Universidad Autónoma de Querétaro

Facultad de Contaduría y Administración

Maestría en Gestión e Innovación Pública Área terminal _____

Ciencia de Datos e AI: Análisis del indicador de reprobación en Educación Secundaria General

Tesis

Que como parte de los requisitos para obtener el grado de

Maestro en Gestión e Innovación Pública

Presenta:

Javier Sosa Franco

Dirigido por:

Dr. Jose Fernando Vasco Leal

Centro Universitario, Querétaro, Qro.
Junio de 2025

RESUMEN

La presente investigación tiene como objetivo proponer un modelo predictivo fundamentado en la ciencia de datos y la inteligencia artificial, que permita identificar de manera anticipada estudiantes que estén en riesgo de reprobar tercer grado de secundaria general en el estado de Querétaro. Esta problemática educativa se aborda desde una perspectiva analítica, considerando que la reprobación escolar impacta negativamente otros indicadores clave como la eficiencia terminal, la deserción y el rezago educativo. La metodología empleada se basó en el enfoque CRISP–DM, que permitió estructurar el análisis en fases bien definidas: compresión de la institución, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Se trabajó con un conjunto de datos históricos proporcionado por la USEBEQ, que incluyó registros académicos de primero y segundo grado de secundaria, datos institucionales, demográficos y variables complementarias. Como parte del tratamiento previo a la modelación, se aplicaron técnicas de ingeniería de características, balanceo mediante SMOTE, codificación de variables categóricas, selección de variables por relevancia estadística (SelectBest), escalado de características y validación cruzada estratificada. Se evaluaron múltiples algoritmos de clasificación supervisada, entre ellos, regresión logística, Random Forests, XGBoost, LightGBM, redes neuronales (Keras y Scikit–learn), así como un algoritmo de ensamble. Los modelos fueron comparados a partir de métricas centradas en la detección efectiva de la clase minoritaria (reprobación), entre las que se incluyeron: F1–score, precision, recall, MCC (Matthews Correlation Coefficient), coeficiente de Kappa, exactitud balanceada (Balanced Accuracy), área bajo la curva ROC (ROC–AUC), área bajo la curva precisión–recall (PR–AUC), mejora relativa frente al azar y una puntuación compuesta tipo WSM (Weighted Sum Model). El modelo seleccionado fue LightGBM por presentar un desempeño destacado en sus métricas, manteniendo un balance adecuado entre sensibilidad y especificidad. Estos resultados sugieren que

LightGBM es una alternativa robusta y eficiente para identificar estudiantes en riesgo y apoyar decisiones pedagógicas oportunas. El modelo puede ser integrado como una herramienta de análisis institucional de la USEBEQ para diseñar intervenciones preventivas y focalizadas que contribuyan a mejorar los indicadores de permanencia y logro educativo.

Palabras clave: aprendizaje automático, ciencia de datos, LightGBM, políticas públicas preventivas, predicción escolar

ABSTRACT

This research aims to propose a predictive model based on data science and artificial intelligence to identify, in advance, students at risk of failing the third grade of general secondary school in the state of Querétaro. This educational challenge is addressed from an analytical perspective, considering that school failure negatively affects other key indicators such as terminal efficiency, school dropout, and educational lag. The methodology employed was grounded in the CRISP–DM framework, which allowed for a structured analysis across clearly defined phases: understanding the institution, data understanding, data preparation, modeling, evaluation, and deployment. The study was conducted using a historical dataset provided by USEBEQ, which included academic records from first and second grades of secondary education, along with institutional, demographic, and complementary variables. As part of the preprocessing pipeline, techniques such as feature engineering, SMOTE balancing, categorical variable encoding, feature selection based on statistical relevance (SelectBest), feature scaling, and stratified cross–validation were applied. Several supervised classification algorithms were evaluated, including logistic regression, Random Forests, XGBoost, LightGBM, neuronal networks (keras and Scikit–learn), as well as an ensemble algorithm. The models were compared using metrics focused on effectively detecting the minority class (school failure), including: F1–score, precision, recall, MCC (Matthews Correlation Coefficient), Kappa coefficient, balanced accuracy, Area Under the ROC Curve (ROC–AUC), Area Under the Precision–Recall Curve (PR–AUC), improvement over random baseline, and a composite score based on the Weighted Sum Model (WSM). The selected model was LightGBM due to its outstanding performance across key metrics, maintaining a suitable balance between sensitivity and specificity. These results suggest that LightGBM is a robust and efficient alternative for identifying students at academic risk and supporting timely pedagogical decisions. The model can be integrated as an institutional analysis tool

within USEBEQ to design preventive and targeted interventions that help improve educational persistence and achievement indicators.

Keywords: data science, educational prediction, LightGBM, machine learning, preventive public policy.

DEDICATORIAS

Este proceso se lo dediqué a las personas que siempre han estado y estarán en mi vida y que las amaré por siempre. Para mí, todas son importantes y las nombraré por orden de aparición en mi vida.

A mi madre Sofía, que siempre estuvo al pendiente de mi salud y de mi persona, gracias.

A mi padre Arturo, que con sus enseñanzas y liderazgo forjó gran parte de mi carácter, gracias.

A mi hermana mayor Azucena, siempre ha sido un ejemplo de rectitud y disciplina, gracias.

A mi hermana Margarita, siempre ha sido un ejemplo de lucha y perseverancia, gracias.

A mi hermano Enrique, que siempre estuvo conmigo con su empatía y sensibilidad enseñándome cómo conducirme, muchas gracias.

A mi hermana menor Nelly, que siempre ha sido un ejemplo de esfuerzo y tenacidad, gracias.

A mi esposa Rosa María, que siempre estuvo al pendiente de la alimentación y cuidado de nuestros hijos, Muchas gracias

A mi hijo Javier, que siempre me da sorpresas y enseñanzas, Muchas gracias.

A mi hija Yanneli, que es un ejemplo de arrojo y valentía, siempre me sorprende, gracias.

A todos, muchas gracias.

AGRADECIMIENTOS

Quiero agradecer de forma especial a mi director de tesis Dr. José Fernando Vasco Leal, por confiar en mí, guiarme y apoyarme para culminar este proyecto.

Agradezco a mis sinodales: Mtro. Carlos Olguín González, Mtra. Larissa Cruz Gutiérrez, Mtra. Elia Socorro Pérez Díaz y Mtro. Alfonso G. Nieto Irigoyen por su noble entrega a la labor docente, que desafortunadamente no se le da el merecido reconocimiento.

De igual manera quiero agradecer a la Mtra. Alejandra Monreal León, por su invaluable apoyo, en especial en este proyecto.

Agradezco a la Unidad de Servicios para la Educación Básica en el Estado de Querétaro por el apoyo que siempre me han brindado.

Agradezco al Lic. Leopoldo Bárcenas Hernández a la Ing. Alma Delia Morales Apolonio, al Maestro Simón Ramírez Olvera, a mis compañeros de trabajo y a todos mis amigos que siempre me han dado su apoyo.

Por último, quiero expresar mi agradecimiento a la Campaña ¡Titúlate Ya! 2025 de la Facultad de Contaduría y Administra

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

La educación es un pilar fundamental para el desarrollo social, económico y humano de cualquier país, en Querétaro, el sistema educativo presenta retos constantes relacionados con los indicadores de aprovechamiento escolar, de forma particular la deserción y el de reprobación. Esta problemática no solo impacta a nivel institucional, sino que genera consecuencias a largo plazo en las trayectorias académicas y personales de los estudiantes. especialmente en aquellos en contextos de vulnerabilidad. Por tal motivo, atender el fenómeno de la reprobación escolar desde una perspectiva preventiva y basada en los datos se vuelve una necesidad apremiante.

El presente trabajo tiene como propósito general proponer un modelo predictivo que permita identificar a los alumnos de tercer grado de secundaria general con alto riesgo de reprobación escolar, utilizando datos académicos e institucionales de primero y segundo grado de secundaria general.

La propuesta se alinea con el Plan Estatal de Desarrollo de Querétaro 2021–2027, específicamente con el Eje 2 – Educación, Cultura y Deporte, que establece retos específicos, como “pasar a los primeros 15 lugares entre las entidades con menor tasa de abandono en secundaria” y “reducir el abandono escolar en 1% en secundaria y educación media superior con respecto al año anterior” (Gobierno del Estado de Querétaro, 2021, pp. 146–147). También se alinea a los objetivos del Programa Institucional 2021–2027 de la USEBEQ, donde se establece como prioridad asegurar trayectorias educativas completas, reducir el abandono escolar y fortalecer la eficiencia terminal (USEBEQ, 2021).

Como se analiza en el apartado 1.1 de esta tesis, los indicadores educativos en el estado de Querétaro muestran una evolución preocupante, según datos estadísticos de la USEBEQ (2025), la reprobación en secundaria bajó a niveles mínimos durante la pandemia de COVID-19, pero volvió a repuntar en el ciclo 2022–

2023. Esta tendencia sugiere una necesidad urgente de establecer mecanismos de alerta temprana. Adicionalmente, el rezago educativo afecta con más del 40% en 10 de los 18 municipios (USEBEQ, 2021 p. 5), aunque el rezago educativo no se mide por el índice de reprobación, sí se ve influido por trayectorias escolares incompletas.

Organismos nacionales e internacionales han documentado que la educación es una de las dimensiones estructurales más importantes para combatir la pobreza. De acuerdo al Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL, 2020), el rezago educativo forma parte de los indicadores que determinan la pobreza multidimensional en México, el Banco Mundial (2018), señala que el bajo nivel educativo se relaciona directamente con el acceso limitado a empleos formales y menores ingresos. Asimismo, la Red de Pobreza Multidimensional (OPHI, s.f.) sostiene que la educación es una de los tres pilares fundamentales del índice de Pobreza Multidimensional, junto con la salud y el nivel de vida. En consecuencia, atender los problemas escolares desde la raíz –como lo es la reprobación escolar– puede tener impactos sociales significativos a largo plazo.

En los últimos años la ciencia de datos ha demostrado ser una herramienta poderosa en el ámbito educativo. La minería de datos educativos (Educational Data Mining) y el aprendizaje automático (Machine Learning) se han utilizado en diferentes sectores para hacer predicciones en diferentes temas, en el educativo, abandono escolar, rendimiento académico y en la optimización de procesos (Medina Romero, M. Á., & Ochoa Figueroa, R. 2025; Ortiz Ocaña, A., 2025; Anaya Benítez, F., 2023; Romero, C., & Ventura, S. 2013). Estas metodologías permiten construir modelos capaces de anticipar el desempeño de los estudiantes y apoyar a los directivos escolares a tomar decisiones informadas.

Para el desarrollo de este trabajo se consultaron fuentes científicas, metodológicas e institucionales públicas entre 2013 y 2025. La revisión abarcó literatura sobre inteligencia artificial, analítica educativa, evaluación institucional,

políticas públicas en educación básica y de modelos de predicciones. Se ha podido identificar que el uso de modelos predictivos aún es escaso en sistemas educativos estatales, los cuales representan una estrategia innovadora con alto potencial de impacto.

Este trabajo se estructura en cuatro capítulos. El capítulo 1, presenta el planteamiento del problema, objetivos, hipótesis y justificación, en este se da un contexto de la institución y la problemática que se tiene. El capítulo 2, presenta el marco teórico, proporcionando un contexto sobre la inteligencia artificial en la actualidad, la ciencia de datos y las principales metodologías empleadas en el desarrollo de estos proyectos. Además, se abordan los diferentes tipos de análisis de datos, los tipos de aprendizaje automático y los modelos más utilizados. Finalmente, en el apartado 2.6, dedicado a las métricas, se describen las herramientas y técnicas empleadas para evaluar el rendimiento, la estabilidad y la robustez de los modelos de aprendizaje automático. El capítulo 3, describe la estructura y diseño de la metodología de la investigación, así como, instrumentos de trabajo y la población objetivo. En este capítulo se detalla cómo fue el desarrollo del proyecto conforme a la metodología que se utilizó para desarrollar la propuesta. Para la presente investigación, la cual está basada en un proyecto de ciencia de datos se describe la metodología CRISP–DM, en sus 6 fases y se desarrolla el proyecto en cada una de ellas conforme a las características que se definen en el marco teórico. En el capítulo 4, se presentan los resultados que se obtuvieron de la aplicación de la metodología en el desarrollo de la propuesta, desde el entendimiento del negocio hasta la selección y evaluación del modelo seleccionado. Finalmente, en el capítulo final, se encuentran las conclusiones donde se hace una recapitulación de lo que se desarrolló en las diferentes fases de la investigación, espacio donde se comenta si el proyecto cumple con los objetivos que planteamos al inicio del proyecto, cumple o no con la hipótesis y la pregunta de investigación.

Con este trabajo se pretende ir más allá de un proyecto técnico, se busca aportar a la construcción de un sistema educativo más justo, eficiente y preventivo.

Al poner este tipo de herramientas al servicio de la toma de decisiones, se abona a la consolidación de una administración pública basada en el conocimiento.

Por último, quiero expresar mi agradecimiento a la Campaña ¡Titúlate Ya! 2025 de la Facultad de Contaduría y Administración, de la Universidad Autónoma de Querétaro (UAQ) por darnos la oportunidad de cumplir y concluir nuestra meta, por el apoyo, atención y empatía de las personas que estuvieron en el diseño, planeación, gestión y operación del programa. Esta iniciativa no solo nos dio una estructura metodológica y acompañamiento académico, también reafirmó el compromiso social que tiene la universidad para darnos la oportunidad de aplicar el conocimiento para transformar la realidad educativa del estado.

1.1 Planteamiento del Problema

La ONU define que “la educación es el fundamento básico para la construcción de cualquier sociedad. Es la inversión única que los países pueden realizar para construir sociedades equitativas, saludables y prósperas”.

La Declaración Universal de los Derechos Humanos de 1948 en su artículo 26, señala que la educación debe de ser gratuita y obligatoria por lo menos en su formación elemental y que toda persona tiene el derecho a recibirla (Naciones Unidas, s.f., art. 26).

Conforme al Artículo tercero de la Constitución Política de los Estados Unidos Mexicanos, toda persona tiene derecho a la educación, la educación básica estará conformada por la educación inicial, preescolar, primaria y secundaria, y será gratuita y obligatoria al igual que la educación media superior.

La rectoría de la educación corresponde al Estado – Federación, Entidades Federativas (Estados) Ciudad de México y Municipios – y velará porque sea universal, inclusiva, pública, gratuita y laica (Cámara de Diputados, 2024, art. 3).

De acuerdo con el reglamento interno de la USEBEQ, el 7 de junio de 1992, se llevó a cabo la desconcentración de los servicios educativos, que pasaron de estar centralizados en la Secretaría de Educación Pública (SEP) a ser administrados por cada uno de los estados. Ese mismo año se firmó el decreto de creación de la Unidad de Servicios Para la Educación Básica en el Estado de Querétaro (USEBEQ), que desde entonces se encarga de gestionar los recursos y la información escolar de la educación básica pública y privada en el estado de Querétaro (USEBEQ, 2020).

De acuerdo a la tabla 1, la estadística publicada en el sitio oficial de la USEBEQ en todo el estado, en el ciclo escolar 2022 – 2023 se tenían 654,719 estudiantes en todos los niveles educativos, 37,144 docentes en 4,336 escuelas. Del total de estudiantes 460,485 corresponden a educación básica (educación Inicial, Preescolar, Primaria y Secundaria), el 70.33%, 20,283 docentes en 3,843 escuelas. Existe un descenso considerable entre el número de estudiantes del nivel educativo de primaria y el de secundaria como lo podemos ver en la tabla 1. Este fenómeno se debe a diferentes factores como pueden ser económicos, sociales, culturales, de disponibilidad, entre otros.

ESTADÍSTICA POR NIVEL EDUCATIVO
(PÚBLICO Y PRIVADO)

NIVEL	ALUMNOS			GRUPOS	DOCENTES	ESCUELAS
	TOTAL	HOMBRES	MUJERES			
TOTAL	**654,719**	**322,284**	**332,435**	**21,149**	**37,144**	**4,336**
Inicial	5,382	2,693	2,689	364	218	131
[2] Especial	2,194	1,395	799	231	284	24
Básica [1]	**455,103**	**229,560**	**225,543**	**17,287**	**20,065**	**3,712**
Preescolar	81,481	41,108	40,373	4,195	4,178	1,565
Primaria	253,497	128,675	124,822	9,200	9,233	1,551
Secundaria	120,125	59,777	60,348	3,892	6,654	596
Media Superior [3]	**91,062**	**42,867**	**48,195**	**3,267**	**5,173**	**330**
Bachillerato General	62,586	28,396	34,190	2,449	5,173	263
Bachillerato Tecnológico	28,434	14,461	13,973	812		65
Profesional Técnico	42	10	32	6		2
Superior [3,4]	**100,978**	**45,769**	**55,209**	**-**	**11,404**	**139**
Técnico Superior	6,064	3,559	2,505	NA	11,404	139
Licenciatura	88,231	39,443	48,788			
Posgrado	6,683	2,767	3,916			
Inicial/Especial/Básica	**462,679**	**233,648**	**229,031**	**17,882**	**20,567**	**3,867**

Tabla 1
Estadística por nivel educativo ciclo escolar 2022–2023 (público y privado)
Nota. Adaptada de Estadística por nivel educativo, por USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/PaginWEB/Estadisitica/indexEstadisticas

Por su parte, en la tabla 2, el nivel de educación secundaria la conforman secundaria general, técnica, telesecundaria y comunitaria. Para este nivel educativo se tienen 120,125 estudiantes, 6,654 docentes y 596 escuelas, como se muestra en la tabla 2. En Educación secundaria comunitaria el número de estudiantes es 852 y son gestionadas por el Consejo Nacional de Fomento Educativo (CONAFE) aunque se integran a información estatal. Secundaria general es el servicio que cuenta con el mayor número de estudiantes 62,370.

ESTADÍSTICA POR NIVEL EDUCATIVO / SERVICIO

NIVEL / SERVICIO	ALUMNOS			GRUPOS	DOCENTES	ESCUELAS
	TOTAL	HOMBRES	MUJERES			
TOTAL	**654,719**	**322,284**	**332,435**	**21,149**	**37,144**	**4,336**
INICIAL	**5,382**	**2,693**	**2,689**	**364**	**218**	**131**
Lactantes	1,807	901	906	142	64	131
Maternal	3,515	1,766	1,749	218	151	
Indígena	60	26	34	4	3	
ESPECIAL - CAM	**2,194**	**1,395**	**799**	**231**	**284**	**24**
Inicial	9	8	1	1	284	24
Preescolar	167	115	52	24		
Primaria	922	577	345	113		
Secundaria	750	482	268	73		
Formación para el Trabajo	278	163	115	20		
Apoyo Complementario	68	50	18	NA		
BÁSICA	**455,103**	**229,560**	**225,543**	**17,287**	**20,065**	**3,712**
PREESCOLAR	**81,481**	**41,108**	**40,373**	**4,195**	**4,178**	**1,565**
General	73,328	36,958	36,370	3,425	3,408	927
Indígena	2,910	1,501	1,409	147	147	84
Comunitario	5,243	2,649	2,594	623	623	554
PRIMARIA	**253,497**	**128,675**	**124,822**	**9,200**	**9,233**	**1,551**
General	245,374	124,528	120,846	8,706	8,706	1,263
Indígena	5,862	2,961	2,901	280	280	74
Comunitario	2,261	1,186	1,075	214	247	214
SECUNDARIA	**120,125**	**59,777**	**60,348**	**3,892**	**6,654**	**596**
General	62,370	30,900	31,470	1,910	4,084	223
Técnica	29,039	14,540	14,499	765	1,353	56
Telesecundaria	27,864	13,916	13,948	1,125	1,125	241
Comunitario	852	421	431	92	92	76
MEDIA SUPERIOR	**91,062**	**42,867**	**48,195**	**3,267**	**5,173**	**330**
Bachillerato General	62,586	28,396	34,190	2,449	5,173	263
Bachillerato Tecnológico	28,434	14,461	13,973	812		65
Profesional Técnico	42	10	32	6		2
SUPERIOR	**100,978**	**45,769**	**55,209**	**-**	**11,404**	**139**
Técnico Superior	6,064	3,559	2,505	NA	11,404	139
Licenciatura	88,231	39,443	48,788			
Posgrado	6,683	2,767	3,916			

Tabla 2

Estadística por nivel educativo /servicio 2022–2023

Nota. Adaptada de *Estadística por nivel educativo / servicio*, por USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/PaginWEB/Estadisitica/indexEstadisticas

Analizando la tabla 3, de los principales indicadores de reprobación, deserción, egreso y eficiencia terminal de fin de ciclo de los años del 2018 a 2023, se puede decir que todos se ven afectados de forma negativa. En los ciclos 2020–

2021 y 2021–2022 debido a la pandemia por COVID-19, se presentó una situación atípica en los indicadores por lo que no es conveniente considerarlos para el análisis. Para los otros ciclos se puede ver que en 2018–2019 y 2019–2020 los indicadores en general presentan una caída y en el ciclo 2022–2023 se tiene una mejoría conforme a su dirección, si los comparamos con los indicadores del ciclo 2018–2019 presentan una caída.

NIVEL +	INDICADOR		FIN DE CICLO				
		Dirección del indicador	2018-2019 %	2019-2020 %	2020-2021 %	2021-2022 %	2022-2023 %
PRIMARIA	REPROBACIÓN	Desc.	1.2	0.5	0.21	0.08	0.64
	DESERCIÓN	Desc.	-0.22	0.47	-0.09	0.69	-0.63
	EGRESIÓN	Asc.	98.95	99.65	99.56	98.66	99.07
	EFICIENCIA TERMINAL	Asc.	101.81	103.44	103.13	102.67	101.99
SECUNDARIA	REPROBACIÓN	Desc.	7.92	4.09	0.7	0.46	5.22
	DESERCIÓN	Desc.	4.63	3.17	1.98	5.13	3.33
	EGRESIÓN	Asc.	93.71	95.76	98.06	94.91	94.74
	EFICIENCIA TERMINAL	Asc.	84.37	87.95	92.79	91.71	87.52

Tabla 3
Comparativo de indicadores de desempeño escolar por nivel educativo de los ciclos escolar 2018–2019 a 2022–2023
Nota. Elaboración propia con base en la información de los ciclos escolares 2018–2019 a 2022–2023, por USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/PaginWEB/Estadisitica/indexEstadisticas

Estos indicadores están estrechamente relacionados sin considerar la causa que los propicia en cada estudiante, pues para cada uno de ellos es diferente la situación que lo genera. Se puede decir que la reprobación es el origen que provoca los demás indicadores ya que a mayor reprobación se incrementa la deserción, disminuye el egreso y la eficiencia terminal. Se puede ver claramente en la tabla de indicadores.

Por ejemplo, regresando al análisis y comparando los indicadores del ciclo 2018–2019 contra 2020–2023 se puede ver que baja la reprobación de 7.92 a 5.22 al igual que la deserción, baja de 4.63 a 3.33 y se incrementan el egreso de 93.71 a 94.74 y la eficiencia terminal de 84.37 a 87.52.

Como se mencionó anteriormente los indicadores se relacionan entre sí de una manera significativa, como se aprecia en la figura 1. De acuerdo con datos del

Programa Institucional de la USEBEQ 2021–2027, podemos percatar como el rezago educativo está marcado en la mayoría de los municipios del estado, de 18 municipios 10 tienen más del 40% de rezago educativo (USEBEQ, 2021, p. 5). Si bien ese porcentaje no se calcula con el número de estudiantes reprobados si lo hace con el número de personas que no concluyeron la educación básica.

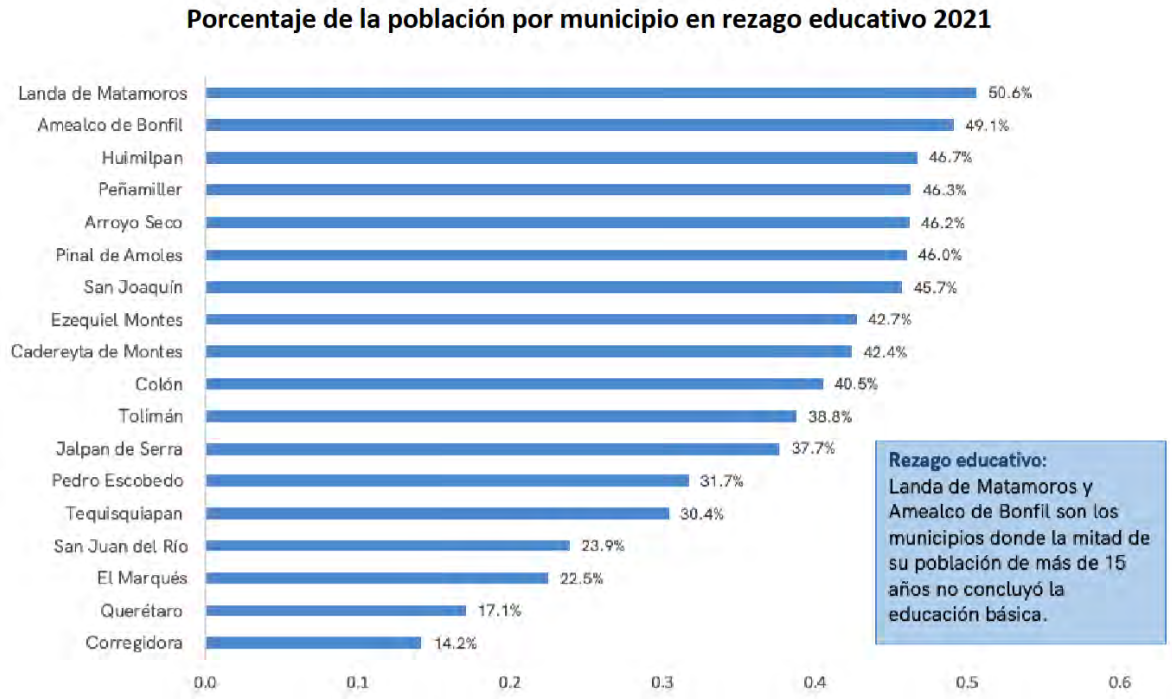


Figura 1
Porcentaje de la población por municipio en rezago educativo, 2021
Nota. Adaptado de datos de USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/Content/Intranet/PROG%20INSTITUCIONAL%20DE%20USEBEQ%2021-27%20vf.pdf

El rezago educativo de igual manera impacta en el indicador de pobreza y de pobreza extrema como lo mencionan diferentes organizaciones como el CONEVAL (2009), el Banco Mundial (Banco Mundial, s.f.–b) o la Red de Pobreza Multidimensional (IPM, s.f.), por nombrar algunas, ellas definen que la pobreza tiene un enfoque multidimensional pues para que se presente intervienen diferentes factores y entre ellos está la educación.

Por lo que se puede ver, para que se mejore el bienestar de las personas y sus familias, se tienen que abordar estos problemas por su origen, si bien, la reprobación escolar es una consecuencia de otros factores, se deben de realizar acciones para contribuir de una manera positiva en el mejoramiento de estos.

1.2. Justificación de la Investigación

De acuerdo con datos del CONEVAL, en 2024 se informó que la pobreza en México en el año 2022 era de 36.3% (CONEVAL, 2023, p. 24). El Banco Mundial (Banco Mundial, s.f.–b), así como la Organización para la Cooperación y el Desarrollo Económicos (OCDE), reconocen que la pobreza es un fenómeno multidimensional, es decir, que son varios factores los que la provocan, uno de ellos es el rezago educativo. El CONEVAL considera que una persona tiene esta carencia cuando no se tiene la escolaridad obligatoria en la edad estipulada para asistir a los niveles educativos vigentes (2009).

En otro informe del año 2021, el CONEVAL señala que, en 2020 en Querétaro, 10 de 18 municipios tienen más del 40% de rezago educativo en su población. Esto quiere decir que estas personas pueden pasar a formar parte del grupo de la pobreza, si su ingreso está por debajo de la línea de bienestar.

Conforme se plantea en el Programa Institucional 2021–2027, la USEBEQ busca consolidar un sistema educativo centrado en el aprendizaje, con un enfoque humanista, inclusivo y orientado al desarrollo de habilidades digitales, socioeconómicas, socioemocionales y cognitiva, promoviendo trayectorias escolares completas para niñas, niños y adolescentes.

En su visión estratégica, USEBEQ plantea que para el año 2027 la educación básica en el estado de Querétaro debe posicionarse entre los primeros cinco lugares nacionales, sustentada en indicadores clave como la cobertura, infraestructura y conectividad. Este propósito se alinea con el Plan Estatal de

Desarrollo 2021–2027 y el Programa Sectorial de Educación, alineando así los objetivos de la institución con los objetivos estratégicos del gobierno estatal.

La misión de USEBEQ se centra en “ofrecer una educación de calidad centrada en el aprendizaje de alumnos de 3 a 14 años que propicie condiciones para impulsar las habilidades digitales y socioemocionales en la construcción del conocimiento” (USEBEQ, 2021, p. 13). Busca asegurar el desarrollo integral de las niñas, niños y adolescentes, con respecto a la dignidad humana y a la equidad en el acceso a la educación.

La planeación está organizada en seis objetivos (USEBEQ, 2021, pp. 13–15).

Fomentar y reactivar la práctica deportiva

Fomentar la cultura a favor de la educación inicial.

Generar oportunidades equitativas de acceso mediante la ampliación de la cobertura de la educación obligatoria,

Asegurar trayectorias educativas completas con aprendizajes significativos, promoviendo la permanencia escolar.

Impulsar que las instituciones del sector educativo tengan buena gobernanza y eficiente administración.

Coordinar acciones interinstitucionales para reducir el rezago educativo y el analfabetismo.

De acuerdo a lo anterior, el objetivo 4 –asegurar trayectorias educativas completas– tiene una relación directa con el propósito del proyecto, pues la reprobación escolar es una de los principales problemas por lo que los estudiantes no concluyen de forma exitosa sus estudios. Asimismo, el uso de herramientas tecnológicas como los modelos predictivos está alineada con la estrategia 5.2 del programa, la cual busca “utilizar los resultados de las evaluaciones para detectar

áreas de oportunidad e implementar acciones que coadyuven en la mejora de recursos y capacidades de los distintos actores del sistema educativos" (USEBEQ, p.15).

El análisis institucional muestra que se tiene un marco estratégico preparado para incorporar modelos predictivos como el que se presenta en esta tesis. La propuesta de detección temprana de estudiantes en riesgo de reprobación se alinea con los objetivos de eficiencia, equidad y mejora educativa planteados en el plan, también, es una muestra clara de los principios de gobernanza basado en datos que la institución pretende consolidar.

Para identificar las fortalezas y debilidades institucionales se muestra la tabla 4 que es un análisis FODA que contiene el Manual de Calidad de la USEBEQ (USEBEQ, s.f.). En el FODA se identifican factores externos e internos que afectan el desempeño educativo estatal. Las principales fortalezas son la disposición del personal, la vocación de los docentes, la estructura de recursos humanos suficiente y la capacidad para realizar cambios. También, se tiene como fortaleza un grupo directivo con experiencia y una cultura organizacional de seguridad, de igual manera una creciente cultura institucional sobre el uso de tecnologías. En debilidades se enlista la obsolescencia de equipo, instalaciones escolares en mal estado, problemas de conectividad y procesos administrativos con áreas de oportunidad. Se identifica un elevado ausentismo del personal y la necesidad de fortalecer la convivencia escolar.

En cuanto al entorno externo, se identifican como amenazas factores como la violencia intrafamiliar, la desintegración familiar, la migración poblacional, la disminución de puestos federales, las adicciones, el cambio climático, los efectos de la inseguridad y la crisis económica. También, se tienen oportunidades, el acceso a programas federales y estatales de apoyo a la educación, el crecimiento de la conectividad digital, la buena relación entre instituciones públicas y privadas, y el tener plataformas tecnológicas para poner al alcance de todos los servicios (USEBEQ, s.f., pp. 24–26).

EXTERNO	INTERNO
AMENAZAS	**OPORTUNIDADES**
Migración poblacional al Estado	Disposición del personal
Disminución del presupuesto federa	Sistema de educación pequeño en comparación con otros estados
Desintegración familia	Grupo Directivo con experiencia
Violencia intrafamiliar	Vocación de los docentes
Influencia negativa de los medios	Estructura de recursos humanos suficientes
Incremento de inseguridad	Buena relación con los sindicatos
Incremento de adicciones	
Cultura de migración	
Falta de cultura alimenticia saludable	
Pandemia de COVID -19	Capacidad para implementar cambios
Contingencias sanitarias	Adecuación para uso de tecnología (teletrabajo)
Crisis económica	Cultura Organizacional de Seguridad
Cambio climático	
OPORTUNIDADES	**DEBILIDADES**
Programas Federales de Apoyo a la Educación	Comunicación
Buena relación con la Secretaria de Educación Estata	Instalaciones escolares en mal estado
Acceso al conocimiento global a través de las tecnologías de la Información y la Comunicación	Conectividad
Legislación nueva y cambiante	Obsolesencia y falta de equipo
Cambios en los hábitos del usuario o consumidor	Equipamiento y mantenimiento de los equipos de computo
Servicio de Internet	Proceso de supervisión e inspección
Programas Estatales de Apoyo a la Educación	Proceso de asignación de claves y plazas para sustitución de personal por cualquier eventualidad o causa
Buena relación con la Estructura Directiva (Vinculación)	Convivencia escolar
Buena relación con instituciones públicas (Vinculación)	Elevado ausentismo de personal por diferentes motivos
	Densidad de empleados
	Infraestructura más o menos adecuada
Diferenciarnos de las instituciones adscritas a la Secretaría de Educación Pública en el Estado de Querétaro estableciendo compromisos y acciones con relación al cambio climático.	

Tabla 4
Análisis FODA de USEBEQ
Nota. Adaptado de datos de USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/Content/SGC/Documentos/ManualCalidad.pdf

De acuerdo a la información anterior, se presenta un formato CANVAS, en la figura 2, para esquematizar el diagnóstico, se presenta el funcionamiento estratégico y operativo de la institución. La propuesta de valor se centra en brindar una educación de calidad, gratuita y organizada, con un enfoque que se centra en la escuela, que promueva la inclusión, la equidad, y el desarrollo integral de las niñas, niños y jóvenes. Esto se logra mediante la capacitación docente y administrativa, automatización de procesos, el fortalecimiento de la comunicación y la atención a las comunidades escolares.

Los socios claves que colaboran se tienen al Gobierno Estatal y Federal, padres de familia, personal docente, universidades, ONGs y organismos sindicales.

Los segmentos de clientes a los que se orientan los servicios incluyen estudiantes, padres de familia, docentes, personal administrativo, escuelas particulares y públicas, así como organismos gubernamentales.

También, contempla los recursos claves que sustentan la operación: infraestructura educativa, recursos humanos. tecnologías. materiales y financieros. Lista los canales de atención (plataformas digitales, oficinas físicas, redes sociales), las fuentes de ingreso, y a la estructura de costos, destacando el gasto de nómina, infraestructura, tecnología y programas sociales.

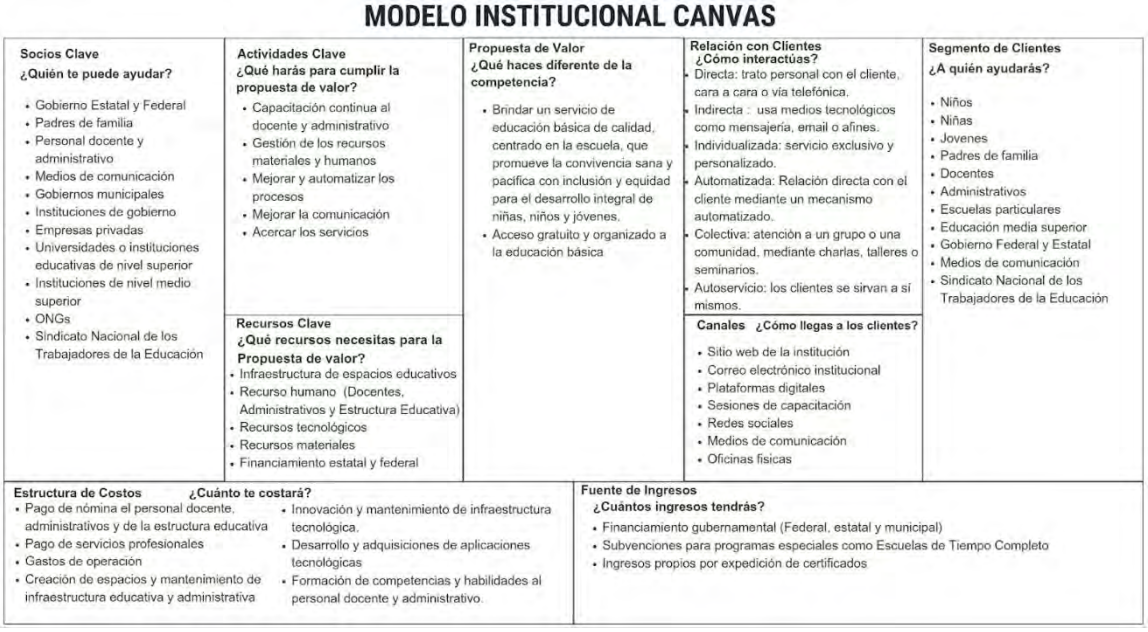


Figura 2
Modelo institucional CANVAS
Nota. Elaboración propia con datos de USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/

Como se analiza en el planteamiento del problema los principales indicadores del nivel secundaria muestran una evolución preocupante en los ciclos escolares recientes, comparando los ciclos del 2018 al 2023. Aunque hubo una aparente mejora durante los ciclos afectados por la pandemia, los datos reportan un repunte en la reprobación, esto requiere de una intervención temprana y precisa.

De acuerdo a los datos oficiales por la USEBEQ (2025) muestran una evolución preocupante de los indicadores en los últimos ciclos escolares. El indicador de reprobación en secundaria en el ciclo escolar 2018–2019 fue de 7.92%, se redujo considerablemente en los dos ciclos siguientes hasta 0.46 en el ciclo 2021–2022, probablemente por las políticas educativas que se dieron durante el periodo de la pandemia. Para el ciclo escolar 2022–2023 sube nuevamente a 5.22%, esto refleja un rebote y abre la preocupación por el rezago académico acumulado.

En esta tabla se observa que los valores de los diferentes indicadores presentan un comportamiento similar, lo que indica la existencia de una correlación entre ellos. A simple vista, se puede deducir que un aumento en la reprobación está asociado con un incremento en la deserción escolar, una disminución en el número de egresados y en consecuencia, una menor eficiencia terminal.

Esto evidencia que la reprobación es un indicador crítico y estratégico, ya que actúa como detonante de otros problemas educativos. La tabla 5, tomada del Programa Institucional 2021–2027, presenta los principales indicadores correspondientes al año 2021.

Principales indicadores educativos, 2021. Porcentajes

INDICADOR	QUERÉTARO	NACIONAL
Grado Promedio de Escolaridad (Años) 2020	10.5	9.7
Rezago Educativo 2020	23.8	30.4
Analfabetismo 2020 (censo 2020, INEGI)	3.5	4.7
Cobertura de Preescolar inicio de ciclo 2020-2021	67.1	65.9
Cobertura de Primaria inicio de ciclo 2020-2021	100.0	100.0
Cobertura de Secundaria inicio de ciclo escolar 2020-2021	100.0	95.8
Absorción en Secundaria inicio de escolar 2020-2021	97.9	94.5
Abandono Escolar Primaria fin de ciclo 2019-2020	0.5	0.4
Abandono Escolar Secundaria fin de ciclo 2019-2020	3.2	2.7
Eficiencia Terminal Primaria fin de ciclo 2019-2020	103.4	96.0
Eficiencia Terminal Secundaria fin de ciclo 2019-2020	88.00	88.5

Tabla 5
Principales indicadores educativos, 2021. Porcentajes
Nota. Adaptado de datos de USEBEQ, 2025. Recuperado de https://www.usebeq.edu.mx/Content/Intranet/PROG%20INSTITUCIONAL%20DE%20USEBEQ%2021-27%20vf.pdf

Se pueden ver 3 indicadores que son de llamar la atención, abandono escolar (2019–2020), Querétaro tiene un valor 3.2%, media nacional valor 2.7%, esto indica que estamos por arriba de la media nacional. La eficiencia terminal Secundaria (2019–2020), Querétaro tiene un valor de 88.0%, media nacional tiene el valor de 88.5%, estamos por debajo de la media nacional. Rezago Educativo (2020), Querétaro tiene un valor de 23.8%, media nacional 30.4%, estamos por arriba de la media nacional, pero si se analiza ese valor se puede decir que 1 de cada 4 personas se encuentran en esa situación de vulnerabilidad.

También, como se mencionó en el planteamiento del problema en el Programa Institucional 2021–2027 se puede ver que 10 de los 18 municipios que se tienen en el estado tienen más del 40% de rezago educativo. Aunque este indicador se calcula con base en la conclusión de niveles educativos, sus causas pueden ser el resultado de diferentes fenómenos y uno de ellos puede ser la reprobación.

Ante esta situación, se justifica la creación de políticas públicas que busquen disminuir este fenómeno, una de ellas podría ser el crear modelos de prevención anticipada, que permitan detectar estudiantes que estén en esa situación de vulnerabilidad. Aunque la reprobación escolar, igualmente puede ser consecuencia de múltiples factores (sociales, económicos, personales, entre otros), contar con una herramienta que ayude a identificar casos en riesgo antes de que inicie el ciclo escolar, permite diseñar estrategias de atención dirigidas, intervenciones pedagógicas específicas y una gestión de recursos más eficiente.

El desarrollo de esta propuesta se sustenta en el uso de herramientas de análisis de datos que permiten modelar comportamientos y anticipar resultados. Como se menciona en el libro Inteligencia Artificial de Ponce Gallegos y Torres Soto (2014), los sistemas inteligentes permiten identificar patrones complejos en grandes volúmenes de datos y transformarlos en conocimiento útil para la toma de decisiones, lo que presenta un valor significativo en contactos públicos (p. 79).

Una vez considerado todo lo anterior de igual manera se determina que el presente trabajo se encuentra alineado con el objetivo general de la Maestría en Gestión e Innovación Pública (MGIP), perteneciente a la Facultad de Contaduría y Administración de la Universidad Autónoma de Querétaro, que es formar profesionales con las competencias necesarias para desarrollar, ejecutar y financiar políticas públicas que respondan a los retos actuales de las instituciones y mejoren la calidad de vida de la población (UAQ, s.f.). Bajo esta orientación, el estudio plantea una propuesta de intervención pública con base tecnológica, la implementación de un modelo predictivo que permita identificar los estudiantes de tercer grado de secundaria que están en riesgo de reprobación.

El proyecto se alinea con las competencias de la MGIP, al emplear métodos cuantitativos modernos para analizar un problema público y realizar pronósticos sobre variables clave –como la reprobación escolar– a partir de un enfoque basado en evidencias. Asimismo, considera los principios de ética, transparencia y

eficiencia que el programa promueve como ejes de formación del servidor público (UAQ. s.f.)

Este trabajo se justifica desde diferentes ejes del programa, en primer lugar, atiende el eje de gestión e innovación al poner el uso de la inteligencia artificial para mejorar la eficiencia institucional. En segundo lugar, se vincula al eje profesionalizante, al identificar una problemática real –la reprobación escolar– y diseñar una solución innovadora que se puede escalar como política pública preventiva.

1.3 Pregunta de investigación

¿Es posible predecir con precisión qué estudiantes de tercer grado de secundaria tienen alto riesgo de reprobar, utilizando información académica e institucional de primero y segundo grado de secundaria, mediante técnicas de ciencia de datos e inteligencia artificial?

1.4 Objetivo

Proponer un modelo de ciencia de datos para predecir el riesgo de reprobación escolar en tercer año de secundaria en el estado de Querétaro, a partir del análisis de información histórica de estudiantes de primero y segundo año de secundaria correspondiente a 2017 y 2018, respectivamente.

1.4.1 Objetivos específicos:

O1: Conocer la institución y el problema de reprobación del nivel de secundaria general para que la propuesta de solución esté alineada a las estrategias y objetivos institucionales.

O2: Analizar la información que se utilizará en el desarrollo de la propuesta de atención del problema de reprobación en el nivel educativo de secundaria general, por mencionar, ciclo escolar, periodos de evaluación o nivel de marginación del centro de trabajo para que de los modelos que se utilicen se obtengan los resultados esperados.

O3: Preparar la información que se proporcione adecuándose a los requerimientos que demanda las técnicas de modelado de aprendizaje automático para que la generalización y sus métricas cumplan los requisitos esperados.

O4: Aplicar técnicas de aprendizaje automático o redes neuronales para identificar el modelo que proporcione el mejor desempeño en las métricas de evaluación y la mayor capacidad predictiva para detectar estudiantes en riesgo de reprobación.

O5: Evaluar el desempeño de las métricas de los modelos de aprendizaje automático o redes neuronales que se someterán a pruebas, para seleccionar el que tenga mejores características predictivas para detectar estudiantes de tercer grado de secundaria que estén en riesgo de reprobar.

O6: Implementar el modelo de aprendizaje automático que tenga mejor rendimiento en la evaluación de rendimiento de sus métricas, y proponer una estrategia de despliegue que considere desde la ingesta, hasta la entrega de información a las autoridades educativas, incluyendo la supervisión, monitoreo, mantenimiento y actualización continua del modelo.

1.5 Hipótesis:

Si se utiliza información académica e institucional de los alumnos de primero y segundo grado de secundaria, entonces es posible construir un modelo predictivo que identifique a los estudiantes de alto riesgo de reprobar el tercer grado de secundaria general, aún en un contexto de datos desbalanceados.

2. MARCO TEÓRICO

2.1 Antecedentes

Se revisó la tesis doctoral titulada Predicciones del fracaso y abandono escolar mediante técnicas de minería de datos, Marqués Vera (2015) propone una metodología basada en técnicas de clasificación y minería de datos para anticipar la reprobación y deserción escolar en el nivel medio superior en México, específicamente del programa II de la Unidad Académica Preparatoria de la Universidad de Zacatecas.

Presenta dos metodologías principales: una que utiliza algoritmos clásicos de clasificación y otra desarrollada por el autor, denominada Interpretable Classification Rule Mining (ICRM), basada en programación genética y gramáticas libres de contexto.

Muestra cómo el uso de técnicas de procesamiento y transformación de los datos como la selección de atributos, balanceo de datos con SMOTE y modelos sensibles al costo, permiten mejorar significativamente la identificación de los estudiantes en riesgo de abandonar la escuela.

Las variables más significativas en los modelos destacan las calificaciones en las materias clave (matemáticas, física e inglés), el número de hermanos, edad, hábitos como el consumo de alcohol y tabaco, y el nivel de motivación del estudiante.

La contribución más destacada es la propuesta de una metodología de predicción temprana, organizada en fases durante todo el semestre escolar. Esta estructura permite construir un sistema de alerta temprana (SIAT) que se puede activar en distintos momentos para anticipar el riesgo de abandono o reprobación.

El trabajo de Márquez Vera (2015) representa un antecedente relevante para la presente investigación, ya que demuestra que es posible aplicar técnicas avanzadas de minería de datos para realizar identificación temprano de estudiantes en riesgo de deserción en un ambiente real.

En la tesis doctoral Análisis del Sistema Educativo de la Universidad Técnica Estatal de Quevedo Mediante Ciencia de Datos, se emplean técnicas de analítica educativa para evaluar el desempeño institucional de una universidad pública ecuatoriana, utilizando fuentes de datos administrativas y académicas.

La importancia de este estudio radica en su capacidad de utilizar herramientas de minería de datos, visualizaciones y modelos de predicción con el fin de mejorar la toma de decisiones. Sus enfoques metodológicos y técnicos son aplicables a otros niveles educativos.

El autor adopta un enfoque estructurado basado en el análisis de información histórica de los estudiantes y de los procesos administrativos, lo cual permite identificar patrones relevantes que impactan en la eficiencia académica, la deserción y el rezago. Su propuesta se alinea con los objetivos del presente trabajo, al demostrar se pueden utilizar estas herramientas para la evaluación preventiva del riesgo académico.

(Presentación de la revisión de bibliografías relevantes tanto clásicas como actuales que tengan que ver con el problema específico que aborda la tesis.)

2.2 Criterios normativos de acreditación y reprobación en educación secundaria

La normatividad oficial para la educación básica en México establece lineamientos claros sobre los umbrales académicos, el procedimiento de acreditación y los requisitos para ser considerado como aprobado o reprobado en

un grado escolar. Estos criterios son fundamentales para definir la variable objetivo en la presente investigación.

Para que el modelo predictivo de estudiantes que estén en riesgo de reprobación en tercer grado de secundaria, sea válido desde un punto de vista institucional, es necesario definir con claridad las condiciones con las cuales un estudiante es considerado acreditado o reprobado. Estos criterios están establecidos en el documento oficial titulado Normas Específicas de Control Escolar Relativas a la Inscripción, Reinscripción, Acreditación, Promoción, Regularización y Certificación en la Educación Básica, Publicado por la Dirección de Acreditación, Incorporación y Revalidación de la SEP.

Se establece que la educación secundaria, la calificación para las asignaturas que conforman el componente curricular Campos de Formación Académica, así como las áreas de Artes y Educación Física, debe expresarse en una escala de 5 a 10, usando únicamente número enteros, donde la calificación de 5 se considera reprobatoria y las calificaciones de 6 a 10, aprobatorias (SEP, 2023, p. 28).

El promedio final se definió como “El resultado de sumar los promedios finales de las asignaturas y áreas, dividiendo entre el número total de asignaturas y áreas que se establezcan para cada grado” (SEP, 2023, p. 29).

Para educación secundaria, el promedio de nivel educativo se obtiene dividiendo la suma de los promedios finales de los tres grados entre tres.

Para acreditar tercer grado de secundaria, el documento señala como requisitos indispensables:

Tener un mínimo de 80% de la asistencia en el ciclo escolar.

Obtener un promedio final de 6 en cada una de las asignaturas (SEP, 2023 p. 30).

Específicamente indica que “El alumno de tercer grado será promovido al siguiente nivel educativo cuando haya acreditado en términos de la Norma 4,6,3 b)” (SEP, 2023, p. 31).

Con base en lo anterior se puede determinar que la variable dependiente del modelo, la variable objetivo o etiqueta, se construye bajo el criterio:

Se considera reprobado un estudiante de tercer grado de secundaria que no cumpla con el mínimo de 80% de asistencia y/o que no tenga una o más asignaturas con calificación de 5, es decir, no acreditadas conforme a la escala normativa.

Un estudiante de tercer grado de secundaria se considera aprobado cuando cumple ambos criterios, sin necesidad de repetir el grado ni acudir a regularización adicional.

Esta premisa garantiza que el modelo predictivo refleje fielmente los criterios institucionales definidos por la SEP y sea útil para su aplicación práctica en contextos escolares.

2.3 La inteligencia artificial

La Inteligencia Artificial (IA) es un campo de la informática que estudia la creación de programas o máquinas capaces de similares procesos inteligentes propios del ser humano, como el razonamiento, el aprendizaje, la planificación y la resolución de problemas. Este campo ha evolucionado desde enfoques deterministas hasta modelos adaptativos que permiten a las máquinas modificar su comportamiento en función de la experiencia (Ponce, Gallegos et al., p. 16)

La inteligencia artificial (IA) es una rama de la ciencia computacional que está enfocada en la creación de agentes, los cuales actúan de manera inteligente en diversos ambientes son capaces de tomar decisiones autónomas y adaptarse a

su entorno. De acuerdo con Poole y Mackworth (2023), la IA estudia la construcción y el análisis de agentes computacionales capaces de tomar decisiones adecuadas, aprender de la experiencia, adaptarse a nuevos entornos y considerar tanto consecuencias a corto como a largo plazo. Estos agentes son evaluados por su comportamiento externo: si actúan inteligentemente, son considerados inteligentes, independientemente de cómo estén estructurados internamente.

Un agente computacional es aquel cuya toma de decisiones puede explicarse en términos de operaciones básicas ejecutadas por un dispositivo físico. Esta perspectiva se alinea con el enfoque de la ingeniería, que busca construir sistemas útiles y eficientes, como asistentes virtuales, sistemas de recomendación, traductores automáticos, entre otros, que amplían la inteligencia y creatividad humana en tareas cotidianas (Poole y Mackworth, 2023).

El propósito científico central de la IA es comprender los principios que hacen posible el comportamiento inteligente, tanto en sistemas naturales como artificiales. Para lograrlo, se formulan y prueban hipótesis mediante el diseño de sistemas computacionales que simulan aspectos de la inteligencia. En este contexto, la IA puede entenderse como una forma de psicología sintética o filosofía experimental, dado que permite experimentar con modelos ejecutables de comportamiento inteligente.

Existen múltiples enfoques para lograr esto, siendo uno de los más tradicionales la IA simbólica o basada en reglas, también conocida como “IA clásica”. Este enfoque funciona a través de estructuras lógicas tipo SI... ENTONCES…, organizadas en sistemas expertos capaces de tomar decisiones, por ejemplo, en medicina o en procesos de manufactura. Aunque estos sistemas han sido útiles, su principal limitación es la dificultad para escalar y mantener el conocimiento codificado, ya que dependen de reglas estáticas que se vuelven complejas al interactuar entre sí.

En contraste, los avances recientes en inteligencia artificial han sido impulsados por el aprendizaje automático, una rama de la IA que permite que los algoritmos aprendan a partir de grandes volúmenes de datos, identificando patrones y haciendo predicciones sin necesidad de ser programados con reglas explícitas.

El aprendizaje automático se clasifica en tres tipos principales:

Supervisado (cuando los datos tienen etiquetas)

No supervisado (cuando los datos no están categorizados)

Por refuerzo (donde el sistema mejora continuamente a partir de la retroalimentación que recibe).

El aprendizaje profundo (deep learning), una parte clave del aprendizaje automático, emplea redes neuronales artificiales para modelar. Estas redes imitan la estructura del cerebro humano a través de capas de neuronas artificiales, permitiendo un procesamiento complejo y no lineal. Son la base de avances como el reconocimiento facial, los asistentes virtuales, la visión por computadora y los vehículos autónomos. Sin embargo, a pesar de su precisión, las redes neuronales profundas suelen ser poco transparentes, ya que no siempre se puede explicar cómo llegaron a una decisión, lo que genera desafíos éticos y técnicos en contextos que pueden llegar a ser sensibles por su naturaleza.

En la práctica, la mayoría de las aplicaciones modernas de IA combinan componentes de aprendizaje automático con reglas definidas por el factor humano. Esto se observa, por ejemplo, en los chatbots, donde parte de las respuestas sigue reglas preprogramadas. Además, la efectividad del aprendizaje automático depende en gran medida del trabajo humano, como la selección y el etiquetado de datos, tarea frecuentemente subcontratada a trabajadores en distintas partes del mundo.

En conjunto, estas técnicas han permitido desarrollar tecnologías aplicables a múltiples campos como el procesamiento del lenguaje natural, el reconocimiento de imágenes, la predicción de fraudes y hasta la generación de contenido creativo.

Aunque el aprendizaje automático es comúnmente asociado como sinónimo de IA, está en realidad representa sólo un fragmento más de un entorno en constante evolución (UNESCO, 2021, pp. 11–13).

En los entornos educativos representa una transformación significativa en los procesos de enseñanza aprendizaje. A través de algoritmos capaces de aprender de los datos y realizar inferencias complejas, la IA permite desarrollar modelos de enseñanza personalizados, fomentar la creatividad estudiantil y mejorar el monitoreo del rendimiento académico (Luna Rizo, Daza Ramírez & Lozoya Arandia, 2024). Ha evolucionado desde enfoques simbólicos y sistemas expertos hasta métodos basados en aprendizaje profundo (Deep Learning), lo que ha permitido su aplicación efectiva en áreas como la retroalimentación automatizada, la tutoría inteligente y la generación de contenido educativo (Luna Rizo, Daza Ramírez & Lozoya Arandia, 2024).

Sin embargo, la IA no solo es un campo técnico. La UNESCO (2021) advierte que su rápida expansión plantea desafíos éticos, sociales, culturales y económicos que requieren una gobernanza responsable. La IA influye en la forma en que interactuamos con el conocimiento, la verdad, la privacidad, la equidad y la toma de decisiones. Por ello, su desarrollo debe estar guiado por principios como la inclusión, la transparencia, la equidad, la rendición de cuentas y la sostenibilidad.

De hecho, la UNESCO (2021) plantea que la inteligencia artificial debe centrarse en el ser humano, promoviendo el desarrollo inclusivo y el respeto a los derechos humanos. Para lograrlo, propone la creación de marcos normativos nacionales, la evaluación de impacto ético de los sistemas de IA y el fortalecimiento de capacidades digitales en todos los niveles sociales.

2.4 Ciencia de datos

La ciencia de datos es un campo interdisciplinario pues combina conocimientos de estadística, matemáticas, programación y análisis avanzado con herramientas de inteligencia artificial y aprendizaje automático. Su principal propósito es encontrar patrones y con ellos generar conocimiento útil a partir de grandes volúmenes de datos, lo que permite facilitar la toma de decisiones al hacerlo de manera más informada y estratégica dependiendo del contexto.

El concepto “ciencia de datos” comenzó a utilizarse en los años 1960 como una alternativa al término estadística. Sin embargo, fue a finales de la década de 1990 cuando empezó a consolidarse como una disciplina autónoma, caracterizada por un enfoque más integral que abarca desde el diseño y la recolección de datos hasta su análisis y modelado. Su adopción más allá del entorno académico se produjo durante los años 2010, impulsada por el crecimiento del big data y los avances en inteligencia artificial (IBM, s.f.).

De acuerdo a Szymkowiak, 2019, el concepto de ciencia de datos emerge como una disciplina en la era digital cuyo propósito principal es generar conocimiento a partir de grandes volúmenes de datos. El enfoque ha evolucionado desde sus orígenes con el procesamiento de datos hasta ser un campo interdisciplinario que integra diferentes áreas, estadística, aprendizaje automático, gestión de datos, visualizaciones y algoritmos computacionales. En la figura 4 se muestra cómo se interrelacionan la inteligencia artificial, el aprendizaje automático y la ciencia de datos.

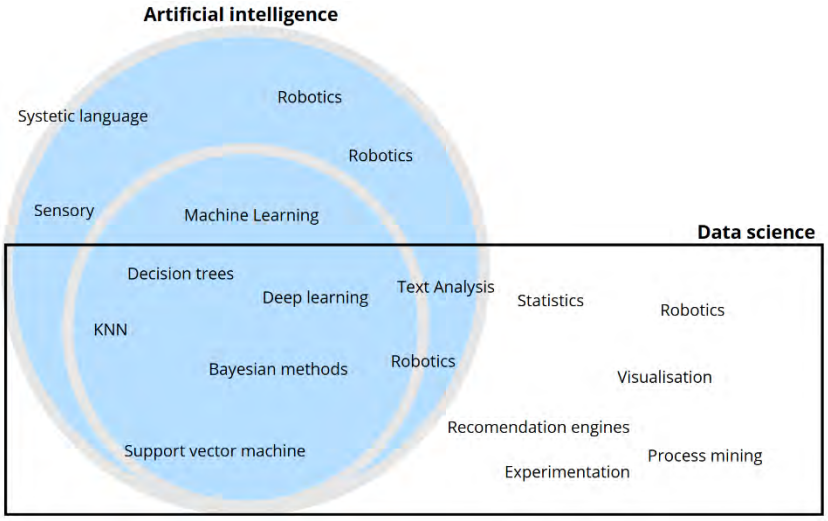


Figura 3
Interrelación entre Inteligencia artificial, aprendizaje automático y ciencia de datos
Nota. Elaborado con información de Szymkowiak, A. (2019). Recuperado de https://doi.org/10.12657/9788379862788

La ciencia de datos permite examinar la información desde diversas perspectivas, dependiendo del propósito del caso que se estudia. De acuerdo con AWS (s.f.), existen cuatro enfoques fundamentales para analizar datos y obtener valor de ellos:

Análisis descriptivo: Este busca responder a la pregunta ¿qué sucedió? Muestra una visión general de los datos históricos lo que ayuda a detectar patrones o tendencias. Para ello, se emplean herramientas de visualización de datos como gráficos de barras, líneas o tablas.

Diagnóstico: Este análisis busca explicar por qué sucedió algo. Se basa en indagar en los datos para encontrar relaciones, causas o elementos asociados a ciertos fenómenos. Incluye técnicas como el análisis exploratorio de los datos.

Predictivo: Este análisis intenta contestar ¿qué podría pasar en el futuro? Se basa en datos anteriores para construir modelos que permitan anticiparse a comportamientos o resultados. Utiliza métodos como proyecciones estadísticas y modelos predictivos.

Prescriptivo: El análisis prescriptivo va más allá al tratar de responder: ¿qué acción(es) tomar? No solo se predicen resultados, también realiza recomendaciones. Usa técnicas como, redes neuronales o sistemas de recomendación.

Cada uno de estos enfoques cumple un papel distinto en el apoyo al proceso de la toma de decisiones.

El análisis de datos ha transformado profundamente la forma en que operan las organizaciones. Hoy en día, tanto grandes como pequeñas empresas requieren una estrategia robusta de análisis de datos para fomentar su crecimiento y conservar su competitividad en el mercado.

El punto de partida de un proyecto de análisis de datos suele ser una necesidad o desafío empresarial específico. El profesional encargado del análisis colaborará estrechamente con los responsables del negocio para comprender a fondo el problema. Luego de identificar con claridad el objetivo, se pone en marcha un proceso con el que comienza el ciclo de vida de la ciencia de datos.

Según IBM (s.f.), usualmente un proyecto de análisis de datos sigue una serie de etapas bien definidas:

- Obtención de datos: Todo proyecto de ciencia de datos comienza con la recolección, los datos pueden proceder de distintas fuentes relevantes. Esta información se obtiene mediante distintos métodos como la introducción manual o captura, la extracción automática desde sitios web o la recopilación en tiempo real desde dispositivos y sistemas. Entre las fuentes comunes se encuentran bases de datos, registros de sistemas, archivos de audio y video, imágenes, dispositivos conectados (IoT) y plataformas sociales.
- Almacenamiento y procesamiento: Dado que los datos difieren en formato y estructura, la organización debe seleccionar el sistema de almacenamiento adecuado para cada tipo. En esta etapa se realiza la limpieza, transformación

y combinación de datos, normalmente mediante procesos de extracción, transformación y carga, u otras soluciones de integración.

- Exploración y análisis: En esta etapa se lleva a cabo un análisis exploratorio para identificar patrones, sesgos, y las distribuciones de los datos. También permite decidir si los datos son adecuados para modelos predictivos, machine learning o aprendizaje profundo. Si los modelos alcanzan niveles altos en sus métricas, los resultados obtenidos pueden servir como base para la toma de decisiones estratégicas dentro de la organización.
- Comunicación de resultados: Los hallazgos se traducen en informes y visualizaciones diseñadas para facilitar su interpretación. Para generar estos informes gráficos se pueden utilizar diferentes lenguajes de programación, aunque también se pueden usar herramientas especializadas en visualización de datos para presentar los resultados de manera más accesible y clara.

Figura 4
Etapas de un proyecto de datos
Nota. Elaborado con información de IBM (s.f.). Recuperado de https://www.ibm.com/mx-es/topics/data-science

De acuerdo a Szymkowiak (2019), el avance de herramientas analíticas, el acceso a grandes volúmenes de datos y el desarrollo computacional ha dado lugar al concepto de data–driven, en el que los datos no solo validan hipótesis, sino que generan nuevas teorías. Considera al investigador como un elemento principal en la interpretación y el uso de resultados y los modelos automatizados exploran las asociaciones complejas para descubrir nuevos patrones.

La ciencia de datos se ha convertido en un pilar clave para convertir grandes cantidades de información en conocimiento práctico. Esta disciplina permite analizar hechos pasados, comprender sus causas, anticipar escenarios futuros y sugerir acciones efectivas. Gracias a su naturaleza, se ha transformado en una herramienta importante para abordar desafíos complejos en múltiples áreas, incluyendo la educativa.

En el sector educativo, el análisis de datos adquiere una importancia particular al facilitar el estudio anticipado de problemáticas como la deserción o el bajo rendimiento académico. Mediante el uso de técnicas estadísticas y algoritmos de aprendizaje automático, es posible descubrir patrones en los datos históricos que permiten identificar con antelación a los alumnos que podrían estar en riesgo, favoreciendo así intervenciones más eficaces por parte de las instituciones. Más allá de proporcionar soluciones tecnológicas, la ciencia de datos promueve una cultura basada en la ética y la evidencia, contribuyendo a la creación de entornos educativos más inclusivos, eficaces y equitativos.

2.5 Metodologías de proyectos de ciencia de datos

Hoy en día el gran desafío que se tiene en la industria de la ciencia de datos es organizacional y metodológico, pero fundamentalmente en la gestión de los proyectos.

Existen diferentes metodologías específicas para gestionar los proyectos dependiendo del sector. En el desarrollo de proyectos de ciencia de datos se utilizan marcos como SEMMA, KDD o CRISP–DM. Estas metodologías son ampliamente reconocidas y son objeto de comparación en diferentes estudios. El análisis realizado por Santos y Azevedo (2005) concluye que KDD y CRISP–DM son metodologías más completas ya que incluyen etapas como la compresión del negocio y el despliegue del modelo, fases que no considera SEMMA.

Según DataScience–PM (s.f.), CRISP–DM es reconocida por ser metodológicamente más robusta, flexible, no está asociada a ninguna herramienta y se adapta fácilmente a cualquier sector.

Conforme a lo anterior se detalla la metodología CRISP–DM para usarla como base en este proyecto de propuesta de un modelo predictivo para un problema educativo.

CRISP–DM

La metodología CRISP–DM (Cross Industry Standard Process for Data Mining) es la metodología que ha sido aceptada como un estándar en el desarrollo de proyectos de ciencia de datos debido a su estructura modular ya que su enfoque trabaja en función a iteraciones, como se muestra en la figura 5. Debido a la flexibilidad en su diseño ha permitido que su uso se haya diversificado en una gama de entornos que van desde lo académico hasta lo comercial, su enfoque parte de lo básico, que es la alineación de objetivos con el entendimiento del negocio, pasa por el análisis de la información hasta la puesta en marcha de los productos haciendo esto un aprendizaje continuo de tal manera que se pueda reproducir. Consta de 6 fases:

1. Compresión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Despliegue

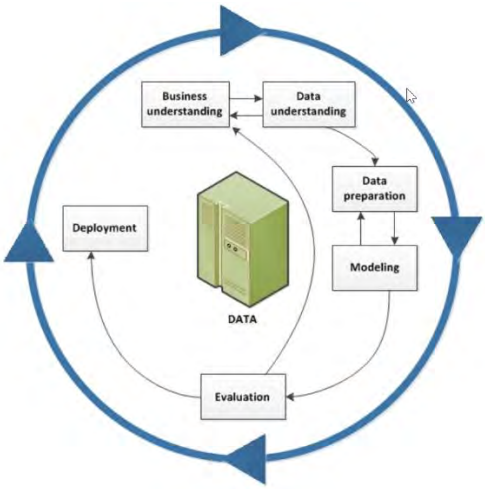


Figura 5
Ciclo de vida de una minería de datos
Nota. Adaptado de datos de IBM, 2015. Recuperado de https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

Cada una de ella se divide en tareas que generan productos definidos para que inicie la siguiente fase. Sigue el flujo natural como la hace cualquier proyecto y además se adapta al contexto del negocio. La forma en que está estructurada la metodología asegura que los proyectos se realicen de una forma eficiente y ordenada, además promueve la colaboración entre equipos, la mejora continua y el reutilizar procesos (IBM, 2015, pp. 1–3).

Fases teóricas de la investigación

Fase 1.

La primera fase de la metodología es la comprensión del negocio, enfatiza que antes de que se haga cualquier esfuerzo se debe entender claramente los objetivos del negocio. Es necesario que se involucre el mayor número de personas que están relacionadas con el proyecto y lo más importante, se deben de documentar los resultados, lo que se pretende es que estén alineados todos los

esfuerzos técnicos con las necesidades de la organización. Está dividido en 4 tareas principales

1. Determinación de los objetivos del negocio
2. Evaluación de la situación
3. Determinación de los objetivos de la minería de datos
4. Elaboración de un plan de proyecto

Esta fase es primordial para el éxito del proyecto, aquí es donde se establecen las bases para delinear el valor que se le dará al negocio. Fomenta la interlocución entre los responsables del proyecto y los responsables de la toma de decisiones (IBM, 2015, pp. 5–12).

Fase 2.

La segunda fase en la metodología es la comprensión de los datos, es donde el personal técnico se encarga de la recolección, familiarización y entendimiento de los datos. Se puede decir que es la labor más importante y la que más tiempo consume, el éxito del proyecto depende de que el trabajo que se haga en esta fase se haga de una manera correcta. Aquí es donde se van a identificar anomalías en los datos, se deben de identificar patrones y formular hipótesis. Esta fase tiene 4 tareas:

1. Recolección de datos iniciales
2. Descripción de los datos
3. Exploración de los datos
4. Verificación de la calidad de los datos

En esta fase se generará un diagnóstico inicial de la información, las siguientes fases dependen de la buena realización de las actividades. El

comprender y entender los datos da como consecuencia un tratamiento preciso, una selección adecuada de variables y lo más importante permite tomar decisiones informadas (IBM, 2015, pp. 13–17).

Fase 3.

La fase de preparación de los datos es una fase crítica, es donde se construye el set de datos final que se utilizará en la fase de modelado para que entren en los modelos. Aquí es donde se transforman los datos para darles utilidad, se eliminan a los que no se le pueda dar un tratamiento o no sean significativos para los modelos, se limpian, se le aplican técnicas para mejorar su interpretabilidad, se construyen o derivan nuevos datos o variables, se reestructuran tablas, se unen fuentes, orígenes de datos o bases de datos diferentes priorizando la integridad y consistencia. Esta fase consta de 5 tareas:

1. Selección de los datos
2. Limpieza de los datos
3. Construcción de los datos
4. Integración de los datos
5. Formato de los datos

Al igual que la fase anterior en esta se le tiene que invertir mucho tiempo y esfuerzo, aunque las actividades pueden ser rutinarias u operativas el desarrollarlas con calidad impacta en el buen diseño y desempeño de los modelos. Conforme al esquema de esta metodología esta etapa es iterativa con la fase de modelado, si se en el modelado se encuentra algunas áreas de oportunidad se tienen que regresar a esta fase para atenderlas y volver a generar los proceso, esto es un círculo virtuoso hasta llegar al mejor modelo (IBM, 2015, pp. 19–23).

Fase 4.

La fase de modelado es en la que se tiene el propósito de aplicar técnicas de análisis que permitan construir modelos predictivos a partir de los datos previamente preparados en la fase anterior. Para lograrlo se hace una selección previa de los modelos que se van a utilizar de acuerdo a los datos y a la naturaleza del problema, pueden ser de clasificación, regresión o segmentación. Una vez seleccionado el modelo se tienen que definir las métricas que se utilizarán para medir su rendimiento, pueden ser, recall, precision o F1–score por mencionar alguna, puede ser necesario la aplicación de búsquedas de hiperparámetros o técnicas de validación cruzada. Una vez que se tiene bien definida la estructura de pruebas y evaluación se procede al entrenamiento del modelo por medio de pruebas sucesivas, en donde se ajustaran parámetros y se optimizará su rendimiento hasta que cumpla con las expectativas del negocio. Esta fase tiene 4 tareas:

1. Selección de la técnica de modelado
2. Diseño del plan de prueba
3. Construcción del modelo
4. Evaluación del modelo

Esta fase es altamente iterativa con la fase anterior, requiere una colaboración entre los especialistas de los datos y el equipo del negocio. Es muy importante que se prueben diferentes algoritmos y elegir el que más se acerque a las métricas definidas. Los resultados se deben de interpretar más allá de los números, debe de ser claro para el equipo del negocio y debe ser útil para la toma de decisiones (IBM, 2015, pp. 25–30)

Fase 5.

Una vez que se tienen construidos y pre evaluados los modelos se inicia esta fase de evaluación que tiene como objetivo verificar que cumplan con los objetivos del negocio. Aunque los modelos hayan tenido un buen rendimiento técnico, se necesita hacer un análisis más amplio y considerar la relevancia, confiabilidad, aplicabilidad y la factibilidad de su implementación.

Se tiene que hacer una revisión cuidadosa de los resultados del modelo, se deben buscar posibles deficiencias que se vienen arrastrando de fases anteriores que impliquen ajustes técnicos o regresar a etapas anteriores.

Las métricas no se deben de interpretar de forma aislada, se deben de relacionar entre ellas y combinarla con los sucesos del negocio para que pueda tener un mejor aporte para el negocio. Es importante que se tenga una interacción con el equipo del negocio y asegurarse de que hayan recibido los resultados y que se hayan interpretado adecuadamente.

Una vez que se ha concluido con esta fase se determina si el modelo está listo para salir a producción o si se requiere realizar más iteraciones o un nuevo planteamiento. Esta fase tiene 3 tareas:

1. Evaluación de los resultados
2. Revisión del proceso
3. Determinación de los próximos pasos

Es muy importante que para tomar la decisión de selección del modelo se utilicen varias métricas y analicen diferentes gráficas para que la decisión sea lo mejor informada posible. Si los resultados no son los esperados es mejor hacer un replanteamiento y si el tiempo y los recursos lo permiten, regresar a las fases iniciales (IBM, 2015, pp. 31–33).

Fase 6.

La fase de despliegues se enfoca en trasladar los resultados de la fase de modelado hacia un entorno productivo dentro del negocio, no se limita a la instalación de la solución, debe incluir información útil para que el equipo del negocio pueda tomar decisiones. El despliegue puede adoptar múltiples formas, dependiendo del entorno en que se encuentre y puede ser desde la entrega de reportes periódicos hasta la integración en sistemas productivos. Lo más importante es que los resultados que se generen o se obtengan del modelo sean comprensibles para el público objetivo, de fácil acceso y ayuden a mejorar la situación del problema por el que fueron desarrollados.

Debe considerarse las herramientas que se utilizaran, los responsables de utilizarlos, la frecuencia de ejecución, los recursos tecnológicos y el soporte para su buen funcionamiento. También se debe incluir la forma de supervisión y el mantenimiento para asegurar que su funcionamiento sigue estable con los nuevos datos.

Esta fase tiene 4 tareas

1. Plan de despliegue
2. Plan de monitoreo y mantenimiento
3. Generación de informes finales
4. Revisión del proyecto

No se debe de asumir que el modelo será permanente ya que los datos, la información y el entorno cambia continuamente, es importante planear revisiones periódicas y sus respectivos ajustes (IBM, 2015, pp. 35–38).

2.6 Tipos de aprendizaje supervisado y aprendizaje profundo

El aprendizaje automático (machine learning) constituye una de las áreas fundamentales de la inteligencia artificial, enfocada en el diseño de algoritmos capaces de aprender patrones a partir de datos sin necesidad de ser explícitamente programados para cada tarea específica. A diferencia de los algoritmos tradicionales, el aprendizaje automático permite que los sistemas generen modelos predictivos que se adaptan dinámicamente conforme se incrementa la información observada, incrementando así su precisión en la toma de decisiones (Navlani, Fandango & Idris, 2021).

2.6.1 Aprendizaje supervisado

El aprendizaje supervisado aprende a partir de un conjunto de datos etiquetados, lo que les permite realizar tareas de clasificación regresión con base en ejemplos previos. Se tienen varios algoritmos en este tipo de aprendizaje que describiremos a continuación.

La regresión Logística es una técnica fundamental en el aprendizaje automático supervisado y en el análisis estadístico, utilizada principalmente para resolver problemas de clasificación binaria. A diferencia de la regresión lineal, cuyo objetivo es predecir una variable continua, la regresión logística estima la probabilidad de que una observación pertenece a una de las clases mutuamente excluyentes, transformando una combinación lineal de predictores mediante una función lógica o sigmoidea.

Matemáticamente, la función logística se expresa como:

$$f(x) = \frac{1}{1 + e^{-x}}$$

donde x es la combinación lineal de variables independientes y sus respectivos coeficientes, Esta transformación garantiza que la salida esté restringida al rango (0,1), lo cual permite que su interpretación como probabilidad.

El modelo se ajusta utilizando el método de máxima verosimilitud, ya que la función de pérdida utilizada no es cuadrática como en la regresión lineal, sino logarítmica. Este enfoque permite encontrar los parámetros que maximizan la probabilidad de observar los datos de entrenamiento bajo el modelo.

Entre las variantes de este algoritmo destacan la regresión logística binaria –aplicada a problemas de dos clases–, la regresión logística multinominal –cuando existen más de dos clases sin orden específico–, y la regresión logística ordinal –cuando las clases tienen un orden natural–. Cada una de estas variantes se adaptan a distintos contextos de clasificación, como diagnóstico médico, análisis de riesgo financiero o predicción de reprobación escolar, como es el caso de esta tesis.

Sus ventajas incluyen interoperabilidad, eficiencia computacional y facilidad de implementación en librerías como scikit–learn, lo cual la hace atractiva para escenarios donde la transparencia del modelo es crítica. Sin embargo, presenta limitaciones problemas con relaciones no lineales complejas o alta multicolinealidad entre predictores, lo que puede comprometer la generalización del modelo (Navlani, Fandango & Idris, 2021).

Random Forests, también conocido como Bosques Aleatorios, es un método de aprendizaje supervisado que pertenece a la familia de los ensambles de modelo y que se basa en la construcción de múltiples árboles de decisión para mejorar la precisión predictiva y la robustez del modelo. Fue introducido como una

extensión del método de bagging, con el objetivo de reducir la varianza y evitar el sobreajuste característico de los árboles de decisión individuales,

Su principio fundamental es la combinación de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios del conjunto de datos, tanto en términos de instancias como de características. En cada petición del árbol, se selecciona aleatoriamente un subconjunto de variables en lugar de considerar todas las características posibles. Este enfoque introduce diversidad en los árboles, lo que reduce la correlación entre ellos y mejora la capacidad de generalizar del modelo.

Durante la predicción, el algoritmo realiza un proceso de votación (en clasificación) o promedio (en regresión) entre todos los árboles que componen el bosque. Esta estrategia reduce el riesgo de que decisiones erróneas de árboles individuales afecten la predicción final, dando una gran estabilidad ante los datos ruidosos y una menor sensibilidad a las anomalías o outliers.

Una de las ventajas más notable es su capacidad para estimar la importancia relativa de las variables predictoras. Esto se logra al calcular la disminución media de la impureza (por ejemplo, entropía o gini) en cada nodo de decisión, lo cual proporciona una métrica útil para la selección de características en problemas complejos y de alta dimensionalidad (RaschKa & Mirjalili, 2017).

XGBoost, uno de los algoritmos más utilizados en problemas de clasificación y regresión por su eficacia y rendimiento es XGBoost (eXtreme Gradient Boosting), desarrollado por Chen & Guestrin (2016). Este modelo se basa en la técnica de gradient boosting, que cambia múltiples árboles de decisión secuencialmente, donde cada nuevo árbol intenta corregir los errores del anterior. Lo que distingue a XGBoost es su diseño optimizado para lograr mayor eficiencia computacional y escalabilidad sin sacrificar precisión.

El algoritmo incorpora una función objetivo regularizada que no solo minimiza el error, sino que también penaliza la complejidad del modelo, lo que contribuye a reducir el sobreajuste. Esta función combina el término de pérdida (por

ejemplo, log–loss) con términos de regularización L1 y L2, lo que proporciona un mayor control sobre la estructura del modelo (Chen & Guestrin, 2016).

Además, implementa una estrategia de crecimiento de árboles por profundidad (depth–first), con una evaluación heurística que selecciona los mejores puntos de división mediante un cálculo eficiente de gain. También utiliza técnicas como el procesamiento en bloques para reducir el uso de memoria y admite el entrenamiento distribuido en clústers, lo que lo hace apto para grandes volúmenes de datos.

En el contexto educativo, este tipo de modelo resulta especialmente valioso al permitir la detección anticipada de estudiantes en riesgo de reprobación, al identificar patrones ocultos en información académica e institucional. Su capacidad para trabajar con conjuntos de datos desbalanceados y su compatibilidad con métricas personalizadas lo convierte en una herramienta robusta y flexible para tareas de predicción orientadas a la toma de decisiones basadas en evidencia.

Las Redes Neuronales con Scikit–learn proporcionan una implementación sencilla y eficaz de redes neuronales multicapa mediante la clase MLPClassifier, la cual permite construir modelos de clasificación supervisada a partir de redes neuronales artificiales profundas (ANN). Aunque no están diseñadas para arquitecturas profundas como las que se implementan con TensorFlow y PyTorch, Scikit–learn permite construir modelos suficientemente robustos para tareas predictivas de complejidad moderada y es particularmente útil en contextos educativos y de prototipado rápido.

El modelo implementa una red neuronal de tipo feedforward, en la que las señales fluyen en una única dirección, desde la capa de entrada hasta la capta de salida, sin retroalimentación entre capas. El entrenamiento del modelo se lleva a cabo mediante el algoritmo de retropropagación (backpropagation), que calcula el gradiente del error y ajusta los pesos sinápticos utilizando un algoritmo de optimización, como descenso del gradiente estocástico (SGD), L–BFGS o Adam.

La clase permite configurar múltiples aspectos del modelo, como el número de capas ocultas, el tipo de función de activación (sigmoidea, ReLu o Tangente Hiperbólica), la función de pérdida (log_loss), y los parámetros de regularización (penalización L2) que ayudan a prevenir el sobreajuste.

Una característica relevante es la capacidad de integración con otras herramientas de Scikit–learn, lo que permite construir pipelines completos que incluyan preprocesamiento, reducción de dimensionalidad, selección de variables, validación cruzada. Esta modularidad y compatibilidad hacen de Scikit–learn una biblioteca ideal para quienes buscan evaluar modelos de redes neuronales sin la complejidad adicional de librerías especializadas en deep learning (RaschKa & Mirjalili, 2017).

LightGBM (Light Gradient Boosting Machine) es un modelo de aprendizaje automático basado en técnicas de boosting que utiliza árboles de decisión como clasificadores base. Es desarrollado por Microsoft, este algoritmo ha ganado amplia aceptación debido a su alta eficiencia computacional, su capacidad para manejar grandes volúmenes de datos y si rendimiento competitivo en tareas de clasificación y regresión,

Una de sus características más destacadas es su capacidad para procesar variables categóricas de forma nativa, sin necesidad de transformarlas previamente mediante codificaciones como one–hot–encoding, lo cual reduce el tiempo de procesamiento y preserva la estructura semántica de los datos. Además, implementa una estrategia de crecimiento de árboles por hojas (leaf–wise), que a diferencia del crecimiento nivel por nivel (level–wise), permite una mejor reducción del error en cada iteración, aunque puede requerir mecanismos adicionales de regularización para evitar el sobreajuste.

El modelo también ofrece soporte para técnicas de regularización L1 y L2, manejo eficiente de valores faltantes y operaciones de paralelización, lo cual lo convierte en una opción adecuada para proyectos con restricciones de tiempo o

procesamiento. En entornos con gran cantidad de variables y clases desbalanceadas –como ocurre en los sistemas educativos al predecir eventos poco frecuentes como la reprobación escolar– LightGBM ha demostrado ser eficaz al combinar precisión, escalabilidad y velocidad de entrenamiento (Microsoft, s.f.).

Por estas razones LightGBM ha sido seleccionado como unos de los algoritmos para el análisis predictivo del riesgo de reprobación de estudiantes de tercer grado de secundaria general, evaluando su desempeño frente a otros modelos.

2.6.2 Aprendizaje profundo

El aprendizaje profundo (deep learning) es una subdisciplina del aprendizaje automático que se basa en el uso de redes neuronales de múltiples capas para modelar representaciones jerárquicas y complejas de datos. Su principal fortaleza radica en la capacidad de estas redes profundas para aprender características automáticamente a partir de grandes volúmenes de datos sin necesidad de intervención manual significativa en la ingeniería de atributos.

Ha sido impulsado por tres factores principales: el aumento en la capacidad de cómputo (particularmente GPU), la disponibilidad de grandes conjuntos de datos etiquetados (big data), y los avances con algoritmos de entrenamiento como el descenso de gradiente estocástico con regularización, normalización por lotes y técnicas como Dropout. Entre las arquitecturas más utilizadas se encuentran las redes convolucionales (CNN), recurrentes (RNN), y sus variantes como LSTM, GRU y transformers (Goodfellow, Bengio & Courville, 2016).

Las redes neuronales artificiales (ANN) son un paradigma computacional basado en la estructura del sistema nervioso humano, que ha sido ampliamente adoptado en el campo del aprendizaje automático. Están compuestas por nodos

denominados neuronas artificiales, las cuales se agrupan en capas interconectadas: una capa de entrada, una o más capas ocultas, y una de salida. Cada conexión entre neuronas posee un pero que determina la importancia relativa de la señal transmitida, permitiendo a la red ajustar su comportamiento durante el proceso de entrenamiento.

En una red neuronal, cada neurona procesa la información aplicando una función de activación a una combinación lineal de sus entradas. Esta función introduce no linealidad al modelo, lo que habilita a la red para aprender representaciones complejas de los datos. Las funciones de activación más comunes incluyen la función logística (sigmoidea), la tangente hiperbólica y la ReLu (Rectified Linear Unit), cada una con propiedades matemáticas específicas que influyen en la convergencia y capacidad de aprendizaje de la red.

El entrenamiento de la ANN se realiza mediante el algoritmo de retropropagación (backpropagation), el cual ajusta los pesos de la red utilizando el método de descenso del gradiente. Este procedimiento consiste en calcular el error de la predicción de la red, propagarlo hacia atrás desde la salida hasta la entrada y actualizar los pesos de manera proporcional al gradiente de la función de pérdida. Este mecanismo iterativo busca minimizar el error global de la red a lo largo del conjunto de entrenamiento, permitiendo que el modelo generalice adecuadamente a nuevos datos.

Además de su capacidad para modelar relaciones no lineales, las ANN son altamente flexibles y adaptables, lo que les permite resolver una amplia variedad de tareas en clasificación, regresión, reconocimiento de patrones y series temporales. Una ventaja importante de este enfoque es su habilidad para aprender directamente de los datos sin requerir una estructura funcional explícita del problema, como ocurre con los modelos estadísticos tradicionales. Esto las convierte en herramientas idóneas para contextos con múltiples variables interrelacionadas o con datos no estructurados.

En términos de su implementación práctica, aunque las ANN pueden presentar desafíos como el sobre ajuste o la necesidad de sintonización cuidadosa de los hiperparámetros, estos pueden ser mitigados mediante estrategias como la regularización, el uso de funciones de activación adecuadas y técnicas de optimización avanzadas. Gracias a estas propiedades, las ANN han demostrado ser modelos robustos y versátiles en múltiples dominios, incluyendo el educativo, social, finanzas y procesamiento de lenguaje natural (RaschKa & Mirjalili, 2017).

2.7 Métricas

Se utilizan métricas para la evaluación del rendimiento de los modelos de clasificación supervisada binaria, se pueden evaluar desde distintas perspectivas. Cada una de ellas proporciona información cuantitativa sobre la capacidad predictiva del modelo, su comportamiento ante diferentes clases y su utilidad en la toma de decisiones, principalmente si existe un fuerte desbalance de clases.

Las métricas que se basan en información de la matriz de confusión se calculan a partir de los verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos. Estás métrica son útiles cuando se requiere evaluar a detalle el comportamiento del modelo sobre cada clase y son relevantes en escenarios con clases desproporcionadas (Chicco & Jurman, 2020)

La métrica exactitud (accuracy) es para evaluar el rendimiento global del modelo, pero se debe de tener cuidado cuando las clases están desbalanceadas ya que su resultado puede ser engañoso en la detección de la clase minoritaria (McHugh, 2012)

Las métricas que se basan en curvas como el área bajo la curva ROC (ROC AUC) y el área bajo la curva Precisión–Recall (PR AUC), evalúan el modelo en función de múltiples umbrales de decisión. PR AUC suele ser más informativa en

casos de escenarios con clases desbalanceadas, pues se focaliza en la clase minoritario o positiva (Davis & Goadrich, 2006; Scikit–learn, s. f.).

2.7.1 Matriz de confusión

De acuerdo a Fawcett, T. (2006), es una herramienta fundamental que se utiliza para evaluar el rendimiento de los modelos de clasificación supervisada. Permite visualizar de forma detallada los errores y los aciertos que comete el modelo en el proceso de predicción. Se compone de 4 elementos.

	Predicción positiva	**Predicción negativa**
Real positiva	Verdadero positivo (**TP**)	Falso negativo (**FN**)
Real negativa	Falso positivo (**FP**)	Verdadero negativo (**TN**)

TP (True Positive): Casos positivos clasificados como positivos.

TN (True Negatives): Casos positivos clasificados como negativos.

FP (False Positives): Casos negativos identificados erróneamente positivos.

FN (False Negatives): Casos positivos clasificados incorrectamente como negativos.

La matriz de confusión es una herramienta muy útil porque además de mostrar las tasas de acierto proporciona los tipos de errores en los que el modelo incurre. Su uso presta mayor relevancia en los casos donde se tienen clases desbalanceadas.

2.7.2. Precision

Según Davis y Goadrich (2006) la precisión es una métrica básica para los modelos de clasificación y principalmente en los que sus datos están desbalanceados. Es la proporción de casos identificados como positivos que realmente lo son. Su fórmula es:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

En donde TP son verdaderos positivos (realmente positivos) y FP son falsos positivos (casos negativos etiquetados como positivos)

2.7.3. Recall (Sensibilidad)

De acuerdo a Davis y Goadrich (2006) presentan el recall(verdaderos positivos o sensibilidad) como una importante en los problemas de clasificación binaria. Se representa como la proporción de casos positivos que son identificados correctamente por el modelo

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Donde:

- **TP**: Verdaderos positivos (casos positivos correctamente identificados).
- **FN**: Falsos negativos (casos positivos que el modelo no detectó).

Mide la capacidad del modelo para detectar los casos relevantes y de manera especial en escenarios en los que no detectar un caso positivo tienen graves consecuencias.

2.7.4. F1–score

El sitio Scikit-learn developers. (s. f.) define la métrica F1–score, también llamada F–measure, es identificada como la media armónica entre precisión(precision) y recall (sensibilidad), busca el balance entre ambas métricas.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Penaliza los valores desiguales de tal manera que será más alto si ambas métricas tienen valores altos.

2.7.5. La Curva ROC

La Curva ROC (Receiver Operating Characteristic) es una herramienta fundamental en la evaluación de los modelos de clasificación binaria. Muestra gráficamente la comparación de la Tasa de Verdaderos Positivos (TPR), conocida como recall, frente a la Tasa de Falsos Positivos.

Un modelo ideal se acerca al punto (0,1) que son 0% de falsos positivos y 100% de verdaderos positivos.

Se utiliza como una métrica (AUC–ROC) y se calcula como se muestra en la fórmula.

$$\text{AUC}_{ROC} = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2}$$

- **TPR (Recall)** = TP / (TP + FN)
- **FPR** = FP / (FP + TN)

El valor oscila entre 0.5 (modelo aleatorio) y 1.0 (modelo perfecto), entre más cercano esté el valor a 1.0, mejor será el desempeño del modelo para separar las clases (Fawcett, T., 2006).

2.7.6. Balanced–accuracy

Según la documentación oficial en el portal de Scikit–learn developers. (s. f.), Balanced–accuracy es “la medida de las tasas de recall obtenidas por cada clase”, esto permite evaluar el rendimiento modelo de una manera justa en entornos donde existe un desequilibrio de las clases.

Es muy útil cuando una clase es significativamente más frecuente que la otra, toma por igual el desempeño de cada clase conforme la fórmula.

$$\texttt{balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Donde:

TP: verdaderos positivos

FN: falsos negativos

TN: verdaderos negativos

FP: falsos positivos

Esta métrica busca identificar de una manera correcta la clase minoritaria sin que la clase mayoritaria oculte los errores.

2.7.7. MCC

Según Chicco, D., & Jurman, G. (2020), es una métrica de evaluación de modelos que mide la calidad de una clasificación binaria y considera todos los elementos de la matriz de confusión, TP, TN, FP y FN.

Considera que es una de las métricas más balanceadas y confiables cuando se usa en entornos de clases desbalanceadas.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

Se maneja entre los parámetros de 1(clasificación perfecta), 0(clasificación aleatoria) y –1(clasificación incorrecta)

No favorece ninguna de las clases, ofrece una única, simétrica y robusta.

2.7.8. Kappa

De acuerdo a McHugh, M. L. (2012) el coeficiente Kappa de Cohen (k) es una métrica estadística que mide el grado de acuerdo entre dos clasificadores. Es útil para evaluar la confiabilidad de predicciones, considerando tanto el acuerdo observado como el esperado.

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

< 0.00 (Sin acuerdo)

0.00 – 0.20 (Leve)

0.21 – 0.40 (Justo)

0.41 – 0.60 (Moderado)

0.61 – 0.80 (Sustancial)

0.81 – 1.00 (Casi perfecto)

El valor

k = 1 Acuerdo perfecto

k = 0 Acuerdo equivale al azar

k = -1 Desacuerdo sistemático (peor que el azar)

2.7.9. PR–AUC

Según Davis, J., & Goadrich, M. (2006) es una métrica que representa el área bajo la curva Precisión–Recall, evalúa el rendimiento de la clase positiva.

$$\text{PR AUC} = \sum_{i=1}^{n-1} (R_{i+1} - R_i) \cdot \frac{P_{i+1} + P_i}{2}$$

Donde:

- P_i = Precisión en el punto i
- R_i = Recall en el punto i
- Se usa una aproximación numérica (como la regla del trapecio) para estimar el área.

Interpretación de los valores.

PR: 1 (Modelo perfecto)

PR: 0.5 (Rendimiento equivale al azar (modelos balanceados)

PR: < 0.5 (Mal rendimiento)

Estos valores son para evaluar modelos que entrenan datos balanceados donde la línea de azar es 0.5.

Cuando se entrena con modelos datos desbalanceados la línea base se obtiene con:

$$\text{PR AUC}_{\text{base}} = \frac{\text{positivos}}{\text{positivos} + \text{negativos}}$$

Conforme a esta fórmula, cualquier modelo que obtenga un PR AUC por arriba de este umbral está mostrando capacidad real de predicción de casos críticos.

3. METODOLOGÍA

3.1 Metodología de la investigación

La presente investigación se enmarca dentro del enfoque cuantitativo, con un diseño no experimental, de tipo aplicado y alcance descriptivo y predictivo. Es cuantitativo ya que utiliza datos numéricos y recursos estadísticos para analizar el fenómeno de la reprobación escolar. Como señalan Sampieri, Collado y Lucio (2022), este enfoque se centra en la medición objetiva de los datos, en el análisis estadístico y la búsqueda de variables correlacionadas. A través de modelos matemáticos y técnicas de aprendizaje automático, se identifican patrones relevantes en la trayectoria académica de los estudiantes.

La investigación es no experimental pues trabaja con datos históricos sin manipulación de las variables. De acuerdo a Arias (2012), este diseño nos permite ver los hechos tal como ocurrieron en su entorno natural, sin intervención del investigador. Además, el estudio es de carácter aplicado, dado que persigue resolver un problema concreto del sistema educativo del estado utilizando herramientas de ciencia de datos. Tamayo y Tamayo (2005) afirman que la investigación aplicada pretende generar conocimiento con una utilidad y práctica inmediata y práctica.

El alcance es descriptivo, porque identifica y caracteriza los factores asociados a la reprobación. Es también predictivo, ya que se pretende por medio del modelo anticipar si un estudiante está o no en riesgo de reprobación. Hernández-Sampieri, Fernández y Baptista (2014) señalan que la descripción de fenómenos y la predicción de su comportamiento son fundamentales para generar acciones basadas en evidencias. Como lo menciona Fernández Naranjo (2016), este tipo de estudios fortalece la toma de decisiones institucionales, al traducir los datos en insumos estratégicos para intervención educativa.

(Indicar qué metodología se usará en la tesis: investigación aplicada, investigación documental, investigación de campo, investigación experimental, investigación no experimental, investigación exploratoria o investigación descriptiva)

3.2 Diseño de la investigación

El éxito de la implementación de los proyectos de Ciencia de Datos, aprendizaje automático o profundo depende en gran medida de la forma en que se realice todo el proceso. Para este caso utilizaremos la metodología CRISP–DM. Como se mencionó, es una de las más utilizadas y de acuerdo a su diseño, permite tener adecuaciones o modificaciones durante todo su ciclo de desarrollo.

3.2.1 Metodología para el objetivo específico 1.

3.2.1.1 Conocimiento del negocio (la institución) o el problema

De acuerdo al objetivo O1 y a la primera fase de la metodología CRISP–DM se iniciará esta fase, para obtener el conocimiento y entendimiento de la Institución y del problema.

En el apartado 1.2 Justificación de este documento se abordan elementos que también son pertinentes en esta fase. Con la finalidad de no redundar, en esta sección se tocarán de manera general los puntos relevantes.

De acuerdo al reglamento interno de USEBEQ el 7 de junio de 1992 se realiza la desconcentración de los servicios educativos y pasan de ser centralizados en la Secretaría de Educación Pública (SEP) a administrados por cada uno de los estados, En ese año se firma el decreto de creación de la Unidad de Servicios Para

la Educación Básica en el Estado de Querétaro (USEBEQ), a partir de ese momento, se encarga de gestionar los recursos de educación básica pública y la información escolar de educación básica pública y priva (USEBEQ, 2020).

El Programa Institucional 2021–2027 de la USEBEQ, define una visión centrada en el aprendizaje integral, con un enfoque humanista, orientado al desarrollo de habilidades cognitivas y digitales. Esta orientación promueve trayectorias escolares completas y busca posicionar al sistema educativo de Querétaro entre los primeros cinco lugares a nivel nacional (USEBEQ, 2021, pp. 13–15).

El proyecto se alinea con el objetivo 4 del programa institucional, el cual establece como prioridad asegurar trayectorias educativas completas con aprendizajes significativos y permanencia escolar. La reprobación en secundaria general, al constituir una causa directa de interrupción en la trayectoria académica de los estudiantes, se convierte en un problema clave para ser atendido desde la gestión pública. La propuesta de este modelo predictivo se configura como una herramienta que permite identificar de manera anticipada a los alumnos con mayor riesgo de reprobación, aportando información concreta para la toma de decisiones educativas con base en evidencia.

De igual manera se vincula con la estrategia 5.2 del mismo programa, que propone utilizar resultados de evaluaciones para detectar áreas de oportunidad e implementar acciones para eficientar los recursos y mejoren las capacidades de las autoridades educativas.

De esta forma, el conocimiento institucional recopilado en esta primera fase no solo proporciona el contexto organizacional y normativo, sino que valida la viabilidad de incorporar herramientas tecnológicas basadas en inteligencia artificial al quehacer institucional.

Como parte del análisis contextual de la institución se consideró el diagnóstico organizacional contenido en el Manual de Calidad de la USEBEQ, el

cual incluye un análisis FODA. En este análisis se identifican factores internos y externos que influyen en el desempeño del sistema educativo estatal (USEBEQ, pp 24–26). Este instrumento permite conocer con mayor claridad el marco operativo, los recursos disponibles y las áreas de oportunidad que inciden directa o indirectamente en la ejecución de proyectos estratégicos.

Este diagnóstico institucional sustente la viabilidad y pertinencia del modelo predictivo propuesto, al identificar un entorno organizacional con capacidades técnicas, humanas y estructurales que, a pesar de ciertas limitaciones, ofrece las condiciones mínimas necesarias para implementar soluciones innovadoras centradas en la mejora del desempeño educativo y una toma de decisiones preventiva.

De acuerdo al modelo CANVAS que se muestra en la justificación, la propuesta de valor institucional se centra en ofrecer una educación de calidad, gratuita y organizada, con un enfoque centrado en la escuela que promueva la inclusión, la equidad y el desarrollo integral de las niñas, niños y adolescentes. Esta misión se materializa a través de estrategias como la capacitación del personal docente y administrativo, la automatización de procesos, el fortalecimiento de los canales de comunicación y la atención directa a las comunidades escolares.

Dentro de los socios claves se identifican entidades gubernamentales en los niveles estatal y federal, madres y padres de familia, personal docente, universidades, organizaciones no gubernamentales (ONGs) y organismos sindicales. Cumplen un papel fundamental en el sostenimiento, fortalecimiento y evaluación de los servicios educativos.

Los servicios van dirigidos a estudiantes, familias, docentes, personal administrativo, instituciones educativas públicas como privadas y dependencias gubernamentales vinculadas al sector educativo.

Definen los recursos clave que permiten la operación institucional: infraestructura educativa, capital humano, tecnologías, materiales y financiamiento

público. Los canales de atención se articulan a través de plataformas digitales, oficinas físicas y redes sociales, buscando garantizar la accesibilidad y cercanía con los distintos públicos.

Las fuentes principales de ingresos son por parte del presupuesto del gobierno estatal y federal. La estructura de costos contempla principalmente el gasto de nómina, mantenimiento y expansión de infraestructura, incorporación de tecnologías y programas sociales orientados al bienestar estudiantil.

También se pudo ver en la justificación se pudo ver el comportamiento del indicador de reprobación en secundaria y se describe el posible rebote post pandemia que pone en evidencia el rezago académico acumulado y el riesgo de que la reprobación continúe desencadenando efectos negativos en otros indicadores.

Los datos que se tienen en el Programa Institucional USEBEQ 2021–2027 –referenciados en la justificación– ilustran claramente que el abandono escolar supera al de la media estatal, Querétaro fue de 3.2% y la media nacional 2.7% para el ciclo 2019–2020. Para ese mismo ciclo la eficiencia terminal en secundaria en el estado fue de 88.0% por debajo de la media nacional. En el rezago educativo Querétaro no supera a la media, pero tiene un porcentaje nada halagador, 23.8 contra 30.4 de la media nacional, es alarmante que 1 de cada 4 personas están en esa situación.

En ese mismo Programa Institucional también se menciona que 10 de los 18 municipios presentan niveles de rezago educativos superior al 40%, este valor es alarmante, muestra una desigualdad clara entre la sociedad, lo que acentúa la urgencia de atender las causas estructurales y escolares de dicho fenómeno.

Ante esta situación se justifica que se creen políticas públicas que busquen disminuir este fenómeno, una de ellas podría ser el crear modelos de prevención anticipada, que permitan detectar estudiantes que estén en esa situación de vulnerabilidad. Aunque la reprobación igualmente puede ser consecuencia de

múltiples factores (sociales, económicos, personales), contar con una herramienta que ayude a identificar casos en riesgo antes de que inicie el ciclo escolar, permite diseñar estrategias de atención dirigidas, intervenciones pedagógicas específicas y una gestión de recursos más eficiente.

Por tal motivo se hace una propuesta de modelo predictivo de reprobación escolar CANVAS ML, se muestra en la figura 6, el cuál traduce los principios de la organización en un marco de trabajo técnico alineado con los objetivos de la institución. La propuesta se centra en desarrollar un modelo de clasificación binaria que, usando datos históricos de primero y segundo grado de secundaria general, permita identificar de manera anticipada a los estudiantes que están en riesgo de reprobar tercer grado de secundaria general previa al inicio del ciclo escolar.

El modelo se alimenta de información institucional y académica, inasistencias, calificaciones, turno, nivel de marginación del lugar en donde está la escuela y categoría de la comunidad, se cargará la información de ese ciclo cada que finalice el ciclo. Las predicciones se realizan cada que se termina segundo grado de secundaria y se les proporcionará a las autoridades educativas para que estas antes de que inicie el ciclo escolar de tercer grado de secundaria ya tengan elaboradas las estrategias para atender a los alumnos que hayan sido identificados.

La propuesta contempla la fase de validación, entrenamiento con datos históricos y monitoreo en vivo. Se da prioridad a métricas como Recall, F1–score, PR AUC y MCC, con el fin de minimizar los falsos negativos y no dejar fuera a estudiantes que están en riesgo de reprobación.

Con esta propuesta la institución no solo contará con sistema de alertas temprana, sino que contará con una herramienta robusta que los apoye a la toma de decisiones, asignación de recursos y planificación de intervenciones. Esta propuesta se alinea a los objetivos del Programa Institucional, específicamente los que se refieren con mejorar la eficiencia del sistema educativo, prevenir el rezago escolar, y consolidar una administración orientada a resultados.

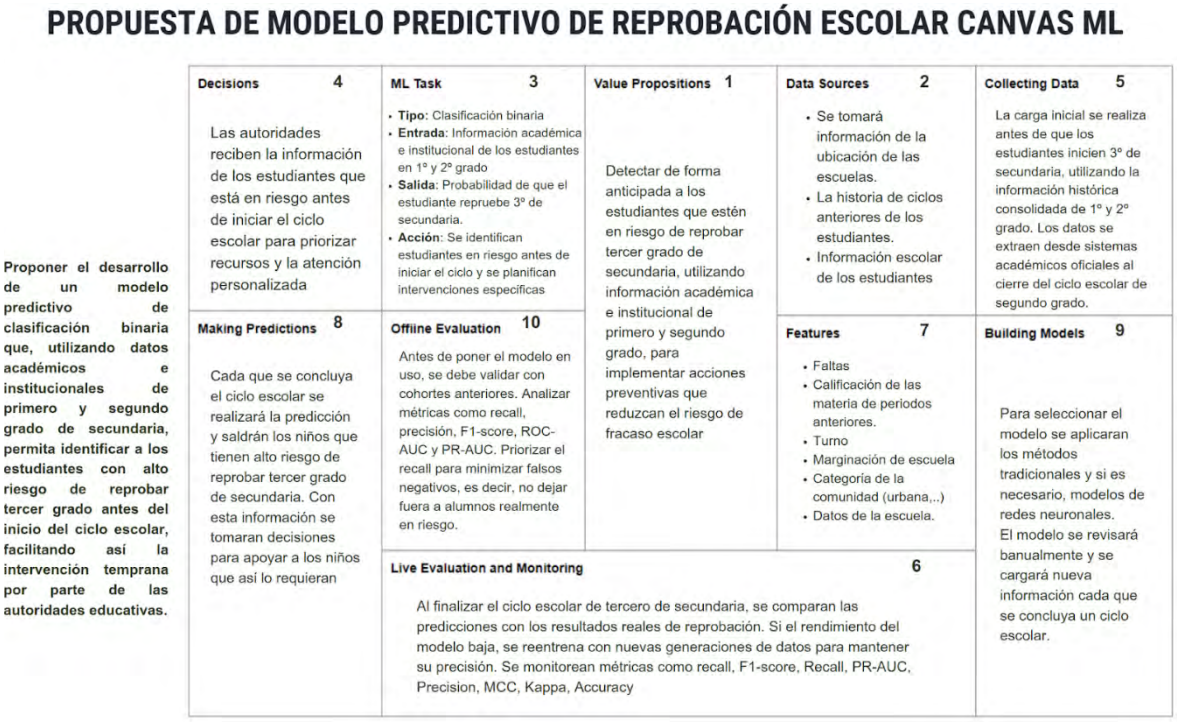


Figura 6

Propuesta de modelo predictivo de reprobación escolar CANVAS ML

Nota. Elaboración propia

La propuesta se planea se realice de acuerdo a un plan general en donde se describen las 6 fases de la metodología CRISP–DM. Se muestran los tiempos estimados en los que se realizarán las actividades en semanas. También, se muestra la columna de recursos que describe las personas y/o los equipos de trabajo que se requerirán para cubrir la fase. En la parte de riesgo igualmente se hace un listado general de lo que se prevé puede ocasionar retrasos o contratiempos en el desarrollo de las tareas de cada fase.

Fase	Tiempo	Recursos	Riesgo
Conocimiento del negocio	2 semanas	Analistas Equipo directivo Especialistas en educación	Personal no conosca el sector educativo Levantamiento de información incorrecta Equipo directivo no disponible Comunicación deficiente No tener acceso a los suficientes recursos
Comprensión de los datos	4 semanas	Científicos de datos Ingenieros de datos Analistas Especialistas en educación Equipo directivo	Datos con anomalias Interpretación inadecuada Comunicación deficiente Recursos insuficientes
Preparación de los datos	6 semanas	Científicos de datos Ingenieros de datos Especialistas en educación	Datos con anomalias que no se puedan tranformar Interpretación inadecuada de información Comunicación deficiente Recursos insuficientes
Modelado	3 semanas	Científicos de datos Especialistas en educación	Interpretación inadecuada de información Comunicación deficiente Recursos insuficientes Personal sin experiencia
Evaluación	2 semanas	Ingeniero en ML Científicos de datos Especialistas en educación	Interpretación inadecuada de información Comunicación deficiente Recursos insuficientes Personal sin experiencia
Despliegue	2 semanas	Arquitecto de soluciones Ingeniero en ML Especialistas en educación	Interpretación inadecuada de información Comunicación deficiente Recursos insuficientes Personal sin experiencia

Tabla 6
Fases de desarrollo de propuesta de modelo predictivo de reprobación escolar de tercer grado de secundaria general
Nota. Elaboración propia

En las diferentes fases la comunicación se encuentra como riesgo ya que se considera un punto neurálgico para el éxito del proyecto.

3.2.2 Metodología para el objetivo específico 2

3.2.2.1 Conocimiento de la información:

El objetivo O2 menciona que se debe analizar la información para conocerla y poder tomar acciones en las siguientes fases. Para el inicio del proceso, la información se proporcionará debidamente organizada en un archivo en formato CSV.

La información que contiene este archivo debe estar conformada por los datos de los estudiantes que cursaron en 2019 el tercer año de secundaria general,

así como, la información de estos cuando cursaron segundo año de secundaria en 2018 y primero de secundaria en 2017, para que se pueda hacer el análisis adecuadamente. Como se comentó el análisis se realizará únicamente con datos del nivel de secundaria general, los datos que no cumplan con estos criterios serán omitidos.

La información no debe estar personalizada, es decir, cada registro no debe identificar la identidad de a quién pertenece, por lo tanto, debe de omitir los datos sensibles que puedan poner en riesgo la identificación de cada estudiante.

El archivo está conformado por 19,159 registros con 61 columnas o variables, a continuación, se describen.

La data está agrupada en 5 segmentos conforme a la información que está asociada. Alumnos (9), son datos que pertenecen a los estudiantes, como: el ciclo en el que estudió, el grado o el grupo. Desempeño (25), define lo referente al aprovechamiento y principalmente son variables de las materias. Escuela (12), son datos que se asocian a la escuela como su ubicación geográfica, el turno o la zona. Comunidad de la escuela (2), es la clasificación al lugar donde se ubica, como son el nivel de marginación o si es rural o urbana. Estadística (13), es la numeralia básica que se asocia a las materias y al estudiante, como es la variable que define si estuvo en una escuela pública o privada o cuantos periodos reprobó en primero de secundaria.

Las variables tienen 3 formatos o tipo de dato, int64 con 46 variables, float64 con 6 variables y objet (objeto) con 9 variables, en total son 61.

Los nombres de algunas variables tienen un sufijo que indica el grado al que pertenece esa información: ANTE, corresponde al primer grado de secundaria, PASA, corresponde al segundo grado de secundaria.

Dado el número de variables se dividirá la tabla en dos partes, en cada una se explica brevemente el esquema de cada columna.

Variable	Descripción	Referencia de Información	Tipo de Dato	Posibles Valores
Alumno	Identificador del alumno	Alumno	int64	Entero
rado	Grado en el que cursa el estudiante en ese ciclo escolar	Alumno	int64	3
rupo	Grupo al que pertenece el estudiante en ese grado y ciclo escolar estudiante	Alumno	object	Letras
exo	Genero del estudiante	Alumno	object	H, M
exold	Clave del genero del estudiante	Alumno	int64	0 = M (mujer), 1 = H (hombre)
echaNacimiento	Fecha de Nacimiento del estudiante	Alumno	object	fecha
dad	Edad del estudiante	Alumno	int64	13 a 20
LUMNOLAT	Latitud de ubicación donde vive el estudiante	Alumno	float64	Decimal
LUMNOLON	Longitud de ubicación donde vive el estudiante	Alumno	float64	Decimal
eproboGrado1	Reprobó Primer Grado 0=No, 1= Sí	Desempeño	int64	0=No, 1= Sí
eproboGrado2	Reprobó Segundo Grado 0=No, 1= Sí	Desempeño	int64	0=No, 1= Sí
eproboGrado3	Reprobó Tercer Grado 0=No, 1= Sí	Desempeño	int64	0=No, 1= Sí
altas_ANTE	Faltas de primer grado de secundaria	Desempeño	int64	Entero
altas_PASA	Faltas de segundo grado de secundaria	Desempeño	int64	Entero
ISTOR_ANTE	Calificación de primer grado de secundaria de HISTORIA	Desempeño	int64	5 a 10
ORMACIANTE	Calificación de primer grado de secundaria de FORMACIÓN CÍVICA Y ÉTICA	Desempeño	int64	5 a 10
SPANO_ANTE	Calificación de primer grado de secundaria de LENGUA MATERNA (ESPAÑOL)	Desempeño	int64	5 a 10
ATEMATANTE	Calificación de primer grado de secundaria de MATEMÁTICAS	Desempeño	int64	5 a 10
IENCIAANTE	Calificación de primer grado de secundaria de CIENCIAS (BIOLOGÍA)	Desempeño	int64	5 a 10
EOGRAFANTE	Calificación de primer grado de secundaria de ARTES	Desempeño	int64	5 a 10
NGLES_ANTE	Calificación de primer grado de secundaria de LENGUA EXTRANJERA (INGLÉS)	Desempeño	int64	5 a 10
SPANIIPASA	Calificación de segundo grado de secundaria de ESPAÑOL II	Desempeño	int64	5 a 10
ATEMIIPASA	Calificación de segundo grado de secundaria de MATEMÁTICAS II	Desempeño	int64	5 a 10
IENCIIPASA	Calificación de segundo grado de secundaria de CIENCIAS II (ÉNFASIS EN FÍSICA)	Desempeño	int64	5 a 10
ISTORIPASA	Calificación de segundo grado de secundaria de HISTORIA	Desempeño	int64	5 a 10
ORMAIIPASA	Calificación de segundo grado de secundaria de FORMACIÓN CÍVICA Y ÉTICA I	Desempeño	int64	5 a 10
NGLEIIPASA	Calificación de segundo grado de secundaria de SEGUNDA LENGUA(INGLÉS)	Desempeño	int64	5 a 10
DUCACIPASA	Calificación de segundo grado de secundaria de EDUCACIÓN FÍSICA	Desempeño	int64	5 a 10
RTES__PASA	Calificación de segundo grado de secundaria de ARTES	Desempeño	int64	5 a 10
MBIT__PASA	Calificación de segundo grado de secundaria de ÁMBITO O CLUB	Desempeño	int64	5 a 10
RADO1	Cursó el primero de secundaria 0=Sí, 1= No	Desempeño	int64	0=Sí, 1= No
RADO2	Cursó segundo de secundaria 0=Sí, 1= No	Desempeño	int64	0=Sí, 1= No
RADO3	Cursó tercero de secundaria 0=Sí, 1= No	Desempeño	int64	0=Sí, 1= No
OTALGRADOS	Total de grados cursados	Desempeño	int64	1 a 3

Tabla 7a

Diccionario de variables del conjunto de datos: características del alumno y desempeño académico

Nota. Elaboración propia con base en datos institucionales de USEBEQ (2025).

De acuerdo a la información analizaremos las características de las principales variables. Para ver su comportamiento utilizaremos gráficas de distribución de densidad por clase (Kernel Density Estimation, KDE). Este tipo de gráfica además de identificar los valores que tiene cada variable se pueden ver la densidad estimada, en donde se agrupan y cómo se dispersan los valores de cada clase, en este caso la variable ReproboGrado3 (etiqueta, objetivo o target) que al ser un problema de clasificación binaria solo tiene los valores 0 y 1 para cada una de ellas. Para nuestro tema, en las gráficas siguientes se mostrarán en el eje de las

X los valores que se tienen en las variables y en el eje de las Y la densidad estimada de concentración de los valores, también se visualizará una etiqueta en la parte superior izquierda o derecha donde se muestra el nombre de la variable objetivo y el color que se tiene para cada valor, en este caso no reprobó color azul, clase 0 y reprobó color naranja, clase 1.

Variable	Descripción	Referencia de Información	Tipo de Dato	Posibles Valores
IdClavecct	Id del Centro de trabajo (CT), que es la escuela	Escuela	int64	Entero
Clavecct	Clave del CT, o clave de la escuela	Escuela	object	Clave
Nombrect	Nombre del CT, nombre de escuela	Escuela	object	Nombre
Turno	Turno del CT, turno de la escuela	Escuela	object	MAT, VES, DIS
TurnoId	Clave del turno del CT	Escuela	int64	0 = MAT (matutino), 1 = VES (vespertino), 2 = DIS (discontinuo)
Municipio	Nombre del Municipio en el que se encuentra la escuela	Escuela	object	Nombre
IdMunicipio	Id del Municipio	Escuela	int64	1 al 18
Region	Región en donde se encuentra la escuela	Escuela	int64	1 al 4
Sector	Sector al que la escuela	Escuela	int64	Entero
Zona	Zona a la que pertenece la escuela	Escuela	int64	Entero
ESCUELALAT	Latitud de ubicación de la escuela	Escuela	float64	Decimal
ESCUELALON	Longitud de ubicación de la escuela	Escuela	float64	Decimal
CATEGORIA	Categoría que defina la CONAPO a la comunidad en donde está ubicada la escuela	Comunidad de la escuela	object	URBANA, RURAL Y S/D CONAPO.
MARGINACION	Nivel de marginación en la que se encuentra la ubicación de la escuela conforma a la CONAPO.	Comunidad de la escuela	object	Nombre
escPriv_ANTE	Estuvo en escuela privada	Estadística	int64	Entero
escPub_ANTE	Estuvo en escuela pública	Estadística	int64	Entero
numescEstuvo_ANTE	Número de escuelas en las que estuvo	Estadística	int64	Entero
bimRepr_ANTE	Número de periodos con calificación menor a 6 en primer grado de secundaria	Estadística	int64	Entero
bimSinCali_ANTE	Número de periodos que no tiene calificación en primero de secundaria	Estadística	int64	Entero
bim6_8_ANTE	Número de periodos con calificación mayor igual a 6 y menor que 8 en primero de secundaria	Estadística	int64	Entero
bim8_10_ANTE	Número de periodos con calificación mayor igual a 8 en primero de secundaria	Estadística	int64	Entero
bimRepr_PASA	Número de periodos con calificación menor que 6 en segundo de secundaria	Estadística	int64	Entero
bimSinCali_PASA	Número de periodos sin calificación en segundo de secundaria	Estadística	int64	Entero
bim6_8_PASA	Número de periodos con calificación mayor o igual a 6 y menor a 8 en segundo de secundaria	Estadística	int64	Entero
bim8_10_PASA	Número de periodos con calificación mayor o igual a 8 en segundo de secundaria	Estadística	int64	Entero
PROMEDIO_ANTE	Promedio de las materias de primer grado de secundaria general	Estadística	float64	Decimal
PROMEDIO_PASA	Promedio de las materias de segundo grado de secundaria general	Estadística	float64	Decimal

Tabla 7b

Diccionario de variables del conjunto de datos: características escolares, geográficas y estadísticas

Nota. Continuación de la Tabla 5.a. Elaboración propia con base en datos institucionales de USEBEQ (2025).

La variable HISTOR_ANTE representa la materia de historia de primer año de secundaria, en ella se tienen las calificaciones que los alumnos obtuvieron en ese grado. En la figura 7 se puede apreciar que existen valores en 0 y sus valores alcanzan un poco más del 0.05 para ambas clases (no reprobó 0, reprobó 1). Se puede ver que la densidad estimada de los alumnos que reprueban tercer año (clase

1) en esta materia alcanza el valor de casi el 0.40 en las calificaciones entre 5 y 7, los que no reprueban, tienen una mayor concentración entre 7 y 9, alcanzan una densidad estimada un poco mayor al 25% y los valores se dispersan entre 6 y 10, como se puede ver, la curva es más ancha en entre estos valores. También, existe una superposición entre ambas curvas, pero hay una diferencia suficiente entre sus valores para considerarla útil en el modelo.

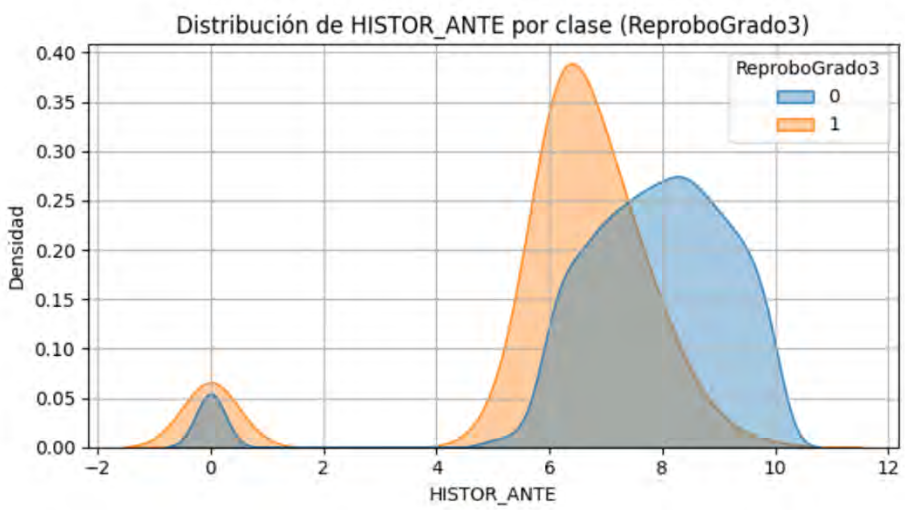


Figura 7
Distribución de la variable HISTOR_ANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable FORMACIANTE representa las calificaciones de los alumnos en la materia de formación cívica y ética de primer grado de secundaria.

Como se puede la en la figura 8 la gráfica de la materia de HISTOR_ANTE son similares y tienen los mismos valores en la calificación de 0. Las pequeñas diferencias están en la densidad estimada que tienen la clase 1 (reprobó tercero) que sobrepasa 0.40 lo que se podría suponer que los alumnos que reprueban tercero tienen calificaciones más bajas en esta materia. Se ve que la curva de los que no reprueba se tiende un poco más a la izquierda lo que podría suponer lo que se menciona respecto a la clase 1 (reprobados en tercero). Hay una separación

entre las curvas por lo que podemos decir que esta variable puede ser útil para el modelo.

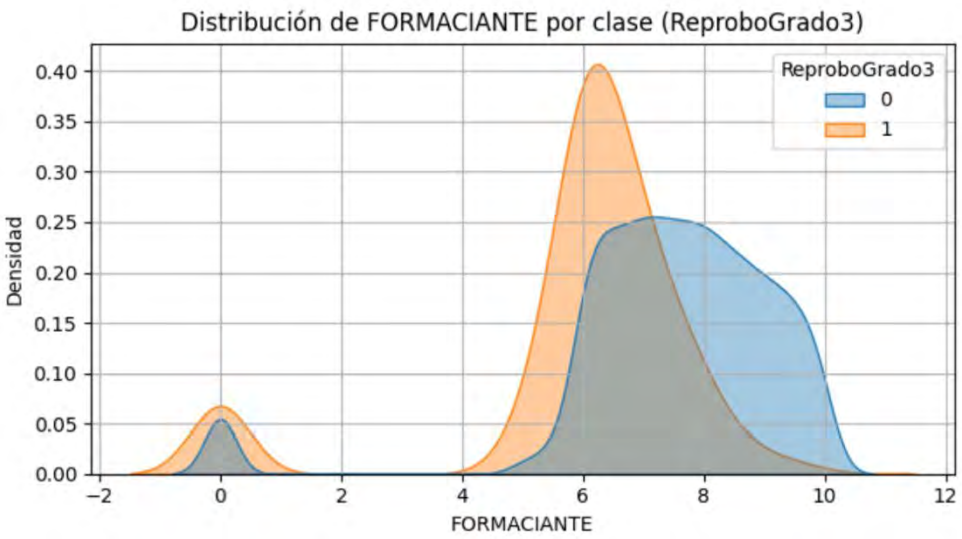


Figura 8
Distribución de la variable FORMACIANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

Las variables ESPANO_ANTE que refiere a las calificaciones de los alumnos en español de primer grado y la variable MATEMATANTE que se refiere a las calificaciones de los alumnos en matemáticas de primer grado son similares a las dos gráficas anteriores por lo que las omitiremos.

La variable CIENCIAANTE representa las calificaciones de los alumnos de primer grado de la materia de ciencias. Como se muestra en la figura 9 al igual que en las gráficas anteriores se presentan valores en 0 con los mismos valores de densidad estimada. Para la clase 0 (no reprobó) la curva se ve alargada y delgada, toma valores de calificación entre 9 y 10 y de densidad es más de 0.4. La clase 1 (repruebó) tiene calificaciones entre 7 y 9 y el valor de la densidad alcanza el 0.3. Se puede decir que los alumnos que reprueban tercer año tiene aceptables calificaciones en esta materia, Las curvas muestran una diferencia lo que indica que esta variable puede ser candidata para utilizarse en el modelo.

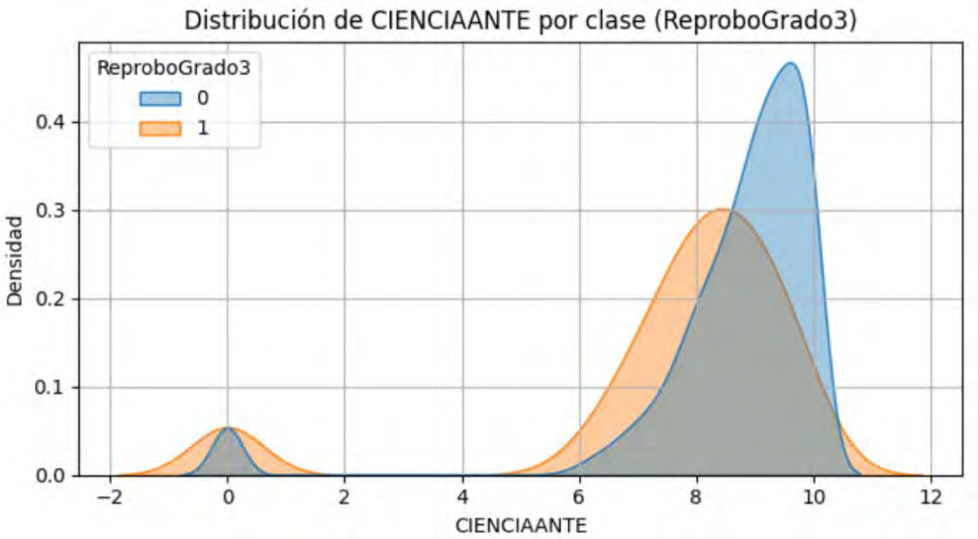


Figura 9
Distribución de la variable CIENCIAANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable GEOGRAFANTE representa las calificaciones de los alumnos que cursaron primer grado en la materia de geografía. La gráfica de esta variable es similar a la de ciencias por lo que no se mostrará.

La variable INGLES_ANTE representa las calificaciones de los alumnos que cursaron primer grado en la materia de inglés. Como se muestra en la figura 10 tienen valores en 0 y la densidad es la misma para las otras gráficas. Se puede ver que ambas curvas tienen alturas similares cercana a 0.30, para la clase 0(no reprobó) las calificaciones de mayor concentración están entre 8 y 9. Para los que reprobaron tercer grado (clase 1) está entre 6.5 y 7.5. Se puede decir que los alumnos que tienen calificaciones bajas en inglés tienden más a reprobar tercer grado. La separación de las curvas muestra que está variable puede ser útil para el modelo.

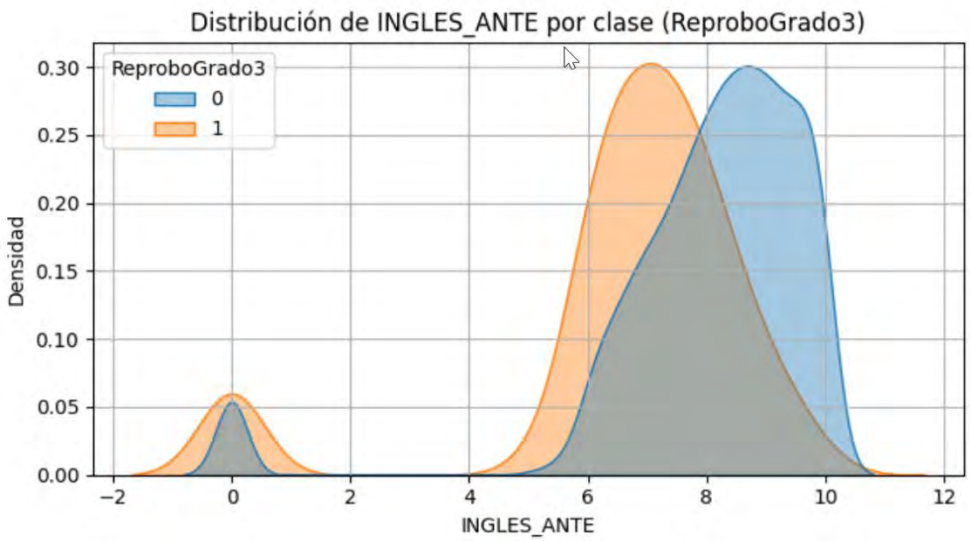


Figura 10
Distribución de la variable INGLES_ANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable ESPANIIPASA representa las calificaciones de los alumnos que cursaron segundo grado en la materia de español. Como se muestra en la figura 11, la curva para los alumnos que no reprobaron (clase 0) alcanza una densidad estimada cercana a 0.30 fluctuando en picos menores, los valores en las calificaciones están entre 6.5 y 10, esto indica que los alumnos que reprueban tercero tienen calificación variada.

Para los alumnos que reprobaron tercer grado (clase 1) la densidad estimada alcanza un valor cercano a 0.55 y con calificación entre 6 y 7. Esto quiere decir que los alumnos que reprobaron tercer grado es muy probable que tengan baja calificación en esta materia. Por la forma de las curvas podemos decir que esta variable es útil para el modelo.

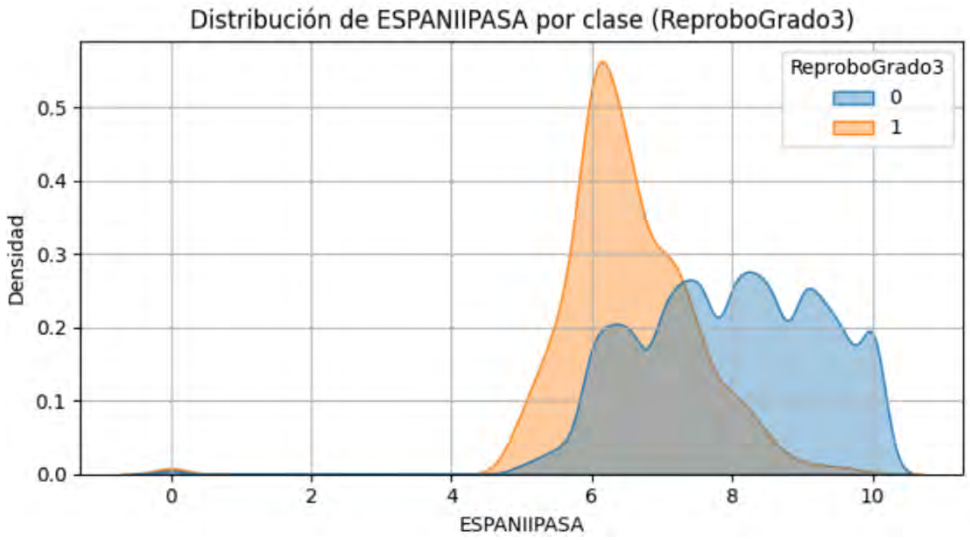


Figura 11
Distribución de la variable ESPANIIPASA según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

Las materias MATEMIIIPASA, CIENCIIIPASA, HISTORIPASA, FORMAIIIPASA e INGLESIIIPASA son similares por lo que no se mostrarán. Tienen curvas similares, lo que las difiere es la altura de las curvas, pero en lo demás son muy parecidas por lo que también son prospectas a considerar para el modelo.

La variable EDUCACIPASA representa las calificaciones de los alumnos de segundo grado de la materia de educación física. Se puede ver en la figura 12 que para la clase 0 (no reprobó) la gráfica muestra una densidad estimada mayor a 0.50 con una variación de picos en diferentes valores de calificaciones. Las calificaciones para esta clase están entre 9 y 10, lo que indica que los alumnos que no reprobaron tienen altas calificaciones. Para la clase 1 (reprobó) tiene más dispersos los valores de las calificaciones, entre 6 y 8.5, la densidad estimada no alcanza el 0.30, esto quiere decir que los alumnos que reprobaron tercer grado tienen un desempeño moderado en esta materia. Por la separación de las curvas esta variable se puede considerar para el modelo.

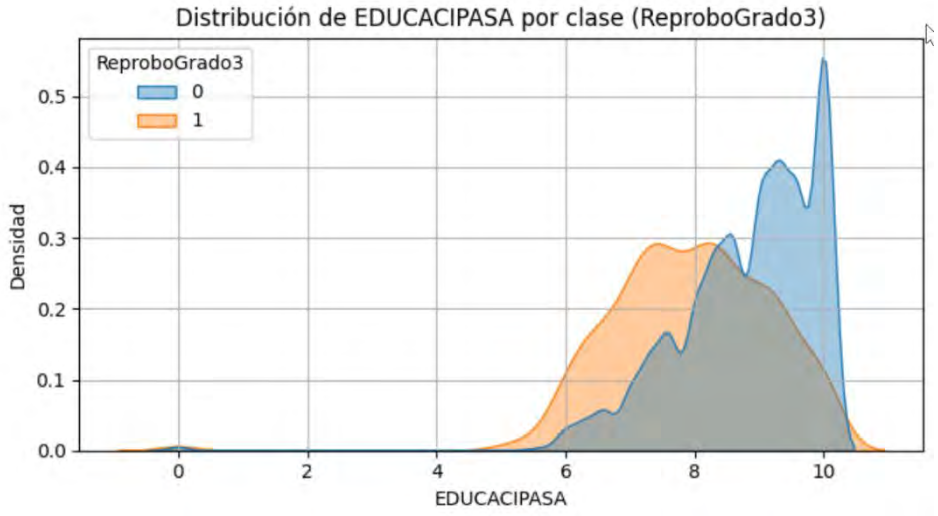


Figura 12
Distribución de la variable EDUCACIPASA según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable ARTES__PASA representa las calificaciones de los alumnos de segundo grado que cursaron la materia de artes.

En la figura 13 muestra que para los alumnos que no reprobaron (clase 0) sus calificaciones están en un rango de 7 y 10 con una densidad estimada 0.40 aproximadamente con varios picos en diferentes calificaciones. Esto indica que los alumnos que no reprobaron tercer grado tienen buenas notas. En la clase 1 (reprobó) la densidad estimada alcanza los 0.35 con un rango de calificaciones entre 6 y 8, lo que indica que los alumnos que reprobaron tercer grado tienen bajas calificaciones en esta materia.

Las curvas muestran separación de clases por lo que se puede decir que esta variable es útil para el modelo.

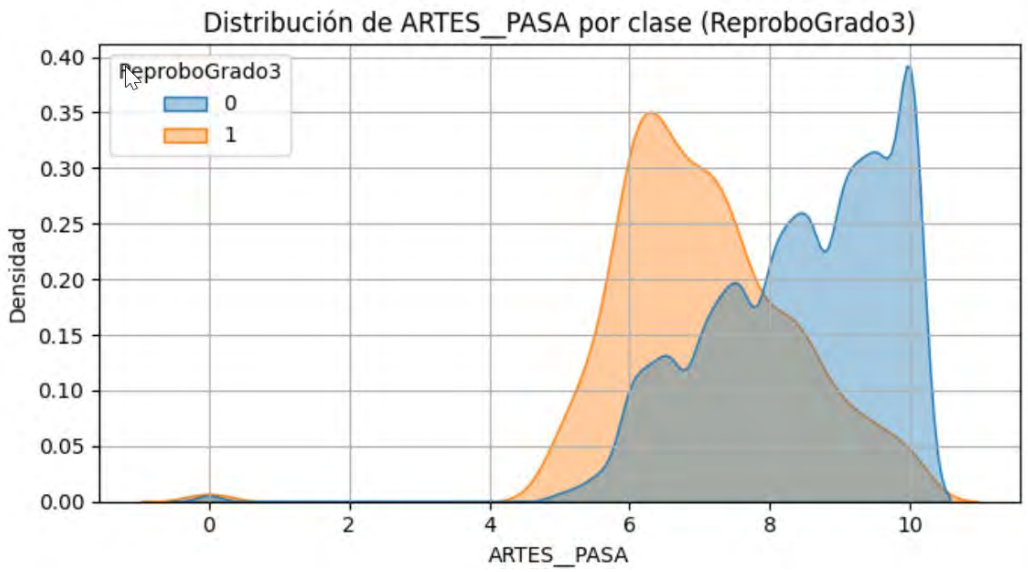


Figura 13

Distribución de la variable ARTES__PASA según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable AMBIT__PASA representa a los alumnos de tercer grado que cursaron la materia de ámbito o club, su gráfica es similar a la variable de artes por lo que no se mostrará. Tiene los valores más bajos en la densidad estimada de las clases, para 0 el valor de 32 y para la 1 el valor de 31 aproximadamente. Igualmente es candidata a elegirse para el modelo.

La variable escPriv muestra si los alumnos que cursaron primero y segundo año y que están en tercero estuvieron en una escuela privada.

Se puede ver en la figura 14 que esta es una variable dicotómica, muestra el valor de 0 para los que no estuvieron en escuela privada y 1 para los que sí. Se puede ver que la densidad estimada para los que no estudiaron en escuela privada es mayor para ambas clases y que los alumnos que más reprobaron son los que estudiaron en escuelas públicas. De acuerdo a las curvas de las clases se puede ver que no es útil para el modelo.

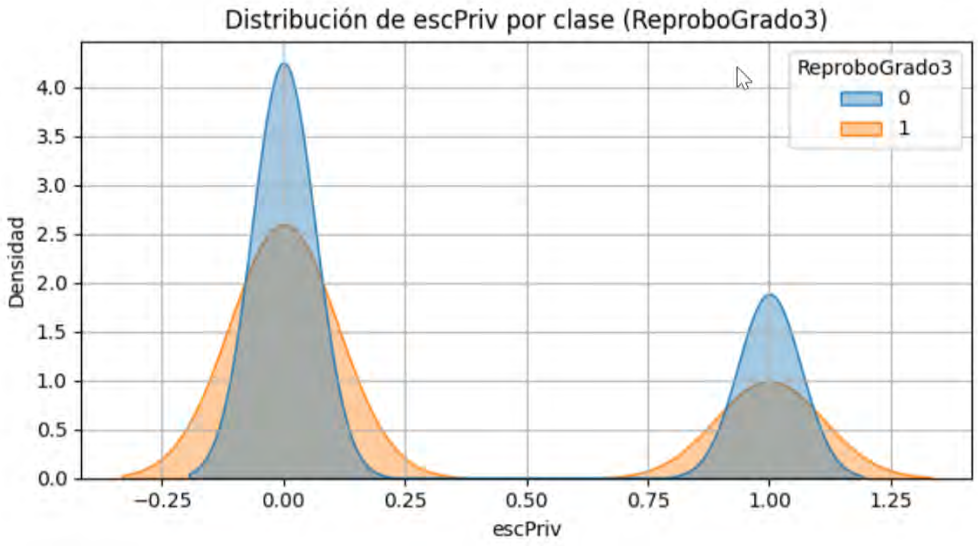


Figura 14
Distribución de la variable escPriv según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable escPub muestra si los alumnos que cursaron primero y segundo año y que están en tercero estuvieron en una escuela pública.

Esta variable es el inverso de la variable anterior por lo que no se muestra y al igual que la exterior no es útil para el modelo.

La variable bimRepr_ANTE representa los alumnos que tuvieron periodos reprobados en primer grado. Como se muestra en la figura 15 para la clase 0 (no reprobó) se tiene un valor de densidad estimada cercano a 0.70, lo que indica que la mayoría de los no reprobaron bimestres en primer grado tampoco reprobaron tercero.

Para la clase 1 (reprobó) se tiene una curva dispersa, tiene valores de 0 a 17. Lo que indica que los que reprobaron tercero tienen periodo reprobados en primero. Dada las curvas que se tienen esta es una variable candidata a utilizarse en el modelo.

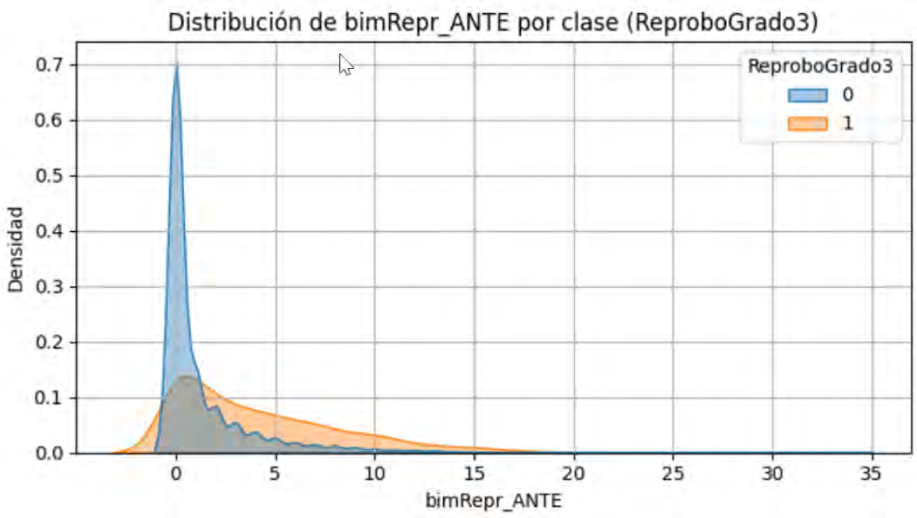


Figura 15
Distribución de la variable bimRepr_ANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable bimRepr_PASA representa los alumnos que tuvieron periodos reprobados en segundo grado.

La gráfica de esta variable es similar a la de bimRepr_ANTE, las variantes que tiene son los valores de la densidad estimada, para la clase 0 tienen el valor cercano a 1 y para la clase y alrededor de 0.16. El número de periodos reprobados para la clase 1 es 13. Como la variable anterior se considera que es muy útil para el modelo.

La variable bim6_8_ANTE representa los periodos de primer grado que el alumno obtuvo calificaciones entre 6 y menor o igual a 8. Para la clase 0 (no reprobó) la curva tiene dos cimas, la primer que está cerca del valor de 0.06 y la segunda que está por encima del valor de 0.03 lo que indica que para los alumnos que no reprobaron tercero tienden a tener menos periodos con calificaciones bajas. Para la clase 1 (reprobó) la curva es a la inversa de la otra clase, igualmente se tiene dos cimas la primera que está cercana al cero periodos y densidad estimada cercana a 0.02 y la segunda con más de 20 periodos y densidad estimada sobrepasando los 20 periodos. Esto nos indica que los alumnos con mayor número de periodos con

esas calificaciones tienen mayor riesgo de reprobar. De acuerdo a la separación de la gráfica, es un claro indicativo que esta variable es muy útil para el modelo.

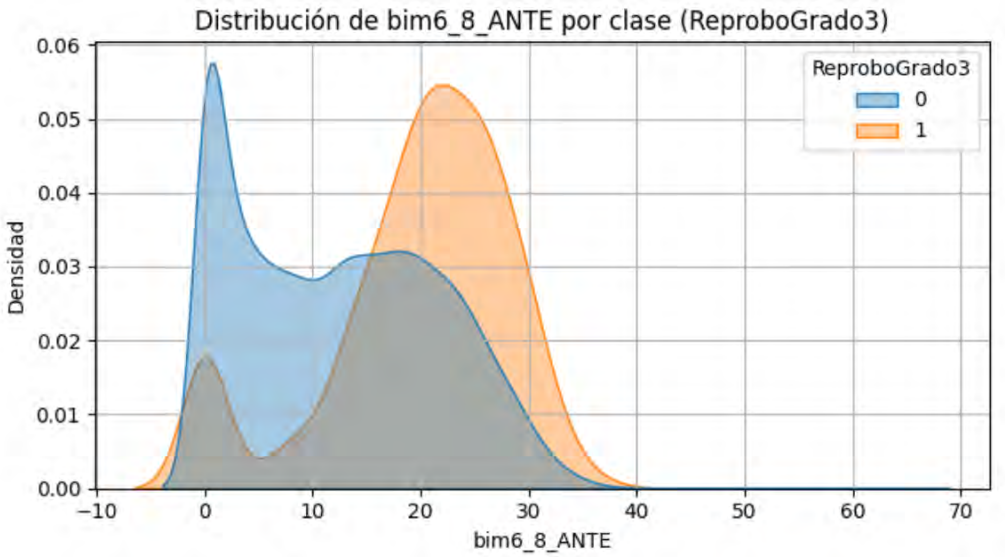


Figura 16
Distribución de la variable bim6_8_ANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable bim6_8_PASA representa los periodos de segundo grado que el alumno obtuvo calificaciones entre 6 y menor o igual a 8.

La gráfica de esta variable es similar a la anterior por lo que no se mostrará. La diferencia radica en los valores de la densidad estimada que para la clase 0 está cercana a 0.09 con periodos entre 0 y 16 y para la clase 1 pasa el valor de 0.10 con periodos entre 5 y 20. Al igual que la variable anterior es útil para el modelo.

La variable bim8_10_ANTE representa los periodos de primer grado que el alumno obtuvo calificaciones mayores o igual a 8. Se muestra en la figura 17 que para la clase 0 (no reprobó) tiene una densidad que va subiendo hasta 0.04 e inicia de 0 periodos hasta 40 en el punto más alto de la densidad estimada. Esto quiere decir que los alumnos que tienen mayor número de bimestres con esas calificaciones en primer grado no reprueban tercer grado. Para la clase 1(reprobó) se tiene un punto alto de densidad estimada de 0.05 entre los periodos 5 y 10. Esto

quiere decir que los alumnos que reprobaron tercer grado obtuvieron menos de 30% de calificaciones altas. Conforme a la separación de las curvas esta variable se define como útil para el modelo.

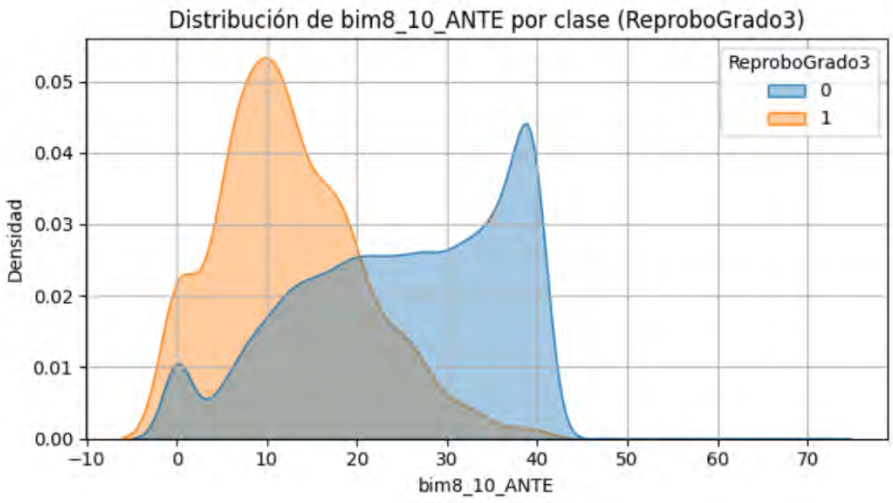


Figura 17
Distribución de la variable bim8_10_ANTE según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable bim8_10_PASA representa los periodos de segundo grado que el alumno obtuvo calificaciones mayores o igual a 8. Esta variable es similar a la anterior en forma por lo que no se mostrará.

Para la clase 0 (no reprobó) tiene un pico de densidad estimada de 0.09 y está entre los períodos entre 15 y 25. La curva se desplaza hacia la derecha lo que indica que entre más periodos tenga con esa calificación en ese grado es más probable que no repruebe tercero.Para la clase 1 (reprobó) tiene su mayor densidad entre los periodos 3 y 7 con un valor de densidad cercano a los 0.11 lo que indica que los alumnos que tuvieron menos de esas calificaciones reprobaron tercer grado.

La variable TurnoId representa el turno de la escuela. Se puede ver en la figura 18 que en el turno 0 que es matutino se tiene la mayor concentración para ambas clases, se puede ver con la densidad de las curvas.

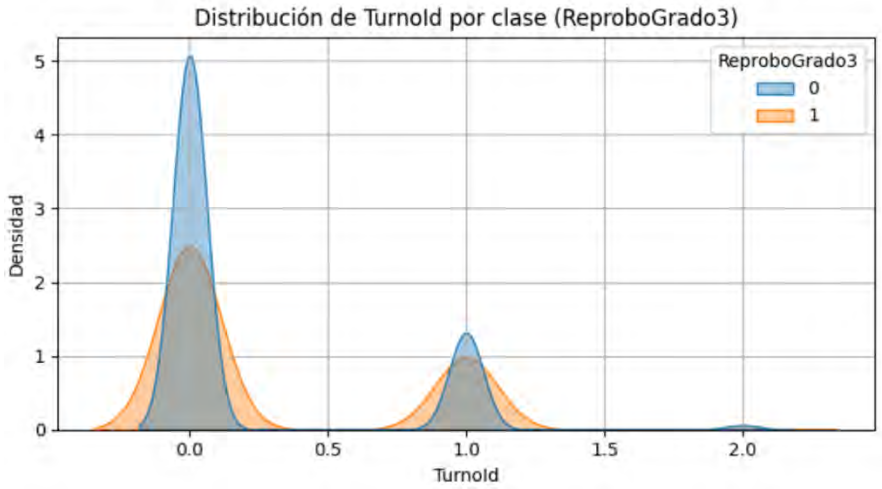


Figura 18
Distribución de la variable TurnoId según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La clase 1 muestra una mayor concentración en los turnos 1, vespertino y 2, discontinuo lo que puede sugerir que los alumnos que están en estos turnos están en mayor riesgo de reprobación. Estos valores nos indican que la variable se puede considerar para incluirla al modelo.

La variable SexoId representa el género del alumno. Esta variable es dicotómica, tiene el valor de 0 para mujer y 1 para hombre.

La figura 19 muestra que la clase 0 (no reprobó) tiene un comportamiento similar para ambos valores de la variable (0 y 1) no así para la clase 1 (reprobó) en la densidad estimada que tiene una mayor concentración para los casos que tienen el valor de 1 (hombre). Podemos decir que la variable es útil para el modelo.

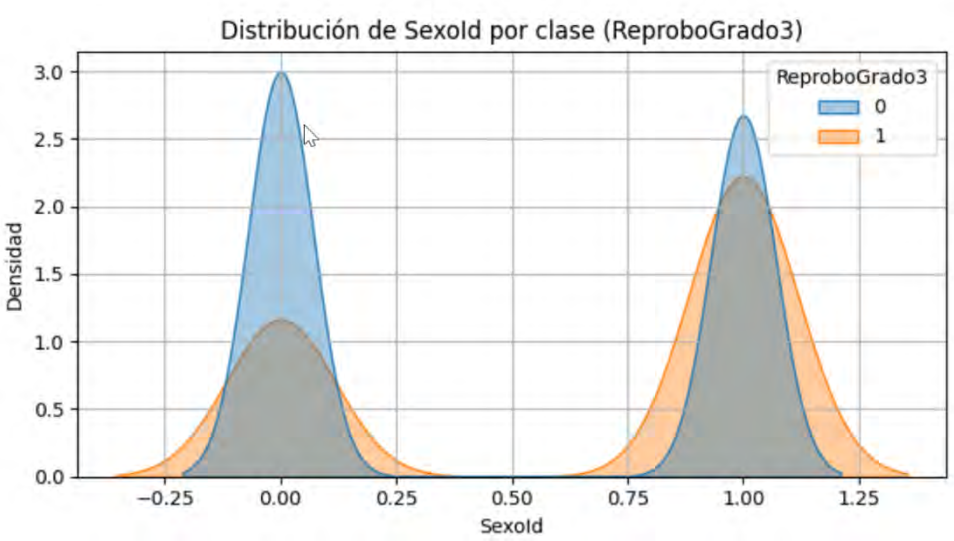


Figura 19

Distribución de la variable SexoId según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable Edad muestra la edad que tienen los jóvenes. Se muestra en la figura 20 que la mayoría de alumnos se encuentran con 14 años para ambas clases. Los estudiantes que tienen más de 14 años, existe una marcada concentración, se puede ver que la clase 0 tiene una mayor densidad estimada para estos casos, lo que indica que tienen un mayor riesgo de reprobar. La variable se considera importante para utilizarla en el modelo.

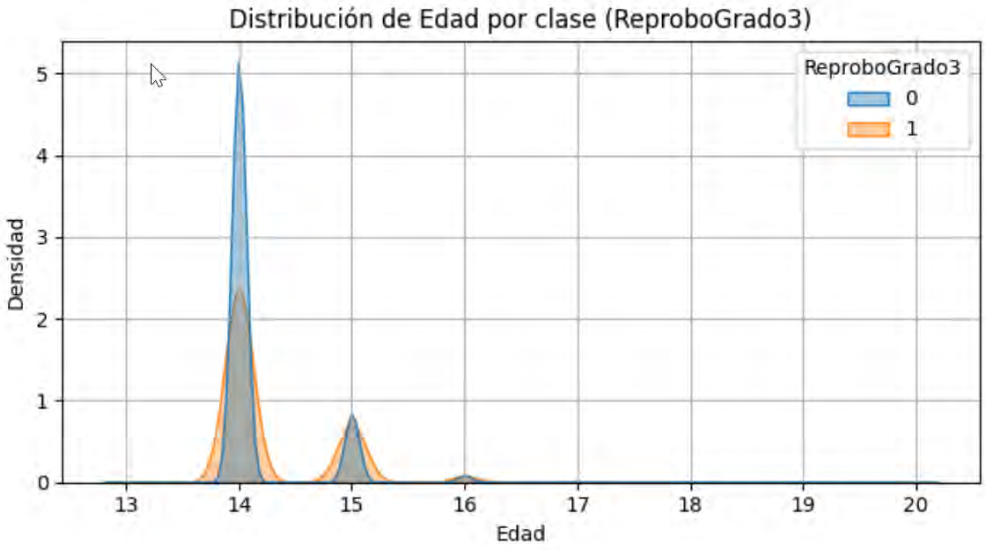


Figura 20

Distribución de la variable Edad según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable IdMunicipio representa la identificación del municipio en donde se encuentra la escuela.

En la figura 21 se puede ver que hay una mayor concentración en los municipios mayores a 13 para ambas clases, para los municipios menores a 14 existe una mayor densidad estimada para la clase 0 aunque está dispersa en los diferentes valores. Existe una superposición o solapamiento entre clases para esta variable y no muestra una clara separación entre ellas lo que indica que puede ser que no sea de gran utilidad para el modelo.

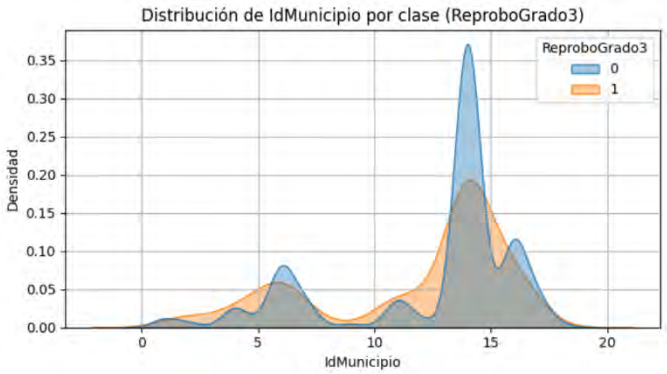


Figura 21

Distribución de la variable IdMunicipio según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable Zona identifica el lugar en la estructura organizacional en que se encuentra la escuela. En la figura 22 se muestra una densidad estimada dispersa entre todos los valores, se puede ver que existen algunos picos de la clase 0 (no reprobó) lo que indica que los alumnos no reprobaron, igualmente sobresalen unos picos de la clase 1 (reprobó) en los valores de 6 y 8. Se tiene una gran superposición en las curvas y no existe una separación clara entre las clases lo que indica que está variable no es útil para el modelo.

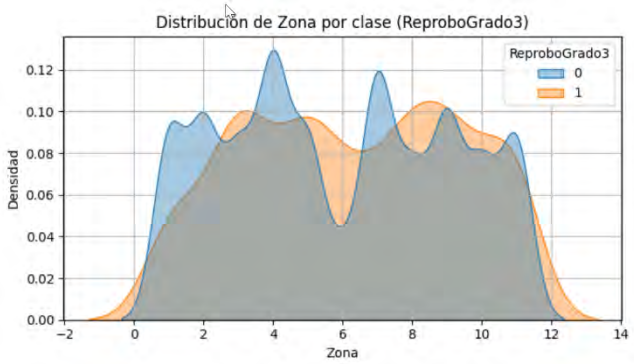


Figura 22

Distribución de la variable Zona según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable ReproboGrado1 indica si el estudiante reprobó el primer grado de secundaria. En la figura 23 se muestra que hay una gran densidad estimada de alumnos que no reprobaron primer grado, esto lo muestra la curva azul en el valor 0, también son mayoría los que no reprobaron tercero. Igualmente se puede ver que hay menos alumnos que reprobaron primero, pero la densidad estimada es alta para los que también reprobaron tercero. En la imagen está claro que la variable puede ser de gran utilidad en el modelo.

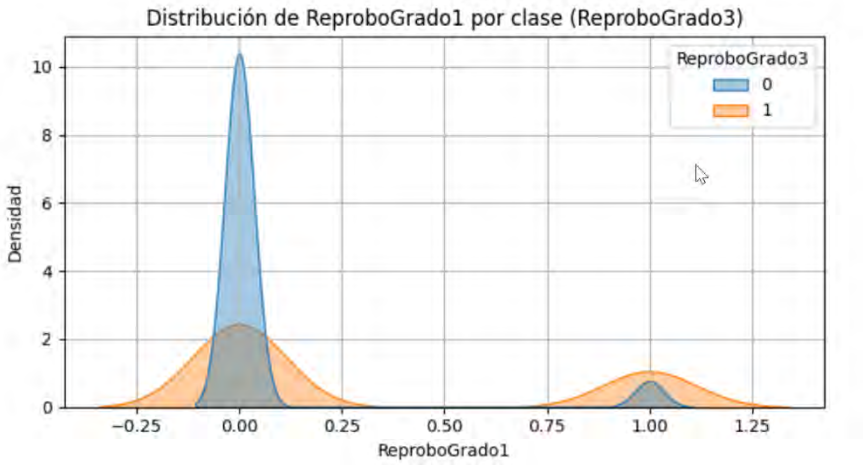


Figura 23

Distribución de la variable ReproboGrado1 según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable ReproboGrado2 indica si el estudiante reprobó el segundo grado de secundaria.

En la figura 24 se puede ver que la mayoría de los alumnos aprobaron segundo grado y también aprobaron tercero. Se tiene una densidad estimada muy alta para los alumnos que reprobaron segundo grado y que también reprobaron tercero. La variable muestra una separación clara entre clases y el comportamiento de las curvas muestra que la variable se debe de considerar para el modelo.

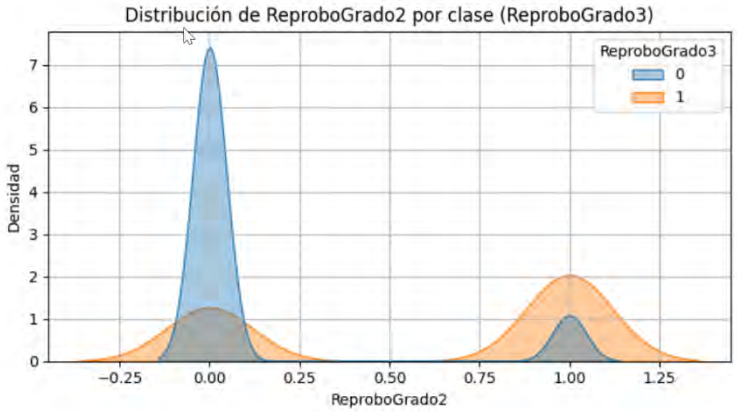


Figura 24

Distribución de la variable ReproboGrado2 según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable MARGINACION es la clasificación (ALTO, MEDIO, BAJO, MUY BAJO y SIN DATO CONAPO) que define el Consejo Nacional de Población (CONAPO) a la comunidad en la que se encuentra la escuela.

Se puede ver en la figura 25 que existe una concentración de estudiantes en los niveles bajo y muy bajo. Conforme a la gráfica existe un porcentaje más elevado de alumnos que reprobaron tercer grado, clase 1, en los niveles de marginación media (9.4) y sin datos (8.6) en comparación con los demás. No existe una diferencia marcada por lo que se puede decir que esta variable no es relevante para el modelo.

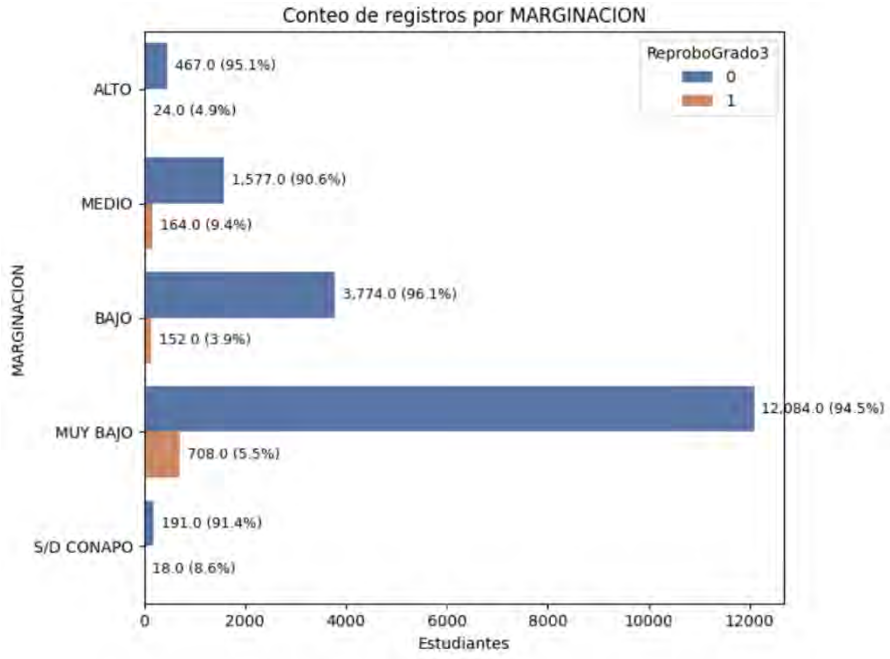


Figura 25
Distribución de la variable MARGINACION según el CONAPO y condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable de CATEGORIA representa la forma en que están catalogadas las comunidades donde se encuentran las escuelas, de acuerdo al Instituto Nacional

de Estadística y Geografía (INEGI): Urbana, Rural y S/D CONAPO. Como se ve en la figura 26 el mayor número de alumnos se encuentran en comunidades urbanas y el porcentaje de reprobados es un poco mayor en las zonas rurales con 8.8% y 5.4% en las urbanas. Esto se puede interpretar como que estudiantes que están en una zona rural tienen un poco más de riesgo de reprobar terror grado. Esta variable puede ser moderadamente útil para el modelo.

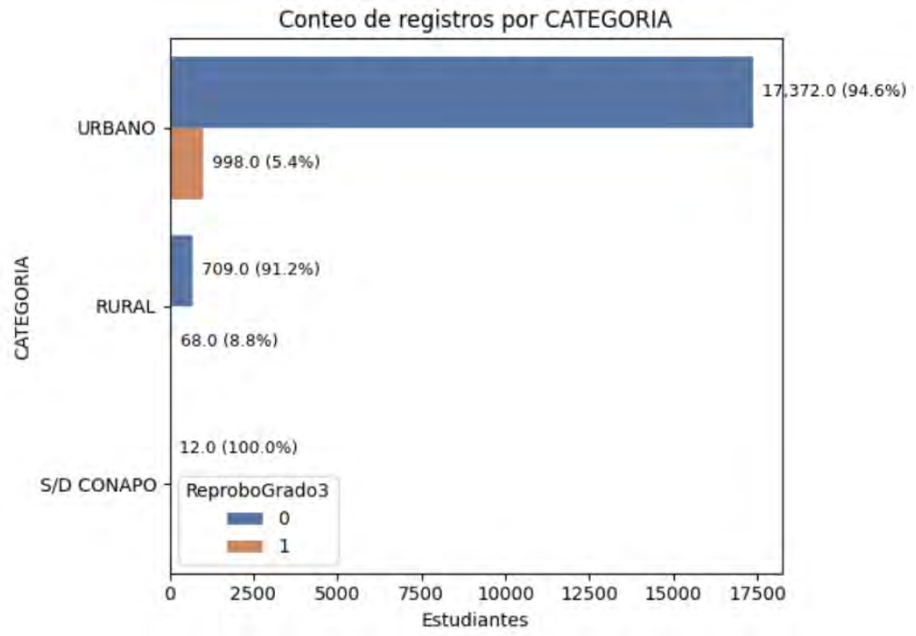


Figura 26
Distribución de la variable CATEGORIA según categoría CONAPO y condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

En la figura 27 se puede ver que los municipios en los que se concentran mayor número de estudiantes son: Querétaro, San Juan del Río y Corregidora. Los municipios que tienen mayor proporción de reprobados (clase 0) son Colón (30.6%), Pinal de Amoles (18%), Pedro Escobedo (10.2) y El Marqués (9,5%). El resto de los municipios tienen porcentajes que rondan el 5%. Aunque están dispersos los porcentajes entre todos los valores esta variable se puede considerar útil para el modelo.

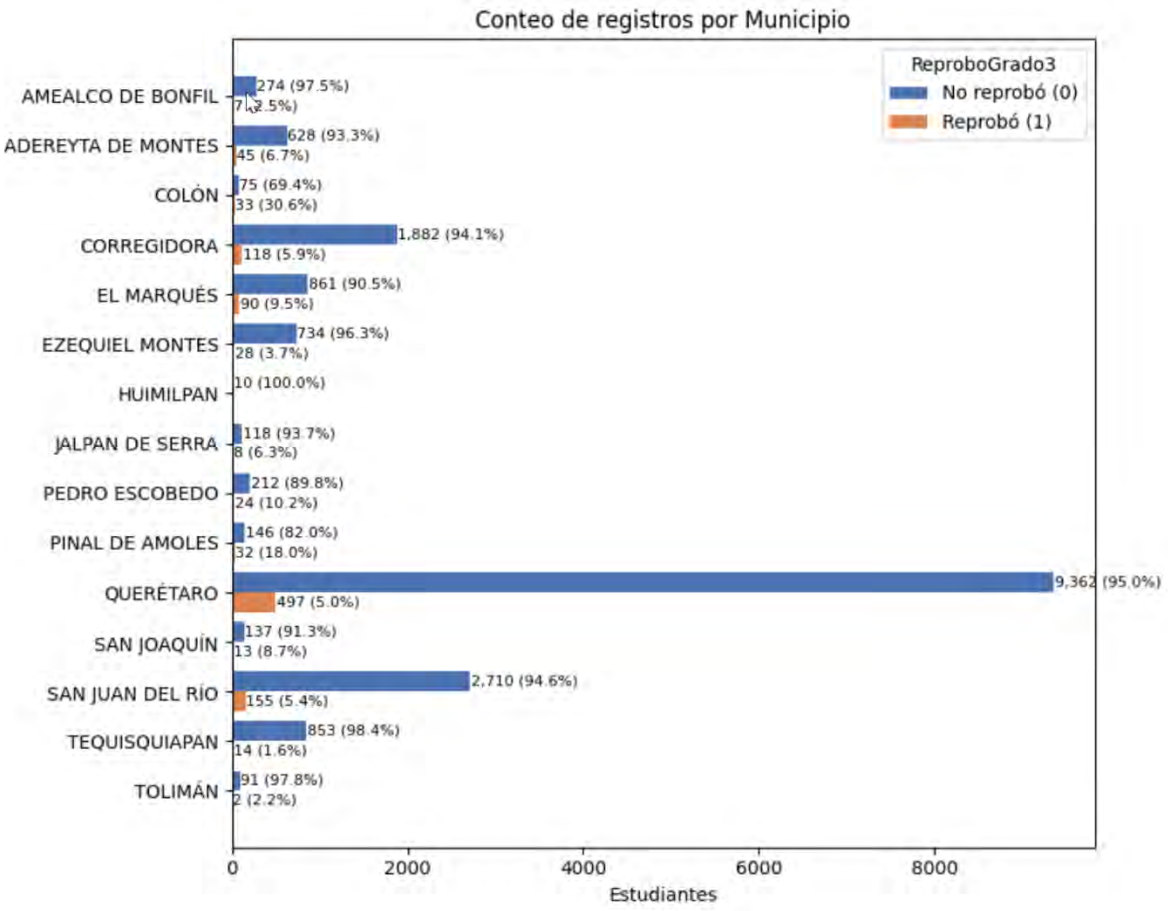


Figura 27

Distribución de la variable Municipio según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable de Sexo (Genero) representa el género de los estudiantes. En la figura 28 se puede ver que la diferencia entre el número de hombres y mujeres no es significativa, pero el porcentaje de reprobación entre uno y otro si lo es 7.6% y 3.7% respectivamente, un poco más del doble. Esto indica que el género puede estar relacionado con el riesgo de reprobación. Esta variable se puede considerar de gran utilidad para el modelo.

Como se muestra en la figura 29 la diferencia de los porcentajes entre hombres y mujeres no es significativa 51.8% para las mujeres y 48.2% para los hombres, una diferencia de 3.6% a favor de las mujeres.

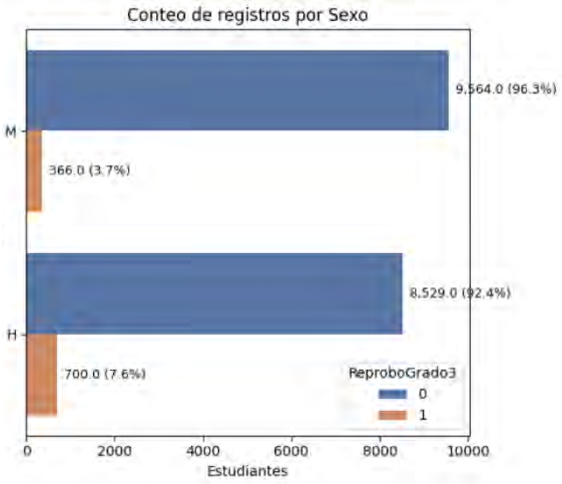


Figura 28
Distribución de la variable Sexo según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

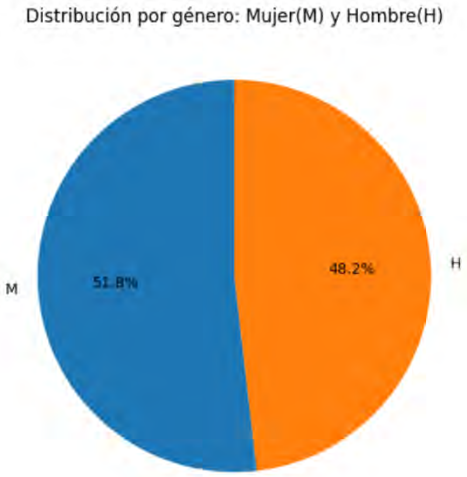


Figura 29
Distribución de la variable Sexo según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

La variable Turno se muestra en la figura 30 indica el turno de la escuela en donde los estudiantes cursaron tercer grado, se concentran mayormente en el turno matutino. La diferencia de porcentaje de los estudiantes que reprobaron en los diferentes turnos es poco por lo que se puede considerar que esta variable no es útil para el modelo.

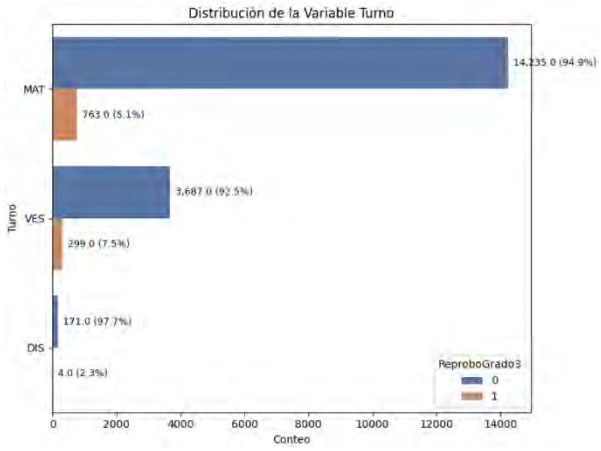


Figura 30
Distribución de la variable Turno según condición de reprobación en tercer grado

Nota. Elaboración propia con base en datos de USEBEQ (2025).

En la figura 31 se muestra la variable ReproboGrado3 que representa el indicador que nos dice si el estudiante reprobó tercer grado de secundaria. Como se mencionó, la utilizaremos como target, objetivo o etiqueta para el entrenamiento y pruebas del modelo. Claramente se puede ver que el conjunto de datos tiene un alto porcentaje de desbalance 94.4% para la clase mayoritaria, el valor de 0 y 5.6% para la clase minoritaria, el valor de 1.

Figura 31
Distribución de la variable ReproboGrado3 según condición de reprobación en tercer grado
Nota. Elaboración propia con base en datos de USEBEQ (2025).

En la figura 32 se utilizan 4 variables para su representación. Tenemos la variable de ReproboGrado3 que es la variable objetivo, nos dice si el estudiante reprobó (clase 1) o no reprobó (clase 0), igualmente se utiliza la variable CATEGORIA que nos dice si la comunidad en donde se encuentra la escuela es urbana, rural o sin definición, por último, se utilizan las variables de georreferenciación de la escuela (ESCUELALON y ESCUELALAT).

La combinación de todas nos muestra la concentración espacial de la escuela y la categoría, están separadas en dos planos, uno por cada clase de la variable ReproboGrado3.

Con esta gráfica se confirma lo que se vio en las otras. Los municipios que concentran más estudiantes son Querétaro, Corregidora y San Juan del Río y están

en las zonas urbanas. Se puede ver que los alumnos que no reprueban están igualmente en estas zonas, no así para los estudiantes que reprueban, hay menos en las zonas urbanas en proporción, concentrando un poco en la zona rural.

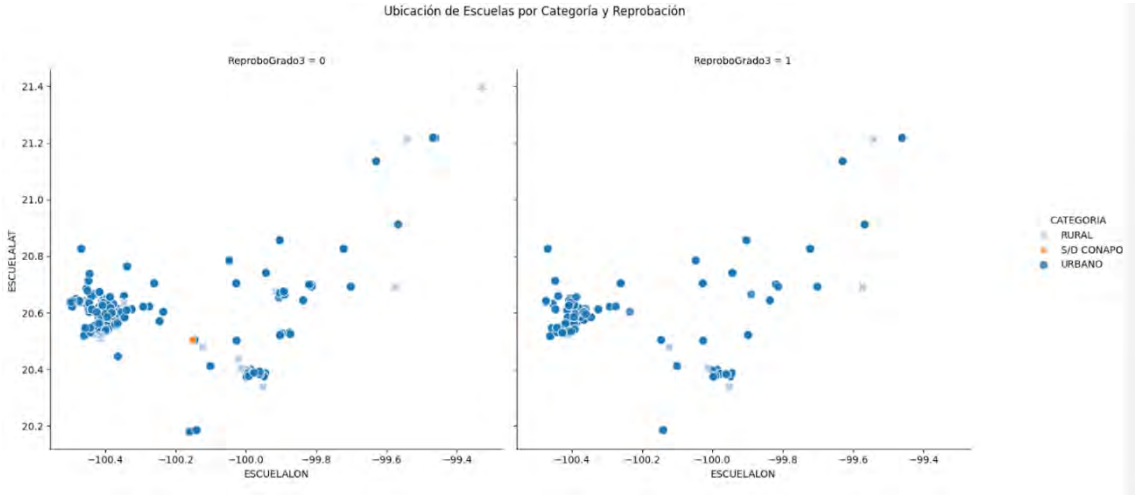


Figura 32
Distribución geográfica de escuelas por categoría CONAPO y condición de reprobación en tercer grado
Nota. Elaboración propia a partir de la georreferenciación de escuelas con base en su categoría CONAPO y la variable de reprobación en tercer grado (ReproboGrado3), con datos de USEBEQ (2025)

En la figura 33 de ubicación de las escuelas (variables ESCUELALON y ESCUELALAT) por municipio separada por clase, alumnos que no reprobaron (clase 0) y alumnos que reprobaron (clase 1), en la gráfica se separan los municipios por forma y color. Se puede ver que existe mayor concentración de estudiantes en los municipios de Querétaro, Corregidora, El Marqués y San Juan del Río.

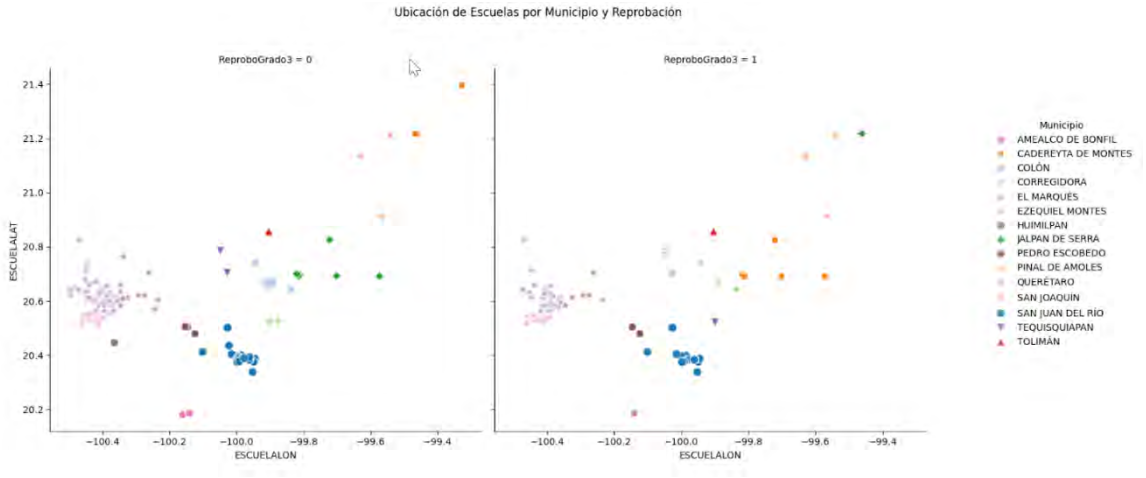


Figura 33

Distribución geográfica de escuelas por municipio y condición de reprobación en tercer grado

Nota. Elaboración propia a partir de la georreferenciación de escuelas y la variable de reprobación en tercer grado (ReproboGrado3), con datos de USEBEQ (2025)

Una vez que se analizaron las variables conforme a las gráficas se muestra un resumen estadístico, en la tabla 8a, con las principales medidas de tendencia central: media, desviación estándar, mínimo, máximo y los cuartiles.

Haremos un breve análisis de manera general de las variables. Por cuestión de formato separamos el número de variables en dos segmentos.

Se puede ver que existen valores mínimos de 0 en las calificaciones, eso no es normal, ya que, conforme a los criterios, las calificaciones aprobatorias son mayores o iguales a 6 y las reprobatorias son con el valor de 5. Igualmente, las faltas tienen unos valores muy elevados, La variable Grado tienen el mismo valor de 3 para todos los registros porque es el grado del que se tomó la información. El desbalance alto que tiene la data se puede ver claramente en las variables dicotómicas que tienen valor de 0 en los dos o tres cuartiles. Se tiene un complemento de medidas que nos ayudarán a tener un mejor análisis de las variables.

📊 **Estadística de variables numéricas**

Variable	count	mean	std	min	25%	50%	75%	max
IdAlumno	19159.00	320619.33	180453.72	38812.00	235670.50	297314.00	387992.00	859483.00
IdClavecct	19159.00	2639.60	1682.60	1200.00	1484.00	1534.00	4468.00	8019.00
TurnoId	19159.00	0.23	0.44	0.00	0.00	0.00	0.00	2.00
Grado	19159.00	3.00	0.00	3.00	3.00	3.00	3.00	3.00
SexoId	19159.00	0.48	0.50	0.00	0.00	0.00	1.00	1.00
Edad	19159.00	14.19	0.47	13.00	14.00	14.00	14.00	20.00
IdMunicipio	19159.00	12.43	3.97	1.00	11.00	14.00	14.00	18.00
Region	19159.00	3.54	0.73	1.00	3.00	4.00	4.00	4.00
Sector	19159.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zona	19159.00	5.90	3.19	1.00	3.00	6.00	9.00	11.00
ReproboGrado1	19159.00	0.12	0.33	0.00	0.00	0.00	0.00	1.00
ReproboGrado2	19159.00	0.17	0.37	0.00	0.00	0.00	0.00	1.00
ReproboGrado3	19159.00	0.06	0.23	0.00	0.00	0.00	0.00	1.00
Faltas_ANTE	19159.00	10.36	20.00	0.00	0.00	3.00	12.00	767.00
Faltas_PASA	19159.00	12.30	20.54	0.00	0.00	5.00	16.00	819.00
HISTOR_ANTE	19159.00	7.59	1.90	0.00	6.80	7.80	8.80	10.00
FORMACIANTE	19159.00	7.39	1.90	0.00	6.60	7.60	8.60	10.00
ESPANO_ANTE	19159.00	7.65	1.90	0.00	6.90	7.90	8.80	10.00
MATEMATANTE	19159.00	7.72	1.92	0.00	7.00	8.00	9.00	10.00
CIENCIAANTE	19159.00	8.55	1.92	0.00	8.20	9.00	9.60	10.00
GEOGRAFANTE	19159.00	8.21	1.97	0.00	7.60	8.60	9.40	10.00
INGLES_ANTE	19159.00	7.95	1.94	0.00	7.30	8.30	9.20	10.00
ESPANIIPASA	19159.00	7.88	1.33	0.00	7.00	8.00	9.00	10.00
MATEMIIPASA	19159.00	7.57	1.36	0.00	6.30	7.60	8.60	10.00
CIENCIIPASA	19159.00	7.68	1.35	0.00	6.60	7.60	8.60	10.00
HISTORIPASA	19159.00	7.89	1.35	0.00	7.00	8.00	9.00	10.00

Tabla 8a

Estadística descriptiva (1/2) de variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

En la siguiente tabla se muestran otras medidas de tendencia, dispersión y forma. Se muestra la columna de coeficiente de variación (CV) que se calcula dividiendo la desviación estándar (std) entre la media, solo si la media es diferente de cero, nos muestra la dispersión de los datos con respecto a la media, normaliza la medida para compararla entre las demás variables, sobre todo cuando tienen diferente escala. La variable de Utilidad clasifica la variable con forma a sus datos,

usa el CV para su definición. El sesgo nos muestra la asimetría que tiene la curva. Los valores Atípicos u Outliers, son valores que se tienen a la izquierda o derecha y quedan fuera del rango que se tiene al aplicar la fórmula de 1.5 por el rango intercuartil (IQR). Los valores Nulos son datos faltantes que trae la información, nulos no quiere decir valores de cero o espacio en blanco, es ausencia de valor. El IQR es el rango intercuartil y es la resta del tercer cuartil menos el segundo cuartil, en proporción de información corresponde al 50% de los datos. Valores Únicos, son los valores distintos que se tienen en esa columna. Mediana, es el valor central de los datos siempre y cuando estén ordenados. La columna de %Dominante representa el porcentaje del valor más frecuente, está dado por la fórmula de la frecuencia entre el total por 100.

En la tabla se muestran las variables que pueden ser de manera preliminar útiles para el modelo, en este caso, los casos que no son descartables. Las variables que se considera que tiene valores atípicos son mayores al 5%. En este caso para la columna de %Nulos no tiene estos valores, lo que sí tiene, son valores en 0, que de acuerdo a los criterios descritos no corresponden.

Estadística enriquecida de variables numéricas

Variable	CV	Utilidad	Sesgo	Atípicos	% Nulos	IQR	Valores únicos	Mediana	% Dominante
IdAlumno	0.56	Útil	Derecha	0.55	0.00	152321.50	19159	297314.00	0.01
IdClavecct	0.64	Útil	Derecha	-0.29	0.00	2984.00	220	1534.00	1.50
TurnoId	1.94	Alta variabilidad	Derecha	1.47	0.00	0.00	3	0.00	78.28
Grado	0.00	Descartable	Simétrico	nan	0.00	0.00	1	3.00	100.00
SexoId	1.04	Alta variabilidad	Simétrico	-1.99	0.00	1.00	2	0.00	51.83
Edad	0.03	Descartable	Derecha	13.40	0.00	0.00	8	14.00	83.73
IdMunicipio	0.32	Útil	Izquierda	0.32	0.00	3.00	15	14.00	51.46
Region	0.20	Útil	Izquierda	1.43	0.00	1.00	4	4.00	66.91
Sector	nan	Sin información	Simétrico	nan	0.00	0.00	1	0.00	100.00
Zona	0.54	Útil	Simétrico	-1.25	0.00	6.00	11	6.00	12.81
ReproboGrado1	2.69	Alta variabilidad	Derecha	3.37	0.00	0.00	2	0.00	87.85
ReproboGrado2	2.25	Alta variabilidad	Derecha	1.24	0.00	0.00	2	0.00	83.46
ReproboGrado3	4.12	Alta variabilidad	Derecha	13.03	0.00	0.00	2	0.00	94.44
Faltas_ANTE	1.93	Alta variabilidad	Derecha	139.21	0.00	12.00	179	3.00	34.10
Faltas_PASA	1.67	Alta variabilidad	Derecha	144.80	0.00	16.00	169	5.00	25.97
HISTOR_ANTE	0.25	Útil	Izquierda	7.09	0.00	2.00	60	7.80	3.84
FORMACIANTE	0.26	Útil	Izquierda	6.19	0.00	2.00	59	7.60	4.68
ESPANO_ANTE	0.25	Útil	Izquierda	7.42	0.00	1.90	60	7.90	3.83
MATEMATANTE	0.25	Útil	Izquierda	7.39	0.00	2.00	60	8.00	3.82
CIENCIAANTE	0.22	Útil	Izquierda	12.16	0.00	1.40	53	9.00	10.08
GEOGRAFANTE	0.24	Útil	Izquierda	8.82	0.00	1.80	52	8.60	8.45
INGLES_ANTE	0.24	Útil	Izquierda	8.08	0.00	1.90	51	8.30	4.87
ESPANIIPASA	0.17	Útil	Simétrico	1.53	0.00	2.00	17	8.00	8.35
MATEMIIPASA	0.18	Útil	Simétrico	0.93	0.00	2.30	19	7.60	11.36
CIENCIIPASA	0.18	Útil	Simétrico	1.10	0.00	2.00	17	7.60	9.40
HISTORIPASA	0.17	Útil	Simétrico	1.32	0.00	2.00	17	8.00	8.57

Tabla 9a

Estadística enriquecida (1/2) de variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

Conforme a lo que se explicó en el párrafo anterior se genera la siguiente tabla que tiene la columna de alerta donde se muestran las observaciones una vez que se aplican los criterios anteriores. Estas descripciones de alerta se tratarán de disminuirlas o eliminarlas en la siguiente fase.

Alertas estadísticas de variables numéricas

Variable	Alerta
IdAlumno	Sin alerta
IdClavecct	Sin alerta
TurnoId	Sin variación (IQR=0)
Grado	Alta redundancia, Sin variación (IQR=0)
SexoId	Sin alerta
Edad	Sin variación (IQR=0), Atípicos alta (outliers)
IdMunicipio	Sin alerta
Region	Sin alerta
Sector	Alta redundancia, Sin variación (IQR=0)
Zona	Sin alerta
ReproboGrado1	Sin variación (IQR=0), Atípicos alta (outliers)
ReproboGrado2	Sin variación (IQR=0)
ReproboGrado3	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
Faltas_ANTE	Alta dispersión, Atípicos alta (outliers)
Faltas_PASA	Alta dispersión, Atípicos alta (outliers)
HISTOR_ANTE	Atípicos alta (outliers)
FORMACIANTE	Atípicos alta (outliers)
ESPANO_ANTE	Atípicos alta (outliers)
MATEMATANTE	Atípicos alta (outliers)
CIENCIAANTE	Atípicos alta (outliers)
GEOGRAFANTE	Atípicos alta (outliers)
INGLES_ANTE	Atípicos alta (outliers)
ESPANIIPASA	Sin alerta
MATEMIIPASA	Sin alerta
CIENCIIPASA	Sin alerta
HISTORIPASA	Sin alerta

Tabla 10a

Alertas estadísticas (1/2) en variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado

La tabla 8b se muestra la continuación de la lista de variables que se tienen en el set de datos, las columnas presentan estadística descriptiva con las principales medidas de tendencia central, media, desviación estándar, valor mínimo, los 3 cuartiles y el valor máximo. Se puede ver que en análisis del primer grupo de variables se mantiene con este segundo grupo, por ejemplo, el valor de 0 para las materias, los promedios de las calificaciones son aceptables basándose en los promedios. Las variables de grado1, grado2 y grado3 tienen los valores muy alto cercano a 1 considerando que son variables binarias, lo que indica que casi todos los niños cursaron ese grado, esto se puede corroborar con el promedio de años cursados con 2.96% en la variable TOTALGRADOS. También en la variable numescEstuvo podemos ver que casi la mayoría de los alumnos estuvieron en una sola escuela, mismo caso para los que estuvieron en escuelas públicas (escPub) que tienen un 68% sobre 31% para escuelas privadas (escPriv). Se tiene poca

dispersión en las coordenadas de georeferenciación de las escuelas, lo que indican que se concentra una cantidad de estudiantes en cada una de ellas, no así en los datos de georeferencia de la vivienda de los alumnos (as). El valor que se tiene en la media de las variables bimRepr_ANTE y bimRepr_PASA, es bajo, es recomendable ver más a detalle la información en general pues tener un desbalance tan grande puede estar creando falsa apreciaciones en el análisis.

Estadística de variables numéricas

Variable	count	mean	std	min	25%	50%	75%	max
FORMAIIPASA	19159.00	7.96	1.35	0.00	7.00	8.00	9.00	10.00
INGLEIIPASA	19159.00	7.79	1.34	0.00	6.60	8.00	9.00	10.00
EDUCACIPASA	19159.00	8.76	1.12	0.00	8.00	9.00	9.60	10.00
ARTES__PASA	19159.00	8.35	1.36	0.00	7.30	8.60	9.60	10.00
AMBIT__PASA	19159.00	8.23	1.54	0.00	7.30	8.50	9.30	10.00
GRADO1	19159.00	0.96	0.19	0.00	1.00	1.00	1.00	1.00
GRADO2	19159.00	1.00	0.05	0.00	1.00	1.00	1.00	1.00
GRADO3	19159.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
TOTALGRADOS	19159.00	2.96	0.20	2.00	3.00	3.00	3.00	3.00
ESCUELALAT	19159.00	20.58	0.14	20.18	20.53	20.60	20.64	21.40
ESCUELALON	19159.00	-100.24	0.25	-100.50	-100.41	-100.39	-99.99	-99.33
ALUMNOLAT	19159.00	0.42	2.92	0.00	0.00	0.00	0.00	21.22
ALUMNOLON	19159.00	-2.05	14.21	-100.49	0.00	0.00	0.00	0.00
escPriv	19159.00	0.31	0.46	0.00	0.00	0.00	1.00	1.00
escPub	19159.00	0.68	0.47	0.00	0.00	1.00	1.00	1.00
numescEstuvo	19159.00	1.01	0.11	1.00	1.00	1.00	1.00	2.00
bimRepr_ANTE	19159.00	1.44	2.72	0.00	0.00	0.00	2.00	34.00
bimSinCali_ANTE	19159.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bim6_8_ANTE	19159.00	12.71	9.39	0.00	4.00	13.00	20.00	65.00
bim8_10_ANTE	19159.00	24.22	11.62	0.00	15.00	25.00	35.00	70.00
bimRepr_PASA	19159.00	1.07	2.04	0.00	0.00	0.00	1.00	15.00
bimSinCali_PASA	19159.00	0.00	0.04	0.00	0.00	0.00	0.00	3.00
bim6_8_PASA	19159.00	7.76	6.00	0.00	2.00	8.00	13.00	26.00
bim8_10_PASA	19159.00	15.11	7.08	0.00	9.00	16.00	22.00	27.00
PROMEDIO_ANTE	19159.00	7.87	1.81	0.00	7.37	8.14	8.89	10.00
PROMEDIO_PASA	19159.00	8.01	1.09	0.00	7.20	8.01	8.86	10.00

Tabla 8b

Estadística descriptiva (2/2) de variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

Para definir si una variable es útil se utiliza la variabilidad, la diferencia entre los valores de la variable, y la distribución es la forma en cómo se organizan o agrupan los datos. En el complemento de los datos se muestra en la tabla 9b y se observan que no se tienen valores nulos, existen variables que aparecen como Descartable, se debe a la baja variabilidad de los datos o tienen un solo valor, por ejemplo la variable GRADO2, se tiene

como Descartable porque su variabilidad es baja, su CV es menor o igual 0.1, además es Dominante porque tiene muy pocos valores (solo 1 y 0) y uno de ellos tiene más del 90%, estás tienen poco valor predictivo para el modelo, por el contrario las que tienen un Alta Variabilidad.

Estadística enriquecida de variables numéricas

Variable	CV	Utilidad	Sesgo	Atípicos	% Nulos	IQR	Valores únicos	Mediana	% Dominante
FORMAIIPASA	0.17	Útil	Izquierda	1.39	0.00	2.00	17	8.00	9.15
INGLEIIPASA	0.17	Útil	Simétrico	1.33	0.00	2.40	18	8.00	8.79
EDUCACIPASA	0.13	Útil	Izquierda	7.50	0.00	1.60	17	9.00	20.36
ARTES__PASA	0.16	Útil	Izquierda	2.65	0.00	2.30	17	8.60	16.36
AMBIT__PASA	0.19	Útil	Izquierda	7.64	0.00	2.00	22	8.50	15.28
GRADO1	0.20	Útil	Izquierda	21.25	0.00	0.00	2	1.00	96.18
GRADO2	0.05	Descartable	Izquierda	430.43	0.00	0.00	2	1.00	99.77
GRADO3	0.00	Descartable	Simétrico	nan	0.00	0.00	1	1.00	100.00
TOTALGRADOS	0.07	Descartable	Izquierda	19.76	0.00	0.00	2	3.00	95.95
ESCUELALAT	0.01	Descartable	Derecha	5.61	0.00	0.11	194	20.60	2.96
ESCUELALON	-0.00	Descartable	Derecha	0.02	0.00	0.41	194	-100.39	2.96
ALUMNOLAT	6.92	Alta variabilidad	Derecha	43.90	0.00	0.00	386	0.00	97.95
ALUMNOLON	-6.92	Descartable	Izquierda	43.90	0.00	0.00	386	0.00	97.95
escPriv	1.51	Alta variabilidad	Derecha	-1.29	0.00	1.00	2	0.00	69.45
escPub	0.68	Útil	Izquierda	-1.38	0.00	1.00	2	1.00	68.38
numescEstuvo	0.11	Útil	Derecha	79.04	0.00	0.00	2	1.00	98.81
bimRepr_ANTE	1.89	Alta variabilidad	Derecha	8.82	0.00	2.00	23	0.00	60.53
bimSinCali_ANTE	nan	Sin información	Simétrico	nan	0.00	0.00	1	0.00	100.00
bim6_8_ANTE	0.74	Útil	Simétrico	-1.04	0.00	16.00	43	13.00	11.88
bim8_10_ANTE	0.48	Útil	Simétrico	-0.92	0.00	20.00	44	25.00	7.73
bimRepr_PASA	1.91	Alta variabilidad	Derecha	7.33	0.00	1.00	16	0.00	63.96
bimSinCali_PASA	79.91	Alta variabilidad	Derecha	6381.33	0.00	0.00	2	0.00	99.98
bim6_8_PASA	0.77	Útil	Simétrico	-1.12	0.00	11.00	26	8.00	15.66
bim8_10_PASA	0.47	Útil	Simétrico	-1.20	0.00	13.00	26	16.00	15.25
PROMEDIO_ANTE	0.23	Útil	Izquierda	10.81	0.00	1.51	800	8.14	3.82
PROMEDIO_PASA	0.14	Útil	Izquierda	5.25	0.00	1.66	929	8.01	0.50

Tabla 9b

Estadística enriquecida (2/2) de variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

Se puede ver que la columna de sesgo se tiene en casi todas las variables con sesgo a la izquierda a la derecha, lo ideal es tener variables que no tengan sesgo, es decir, que su gráfica tenga una distribución normal, ese es el objetivo de la fase siguiente.

En la tabla 10b se tiene la columna de alerta que muestra una descripción simple de los hallazgos que se encontraron de cada variable. Se puede ver que son pocas las que tienen la leyenda “Sin Alerta”.

Alertas estadísticas de variables numéricas

Variable	Alerta
FORMAIIPASA	Sin alerta
INGLEIIPASA	Sin alerta
EDUCACIPASA	Atípicos alta (outliers)
ARTES__PASA	Sin alerta
AMBIT__PASA	Atípicos alta (outliers)
GRADO1	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
GRADO2	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
GRADO3	Alta redundancia, Sin variación (IQR=0)
TOTALGRADOS	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
ESCUELALAT	Atípicos alta (outliers)
ESCUELALON	Sin alerta
ALUMNOLAT	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
ALUMNOLON	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
escPriv	Alta dispersión
escPub	Sin alerta
numescEstuvo	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
bimRepr_ANTE	Alta dispersión, Atípicos alta (outliers)
bimSinCali_ANTE	Alta redundancia, Sin variación (IQR=0)
bim6_8_ANTE	Sin alerta
bim8_10_ANTE	Sin alerta
bimRepr_PASA	Alta dispersión, Atípicos alta (outliers)
bimSinCali_PASA	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)
bim6_8_PASA	Sin alerta
bim8_10_PASA	Sin alerta
PROMEDIO_ANTE	Atípicos alta (outliers)
PROMEDIO_PASA	Atípicos alta (outliers)

Tabla 10b

Alertas estadísticas (2/2) en variables numéricas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado

Como resumen de este análisis se puede concluir que el historial académico es un buen indicador para poder predecir si los estudiantes van a reprobar tercer grado, el sexo puede ser un factor determinante para que un estudiante repruebe tercer grado, los hombres tiene mayor riesgo.

En la categoría, tipo de escuela y la marginación es información que está relacionada directamente a los estudiantes que están en riesgo de reprobación en el tercer grado de secundaria, se debe considerar la integración de estas variables en el modelo.

3.2.3 Metodología para el objetivo específico 3.

3.2.3.1 Preparación de la información:

Una vez que se analizó la información, el siguiente paso es prepararla y adecuarla y con ello atender el objetivo O3 con el que buscamos tenerla lista para empezar con el entrenamiento de los modelos.

De acuerdo a las tablas de estadística enriquecida se identifican alertas en cada variable, primeramente, trabajaremos con las variables que tienen una pequeña o nula diferencia en el rango intercuartílico (IQR), valor de 0, ya que tienen el mismo valor en todos los registros o la mayoría de los registros tienen el mismo. Esto se corrige eliminando las variables.

Las variables que se eliminaron son: ALUMNOLAT, ALUMNOLON, Edad, GRADO1, GRADO2, ReproboGrado1, ReproboGrado2, TOTALGRADOS, TipoCategoria, TurnoId, bimSinCali_PASA y numescEstuvo. Quedando solo 38 variables como se muestra en las tablas que contienen las estadísticas.

A los registros que tienen valores atípicos les asigna el valor mínimo o máximo permitido, es decir, se toma el valor máximo que tiene el rango intercuartil y los valores que sean mayor de ese valor se les asignan el valor máximo. Para los casos que tienen valor más bajo que el valor mínimo del rango intercuartil se les asignará el valor mínimo.

Cuando tiene una alta dispersión, es decir, para los casos que el coeficiente de variación es mayor que 1, se aplica una técnica de escalamiento RobustScaler. En todas las materias de primer grado y segundo grado se identificaron valores en 0, esto provocaba que en las gráficas afectarán las medidas estadísticas, por ejemplo, es sesgo, la distribución, por mencionar algunas. Se eliminaron los registros que en alguna de las materias tuvieran valores menores a 5. Esto dejó un set de datos de datos de 18,176 registros.

En la tabla 11 a se muestra la estadística con las variables transformadas. Se pueden ver las variables FALTAS_ANTE y FALTAS_PASA que tienen valores negativos, esto debido al escalado que se realizó. La media en todas las métricas se incrementó, la desviación estándar ha disminuído, los valores mínimos igualmente se han incrementado.

Estadística de variables numéricas transformadas

Variable	count	mean	std	min	25%	50%	75%	max
IdAlumno	18176.00	303442.72	153342.32	38812.00	233310.50	292810.50	385704.25	614294.88
IdClavecct	18176.00	2623.2'	1668.64	1200.00	1484.00	1534.00	4466.00	8019.00
SexoId	18176.00	0.48	0.50	0.00	0.00	0.00	1.00	1.00
IdMunicipio	18176.00	12.68	3.41	6.50	11.00	14.00	14.00	18.00
Region	18176.00	3.55	0.71	1.50	3.00	4.00	4.00	4.00
Zona	18176.00	5.90	3.20	1.00	3.00	6.00	9.00	11.00
Faltas_ANTE	18176.00	0.30	0.77	-0.38	-0.38	0.00	0.62	2.12
Faltas_PASA	18176.00	0.23	0.72	-0.42	-0.42	0.00	0.58	2.08
HISTOR_ANTE	18176.00	7.89	1.17	5.00	7.00	7.90	8.80	10.00
FORMACIANTE	18176.00	7.69	1.22	5.00	6.70	7.60	8.70	10.00
ESPANO_ANTE	18176.00	7.95	1.15	5.00	7.00	8.00	8.90	10.00
MATEMATANTE	18176.00	8.03	1.16	5.00	7.10	8.00	9.00	10.00
CIENCIAANTE	18176.00	8.90	0.88	6.60	8.40	9.00	9.60	10.00
GEOGRAFANTE	18176.00	8.54	1.11	5.25	7.80	8.70	9.50	10.00
INGLES_ANTE	18176.00	8.27	1.13	5.00	7.40	8.40	9.20	10.00
ESPANIIPASA	18176.00	7.91	1.27	5.00	7.00	8.00	9.00	10.00
MATEMIIPASA	18176.00	7.60	1.31	5.00	6.60	7.60	8.60	10.00
CIENCIIPASA	18176.00	7.71	1.30	5.00	6.60	7.60	8.60	10.00
HISTORIPASA	18176.00	7.92	1.30	5.00	7.00	8.00	9.00	10.00

Tabla 11a

Estadística descriptiva (1/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

En la tabla 12a se puede ver el resultado de la aplicación de la transformación de los datos, el Coeficiente de Variación ha disminuido en la mayoría de las variables. Por consecuencia se reducen las variables que tienen el estatus de Alta variabilidad, el sesgo se modifica en la mayoría de las variables y queda en casi todas en simétrica. Los valores atípicos se reducen llegando en algunos casos a 0. La mediana en su mayoría se mantiene al igual que los valores únicos y los valores del % Dominante.

Como se puede ver el tratamiento dio un buen resultado en las variables esto se ve claramente en el sesgo.

Estadística enriquecida de variables numéricas transformadas

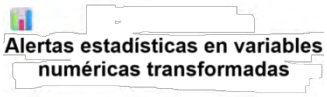
Variable	CV	Utilidad	Sesgo	Atípicos	% Nulos	IQR	Valores únicos	Mediana	% Dominante
IdAlumno	0.51	Útil	Simétrico	-0.38	0.00	152393.75	16918	292810.50	6.93
IdClavecct	0.64	Útil	Derecha	-0.29	0.00	2982.00	220	1534.00	1.58
SexoId	1.05	Alta variabilidad	Simétrico	-1.99	0.00	1.00	2	0.00	52.26
IdMunicipio	0.27	Útil	Izquierda	-0.54	0.00	3.00	11	14.00	51.68
Region	0.20	Útil	Izquierda	0.55	0.00	1.00	4	4.00	67.04
Zona	0.54	Útil	Simétrico	-1.26	0.00	6.00	11	6.00	13.13
Faltas_ANTE	2.62	Alta variabilidad	Derecha	0.22	0.00	1.00	22	0.00	32.74
Faltas_PASA	3.08	Alta variabilidad	Derecha	0.48	0.00	1.00	32	0.00	25.89
HISTOR_ANTE	0.15	Útil	Simétrico	-0.95	0.00	1.80	56	7.90	3.58
FORMACIANTE	0.16	Útil	Simétrico	-0.98	0.00	2.00	53	7.60	4.87
ESPANO_ANTE	0.14	Útil	Simétrico	-0.89	0.00	1.90	55	8.00	3.46
MATEMATANTE	0.14	Útil	Simétrico	-0.95	0.00	1.90	55	8.00	3.61
CIENCIAANTE	0.10	Descartable	Izquierda	-0.15	0.00	1.20	37	9.00	10.39
GEOGRAFANTE	0.13	Útil	Izquierda	-0.56	0.00	1.70	50	8.70	8.69
INGLES_ANTE	0.14	Útil	Simétrico	-0.80	0.00	1.80	50	8.40	5.01
ESPANIIPASA	0.16	Útil	Simétrico	-1.01	0.00	2.00	16	8.00	8.33
MATEMIIPASA	0.17	Útil	Simétrico	-1.03	0.00	2.00	18	7.60	11.31
CIENCIIPASA	0.17	Útil	Simétrico	-1.03	0.00	2.00	16	7.60	9.26
HISTORIPASA	0.16	Útil	Simétrico	-1.03	0.00	2.00	16	8.00	8.68

Tabla 12a

Estadística enriquecida (1/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

En la tabla 13 a se muestra el resultado de la aplicación de la transformación en las variables de tener casi todas descripciones de alertas, en esta tabla ya no existen.



Alertas estadísticas en variables numéricas transformadas

Variable	Alerta
IdAlumno	Sin alerta
IdClavecct	Sin alerta
SexoId	Sin alerta
IdMunicipio	Sin alerta
Region	Sin alerta
Zona	Sin alerta
Faltas_ANTE	Alta dispersión
Faltas_PASA	Alta dispersión
HISTOR_ANTE	Sin alerta
FORMACIANTE	Sin alerta
ESPANO_ANTE	Sin alerta
MATEMATANTE	Sin alerta
CIENCIAANTE	Sin alerta
GEOGRAFANTE	Sin alerta
INGLES_ANTE	Sin alerta
ESPANIIPASA	Sin alerta
MATEMIIPASA	Sin alerta
CIENCIIPASA	Sin alerta
HISTORIPASA	Sin alerta

Tabla 13a

Alertas estadísticas (1/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

Al igual que la primera parte de las variables en la tabla 11a se muestra que los valores de la media, la desviación estándar, el mínimo y el primer cuartil se modifican.

Estadística de variables numéricas tranformadas

Variable	count	mean	std	min	25%	50%	75%	max
FORMAIIPASA	18176.00	7.99	1.30	5.00	7.00	8.00	9.00	10.00
INGLEIIPASA	18176.00	7.82	1.28	5.00	6.60	8.00	9.00	10.00
EDUCACIPASA	18176.00	8.80	1.03	5.60	8.00	9.00	9.60	10.00
ARTES__PASA	18176.00	8.38	1.28	5.00	7.30	8.60	9.60	10.00
AMBIT__PASA	18176.00	8.34	1.25	5.00	7.50	8.50	9.30	10.00
ESCUELALAT	18176.00	20.58	0.11	20.37	20.53	20.60	20.64	20.80
ESCUELALON	18176.00	-100.24	0.25	-100.50	-100.41	-100.39	-99.99	-99.37
escPriv	18176.00	0.30	0.46	0.00	0.00	0.00	1.00	1.00
escPub	18176.00	0.68	0.46	0.00	0.00	1.00	1.00	1.00
bimRepr_ANTE	18176.00	0.58	0.87	0.00	0.00	0.00	1.00	2.50
bim6_8_ANTE	18176.00	13.22	9.22	0.00	5.00	13.00	21.00	45.00
bim8_10_ANTE	18176.00	25.21	10.76	0.00	16.00	26.00	35.00	63.50
bimRepr_PASA	18176.00	0.65	0.96	0.00	0.00	0.00	1.00	2.50
bim6_8_PASA	18176.00	7.72	6.00	0.00	2.00	7.00	13.00	26.00
bim8_10_PASA	18176.00	15.23	7.03	0.00	9.00	16.00	22.00	24.00
PROMEDIO_ANTE	18176.00	8.18	0.92	5.31	7.47	8.20	8.91	10.00
PROMEDIO_PASA	18176.00	8.05	1.02	5.11	7.24	8.04	8.88	10.00
TipoMarginacion	18176.00	1.45	0.74	0.00	1.00	1.00	2.00	3.50
ReproboGrado3	18176.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00

Tabla 11b

Estadística descriptiva (2/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

En la siguiente tabla se puede ver que existe unas pequeñas mejoras en las estadísticas. En las que se ve el cambio es en la de Utilidad y en la de Sesgo, la mayoría son útiles y el sesgo es casi todas en simétrico.

Estadística enriquecida de variables numéricas transformadas

Variable	CV	Utilidad	Sesgo	Atípicos	% Nulos	IQR	Valores únicos	Mediana	% Dominante
FORMAIIPASA	0.16	Útil	Simétrico	-1.05	0.00	2.00	16	8.00	9.40
INGLEIIPASA	0.16	Útil	Simétrico	-1.03	0.00	2.40	17	8.00	8.87
EDUCACIPASA	0.12	Útil	Izquierda	-0.19	0.00	1.60	14	9.00	20.73
ARTES__PASA	0.15	Útil	Simétrico	-0.84	0.00	2.30	16	8.60	16.65
AMBIT__PASA	0.15	Útil	Simétrico	-0.71	0.00	1.80	21	8.50	15.65
ESCUELALAT	0.01	Descartable	Simétrico	-0.05	0.00	0.11	183	20.60	4.12
ESCUELALON	-0.00	Descartable	Derecha	0.01	0.00	0.41	194	-100.39	3.09
escPriv	1.51	Alta variabilidad	Derecha	-1.28	0.00	1.00	2	0.00	69.55
escPub	0.68	Útil	Izquierda	-1.37	0.00	1.00	2	1.00	68.47
bimRepr_ANTE	1.50	Alta variabilidad	Derecha	0.23	0.00	1.00	6	0.00	58.97
bim6_8_ANTE	0.70	Útil	Simétrico	-1.05	0.00	16.00	43	13.00	8.31
bim8_10_ANTE	0.43	Útil	Simétrico	-1.07	0.00	19.00	43	26.00	8.02
bimRepr_PASA	1.48	Alta variabilidad	Derecha	-0.58	0.00	1.00	4	0.00	64.39
bim6_8_PASA	0.78	Útil	Simétrico	-1.11	0.00	11.00	26	7.00	15.65
bim8_10_PASA	0.46	Útil	Simétrico	-1.20	0.00	13.00	25	16.00	15.48
PROMEDIO_ANTE	0.11	Útil	Simétrico	-0.86	0.00	1.44	779	8.20	0.57
PROMEDIO_PASA	0.13	Útil	Simétrico	-0.96	0.00	1.63	905	8.04	0.49
TipoMarginacion	0.51	Útil	Derecha	0.59	0.00	1.00	5	1.00	66.23
ReproboGrado3	4.22	Alta variabilidad	Derecha	13.87	0.00	0.00	2	0.00	94.69

Tabla 12b

Estadística enriquecida (2/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

En la tabla 13b se muestra el resultado de la aplicación de la transformación. La alerta en todas es Sin Alerta excepto en la variable objetivo que por su naturaleza a esa variable no se le hace ninguna modificación.

Alertas estadísticas de variables numéricas transformadas

Variable	Alerta
FORMAIIPASA	Sin alerta
INGLEIIPASA	Sin alerta
EDUCACIPASA	Sin alerta
ARTES__PASA	Sin alerta
AMBIT__PASA	Sin alerta
ESCUELALAT	Sin alerta
ESCUELALON	Sin alerta
escPriv	Alta dispersión
escPub	Sin alerta
bimRepr_ANTE	Sin alerta
bim6_8_ANTE	Sin alerta
bim8_10_ANTE	Sin alerta
bimRepr_PASA	Sin alerta
bim6_8_PASA	Sin alerta
bim8_10_PASA	Sin alerta
PROMEDIO_ANTE	Sin alerta
PROMEDIO_PASA	Sin alerta
TipoMarginacion	Sin alerta
ReproboGrado3	Alta redundancia, Sin variación (IQR=0), Atípicos alta (outliers)

Tabla 13b

Alertas estadísticas (2/2) de variables numéricas transformadas del conjunto de datos escolares

Nota. Elaboración propia con base en datos del ciclo escolar analizado.

Después de hacer el tratamiento en las variables continúa la selección de las variables significativas, esto es, las que tienen más poder predictivo para el modelo.

En la figura 34 del mapa de calor (matriz de calor de Pearson, valores de -1 a 1) de correlación se puede ver la alta relación que existe en gran número de sus variables. Los cuadros que tienen los colores más intensos son las variables que tienen más relación, el rojo es una correlación negativa y el azul es la correlación positiva.

Esto quiere decir que, si dos variables están relacionadas, cuando una se incremente o disminuya la otra también lo hará en proporción de su relación.

La gráfica muestra en cada cuadro el color, valor y el sentido de su relación. Por ejemplo, la variable PROMEDIO_PASA está fuertemente relacionado con 7 variables bimRepr_ANTE, bim6_8_ANTE, bim8_10_ANTE, bimRepr_PASA, bim6_8_PASA, bim8_10_PASA y PROMEDIO_ANTE, esto quiero decir que, si se incrementa la variable bimRepr_ANTE se disminuye la variable PROMEDIO_PASA, si se incrementa la variable bim8_10_PASA, se incrementa el variable PROMEDIO_PASA.

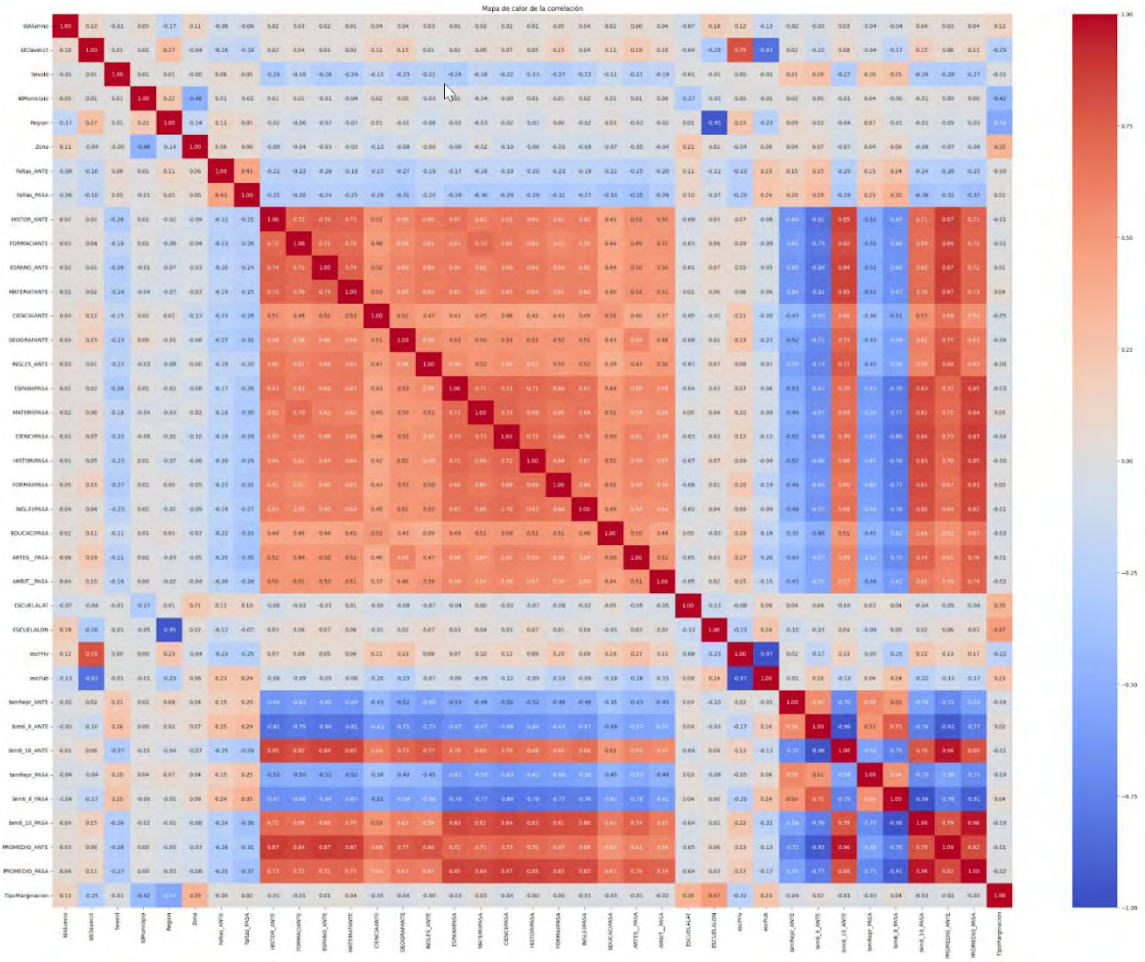


Figura 34
Mapa de calor de correlaciones entre variables numéricas del conjunto de datos escolares
Nota. Elaboración propia.

Después de identificar la correlación entre variables se dejan las más significativas. Se muestra la tabla 14 con la lista de las variables.

Variables Significativas

Variable	Columnas
0	SexoId
1	Region
2	Zona
3	Faltas_ANTE
4	Faltas_PASA
5	escPriv
6	PROMEDIO_ANTE
7	PROMEDIO_PASA
8	TipoMarginacion
9	MateriasRep_ANTE
10	MateriasRep_PASA
11	ReproboGrado3

Tabla 14
Variables significativas seleccionadas para el análisis predictivo
Nota. Elaboración propia.

En resumen, se identificaron variables con alta redundancia, sin variabilidad, con valores atípicos, con alta dispersión y alta correlación.

Se eliminaron registros de calificaciones con valor de 0 al igual que las que tenían rango intercuartil de 0. Se les aplicó tratamiento a las variables que tenían una alta correlación identificadas en la gráfica de calor de Pearson.

A las variables que tenían una alta dispersión se les aplicó una técnica de escalamiento de valores llamada RobustScaler.

El resultado fue un dataset con 12 variables y 18,176 registros.

3.2.4 Metodología para el objetivo específico 4

3.2.4.1 Entrenamiento de los modelos:

Para atender al objetivo O4, en esta fase se aplicaron técnicas de aprendizaje automático y profundo para encontrar el modelo más estable y robusto que tengan el mejor desempeño en sus métricas para clasificar a los estudiantes que están en riesgo de reprobación escolar.

Una vez que al set de datos se le aplicaron técnicas de tratamiento y depuración de información en la fase anterior, se tomó la data resultante con las variables más significativas, en este caso son 12 y 18,176 registros. De acuerdo al problema y a los datos que se presentaron se utilizó el aprendizaje supervisado ya que se tiene una variable objetivo o etiqueta que en el set de datos se conoce como ReproboGrado3, como se definió en la fase 2 en esta variable se identifican si el estudiante reprobó (clase 1) o no reprobó (clase 0), las demás variables se utilizaron para entrenar los modelos.

Se utilizaron 7 modelos para el entrenamiento, 6 modelos clásicos y 1 de aprendizaje profundo (Deep Learning): Red Neuronal Artificial (Keras, aprendizaje profundo), Ensamble, LightGBM, Red Neuronal (SKL), Random Forest, XGBoost y Regresión logística.

Para todos los modelos se utilizó el dataset con 18,176 registros y 11 variables además de la variable objetivo.

Al inicio se prepararon las variables en conjuntos de datos diferentes y se clasificaron las variables por tipo de datos, escalares, categóricos y otros. Se construyó un set de procesamiento (ColumTransformer) en el que se configura para estandarizar su escalar y a las variables categóricas se les asigna un valor numérico. Este paso no se realizó en la Red Neuronal (Keras).

Para el caso específico del modelo de Ensamble en esta parte se definieron los modelos que se integraron en un método que se llama Votación Blanda (Soft Voting), en este caso fueron Regresión Logística, Random Forest y XGBoost.

Se construyó un pipeline (flujo de procesamiento encapsulado) con el objetivo de que los procesos se realicen en una forma ordenada, se balanceo la

data con la técnica de SMOTE (Synthetic Minority Oversampling Technique), se seleccionaron las mejores variables de acuerdo a cada modelo y se configuró el modelo con el que se va a entrenar. Este paso no se realizó en la Red Neuronal (Keras).

La data se dividió en un set de para el entrenamiento con el 80% de los datos y el 20% restante para pruebas, este conjunto de datos nos servirá para mostrar cómo se comporta el con información real una vez entramado y se realizará por medio de gráficas específicas para ver su rendimiento.

En esta parte para el modelo específico de la Red Neuronal (Keras) se realizó el balance de la data con la técnica de SMOTE (Synthetic Minority Oversampling Technique) y se configuró la arquitectura de la red.

Para obtener un mejor rendimiento de los modelos se utilizó una búsqueda de hiperparámetros utilizando validación cruzada de 5 pliegues. También, se definieron las diferentes métricas que se utilizaron en el entrenamiento para hacer la evaluación de los modelos y poder definir el que mejor desempeño obtuvo. Estas son las métricas que se definieron: Recall, F1–score, Precision, PR AUC, Mejora vs azar, MCC, Kappa, Balanced Accuracy y ROC AUC.

Una vez definidos los parámetros para cada modelo se entrenaron y se generó la información para poder hacer el análisis de sus métricas.

En la tabla 15 se pueden ver los valores para cada métrica por modelo. Se hizo el análisis de los 3 modelos que tuvieron las métricas más sobresalientes de manera general: Red Neuronal (Keras), Ensamble y LightGBM.

	Modelo	F1-score	Precision	Recall	MCC	Kappa	Balanced Acc	ROC AUC	PR AUC	Mejora vs azar
0	Red Neuronal (Keras)	0.3855	0.3427	0.4404	0.3591	0.3130	0.7786	0.8502	0.2660	5.3814
1	Ensamble	0.3699	0.2761	0.5602	0.3456	0.3216	0.7389	0.8854	0.2888	5.4410
2	LightGBM	0.3520	0.2474	0.6107	0.3364	0.2989	0.7531	0.8787	0.3218	6.0616
3	Red Neuronal SKL	0.3332	0.2379	0.5614	0.3111	0.2794	0.7296	0.8513	0.2475	4.6637
4	Random Forest	0.3301	0.2250	0.6197	0.3173	0.2735	0.7500	0.8710	0.3141	5.9168
5	XGBoost	0.3126	0.3391	0.2911	0.2784	0.2772	0.6296	0.8768	0.2636	4.9668
6	Logistic Regression	0.2762	0.1661	0.8202	0.3003	0.2060	0.7945	0.8742	0.2888	5.4410

Tabla 15
Comparativo de desempeño de modelos de clasificación para predecir reprobación en tercer grado de secundaria
Nota. Elaboración propia.

Modelo de Red Neuronal (Keras)

La red neuronal que se utilizó fue una Red Neuronal Artificial (ANN por sus siglas en inglés) de tipo feedforward, específicamente es una arquitectura densa (fully connected).

La arquitectura que se utilizó para entrenar el modelo es como se muestra en la tabla. Consta de 5 capas y se diseñó en forma de embudo teniendo en la capa superior 128 neuronas y en la última 1 neurona. Se utilizó la función de activación ReLu para algunas neuronas en la capa ocultas y para la capa de salida la función sigmoide. Se utilizó la normalización por lotes en 3 neuronas de la capa oculta. Dropout es una técnica de regularización, activa y desactiva neuronas en el entrenamiento con el porcentaje que se muestra en la tabla.

Arquitectura de la Red Neuronal Configurada

Capa	Tipo	Neurona(s)/Salida	Activación	Regularización
dense	Dense	128	relu	—
batch_normalization	BatchNormalization	—	—	BatchNormalization
dropout	Dropout	—	—	Dropout (30%)
dense_1	Dense	64	relu	—
batch_normalization_1	BatchNormalization	—	—	BatchNormalization
dropout_1	Dropout	—	—	Dropout (20%)
dense_2	Dense	32	relu	—
batch_normalization_2	BatchNormalization	—	—	BatchNormalization
dropout_2	Dropout	—	—	Dropout (10%)
dense_3	Dense	8	relu	—
dense_4	Dense	1	sigmoid	—

Tabla 16
Arquitectura de la red neuronal densa configurada para la predicción de reprobación escolar
Nota. Elaboración propia.

En las figuras 35 y 36 se muestran las gráficas de Evolución de las Métricas representan cómo se comportó el modelo en el entrenamiento y se muestran las principales.

La evolución de Accuracy en la primera época inicia en 0.65 en la primera época, asciende rápidamente y empieza a disminuir poco a poco hasta que se mantiene en el 0.80. en la época 11. El val_accuracy tiene el mismo comportamiento, pero en una diferencia menor a 0.1 en las 11 épocas. Lo que indica que el modelo generaliza bien y no tiene sobreajuste.

La evaluación de Precision asciende rápidamente de 0.5 a 0.7 y a partir de la época 3 se mantiene estable en 0.72 aproximadamente. La validación se mantiene constante en 1 durante todas las épocas. Indica que el modelo es conservador al predecir, quiere decir que cuando lo hace, acierta.

La gráfica de la Evaluación de Recall desciende rápidamente en la época 1 de 0.85 a 0.72 en la época 3, se estabiliza en el valor de 0.75 aproximadamente y se mantiene hasta la época 11. La validación inicia en 0.55 asciende gradualmente hasta llegar 0.75 en la época 11. Indica que conforme se incrementa las iteraciones

el modelo está aprendiendo e identifica correctamente a los alumnos que están en riesgo de reprobación.

La evaluación de Loss y val_loss tiene una curva similar iniciando en valores altos en la primera época y van disminuyendo gradualmente conforme se incrementan las épocas.

Como se ha mencionado en varias ocasiones la métrica de F1–score la conforman la combinación de la precisión y el recall, muestra el equilibrio entre esas dos métricas. Su curva al igual que la de val_f1–score al igual que las anteriores gráficas muestran consistencia en sus valores, aquí los valores de la validación son más altos, lo que indica que existe un buen equilibrio entre las métricas, el valor que alcanza F1–score es de 0.8691 en la época 11.

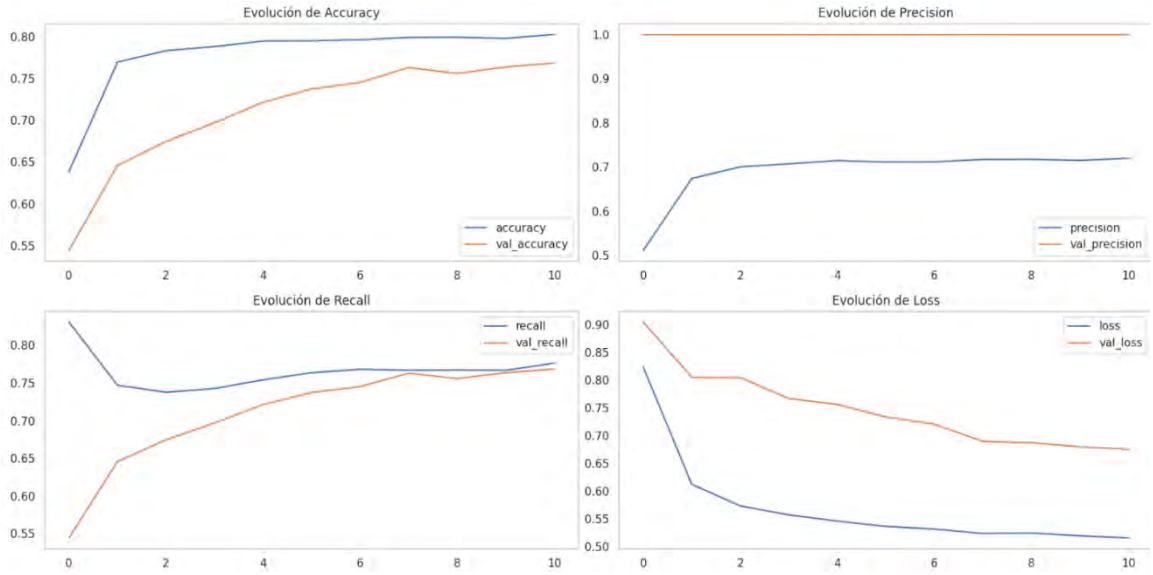


Figura 35
Evolución de métricas de desempeño de la red neuronal durante el entrenamiento
Nota. Elaboración propia.

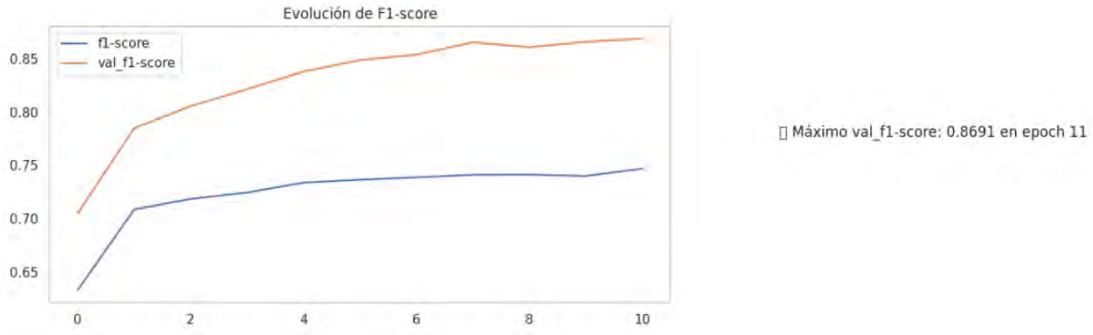


Figura 36
Evolución del F1-score durante el entrenamiento de la red neuronal
Nota. Elaboración propia.

De las métricas, a la arquitectura de la red y a la evolución de las métricas se puede ver que el modelo en lo general tiene un buen desempeño, las técnicas usadas en su configuración lo hacen que sea robusto y estable. La evolución progresiva que se muestran en las curvas de Recall y F1–score indican que el modelo puede ser útil para usarse como modelo predictivo de alumnos que están en riesgo de reprobar tercer grado de secundaria.

Modelo de Ensamble

El proceso de entrenamiento de un modelo de aprendizaje profundo y uno modelo clásico es un poco diferente, en el caso del modelo de Ensamble se integraron 3 modelos, XGBoost, Random Forest y Regresión Logística. Se optimizó el modelo, se usó validación cruzada y un pipeline. Las curvas que se ven son estáticas y no un proceso secuencial donde se van guardando los valores conforme se van iterando las épocas como en las redes neuronales.

La gráfica de Accuracy muestra 0 en val_accuracy durante la combinación de evaluaciones. Esto indica que el modelo prioriza otras métricas y este valor puede estar dado por el desbalance de los datos.

La Evolución de Precision asciende rápidamente y tiene un valor máximo de 0.276, lo que quiere decir que de cada 100 predicciones de reprobación que

realiza solo 27 son correctas. Este valor es subjetivo si se toma de manera individual, dependiendo del tipo de problema y el contexto completo de la situación, se le dará el justo valor al combinarse con otras métricas, por ejemplo, para el caso de reprobación donde lo más importante es tener una detección temprana de alumnos en riesgo este valor de precisión es aceptable.

La Evolución de Recall inicia en 0.548 y tiene un valor máximo de 0.56 en la primera combinación lo hace en una etapa temprana. Esto representa que de cada 100 estudiantes que reprueban 56 casos lo hacen de manera correcta. Para un caso con el desbalance que tenemos en el set de datos es aceptable ya que está por encima del azar.

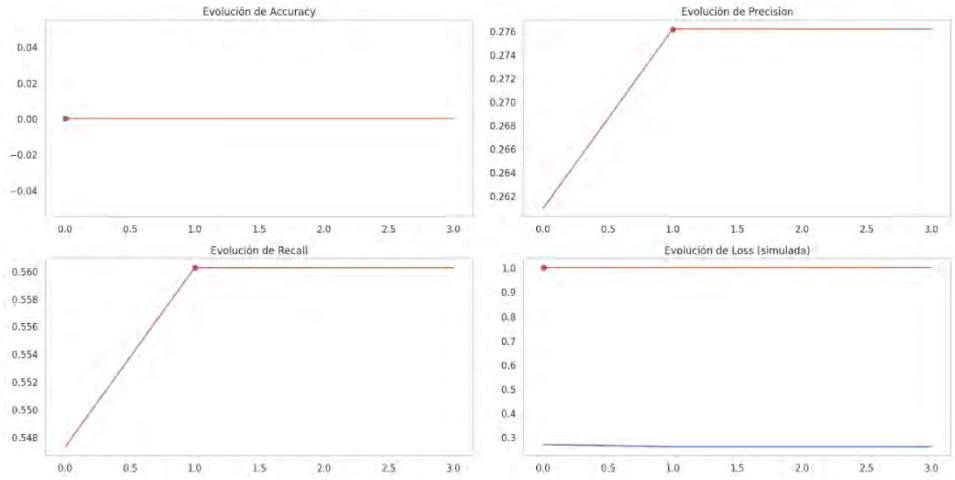


Figura 37
Evolución de métricas durante el entrenamiento de un modelo con desempeño limitado
Nota. Elaboración propia.

En la curva de Loss (simulada) se mantiene constante en 1, este valor puede ser un error ya que si vemos el valor que se generó en el entrenamiento es de 0.27.

En la figura 38 se muestra que para la gráfica de F1–score se tiene un valor máximo de 0.3699 que se alcanza en la primera combinación, refleja un buen equilibrio entre la precision y el recall, eso es importante para el tipo de modelos que tenemos.

La evolución de MCC (Matthews Correlation Coeffcient) muestra un valor de 0.3456 para la combinación 1, este un valor bueno, ya que está por arriba del azar que es 0 y 1 el máximo valor.

La Evolución de Cohen's Kappa tiene un valor máximo de 0.3216 en la primera combinación, ese valor indica que el modelo tiene un buen poder de predictivo y más considerando el desbalance que se tiene. Los valores para está métrica oscilan entre -1 identifica un modelo muy malo 0 para un modelo que predice igual al azar y 1 para un modelo excelente. En este caso su valor está por encima del azar.

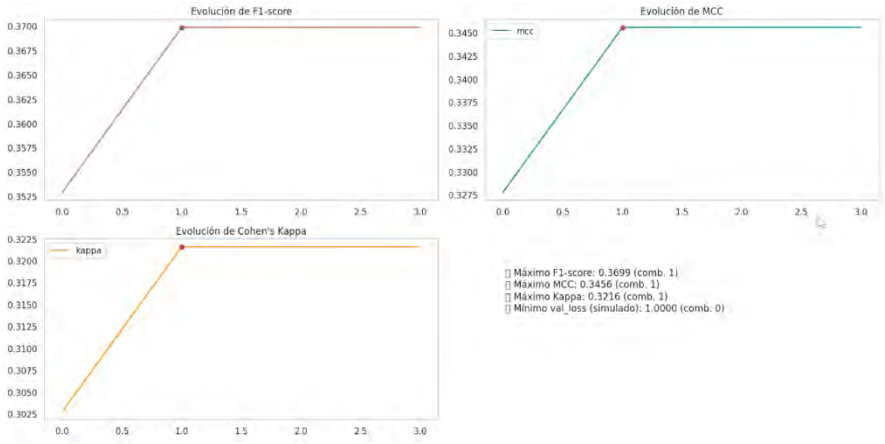


Figura 38
Evolución de F1-score, MCC y Kappa en un modelo con bajo número de combinaciones
Nota. Elaboración propia.

Podemos resumir que las métricas importantes como F1–score, MCC y Kappa tienen valores aceptables considerando el desbalance de las clases. La consistencia en las métricas refuerza el resultado que indica que tiene un buen balance, es estable y tiene una buena capacidad de discriminación, principalmete lo hace por arriba del azar.

En la figura 39 se muestra el rendimiento conforme se incrementa el número de variables. Recordemos que para los modelos de aprendizaje clásico se utilizaron técnicas de transformación, sobre todo para las variables categóricas. El inicio del

modelado se empezó con una data con 12 variables y esta se puede incrementar notablemente conforme se aplican técnicas como la de transformación. En la gráfica se tiene en el eje Y los valores para la métrica f1–score y en el eje de las X se tiene el número de variables. Se puede ver que el valor inicial es de 0.37 cuando inicia el entrenamiento y desciende a un poco más de 35 con diez variables. Después empieza incrementarse el valor hasta llegar a su máximo con 15 variables y se mantiene hasta la variable 20. Lo que quiere decir que de la variable 16 en adelante no aportan el modelo o porque son redundantes.

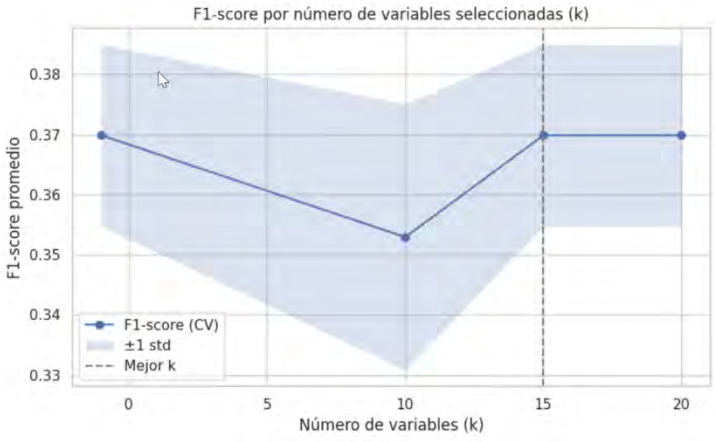


Figura 39
Desempeño promedio del modelo según el número de variables seleccionadas
Nota. Elaboración propia.

Se puede concluir que el modelo de ensamble es un modelo estable y robusto que tiene un rendimiento equilibrado y una capacidad de generalización adecuada para problemas educativos. Tiene un buen desempeño en sus métricas sobre todo en la F1–score, MCC y Kappa que son métricas importantes para una data con un desbalance alto.

Modelo LightGBM

El Modelo LightGBM (Light Gradient Boosting Machine) es muy útil para modelos que tienen un alto nivel de desbalance en sus clases.

Como se mencionó el modelo se configuró por medio de pipeline, con balanceo y transformación de datos, se aplicó validación cruzada buscando los mejores hiperparámetros. De acuerdo a las métricas de la tabla de métricas de los modelos se pueden ver que los valores que tiene este modelo están por encima de los demás modelos, a continuación, se analizarán gráficas para ver su desempeño que se muestran en la figura 40.

Se puede ver que la gráfica de Accuracy la métrica no se registró y tiene un valor de 0, probablemente debido al desbalance de clases que existe.

En la gráfica de precisión se puede ver que existen variaciones importantes en los diferentes parámetros, es una curva accidentada llegando a su valor máximo en la combinación 7 con un valor aproximado de 0.34.

La gráfica de Evolución de Recall igualmente muestra una curva errática, inicia con el valor por arriba del 6 y llega a su punto máximo en la combinación 1 y tiene un valor alrededor de 0.75. Existe un descenso importante en la combinación 7 que puede ser consecuencia del máximo valor en esa combinación en la gráfica de precision. Estas dos métricas están muy relacionadas.

La curva de val_loss (simulada) se mantiene en una línea continua en 1, puede ser un error por el desbalance de las clases. Loss tiene un comportamiento normal con un pequeño sobresalto en la combinación 7.

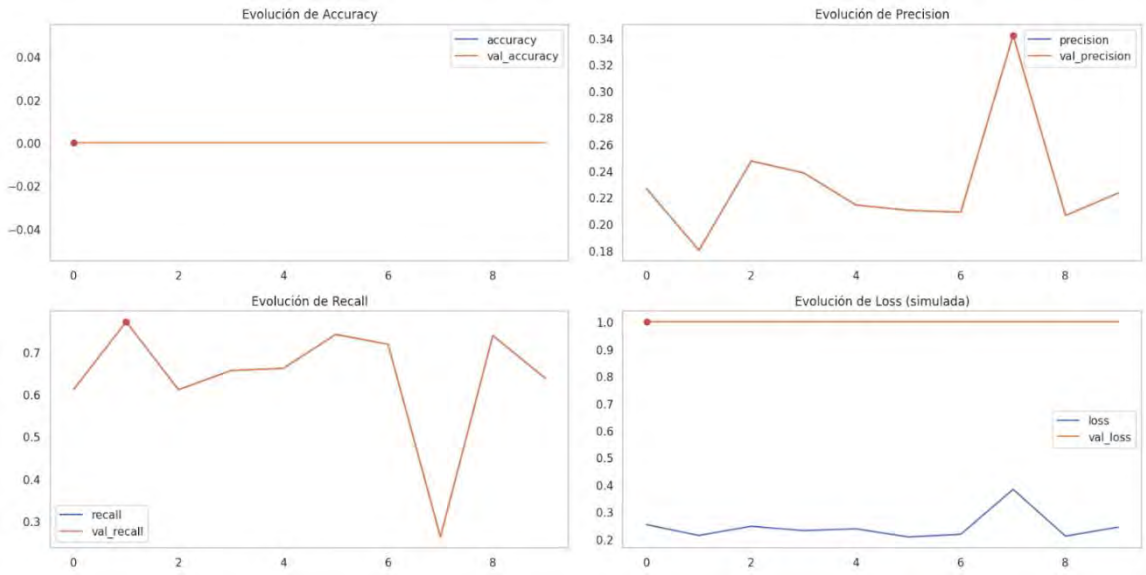


Figura 40
Evolución de métricas (Accuracy, precicison, recall, loss) de desempeño del modelo LightGBM
Nota. Elaboración propia.

La figura 41 muestra diferentes gráficas. La gráfica de Evolución de F1–score muestra una curva irregular que inicia con un descenso pronunciado al nivel más bajo en la combinación 1 y de manera inversa sube abruptamente al máximo valor en la combinación 2 con el valor de 0.3520. La forma de las curvas en las diferentes gráficas indica que el modelo es sensible a las combinaciones de hiperparámetros.

La evolución de MCC muestra su máximo valor en la combinación 3 con 0.3417. Esta métrica mide la correlación entre las predicciones y los valores reales, tomando en cuenta los valores de la matriz de confusión.

En la Evolución Cohen´s Kappa el valor máximo es el de 0.2989 en la combinación 2. Igualmente tiene una curva irregular que tiene ascensos y descensos abruptos en las primeras combinaciones.

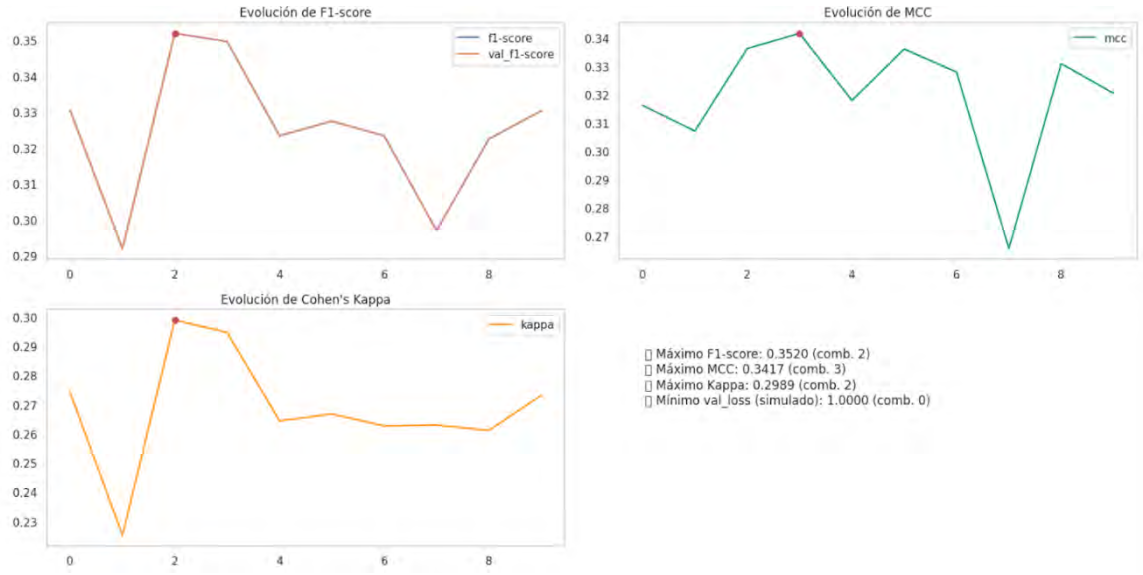


Figura 41
Evolución de métricas (Fi–score, MCC, Kappa) por combinación de variables seleccionadas
Nota. Elaboración propia.

Se puede ver que las gráficas presentan una relación entre las curvas, son sensibles a los hiperparámetros y muestran un pico importante en las combinaciones 1 y 7.

En la figura 42 se puede ver la gráfica de Importancia y dirección del impacto de las variables y se muestra las variables significativas en el modelo, están del lado izquierdo. En la imagen se muestran nombres de variables con un prefijo escalado, esto quiere decir que son variables que ya son tratadas en el modelo y el modelo le asigna un nombre para diferenciarlas. También, existen nombres duplicados lo que sugiere que el modelo la manda llamar en diferentes partes del entrenamiento. El orden de las variables representa la influencia que tiene la variable en el modelo. En el eje X se muestran valores positivos y negativos, del lado derecho una escala de tonos que va del color azul (bajo) al color rojo (alto).

En la gráfica se muestran las métricas más influyentes en el modelo son PROMEDIO_PASA y PROMEDIO_ANTE. Estas variables representan los promedios de calificaciones de primer año de secundaria y segundo año de

secundaria, indican que cuanto mayor sea la calificación en esos grados menor será la probabilidad de que repruebe el estudiante.

Las variables de MateriasRep_PASA, Sexp, Zona, escPriv tienen un comportamiento similar entre ellas puede ser que exista relación entre ellas. Lo que indican que entre más materias reprobaron en segundo de secundaria es más probable que reprueben tercer grado.

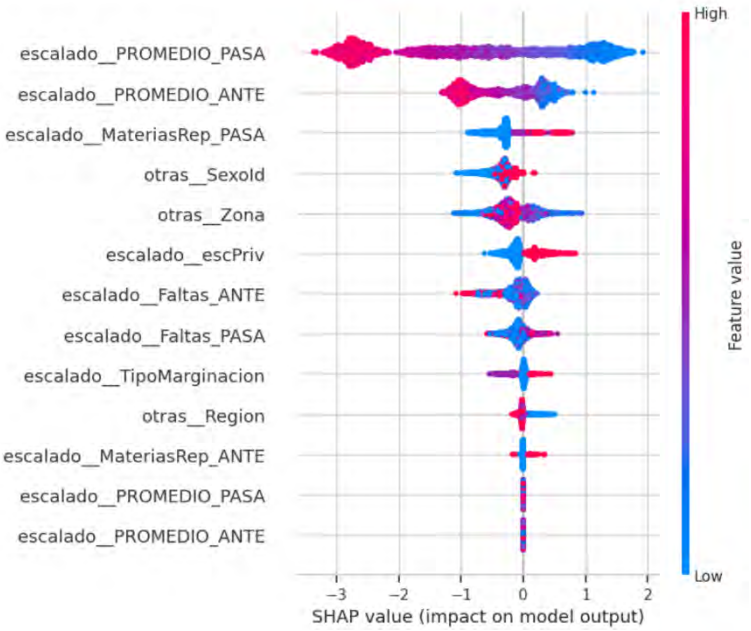


Figura 42
Importancia y efecto de las variables sobre la predicción del modelo (valores SHAP)
Nota. Elaboración propia.

Haciendo un análisis de la tabla de valores de las métricas, de las gráficas de evolución y de la gráfica de variables, se puede ver claramente que el modelo sobresale en varias métricas que se pueden considerar importantes para el caso que se está abordando.

La figura 43 muestra la gráfica de Radar de Métricas por modelo, ahí se evaluaron 4 métricas F1–score, Recall, Precision y PR AUC, que son relevantes para un problema que tiene un desbalance alto, como es el caso. En esta gráfica se

compara el desempeño de los 3 modelos con respecto a las 4 métricas. Para poder hacer una comparativa adecuada los datos se normalizaron entre 0 y 1.

El área que muestra la figura define el modelo que mejor desempeño tiene, los valores que tienen en cada métrica el modelo acerca el vértice del polígono a la orilla del círculo.

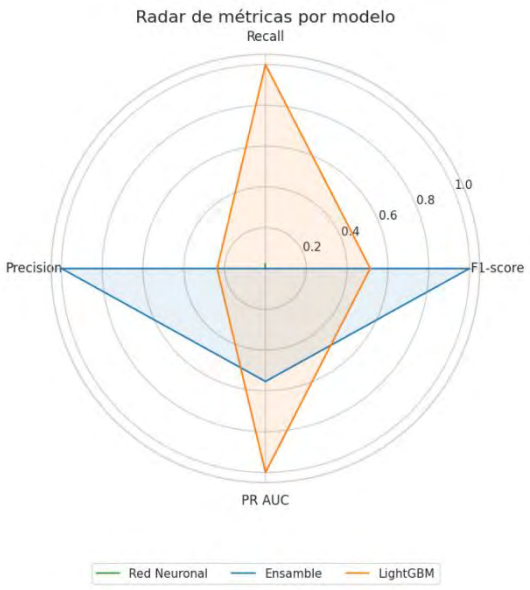


Figura 43
Comparación de modelos mediante radar de métricas normalizadas
Nota. Elaboración propia.

El modelo que se muestra en el centro de la figura como un punto es la Red Neuronal (Keras) y son tan bajos sus valores normalizados que casi es imperceptible. En el caso del modelo de Ensamble se ve que domina las métricas de Precision y F1-score y un poco PR AUC, el valor de Recall es muy bajo. Para el caso del modelo de LightGBM tiene un área más grande y su fortaleza en comparación de los otros modelos son en el Recall y PR AUC que para el tipo de problema de reprobación son las métricas que tienen mayor relevancia. El alargamiento hacia las otras métricas es moderado.

Esta gráfica es determinante para seleccionar el modelo que tiene el mejor desempeño conforme a sus métricas, en este caso es el LightGBM. En la siguiente fase de evaluación se definirá el modelo de acuerdo a otros métodos de selección.

En resumen, en la fase de modelado se entrenaron diferentes modelos de características diferentes, Ensambles, aprendizaje clásico, aprendizaje profundo y un modelo relativamente nuevo como LightGBM. Para cada uno se les aplicaron técnicas particulares de configuración automática o manualmente para encontrar la mejor configuración con el que el modelo fuera más eficiente y tuviera un mejor rendimiento en sus métricas.

De los 7 modelos que se entrenaron se presentaron gráficas adicionales de los 3 que tenían mejores métricas, Redes Neuronales (Keras), Ensamble y LightGBM, con ellas se pudo mostrar cuál es el que tiene las mejores métricas conforme al problema que se está abordando. La última gráfica que se mostró (Radar de Métricas por Modelo) ratifica la decisión documentada de que el mejor modelo es el de LightGBM.

3.2.5 Metodología para el objetivo específico 5

3.2.5.1 Evaluación y selección del modelo:

Conforme el objetivo O5 y una vez entrenado los 7 modelos se realizó la elección del más estable, robusto y el que de manera general obtuvo el mejor desempeño en las principales métricas.

El resultado de la fase anterior generó una tabla con los modelos que se entrenaron para atender el problema de reprobación escolar, esta tiene 9 columnas que son las métricas con las que se evaluó el modelo.

Es importante hacer notar que los valores que se presentan en la tabla son el promedio de los cálculos que se hacen con la validación cruzada (cross–

validation) en el proceso y con los datos de entrenamiento del modelo. Los valores que se presentan en las demás gráficas son resultado de los cálculos que se realizan con el set de datos de pruebas que se designó para este motivo y se comentó en la fase anterior, son datos que el modelo aún no ha visto.

Para hacer la selección del mejor modelo se optó por utilizar la evaluación multicriterio utilizando la técnica WSM (Weighted Sum Model), dando pesos a las métricas de acuerdo a la importancia y particularidad del problema, pero sobre todo al desbalance de las clases que se tienen, dando prioridad a la sensibilidad y a la precisión del modelo. El orden de ponderación es el siguiente:

Recall = 1, F1–score = 2, Precision = 3, PR AUC = 4, Mejora vs azar = 5, MCC = 6, Kappa = 7, Balanced Accuracy = 8 y ROC AUC = 9

Después se ordenaron los valores de la métrica dando 1 al valor más alto y así de forma ascendente hasta el de menor valor que se le dió el número 7, después, ese valor de orden del modelo en cada métrica se multiplicó por el peso de ponderación de la métrica dando como resultado la columna en la tabla de WSM Score. Conforme a los criterios que se definieron el valor más bajo que se tiene en esa columna es el que mejores características tiene y es el valor de 111 que corresponde al modelo LightGBM, se muestra en la tabla 17.

	Modelo	F1-score	Precision	Recall	MCC	Kappa	Balanced Acc	ROC AUC	PR AUC	Mejora vs azar	WSM_Score
2	LightGBM	0.3520	0.2474	0.6107	0.3364	0.2989	0.7531	0.8787	0.3218	6.0616	111
1	Ensamble	0.3699	0.2761	0.5602	0.3456	0.3216	0.7389	0.8854	0.2888	5.4410	113
0	Red Neuronal (Keras)	0.3855	0.3427	0.4404	0.3591	0.3130	0.7786	0.8502	0.2660	5.3814	146
4	Random Forest	0.3301	0.2250	0.6197	0.3173	0.2735	0.7500	0.8710	0.3141	5.9168	191
6	Logistic Regression	0.2762	0.1661	0.8202	0.3003	0.2060	0.7945	0.8742	0.2888	5.4410	192
5	XGBoost	0.3126	0.3391	0.2911	0.2784	0.2772	0.6296	0.8768	0.2636	4.9668	230
3	Red Neuronal SKL	0.3332	0.2379	0.5614	0.3111	0.2794	0.7296	0.8513	0.2475	4.6637	241

Tabla 17
Comparativa de modelos con WSM_Score como criterio de selección
Nota. Elaboración propia.

Como se comentó, la evaluación de los modelos y la comparación entre ellos se realiza con promedios de métricas, por tal motivo las gráficas que veremos

adelante va a variar sus números esto debido a que ellas se hacen tomando en consideración un set de datos de pruebas (test) que se separa al inicio de la ejecución de los modelos, se recomienda que sea de 10% a 30% dependiendo del volumen de datos que se tengan.

En la figura 44 se puede ver el reporte de clasificación muestra el comportamiento de las métricas en las diferentes clases, no reprobó (clase 0) y reprobó (clase 1). Se puede ver que para la clase 0 las métricas son muy buenas, estos valores se deben gran parte al desbalance que tienen los datos cerca del 94%. Como se comentó para el caso de reprobación la métrica que se debe considerar en la de recall, que es la que define los verdaderos positivos (clase 1) y verdaderos negativos (clase 0) que tuvo el modelo, es decir, clasificó correctamente al estudiante que reprobó (clase 0) o al que no reprobó (clase 0).

Para el caso que nos atiende, la reprobación, revisará la clase 1. En este caso en el recall se tiene que el 75% de los alumnos que reprobaron fueron clasificados correctamente, este es un valor alto y nos permite identificar de manera temprana a los alumnos para poderles dar atención y evitar que caigan de manera real en ese escenario.

La precision es la métrica que nos dice de los estudiantes que el modelo clasifica cuántos fueron correctos, tiene un valor de 27%, es decir, un poco más de 1 de cada 4 estudiantes no fueron clasificados correctamente. Para los casos que no se clasifiquen adecuadamente se tendrá que desarrollar programas de validación donde se utilicen filtros con docentes que nos ayudarán a asegurarnos en la confiabilidad del modelo.

F1–score es la métrica que se toma como balance de las otras dos, es la que muestra que tan equilibrado está el modelo, tiene 39% que para tener un desbalance tan elevado se considera aceptable ese porcentaje.

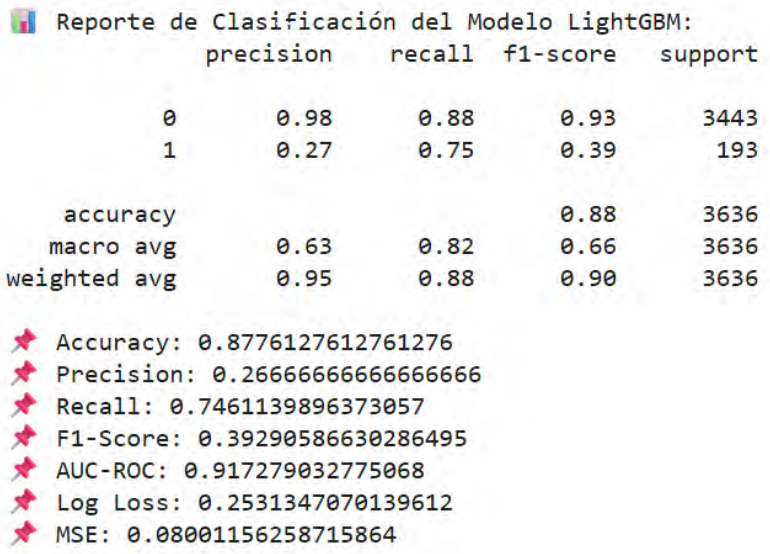


Figura 44
Reporte de clasificación del modelo LightGBM
Nota. Elaboración propia.

AUC–ROC muestra que tan eficiente el modelo separa las clases, estudiantes que no reprobaron de los que reprobaron, en este caso con casi 92% nos dice que está trabajando de una manera eficiente.

Log Loss (Logarithmic Loss o Binary cross–entropy) evalúa la calidad de las probabilidades que predijo el modelo. No mide solo si acertó o no. sino qué confiado estaba el modelo de la predicción y qué tan cerca estuvo de hacerlo. Por ejemplo: si el modelo predice que el alumno no reprobó y tiene una probabilidad del 99% se le da una penalización alta. Por el contrario, si tiene una probabilidad de 55% de reprobar y si reprueba da una penalización baja. En este caso para el modelo tiene un valor de 0.25, es un valor bajo, lo que indica que está trabajando bien.

MSE (Error Cuadrático Medio) tiene el valor de 0.08, nos dice que tan alejada está la probabilidad predicha de los valores reales. En este caso el modelo se equivoca por muy poco.

En la figura 45 se puede ver la matriz de confusión que muestra los valores en los que el modelo clasificó correctamente y en los que se equivocó.

Verdaderos positivos (TN) 3.047, Falsos Negativos (FP) 396, Falsos Negativos (FN) 49 y Verdaderos Positivos (TP) 144.

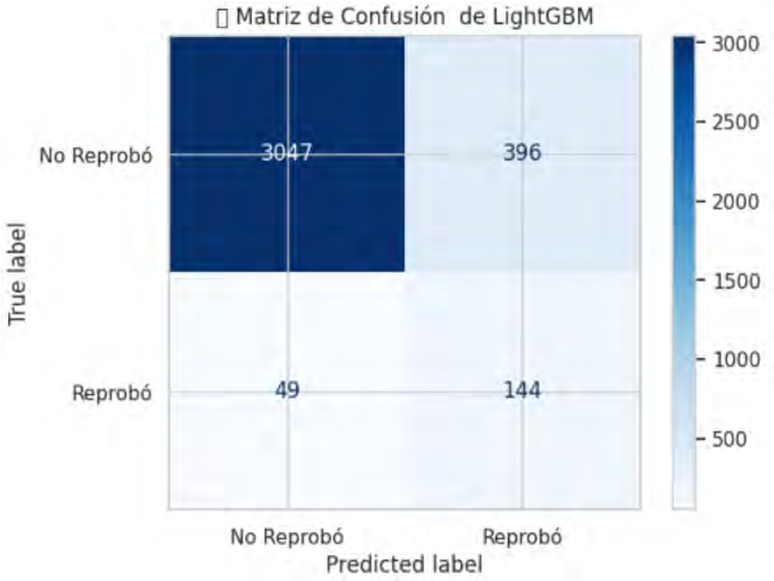


Figura 45
Matriz de confusión del modelo LightGBM
Nota. Elaboración propia.

Estos valores nos arrojan las métricas que se mostraron en el Reporte de Clasificación. Como se pueden ver las mejoras son significativas probando el modelo con datos no vistos, en este caso con el set de pruebas.

Se puede ver en la figura 46 la curva ROC (Receiver Operating Characteristic), muestra la relación que existe entre la Tasa de Verdaderos Positivos (TPR), proporción de casos positivos reales fueron correctamente identificados, y la Tasa de Falsos Positivos (FRP), mide la proporción de los casos negativos reales que fueron identificados como positivos. Como se mencionó muestra la capacidad del modelo de diferenciar las clases, los casos que reprobaron de los que no reprobaron. El valor que tiene esta métrica es muy alto lo que indica que el modelo está trabajando correctamente. Este valor no se debe de tomar como único para el análisis ya que por el desbalance de los datos puede resultar engañoso.

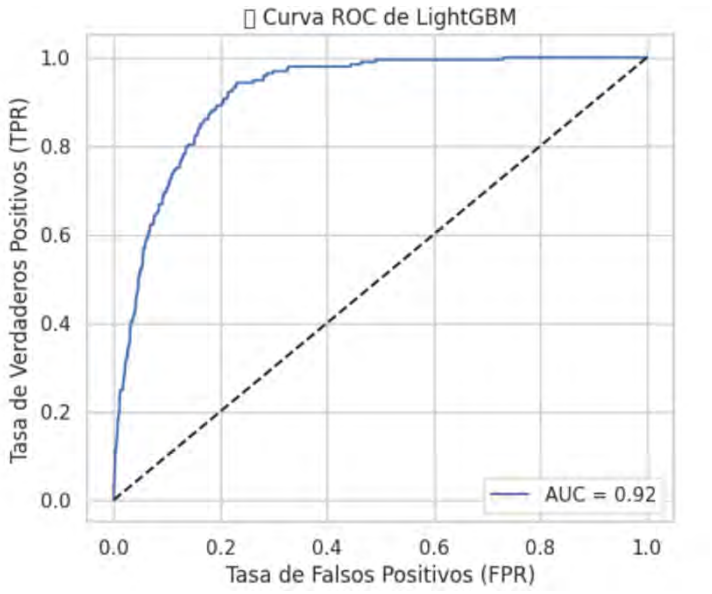


Figura 46
Curva ROC–AUC del modelo LightGBM
Nota. Elaboración propia.

La curva KS (Kolmogorov–Smirnov) se puede ver en la figura 47 y muestra la comparación acumulada de probabilidades de clases, estudiantes que no reprobaron (clase 0) y estudiantes que reprobaron (clase 1)

Conforme avanza el umbral muestra la diferencia entre las Tasas (TPR – FPR), el punto máximo en el umbral en donde encuentra la diferencia más grande.

En la curva se puede ver que la distancia más grande se obtiene en el umbral 0.27 y es el valor de 0.71. Este valor es muy bueno e indica que el modelo separa muy bien las clases.

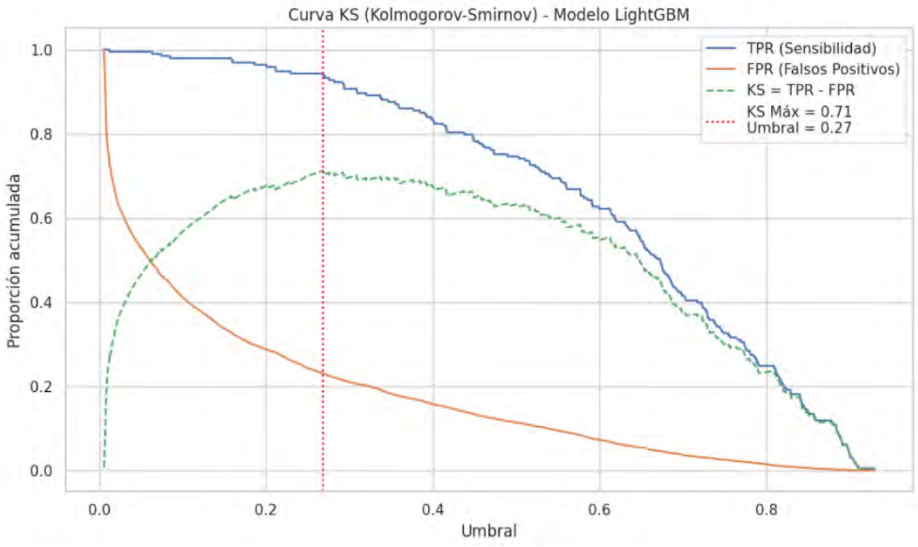


Figura 47
Curva KS (Kolmogorov–Smirnov) del modelo LightGBM
Nota. Elaboración propia.

La curva de Precision–Recall la podemos ver en la tabla 48, se puede decir, que es el indicador más útil cuando la data se encuentra tan desbalanceada, como es el caso que estamos tratando, aproximadamente 5% para la clase 1.

La línea roja punteada muestra el desempeño de un modelo aleatorio, es decir, lo que comúnmente decimos echar un volado, dejarlo al azar, no hacer nada. Para nuestro modelo este valor es aproximadamente 0.05.

La línea azul muestra como es la variación entre precision y recall. Esta curva está enfocada en tratar solo la clase minoritaria en este caso los que reprobaron.

El valor que se tiene es de 0.393 para la curva, es un valor muy alto, representa más de 7 veces el tener un modelo aleatorio. Este valor fortalece la decisión de mantener la propuesta del modelo pues detecta los estudiantes que están en riesgo con una precisión aceptable.

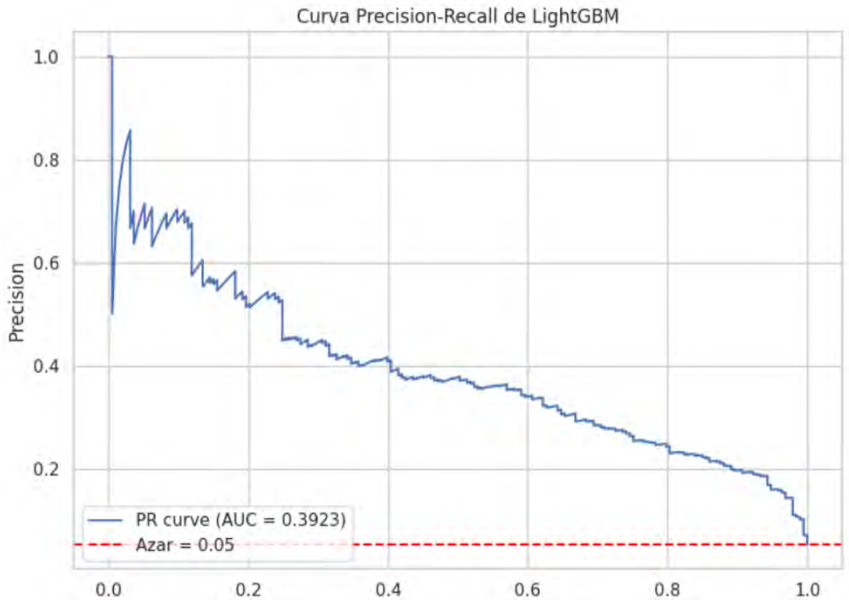


Figura 48
Curva de Precision–Recall (PR-AUC) del modelo LightGBM
Nota. Elaboración propia.

En la figura 49 se puede ver la curva de ganancia acumulada, es una herramienta que nos permite evaluar la eficacia que tiene un modelo al priorizar la población en riesgo, representa el porcentaje acumulado de los casos de la clase minoritaria en este caso los estudiantes que están en riesgo que se pueden detectar sobre un porcentaje creciente de la población con mayor probabilidad.

La gráfica nos dice que, si ordenamos el resultado de alumnos clasificados en riesgo y los ordenamos por orden de probabilidad, si tomamos el 20% de los valores más altos estaremos atendiendo más del 80% de los casos reales. Se puede ver claramente como sube en 20% y en 40% ya se tiene casi el 100%. Esto nos da una ventaja importante porque podemos localizar las intervenciones y dirigir los recursos adecuadamente. Esta gráfica nos muestra que el modelo clasifica adecuadamente.

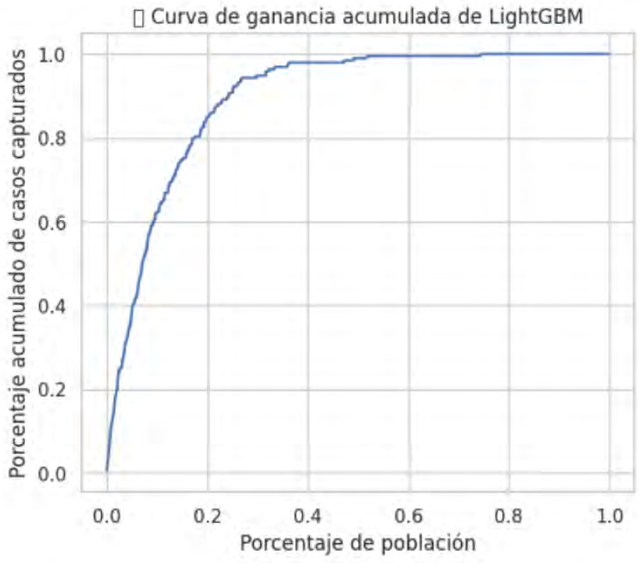


Figura 49
Curva de ganancia acumulada del modelo LightGBM
Nota. Elaboración propia.

Una vez que se realizó la comparación entre 7 modelos por medio de la evaluación WSM dando pesos a cada métrica y multiplicando por valor de cada modelo se determinó que el modelo más estable, robusto y que cumple de manera global con las características es el LightGBM.

Se mostró el reporte de clasificación y conforme a los valores de la matriz de confusión se pudo determinar que las métricas que se tienen con los datos de prueba son significativamente mejores que los que se muestran en la tabla comparativa de modelos que se genera con el promedio de las métricas de la validación cruzada del entrenamiento de cada modelo. La medida de Recall que se tiene es de 0.7461, la de F1–score es de 0.3929 y una precision de 0.2667, para una data con un desbalance de esa naturales son métricas aceptables con tendencia a buenas.

La curva ROC–AUC reportó 0.92, este valor es excepcional y demuestra que el modelo tiene una buena capacidad para diferenciar entre la clase 0 (no reprobó) y la clase 1 (reprobó) aunque su valor puede ser engañoso. También, la curva KS muestra un alto valor entre las líneas en la gráfica de las tasas de

verdaderos positivos (TPR) y la línea de los falsos positivos (FPR) esta diferencia alcanza su valor máximo en el umbral 0.27 y es 71%.

Para un set de datos que tiene el porcentaje de desbalance tan grande, una de las gráficas que debe ser más importante es la de Precision–Recall que en la gráfica mostró un AUC de 0.3923, un valor aceptable tomando en cuenta que el azar en el modelo es de 0.05, esto equivale a más de 7 veces ese valor.

La última gráfica que se analizó fue la de curva de ganancia acumulada, esta gráfica es determinante para definir si el modelo cumple las características y está listo para salir a producción. En la gráfica muestra que el 90% de los estudiantes que realmente reprueban se pueden identificar ordenándose por el porcentaje de probabilidad y tomando el 30% de los más alto. Esto en cuestión de recurso es sumamente importante pues con menos recursos se van a poder atender más alumnos.

Una vez hecho este análisis se puede determinar que el modelo que se propone es un modelo que tiene un desempeño moderado tanto técnico como operativo. Con él se podrán focalizar los recursos, orientar las decisiones pedagógicas, pero, sobre todo, lo principal, se van a poder apoyar a los estudiantes que estén en riesgo de reprobar.

3.2.6 Metodología para el objetivo específico 6

3.2.6.1 Implementación y despliegue del modelo:

De acuerdo al objetivo O6 en donde se plantea implementar y desplegar el modelo y una vez que se definió el modelo predictivo que se usará para identificar a estudiantes en riesgo de reprobar tercer grado de secundaria, se hace la propuesta de arquitectura para el despliegue del modelo predictivo de reprobación escolar que permite su implementación, supervisión, monitoreo, mantenimiento,

actualización y entrega oportuna de resultados a las autoridades educativas. Esta propuesta la describiremos en dos fases, la primera que es la de despliegue y puesta en marcha del modelo y la segunda en donde se describe el esquema de seguimiento, monitoreo, mantenimiento del modelo y entrega de información a las autoridades educativas.

La arquitectura de despliegue que se propone utiliza como base la plataforma de nube Microsoft Azure, integra procesos de ingesta de datos, transformación, entrenamiento, visualización y gobernanza.

En la obtención de fuentes de datos se integran datos de las escuelas, comunidades donde se ubican las escuelas, históricos académicos de los estudiantes de primero y segundo grado de secundaria de los estudiantes que ingresaran a tercer grado de secundaria. Los datos se obtendrán en un formato estructurado provenientes de la base de datos SQL de los sistemas educativos institucionales. La ingesta y transformación se propone se realice mediante Azure Synapse en donde un proceso automatizado limpiará, enriquecerá y transformará los datos para alimentar al modelo. En el entrenamiento y validación del modelo se propone utilizar Azure Machine Learning y AutoML para entrenar, evaluar el rendimiento del modelo y llevar el control de las versiones de las modificaciones y actualizaciones que se le realicen al modelo.

Una vez que el modelo se haya probado y cumpla con las especificaciones definidas se realizará el despliegue, el modelo se publica como servicio en producción, disponible para realizar predicciones sobre los estudiantes que estén próximos a ingresar a tercer grado de secundaria. La visualización y entrega de información de resultados se realizará con un Power BI, por medio de los sistemas escolares e institucionales y una API que estará a disposición de su uso para otras aplicaciones. Con estás interfaces se podrá consultar los estudiantes en riesgo, estadísticas y reportes para apoyar a la toma de decisiones para una intervención temprana en los alumnos por medio de políticas públicas que ayuden a mitigar esta situación de vulnerabilidad de los estudiantes.

Para garantizar la gobernanza y seguridad de los datos se incorporan los servicios de Azure Purview, Key Vault, App Registration, Entra ID y Azure Monitor, nos ayudan a mantener una trazabilidad de los datos, seguridad y control en el acceso y en el monitoreo de toda la arquitectura.Para garantizar la utilidad y precisión del modelo a lo largo del tiempo, se hace una propuesta de estrategia para el esquema de seguimiento, monitoreo, mantenimiento y actualización continua del modelo, como se muestra en el diagrama de propuesta de arquitectura.

Cada ciclo escolar se comparan las predicciones que ha realizado el modelo con la información de los datos reales de estudiantes que reprobaron tercer grado de secundaria para validar el rendimiento del modelo. Se verifican los resultados de las métricas.

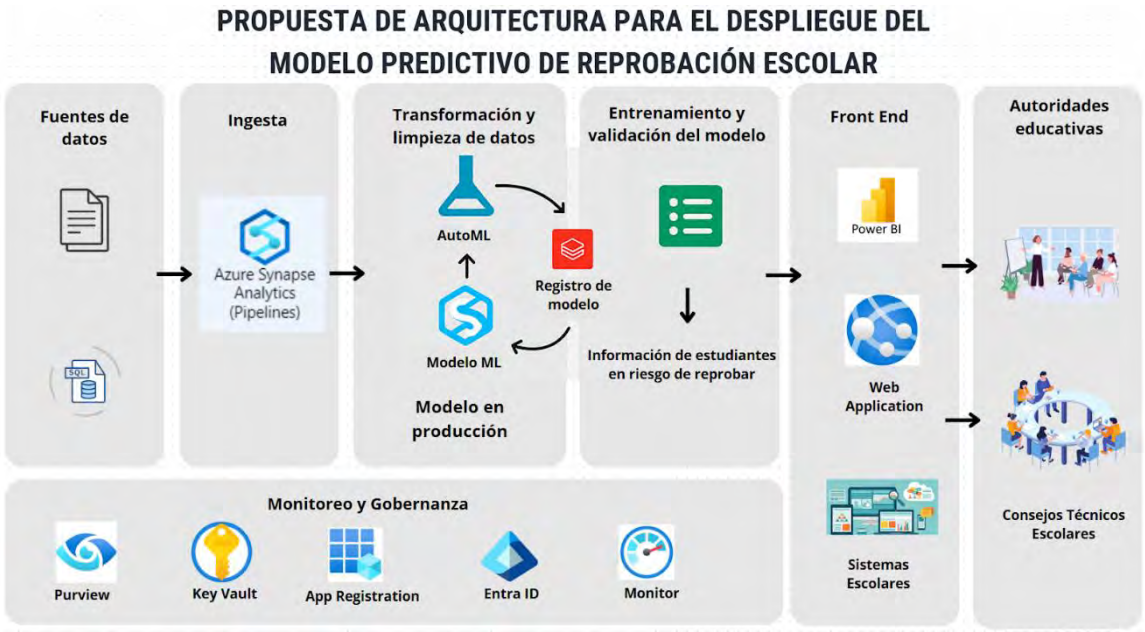


Figura 50
Propuesta de arquitectura para el despliegue del modelo predictivo de reprobación escolar
Nota. Elaboración propia.

Conforme al diagrama se aplicarán técnicas de data drift y concept drift, son técnicas de monitoreo de variables y sirven para identificar cualquier cambio que se presente en los datos de ingesta o durante en proceso del modelado, los servicios

que se proponen tienen estas características y permiten que se configuren diferentes parámetros incluido los tiempos de revisión.

El modelo se actualizará cada ciclo escolar conforme a los resultados de las evaluaciones y al nuevo conjunto de datos, los ajustes de hiperparámetros se realizará automáticamente con las herramientas propuestas.

Cada versión del modelo será registrada con sus respectivas configuraciones y desempeño, registrando los elementos como variables utilizadas, fecha de entrenamiento, rendimiento base, identificación del modelo, así como los usuarios responsables.

Las autoridades educativas tendrán acceso a la información por medio de tableros y reportes que serán diseñados y organizados conforme a la estructura educativa que se quiera informar. El primer informe que se realizará será antes del inicio escolar de cada año, previo a la reunión de los Consejos Técnicos Escolares de inicio de ciclo ya que ahí es donde se definirán las estrategías para darle la atención al problema.

Al final del ciclo escolar se realizan reuniones de retroalimentación entre los expertos en educación y el equipo técnico para ver los resultados que se obtuvieron aplicando políticas públicas a este grupo de estudiantes y con ello evaluar el resultado que se tiene al aplicar este tipo de herramientas tecnológicas.

3.3 Instrumentos a trabajar

Infraestructura: La información histórica de los alumnos se tiene almacenada en infraestructura de la nube en Amazon Web Services (AWS) en un servicio llamado EC2 que es el equivalente a un servidor en sitio. Cuenta con sistema operativo Windows Server. La aplicación con la que se consulta esa

información está ubicada en la nube de Azure y utiliza diferentes servicios como: service plan, Key Vault, App Registration, Entra ID, por mencionar algunos.

Bases de datos: En la institución se utiliza la base de datos SQL Server Enterprise Edition versión 2022. Como se comentó, la información y las aplicaciones se encuentran alojadas en la nube de AWS y AZure, ahí es donde se encuentra la información histórica de los estudiantes, de su rendimiento académico, así como los datos de los centros educativos a donde asisten. Es una base de datos relacional que almacena el registro de más de 400,00 registros de alumnos al año, así como la evaluación de cada periodo y la información concerniente a su desempeño escolar.

Herramientas de software: Para la codificación del modelo se utiliza Google Colab como entorno de desarrollo, como lenguaje de programación se utiliza Python con las respectivas librerías para el procesamiento del modelo como: Numpy, pandas, scikit–learn, Matplotlib, TensorFlow. Para poder entrenar y procesar el modelo se requiere un entorno de ejecución en google colab y dependiendo de la carga de proceso se requerirá utilizar CPU, GPU o TPU.

NumPy (Numerical Python) es una biblioteca fundamental para la computación científica en Python utilizada por su capacidad para manejar eficientemente arreglos y matrices multidimensionales, junto con una amplia colección de funciones matemáticas de alto rendimiento. La estructura central de MumPy es el narray, que permite realizar operaciones vectorizadas, lo cual optimiza significativamente los tiempos de cómputo al evitar bucles explícitos en Python puro. Esta característica resulta especialmente útil en tareas de preprocesamiento de datos, transformación de variables y ejecución de algoritmos de aprendizaje automático que requieren manipulación intensiva de matrices (NumPy Developers, s.f.).

Pandas es una biblioteca de código abierto escrita en Python que proporciona estructura de datos y herramientas de análisis de alto rendimiento, diseñadas para facilitar la manipulación, limpieza, transformación y análisis de datos estructurados. Su estructura fundamental, el DataFrame, permite trabajar con datos tabulares de forma flexible, similar a las hojas de cálculo o tablas relacionales, ofreciendo operaciones optimizadas para filtrado, agrupamiento, agregación y fusión. Se integra de manera fluida con otras bibliotecas del ecosistema científico de Python, como Numpy, Matplotlib u Scikit–learn, lo que la convierte en un componente esencial en proyectos de ciencia de datos, análisis estadísticos y aprendizaje automático (Pandas Development Team, s.f.).

Matplotlib es una biblioteca de visualización de datos de Python diseñada para crear gráficos estáticos, animados e interactivos de alta calidad. Su arquitectura flexible permite representar información de manera gráfica mediante diagramas de líneas, barras, dispersión, histogramas, mapas de calor, gráficos 3D y más. Es ampliamente utilizada en el ámbito científico y académico debido a su capacidad de integrarse con bibliotecas como NumPy, Pandas y SciPy, lo que la convierte en una herramienta central del ecosistema de ciencia de datos de Python. La interfaz principal de Matplotlib, conocida como pyplot, proporciona una estructura similar a la de MATLAB, permitiendo a los usuarios generar gráficos complejos con comandos simples, sin sacrificar el control sobre cada elemento visual. Es una herramienta fundamental para la representación visual de resultados estadísticos, análisis de datos y comunicación de hallazgos en investigaciones basadas en aprendizaje automático (Matplotlib Developers, s.f.).

Scikit–learn es una biblioteca de código abierto desarrollada para el lenguaje de programación Python, especializada en algoritmos de aprendizaje automático. Desde su creación, ha sido diseñada con el propósito de facilitar la implementación de modelos estadísticos y de machine learning a través de un API coherente, accesible tanto para investigadores como para profesionales de la industria. Su arquitectura modular está construida sobre bibliotecas científicas

ampliamente adoptadas como Numpy, Scipy y matplotlib, lo que garantiza eficiencia computacional y compatibilidad con ecosistemas analíticos consolidados. El desarrollo del proyecto. Se distingue por su capacidad de ofrecer herramientas robustas para tareas de clasificación, regresión agrupamiento, reducción de dimensionalidad, selección de características y validación cruzada (Scikit–learn, s.f.).

Seaborn es una biblioteca de visualización estadística en python que proporciona una interfaz de alto nivel para construir gráficos atractivos y bien estructurados. Permite representar relaciones entre variables, distribuciones y categorizaciones de forma eficiente, integrando funcionalidad con estructuras de datos tipo Dataframe de Pandas. Su uso es común en proyectos de ciencia de datos debido a su capacidad de mostrar patrones, correlaciones y variabilidad en los datos, lo que facilita la interpretación exploratoria previa al modelo (Waskom, 2021).

3.4 Población

Como se comentó en el estado de Querétaro a finales del ciclo escolar 2022–2023 se tenían registrados 654,719 estudiantes en todos los niveles educativos, de los cuales 460,485 corresponden a educación básica que son educación inicial, preescolar, primaria y secundaria.

USEBEQ atiende el nivel de educación básica y la propuesta va dirigida a atender a este sector, aunque el análisis que se realizará será de la información histórica de los años de 2017, 2018 y 2019.

3.5 Muestra: Subconjunto de la población

Con la implementación de esta propuesta se pretende que se identifiquen los estudiantes que están en riesgo de reprobación en el tercer grado de secundaria para que por medio de programas y políticas públicas se les pueda ayudar a salir de la situación en la que se encuentran y puedan aprobar y concluir satisfactoriamente este nivel educativo.

De igual manera se pretende que de forma indirecta los demás indicadores, incluso a la larga el de rezago educativo que no forma parte de los indicadores que reporta la USEBEQ, se vean impactados de forma positiva y se mejoren gradualmente.

Por consiguiente, la propuesta que se realizará contempla el análisis de la información histórica de los años 2017, 2018 y 2019 de los estudiantes que cursaron en 2019 tercer año de secundaria, lo que determina la muestra que será utilizada.

4. RESULTADOS

El presente trabajo busca proponer un modelo de ciencia de datos que nos ayude a predecir la probabilidad de riesgo que tiene un estudiante de tercer grado de secundaria general.

La información histórica proporcionada es la que cuenta la institución y corresponde al cierre del ciclo escolar 2018–2019 de los estudiantes que cursaron tercer grado de secundaria en el año 2019 y la información histórica de primer grado y segundo grado de secundaria de estos alumnos.

Se utilizó la metodología CRISP–DM que es la que tiene más aceptación en el desarrollo de proyectos de minería de datos y de ciencia de datos.

Una vez concluido con las fases de la metodología nos hemos dado cuenta que la propuesta del modelo puede ser viable ya que el análisis ha sido riguroso y se han aplicado diferentes técnicas de evaluación del rendimiento y los resultados, sino son los mejores dejan ver que el modelo tiene mucho potencial, sobre todo por la naturaleza de problema, que corresponde al sector que se atiende, el educativo.

4.1 Resultados obtenidos para el objetivo específico 1.

Se llevó a cabo un análisis del entorno institucional y del problema educativo a resolver. Esta fase permitió comprender el marco normativo, organizacional y operativo de la Unidad de Servicios para la Educación Básica en el Estado de Querétaro (USEBEQ), así como la pertinencia de aplicar modelos predictivos basados en inteligencia artificial para mejorar la toma de decisiones en el ámbito educativo.

Desde su creación en 1992, USEBEQ es responsable de administrar los servicios de educación básica en el estado. Su Programa Institucional 2021–2027

establece como uno de sus objetivos prioritarios asegurar trayectorias educativas completas, permanencia escolar y aprendizajes significativos. En este contexto, la reprobación en secundaria general representa un problema estructural que compromete dichas trayectorias y requiere atención mediante herramientas preventivas y basadas en evidencia.

El proyecto se alinea con el Objetivo 4 y la Estrategia 5.2 del programa institucional, que enfatizan el uso de datos para identificar áreas de oportunidad y optimizar los recursos educativos. Asimismo, se recuperó el diagnóstico FODA contenido en el Manual de Calidad de USEBEQ, el cual permitió identificar fortalezas, limitaciones y condiciones mínimas necesarias para la implementación de soluciones tecnológicas innovadoras.

La propuesta de solución se basa en el modelo CANVAS ML, cuya propuesta de valor está centrada en ofrecer una educación de calidad, inclusiva y orientada al desarrollo integral. Este modelo, representado en la Figura 6, traduce los principios institucionales en un marco técnico orientado a la prevención de la reprobación escolar, mediante un enfoque de clasificación binaria que permite anticipar, con base en datos históricos de primero y segundo grado de secundaria, a los estudiantes en riesgo de reprobar tercer grado.

El modelo incorpora variables como inasistencias, calificaciones, turno escolar, nivel de marginación y categoría de la comunidad. Las predicciones se realizan al concluir el segundo grado y son entregadas a las autoridades educativas antes del inicio del siguiente ciclo escolar, permitiendo planificar estrategias de intervención focalizadas.

En cuanto a su diseño metodológico, la propuesta contempla el entrenamiento con datos históricos, validación y monitoreo continuo, priorizando métricas como Recall, F1-score, PR AUC y MCC, con el fin de minimizar los falsos negativos. De este modo, la institución contará no solo con un sistema de alerta temprana, sino con una herramienta robusta que respalde la asignación eficiente de

recursos, la planificación de acciones pedagógicas y la mejora de indicadores clave del sistema educativo.

El plan de implementación se estructura en seis fases según la metodología CRISP–DM, como se detalla en la Tabla 6. Este plan incluye tiempos estimados en semanas, recursos humanos y técnicos requeridos, así como posibles riesgos que podrían afectar el desarrollo, destacando la comunicación como un factor crítico para el éxito del proyecto.

4.2 Resultados obtenidos para el objetivo específico 2.

Durante las fases de conocimiento de la información se pudieron identificar por medio de análisis exploratorio de datos algunos hallazgos que fueron de utilidad para la preparación y transformación de la información. Primeramente, se tiene un archivo con 19,156 registros y 64 variables o campos en formato CSV. Se cargó al ambiente de desarrollo que es una plataforma de desarrollo de Google llamada Colab. Se agruparon las diferentes variables de acuerdo a su origen.

Respecto al entorno de la escuela se tienen las variables de municipio, región, categoría y marginación. En la variable de municipio se pudo ver que Colón (30%) y Pinal de Amoles (18%) son los que tienen los porcentajes más altos de reprobación y los municipios de la región 4 o metropolitana tienen porcentajes de 6% (moderado) a pesar de que esta tiene el porcentaje más alto de estudiantes.

La región 1 que corresponde a Sierra del estado tiene un porcentaje de 13.2% de reprobados con respecto a la población que tiene, es un valor alto pues las otras 3 regiones rondan en 5% de reprobación. Respecto a la variable CATEGIRIA, en esta se define la clasificación que INEGI les da a las localidades donde se encuentra la escuela y se tienen los valores de rural, urbana y S/D. Conforme a el análisis, las escuelas rurales tienen un porcentaje más alto (8.8%) de estudiantes que reprobaron y es menor a las urbanas (5.4%). Estos valores que se

tienen en estas variables marcan una desigualdad de condiciones de aprendizaje, de acceso a los recursos educativos, infraestructura educativa, de comunicaciones, de oportunidades, por mencionar algunas.

Se tienen variables que corresponde al niño como en es turno en el que asiste a la escuela, el sexo y la edad. En la edad se pudo ver que la mayoría de los estudiantes están en tercer grado de secundaria a la edad de 14 años y el porcentaje de reprobación es moderado (5%). Las edades donde existe mayor reprobación son 15 y 16 años con 8.8% y 8.6%. Esto nos da un indicativo que se puede tomar de diferentes puntos de vista, el educativo, donde se puede deducir que estos estudiantes ya traen arrastrando rezago de ciclos anteriores, también se puede ver desde el familiar y económico, donde los jóvenes tienen que trabajar para mantener a la familia, emocional o social. La variable de sexo tiene un porcentaje equilibrado entre mujeres y hombres con una pequeña diferencia a favor de las mujeres. En cuestión de reprobación no se muestran los mismos números ya que los hombres tienen 7.6% de porcentaje de reprobación y las mujeres 3.7%, esto es un poco más del doble. Este fenómeno puede estar relacionado a diferentes factores, por ejemplo, la edad, familiar o el lugar donde está la escuela

El turno también muestra un poco de diferencia entre los estudiantes que reprobaron que tienen turno matutino y vespertino, para el primero, aunque tiene una mayor concentración de alumnos su porcentaje es de 5.1% para el vespertino muestra en porcentaje de 7.5%.

También se tienen variables que están asociadas al rendimiento, como son las calificaciones de las materias, faltas, número de materias reprobadas, número de periodos reprobados, número de periodos con calificaciones de 6 a 8, número de calificaciones mayores a 8, por mencionar algunas.

Estas variables tienen las variables que cuentan los periodos entre 6 y 8 de primero y segundo grado entre más calificaciones tenga en ese rango mayor será el riesgo de reprobación y por el contrario si tienen más periodos con calificaciones

mayores a 8 menor es el riesgo de reprobación. Pasa exactamente lo mismo con las variables que cuentan los bimestres reprobaron en primero y segundo grado, entre más periodos reprueben mayor es el riesgo de reprobar tercer grado.

En cuestión de calidad de la información se pudo ver que la data tiene valores en las calificaciones que tienen valor de 0 y las calificaciones permitidas son de 5 en adelante. Estos valores afectan directamente en la distribución de los datos y en las medidas estadísticas de cada variable, esto provoca que el sesgo se tienda al lado izquierdo principalmente o que la media, la mediana o la desviación estándar cambien su valor. También se encuentran variables que tienen una diferencia muy bajo o que su valor es 0 en el rango intercuartil, estás tienen el mismo valor o la proporción de valores es muy grande.

Se identificó la utilidad de las variables según su dispersión se identificaron 8 variables para ser descartadas, entre ellas la edad, grado, grado2, grado3, la latitud y longitud de la escuela.

La variable que se utilizó como etiqueta u objetivo para que los modelos hicieran su entrenamiento y pudiera comprobar su rendimiento fue ReproboGrado3, esta variable es dicotómica con valores de 0 (no reprobó) y 1 (reprobó), tiene un desbalance muy alto con aproximadamente 96% para la clase mayoritario (no reprobó) y aproximadamente 4% para la clase minoritaria (reprobó). Con estos hallazgos se procedió a realizar el tratamiento de la información con técnicas de eliminación, transformación y escalado de los datos.

4.3 Resultados obtenidos para el objetivo específico 3.

De acuerdo a la metodología CRISP–DM después de la fase de conocimiento de la información continúa la fase de preparación de los datos, tiene el propósito de darle calidad, integridad y valor significativo a la información para que los modelos se puedan entrenar adecuadamente y proporcionen las mejores

métricas en su rendimiento. Se pretende mantener la mayor información de la data porque entre más registros se tengan mejores resultados se tendrán en el entrenamiento.

Primero se realizó la depuración de las variables redundantes y sin valor. Las variables que tenían poco o nula variabilidad, rango intercuartil cercano a cero se eliminaron ya que no tenían variabilidad, su valor era el mismo o tenían la misma información que otras variables, por ejemplo las variable GRADO1, GRADO2, TOTALGRADOS, ALUMNOLAT, ALUMNOLON, bimSinCali_ANTE y binSinCali_PASA.

A los valores que tienen las variables más allá del límite del rango aceptable a la derecha y a la izquierda se le conoce como valores atípicos(outliers). Se detectaron atípicos en las variables bimRepr_ANTE,bimRepr_PASA, PROMEDIO_ANTE, PROMEDIO_PASA, EDUCACPASA, AMBIT__PASA, entre otras. Como se comentó lo que se pretende es tratar de eliminar la menor información posible por lo que se realiza la sustitución de esos valores, para el caso se reemplazaron los valores atípicos por el valor del límite aceptable ya sea de la derecha o de la izquierda. Este procedimiento evita la eliminación de información y mantiene la integridad de los registros, además, normaliza la dispersión de los datos de la variable.

Se identificaron las variables que tienen un coeficiente de variación alto, esto quiere decir que tienen una alta dispersión, se calcula tomando la desviación estándar entre la mediana, los valores que son mayores a 1 son los que tienen una dispersión alta. Las variables que caen en este supuesto son escPriv, bimRepr_ANTE y bimRepr_PASA y se consideraron como variables de alta dispersión. A estas se les aplicó la técnica de escalado RobustScaler, toma la mediana y se escala conforme al valor del rango intercuartílico. Con esta técnica evitas el problema de valores atípicos y se mantiene la distribución de la variable.

Una vez concluido el proceso de eliminación, depuración y transformación en las alertas de anomalías en la información se tiene que identificar las variables que tienen un alto grado de correlación, esto quiere decir, que si se incrementa o disminuye el valor de una variable la otra variable se modifica, por ejemplo: si baja el valor de la variable promedio la variable de que indica el porcentaje de probabilidad de reprobación también se modifica.

Esta correlación se identifica en una gráfica de calor de Pearson donde las variables que tienen un color más intenso o un número cercano a 1 o -1 tiene una correlación con la otra variable. Los valores que son 0 o cercanos a él no tienen correlación. Ya que se identificaron las variables que se correlacionan se tiene que eliminar.

Ya que se eliminaron las variables se aplica una técnica de selección de variables que selecciona las más significativas, es decir, las que tengan mayor predictivo para el modelo. El resultado final de la selección fue un dataset con 12 variables y 18,176 registros.

4.4 Resultados obtenidos para el objetivo específico 4.

Una vez que se tiene el dataset en condiciones de trabajar con los modelos se procede hacer las configuraciones para cada uno de ellos. Para este caso se entrenaron 7 modelos Red Neuronal (Keras), Ensamble (Random Forest, XGBoost, LightGBM), Red Neuronal (SciKit–learn), Random Forest, XGBoost y Regresión Logística.

Para ver el rendimiento de los modelos se utilizaron 8 métricas que son las más importantes conforme al desbalance que se tiene en el set de datos y son: F1–score, precision, recall, ROC AUC, PR AUC, Kappa y MCC y Mejor vs azar, es la columna que indica cuántas veces es mejor el modelo conforma a la curva PR AUC considerando la línea de referencia o línea de azar.

Se generó una tabla con los valores de cada métrica para cada modelo, en ella se muestra por medio de colores cuál es el modelo que tiene el valor más alto en cada métrica. A simple vista se puede ver que los valores que sobresalen son los modelos de Red Neuronal (Keras), Ensamble y LightGBM.

Para la Red Neuronal se creó una arquitectura de 5 capas, conforme a las gráficas que se mostraron en la etapa de modelado se tiene una evolución favorable. Sus métricas son F1–score con 0.3855 que para el desbalance que tenemos es un valor moderado, recall de 0.4404, es un valor bajo, está por debajo de la línea de referencia o la línea de azar, precision es de 0.3427, que para el tipo de datos que se tienen se puede considerar moderado. MCC tiene un valor de 0.3591, lo podemos considerar aceptable considerando el desbalance, val_f1–score tiene el valor de 0.8691en la época 11, es una métrica de validación y lo hace con un set de datos que el modelo no ha visto, esto indica que en las pruebas el modelo se comportó con un muy buen rendimiento con una buena capacidad de generalización. En lo general el modelo tiene un buen rendimiento el recall aunque es bajo muestra la capacidad para detectar alumnos en riesgo de reprobación. La captura de casos positivos permite usarlo para la identificación temprana de estudiantes en riesgo.

Para el modelo de Ensamble se integraron 3 modelos, Random Forest, XGBoost y Regresión Logística. El modelo se entrenó con una técnica que se llama de validación cruzada y lo hace realizando iteraciones, en cada una realiza una combinación de parámetros buscando encontrar los mejores.

Las mejores métricas que obtuvo fueron F1–score con 0.3699, moderado de acuerdo al dataset, precisión con valor de 0.2761, bajo, pero considerando las condiciones se puede decir que es aceptable, recall con 0.5602 y Kappa con valor de 0.3216, este valor fue el más alto en está métrica y el PR AUC que tiene el valor de 0.2888, es un valor muy bueno ya que si lo comparamos contra el azar muestra el valor que da es alto. Las métricas que tiene muestran un modelo robusto con un buen desempeño. Su capacidad para capturar estudiantes en riesgo es buena por

su buen recall. Integrar los tres modelos hace que sea equilibrado. Muestra un comportamiento estable en su curva de validación. Demuestra que puede ser útil para la detección temprana de estudiantes que se encuentran en riesgo de reprobación.

El modelo de LightGBM muestra el mejor recall de los 3 modelos con 0.6107 y el más alto valor para PR AUC con 0.3218, con este valor se supera por mucho el valor que se tiene en la línea de azar. Son métricas de suma importancia donde es crítico la detención acertada de verdaderos positivos. Las métricas que tiene son equilibradas y por consecuencia tiene esos valores en MCC de 0.3456 y Kappa de 0.2989. La gráfica de SHAP en donde se muestran las variables más significativas en el modelo mostró cómo las variables PROMEDIO_PASA y PROMEDIO_ANTE corresponden al promedio de primer grado y segundo grado de secundaria y materiasRep_PASA, cantidad de materias reprobadas en segundo, influyen en las predicciones para la detección temprana de estudiantes en riesgo de reprobación. En general el modelo es robusto y tiene un desempeño equilibrado, muestra que tiene un alto nivel para capturar casos de verdaderos positivos que están en riesgo de reprobación.

4.5 Resultados obtenidos para el objetivo específico 5.

Se realizó la evaluación de los modelos haciendo una ponderación de las métricas conforme a la importancia del caso que estamos tratando en este orden: recall, F1–score, precision, PR AUC, Mejora vs azar, MCC, Kappa, Balanced Accuracy y ROC AUC. Se ordenaron las métricas de menor a mayor y dio como resultado el mejor modelo. Conforme a este ordenamiento el modelo más equilibrado es LightGBM, en la fase de modelado la gráfica de Radar ya había mostrado que ese modelo era el que mejor rendimiento tenía.

El recall que tiene es alto (0.6107), lo que indica que tiene una buena detección de estudiantes en riesgo de reprobación. PR AUC con valor de 0.3218 muestra la capacidad que tiene para identificar casos positivos en un entorno de datos desbalanceado. Aunque la curva ROC AUC no es una métrica confiable para set de datos desbalanceados esta métrica muestra un valor alto (0.8787). MCC y Kappa estás métricas están asociadas a medir el rendimiento global del modelo conforme a la predicción y al valor real, también tienen valores aceptables que indican lo robusto del modelo, sus valores son: 0.3364 y 0.2989 respectivamente. La columna Mejor vs Azar es un valor que agregué a la tabla para poder visualizar de una manera clara el rendimiento del modelo comparándolo contra una predicción aleatoria, para este modelo se puede ver que es 6.0616 veces mejor.

En la gráfica de la curva PR AUC se muestra la relación que existe entre la precision y el recall en un modelo de clasificación en diferentes umbrales. Esta métrica se focaliza en la clase minoritaria, en nuestro caso los alumnos en riesgo de reprobación. La línea aleatoria es la proporción de los casos positivos en el conjunto de datos. Lo que está por arriba de esa línea es la identificación correcta de los casos positivos y entre más se separe de esa línea mejor desempeño tiene el modelo. De acuerdo a la curva y al valor de la métrica nos indican que el modelo está capturando bien los casos verdaderos positivos sin generar demasiados falsos positivos.

La curva KS muestra la capacidad que tiene el modelo para diferenciar entre las clases, se usa en modelos de clasificación. El valor de KS que se mostró en la gráfica es del 0.71, indica la alta capacidad para separar las clases comando en consideración que el máximo valor puede ser 1, este valor sucede en el umbral de 0.27, quiere decir que en este valor el modelo tiene su máxima capacidad para identificar estudiantes en riesgo de reprobación.

La curva de Ganancia Acumulada muestra la relación entre el porcentaje de población contra el porcentaje acumulado de casos capturados. Conforme a la gráfica podemos ver que si tomamos el 30% de los estudiantes con mayor

probabilidad se logrará capturar más del 90% de los estudiantes que efectivamente reprobaron.

Esta gráfica demuestra realmente el poder que tiene el modelo porque si se focaliza los esfuerzos en atender principalmente al 30% de los estudiantes que están en riesgo se podrán abatir este problema en un gran porcentaje invirtiendo menos recursos.

4.6 Resultados obtenidos para el objetivo específico 6.

Se diseñó una arquitectura tecnológica para implementar y desplegar el modelo predictivo de reprobación escolar para estudiantes de tercer grado de secundaria. Esta solución permite su ejecución, monitoreo, mantenimiento, actualización y entrega de resultados a las autoridades educativas.

La propuesta contempla dos fases: el despliegue inicial del modelo y un esquema de seguimiento y mejora continua. La arquitectura se basa en Microsoft Azure e integra procesos de ingesta, transformación, entrenamiento, visualización y gobernanza de datos.

Los datos provienen de bases institucionales en SQL, que incluyen antecedentes académicos de los alumnos y características del entorno escolar. La ingesta y transformación se automatizan con Azure Synapse, mientras que el entrenamiento y validación se realizan con Azure Machine Learning y AutoML.

Una vez validado, el modelo se publica como servicio para predecir el riesgo de reprobación en nuevos estudiantes. Los resultados se consultan mediante Power BI y una API que permite generar reportes y estadísticas para decisiones educativas.

La gobernanza se garantiza con servicios como Azure Purview, Key Vault y Monitor, que aseguran trazabilidad, seguridad y control de accesos. Cada ciclo

escolar se comparan las predicciones con los resultados reales y se monitorean cambios en los datos mediante técnicas de data drift y concept drift, permitiendo ajustes automáticos del modelo.

El acceso a los resultados por parte de las autoridades educativas se garantiza mediante tableros de control dinámicos y reportes personalizados, estructurados conforme a los niveles de la organización educativa. El primer informe se genera antes del inicio del ciclo escolar, con el fin de presentar los resultados en las reuniones del Consejo Técnico Escolar, donde se definen las estrategias de intervención. Al cierre del ciclo, se realizan sesiones de retroalimentación entre equipos técnicos y expertos en educación, con el propósito de evaluar el impacto de las políticas aplicadas a partir de las predicciones del modelo y fortalecer su uso como herramienta tecnológica de apoyo a la toma de decisiones.

4.7 Principales hallazgos y propuestas derivadas de la implementación del modelo.

De acuerdo con los resultados obtenidos, el principal hallazgo de esta investigación es que el uso de modelos de ciencia de datos, en específico de clasificación binaria, permite identificar de manera anticipada a estudiantes en riesgo de reprobación, habilitando una toma de decisiones oportuna y basada en evidencia dentro del sistema educativo.

La implementación de esta propuesta puede incidir en distintos ámbitos de acción institucional y de política pública, como se detalla a continuación:

- **Diseño de políticas públicas preventivas:** Posibilita la creación de estrategias específicas para mitigar el riesgo de reprobación, dirigidas a estudiantes en situación de vulnerabilidad académica o social.

- **Planeación escolar basada en datos:** Facilita la asignación estratégica de recursos humanos, pedagógicos y tecnológicos en función del perfil de riesgo de los estudiantes antes de iniciar el ciclo escolar.
- **Focalización de intervenciones pedagógicas:** Permite implementar tutorías, programas de acompañamiento o reforzamiento escolar enfocados en alumnos con alta probabilidad de reprobación.
- **Fortalecimiento de la gestión institucional:** Mejora la capacidad operativa de las autoridades educativas mediante tableros y reportes dinámicos, apoyando decisiones informadas en los Consejos Técnicos Escolares.
- **Reducción de brechas educativas territoriales:** Identifica desigualdades entre regiones, municipios y comunidades (Colón, Pinal de Amoles, escuelas rurales), visibilizando territorios con mayores necesidades educativas.
- **Priorización de atención según características individuales:** Muestra patrones diferenciales de reprobación por edad, sexo, turno escolar o historial académico, lo que permite personalizar las estrategias de intervención.
- **Automatización y actualización continua del modelo:** Incorpora una arquitectura tecnológica que posibilita su actualización anual, detección de cambios en los datos (data drift) y mejora continua de su desempeño.
- **Integración del enfoque de equidad educativa:** Al visibilizar la disparidad en los indicadores entre estudiantes hombres y mujeres, zonas rurales y urbanas, o grupos con rezago acumulado, se promueve la atención diferenciada y justa.
- **Optimización de recursos con base en curvas de ganancia acumulada.** La focalización de esfuerzos en el 30% de estudiantes con mayor riesgo permite capturar más del 90% de los casos reales de reprobación, con una alta eficiencia de inversión.

- **Sustento técnico para modernizar la gestión educativa:** Este proyecto muestra que es viable integrar soluciones de inteligencia artificial (como redes neuronales, ensambles y modelos interpretables como LightGBM) en la toma de decisiones de instituciones públicas.
- **Modelo replicable y escalable:** Dada su arquitectura en Microsoft Azure y el uso de herramientas estandarizadas, el sistema puede replicarse a otras entidades federativas o niveles educativos, adaptándose a sus bases de datos e infraestructura.
- **Generación de evidencia para la formulación de nuevas líneas de investigación:** El análisis de variables predictivas abre la puerta al estudio de factores estructurales asociados al rezago escolar, abandono y desigualdad en la trayectoria educativa.

CONCLUSIONES

La presente tesis tuvo como principal objetivo diseñar, entrenar y evaluar un modelo predictivo que tenga la capacidad de identificar a estudiantes de tercer grado de secundaria general con alto riesgo de reprobación, utilizando con base para su entrenamiento y pruebas el historial académico de primero y segundo grado de secundaria, así como otras variables institucionales que tiene en su contexto el estudiante. Esta propuesta se contextualiza dentro de una problemática social y educativa de gran importancia en el estado de Querétaro, para que los estudiantes logren con éxito la permanencia y conclusión de su educación básica.

Esta propuesta está estructurada bajo la metodología de desarrollo CRISP–DM (Cross–Industry Standard Process for Data Mining), lo que permitió abordar el problema de una forma ordenada y sistémica, desde la comprensión del negocio hasta la evaluación e implementación del modelo. Se pudo demostrar que esta metodología es adecuada para este tipo de problemas no solo por su enfoque iterativo, sino porque permitió la alineación del modelo con las necesidades reales de educativo que se abordó, sobre todo en un contexto de recursos limitados.

Durante la compresión del negocio se identificaron los principales factores asociados a la reprobación escolar, basados en evidencias empíricas y reportes educativos escolares. Se definió que un alumno se considera reprobado si reprueba al menos una materia al final del ciclo escolar. Asimismo, se reconoció la importancia de desarrollar una solución que permita anticipar dichos casos, permitiendo intervenir antes de que el estudiante pierda el ciclo escolar.

En la fase de preparación de los datos se hizo la depuración de la información académica histórica de primero y segundo grado de secundaria de los estudiantes que cursaron el tercer grado de secundaria en 2019. Se integraron diferentes variables asociadas al alumno, a su rendimiento académico, a la escuela, al entorno de la escuela, así como derivaciones con alguna de ellas. Este proceso

implicó análisis de distribución, detección de valores atípicos, manejo de valores faltantes y el tratamiento de variables correlacionadas. Se aplicaron técnicas para normalizar los datos, escalado, codificación de variables categóricas y balanceo del set de datos mediante la técnica de SMOTE. Esto propició dejar un conjunto de datos, limpio, estructurado y con el número de variables adecuado para los modelos de clasificación.

En la fase de modelado se utilizaron diferentes modelos de clasificación como: regresión logística, Random Forest, XGBoost, LightGBM, Red Neuronal (Keras), Red Neuronal (Scikit–learn) y un modelo de ensamble que integró 3 modelos (regresión logística, Random Forest y XGBoost). De acuerdo al alto grado de desbalance de los datos (aproximadamente 96% no reprobados y 4% reprobados), se le dio prioridad a la optimización de las métricas de Recall, F1–score y PR–ACU debido a su sensibilidad para detectar correctamente a los casos positivos y su utilidad en contextos donde es preferible errar por exceso que, por omisión, como es el caso de la reprobación de estudiantes.

En la fase de evaluación se hizo la comparación de los diferentes modelos a partir de una tabla de métricas, se utilizó un enfoque multicriterio (Weighted Sum Model – WSM) para acumular los resultados de manera objetiva. Conforme a este enfoque el modelo que obtuvo mejor evaluación fue LightGBM, las métricas que obtuvo fueron: Recall de 0.6107, F1–score con 0.3699, PR AUC de 0.3218, ROC AUC de 0.8787 y un incremento de 6.06 veces en la detección de casos respecto a un modelo aleatorio. Este desempeño fue corroborado mediante la curva Kolmogorov–Smirnov (KS), cuyo valor obtenido fue de 0.71 con un umbral de 0.27 y mediante la curva de ganancia acumulada, y con la curva de ganancia acumulada, que dio como resultado que al intervenir al 30% de los con mayor probabilidad de riesgo, se tenía el 90% de los casos reales de reprobación.

También, se presentó un análisis por medio de la gráfica SHAP, el cual arrojó la identificación de las variables más significativas para las predicciones del modelo y fueron: el promedio de primer grado de secundaria, el promedio de

segundo grado de secundaria y el número de materias reprobadas en segundo año de secundaria. Esto no sólo valida la lógica en los resultados del modelo desde una perspectiva educativa, sino también genera información relevante para orientar las políticas educativas, seguimiento escolar y la asignación de los recursos.

Esta tesis demuestra que el uso de técnicas de ciencia de datos y aprendizaje automático puede tener aplicaciones de alto impacto en el ámbito educativo. Un modelo como el que se propone permite a las autoridades anticiparse y planificar intervenciones para focalizar sus esfuerzos y prevenir no solo el que reprueben los estudiantes, sino también todos los sucesos que puede desencadenar este hecho. También, evidencia que es posible construir modelos equilibrados, robustos y explicables incluso en contextos con limitaciones de infraestructura y datos.

Los resultados alcanzados permiten concluir que el modelo LightGBM es una herramienta viable, confiable, eficiente y explicable para la detección temprana de estudiantes que se encuentran en riesgo de reprobación. Las técnicas y la metodología que se aplicaron, fueron rigurosas. adaptadas a las condiciones reales del sistema educativo y los hallazgos que se obtuvieron dan pie a que se tiendan las áreas de oportunidades similares al contexto de esta tesis con este tipo de herramientas, como pudieran ser la deserción, la reprobación por periodo o por materia, entre otros muchos temas.

Si bien el modelo que se propone tiene métricas aceptables, es un hecho que tiene un área de oportunidad muy grande y se puede mejorar con el reentrenamiento y la configuración de hiperparámetros durante el proceso de seguimiento, monitoreo, mantenimiento y actualización que se plantea en la fase de despliegue de la metodología CRISP–DM.

Es evidente que el rendimiento escolar no solo lo determinan las variables escolares como las que se utilizaron en esta propuesta. Existe información que afirma que variables asociadas a contexto familiar, social, económico, físico, por

citar algunos, influyen de igual manera, sino es que más, en la forma en que se desenvuelven los estudiantes en las escuelas y por consiguiente en su rendimiento académico.

Se recomienda que la institución realice un estudio al respecto para encontrar los mecanismos para obtener esa información y hacerla accesible para el modelo y se haga un reentrenamiento para intentar mejorar su rendimiento.

Igualmente se recomienda que se considere la implementación de modelos de ciencia de datos e inteligencia artificial, como el que se desarrolló en este estudio, y que los integre en el diseño de su planeación estratégica, políticas públicas y los mantenga en la agenda como una actividad cotidiana, con esto lograrán que la institución emprenda el camino para convertirse en una institución que tome decisiones operativas, estratégicas y tácticas sustentadas en el análisis sistemático de datos.

a) Se cumple la hipótesis (sí, no, porqué)

Sí, se cumple la hipótesis.

La hipótesis proponía que, Si se utiliza información académica e institucional de los alumnos de primero y segundo grado de secundaria, entonces es posible construir un modelo predictivo que identifique con una buena precisión a los estudiantes de alto riesgo de reprobar el tercer grado de secundaria general, aún en un contexto de datos desbalanceados.

Conforma a los resultados que se han mostrado en el desarrollo de la propuesta se define que la hipótesis si se cumple:

- Se logró entrenar un modelo LightGBM con un rendimiento aceptable con métricas de Recall de 0.6107, F1–score con 0.3699, ROC AUC de 0.8787, PR AUC de 0.3218, y un desempeño superior al azar en la identificación de estudiantes en riesgo de reprobación.

- El modelo mostró una alta capacidad de priorización, capturando el 90% de los estudiantes con alto riesgo de reprobación al intervenir al 30% con mayor probabilidad.
- El análisis de la gráfica SHAP confirmó que las variables más relevantes están relacionadas con el desempeño previo del estudiante, promedios de los grados anteriores y el número de materias reprobadas.

En consecuencia. de los antes planteado se valida empíricamente que la hipótesis es válida

b) Responde a los objetivos/pregunta de investigación (sí, no, porqué)

Sí, responde tanto a la pregunta como a los objetivos de manera integral.

Al aplicar el enfoque de la metodología CRISP–DM, el conocimiento de la institución, problema y datos, la preparación rigurosa de los datos, la aplicación de técnicas avanzadas para la búsqueda de hiperparámetros, el entrenamiento iterativo de los modelos y su evaluación objetiva demostraron que, sí es posible predecir con precisión y anticipación el riesgo de reprobación de estudiantes de tercer grado de secundaria utilizando información académica e institucional de primero y segundo grado de secundaria, mediante técnicas de ciencia de datos e inteligencia artificial.

El Objetivo O1, sí se cumplió. Se contextualiza el problema educativo en el nivel educativo de secundaria general y se alineó la propuesta al objetivo 5 del Programa Institucional 2012 – 2027, que dice “Impulsar que las diferentes instituciones del sector educativo tengan buena gobernanza y eficiente administración” (USEBEQ, 2021, p. 15).

El Objetivo O2, sí se cumplió. Se realizó un análisis detallado de los ciclos escolares, evaluación, promedio, y nivel de marginación, considerando variables del entorno escolar.

El Objetivo O3, sí se cumplió. Se aplicaron técnicas de limpieza, transformación, normalización, escalado, codificación y balanceo para preparar los datos conforme estándares de modelado.

El Objetivo O4, sí se cumplió. Se entrenaron modelos de aprendizaje automático y redes neuronales, evaluando métricas claves, seleccionando el mejor modelo conforme a su mejor rendimiento y desempeño.

El Objetivo O5, sí se cumplió. Se evaluaron los modelos conforme a las métricas: F1–score, Recall, PR AUC, ROC AUC, KS, MCC, Kappa; se seleccionó LightGBM como modelo final por su equilibrio y capacidad predictiva.

El Objetivo O6, sí se cumplió. Se propuso una estrategia de despliegue que incluye ingesta de datos, monitoreo, supervisión y actualización continúa orientada al uso institucional.

Se concluye, que la hipótesis es válida, la pregunta de investigación se responde afirmativamente y los seis objetivos específicos se cumplen, tanto en el plano metodológico, técnico y el aplicado.

REFERENCIAS

Anaya Benítez, F. (Ed.). (2023). Caminando hacia la innovación en educación: De la teoría a la práctica (1.ª ed.). Dykinson S.L.

Arias, F. (2012). El proyecto de investigación: Introducción a la metodología científica (6.ª ed.). Editorial Episteme.

Banco Mundial. (s.f.–b). Multidimensional poverty measure (MPM). Recuperado de https://www.worldbank.org/en/topic/poverty/brief/multidimensional-poverty-measure

Brownlee, J. (s.f.). ROC curves and precision–recall curves for classification in Python. Machine Learning Mastery. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21 (6), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21 (6). https://doi.org/10.1186/s12864-019-6413-7

Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2009). Metodología para la medición multidimensional de la pobreza en México (3ª ed.). Recuperado de https://www.coneval.org.mx/InformesPublicaciones/InformesPublicaciones/Documents/Metodologia-medicion-multidimensional-3er-edicion.pdf

Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2023). Informe de pobreza multidimensional en México 2022. https://www.coneval.org.mx/InformesPublicaciones/Documents/Pobreza_Multidimensional_2022.pdf

Davis, J., & Goadrich, M. (2006). The relationship between Precision Recall and ROC curves. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández Naranjo, S. (2016). Metodología de la investigación aplicada a la educación. Universidad de Sevilla.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Recuperado de https://www.deeplearningbook.org

Guanin Fajardo, J. H. (2024). Análisis del sistema educativo de la Universidad Técnica Estatal de Quevedo mediante ciencia de datos [Tesis doctoral, Universidad de Granada].

Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2014). Metodología de la investigación (6.ª ed.). McGraw-Hill Educación.

IBM. (2015). CRISP-DM: Cross Industry Standard Process for Data Mining (Versión 1.0). IBM SPSS Modeler Documentation. Recuperado de

https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

IBM. (s.f.). ¿Qué es la ciencia de datos?. https://www.ibm.com/mx-es/topics/data-science

Luna Rizo, M., Daza Ramírez, M. T., & Lozoya Arandia, J. (Coords.). (2024). Tendencias de la inteligencia artificial en educación. Universidad de Guadalajara. ISBN: 978-607-581-379-0

Márquez Vera, C. (2015). Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos [Tesis doctoral, Universidad de Córdoba]. Helvia Repositorio Institucional. https://helvia.uco.es/handle/10396/12418

Martínez, C. (2022). *Fundamentos de investigación educativa*. Universidad Nacional Abierta y a Distancia.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22 (3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Medina Romero, M. Á., & Ochoa Figueroa, R. (2025). Inteligencia artificial en educación: Innovaciones y su impacto pedagógico (1.ª ed.). Know Press. https://doi.org/10.70180/978-9942-7389-0-5

Microsoft. (s.f.). Documentación de Microsoft Azure. Microsoft Learn. https://learn.microsoft.com/es-mx/azure/?product=popular

Microsoft. (s.f.). Features — LightGBM documentation. ReadTheDocs. Recuperado el 13 de mayo de 2025, de https://lightgbm.readthedocs.io/en/latest/Features.html

Matplotlib Developers. (s.f.). Matplotlib: Visualization with Python. Recuperado el 19 de mayo de 2025, de https://matplotlib.org

Naciones Unidas. (s.f.). Declaración Universal de los Derechos Humanos. https://www.un.org/es/about-us/universal-declaration-of-human-rights

Navlani, A., Fandango, A., & Idris, I. (2021). Python data analysis (3rd ed.). Packt Publishing.

NumPy Developers. (s.f.). NumPy. Recuperado el 23 de mayo de 2025, de https://numpy.org

Ortiz Ocaña, A. (2025). Inteligencia artificial aplicada a la educación: Manual para docentes, estudiantes y directivos. Ecoe Ediciones.

Pandas Development Team. (s.f.). Pandas: Python data analysis library. Recuperado el 23 de mayo de 2025, de https://pandas.pydata.org

Ponce Gallegos, J. C., & Torres Soto, A. (2014). Inteligencia artificial. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn). http://www.proyectolatin.org/

Poole, D., & Mackworth, A. (2023). Artificial Intelligence: Foundations of Computational Agents (3rd ed.). Cambridge University Press.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning (2nd ed.). Packt Publishing

Red de Pobreza Multidimensional (MPPN). (s.f.). ¿Qué es el IPM?. https://www.mppn.org/es/pobreza-multidimensional/que-es-el-ipm/

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3 (1), 12–27.

Sampieri, R. H., Collado, C. F., & Lucio, M. P. B. (2022). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta (7.ª ed.). McGraw-Hill Education.

Santos, M., & Azevedo, C. (2005). KDD, SEMMA and CRISP–DM: A Parallel Overview. Repositório Científico do Instituto Politécnico do Porto. https://core.ac.uk/download/pdf/47135941.pdf

Scikit learn developers. (s. f.). Precision, recall and F–measures. En Model evaluation (Sección 3.4.4.9)

Scikit–learn. (s.f.). Supervised learning. https://scikit-learn.org/stable/supervised_learning.html

Secretaría de Educación Pública (SEP). (2023). Normas específicas de control escolar relativas a la inscripción, reinscripción, acreditación, promoción, regularización y certificación en la educación básica (pp. 27–33). Dirección General de Acreditación, Incorporación y Revalidación.

Secretaría de Educación Pública (SEP). (s.f.). ¿Sabes qué es el Consejo Técnico Escolar (CTE)? Gobierno de México. https://www.gob.mx/sep/articulos/sabes-que-es-el-consejo-tecnico-escolar-cte?idiom=es

Szymkowiak, A. (2019). Discovering knowledge from data. Bogucki Wydawnictwo Naukowe. https://doi.org/10.12657/9788379862788

Tamayo y Tamayo, M. (2005). El proceso de la investigación científica (5.ª ed.). Limusa.

UNESCO. (2021). Recomendación sobre la ética de la inteligencia artificial* Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.

Unidad de Servicios para la Educación Básica en el Estado de Querétaro (USEBEQ). (2020). Reglamento interior de la Unidad de Servicios para la Educación Básica en el Estado de Querétaro (septiembre 2020) (p. 6). Periódico Oficial “La Sombra de Arteaga”.

Unidad de Servicios para la Educación Básica en el Estado de Querétaro (USEBEQ). (2021). Programa Institucional USEBEQ 2021–2027 (p. 5). https://www.usebeq.edu.mx/Content/Intranet/PROG%20INSTITUCIONAL%20DE%20USEBEQ%2021-27%20vf.pdf

Unidad de Servicios para la Educación Básica en el Estado de Querétaro (USEBEQ). (s.f.). Manual de calidad del Sistema de Gestión de la Calidad. https://www.usebeq.edu.mx/Content/SGC/Documentos/ManualCalidad.pdf

Unidad de Servicios para la Educación Básica en el Estado de Querétaro (USEBEQ). (2021). Programa Institucional de la USEBEQ 2021–2027. https://www.usebeq.edu.mx/Content/Intranet/PROG%20INSTITUCIONAL%20DE%20USEBEQ%2021-27%20vf.pdf

Universidad Autónoma de Querétaro (UAQ). (s.f.). Maestría en Gestión e Innovación Pública. Recuperado el 18 de mayo de 2025, de https://www.uaq.mx/index.php/nivel-posgrados/maestrias/fcya/maestria-en-gestion-e-innovacion-publica

Waskom, M. (2021). Seaborn: Statistical data visualization (versión 0.11.2) [Software]. https://seaborn.pydata.org

ANEXOS

ANEXO A: Siglas y Abreviaturas

Sigla / Abreviatura	Significado
AI	Inteligencia Artificial
ANN	Artificial Neural Network (Red Neuronal Artificial)
API	Application Programming Interface (Interfaz de Programación de Aplicaciones)
AUC	Area Under the Curve (Área Bajo la Curva)
CANVAS	Lienzo de Modelo de Negocio
CRISP–DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Values (Valores Separados por Comas)
CTE	Consejo Técnico Escolar
DTIC	Dirección de Tecnologías de la Información y Comunicaciones
EDA	Exploratory Data Analysis (Análisis Exploratorio de Datos)
F1–score	Medida F1 (Media armónica entre precisión y recall)
FODA	Fortalezas, Oportunidades, Debilidades y Amenazas
IA	Inteligencia Artificial
IQR	Interquartile Range (Rango Intercuartílico)
KS	Kolmogorov–Smirnov (Estadístico de comparación de distribuciones)
MCC	Matthews Correlation Coefficient (Coeficiente de Correlación de Matthews)
ML	Machine Learning (Aprendizaje Automático)
ONGs	Organizaciones No Gubernamentales
PR AUC	Área bajo la curva Precisión–Recall
ROC AUC	Área bajo la curva ROC (Receiver Operating Characteristic)

SHAP	SHapley Additive exPlanations (Explicaciones Aditivas de Shapley)
SMOTE	Synthetic Minority Over-sampling Technique (Técnica de sobremuestreo sintético)
SQL	Structured Query Language (Lenguaje de Consulta Estructurado)
UAQ	Universidad Autónoma de Querétaro
USEBEQ	Unidad de Servicios para la Educación Básica en el Estado de Querétaro