



**UNIVERSIDAD LAICA ELOY ALFARO DE MANABÍ**

**Facultad:**

**CIENCIAS DE LA VIDA Y TECNOLOGÍAS**

**Carrera:**

**TECNOLOGÍAS DE LA INFORMACIÓN**

**Tema:**

Minería de datos educacional para la detección temprana de barreras en el aprovechamiento estudiantil en la Unidad Educativa Fiscomisional Juan Montalvo.

Trabajo de titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en Tecnologías de la Información.

---

**Autor/es**

Campozano López Samuel Segundo

Vera Parrales Jonathan Marcos

**Tutor**

Ing. Pincay Ponce Jorge Ivan, Ph.D.

2026



## **Dedicatoria**

A nuestros padres, por ser el ejemplo de esfuerzo y perseverancia que nos ha guiado hasta aquí. Gracias por su apoyo silencioso, por su paciencia en los momentos difíciles y por creer en nosotros incluso cuando dudábamos. Este logro es el resultado de su sacrificio y amor incondicional; cada página de este trabajo lleva implícita su esencia y sus enseñanzas.

A nuestras familias, por ser nuestro refugio y fuente de motivación constante. Gracias por comprender nuestras ausencias, por alentarnos a no rendirnos y por celebrar cada pequeño avance como si fuera el más grande. Hoy, este triunfo es tan suyo como nuestro.



## **Agradecimiento**

A nuestra alma máter, la Universidad Laica Eloy Alfaro de Manabí, y en especial a la Facultad de Ciencias de la Vida y Tecnologías. Gracias por brindarnos los espacios de aprendizaje, las herramientas tecnológicas y el entorno académico necesarios para nuestra formación profesional. En sus aulas no solo adquirimos conocimientos, sino también los valores que hoy nos definen como futuros profesionales.

Expresamos nuestra profunda gratitud a nuestro tutor de tesis, por su acertada dirección y acompañamiento durante todo este proceso. Gracias por habernos guiado con paciencia, por su rigor académico y por compartir generosamente su experiencia y conocimiento con nosotros, lo cual fue fundamental para llevar este proyecto a buen puerto.

De igual manera, a la Unidad Educativa Fiscomisional Juan Montalvo, extendemos nuestro agradecimiento por abrirnos sus puertas con confianza. Gracias a sus autoridades y personal por facilitarnos el acceso a la información y los recursos necesarios para validar este estudio, permitiendo que nuestra investigación tenga un impacto real en su comunidad educativa.



## CERTIFICACIÓN

En calidad de docente tutor de la Facultad de Ciencias de la Vida y Tecnología de la Universidad Laica “Eloy Alfaro” de Manabí, CERTIFICO:

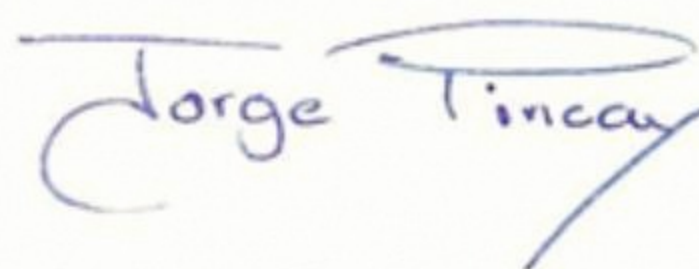
Haber dirigido, revisado y aprobado preliminarmente el Trabajo de Integración Curricular bajo la autoría del estudiante **VERA PARRALES JONATHAN MARCOS**, legalmente matriculado en la carrera de tecnologías de la información, período académico 2025 (2), cumpliendo el total de 384 horas, cuyo tema del proyecto es **“MINERÍA DE DATOS EDUCACIONAL PARA LA DETECCIÓN TEMPRANA DE BARRERAS EN EL APROVECHAMIENTO ESTUDIANTIL EN LA UNIDAD EDUCATIVA FISCOMISIONAL JUAN MONTALVO”**.

La presente investigación ha sido desarrollada en apego al cumplimiento de los requisitos académicos exigidos por el Reglamento de Régimen Académico y en concordancia con los lineamientos internos de la opción de titulación en mención, reuniendo y cumpliendo con los méritos académicos, científicos y formales, y la originalidad de este, requisitos suficientes para ser sometida a la evaluación del tribunal de titulación que designe la autoridad competente.

Particular que certifico para los fines consiguientes, salvo disposición de Ley en contrario.

Manta, 23 de enero de 2026.

Lo certifico,



Ing. Pincey Ponce Jorge Ivan, PhD.

**Docente Tutor**

**Facultad de Ciencias de la Vida y Tecnología**  
**Carrera Tecnologías de la Información**



## CERTIFICACIÓN

En calidad de docente tutor de la Facultad de Ciencias de la Vida y Tecnología de la Universidad Laica "Eloy Alfaro" de Manabí, CERTIFICO:

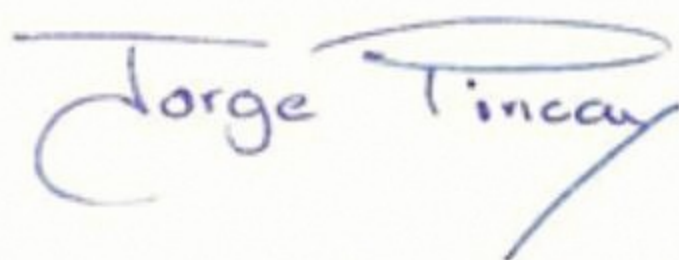
Haber dirigido, revisado y aprobado preliminarmente el Trabajo de Integración Curricular bajo la autoría del estudiante **CAMPOZANO LOPEZ SAMUEL SEGUNDO**, legalmente matriculado en la carrera de tecnologías de la información, período académico 2025 (2), cumpliendo el total de 384 horas, cuyo tema del proyecto es **"MODELO DE MINERIA DE DATOS EDUCACIONAL PARA LA DETECCION TEMPRANA DE BARRERAS QUE PUEDEN OBSTACULIZAR EL APROVECHAMIENTO ESTUDIANTIL"**.

La presente investigación ha sido desarrollada en apego al cumplimiento de los requisitos académicos exigidos por el Reglamento de Régimen Académico y en concordancia con los lineamientos internos de la opción de titulación en mención, reuniendo y cumpliendo con los méritos académicos, científicos y formales, y la originalidad de este, requisitos suficientes para ser sometida a la evaluación del tribunal de titulación que designe la autoridad competente.

Particular que certifico para los fines consiguientes, salvo disposición de Ley en contrario.

Manta, 23 de enero de 2026.

Lo certifico,



Ing. Pincay Ponce Jorge Ivan, PhD.  
**Docente Tutor**  
**Facultad de Ciencias de la Vida y Tecnología**  
**Carrera Tecnologías de la Información**



### **Declaración expresa de autoría**

Nosotros, Campozano López Samuel Segundo, con cédula de ciudadanía 172959632-8 y Vera Parrales Jonathan Marcos, con cédula de ciudadanía 131636114-4, en calidad de autores del trabajo de titulación:

"Minería de datos educacional para la detección temprana de barreras en el aprovechamiento estudiantil en la Unidad Educativa Fiscomisional Juan Montalvo".

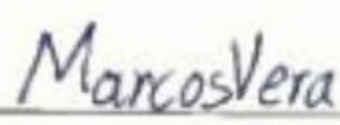
Declaramos que este trabajo es original y de nuestra autoría, hemos citado correctamente las fuentes utilizadas y respetado las leyes vigentes sobre derechos de autor.

Autorizamos a la Universidad Laica "Eloy Alfaro" de Manabí a usar este trabajo con fines académicos o investigativos, así como a publicarlo en su repositorio digital.



---

Campozano López Samuel Segundo  
CI: 1729596328



---

Vera Parrales Jonathan Marcos  
C.I: 1316361144





## ÍNDICE

|                                                                             |           |
|-----------------------------------------------------------------------------|-----------|
| <b>Capítulo I: Introducción .....</b>                                       | <b>15</b> |
| 1.1    Introducción .....                                                   | 15        |
| 1.2    Presentación del tema .....                                          | 15        |
| 1.3    Ubicación y contextualización de la problemática .....               | 16        |
| 1.4    Planteamiento del problema .....                                     | 17        |
| 1.4.1    Problematización .....                                             | 17        |
| 1.4.2    Génesis del problema .....                                         | 17        |
| 1.4.3    Estado actual del problema .....                                   | 17        |
| 1.5.    Diagrama causa – efecto del problema .....                          | 19        |
| 1.6    Objetivos .....                                                      | 20        |
| 1.7    Justificación.....                                                   | 21        |
| 1.8    Impactos esperados .....                                             | 22        |
| 1.8.1    Impacto tecnológico .....                                          | 22        |
| 1.8.2    Impacto social .....                                               | 22        |
| 1.8.3    Impacto ecológico .....                                            | 23        |
| <b>Capítulo II: Marco teórico de la investigación.....</b>                  | <b>24</b> |
| 2.1    Antecedentes históricos .....                                        | 24        |
| 2.2    Antecedentes de investigaciones relacionadas al tema presentado..... | 25        |
| 2.3    Definiciones conceptuales .....                                      | 27        |
| 2.3.1    Educational Data Mining (EDM).....                                 | 27        |





|                                                                             |           |
|-----------------------------------------------------------------------------|-----------|
| 2.3.2 Técnicas de Machine Learning en Educación .....                       | 27        |
| 2.3.3 Variables Predictivas en Contextos Educativos .....                   | 29        |
| 2.3.4 Explicabilidad en Modelos Educativos (SHAP Values) .....              | 30        |
| 2.4 Conclusiones del marco teórico .....                                    | 31        |
| <b>Capítulo III: Marco investigativo.....</b>                               | <b>32</b> |
| 3.1 Introducción .....                                                      | 32        |
| 3.2 Tipo de investigación .....                                             | 32        |
| 3.3. Método(s) de investigación .....                                       | 32        |
| 3.4 Fuentes de información de datos .....                                   | 33        |
| 3.5 Estrategia operacional para la recolección de datos .....               | 34        |
| 3.5.1 Población y muestra.....                                              | 34        |
| 3.5.2 Análisis de las herramientas de recolección de datos a utilizar ..... | 34        |
| 3.5.3 Plan de recolección de datos .....                                    | 39        |
| 3.6 Análisis y presentación de resultados.....                              | 40        |
| 3.6.1 Tabulación y análisis de los datos .....                              | 40        |
| 3.6.2 Informe final del análisis de los datos.....                          | 45        |
| <b>Capítulo IV: Marco propositivo .....</b>                                 | <b>47</b> |
| 4.1 Introducción.....                                                       | 47        |
| 4.2 Descripción de la propuesta .....                                       | 47        |
| 4.3 Determinación de recursos.....                                          | 48        |
| 4.3.1 Humanos .....                                                         | 48        |





|                                                                     |           |
|---------------------------------------------------------------------|-----------|
| 4.3.2 Tecnológicos.....                                             | 49        |
| 4.3.3 Económicos (presupuesto) .....                                | 50        |
| 4.4 Etapas de acción para el desarrollo de la propuesta .....       | 50        |
| 4.4.1 Fase I Preparación de datos y redefinición del problema ..... | 50        |
| 4.4.2 Fase II Desarrollo y validación del modelo predictivo .....   | 50        |
| 4.4.3 Fase III Desarrollo y validación del modelo predictivo .....  | 51        |
| <b>Capítulo V: Evaluación de resultados.....</b>                    | <b>53</b> |
| 5.1 Introducción.....                                               | 53        |
| 5.2 Presentación y monitoreo de resultados.....                     | 53        |
| 5.3 Interpretación objetiva .....                                   | 55        |
| <b>Capítulo VI: Conclusiones y recomendaciones .....</b>            | <b>57</b> |
| 6.1 Conclusiones .....                                              | 57        |
| 6.2 Recomendaciones .....                                           | 60        |
| <b>6.3 Bibliografía .....</b>                                       | <b>62</b> |





## ÍNDICE TABLAS

|                                                                          |    |
|--------------------------------------------------------------------------|----|
| Tabla 1 Comparación de Modelos Minimalistas vs. Complejos .....          | 25 |
| Tabla 2 Distribución del Sistema Educativo Ecuatoriano.....              | 26 |
| Tabla 3 Comparación de Kernels SVM en EDM.....                           | 28 |
| Tabla 4 Impacto de Variables Socioeconómicas (Valores SHAP).....         | 29 |
| Tabla 5 Distribución de la Muestra por Niveles .....                     | 34 |
| Tabla 6 Variables de Características de la Vivienda .....                | 35 |
| Tabla 7 Variables de Acceso a Servicios .....                            | 35 |
| Tabla 8 Variables de Posesión de Bienes.....                             | 36 |
| Tabla 9 Cronograma de Actividades 2025 .....                             | 37 |
| Tabla 10 Estadísticas Descriptivas de Variables Numéricas .....          | 38 |
| Tabla 11 Resultados del Análisis ANOVA por Grupo Socioeconómico.....     | 38 |
| Tabla 12 Impacto de Tenencia de Laptop en Rendimiento .....              | 39 |
| Tabla 13 Rendimiento por Nivel Educativo del Representante.....          | 39 |
| Tabla 14 Impacto de Barreras Acumulativas en el Rendimiento .....        | 40 |
| Tabla 15 Métricas de Rendimiento del Modelo de Regresión Logística ..... | 40 |
| Tabla 16 : Top 10 Coeficientes del Modelo (Regresión Logística) .....    | 42 |
| Tabla 17 Importancia de Variables en el RiskCalculator.....              | 42 |
| Tabla 18 Impacto de barreras acumulativas en el rendimiento .....        | 43 |
| Tabla 19 Metricas de rendimiento del modelo de regresion logistica ..... | 53 |



## ÍNDICE GRÁFICOS E ILUSTRACIONES

|                                                                                        |           |
|----------------------------------------------------------------------------------------|-----------|
| <b>Ilustración 1 Diagrama causa efecto. ....</b>                                       | <b>19</b> |
| <b>Ilustración 2 Ciclo de vida de la metodología CRIP-DM .....</b>                     | <b>33</b> |
| <b>Ilustración 3 Relación entre el Nivel Educativo del Representante y el GPA.....</b> | <b>43</b> |
| <b>Ilustración 4 Impacto de Barreras Acumulativas en el Rendimiento. ....</b>          | <b>44</b> |
| <b>Ilustración 5 Arquitectura del Sistema (Flujo de datos y motor híbrido) .....</b>   | <b>52</b> |



## Resumen

La presente investigación desarrolla e implementa un modelo integral de Minería de Datos Educacional (EDM) para la detección temprana de barreras académicas en la Unidad Educativa Fiscomisional Juan Montalvo. El estudio adopta la metodología CRISP-DM como marco metodológico principal, analizando 687 estudiantes con 47 variables socioeconómicas para la predicción del riesgo académico.

El desarrollo del modelo no fue lineal, sino que pasó por dos etapas bien diferenciadas. En una primera fase, se trabajó con un modelo CatBoost aplicado al nivel materia–estudiante, con el cual se obtuvo un desempeño alto, alcanzando un ROC-AUC de 0.934 y un Recall del 75%. Pero luego de identificar puntos débiles respecto a esa primera etapa, en una segunda fase se optó por simplificar el enfoque, teniendo que replantear de nuevo el problema como una clasificación binaria a nivel de estudiante (en riesgo y no en riesgo) para comparar luego el rendimiento de 7 algoritmos de aprendizaje de máquina. Donde la Regresión Logística logró mejores resultados que los modelos de ensamble, obteniendo el mayor puntaje de validación cruzada ( $CV\ 0.681 \pm 0.033$ ). Este resultado evidencia que, cuando se trabaja con conjuntos de datos pequeños, los modelos más simples y con regularización suelen generalizar mejor que los algoritmos más complejos.

Del total de estudiantes analizados, el 37% fue identificado como en riesgo académico, considerando como criterio cualquier calificación igual o menor a 7.5. Al optimizar el umbral de decisión en 0.25, el modelo logró un Recall del 92.2%, identificando correctamente a 47 de los 51 estudiantes en riesgo dentro del conjunto de prueba, con únicamente 4 casos no detectados. Para garantizar la validez del sistema, se aplicaron medidas estrictas para evitar la fuga de datos, excluyendo las calificaciones del conjunto de variables predictoras, porque estas componen la variable objetivo.



El sistema se implementó con Flask para el backend y React para el frontend. Además, se integró un motor de reglas heurísticas (RiskCalculator) que permite generar alertas en tiempo real, junto con el modelo de Machine Learning encargado de realizar predicciones anticipadas utilizando únicamente información socioeconómica



## Abstract

This research develops and implements a comprehensive Educational Data Mining (EDM) model for the early detection of academic barriers at Unidad Educativa Fiscomisional Juan Montalvo. The study adopts the CRISP-DM methodology as its primary framework, analyzing 687 students with 47 socioeconomic variables for academic risk prediction.

The development of the model wasn't a linear thing at all; it took two different phases. In Phase 1, a CatBoost model was implemented at the subject-student level, achieving an ROC-AUC of 0.934 with 75% Recall. Then the second part followed a different approach in which it got simplified to a binary student-level classification (at-risk vs. not at risk), this was done by comparing 7 Machine Learning algorithms. Logistic Regression outperformed ensemble models, achieving the best cross-validation score (CV  $0.681 \pm 0.033$ ), this demonstrates that for small datasets, a simple regularized model can generalize better than complex algorithms.

From the total amount of students 37% of them were classified as at-risk (any grade  $\leq 7.5$ ). Through threshold optimization (0.25), the model achieves 92.2% Recall, detecting 47 of 51 at-risk students in the test set, with only 4 students missing. Rigorous data leakage prevention measures were implemented, excluding grades as predictors since they constitute the target variable.

The whole system was deployed using Flask (backend) and React (frontend) as architecture, integrating a heuristic rules engine (RiskCalculator) for prompt alerts and the ML model for proactive prediction based on socioeconomic factors known at enrollment.



## Capítulo I: Introducción

### 1.1 Introducción

Para nadie es sorpresa que la educación se encuentra en un periodo de transformación digital, los centros de enseñanza enfrentan el desafío crítico de identificar tempranamente las barreras que pueden obstaculizar el aprovechamiento estudiantil. La Unidad Educativa Fiscomisional Juan Montalvo, con aproximadamente 900 estudiantes distribuidos en niveles de educación primaria y secundaria, representa un caso típico de las limitaciones que enfrentan las instituciones educativas ecuatorianas en la gestión y análisis de datos académicos y socioeconómicos.

Esta tesis se organiza en seis capítulos principales. Tras esta introducción, el Capítulo II presenta el marco teórico comprehensivo sobre EDM, metodologías de minería de datos y el estado del arte en sistemas predictivos educativos. El Capítulo III detalla el marco investigativo, abarcando la metodología, el análisis y la exploración de datos. El Capítulo IV describe el marco propositivo, incluyendo el desarrollo técnico del sistema, arquitectura, componentes y la integración de modelos. El Capítulo V presenta la evaluación de los resultados experimentales y la validación del modelo. Finalmente, el Capítulo VI presenta las conclusiones y recomendaciones del estudio.

### 1.2 Presentación del tema

La Minería de Datos Educativa (EDM) ha evolucionado significativamente en la última década. Romero & Ventura (2020) definen EDM como “la aplicación de técnicas de minería de datos para explorar los tipos únicos de datos que provienen de contextos educativos”. Estos avances demuestran el potencial transformador de las técnicas de *machine learning* en contextos educativos.



Como todo en el mundo de la tecnología, dicha definición ha evolucionado para incluir la predicción del rendimiento académico y el análisis de la deserción escolar, evidenciando una oportunidad significativa para investigaciones integradas. Por esta razón, los modelos predictivos y de aprendizaje automático aparecen como una alternativa prometedora para capturar las relaciones multifactoriales (como la situación socioeconómica familiar y el acceso a herramientas tecnológicas) que inciden en el riesgo de fracaso académico.

Por lo tanto, la presente investigación desarrolla e implementa un modelo integral de Minería de Datos Educativo (EDM) basado en la metodología CRISP-DM para la detección temprana de barreras académicas en la Unidad Educativa Fiscomisional Juan Montalvo, integrando técnicas avanzadas de *machine learning* con variables socioeconómicas

### **1.3 Ubicación y contextualización de la problemática**

Según datos oficiales del Instituto Nacional de Estadística y Censos (Instituto Nacional de Estadística y Censos, 2022), el 4.1% de la población escolar ecuatoriana, equivalente a 195,188 estudiantes entre 5 y 17 años, abandona sus estudios anualmente, siendo la pobreza identificada como el factor principal de deserción. Esta realidad se agudiza en las instituciones fiscomisionales, que representan el 9.34% del sistema educativo nacional y atienden a una población estudiantil con alta heterogeneidad socioeconómica.

La Unidad Educativa Fiscomisional Juan Montalvo, el colegio objeto de estudio, enfrenta los siguientes desafíos documentados:

- Sistema de gestión académica (Idukay) con capacidades limitadas de análisis predictivo.
- Registro manual de información socioeconómica mediante formularios físicos.
- Ausencia de herramientas para identificación temprana de factores de riesgo.



- Incapacidad para diseñar intervenciones personalizadas basadas en evidencia.

## **1.4 Planteamiento del problema**

Ecuador presenta particularidades importantes en su sistema educativo. Avilés-Salvador et al. (2020) analizaron datos de 281,373 estudiantes del examen Ser Bachiller entre 2016 y 2017, encontrando que el 69.51% provenía de instituciones públicas, el 21.15% de privadas y el 9.34% de fiscomisionales. Estos diferentes tipos de establecimientos mostraron patrones de rendimiento distintos, lo que justifica desarrollar modelos específicos para cada realidad

### **1.4.1 Problematicación**

Los datos iniciales indican que factores como la situación socioeconómica familiar, el acceso a herramientas tecnológicas y la escolaridad de los padres inciden considerablemente en el riesgo de fracaso académico. La pobreza ha sido identificada como el factor principal de deserción, una realidad que se agudiza fuertemente en las instituciones fiscomisionales que atienden a una población estudiantil con alta heterogeneidad socioeconómica

### **1.4.2 Génesis del problema**

El origen de este problema radica en que las instituciones fiscomisionales ecuatorianas presentan características únicas: tienen una mayor heterogeneidad socioeconómica que las instituciones públicas y cuentan con recursos más limitados que las instituciones privadas. Debido a estas limitaciones, resulta particularmente relevante y necesario adoptar enfoques analíticos o modelos simplificados que se ajusten a la escasez de recursos de las instituciones fiscomisionales ecuatorianas para poder identificar las verdaderas causas del bajo rendimiento

### **1.4.3 Estado actual del problema**



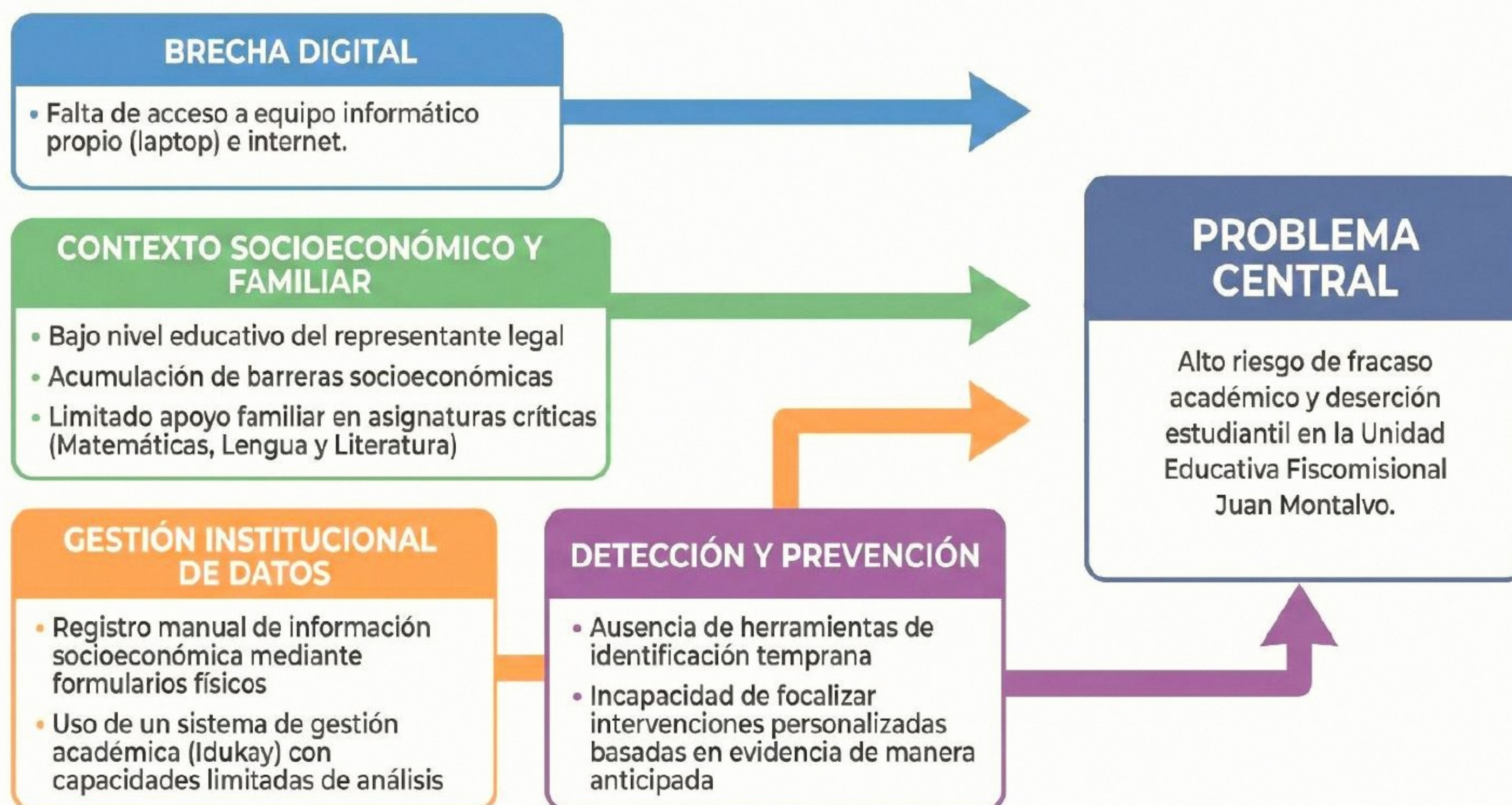
En la actualidad, por esta razón, los modelos de *machine learning* aparecen como una alternativa prometedora para capturar estas relaciones multifactoriales. Con la implementación de este sistema predictivo se busca detectar estudiantes en riesgo académico con 2.5 meses de antelación y disminuir la deserción escolar en al menos un 30%. Además, se pretende mejorar cómo el Departamento de Consejería Estudiantil distribuye sus recursos actuales y crear un modelo que otras instituciones del país puedan adaptar.



## 1.5. Diagrama causa – efecto del problema

### Ilustración 1

*Diagrama causa efecto*



El diagrama causa-efecto permite identificar los principales factores que influyen en el riesgo de fracaso académico y la potencial deserción escolar en la Unidad Educativa Fiscomisional Juan Montalvo. En él se evidencian causas estrechamente relacionadas con la brecha digital de los estudiantes, el contexto socioeconómico familiar (como el nivel educativo del representante legal y el efecto de las barreras acumulativas), así como limitaciones institucionales reflejadas en el registro manual de datos y la falta de un análisis predictivo en los sistemas de gestión actuales. La identificación de estas causas fundamenta la necesidad urgente de implementar un sistema



predictivo basado en Minería de Datos Educacional (EDM) y metodologías ágiles que actúe como una herramienta de alerta temprana para mitigar estos factores de riesgo.

## 1.6 Objetivos

### Objetivo general

Desarrollar e implementar un modelo integral de Minería de Datos Educacional basado en la metodología CRISP-DM para la detección temprana de barreras académicas en la Unidad Educativa Fiscomisional Juan Montalvo, integrando técnicas avanzadas de machine learning con variables socioeconómicas del modelo INEC Ecuador.

### Objetivos específicos

1. Implementar las seis fases de la metodología CRISP-DM adaptadas al contexto educativo ecuatoriano, estableciendo procesos replicables para la comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.
2. Crear un software utilizando las tecnologías de React, Flask y Supabase que permitan la captura, procesamiento y visualización de datos para 687 estudiantes con 47 variables socioeconómicas
3. Realizar pruebas y comparaciones en profundidad entre 7 algoritmos de Machine Learning (Regresión Logística, Random Forest, CatBoost, XGBoost, LightGBM, Gradient Boosting) para identificar el modelo óptimo en términos de validación cruzada y capacidad de generalización
4. Validar el sistema mediante métricas estándar (ROC-AUC, Recall, Precisión, F1-score) y optimización de umbral, con el objetivo de maximizar la detección de estudiantes en riesgo ( $\text{Recall} \geq 90\%$ )



5. Analizar e interpretar los coeficientes del modelo final para detectar qué factores incrementan o reducen el riesgo académico, y con esa información se propondrán acciones específicas para el Departamento de Consejería Estudiantil.

### 1.7 Justificación

La Minería de Datos Educacional ha evolucionado significativamente en la última década. Malik et al. (2025) introdujeron recientemente el concepto de Dynamic Feature Ensemble Evolution (DE-FS), logrando precisiones del 95.8% en matemáticas y 95.3% en portugués mediante la combinación de análisis de correlación dinámico, information gain adaptativo y validación cruzada estratificada. Estos avances demuestran el potencial transformador de las técnicas de machine learning en contextos educativos.

De igual manera, Yağcı (2022) demostró que modelos simplificados con solo tres variables (calificaciones parciales, departamento y facultad) pueden alcanzar precisiones del 70-75%, proporcionando 2.5 meses para intervenciones correctivas. Esta aproximación minimalista resulta particularmente relevante para instituciones con recursos limitados como las fiscomisionales ecuatorianas.

La adopción de CRISP-DM como metodología principal responde a su consolidación como estándar de facto en proyectos de minería de datos. Según Martínez-Plumed et al. (2021), CRISP-DM posee una utilización que se segmenta de la siguiente manera, 65% de los proyectos industriales y el 58% de los proyectos académicos de EDM a nivel mundial. Su estructura de seis fases proporciona un marco sistemático y replicable que garantiza la rigurosidad científica del proceso.



Este trabajo se conecta con el marco KDD propuesto por Fayyad et al. en 1996, que marcó las bases del descubrimiento de conocimiento en bases de datos. También incorpora elementos del enfoque OSEMN de Mason y Wiggins (2010), que promueve ciclos iterativos y experimentación en proyectos de datos.

## **1.8 Impactos esperados**

La implementación de este modelo integral de Minería de Datos Educacional en la Unidad Educativa Fiscomisional Juan Montalvo busca generar impactos significativos en el ámbito tecnológico, social y ecológico, transformando la manera en que la institución gestiona la información y previene el fracaso académico.

### **1.8.1 Impacto tecnológico**

El proyecto representa una innovación tecnológica importante al sustituir la dependencia de un sistema de gestión académica (Idukay) con capacidades limitadas de análisis, por un sistema predictivo de alerta temprana avanzado. Al integrar técnicas de *machine learning* y desarrollar un software utilizando React, Flask y Supabase, la institución adopta herramientas de última generación para la captura, procesamiento y visualización de datos. Además, crea un modelo tecnológico escalable y estandarizado bajo CRISP-DM que otras instituciones del país podrán adaptar fácilmente a sus realidades.

### **1.8.2 Impacto social**

Desde una perspectiva social, con este sistema predictivo se busca detectar estudiantes en riesgo académico con 2.5 meses de antelación y disminuir la deserción escolar en al menos un



30%. Además, pretende mejorar cómo el Departamento de Consejería Estudiantil (DCE) distribuye sus limitados recursos. Esto permitirá focalizar las intervenciones pedagógicas y psicosociales directamente en los estudiantes que presentan mayores barreras socioeconómicas acumulativas, garantizando una educación más equitativa e inclusiva

### **1.8.3 Impacto ecológico**

El impacto ecológico del proyecto es sumamente favorable, ya que promueve la digitalización de los procesos de recolección de información institucional. Anteriormente, el colegio realizaba el registro manual de información socioeconómica y académica mediante formularios físicos. Con la implementación de la plataforma web y la digitalización del Formulario Socioeconómico del INEC, se elimina drásticamente la necesidad de utilizar papel impreso y otros consumibles, fomentando una gestión administrativa cero-papel y sustentable con el medio ambiente.



## Capítulo II: Marco teórico de la investigación

### 2.1 Antecedentes históricos

El campo de la Minería de Datos Educativa se origina del proceso KDD (Knowledge Discovery in Databases), que Fayyad et al. (1996) propusieron como un proceso de nueve pasos que se repite de manera iterativa con el objetivo de extraer conocimiento útil desde grandes bases de datos. Este proceso KDD fue la base que después evolucionó hacia metodologías con mayor estructura y que estaban más orientadas hacia aplicaciones industriales.

Para este campo una transición desde KDD a CRISP-DM es una demostración de la profesionalización de este. Cuando por un lado KDD proporcionó el marco teórico con sus fases de selección, preprocesamiento, transformación, minería de datos y evaluación, del otro lado tenemos a CRISP-DM que emergió en 1996 como respuesta a la necesidad de una metodología estandarizada y práctica. Desarrollado por un consorcio liderado por NCR, DaimlerChrysler (ahora Mercedes-Benz), SPSS (ahora IBM) y OHRA, CRISP-DM se diseñó específicamente para proyectos industriales de minería de datos (Chapman et al., 2000; Wirth & Hipp, 2000).

Adicionalmente a esta evolución, existe una metodología llamada OSEMN, introducida más recientemente por Mason & Wiggins (2010), la cual representa una simplificación orientada a la ciencia de datos moderna:

- **Obtain (Obtener):** Adquisición de datos desde múltiples fuentes.
- **Scrub (Limpiar):** Limpieza y preparación de datos.
- **Explore (Explorar):** Análisis exploratorio y visualización.
- **Model (Modelar):** Construcción y entrenamiento de modelos.



- **Interpret (Interpretar):** Interpretación y comunicación de resultados.

## 2.2 Antecedentes de investigaciones relacionadas al tema presentado

Los avances más significativos en EDM se han producido en los últimos cinco años, con mejoras sustanciales en precisión predictiva. Malik et al. (2025) introdujeron DE-FS, logrando precisiones del 95.8% en matemáticas y 95.3% en portugués. Su metodología combina análisis de correlación dinámico, procesos de extracción de información adaptativos, Chi-square con validación cruzada, regularización L1/L2 y un ensemble de Random Forest, XGBoost y Neural Networks.

Modelos Minimalistas para Recursos Limitados Por otro lado, Yağcı (2022) demostró que modelos simplificados pueden ser altamente efectivos:

**Tabla 1**

*Comparación de modelos minimalistas vs complejos*

| Característica           | Minimalista | Complejo   | Diferencia    |
|--------------------------|-------------|------------|---------------|
| Variables utilizadas     | 3           | 47         | -93.6%        |
| Precisión alcanzada      | 75%         | 92%        | -17%          |
| Tiempo de entrenamiento  | 2 minutos   | 45 minutos | 95.6%         |
| Recursos computacionales | Bajo        | Alto       | Significativo |
| Interpretabilidad        | Alta        | Baja       | Mejor         |

*Fuente: Yağcı (2022)*



Sistemas de Alerta Temprana Rienties: et al. (2020) documentaron la implementación más grande reportada hasta la fecha (Open University UK, 170,000+ alumnos), logrando un porcentaje de precisión promedio del 89% y una reducción de deserción del 32%. Asimismo, Angeioplastis et al. (2025) logró identificar factores críticos de éxito a través del análisis de 450 estudiantes: acciones preventivas tempranas (semana 4), personalización de mensajes, combinación de distintas variables y retroalimentación continua.

Contexto Latinoamericano y Ecuatoriano: Huanca-Anquise et al. (2022) documentaron la implementación del sistema nacional de alerta temprana en Chile, con 365,854 estudiantes analizados y una precisión de 78%, logrando una reducción de 25% en abandono escolar tras las intervenciones a tiempo.

Específicamente en Ecuador, Avilés-Salvador et al. (2020) analizaron el sistema educativo con datos de 281,373 estudiantes del examen Ser Bachiller 2016-2017:

**Tabla 2**

*Distribución del sistema educativo ecuatoriano*

| <b>Tipo de Institución</b> | <b>Estudiantes</b> | <b>Porcentaje</b> | <b>Rendimiento Promedio</b> |
|----------------------------|--------------------|-------------------|-----------------------------|
| Pública                    | 195,683            | 69.51%            | 7.2/10                      |
| Privada                    | 59,530             | 21.15%            | 8.1/10                      |
| Fiscomisional              | 26,260             | 9.34%             | 7.6/10                      |

*Fuente.* Avilés-Salvador et al. (2020)

Las instituciones fiscomisionales presentan características únicas: mayor heterogeneidad socioeconómica que las públicas y recursos más limitados que las



instituciones privadas. Esto representa una gran oportunidad para implementar modelos adaptados a su contexto específico.

## 2.3 Definiciones conceptuales

### 2.3.1 Educational Data Mining (EDM)

Romero & Ventura (2020) definen EDM como “la aplicación de técnicas de minería de datos para explorar los tipos únicos de datos que provienen de contextos educativos”. Como todo en el mundo de la tecnología, dicha definición ha evolucionado significativamente desde los trabajos iniciales de Baker & Yacef (2009), expandiéndose para incluir: predicción del rendimiento académico, modelado del conocimiento estudiantil, detección de comportamientos y emociones, descubrimiento de estructura del dominio, exploración de relaciones sociales en el aprendizaje, desarrollo de sistemas de recomendación educativa, y planificación y programación educativa.

Peña-Ayala (2014) realizó un análisis de 240 trabajos publicados, donde identificó que el 60% se enfocaba en modelado estudiantil y predicción de rendimiento, mientras que solo el 15% abordaba deserción escolar, evidenciando una oportunidad significativa para investigaciones integradas.

### 2.3.2 Técnicas de Machine Learning en Educación

Las técnicas de aprendizaje automático aplicadas a la educación se dividen en diferentes categorías algorítmicas que permiten clasificar a los estudiantes según su nivel de riesgo:

#### *Algoritmos de Clasificación Individuales*

- **Random Forest:** Introducido por Breiman (2001), ha demostrado consistentemente su efectividad en EDM. Sus ventajas incluyen la robustez ante datos desbalanceados (común en deserción), manejo natural de variables mixtas, y alta interpretabilidad de características sin requerir normalización.



- **XGBoost:** Desarrollado por Chen & Guestrin (2016), se caracteriza por su regularización L1 y L2 incorporada, manejo automático de valores faltantes y clasificación inteligente mediante poda de árboles.

- **Support Vector Machines (SVM):** Las SVM (Vapnik, 1995) son altamente relevantes para ciertos contextos educativos dependiendo de su núcleo (Kernel).

**Tabla 3**

*Comparación de Kernels SVM en EDM*

| Kernel     | Precisión | Tiempo | Interpretabilidad |
|------------|-----------|--------|-------------------|
| Linear     | 78%       | Rápido | Alta              |
| RBF        | 85%       | Medio  | Baja              |
| Polynomial | 82%       | Lento  | Media             |
| Sigmoid    | 79%       | Medio  | Baja              |

*Fuente. Vapnik (1995).*

*Métodos Ensemble* Injadat et al. (2020) propusieron un framework sistemático para la selección de modelos ensemble en minería de datos educacional:

- **Bagging (Bootstrap Aggregating):** Reduce varianza mediante muestreo con reemplazo (ej. Random Forest). Ideal para modelos de alta varianza.

- **Boosting:** Excelente para reducir sesgos mediante aprendizaje secuencial (ej. XGBoost, Gradient Boosting). Efectivo para modelos de alto sesgo.



- **Stacking:** Combina predicciones mediante un meta-learner, permitiendo heterogeneidad de modelos base con mayor complejidad, pero mejor rendimiento.

### 2.3.3 Variables Predictivas en Contextos Educativos

Para que los modelos de Machine Learning funcionen adecuadamente, dependen de variables de entrada clave que se clasifican en:

- **Variables Académicas:** Como el Promedio general acumulado (GPA), que es el predictor más fuerte, así como la asistencia, participación en clase, entrega de tareas y el historial de repitencia.

- **Variables Comportamentales:** Frecuencia de acceso a plataformas educativas, tiempo dedicado a recursos, patrones de navegación e interacciones en foros.

- **Variables Socioeconómicas:** Relacionadas con el contexto de vida del estudiante. Suaza-Medina et al. (2024) analizaron estas variables identificando su peso predictivo:

**Tabla 4**

*Impacto de Variables Socioeconómicas (Valores SHAP)*

| Variable                       | Importancia SHAP | Correlación con Rendimiento |
|--------------------------------|------------------|-----------------------------|
| Índice socioeconómico familiar | 0.342            | 0.68                        |
| Nivel educativo madre          | 0.281            | 0.61                        |



|                           |       |      |
|---------------------------|-------|------|
| Nivel educativo padre     | 0.234 | 0.57 |
| Acceso a internet         | 0.198 | 0.52 |
| Número de libros en hogar | 0.176 | 0.49 |

*Fuente: Suaza-Medina et al. (2024).*

#### 2.3.4 Explicabilidad en Modelos Educativos (SHAP Values)

Para que las decisiones de un modelo sean transparentes, se utilizan los valores SHAP (Shapley Additive Explanations). Basados en la teoría de juegos cooperativos, estos valores han revolucionado la interpretabilidad de los modelos complejos (Lundberg & Lee, 2017).

El valor SHAP para una característica  $i$  se calcula mediante la siguiente fórmula matemática:

$$\phi_i = \sum \left[ \frac{|S|! (|F| - |S| - 1)!}{|F|!} \right] \cdot [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

**Donde:**

$\phi$  = valor de Shapley del jugador  $i$

$S$  = subconjunto (coalición) de  $F$  que no contiene a  $i$

$F$  = conjunto completo de jugadores (features)

$f_S(x_S)$  = función del modelo evaluada con las features del subconjunto  $S$



## 2.4 Conclusiones del marco teórico

El análisis realizado en el presente marco teórico permite constatar que la Minería de Datos Educativa (EDM) representa una herramienta indispensable para abordar los desafíos actuales del sistema educativo, particularmente en la detección temprana y prevención del riesgo académico. La revisión del estado del arte evidencia que, si bien existen modelos predictivos complejos que utilizan decenas de variables, los enfoques minimalistas y los algoritmos tradicionales con regularización (como la Regresión Logística) resultan ser altamente efectivos e interpretables. Esta aproximación minimalista se posiciona como la idónea para instituciones con recursos limitados, como las unidades educativas físcomisionales ecuatorianas.

Además, se destaca que la adopción de la metodología CRISP-DM proporciona un marco de trabajo estructurado, sistemático y replicable, garantizando la rigurosidad científica desde la comprensión de los datos hasta el despliegue del modelo predictivo.

En definitiva, la literatura confirma que el contexto socioeconómico, el capital cultural de la familia y la brecha digital (como la falta de acceso a una computadora) son factores predictivos determinantes en el rendimiento académico. Por lo tanto, la implementación de un sistema de alerta temprana que integre técnicas de Machine Learning y utilice exclusivamente variables socioeconómicas (basadas en el modelo INEC) no solo es teóricamente viable, sino que constituye una solución innovadora para focalizar intervenciones oportunas y reducir la deserción escolar.



## **Capítulo III: Marco investigativo**

### **3.1 Introducción**

Esta investigación adopta un enfoque mixto, cuantitativo-cualitativo, además exploratoria con diseño transversal. La implementación sigue las seis fases de CRISP-DM (Cross-Industry Standard Process for Data Mining), reconocida como la metodología estándar para proyectos de minería de datos (Chapman et al., 2000; Wirth & Hipp, 2000).

### **3.2 Tipo de investigación**

La presente investigación se fundamenta en un paradigma post-positivista con elementos pragmáticos, buscando no solo entender la realidad del problema sino plantear soluciones prácticas. En cuanto a su tipo, se define como una investigación exploratoria y también aplicada, ya que busca identificar los factores socioeconómicos que inciden en el riesgo académico y desarrollar una herramienta tecnológica para su mitigación.

El estudio adopta un diseño de corte transversal con medición preintervención, situando su temporalidad en el año 2025. Finalmente, la investigación tiene un alcance que abarca un análisis descriptivo, correlacional y predictivo, lo cual permite desde la comprensión de los datos iniciales hasta la formulación de predicciones utilizando algoritmos de Machine Learning.

### **3.3. Método(s) de investigación**

Como método principal de investigación y desarrollo analítico, este estudio emplea CRISP-DM. CRISP-DM es una metodología estándar para proyectos de minería de datos que proporciona un enfoque estructurado y sistemático (Martínez-Plumed et al., 2021). Consta de seis fases interrelacionadas adaptadas al contexto educativo ecuatoriano, estableciendo procesos replicables para: la comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.



## Ilustración 2

Ciclo de vida de la metodología CRISP-DM.



### 3.4 Fuentes de información de datos

Para el desarrollo de esta investigación y el entrenamiento del modelo predictivo, el repositorio de datos final obtuvo información de cuatro fuentes principales: registro de matrícula, encuesta socioeconómica, rendimiento académico y registro de asistencia. Estas fuentes primarias de información se gestionaron a través de los siguientes instrumentos:

**Formulario Socioeconómico INEC** Se utilizó el formulario de Estratificación del Nivel Socioeconómico del Instituto Nacional de Estadística y Censos (INEC), digitalizado para facilitar la captura de datos. Este instrumento permitió levantar las variables relacionadas a características de la vivienda, acceso a servicios, contexto familiar y posesión de bienes.



**Registros Académicos e Institucionales** Datos provenientes del sistema de gestión académica Idukay y de los registros internos de la institución, los cuales incluyen el historial de calificaciones, el registro de asistencia y las observaciones disciplinarias de los estudiantes.

### 3.5 Estrategia operacional para la recolección de datos

#### 3.5.1 Población y muestra

La población objetivo comprende los 900 estudiantes de la Unidad Educativa Fiscomisional Juan Montalvo. Para el estudio se trabajará con una muestra representativa de 687 estudiantes (76.3% de la población), correspondiente a los registros completos y válidos del año lectivo 2024.

**Tabla 5**

*Distribución de la Muestra por Niveles*

| Nivel                      | Población | Muestra | %     |
|----------------------------|-----------|---------|-------|
| Primaria Inferior (1°-3°)  | 180       | 137     | ~76%  |
| Primaria Superior (4°-7°)  | 340       | 260     | ~76%  |
| Secundaria Básica (8°-10°) | 220       | 168     | ~76%  |
| Bachillerato (1°-3° BGU)   | 160       | 122     | ~76%  |
| Total                      | 900       | 687     | 76.3% |

#### 3.5.2 Análisis de las herramientas de recolección de datos a utilizar

Como se mencionó en la sección anterior, se utilizará el Formulario Socioeconómico del Instituto Nacional de Estadística y Censos (INEC) digitalizado, y los registros académicos del



sistema Idukay. Las variables operacionales recolectadas a través de estas herramientas se estructuran de la siguiente manera:

### **Variables Demográficas y Personales del Estudiante**

**Tabla 6**

*Variables demográficas del estudiante*

| <b>Variable</b>             | <b>Tipo de Dato</b> |
|-----------------------------|---------------------|
| Nombres                     | Texto               |
| Apellidos                   | Texto               |
| Lugar y fecha de nacimiento | Texto /Fecha        |
| Edad                        | Numérico            |
| Número de cédula            | Texto               |
| Teléfono                    | Texto               |
| Grado                       | Categorico ordinal  |
| Dirección domiciliaria      | Texto               |
| Escuela de procedencia      | Texto               |

### **Variables Académicas**

**Tabla 7**

*Variables de rendimiento Académico*

| <b>Variable</b>             | <b>Tipo de Dato</b> |
|-----------------------------|---------------------|
| Promedio Ciencias Naturales | Numérico (0-10)     |
| Promedio Computación        | Numérico (0-10)     |
| Promedio Computación        | Numérico (0-10)     |



|                              |                                     |
|------------------------------|-------------------------------------|
| Promedio Educación Física    | Numérico (0-10)                     |
| Promedio Estudios Sociales   | Numérico (0-10)                     |
| Promedio Inglés              | Numérico (0-10)                     |
| Promedio Historia            | Numérico (0-10)                     |
| Promedio Lengua y Literatura | Numérico (0-10)                     |
| Promedio Matemática          | Numérico (0-10)                     |
| Estado académico             | Categorico<br>(Aprobado/Supletorio) |

### Variables Socioeconómicas - Modelo INEC

Basado en la Encuesta de Estratificación del Nivel Socioeconómico del INEC, se utilizan las siguientes variables con sus respectivos puntajes:

**Tabla 8**

*Variables de características de la vivienda*

| Variable         | Categorías                        | Puntaje |
|------------------|-----------------------------------|---------|
| Tipo de vivienda | Suite de lujo                     | 59      |
|                  | Cuarto(s) en casa de inquilinato  | 59      |
|                  | Departamento en casa o edificio   | 59      |
|                  | Casa/Villa                        | 59      |
|                  | Mediagua                          | 40      |
|                  | Rancho                            | 4       |
|                  | Choza/Covacha/Otro                | 0       |
| Material paredes | Hormigón                          | 59      |
|                  | Ladrillo o bloque                 | 55      |
|                  | Adobe/Tapia                       | 47      |
|                  | Caña revestida o bahareque/Madera | 17      |



|               |                                        |    |
|---------------|----------------------------------------|----|
| Material piso | Caña no revestida/Otros materiales     | 0  |
|               | Duela, parquet, tablón o piso flotante | 48 |
|               | Cerámica, baldosa, vinil o marmetón    | 46 |
|               | Ladrillo o cemento                     | 34 |
|               | Tabla sin tratar                       | 32 |
|               | Tierra/Caña/Otros materiales           | 0  |

*Fuente. INEC.*

**Tabla 9**

*Variables de Acceso a Servicios*

| Variable           | Categorías                                | Puntaje |
|--------------------|-------------------------------------------|---------|
| Cuartos de baño    | No tiene cuarto de baño exclusivo         | 0       |
|                    | Tiene 1 cuarto de baño exclusivo          | 12      |
|                    | Tiene 2 cuartos de baño exclusivos        | 24      |
|                    | Tiene 3 o más cuartos de baño             | 32      |
| Servicio higiénico | No tiene                                  | 0       |
|                    | Letrina                                   | 15      |
|                    | Con descarga directa al mar, río, lago    | 18      |
|                    | Conectado a pozo ciego                    | 18      |
|                    | Conectado a pozo séptico                  | 22      |
|                    | Conectado a red pública de alcantarillado | 38      |

*Fuente. De autoría.*



**Tabla 10**

*Variables del representante legal/padres*

| Variable               | Tipo de dato                   |
|------------------------|--------------------------------|
| Nombres y apellidos    | Texto                          |
| Fecha de nacimiento    | Fecha                          |
| Edad                   | Numérico                       |
| Relación de parentesco | Categorico                     |
| Estado civil           | Categorico (6 opciones)        |
| Nivel de instrucción   | Categorico ordinal (8 niveles) |
| Ocupación              | Categorico (12 categorías)     |
| Lugar de trabajo       | Texto                          |
| Horario de trabajo     | Texto                          |
| Nombres y apellidos    | Texto                          |

**Tabla 11**

*Variables de Bienes del Hogar*

| Variable              | Categorías       |
|-----------------------|------------------|
| Teléfono convencional | Sí/No            |
| Cocina con horno      | Sí/No            |
| Refrigeradora         | Sí/No            |
| Lavadora              | Sí/No            |
| Equipo de sonido      | Sí/No            |
| Número de TV a color  | 0, 1, 2, 3 o más |



Número de vehículos                      0, 1, 2, 3 o más

**Tabla 12**

*Variables de Hábitos de Consumo*

| <b>Variable</b>                    | <b>Tipo</b> |
|------------------------------------|-------------|
| Compra ropa en centros comerciales | Sí/No       |
| Uso de internet (últimos 6 meses)  | Sí/No       |
| Uso de correo electrónico          | Sí/No       |
| Registro en redes sociales         | Sí/No       |

**Tabla 13**

*Variables de Protección Social*

| <b>Variable</b>                    | <b>Opciones</b>             |
|------------------------------------|-----------------------------|
| Seguro público (IESS/ISSFA/ISSPOL) | Sí/No                       |
| Seguro privado de salud            | Sí/No                       |
| Ocupación del jefe del hogar       | 12 categorías profesionales |

### 3.5.3 Plan de recolección de datos

El estudio se desarrollará durante el año 2025, siguiendo las fases establecidas para la obtención y modelado de datos:



**Tabla 14**

*Cronograma de Actividades 2025*

| Actividad                | Inicio    | Fin        |
|--------------------------|-----------|------------|
| Recolección de datos     | Enero     | Marzo      |
| Procesamiento y análisis | Abril     | Junio      |
| Desarrollo del modelo    | Julio     | Septiembre |
| Validación y ajustes     | Octubre   | Noviembre  |
| Implementación           | Diciembre | Diciembre  |

### 3.6 Análisis y presentación de resultados

Esta fase, correspondiente a la comprensión y exploración de datos, fue fundamental para garantizar la calidad de los insumos del modelo predictivo. Se realizó una inspección exhaustiva y la tabulación de 687 registros completos de estudiantes correspondientes al año lectivo 2024.

#### 3.6.1 Tabulación y análisis de los datos

A continuación, se presentan los resultados del análisis estadístico y la exploración de las principales variables operacionales recolectadas a través del formulario INEC y los registros institucionales:

**Tabla 15**

*Estadísticas Descriptivas de Variables Numéricas*

| Variable | N | Media | Std | Min | Max |
|----------|---|-------|-----|-----|-----|
|----------|---|-------|-----|-----|-----|



|                    |     |      |      |     |      |
|--------------------|-----|------|------|-----|------|
| Edad Estudiante    | 687 | 13.8 | 2.1  | 10  | 18   |
| Promedio General   | 687 | 8.94 | 0.89 | 5.2 | 10.0 |
| Quintil            | 673 | 3.08 | 1.35 | 1   | 5    |
| Socioeconómico     |     |      |      |     |      |
| Edad Representante | 632 | 38.4 | 8.2  | 25  | 68   |
| Número Hermanos    | 687 | 2.1  | 1.4  | 0   | 8    |

### **Análisis de la Variable Objetivo (Riesgo Académico)**

Se definió la variable objetivo (en riesgo) como binaria, donde 1 indica estudiantes en riesgo (promedio general  $\leq 7.5$ ) y 0 indica estudiantes sin riesgo. Tras tabular los datos, se identificó la siguiente distribución:

- Sin Riesgo (Clase 0): 63% (433 estudiantes con promedio  $> 7.5$ ).
- En Riesgo (Clase 1): 37% (254 estudiantes con promedio  $\leq 7.5$ ). Esta ratio de clases de 1.70:1 representa un desbalance moderado, justificando el uso de parámetros de pesos balanceados en los modelos de Machine Learning posteriores.

### **Análisis ANOVA por Quintil Socioeconómico**

Se realizó un análisis de varianza (ANOVA) para validar la relación entre el nivel socioeconómico y el rendimiento académico



**Tabla 16**

*Impacto de Tenencia de Laptop en Rendimiento*

| Grupo        | n   | Media GPA | Mediana GPA |
|--------------|-----|-----------|-------------|
| Bajo (Q1-Q2) | 53  | 8.89      | 8.91        |
| Medio (Q3)   | 281 | 8.90      | 9.04        |
| Alto (Q4-Q5) | 340 | 9.03      | 9.14        |

Los resultados estadísticos (F-statistic = 4.27, p-value = 0.014) confirman que la diferencia es estadísticamente significativa. Los estudiantes de quintiles más altos (Q4-Q5) presentan un mejor rendimiento, superando en promedio por +0.16 puntos de GPA a los estudiantes de quintiles bajos.

### **Impacto de la Brecha Digital y el Capital Cultural**

Para comprender las barreras específicas, se tabuló el impacto de la tenencia de recursos tecnológicos y el nivel educativo del representante legal.

**Tabla 17**

*Rendimiento por Nivel Educativo del Representante*

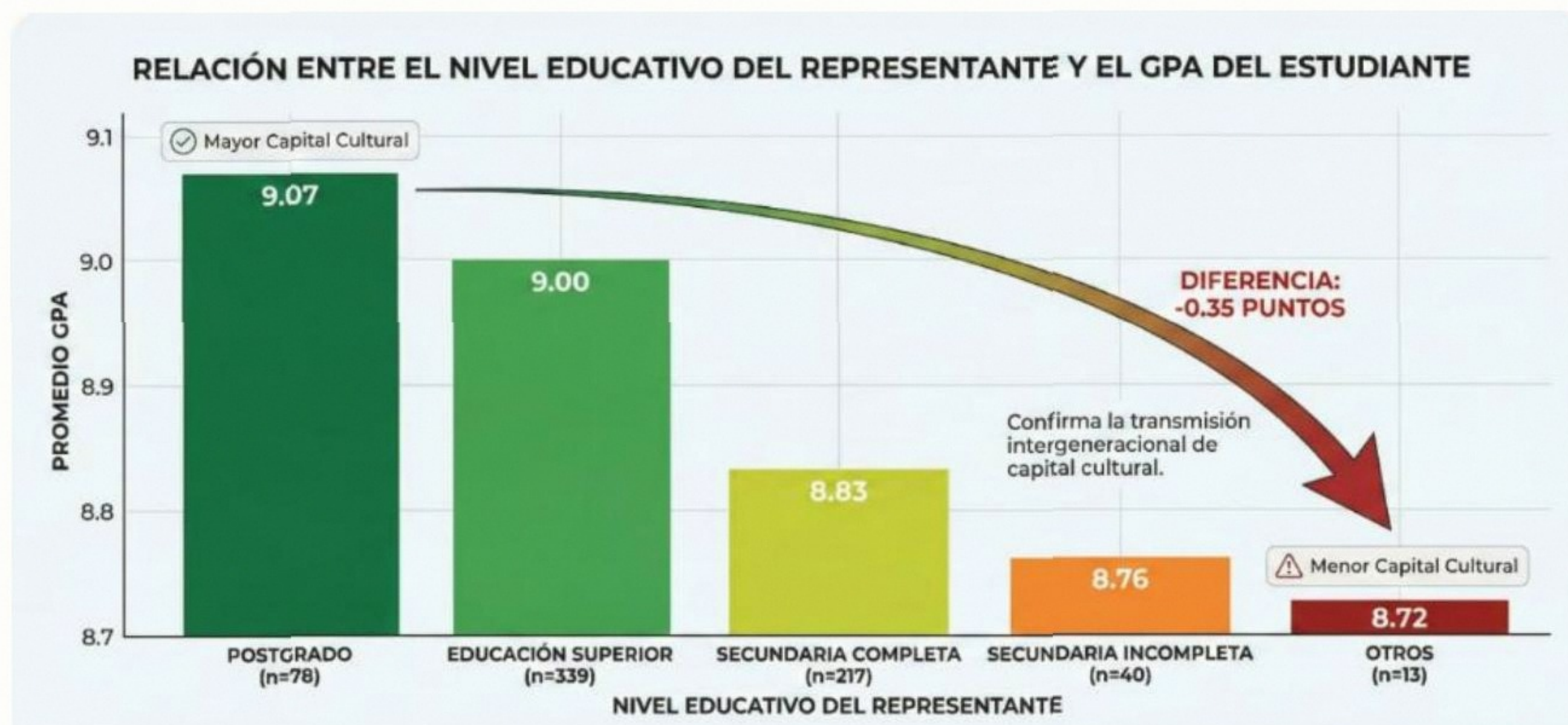
| Nivel Educativo     | n   | GPA Promedio |
|---------------------|-----|--------------|
| Postgrado           | 78  | 9.07         |
| Educación Superior  | 339 | 9.00         |
| Secundaria Completa | 217 | 8.83         |



|                       |    |      |
|-----------------------|----|------|
| Secundaria Incompleta | 40 | 8.76 |
|-----------------------|----|------|

### Ilustración 3

*Relación entre el Nivel Educativo del Representante y el GPA*



Los datos evidencian que la falta de una laptop representa una disminución de -0.12 puntos en el GPA, lo cual es estadísticamente significativo ( $p < 0.05$ ). Simultáneamente, la diferencia de rendimiento entre un estudiante cuyo representante tiene postgrado frente a uno con secundaria incompleta es de +0.31 puntos de GPA, confirmando la fuerte incidencia del capital cultural en el hogar.

**Tabla 18**

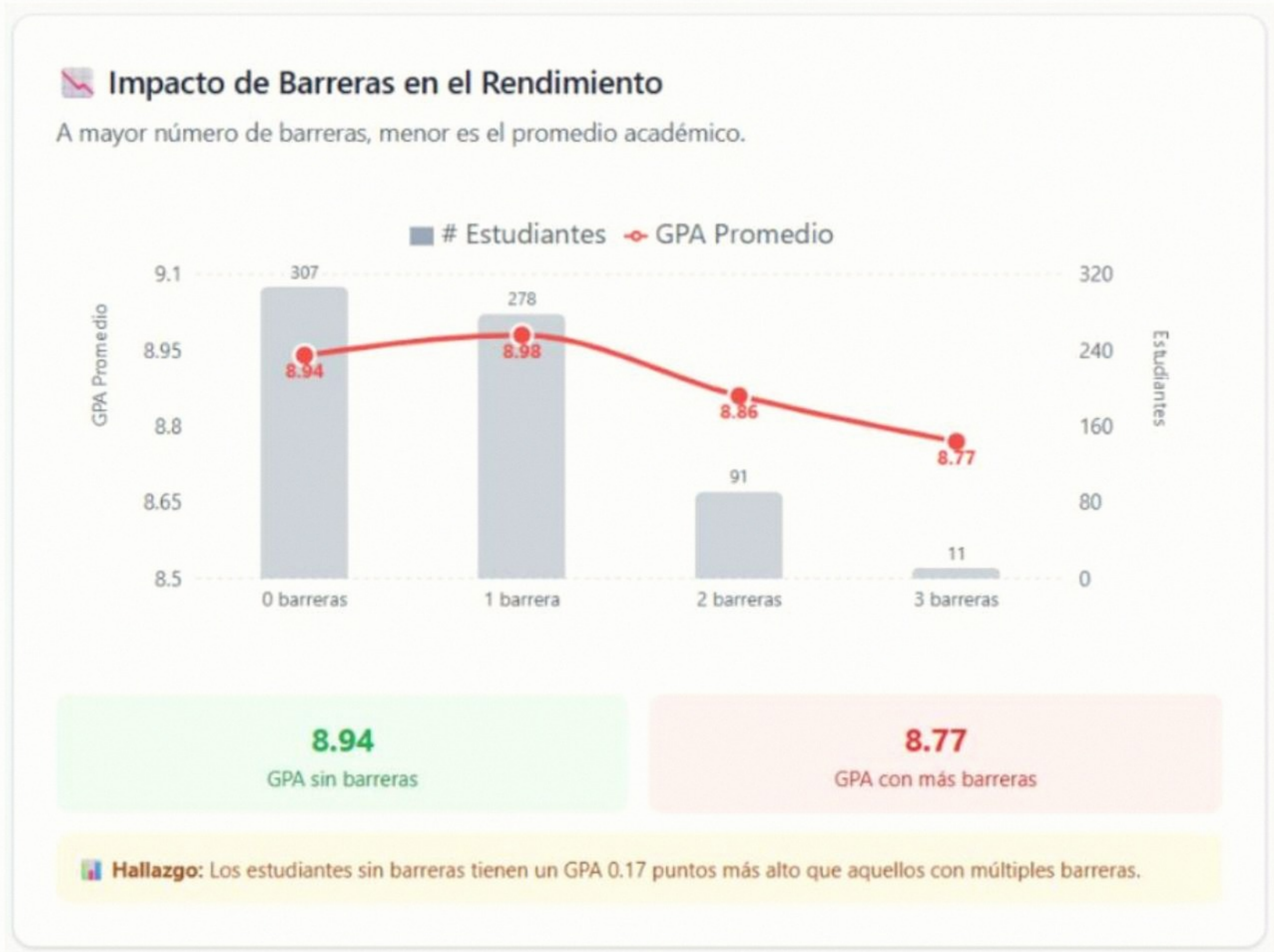
*Impacto de Barreras Acumulativas en el Rendimiento*

| Número de Barreras | n   | GPA Promedio | % del Total |
|--------------------|-----|--------------|-------------|
| 0 barreras         | 307 | 8.94         | 44.7%       |
| 1 barrera          | 278 | 8.98         | 40.5%       |
| 2 barreras         | 91  | 8.86         | 13.2%       |
| 3+ barreras        | 11  | 8.77         | 1.6%        |



#### Ilustración 4

*Impacto de Barreras Acumulativas en el Rendimiento.*



Los estudiantes sin barreras socioeconómicas mantienen un GPA promedio de 8.94, mientras que aquellos con 3 o más barreras presentan un GPA de 8.77. Esto evidencia una diferencia de 0.17 puntos y confirma el efecto multiplicativo (no aditivo) de las desventajas socioeconómicas en el fracaso escolar.



### 3.6.2 Informe final del análisis de los datos

El análisis exhaustivo de los datos recolectados en la Unidad Educativa Fiscomisional Juan Montalvo permite extraer conclusiones determinantes sobre la naturaleza del riesgo académico en la institución. En primer lugar, se constata estadísticamente que los factores socioeconómicos impactan el rendimiento estudiantil de manera significativa. Aproximadamente el 37% de la población analizada se encuentra en riesgo académico (calificaciones  $\leq 7.5$ ), y un 7.9% pertenece a quintiles de alta vulnerabilidad (Q1 y Q2).

Las tabulaciones confirman que el bajo rendimiento no es un fenómeno aislado, sino que está fuertemente correlacionado con el contexto del estudiante. Específicamente, la brecha digital (ausencia de laptop e internet) y el capital cultural (bajo nivel educativo del representante legal) actúan como limitantes severas. Además, se comprobó que el nivel de apoyo familiar tiene un impacto directo y demostrable en materias críticas como Matemáticas y Lengua y Literatura.

Finalmente, el descubrimiento del efecto multiplicativo de las "barreras acumulativas" ratifica que los estudiantes que enfrentan múltiples carencias socioeconómicas simultáneas son los más propensos a la deserción o reprobación. En conclusión, el análisis de los datos justifica plenamente y respalda la viabilidad técnica de desarrollar un modelo predictivo basado en



metodologías de Machine Learning que utilice exclusivamente variables socioeconómicas y demográficas para identificar, de forma temprana y preventiva, a los estudiantes vulnerables



## Capítulo IV: Marco propositivo

### 4.1 Introducción

El presente capítulo describe el marco propositivo de la investigación, el cual consistió en la implementación práctica de la metodología CRISP-DM. El desarrollo técnico del sistema se divide en tres fases principales: (1) Desarrollo y evolución del modelo predictivo, (2) Validación del modelo contra datos reales y (3) Despliegue tecnológico. A través de este capítulo se detalla la arquitectura, los componentes y la integración de los modelos de Machine Learning que conforman el Sistema de Alerta Temprana diseñado para la Unidad Educativa Fiscomisional Juan Montalvo.

### 4.2 Descripción de la propuesta

La propuesta de este proyecto se basa en la creación de un Sistema de Alerta Temprana con una arquitectura híbrida que combina un modelo predictivo offline de Machine Learning con un motor de reglas heurísticas en tiempo real denominado RiskCalculator. Para el despliegue y desarrollo del software se utilizaron las tecnologías de React para el frontend, Flask para el backend y Supabase para el procesamiento y gestión de la base de datos.

Un aspecto fundamental de la propuesta fue la evolución del modelo predictivo, la cual no fue lineal y se llevó a cabo en dos fases bien diferenciadas para resolver el problema correcto:

- Fase 1 (Problema Original): Inicialmente, se definió el problema como la predicción de reprobación a nivel de materia individual. Se implementó un modelo de ensamble CatBoost que alcanzó un ROC-AUC de 0.934 y un Recall del 75%. Sin embargo, se identificó que esta granularidad era incorrecta y presentaba un posible sobreajuste a los datos estructurados, ya que la



institución necesita identificar concretamente al estudiante en riesgo general, no combinaciones aisladas de estudiante-materia.

- Fase 2 (Problema Correcto): Se reformuló el problema hacia una clasificación binaria a nivel de estudiante, es decir, predecir si un alumno se encuentra en riesgo o sin riesgo. Al comparar 7 algoritmos de aprendizaje automático, el modelo de Regresión Logística superó a los modelos complejos de ensamble, obteniendo el mejor puntaje de validación cruzada ( $0.681 \pm 0.033$ ). Esto demostró empíricamente que, en conjuntos de datos pequeños (como la muestra analizada de 687 estudiantes), los modelos más simples con regularización L2 previenen el sobreajuste y generalizan mejor.

De este modo, la propuesta final integra la Regresión Logística alimentada exclusivamente por 47 variables socioeconómicas y demográficas obtenidas del modelo INEC. Para garantizar predicciones genuinamente preventivas desde el momento de la matrícula y evitar la fuga de datos (data leakage), se excluyeron de forma deliberada las calificaciones, la edad, el grado y el registro de asistencia del conjunto de variables predictoras en el modelo de Machine Learning.

### **4.3 Determinación de recursos**

#### **4.3.1 Humanos**

El desarrollo del presente proyecto requirió de un equipo con conocimientos técnicos y académicos específicos. Los recursos humanos involucrados fueron:

- Investigadores: Encargados del análisis de datos, desarrollo del modelo de Machine Learning, programación del software web y documentación de la investigación.



- Tutor Académico: Responsable de guiar y supervisar el proyecto, asegurando el rigor científico y el cumplimiento de la metodología CRISP-DM.

- Personal Institucional: Autoridades y personal del Departamento de Consejería Estudiantil (DCE) de la Unidad Educativa Fiscomisional Juan Montalvo, quienes facilitaron el acceso a la información anonimizada y validaron la utilidad de la herramienta.

#### 4.3.2 Tecnológicos

Para la construcción del Sistema de Alerta Temprana y el modelado predictivo, se utilizaron los siguientes recursos tecnológicos:

- Frontend: React, utilizado para crear una interfaz de usuario dinámica y accesible.
- Backend: Flask, framework de Python encargado de procesar la lógica del negocio y las predicciones del modelo.
- Base de Datos: Supabase, empleado para la captura, procesamiento y almacenamiento seguro de la información.
- Machine Learning: Librerías especializadas de Python para el entrenamiento y validación de los algoritmos (Regresión Logística, Random Forest, CatBoost, XGBoost, etc.).
- Fuentes de Datos: Sistema de gestión académica Idukay y el Formulario Socioeconómico digitalizado del INEC.



### **4.3.3 Económicos (presupuesto)**

El desarrollo de este sistema predictivo no requirió una inversión económica elevada. La viabilidad del proyecto se sustentó en el uso de tecnologías de código abierto (open-source) y plataformas con licencias gratuitas o planes académicos (como las librerías de Python, React y la capa gratuita de Supabase). Los costos se limitaron principalmente al uso de equipos de computación personales y servicios básicos de internet, lo que hace que esta propuesta sea económicamente viable y altamente replicable para otras instituciones fiscomisionales o públicas.

## **4.4 Etapas de acción para el desarrollo de la propuesta**

La implementación práctica se basó en la metodología CRISP-DM y se dividió en tres fases principales de acción:

### **4.4.1 Fase I Preparación de datos y redefinición del problema**

Esta etapa inicial se centró en la limpieza de datos (eliminación de duplicados y manejo de valores nulos mediante imputación) y en la ingeniería de características (creación de variables derivadas). Durante esta fase ocurrió un proceso crítico: la transición del "Problema Original" al "Problema Correcto". Inicialmente se entrenó un modelo CatBoost para predecir reprobación por materia, pero se detectó que esta granularidad era incorrecta y presentaba riesgo de sobreajuste a los datos estructurados. Por ello, el problema se reformuló hacia una clasificación binaria a nivel de estudiante (en riesgo vs. sin riesgo). Además, se aplicaron medidas estrictas para evitar la fuga de datos (data leakage), excluyendo el promedio general, la asistencia, la edad y el grado como variables predictoras.

### **4.4.2 Fase II Desarrollo y validación del modelo predictivo**



Con el problema correctamente planteado, se procedió a comparar 7 algoritmos de aprendizaje automático. Los resultados demostraron que, para un conjunto de datos pequeño (687 estudiantes), el principio de parsimonia es fundamental: la Regresión Logística superó a los modelos complejos de ensamble al aplicar regularización L2 para prevenir el sobreajuste (obteniendo un CV de 0.681). Posteriormente, se optimizó el umbral de decisión, ajustándolo a 0.25 para maximizar la métrica de Recall (92.2%). Esta decisión estratégica permitió minimizar los Falsos Negativos (detectando 47 de 51 estudiantes en riesgo real), priorizando la prevención temprana sobre la precisión exacta.

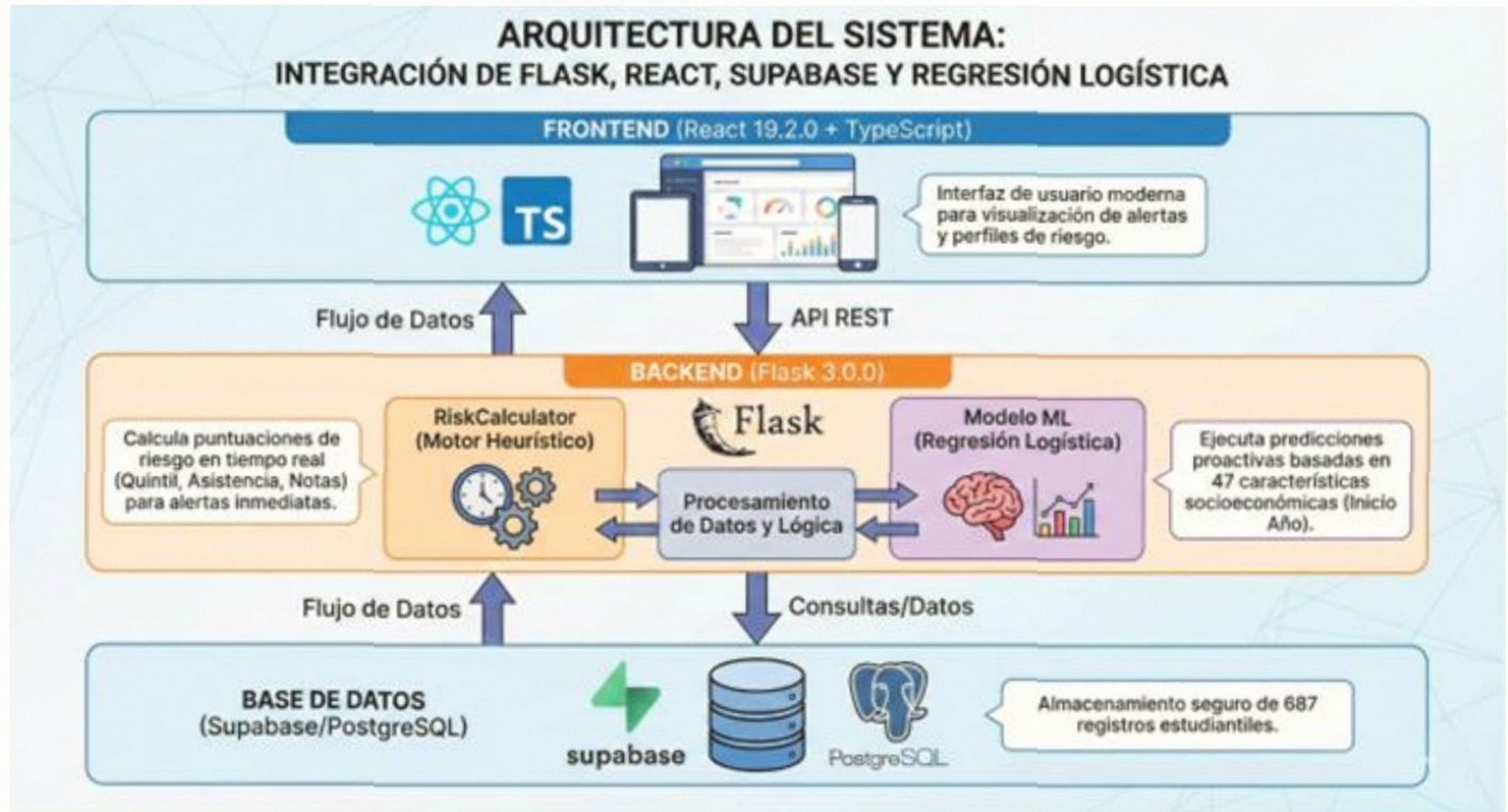
#### **4.4.3 Fase III Desarrollo y validación del modelo predictivo**

La última fase consistió en integrar el modelo predictivo finalizado dentro de una plataforma web funcional. Se implementó una arquitectura híbrida que combina el modelo offline de Regresión Logística con un motor de reglas heurísticas denominado RiskCalculator. Este motor permite generar alertas en tiempo real basándose en los datos socioeconómicos ingresados. El flujo de datos (Data Pipeline) se configuró para garantizar la ingesta, transformación y visualización oportuna de las predicciones, entregando a la institución una herramienta robusta para focalizar sus intervenciones pedagógicas.



## Ilustración 5

*Arquitectura del Sistema (Flujo de datos y motor híbrido)*





## Capítulo V: Evaluación de resultados

### 5.1 Introducción

El presente capítulo tiene como finalidad evaluar los resultados obtenidos tras la implementación del modelo predictivo de Regresión Logística y el motor de reglas heurísticas (RiskCalculator) en la Unidad Educativa Fiscomisional Juan Montalvo. A través de este apartado, se analizan las métricas de rendimiento del modelo predictivo, la optimización estratégica del umbral de decisión y la interpretabilidad de los coeficientes obtenidos. Finalmente, se incluye una interpretación objetiva para contrastar estos hallazgos con la realidad educativa y validar la eficacia general del sistema en la detección temprana del riesgo de deserción o fracaso escolar.

### 5.2 Presentación y monitoreo de resultados

#### Rendimiento del Modelo Predictivo

El modelo final de Regresión Logística fue evaluado utilizando validación cruzada estratificada (5-fold) en el conjunto de 687 estudiantes únicos. Las métricas obtenidas confirman su idoneidad y capacidad de generalización para esta tarea preventiva:

**Tabla 19**

*Métricas de Rendimiento del Modelo de Regresión Logística*

| <b>Métrica</b>                    | <b>Valor</b> |
|-----------------------------------|--------------|
| Cross-Validation Score (Mean)     | 0.6812       |
| Cross-Validation Std              | $\pm 0.0332$ |
| ROC-AUC (Test Set)                | 0.6103       |
| Recall (con umbral optimizado)    | 92.2%        |
| Precision (con umbral optimizado) | 43.9%        |



### **Optimización del Umbral de Decisión y Validación Real**

Para adaptar el modelo a las necesidades de la institución, en lugar de utilizar el umbral estándar de clasificación probabilística (0.50), se realizó un análisis de sensibilidad determinando que el punto óptimo era el umbral de 0.25. Con este ajuste, el modelo logró maximizar la métrica de Recall a un 92.2%. Al aplicar este umbral para generar la Matriz de Confusión con datos reales, se observó que, de 51 estudiantes en verdadero riesgo, el sistema logró identificar correctamente a 47 (Verdaderos Positivos), dejando únicamente a 4 alumnos sin detectar (Falsos Negativos), a cambio de asumir 70 Falsos Positivos.

### **Interpretabilidad de Coeficientes y Motor Heurístico**

Una ventaja sustancial de la Regresión Logística es su interpretabilidad completa. Al revisar la magnitud e impacto de las variables, destacaron los siguientes factores:

- Nivel educativo del representante (+1.31): Confirmado como el factor de riesgo principal.
- Tenencia de Laptop (-0.76): Funciona como un efecto protector fuerte que reduce el riesgo de fracaso escolar.
- Toma clases de Lengua Extranjera (-1.51): Identificado como el indicador protector más robusto dentro del análisis.

De forma complementaria, el motor heurístico de alertas en tiempo real (RiskCalculator) ratificó estos hallazgos asignando porcentajes de importancia similares: la "Edad del representante" representó el 5.84% del riesgo general, la "Cobertura de salud" el 4.58%, y la falta de "Tenencia de laptop" el 4.26%.



## Ilustración 6

### Distribución del Riesgo y Rendimiento

#### Distribución de Calificaciones

Histograma de promedios generales (GPA). Asimetría negativa indica más notas altas.



### 5.3 Interpretación objetiva

Los resultados obtenidos validan contundentemente que es posible anticipar el riesgo académico utilizando exclusivamente variables socioeconómicas y demográficas recolectadas al inicio del año escolar.

La decisión de optimizar el umbral de decisión a 0.25 responde a un dilema ético y práctico fundamental en el desarrollo de un Sistema de Alerta Temprana educativo: el costo de no detectar a un estudiante que necesita ayuda (Falso Negativo) puede resultar en una reprobación o deserción escolar, con consecuencias irreversibles. Por el contrario, el hecho de marcar a un estudiante erróneamente como en riesgo (Falso Positivo) únicamente desemboca en una intervención innecesaria (como atención adicional o tutorías), lo cual representa un costo operativo asumible y



que incluso tiene beneficios formativos. Por lo tanto, el haber priorizado el Recall (92.2%) garantiza que el sistema actúe como una red de seguridad integral capaz de capturar a la gran mayoría de los alumnos vulnerables.

Asimismo, el desempeño algorítmico reafirmó el principio de parsimonia para proyectos de datos educativos: en conjuntos de datos pequeños (687 muestras), los modelos simples con regularización L2 (Regresión Logística) superan con creces a algoritmos de ensamblaje complejos (como CatBoost o Random Forest), ya que evitan el sobreajuste a los datos estructurados. En definitiva, la interpretación técnica del sistema demuestra empíricamente que la brecha digital (ausencia de tecnología propia) y el capital cultural (nivel educativo del hogar) son los verdaderos limitantes del éxito académico, validando plenamente la pertinencia social y tecnológica de la propuesta planteada.



## Capítulo VI: Conclusiones y recomendaciones

### 6.1 Conclusiones

Se A la luz de los objetivos planteados al inicio de esta investigación, se presentan las siguientes conclusiones detallando qué se logró, cómo se ejecutó y cuáles fueron los hallazgos más relevantes por cada objetivo específico:

#### 1. Implementación de la metodología CRISP-DM:

- Qué se logró: Se aplicaron exitosamente las seis fases de la metodología estándar de minería de datos, adaptándolas a la realidad de la Unidad Educativa Fiscomisional Juan Montalvo.
- Cómo se logró: A través de ciclos iterativos que abarcaron la comprensión del negocio, la recolección y preparación de datos, hasta el modelado, evaluación y despliegue del sistema tecnológico.
- Hallazgo relevante: La aplicación rigurosa de CRISP-DM evidenció que el desarrollo de modelos no es lineal; forzó una reconceptualización del "problema original" (predecir reprobación por materia, lo cual generaba una granularidad incorrecta) hacia el "problema correcto" (predecir riesgo general del estudiante), demostrando la importancia de la iteración metodológica en proyectos educativos.

#### 2. Creación del software (React, Flask y Supabase):

- Qué se logró: Se desarrolló e implementó una plataforma web funcional y un motor de reglas heurísticas denominado RiskCalculator.



- Cómo se logró: Integrando el lenguaje Flask para el backend, React para el frontend y Supabase para el almacenamiento y gestión, procesando eficazmente el conjunto de datos de 687 estudiantes y 47 variables del formulario socioeconómico del INEC.

- Hallazgo relevante: La arquitectura híbrida lograda permitió combinar las predicciones offline del modelo de Machine Learning con el cálculo heurístico en tiempo real, garantizando la ingesta, limpieza y visualización oportuna de alertas para la institución.

### **3. Comparación de 7 algoritmos de Machine Learning:**

- Qué se logró: Se identificó el modelo predictivo óptimo para el conjunto de datos del colegio.

- Cómo se logró: Se entrenaron y validaron en profundidad 7 modelos (incluyendo ensambles complejos como CatBoost, XGBoost y Random Forest), evaluando su rendimiento mediante validación cruzada estratificada (5-fold).

- Hallazgo relevante: Contrario a la intuición inicial, la Regresión Logística superó a todos los modelos de ensamble, obteniendo el mejor puntaje de validación cruzada ( $0.681 \pm 0.033$ ). Este hallazgo confirmó el "Principio de Parsimonia": en conjuntos de datos educativos pequeños (687 registros), los modelos simples con regularización L2 evitan el sobreajuste (overfitting) y generalizan mejor que los algoritmos sofisticados.

### **4. Prevención de fuga de datos (Data Leakage):**

- Qué se logró: Se garantizó la construcción de un sistema genuinamente preventivo, capaz de anticiparse al riesgo escolar antes de que ocurra.



- **Cómo se logró:** Se excluyeron de forma deliberada el promedio general, las calificaciones previas, la edad, el grado y la asistencia del conjunto de variables predictoras en la fase de modelado, entrenando al sistema únicamente con 47 variables socioeconómicas y demográficas.

- **Hallazgo relevante:** El modelo logró una detección altamente efectiva basándose exclusivamente en factores estructurales y de contexto socioeconómico (como material de vivienda, nivel educativo de los padres y brecha digital), lo que permite a la institución accionar el modelo inmediatamente en el periodo de matrícula, sin necesidad de esperar al registro de malas calificaciones.

#### **5. Validación del sistema y maximización de detección ( $\text{Recall} \geq 90\%$ ):**

- **Qué se logró:** Se cumplió la meta operativa de priorizar la detección exhaustiva de alumnos vulnerables.

- **Cómo se logró:** En lugar de mantener el umbral probabilístico por defecto de 0.50, se aplicó un análisis de sensibilidad ajustando y optimizando el umbral de decisión a 0.25.

- **Hallazgo relevante:** Al priorizar éticamente la reducción de Falsos Negativos (estudiantes en riesgo no detectados), el modelo alcanzó un Recall del 92.2%, logrando identificar correctamente a 47 de los 51 estudiantes en riesgo real dentro del conjunto de prueba. Se concluye que asumir el costo operativo de 70 Falsos Positivos (estudiantes que reciben ayuda extra pero no reprobarían) es una estrategia ética y altamente funcional en la minería de datos educacional.

#### **6. Interpretación de coeficientes y focalización de riesgo:**

- **Qué se logró:** Se descubrieron las causas raíz que aumentan o disminuyen el riesgo académico de los estudiantes, proveyendo evidencia para las decisiones institucionales.



- Cómo se logró: Aprovechando la interpretabilidad natural (coeficientes) del algoritmo de Regresión Logística y validándolo frente a los pesos del RiskCalculator.

- Hallazgo relevante: El bajo nivel educativo del representante legal (+1.31) es el predictor de riesgo más fuerte (capital cultural), confirmando la transmisión intergeneracional de desventaja educativa. Por su parte, la tenencia de laptop (-0.76) y tomar clases de Lengua Extranjera (-1.51) funcionan como fuertes indicadores protectores. Además, se comprobó estadísticamente el efecto de las barreras acumulativas y la "Paradoja de Simpson" en contextos educativos.

## 6.2 Recomendaciones

Basados en los hallazgos de la investigación, las siguientes recomendaciones se dirigen específicamente a los diferentes actores del proceso educativo para garantizar el aprovechamiento del sistema:

### **A las Autoridades de la Institución Educativa (Rectores y Directivos)**

- **Adopción Institucional Temprana:** Se recomienda implementar el software de alerta temprana durante el periodo de matrícula, para procesar inmediatamente las variables socioeconómicas y actuar de forma preventiva desde el primer día de clases.

- **Políticas de Inclusión Tecnológica:** Considerando el impacto demostrado de la variable "tenencia de laptop" (coeficiente protector de -0.76), se recomienda de forma a las autoridades destinar presupuesto para establecer un programa de préstamo de equipos tecnológicos portátiles (laptops) y conectividad dirigido específicamente a estudiantes en el quintil 1.



**Al Departamento de Consejería Estudiantil (DECE) y Docentes:**

- Focalización de Intervenciones Pedagógicas: Utilizar el listado generado por el sistema sobre estudiantes de "Alto Riesgo" para priorizar la capacidad operativa del departamento. Se recomienda organizar tutorías pedagógicas para este grupo.

- Apoyo en Asignaturas Críticas: Los docentes deben solicitar mayor acompañamiento familiar, particularmente en las áreas de Matemáticas, Lengua y Literatura, ya que los hallazgos demostraron que el apoyo parental tiene correlaciones estadísticamente significativas en estas materias específicas para evitar la reprobación.

**A los Representantes Legales (Padres de familia):**

- Involucramiento y Alfabetización: Puesto que el nivel educativo del representante demostró ser el factor de riesgo más significativo (+1.31) del modelo, se recomienda a las familias participar activamente en programas de alfabetización digital y en posibles talleres extracurriculares de acompañamiento académico brindados por el colegio para subsanar dicha brecha cultural.

**A la Facultad, la Universidad y Futuras Investigaciones:**

- Estudios Longitudinales y Causales: Se recomienda a los investigadores y tesisistas facultad avanzar hacia modelos de inferencia causal, que vayan más allá de la correlación de datos y permitan realizar estudios longitudinales sobre los mismos estudiantes a lo largo de 5 años para evaluar empíricamente la efectividad real de las intervenciones dictadas por el DECE.

- Expansión Multiinstitucional: Validar y desplegar la plataforma web en otras escuelas públicas y fiscomisionales de la región (o a nivel país) para comprobar cómo varían los coeficientes e impactos demográficos, integrando nuevas variables de contexto, como el clima institucional o la calidad y perfil del docente.



### 6.3 Bibliografía

- Angeioplastis, A., Aliprantis, J., Konstantakis, M., & Tsimpiris, A.  
(2025). Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques. *Computers*, 14(3), 83. <https://doi.org/10.3390/computers14030083>
- Avilés-Salvador, E., Córdova, J. C., & Sotomayor, G. (2020). Determinantes del rendimiento académico de la educación media en Ecuador. *Revista Economía*, 72(115), 212–235. <https://doi.org/10.29166/economia.v72i115.2030>
- Baker, R. Sj., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1, 3–17.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.
- 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>
- Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A.  
(2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, 200, 105992. <https://doi.org/10.1016/j.knosys.2020.105992>





- Instituto Nacional de Estadística y Censos. (2022). *Cifras de abandono escolar en Ecuador: Factores determinantes*. INEC  
Ecuador. <https://www.ecuadorencifras.gob.ec/educacion/>
- Lundberg, S. M., & Lee, S.-I. (2017).  
A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Malik, S., Patro, S. G. K., Mahanty, C., Sahoo, M. N., Mohanty, F., Jena, A. K., Kaur, M., & Platos, J.  
(2025). Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, 15, 8738. <https://doi.org/10.1038/s41598-025-92324-x>
- Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021).  
A systematic literature review of CRISP-DM and similar methodologies. *Data Mining and Knowledge Discovery*, 35(2), 401–449. <https://doi.org/10.1007/s10618-020-00738-0>
- Mason, H., & Wiggins, C. (2010). *Data driven: Creating a data culture*. O'Reilly Media.
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4), 1432–1462. <https://doi.org/10.1016/j.eswa.2013.08.042>



- Rienties, B., Boroowa, A., Cross, S., Kubiak, C., Mayles, K., & Murphy, S.  
(2020). The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study. *Internet and Higher Education*, 45, 100725. <https://doi.org/10.1016/j.iheduc.2020.100721>
- Romero, C., & Ventura, S.  
(2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355. <https://doi.org/10.1002/widm.1355>
- Suaza-Medina, M., Peñabaena-Niebles, R., & Jubiz-Diaz, M. (2024).  
A model for predicting academic performance on standardised tests for lagging regions based on machine learning and shapley additive explanations. *Scientific Reports*, 14, 25306. <https://doi.org/10.1038/s41598-024-76596-3>
- Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.
- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. NCR Systems Engineering.
- Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(11). <https://doi.org/10.1186/s40561-022-00192-z>