***				
	(-3.096)				
Con qué frecuencia: los padres participan en la organización de actividades extraescolares					
Frecuentemente (Ref.: Nunca, a veces, siempre)	-5.202*				
	(-1.782)				
Siempre (Ref.: Nunca, a veces, frecuentemente)		-0.291	1.695		
		(-0.08)	(0.443)		
Grado de satisfacción de los padres con los profesores					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)	-2.635	0.288	-2.731		
	(-0.753)	(0.065)	(-0.586)		
Grado de satisfacción de los padres con el trabajo en clase					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)	-3.473				
	(-1.039)				
Grado de satisfacción de los padres con los recursos del colegio					
Muy satisfecho (Ref.: Nada satisfecho, poco satisfecho, satisfecho)	6.429***			2.939	6.658***
	(3.285)			(1.222)	(2.898)
Grado de satisfacción de los padres con el nivel de aprendizaje					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)		-8.186**	-3.398		
		(-2.134)	(-0.845)		
Grado de satisfacción de los padres con la flexibilidad y paciencia del profesor					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)	-2.726				
	(-0.613)				

Expectativas de los padres acerca del nivel de estudios que alcanzará el estudiante					
Finalizar FP		-37.459**	13.141		
		(-2.43)	(0.813)		
Finalizar FP avanzada		-13.346***	-2.282		
		(-2.715)	(-0.44)		
Finalizar la universidad	5.691*			8.066**	11.724***
	(1.751)			(2.249)	(3.431)
Ocupación del padre					
Ocupación baja	-5.37*	-4.464	-7.403*		
	(-1.854)	(-1.239)	(-1.957)		
Ocupación alta		-0.955	0.596		
		(-0.355)	(0.211)		
Frecuencia de uso: ordenador en clase					
A veces (Ref.: Nunca, frecuentemente, siempre)		-2.911	1.282		
		(-1.22)	(0.512)		
Trabajo con los profesores: los profesores hacen sentir bien a los estudiantes					
Frecuentemente (Ref.: Nunca, a veces, siempre)		10.802***	5.002		
		(3.616)	(1.594)		
Material en clase: libros de biblioteca					
En todas las clases (Ref.: Ninguno, en algunas clases, en la mayoría de las clases)		-5.954*	3.248		
		(-1.77)	(0.92)		
Material en clase: revistas y periódicos					
En la mayoría de las clases (Ninguno, en algunas clases, en todas las clases)		-10.119*	-6.411		
		(-1.844)	(-1.117)		
Material en clase: ordenadores y tablets					
En la mayoría de las clases (Ninguno, en algunas clases, en todas las clases)		-2.67	-0.882		
		(-0.743)	(-0.233)		
Con qué frecuencia el estudiante dispone en casa de: un lugar tranquilo para estudiar					
Siempre (Ref.: Nunca, una o dos veces al mes, una o dos veces a la semana)		-5.652	2.453		
		(-1.441)	(0.596)		
Con qué frecuencia el estudiante usa en casa: periódicos					
Una o dos veces a la semana (Ref.: Nunca, una o dos veces al mes, siempre)		-9.605**	0.472		
		(-2.44)	(0.114)		
Número de dispositivos digitales del hogar					
Número de dispositivos (variable continua)		0.821	-0.968*		
		(1.643)	(-1.846)		
En una entrevista con el profesor, los padres se interesan en: puntuaciones del estudiante					
Una vez al mes (Ref.: Nunca, a veces, una vez a la semana)		-8.014**	-0.307		
		(-2.192)	(-0.08)		
En una entrevista con el profesor, los padres se interesan en: desempeño del estudiante					
Una vez a la semana (Ref.: Nunca, a veces, una vez al mes)		-3.097	3.654		
		(-0.747)	(0.842)		
Con qué frecuencia el estudiante dispone en casa de: el material necesario para estudiar					
Una o dos veces a la semana (Ref.: Nunca, una o dos veces al mes, siempre)		9.031	-6.883		
		(1.629)	(-1.182)		
Con qué frecuencia: los padres comprueban que el estudiante haya hecho la tarea					
Siempre (Ref.: Nunca, una o dos veces al mes, una o dos veces a la semana)		-15.481**	-6.066		
		(-2.444)	(-0.909)		
Grado de satisfacción: relación de los padres con los profesores					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)		-0.849	-5.418		
		(-0.182)	(-1.11)		
Grado de satisfacción: relación de los padres con los compañeros del estudiante					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, muy satisfecho)		3.832	-0.113		
		(1.562)	(-0.044)		
Grado de satisfacción: con el colegio en general					
Muy satisfecho (Ref.: Nada satisfecho, poco satisfecho, neutral, satisfecho)		4.861*	-3.464		
		(1.901)	(-1.291)		
Días de estudio					
En una escala del 0-5: 2				-4.169	0.534
				(-1.533)	(0.206)
Grado de acuerdo: los estudiantes participan en la toma de decisiones					

Desacuerdo (Ref.: muy en desacuerdo, de acuerdo, muy de acuerdo)				6.81**	3.331
				(2.024)	(1.039)
Grado de satisfacción: de los padres con el director					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, neutral, muy satisfecho)				4.1	1.794
				(1.634)	(0.748)
Grado de satisfacción: de los padres con el ambiente general del colegio					
Satisfecho (Ref.: Nada satisfecho, poco satisfecho, neutral, muy satisfecho)				1.084	-3.277
				(0.448)	(-1.417)
_cons					
	237.19***	244.01***	265.47***	262.89***	187.86***
	(21.075)	(19.615)	(20.305)	(24.822)	(18.555)
Observaciones	4910	5355	5344	6908	6916
R^2	0.335	0.296	0.192	0.183	0.273
Error MSE	65.319	83.99	87.17	88.06	84.30

t-values en paréntesis
**** p<.01, ** p<.05, * p<.1*

En líneas generales, de la Tabla 1 se puede concluir: primero, los signos de las variables, así como sus magnitudes, son robustas independientemente del modelo que se utilice. Esto indica que la explicación que aporta cada variable, es semejante independientemente de si se está estudiando la calificación de matemáticas, de lengua, o del agregado (no hay inconsistencias significativas entre asignaturas). Segundo, la relevancia que tiene por separada cada una de las variables explicativas, es menor que cuando se estudian en su conjunto. De ahí que los modelos para matemáticas y para lengua no pierdan sus capacidades predictivas, aun alterando las variables dependientes para los que fueron diseñados. Tercero, el alto índice de coincidencia que muestran los signos de las variables independientes cuando los modelos son sometidos a la etapa de predicciones cruzadas, indica la fuerte interseccionalidad que existe entre las diferentes variables recogidas. Cuarto, las principales diferencias que muestran los modelos no se encuentran en los signos, sino en las magnitudes. Esto pone de manifiesto que, en términos generales, las diferentes variables tienen un impacto en la misma dirección, sin importar cuál sea la asignatura objeto de estudio.

De forma más detallada, los resultados más relevantes que desprenden de las estimaciones recogidas en la Tabla 1, son los siguientes:

Puntuaciones de tercero de primaria: esta variable presenta el nivel de significación más alto para todos los modelos. Además, se trata de una variable continua cuyos valores oscilan entre 300 y 700, lo que genera un producto (al multiplicarse por el valor de sus parámetros beta) mayor que el de cualquier otra variable de la tabla. Esto demuestra el fuerte impacto que tiene el efecto inercia en la educación.

Sexo: los resultados parecen mostrar que los hombres obtienen mejores calificaciones en matemáticas y las mujeres en lengua e inglés. Esto explica por qué en el agregado, que recoge las tres asignaturas, ser mujer está asociado positivamente con mayores calificaciones.

Repetidor: pertenecer al grupo de repetidores muestra un efecto negativo sobre las calificaciones en general (así se hace ver en el modelo para el agregado), aunque pierde su efecto cuando se desagrega por asignaturas.

Ausencia a clase: no faltar a las clases sin excusa tiene una repercusión positiva sobre todas las asignaturas. No obstante, esta incidencia es mayor (y más significativa) para matemáticas que para lengua, posiblemente porque las explicaciones en matemáticas son críticas para el correcto seguimiento de la asignatura, mientras que la lengua se presta más al trabajo autónomo.

Uso del internet: un uso intensivo del internet para buscar exclusivamente información relacionada con los deportes tiene un efecto negativo en las calificaciones, mientras que un uso para temas generales tiene un efecto positivo. Esta característica es relevante cuando se analiza el agregado, pero no cuando se estudian las asignaturas de manera individual.

Los profesores explican durante la mayor parte de la clase: recibir siempre clases magistrales tiene un efecto negativo en las calificaciones, mientras que recibirlas de vez en cuando, tiene un efecto positivo. Este fenómeno es sobre todo significativo para matemáticas (que habitualmente precisa de clases más interactivas), y menos en lengua, donde la metodología Lasso ni si ha considerado esta variable como relevante.

Material preparado por el profesor: esta variable tiene un efecto negativo en todos los modelos (específicamente en la categoría “en todas las clases” con respecto a las demás). Este signo, aparentemente contraintuitivo, podría estar explicado por el hecho de que los estudiantes asocian esta pregunta al volumen de tareas que deben realizar en clase.

Pizarra digital: utilizar la pizarra digital tiene un efecto negativo. Esto puede deberse a la brecha digital tanto de estudiantes como de profesores, así como al hecho de que los profesores permiten una menor interacción de los alumnos con ellas. Nuevamente, este efecto es característico de la asignatura de matemáticas, donde se prima la interacción, mientras que en el modelo para lengua, esta variable no resulta significativa.

Los profesores hacen preguntas antes de comenzar un tema: esta variable tiene un efecto positivo si tales preguntas se hacen frecuentemente, pero negativo si se hacen siempre. Esto demuestra el efecto estrés que pueden causar los profesores sobre el alumnado, pero también recalca la importancia de hacer un seguimiento sobre asignaturas con una línea de aprendizaje más continua (como matemáticas), así como de poner a prueba su iniciativa con los conocimientos ya adquiridos. En asignaturas donde el haber comprendido los temas anteriores no es tan relevante para los temas futuros, como sucede en lengua (donde existe linealidad entre temas), este mecanismo desaparece.

Los profesores evalúan la realización de tareas, interés, y escuchan a los estudiantes: la no evaluación de los trabajos realizados en clase, así como del interés y de la motivación de los estudiantes, tiene un efecto negativo en sus calificaciones. Este elemento pone de manifiesto la importancia de valorar la actitud del estudiante, y no sólo su rendimiento.

Soledad del estudiante: si el estudiante se siente solo, o tiene miedo de sus compañeros, tiene un efecto negativo sobre su rendimiento. La incidencia del miedo a los compañeros es particularmente significativa para lengua, posiblemente debido a que las dificultades en conocimientos acerca de matemáticas están más normalizadas y/o generalizadas, que las dificultades en conocimientos sobre lengua, como lo son la lectura. Por otro lado, en el agregado, si el estudiante juega con los compañeros, esto tendrá un efecto positivo.

Los compañeros ayudan al estudiante en clase: los estudiantes que reciben ayuda de sus compañeros en clase tienen peores calificaciones. Una posible razón de este resultado es que, precisamente, son los estudiantes con peor rendimiento los que reciben más ayuda por parte de sus compañeros.

El estudiante querría cambiar de colegio: si el estudiante se encuentra incómodo en su actual colegio, y quiere cambiar, esto tiene un efecto negativo sobre sus calificaciones. Como en el caso del miedo a los compañeros, este efecto es significativo para lengua, pero no para matemáticas.

Educación de los padres: tener educación terciaria (nivel universitario), tanto del padre como de la madre, tiene un efecto positivo. No obstante, la educación de la madre sólo parece tener relevancia en lengua, mientras que la del padre tiene relevancia en todas las asignaturas.

Edad de escolarización: comenzar la escuela tardíamente tiene un impacto negativo sobre las calificaciones, sobre todo en matemáticas. Esta circunstancia podría deberse a la evolución o exigencias de los procesos cognitivos que tiene cada asignatura.

Libros y enciclopedias: un uso escaso de los libros tiene impacto negativo (mucho más acusado para lengua que para matemáticas), mientras que hacer frecuentemente uso de las enciclopedias tiene impacto positivo. Tener mayor número de libros en el hogar, favorece las calificaciones en todas las asignaturas.

Preocupación de los padres: las variables que miden esta preocupación (ayudas de los padres al realizar tareas, visitas al tutor, interés por dificultades en el proceso de aprendizaje, rendimiento del grupo en general) muestran valores beta de signo negativo. Una posible razón de este impacto negativo, similar al resultado sobre los estudiantes que reciben ayuda de sus compañeros, es que los estudiantes con peores calificaciones son los que necesitan mayor número de refuerzos y sobre los que los padres muestran una mayor preocupación.

Expectativas de los padres acerca del nivel de estudios de sus hijos: mayores expectativas de alcanzar titulaciones más altas presenta efectos positivos sobre todas las asignaturas.

Ocupación de los padres: la ocupación de la madre no es significativa para ningún modelo, mientras que la ocupación del padre tiene un efecto significativo negativo cuando es baja.

Material en el aula, libros, revistas y ordenadores: un volumen excesivo en el aula de los materiales descritos tiene un efecto negativo. Una posible razón es que éstos puedan causar distracción o no se utilicen de la manera correcta.

Número de dispositivos del hogar: se trata de una variable continua que tiene un efecto positivo para matemáticas, pero negativo para lengua. Esta circunstancia podría ser debida a la exposición del traductor y autocorrectores, que favorecen las lecturas rápidas y superficiales sobre el análisis crítico de los textos, y la reducción de la escritura a mano, que ayuda a la memoria y la mejora de la caligrafía.

Cabe destacar que el error cuadrático medio del modelo para matemáticas cuando se utiliza para estimar matemáticas se sitúa en los 83,99 puntos, mientras que cuando se utiliza para estimar lengua se sitúa en los 87,17 puntos. Por otro lado, el error cuadrático medio del modelo para lengua cuando se utiliza para estimar lengua se sitúa en los 88,06 puntos, mientras que cuando se utiliza para estimar matemáticas se sitúa en los 84,30 puntos.

Aparentemente, el modelo de matemáticas estaría prediciendo mejor el rendimiento en lengua que el propio modelo para lengua. Sin embargo, es necesario tener en cuenta las diferencias entre modelos, puesto que mientras el modelo de lengua incluye 18 variables explicativas, el de matemáticas tiene 37, lo que podría estar causando over-fitting. Así mismo, la distribución de las calificaciones en matemáticas se comporta de manera más semejante a la distribución normal que la de las de lengua, siendo más consistente con los supuestos de estimación (véase la Figura 1). Finalmente, también es necesario considerar la muestra sobre la que se aplica el modelo, que en matemáticas es de 5.344 y en lengua de 6.908, haciendo que la estimación en este último caso sea más complicada.

Estas diferencias, que impiden la comparación directa de los errores cuadráticos medios, aplica de igual forma para todos los modelos descritos, por lo que no es posible determinar únicamente a través de este indicador cuál es el mejor de los modelos. No obstante, sin perjuicio de lo anterior, en general, los resultados indican que el modelo agregado es el que tiene una mayor eficacia predictiva. Esto se debe en parte por aplicarse a una menor muestra de entrenamiento y predicción, pero también a la propia definición de la variable rendimiento medio (la media de tres rendimientos) ya que, al agregar, se suavizan y minimizan los valores anómalos que se observan en las asignaturas por separado, permitiendo obtener una distribución del rendimiento a predecir que se acerca más a la normal.

6. CONCLUSIONES

El rendimiento académico de los estudiantes y las variables que lo determinan es uno de los principales temas de investigación, no sólo en materia escolar, sino también en el ámbito económico, social, cultural y político, por las importantes implicaciones que tienen los resultados académicos en el desarrollo futuro de los individuos.

En este trabajo se han elaborado tres modelos para predecir el rendimiento escolar de los estudiantes de sexto curso de educación primaria: uno para matemáticas, otro para lengua, y otro para el agregado académico (calculado como la media de calificaciones obtenidas en matemáticas, lengua e inglés). Por tanto, se trata de un trabajo pionero que explora las asignaturas de manera individualizada, en lugar de sólo considerar el agregado, lo que en última instancia permite encontrar las diferencias explicativas que se hallan en cada materia.

Estos modelos se han estimado utilizando un panel de datos obtenido a partir de las evaluaciones de diagnóstico llevadas a cabo por la Agencia Canaria de Calidad Universitaria y Evaluación Educativa (ACCUEE). La metodología que se ha empleado es la *Lasso Cross Validation*, que permite realizar, sobre grandes conjuntos de datos, una selección de las variables más relevantes y con mayor potencial explicativo y predictivo.

Aparte de la estimación de la especificación particular de cada modelo, estas técnicas de validación cruzada se han utilizado en un segundo análisis, en la que los modelos para matemáticas y para lengua han intercambiado sus variables objetivo. La finalidad es realizar un ejercicio de validación que permita garantizar que los resultados de cada modelo son sólidos, y que éstos no pierden su capacidad predictiva cuando se utilizan para predecir sobre otras asignaturas.

En lo que respecta a poder predictivo que se ha obtenido mediante la utilización de mecanismos de *machine learning*, específicamente la regresión Lasso y su sometimiento a técnicas de validación cruzada, se ha logrado obtener un error medio de predicción del 7,21% para el mejor de los modelos y un 10,36% para el peor de ellos. Estos resultados ponen de manifiesto el considerable potencial que tiene esta metodología para la elaboración de mecanismos de anticipación que sean útiles, eficaces y fiables. La precisión alcanzada por los modelos analizados demuestra la enorme utilidad que tienen estas técnicas para mejorar la toma de decisiones y optimizar los resultados de políticas de ayuda educativa.

Además, otra característica que refuerza la fiabilidad de los resultados obtenidos, es la amplia dimensión del análisis exploratorio realizado. El análisis de predicción parte una base de datos inicial compuesta por 541 variables explicativas entre las que se incluyen bloques de preguntas para el estudiante, familias, profesores y del propio centro, de las cuales, el propio procedimiento estadístico selecciona 69 en el conjunto de los cinco modelos resultantes. Esta gran dimensión garantiza que el análisis ha realizado una aproximación holística, sin centrarse en ningún aspecto característico.

En línea con la evidencia empírica existente, los resultados muestran la fuerte incidencia que tienen las calificaciones de tercero sobre las calificaciones de sexto, lo

que ilustra la magnitud del efecto inercia en el ciclo escolar. Además, otras variables que obtienen el signo esperado según la literatura son: la evaluación por parte del profesor del interés y motivación del estudiante; el sentimiento de soledad o aislamiento por parte del resto de compañeros; la educación, expectativas y ocupación de los padres; el entorno en el hogar y de estudio (como número de libros y espacio tranquilo); las ausencias sin justificación a clase; y si el estudiante pertenece al grupo de repetidores o no. Estos resultados podrían constituirse en la guía para la elaboración de políticas educativas dirigidas a disminuir el fracaso académico o promover la excelencia educativa.

Durante el desarrollo de este estudio se han identificado ciertas limitaciones derivadas del análisis y de la propia base de datos. Primero, se han detectado problemas de endogeneidad en los datos. Este aspecto es sobre todo llamativo cuando se analizan variables como la preocupación de los padres por el rendimiento de sus hijos (a mayor preocupación, peores calificaciones), lo que podría estar evidenciando una causalidad inversa. Segundo, el análisis carece de validez externa. Dado que la recogida de datos que componen la muestra se ha realizado única y exclusivamente sobre el territorio canario, esto podría estar sesgando los datos a la realidad canaria y omitir circunstancias relevantes de otros lugares. Tercero, la presencia de valores perdidos o *missing data* en la muestra utilizada obliga a reducir considerablemente el número de individuos a emplear en las etapas de entrenamiento. Además, algunas técnicas de reemplazo de valores perdidos no han resultado ser efectivas por la falta de homogeneidad en la distribución de tales valores, los cuales se concentran en algunas variables explicativas.

Por tanto, como futuras líneas de investigación, este trabajo propone: ahondar en las razones que existen detrás de la falta de información en aquellas variables que no se han podido utilizar en este trabajo; la utilización de la metodología planteada en otras bases de datos, que sean capaces de recoger realidades más generales (como podría ser para la población española, o europea); y explorar mecanismos que permitan reducir el impacto de los problemas de endogeneidad.

7. BIBLIOGRAFÍA

Apaza Pacasi, R. (2015). Influencia económica familiar en el rendimiento académico de estudiantes universitarios*, Caso: Carrera de Economía de la UMSA* (Doctoral dissertation).

Arias Manrique, I. J., y Ávila Carreño, C. A. (2014). Influencia de los padres en el rendimiento académico de los hijos: una aproximación econométrica en el contexto de la educación media colombiana. *Educación y Desarrollo Social*, *8*(2), 184-199.

Berger, C., Álamos, P., Milicic, N., y Alcalay, L. (2014). Rendimiento académico y las dimensiones personal y contextual del aprendizaje socioemocional: evidencias de su asociación en estudiantes chilenos. *Universitas Psychologica*, *13*(2), 627-638.

Broc Cavero, M. Á. (2000). Autoconcepto, autoestima y rendimiento académico en alumnos de 4º de ESO Implicaciones psicopedagógicas en la orientación y tutoría. *Revista de investigación educativa*, *18*(1), 119-146.

Broc Cavero, M. Á. (2006). Motivación y rendimiento académico en alumnos de Educación Secundaria Obligatoria y Bachillerato LOGSE. *Revista de educación*.

Buenrostro Guerrero, A. E., Valadez Sierra, M. D., Soltero Avelar, R., Nava Bustos, G., Zambrano Guzmán, R., y García García, A. (2012). Inteligencia emocional y rendimiento académico en adolescentes. *Revista de educación y Desarrollo*, *20*(1), 29-37.

Cabrera Torres, A., Chacón Luna, A., y Vera Paredes, D. (2015). Incidencia del uso del internet en los adolescentes de las instituciones de educación media. *Revista ciencia UNEMI*, *8*(14), 57-66.

Castrillón, O. D., Sarache, W., y Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación universitaria*, *13*(1), 93-102.

Cedeño Cedeño, R. J., Vásquez Castro, P. C., y Maldonado Palacios, I. A. (2023). Impacto de las Tecnologías de la Información y la Comunicación (TIC) en el Rendimiento Académico: Una Revisión Sistemática de la Literatura. *Ciencia Latina Revista Científica Multidisciplinar*, *7*(4), 10297-10316.

Cerda Etchepare, G., y Pérez Wilson, C. (2014). Competencias matemáticas tempranas y actitud hacia las tareas matemáticas variables predictoras del rendimiento académico en educación primaria: resultados preliminares.

Cerda, G., Pérez, C., Elipe, P., Casas, J. A., y Del Rey, R. (2019). Convivencia escolar y su relación con el rendimiento académico en alumnado de Educación Primaria. *Revista de psicodidáctica*, *24*(1), 46-52.

Chilca Alva, M. L. (2017). Autoestima, hábitos de estudio y rendimiento académico en estudiantes universitarios. *Propósitos y representaciones*, *5*(1), 71-127.

Chinea Obal, C. S. (2018). La pobreza en Canarias. Un desafío económico y social. *Repositorio Universitario de la Universidad de La Laguna*.

Costa-Mendes, R., Oliveira, T., Castelli, M., y Cruz-Jesus, F. (2020). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*, 26, 1527–1547.

Delgado, C., y Estela, Y. (2020). Los determinantes del rendimiento académico en las regiones del Perú: un enfoque de Econometría Espacial.

Đurđević Babić, I. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*, 443-461.

Espinoza Guamán, E. E., Cruz Yaguachi, L. N., y Espinoza Freire, E. E. (2018). Las redes sociales y rendimiento académico. *Revista Metropolitana de Ciencias Aplicadas*, *1*(3), 38-44.

Espinoza, E. (2006). Impacto del maltrato en el rendimiento académico.

Ferrei Ortega, F. R., Vélez Mendoza, J., y Ferrel Ballestas, L. F. (2014). Factores psicológicos en adolescentes escolarizados con bajo rendimiento académico: depresión y autoestima. *Encuentros*, *12*(2), 35-47.

Fonseca, L., Pujals, M., Lasala, E., Migliardo, G., Aldrey, A., Buonsanti, L., y Barreyro, J. P. (2014). Desarrollo de habilidades de comprensión lectora en niños de escuelas de distintos sectores socioeconómicos. *Neuropsicología Latinoamericana*, *6*(1).

Franquesa Oliveres, M., y Zancajo Silla, A. (2010). Descomposición del efecto inmigrante en el rendimiento académico en Cataluña según la zona origen. *Investigaciones de Economía de la Educación volume 5*, *5*, 101-116.

García Fernández, N., Rivero Moreno, M. L., y Ricis Guerra, J. (2020). Brecha digital en tiempo del COVID-19. *Hekademos: revista educativa digital*, (28), 76-85.

González-Pienda García, J. A., Fernández Cueli, M., Suárez Fernández, N., Fernández Alba, M. E., Tuero Herrero, E., García Fernández, T., y Silva, E. H. D. (2012). Diferencias de género en actitudes hacia las matemáticas en la enseñanza obligatoria. *Revista iberoamericana de psicología y salud*.

González-Pienda, J. A. (2003). El rendimiento escolar. Una análisis de las variables que lo condicionan. *Revista Galego-Portuguesa de psicoloxía e educación.* N° 7 (Vol. 8) Ano 7°-2003 ISSN: 1138-1663

Guzmán, B., Véliz, M., y Reyes, F. (2017). Memoria operativa, comprensión lectora y rendimiento escolar. *Literatura y lingüística*, (35), 377-402.

Hernández Carrillo, C. (2018). Las Islas Canarias en PISA: mejora en lectura entre 2009 y 2015. *Repositorio Universitario de la Universidad de La Laguna*.

Hoffait, A. S., y Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1-11.

Hopson, L. M., Schiller, K. S., y Lawson, H. A. (2014). Exploring linkages between school climate, behavioral norms, social supports, and academic success. *Social Work Research*, *38*(4), 197-209.

Issah, I., Appiah, O., Appiahene, P., y Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision analytics journal*, 100204.

Jiménez Morales, M. I., y López Zafra, E. (2013). Impacto de la inteligencia emocional percibida, actitudes sociales y expectativas del profesor en el rendimiento académico. *Electronic Journal of Research in Educational Psychology*, *11*(1), 75-98.

Lamas Rojas, H. (2008). Aprendizaje autorregulado, motivación y rendimiento académico. *Liberabit*, *14*(14), 15-20.

Lamas, H. A. (2015). Sobre el rendimiento escolar. *Propósitos y representaciones*, *3*(1), 313-386.

López Mero, P., Barreto Pico, A., y del Salto Bello, M. W. A. (2015). Bajo rendimiento académico en estudiantes y disfuncionalidad familiar. *Medisan*, *19*(9), 1163-1166.

Martínez Vicente, M., Suárez Riveiro, J. M., y Valiente Barroso, C. (2020). Implicación estudiantil y parental en los deberes escolares: diferencias según el curso, género y rendimiento académico. *Revista de Psicología y Educación*, *15*(2), 151-165.

Masci, C., Johnes, G., y Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072-1085.

Mena, I., Romagnoli, C., y Valdés, A. M. (2008). ¿Cuánto y Dónde Impacta? Desarrollo de habilidades socio emocionales y éticas en la escuela.

Mizala, A., Romaguera, P., y Reinaga, T. (1999). Factores que inciden en el rendimiento escolar en Bolivia. *Centro de Economía Aplicada, Universidad de Chile*. Vol. 61.

Musso, M. F., Cascallar, E. C., Bostani, N., y Crawford, M. (2020). Identifying reliable predictors of educational outcomes through machine-learning predictive modeling. In *Frontiers in Education* (Vol. 5, p. 104).

Navas, L., Sampascual G., y Castejón, J. L. (1992). Atribuciones y expectativas de alumnos y profesores: influencias en el rendimiento escolar. *Revista de psicología general y aplicada: Revista de la Federación Española de Asociaciones de Psicología*, *45*(1), 55-62.

Olivar, A., y Daza, A. (2007). Las tecnologías de la información y comunicación (TIC) y su impacto en la educación del siglo XXI. *Negotium: revista de ciencias gerenciales*, *3*(7), 2.

Otero, M. R., Greca, I. M., y Silveira, F. L. D. (2003). Imágenes visuales en el aula y rendimiento escolar en Física: un estudio comparativo. *Revista electrónica de enseñanza de las ciencias. Ourense. Vol. 2, no. 1 (2003), p. 1-30*.

Paz Loscos, M. (1985). Meta-análisis sobre la predicción del rendimiento escolar.

Pendones Fernández, J. Á., Flores Ramírez, Y., Espino Olivas, G., y Durán Núñez, F. A. (2021). Autoconcepto, autoestima, motivación y su influencia en el desempeño académico. Caso: alumnos de la carrera de Contador Público. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(23).

Perpiñà Martí, G., Sidera Caballero, F., y Serrat Sellabona, E. (2021). Rendimiento académico en educación primaria: relaciones con la Inteligencia Emocional y las Habilidades Sociales. *Revista de educación*.

Robledo Ramón, P., y García Sánchez, J. N. (2009). El entorno familiar y su influencia en el rendimiento académico de los alumnos con dificultades de aprendizaje: revisión de estudios empíricos. *Aula abierta*, *37*(1), 117-128.

Rodriguez Gutierrez, I. M. (2020). Intervenciones educativas con estrategias didácticas bajo el enfoque socio cognitivo, mejora el desarrollo del aprendizaje en el área de historia en los estudiantes del cuarto grado “D” del Colegio de Alto Rendimiento–Ucayali, 2018.

Ros Morente, A., Filella Guiu, G., Ribes Castells, R., y Pérez Escoda, N. (2017). Análisis de la relación entre competencias emocionales, autoestima, clima de aula, rendimiento académico y nivel de bienestar en educación primaria. *Revista española de orientación y psicopedagogía*, *28*(1), 8-18.

Rubin, D. B., & Schenker, N. (1986). Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, 81(394), 366–374.

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., y Tessore, J. P. (2016). Tratamiento masivo de datos utilizando técnicas de Machine Learning.

Russo, V. (2019). Las competencias socioemocionales: su influencia en el rendimiento académico y las relaciones en el aula.

Serrano Díaz, M. N., Aragón Mendizábal, E. L., y Mérida Serrano, R. (2022). Percepción de las familias sobre el desempeño escolar durante el confinamiento por COVID-19.

Serrano Muñoz, A., Mérida Serrano, R., y Tabernero Urbieta, C. (2016). La autoestima infantil, la edad, el sexo y el nivel socioeconómico como predictores del rendimiento académico. *Revista de investigación en educación*, *14*(1), 33-66.

Sothan, S. (2019). The determinants of academic performance: evidence from a Cambodian University. *Studies in Higher Education*, *44*(11), 2096-2111.

Usai, M. G., Goddard, M. E., y Hayes, B. J. (2009). LASSO with cross-validation for genomic selection. *Genetics research*, *91*(6), 427-436.

Usán Supervía, P., y Salavera Bordás, C. (2018). Motivación escolar, inteligencia emocional y rendimiento académico en estudiantes de educación secundaria obligatoria. *Actualidades en psicología*, *32*(125), 95-112.

Vázquez, C., Cavallo, M., Aparicio, S., Muñoz, B., Robson, C., Ruíz, L., Florencia Secreto, M., Sepliarsky, P., y Eugenia Escobar, M. (2012). Factores de impacto en el rendimiento académico universitario. Un estudio a partir de las percepciones de los estudiantes. *Instituto de Investigaciones Teóricas y Aplicadas. Escuela de Contabilidad*.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., y Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, *104*, 106189.

Xu, X., Wang, J., Peng, H., y Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166-173.

Yıldız, M., y Börekci, C. (2020). Predicting academic achievement with machine learning algorithms. *Journal of educational technology and online learning*, *3*(3), 372-392.

Zuluaga González, V. H. (2013). Recesos escolares y rendimiento académico: Evidencia de un experimento natural.