

Minería de Datos y Analítica Visual para la Gestión Integral de Trayectorias Escolares en Ingeniería: Un Enfoque de Design Science Research

Data Mining and Visual Analytics for the Comprehensive Management of Academic Trajectories in Engineering: A Design Science Research Approach

Machado-Díaz, E¹; Serrano-Hernández, A² ✉; Román-Landeros, D.A.³

Datos de Adscripción:

¹ Machado Díaz, Eduardo
Tecnológico Nacional de México – Instituto Tecnológico Superior de Lerdo.
División de Ingeniería en Sistemas Automotrices. Ciudad Lerdo, Durango.
México
ROR ID: <https://ror.org/00davry38>
ROR ID: <https://ror.org/04bjtbt12>
eduardo.md@itslerdo.edu.mx
ORCID: <https://orcid.org/0000-0003-2610-7570>

² ✉ Serrano Hernández, Arturo
Tecnológico Nacional de México – Instituto Tecnológico Superior de Lerdo.
División de Ingeniería en Sistemas Automotrices. Ciudad Lerdo, Durango.
México
ROR ID: <https://ror.org/00davry38>
ROR ID: <https://ror.org/04bjtbt12>
arturo.sh@itslerdo.edu.mx
ORCID: <https://orcid.org/0009-0001-1302-5747>

³ Román Landeros, Diego Alberto
Tecnológico Nacional de México – Instituto Tecnológico Superior de Lerdo.
División de Ingeniería en Sistemas Automotrices. Ciudad Lerdo, Durango.
México
ROR ID: <https://ror.org/00davry38>
ROR ID: <https://ror.org/04bjtbt12>
diego.rl@itslerdo.edu.mx
ORCID: <https://orcid.org/0000-0002-0671-7577>

Resumen - La gestión de trayectorias escolares en los programas de ingeniería enfrenta el reto crítico de procesar volúmenes masivos de datos institucionales para mitigar las altas tasas de deserción temprana. Tradicionalmente, la dependencia de procesos manuales y la fragmentación de los registros en silos administrativos generan un rezago operativo que imposibilita la intervención preventiva. Para resolver esta brecha técnica, el presente artículo expone el diseño y desarrollo de un sistema de minería de datos visual, estructurado bajo la metodología Design Science Research (DSR) y el estándar CRISP-DM. Se construyó una arquitectura computacional en Python mediante un robusto motor de Extracción, Transformación y Carga (ETL) que implementa algoritmos de coincidencia difusa para normalizar catálogos irregulares y unificar la información en un repositorio relacional. La validación técnica comprobó la eficacia del sistema al procesar de manera longitudinal un censo poblacional de 961 expedientes académicos y estructurar 23,722 calificaciones transaccionales. En su aplicación práctica, el tablero interactivo permitió diagnosticar objetivamente una deserción del 72.1% en las cohortes fundadoras, correlacionando visualmente este fenómeno con la severidad en asignaturas filtro y la inestabilidad de la plantilla docente. Durante la evaluación empírica con

usuarios expertos, la interfaz obtuvo un puntaje de 87.5 en la Escala de Usabilidad del Sistema (SUS), confirmando una excelente aceptabilidad. Se concluye que el artefacto tecnológico reduce drásticamente la carga cognitiva, transformando historiales estáticos en un triaje operativo que fundamenta la toma de decisiones para impulsar la retención estudiantil.

Palabras Clave - Analítica visual, Design Science Research, Eficiencia terminal, Minería de datos educativos, Python, Trayectorias escolares.

Abstract - The management of educational trajectories in engineering programs faces the critical challenge of processing massive volumes of institutional data to mitigate high early dropout rates. Traditionally, the reliance on manual processes and the fragmentation of records in administrative silos generate an operational lag that precludes preventive intervention. To bridge this technical gap, this article presents the design and development of a visual data mining system, structured under the Design Science Research (DSR) methodology and the CRISP-DM standard. A Python-based computational architecture was built through a robust Extract, Transform, and Load (ETL) engine that implements fuzzy matching algorithms to normalize irregular catalogs and unify information in a relational repository. Technical validation proved the system's efficacy by longitudinally processing a population census of 961 academic records and structuring 23,722 transactional grades. In its practical application, the interactive dashboard objectively diagnosed a 72.1% dropout rate in founding cohorts, visually correlating this phenomenon with the severity in gatekeeper subjects and teaching staff instability. During the empirical evaluation with expert users, the interface obtained a score of 87.5 on the System Usability Scale (SUS), confirming excellent acceptability. It is concluded that the technological artifact drastically reduces cognitive load, transforming static academic histories into an operational triage that grounds decision-making to boost student retention.

Keywords - Design Science Research, Educational data mining, Python, Student retention, Terminal efficiency, Visual analytics.

I. INTRODUCCIÓN

1.1 Contexto de la gestión de trayectorias en ingeniería

La masificación de la educación superior contemporánea ha transformado a las universidades en entidades generadoras de volúmenes masivos de datos; sin embargo, la retención estudiantil en áreas de Ciencia, Tecnología, Ingeniería y Matemáticas (STEM) persiste como un desafío analítico central (Chen et al., 2021). Los sistemas de control escolar acumulan registros detallados sobre el desempeño académico, pero la integración y el análisis oportuno de esta información representan una limitante operativa constante para las instituciones.

En el sistema de educación superior tecnológica en México, los programas de ingeniería registran altas tasas de abandono durante los primeros semestres. En esta etapa de transición, la complejidad analítica de las asignaturas de ciencias básicas representa una barrera curricular significativa (García & Ruiz, 2023). Aunque la evidencia de este rezago queda documentada en los historiales académicos, la arquitectura de los datos institucionales frecuentemente opera bajo esquemas fragmentados y en silos tecnológicos aislados.

Esta fragmentación estructural genera deficiencias de procesamiento en la gestión administrativa de las academias y coordinaciones. En la práctica, la extracción de expedientes y el seguimiento de las trayectorias dependen predominantemente de procesos manuales y de la consolidación de hojas de cálculo no estandarizadas. El tiempo que el personal docente y administrativo invierte en la depuración de registros inconsistentes retrasa la capacidad de intervención pedagógica, priorizando la organización mecánica de la información sobre su análisis analítico.

El análisis del desempeño de una cohorte en ingeniería requiere un enfoque longitudinal que trascienda la evaluación transversal al final del ciclo escolar, permitiendo la identificación de patrones de riesgo latentes antes de que se materialice la deserción (López-Zambrano et al., 2021). La carencia de herramientas computacionales que automaticen este seguimiento condena a las instituciones a un modelo de gestión reactivo.

Para anticiparse a estos fenómenos críticos, resulta indispensable implementar arquitecturas tecnológicas eficientes. Reemplazar el procesamiento manual de datos en bruto por una infraestructura analítica permite convertir los repositorios históricos en recursos de información accionables. Esto facilita a las divisiones académicas el diseño de estrategias de acompañamiento precisas y fundamentadas en evidencia para impulsar la retención estudiantil.

1.2 El proceso ETL y el estándar CRISP-DM como fundamento analítico

Para transitar de una administración escolar retrospectiva a un entorno de analítica visual, la infraestructura computacional debe resolver la dispersión estructural de sus repositorios de información. En el ámbito universitario, los registros históricos residen en bases de datos transaccionales con estructuras heterogéneas que reflejan la operación aislada de cada departamento, dificultando su integración directa. La literatura

especializada en minería de datos educativos destaca que la consolidación de estos silos administrativos constituye el principal desafío técnico para cualquier despliegue analítico, exigiendo arquitecturas capaces de integrar variables multidimensionales y longitudinales (Sabando & Zambrano, 2024; Wang & He, 2025).

Para dotar de rigor científico a esta consolidación de datos, el modelado informático se fundamenta en las fases de preparación de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) (Martínez-Plumed et al., 2021). Bajo este estándar industrial, el proceso de Extracción, Transformación y Carga (ETL) opera como el núcleo arquitectónico. La extracción recopila los expedientes académicos desde los sistemas heredados de control escolar, los cuales frecuentemente generan reportes transaccionales en formatos no estructurados o de texto plano.

La fase de transformación constituye el esfuerzo metodológico central del pipeline computacional. Los historiales estudiantiles de cohortes pasadas presentan varianza operativa, manifestada en inconsistencias tipográficas en los catálogos docentes, claves de asignaturas discontinuadas y registros huérfanos. Mediante rutinas automatizadas programadas en Python, que incluyen algoritmos de coincidencia difusa (Fuzzy Matching) basados en la distancia matemática de Levenshtein, se estandarizan los catálogos y se depuran las anomalías estructurales. Este procesamiento riguroso permite reconstruir el historial del estudiante de manera longitudinal, transformando calificaciones aisladas en una línea de tiempo continua que documenta el avance real de la cohorte (Romero & Ventura, 2020).

La etapa de carga consolida los registros estandarizados en un repositorio relacional normalizado. Automatizar este ciclo elimina la latencia y el error humano inherentes al cruce manual de hojas de cálculo (Cardona et al., 2023). En el contexto operativo de instituciones de alta matrícula como el Tecnológico Nacional de México, esta arquitectura técnica resulta indispensable. Sin un flujo de datos limpio, parametrizado y auditable, los componentes de exploración interactiva y analítica visual carecerían de la validez estructural necesaria para fundamentar el diseño de estrategias institucionales de retención (De Laet et al., 2020).

1.3 Analítica Visual como herramienta cognitiva

El procesamiento masivo de datos mediante arquitecturas ETL resuelve la centralización de la información, pero plantea un desafío secundario: la capacidad de los usuarios para interpretarla sin experimentar fatiga o saturación informativa. En este contexto, la analítica visual se distancia de los reportes estadísticos estáticos al combinar la potencia del cómputo automatizado con el razonamiento humano estructurado. El objetivo radica en implementar Sistemas de Soporte a la Decisión (DSS) mediante tableros interactivos (dashboards), los cuales facilitan el tránsito desde la exploración de datos en bruto hacia la toma de decisiones asertivas en el entorno académico (Haro, 2018).

Los expedientes históricos de los estudiantes de ingeniería son multidimensionales, conteniendo variables temporales y de rendimiento que cambian de manera constante. Desplegar esta información a través de estimaciones de densidad estadística y gráficos jerárquicos aprovecha los atributos preatentivos del

sistema visual humano (Verbert et al., 2013). Esta representación permite a las coordinaciones académicas detectar anomalías, tendencias de reprobación y estructuras de rezago latentes que pasarían desapercibidas en formatos tabulares convencionales. Al mapear visualmente los flujos curriculares, la interfaz gráfica comunica de manera objetiva los puntos exactos donde se fractura el avance de una cohorte.

La simbiosis entre la arquitectura informática subyacente y la visualización interactiva reduce significativamente la carga cognitiva del personal educativo (De Laet et al., 2020). Al delegar la normalización de datos y el cálculo matemático a la infraestructura backend en Python, y la síntesis de resultados a los componentes visuales, la atención del usuario se desplaza desde la manipulación operativa hacia la interpretación estratégica. De este modo, la analítica visual dota a las instituciones de herramientas interactivas que amplifican la capacidad administrativa para diseñar planes de acompañamiento verdaderamente proactivos y fundamentados en evidencia cuantitativa.

1.4 Planteamiento del problema y contribución

En el Instituto Tecnológico Superior de Lerdo (ITSL), específicamente en la División de Ingeniería en Sistemas Automotrices, los indicadores históricos revelan un desafío crítico en la retención estudiantil. El análisis retrospectivo demuestra que las cohortes fundadoras (ej., generación 2016) registraron tasas de abandono definitivas que alcanzaron hasta el 72.1% de la matrícula inicial. Esta vulnerabilidad estructural exige una evolución inmediata desde un modelo de gestión reactivo hacia un paradigma preventivo fundamentado en evidencia cuantitativa.

La principal barrera tecnológica para consolidar esta transición es la carencia de una infraestructura de datos automatizada. La información académica de las trayectorias reside de manera fragmentada en sistemas legados, almacenada frecuentemente en formatos de texto plano y hojas de cálculo desnormalizadas. La consolidación manual de estos registros genera una latencia operativa severa, impidiendo a las coordinaciones de carrera y academias identificar oportunamente patrones de riesgo temprano, tales como el desgaste acumulativo provocado por la reprobación en asignaturas críticas.

Para resolver esta brecha técnica, la presente investigación expone el diseño y desarrollo de un sistema de minería de datos visual estructurado bajo la metodología Design Science Research (DSR). La contribución central del estudio radica en la construcción de una arquitectura computacional fundamentada en un proceso ETL (Extracción, Transformación y Carga) automatizado en Python, con capacidad comprobada para procesar de manera longitudinal un censo de 961 expedientes y normalizar 23,722 calificaciones transaccionales en un repositorio relacional unificado. El acoplamiento de esta infraestructura de datos con un tablero de control visual (Dashboard) elimina la deficiencia en el procesamiento de la información, proporcionando a la institución un artefacto tecnológico objetivo y accionable para diagnosticar el rezago académico e impulsar la eficiencia terminal.

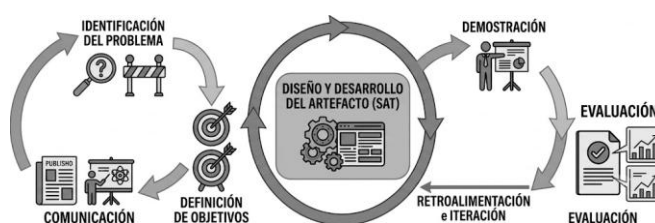
II. PARTE TÉCNICA DEL ARTÍCULO

2.1 Instanciación del marco metodológico Design Science Research (DSR) y CRISP-DM

El rigor científico para la construcción de este sistema de minería visual se fundamenta en una arquitectura metodológica dual que busca acercar la tecnología a las necesidades reales de los docentes. En un primer nivel, la investigación adopta el paradigma de *Design Science Research* (DSR). Este modelo se enfoca en la creación y evaluación empírica de artefactos tecnológicos diseñados para resolver problemas organizacionales complejos. En el entorno de la educación superior, aplicar DSR asegura que el desarrollo del software no sea un simple ejercicio de programación, sino una respuesta genuina a los retos que enfrentan las academias de ingeniería. Para gobernar el procesamiento analítico de los historiales académicos y dar un orden estandarizado a la ingeniería de datos, el núcleo de la herramienta se estructuró apoyándose en la metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM). Investigaciones recientes en analítica educativa confirman que aplicar CRISP DM garantiza que la transición de registros aislados a plataformas de diagnóstico se realice de forma íntegra y libre de sesgos metodológicos (Carneiro et al., 2022). De esta forma, mientras el enfoque DSR garantiza que el sistema completo sea útil para la institución, las fases de preparación de CRISP DM orientan paso a paso la construcción del proceso de Extracción, Transformación y Carga (ETL). Para comprobar que esta propuesta es verdaderamente viable, el diseño de la investigación estableció métricas de evaluación.

La eficacia técnica del artefacto se evalúa observando el rendimiento computacional del proceso ETL, teniendo como meta principal lograr la automatización total en la limpieza de los expedientes. El sistema asume el reto de procesar y normalizar miles de calificaciones, reduciendo el tiempo de espera administrativo, el cual tradicionalmente consume semanas enteras de trabajo manual, a un proceso que se completa en cuestión de segundos. Alcanzar esta agilidad es fundamental para que el tablero visual interactivo pueda responder de inmediato y facilitar la toma de decisiones para frenar la deserción a tiempo mediante evidencia paramétrica (Bustamante et al., 2024). El proceso puede visualizarse en la Figura 1.

Figura 1
Ciclo metodológico de Design Science Research aplicado al desarrollo del artefacto. Fuente: Elaboración Propia



2.2 Especificación y alcance del artefacto computacional

La resolución de la fragmentación operativa descrita exige la construcción de un artefacto tecnológico bidimensional que actúe directamente sobre el flujo de información institucional. La propuesta se compone de un motor de procesamiento backend desarrollado en Python y una interfaz web interactiva orientada a la toma de decisiones. Este diseño permite que el flujo de información transite de manera fluida entre el almacenamiento en bruto y la visualización estratégica, eliminando por completo la necesidad de intervenciones manuales o cruces tradicionales en hojas de cálculo por parte de las academias.

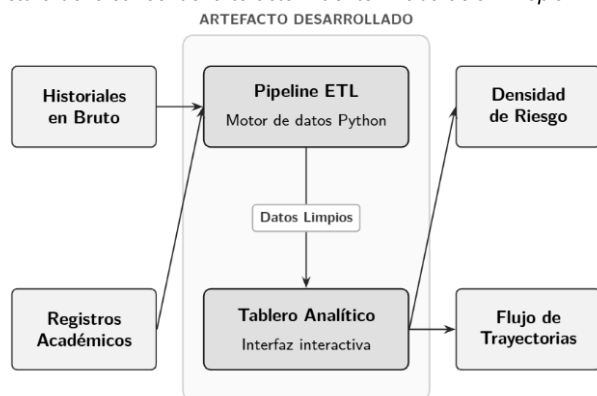
El núcleo operativo del artefacto delimita sus funciones con precisión algorítmica para garantizar la viabilidad del sistema dentro de las restricciones de la organización. El backend asume la responsabilidad exclusiva de ejecutar las tareas de extracción, transformación y carga (ETL), unificando los expedientes históricos de las generaciones fundadoras en una estructura relacional única y longitudinal. Por su parte, el componente de analítica visual traduce estos conjuntos de datos depurados en componentes gráficos interactivos, empleando estimaciones de densidad estadística y diagramas jerárquicos.

La herramienta dota al usuario de capacidades de exploración multivariable, permitiéndole segmentar el rendimiento académico por cohortes, ciclos escolares y líneas curriculares específicas mediante una navegación intuitiva que mitiga la carga cognitiva del docente.

El sistema desarrollado en esta etapa se concentra exclusivamente en las dimensiones descriptivas y diagnóstica del análisis de trayectorias escolares. De este modo, el artefacto se establece como el cimiento tecnológico indispensable; una infraestructura analítica sólida que transforma los registros históricos pasivos en conocimiento táctico inmediato antes de abordar la complejidad de la analítica prescriptiva cuya arquitectura se ve en la Figura 2.

Figura 2

Estructura del alcáncel del artefacto. Fuente: Elaboración Propia



2.3 Arquitectura del sistema y desarrollo de los procesos ETL

El núcleo técnico de la solución propuesta se centra en el diseño e implementación de una arquitectura computacional orientada a automatizar la ingesta y el refinamiento de la información escolar. La problemática inicial en los repositorios institucionales radicaba en su naturaleza estrictamente

transaccional, dispersa y no estructurada. Los registros históricos de las asignaturas procedían principalmente de archivos de texto plano (Bloc de notas) y hojas de cálculo irregulares fragmentadas entre silos administrativos, como se ve en la Figura 3.

Figura 3

Ejemplo de información en bloc de notas. Fuente: Elaboración Propia

Num.Control : 16231010		Nombre : ACOSTA SANCHEZ ANGEL ARMANDO		
Semestre : 1		Carrera : INGENIERIA EN SISTEMAS AUTOMOTRICES		
ACF0901	CALCULO DIFERENCIAL	5	84	1erCur-1er0port00
ACA0907	TALLER DE ETICA	4	81	1erCur-1er0port00
SAB1317	FUNDAMENTOS DE DIBUJO	5	10	1erCur-2da0port00
Promedio : 59.50				

A pesar de su complejidad sintáctica, la extracción y procesamiento de estos archivos de texto resultaba crítica y obligatoria para el éxito del modelo, ya que representaban la única fuente disponible donde residían variables analíticas indispensables para la reconstrucción longitudinal de las trayectorias: la oportunidad o tipo de evaluación en que el alumno cursó y acreditó la materia (v.g., ordinario, extraordinario o recursamiento), el semestre cronológico exacto y la identidad del docente que impartió la asignatura, entre otros metadatos operativos esenciales.

Estos reportes crudos carecían de delimitadores computacionales estándar como comas o tabulaciones uniformes, presentando estructuras complejas orientadas originalmente a la impresión física o la interpretación visual humana, lo que volvía inoperante su análisis estadístico directo. Ejemplo de otro archivo es el visto en la Figura 4.

Figura 4

Información de acuerdo con la materia. Fuente: Elaboración Propia

CARRERA : 10		ING SIST AUTOM	ISAU-2013-240		PERIODO : AGO-DIC-2016	
MATERIA : SAB1317 FUNDAMENTOS DE DIBUJO						
MAESTRO : 297 -		JORGE LUIS HERNANDEZ MAÑON			GRUPO : 5B1A	
HORARIO : Lu 12:00-13:00-LC4 Ma 12:00-13:00-LC4 Mi 12:00-13:00-LC4						
NO.	CONTROL	NOMBRE DEL ALUMNO			10P	20P
1	16231010	ACOSTA SANCHEZ ANGEL ARMANDO				NA
2	16231499	AGUILAR PINEDA JOAQUIN			78	

Para superar este escenario de severa heterogeneidad estructural, se seleccionó Python como el motor analítico principal del backend asíncrono, programando rutinas capaces de procesar la varianza de los formatos de salida originales y centralizar la información.

La fase de transformación constituyó el desafío metodológico y computacional más complejo dentro del *pipeline* de datos. Los métodos aplicados consistieron en el desarrollo de scripts robustos apoyados en las librerías Pandas y NumPy para limpiar el ruido operativo acumulado durante los ciclos escolares de las generaciones fundadoras.

El proceso requirió realizar un *parsing* secuencial explicado en la Figura 6 basado en espaciados dinámicos para extraer las variables independientes aisladas en el encabezado de los bloques de texto plano y mapearlas de forma longitudinal bajo una llave primaria única (el número de control del alumno).

2.4 Arquitectura del Pipeline ETL y Coincidencia Difusa

El núcleo técnico de la solución se centra en una arquitectura computacional programada en Python, diseñada específicamente para automatizar la ingesta y el refinamiento masivo de la información escolar. Para este estudio, la infraestructura procesó de manera asíncrona un censo de 961 expedientes históricos y 23,722 calificaciones transaccionales correspondientes a las generaciones fundadoras. La problemática inicial radicaba en la naturaleza dispersa y no estructurada de los repositorios institucionales, donde los historiales provenían de archivos de texto plano y hojas de cálculo irregulares, fragmentadas entre distintos departamentos administrativos.

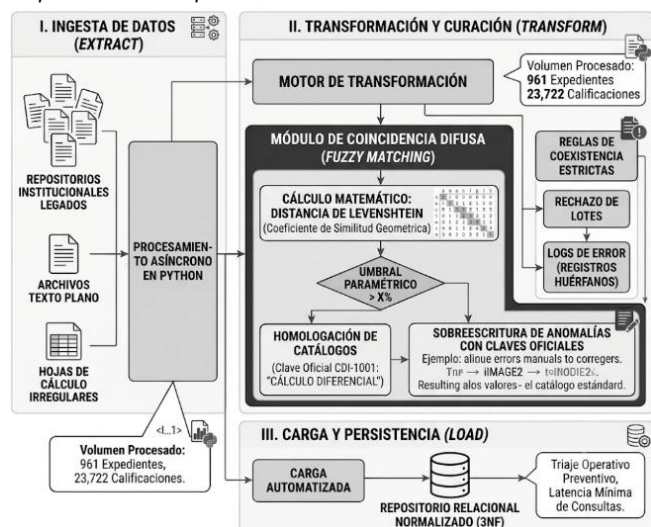
A pesar de su complejidad sintáctica, el procesamiento automatizado de estos documentos resultaba obligatorio, ya que contenían las variables analíticas indispensables para reconstruir longitudinalmente las trayectorias: oportunidades de acreditación, semestres cronológicos exactos y la identidad del docente evaluador. El diseño del artefacto resolvió esta heterogeneidad acotando su alcance exclusivamente a la limpieza, transformación y persistencia de los datos, estableciendo una base sólida para la analítica posterior.

En los sistemas legados universitarios, es común encontrar una alta varianza operativa en los nombres de las asignaturas y docentes debido a errores de captura manual acumulados a lo largo de los años. Para resolver esto sin recurrir a una depuración manual exhaustiva, se implementó un algoritmo de coincidencia difusa (Fuzzy Matching) fundamentado en el cálculo matemático de la distancia de Levenshtein (Cardona et al., 2023). El algoritmo calcula un coeficiente de similitud y, al superar un umbral paramétrico de confianza, sobrescribe la anomalía con la clave institucional oficial unificada.

El motor fue programado para aplicar reglas de coexistencia estrictas, rechazando lotes de información si los expedientes carecían de variables conjuntas obligatorias, garantizando así la integridad analítica. La Figura 5 muestra la arquitectura utilizada.

Figura 5

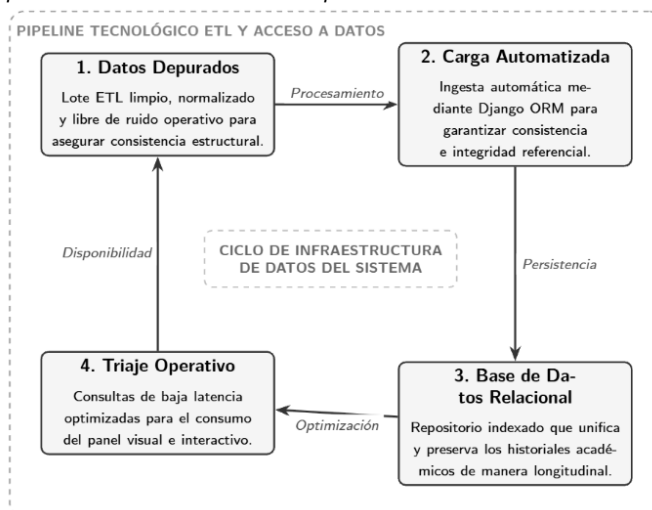
Arquitectura del esquema de coincidencias. Fuente: Elaboración Propia



La etapa de carga consolida de manera automática los registros completamente depurados y normalizados dentro de un repositorio relacional integrado en el entorno web de Django. Al almacenar la información bajo una estructura indexada y libre de ruido operativo, el sistema garantiza la integridad referencial de los historiales académicos a lo largo de la línea del tiempo. Este diseño técnico optimiza el rendimiento del backend para soportar consultas complejas de baja latencia directamente desde la interfaz gráfica, proveyendo un cimiento de datos estable, auditable y de alta disponibilidad para el triaje académico operativo. De este modo, el *pipeline* ETL se establece como la infraestructura analítica fundamental que transforma el rezago de los archivos en bruto en un activo dinámico, listo para ser consumido por los componentes de exploración visual interactiva. Este proceso se ilustra en la Figura 6.

Figura 6

Arquitectura del flujo de persistencia, integridad y consumo para el triaje operativo. Fuente: Elaboración Propia



Este desacoplamiento técnico garantiza consultas inmediatas en la interfaz gráfica del panel de control, proveyendo un cimiento limpio, estable y auditable para el triaje académico operativo.

2.5 Diseño del modelo de datos relacional y esquema de entidades

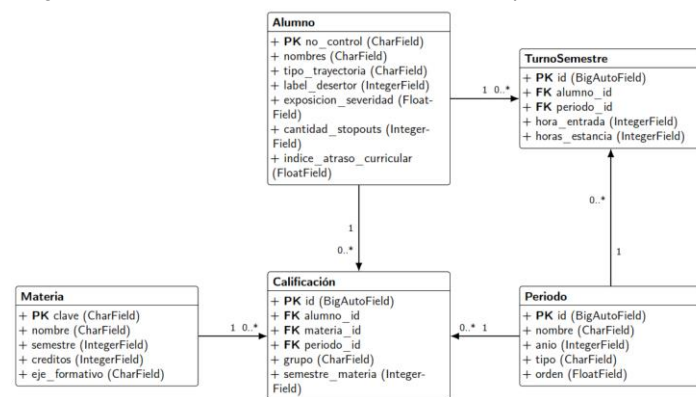
Una vez concluido el proceso de extracción y unificación en el pipeline de Python, el diseño arquitectónico exige estructurar la persistencia de la información de manera eficiente. Los archivos transaccionales de origen presentan una naturaleza redundante y desnormalizada, replicando los datos del alumno de forma indeterminada en cada ciclo escolar. Para resolver esta deficiencia y garantizar la integridad referencial del sistema, se construyó un esquema de base de datos relacional alineado con los principios de la Tercera Forma Normal (3NF). Este nivel de normalización erradica las dependencias transitivas y aísla los componentes del entorno escolar en entidades independientes, lo cual optimiza el rendimiento de las consultas y reduce drásticamente el espacio de almacenamiento físico. Diversas investigaciones en arquitectura

de información establecen que transitar de silos administrativos a almacenes normalizados resulta indispensable para sostener aplicaciones analíticas sin comprometer la consistencia de las transacciones (Elmasri & Navathe, 2010).

La orquestación de este esquema lógico se implementó mediante el Mapeo Objeto Relacional (ORM) del entorno web de Django. Esta tecnología abstrae la complejidad de las consultas y consolida el catálogo de asignaturas, la plantilla docente y el perfil demográfico del estudiante en entidades interconectadas. El núcleo analítico del sistema converge en una entidad asociativa que resguarda el contexto longitudinal de la trayectoria, registrando variables críticas como la calificación numérica final, el periodo cronológico y la oportunidad de acreditación. El diagrama UML de la base de datos se muestra en la Figura 7.

Figura 7

Diagrama UML del Sistema. Fuente: Elaboración Propia



La estructura definió tres entidades fundamentales interconectadas: *Alumno*, *Asignatura* y *Docente*. La entidad *Alumno* se estableció utilizando el número de control institucional (boleta) como llave primaria única, conteniendo los atributos demográficos y el estatus general del estudiante. Por su parte, la entidad *Asignatura* centralizó el catálogo oficial de materias del plan de estudios de ingeniería empleando la clave oficial de la retícula curricular como identificador, lo que resolvió de manera definitiva el problema de las variaciones nominales detectadas en las fases previas de la extracción. La relación de muchos a muchos (*Many-to-Many*) existente entre estudiantes y asignaturas se resolvió mediante la construcción de la entidad asociativa denominada *HistorialAcadémico*.

La entidad asociativa *HistorialAcadémico* constituye el núcleo analítico de la base de datos, ya que es el punto exacto donde convergen los flujos lógicos de la información y donde se aloja el contexto de la trayectoria. Cada registro en esta tabla representa un evento de inscripción único y se vincula mediante llaves foráneas (*Foreign Keys*) a un alumno, una asignatura y al docente específico que impartió la clase.

Para asegurar la velocidad de respuesta del tablero analítico web y mitigar las latencias durante la renderización de las estimaciones de densidad y diagramas de flujo, se implementaron índices compuestos en la base de datos sobre los campos de control combinados de alumno, asignatura y periodo, permitiendo que la interfaz interactiva recupere décadas de historiales de las generaciones fundadoras en milisegundos.

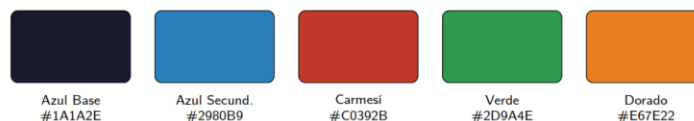
2.6 Componentes de analítica visual y diseño de la interfaz

El diseño de la interfaz de usuario (UI) y el desarrollo de la experiencia de usuario (UX) se centraron rigurosamente en el principio de eficiencia cognitiva para la gestión escolar y el seguimiento de trayectorias. Para mitigar la fatiga informativa y evitar la saturación de tablas de datos crudos, el frontend del artefacto se construyó sobre el framework de diseño Bootstrap en su versión 5.3.2, aprovechando su sistema de rejilla flexible para garantizar una respuesta adaptativa óptima en múltiples dispositivos.

El sistema tipográfico ejerce una función de control estructurado: la familia *Inter* se seleccionó para los cuerpos de texto y narrativas por su elevada legibilidad en pantallas de alta densidad de píxeles, mientras que, para los identificadores numéricos, claves de control estudiantil y registros de calificaciones se implementó la fuente monoespaciada *JetBrains Mono*, reduciendo significativamente el error humano durante la consulta rápida de datos transaccionales. Para consolidar la apropiación tecnológica, se integró una hoja de estilos personalizada (sat.css) que alinea la paleta cromática con los valores institucionales del Instituto Tecnológico Superior de Lerdo (ITSL) y del Tecnológico Nacional de México (TecNM). Como se detalla formalmente en la paleta de diseño de la Figura 8, el entorno visual distribuye sus propiedades cromáticas bajo una jerarquía estricta: un azul institucional base (#1A1A2E) estructurado en la barra de navegación lateral y los encabezados primarios para transmitir estabilidad; un azul secundario e interactivo (#2980B9) implementado en botones de acción primaria (*btn-sat-primary*) e indicadores de progreso; tonalidades carmesí y rojo de alerta (#C0392B y #8E232E) reservadas para umbrales de riesgo crítico; un verde de aprobación (#2D9A4E) asociado a trayectorias estables; y un dorado o naranja preventivo (#E67E22) empleado para variables de riesgo medio o rezago latente. Esta identidad visual da soporte a una arquitectura de *frontend* basada en un modelo de herencia de plantillas de Django, reduciendo la redundancia de código HTML.

Figura 8

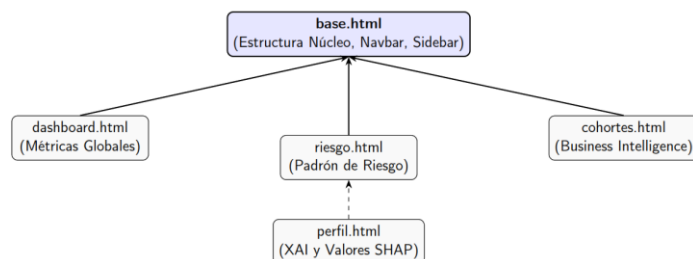
Paleta de colores utilizada en el diseño de la interfaz. Fuente: Elaboración Propia



El archivo maestro base.html actúa como el armazón estructural, conteniendo el menú lateral de acceso rápido y la barra de navegación superior fija, permitiendo que las vistas analíticas especializadas (dashboard.html, riesgo.html, cohortes.html y perfil.html) extiendan dinámicamente este núcleo analítico central. La estructura se ve en la Figura 9.

Figura 9

Topología de herencia de plantillas. Fuente: Elaboración Propia



Las peticiones son dirigidas por el enrutador hacia la capa de negocio (*views.py*), donde los *QuerySets* de Django consultan el repositorio relacional de manera asíncrona, inyectando diccionarios de contexto serializados en formato JSON hacia las plantillas.

El componente visual más innovador implementado mediante esta transferencia de datos es el mecanismo de clasificación cromática de riesgo. Mediante etiquetas personalizadas (*templatetags*) y lógica condicional de plantillas, la interfaz procesa las variables de trayectoria y les asigna reglas CSS condicionales directamente en las vistas del padrón de riesgo (*riesgo.html*) y el directorio docente (*lista.html*), forzando una evaluación donde el ojo humano del usuario clasifica el nivel de vulnerabilidad por el color antes de procesar el texto numérico.

Como se ejemplifica en la Figura 10, la interfaz dispara en rojo intenso ante índices críticos de atraso curricular superiores a tres o promedios deficientes en asignaturas Gatekeepers; el color dorado para riesgos latentes; e inyecta las clases en tonalidades verdes para trayectorias estables. Esta automatización suprime la ceguera institucional e individual, optimizando el flujo de trabajo tutorial al priorizar la atención de los focos rojos sin requerir revisiones manuales previas.

Figura 10

Topología de herencia de plantillas. Fuente: Elaboración Propia

CLAVE	PERIODO	GRUPO	ALUMNOS	% REPRODUCCIÓN HIST.	NIVEL	ACCIÓN
242	1	1	1	100.0%	Crítico	Ver
718	1	1	22	63.64%	Crítico	Ver
5839	1	1	8	62.5%	Crítico	Ver
732	4	5	154	51.3%	Crítico	Ver
5828	1	1	6	50.0%	Crítico	Ver
905	1	1	25	48.0%	Alto	Ver

III. RESULTADOS Y DISCUSIÓN

3.1 Eficiencia operativa y rendimiento computacional del pipeline ETL

La evaluación del artefacto tecnológico comenzó con la validación cuantitativa de la infraestructura informático-analítica encargada de la ingesta y purificación de datos. El procesamiento automatizado abarcó un horizonte temporal de 18 periodos semestrales consecutivos (desde el ciclo de apertura en agosto-diciembre de 2016 hasta agosto-diciembre de 2025), consolidando una masa crítica de información transaccional histórica perteneciente a la carrera de Ingeniería

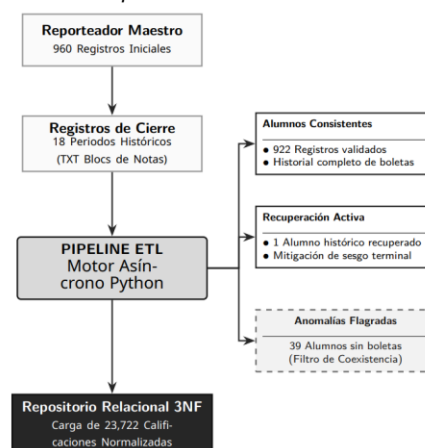
en Sistemas Automotrices del Instituto Tecnológico Superior de Lerdo (ITSL).

En términos de eficiencia operativa, los resultados revelaron una reducción observada de la latencia administrativa institucional. El método manual artesanal utilizado históricamente por las coordinaciones y el personal de apoyo, basado en el cruce celda por celda de reportes aislados y hojas de cálculo desnormalizadas, demandaba semanas enteras de labor burocrática para compilar de forma longitudinal una sola cohorte.

En contraste, la ejecución integrada del pipeline en Python resolvió de manera asíncrona la lectura secuencial, el parsing sintáctico y el balanceo matricial de la totalidad de los ciclos escolares en un tiempo de ejecución inmediato medido en segundos, devolviendo la capacidad de triaje proactivo a la organización antes de que concluyan los periodos críticos de deserción temprana. Este proceso se ve en la Figura 11.

Figura 11

Flujo volumétrico y resultados de auditoría computacional del pipeline ETL. Fuente: Elaboración Propia

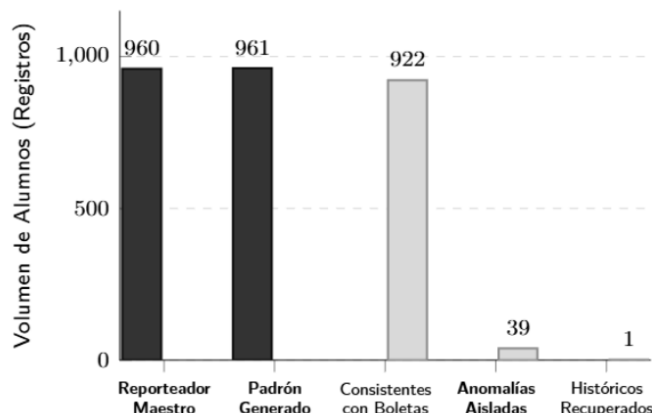


El rigor algorítmico del *backend* se constató en su capacidad de auditoría y diagnóstico preventivo durante la fase de transformación de los datos oscuros. Al procesar el padrón demográfico inicial compuesto por 960 registros provenientes del reporteador maestro, el motor ETL detectó discrepancias críticas de integridad referencial respecto a los archivos transaccionales de texto plano (*AUTO BOLETAS.txt*) y de actas por materia (*AUTO CALIFICACIONES FINALES.txt*). El sistema identificó un subconjunto de 922 alumnos con historiales de cursamiento completamente consistentes y estables.

No obstante, la principal contribución informática para mitigar el sesgo de supervivencia en los modelos analíticos fue la recuperación dinámica de alumnos históricos que habían causado baja definitiva previa al corte administrativo actual, insertándolos de forma automatizada en el padrón oficial generado (elevando la matrícula de análisis a 961 registros) con su respectiva etiqueta de deserción activa. De manera simultánea, el filtro de correspondencia conjunta obligatoria aisló y marcó a 39 "alumnos fantasmas" cuyos folios de matrícula carecían de registros de inscripción o boletas reales en los servidores, emitiendo *logs* de error estructurados que protegen al software de sesgos en el aprendizaje computacional. El resumen de los datos obtenidos en alumnos se muestra en la gráfica de la Figura 12.

Figura 12

Resumen estadístico de datos de alumnos: Elaboración Propia



3.2 Diagnóstico longitudinal de cohortes fundadoras e inestabilidad docente

La disponibilidad de un repositorio relacional estructurado en tercera forma normal (3NF) que integró un volumen masivo de 23,722 registros de calificaciones individuales asociados a una matrícula longitudinal de 961 alumnos y una plantilla histórica de 106 docentes a lo largo de 18 periodos académicos, permitió realizar la primera auditoría retrospectiva de las trayectorias de Ingeniería en Sistemas Automotrices en el ITSL. La problemática del alto abandono temprano en las generaciones fundadoras (que tradicionalmente se atribuía de forma simplista a la falta de vocación estudiantil o a deficiencias socioeconómicas) pudo ser reevaluada científicamente mediante los módulos de analítica visual.

Al cruzar las variables temporales y de rendimiento con la identidad de la plantilla de profesores, el artefacto demostró que el fenómeno de la deserción estuvo fuertemente condicionado por factores estructurales internos de la organización, específicamente por la severidad académica acumulada en asignaturas iniciales y la alta inestabilidad de la plantilla docente durante los años de consolidación de la carrera.

El análisis cuantitativo de las trayectorias aisló un total de 136 asignaturas-docente catalogadas como de alto riesgo o criticidad severa, comúnmente denominadas asignaturas *Gatekeepers*. Estas materias, caracterizadas por registrar índices de reprobación muy superiores a la media institucional del 18.94% (alcanzando promedios del 30.37% en el nivel de riesgo alto y un alarmante 62.44% en el nivel crítico), actúan como verdaderos filtros que modifican el flujo normal de las cohortes.

Como se ilustra en el comportamiento semestral de la Figura 13, esta barrera académica no se distribuye de manera uniforme, sino que se concentra dramáticamente en los tres primeros semestres de la carrera.

El primer semestre registra la mayor acumulación con 33 casos críticos de asignaturas *Gatekeepers*, seguido por el tercer semestre con 20 casos, coincidiendo con la transición hacia las ciencias básicas de la ingeniería. De los 106 docentes evaluados en el histórico, el sistema identificó que una población selecta de 37 profesores únicos estuvo a cargo de impartir estas asignaturas de alta reprobación, exponiendo de

forma sistemática a los estudiantes a un entorno de alta fricción curricular durante su periodo de inducción universitaria.

La inestabilidad de la plantilla docente en estas asignaturas fundamentales constituyó una de las causas del rezago longitudinal. Mediante el rastreo temporal de los identificadores de docentes recuperados de las actas de examen, el artefacto comprobó que, en las primeras generaciones, asignaturas críticas como *Cálculo Diferencial* (Semestre 1) y *Estática* (Semestre 3) fueron impartidas por docentes con altos índices de rotación o contratación transitoria. Esto se resume en el diagrama de la Figura 14.

Figura 13

Distribución semestral de los 136 cursos detectados como asignaturas de alto riesgo y criticidad docente. Fuente: Elaboración Propia

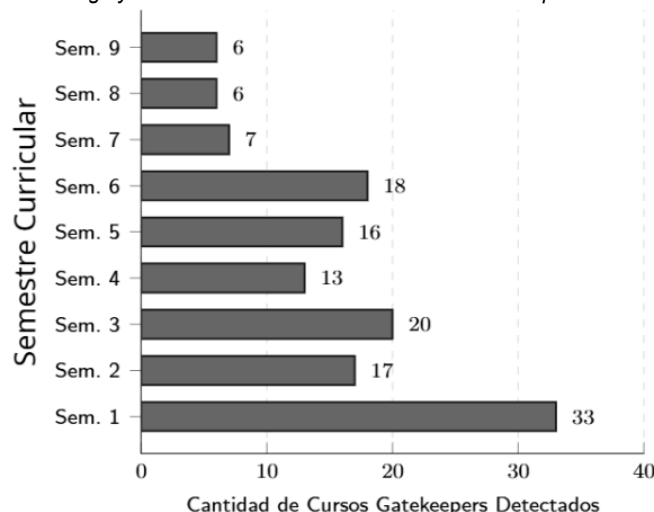
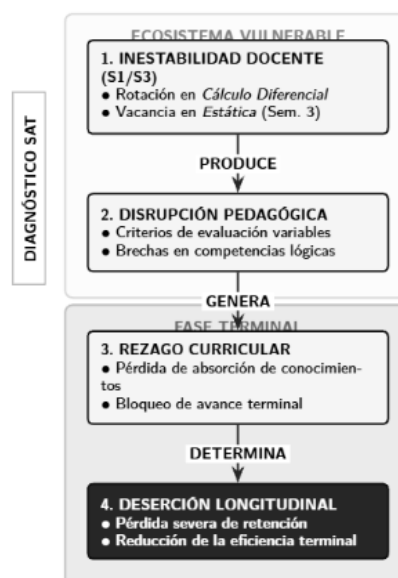


Figura 14

Impacto de la inestabilidad docente en los primeros semestres sobre la eficiencia terminal. Fuente: Elaboración Propia



Al no existir una continuidad pedagógica, el rendimiento de los grupos fluctuó de manera severa entre semestres consecutivos, generando un rezago acumulativo en la adquisición de competencias lógicas básicas. Al llegar a semestres avanzados, este desfase curricular se tradujo en una incapacidad de absorción académica, impactando directamente en la retención terminal. De este modo, la analítica del SAT demostró que el abandono escolar de las generaciones fundadoras no fue un suceso aislado, sino el resultado de un ecosistema académico vulnerable donde la inestabilidad en la asignación docente de los primeros semestres debilitó sistemáticamente las trayectorias escolares.

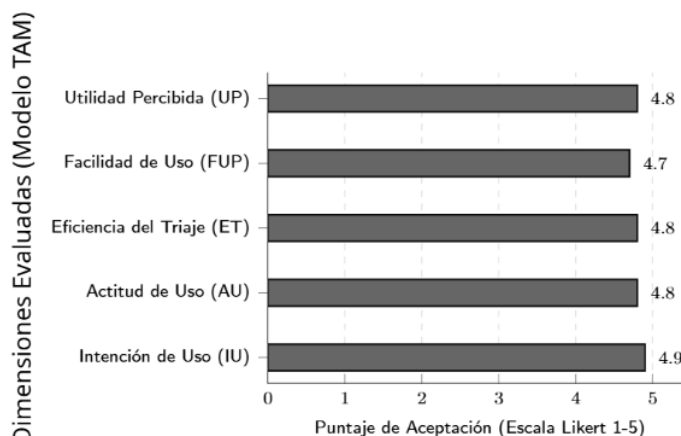
3.3 Evaluación de usabilidad y utilidad operativa con usuarios expertos

La validación del artefacto computacional bajo el marco metodológico de Design Science Research (DSR) requirió evaluar no solo la precisión de su backend, sino también su utilidad operativa y facilidad de uso en un entorno organizativo real. Para este propósito, se implementó un protocolo de evaluación sistemático mediante un grupo de enfoque (focus group) compuesto por diez usuarios participantes de la comunidad académica del ITSL, integrando de manera estratégica a docentes pertenecientes a la academia de Ingeniería en Sistemas Automotrices, al personal de la jefatura de carrera y a una muestra seleccionada de alumnos vigentes de la misma especialidad. La incorporación activa de los estudiantes en este proceso se realizó con la firme intención de que percibieran de forma directa, visual e interactiva el estado real de su trayectoria escolar y el comportamiento general de su carrera a lo largo de los ciclos cursados. El proceso de evaluación se dividió en dos fases: una sesión de interacción guiada con el panel de control utilizando los datos reales ya depurados de las generaciones fundadoras, seguida de la aplicación de un instrumento de recolección de datos cuantitativo basado en la adecuación del Modelo de Aceptación Tecnológica (TAM) y la Escala de Usabilidad del Sistema (SUS).

El análisis cuantitativo de la dimensión de usabilidad arrojó un puntaje de evaluación sobresaliente, alcanzando un valor promedio de 87.5 en la escala SUS, lo que posiciona al software en la categoría de usabilidad "Excelente" (Aceptabilidad Alta). Los docentes, la jefatura de carrera y los estudiantes coincidieron de forma unánime en que la unificación de la interfaz bajo una hoja de estilos común (sat.css) y el modelo de herencia de plantillas (base.html) suprimen por completo el esfuerzo de navegación errática característico de los sistemas tradicionales. Como se ilustra en el desglose dimensional de la Figura 15, las variables evaluadas mediante el modelo TAM confirmaron que la Utilidad Percibida (UP) y la Facilidad de Uso Percibida (FUP) obtuvieron valoraciones promedio de 4.8 y 4.7, respectivamente, sobre una escala Likert de 5 puntos.

Figura 15

Resultados cuantitativos de la evaluación de aceptación tecnológica (TAM). Fuente: Elaboración Propia



IV. CONCLUSIONES

El estudio permitió diseñar y evaluar un artefacto computacional para integrar, depurar y visualizar información académica de la carrera de Ingeniería en Sistemas Automotrices del Instituto Tecnológico Superior de Lerdo. La articulación de Design Science Research con el proceso CRISP-DM proporcionó una estructura metodológica para transformar registros institucionales heterogéneos en un repositorio relacional orientado al análisis longitudinal de las trayectorias escolares. El sistema procesó 23,722 registros de calificaciones correspondientes a 961 estudiantes, 106 docentes y 18 periodos académicos comprendidos entre 2016 y 2025.

La auditoría del proceso ETL mostró que, de los 960 registros procedentes del reporteador maestro, 922 contaban con historiales consistentes, 39 presentaban anomalías de correspondencia y un registro histórico fue recuperado e incorporado al padrón de análisis. Estos resultados evidencian la capacidad del artefacto para identificar inconsistencias de integridad y consolidar información dispersa. No obstante, el desempeño computacional debe interpretarse con cautela, ya que en esta etapa no se documentaron el tiempo exacto de ejecución, su variabilidad entre repeticiones, el porcentaje de automatización ni métricas de precisión calculadas frente a un conjunto de referencia. Por ello, los resultados respaldan la funcionalidad operativa del sistema, pero no permiten establecer todavía su superioridad respecto de otros procedimientos de procesamiento de datos.

El análisis descriptivo identificó 136 combinaciones asignatura-docente clasificadas en niveles de riesgo alto o crítico. La reprobación institucional promedio fue de 18.94 %, mientras que las combinaciones clasificadas en riesgo alto y crítico presentaron promedios de 30.37 % y 62.44 %, respectivamente. La mayor concentración se observó en el primer semestre, con 33 casos, seguida del tercero, con 20.

Asimismo, 37 de los 106 docentes registrados estuvieron asociados con estas asignaturas durante el periodo analizado. En conjunto, los hallazgos muestran una asociación entre la reprobación en asignaturas iniciales, la variación en la asignación docente y la acumulación de rezago curricular.

Debido al carácter retrospectivo y descriptivo del estudio, esta relación no debe interpretarse como evidencia causal; factores académicos, personales, económicos e institucionales no incorporados al análisis también podrían intervenir en la deserción.

La evaluación con diez integrantes de la comunidad académica aportó evidencia inicial sobre la aceptación del sistema. La puntuación promedio de 87.5 en la Escala de Usabilidad del Sistema indicó una valoración favorable de la interfaz. De manera complementaria, las dimensiones del Modelo de Aceptación Tecnológica obtuvieron promedios de entre 4.7 y 4.9 en una escala de cinco puntos.

Estos resultados sugieren que la visualización de trayectorias y la clasificación cromática por niveles de riesgo pueden facilitar la consulta de información y la priorización de casos. Sin embargo, el tamaño reducido y la selección intencional de los participantes limitan la generalización de los resultados; también se requiere informar la variabilidad de las respuestas y evaluar el sistema con grupos de usuarios más amplios y diferenciados. En consecuencia, la principal aportación del trabajo consiste en una infraestructura informática que integra datos escolares longitudinales y los convierte en información diagnóstica para apoyar la gestión académica. Su alcance actual es descriptivo y diagnóstico. Como trabajo futuro, se propone realizar pruebas controladas de rendimiento con tiempos exactos y múltiples ejecuciones; calcular porcentajes de automatización y métricas de precisión mediante datos previamente etiquetados; evaluar la confiabilidad de los instrumentos de aceptación; ampliar la validación a otras cohortes, carreras e instituciones; y analizar longitudinalmente si las intervenciones derivadas del sistema se relacionan con mejoras en la retención y la eficiencia terminal. Estas acciones permitirán determinar con mayor rigor la reproducibilidad, validez externa y utilidad institucional del artefacto.

V. AGRADECIMIENTOS

Agradecimientos al Tecnológico Nacional de México – Instituto Tecnológico Superior de Lerdo, Dgo., especialmente a la División de Ingeniería en Sistemas Automotrices y a la Subdirección de Posgrado e Investigación por las facilidades dadas para esta investigación.

VI. REFERENCIAS

- Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, 8(2), 29–40. http://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Brooke_February_2013.pdf
- Bustamante, A., Valdivia, L., & García, M. (2024). Data mining to identify university student dropout factors. *Applied Sciences*, 14(7), 2423. <https://doi.org/10.3390/app14072423>
- Cardona, M. A., Cudney, E. A., & Snyder, J. E. (2023). Predicting student retention using support vector machines and random forests. *Engineering Management Journal*, 35(1), 1–13. <https://doi.org/10.1080/10429247.2022.2043336>
- Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., & Araújo, R. D. (2022). Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. *Revista Brasileira de Informática na Educação*, 30, 379–395. <https://doi.org/10.5753/rbie.2022.2518>

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Holovaty, A., & Kaplan-Moss, J. (2009). *The definitive guide to Django: Web development done right* (2nd ed.). Apress. <https://doi.org/10.1007/978-1-4302-1937-8>
- Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. In *Visual Data Mining* (pp. 76–90). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71017-2_6
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-020-0143-6>
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*, 54(2), 588–599. <https://doi.org/10.1016/j.compedu.2009.09.008>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125. <https://doi.org/10.3102/00346543045001089>

No.	Rol	Descripción <small>Nota 1: favor de eliminar los corchetes una vez colocado su nombre Nota 2: elimine el todo texto de apoyo que se proporcionó en los asteriscos de cada fila y debajo de cada nombre.</small>
1	Conceptualización [Conceptualization]	Eduardo Machado Díaz
2	Curación de datos [Data curation]	Eduardo Machado Díaz
3	Análisis formal [Formal analysis]	Arturo Serrano Hernández
4	Obtención de fondos [Funding acquisition]	Arturo Serrano Hernández
5	Investigación [Investigation]	Eduardo Machado Díaz
6	Methodología [Methodology]	Eduardo Machado Díaz.
7	Administración del proyecto [Project administration]	Diego Alberto Román Landeros
8	Recursos [Resources]	Diego Alberto Román Landeros
9	Software [Software]	Eduardo Machado Díaz
10	Supervision [Supervision]	Diego Alberto Román Landeros
11	Validación [Validation]	Arturo Serrano Hernández
12	Visualización [Visualization]	Eduardo Machado Díaz
13	Redacción - borrador original [Writing – original draft]	Eduardo Machado Díaz
14	Redacción - revisión y edición [Writing – review & editing]	Arturo Serrano Hernández