proyecto CDA

n.klopstock

October 2025

1 Contexto de negocio

El proyecto propuesto está enfocado en una oficina de consejería académica universitaria. Esta oficina está encargada de orientar a los estudiantes en su planeación de su trayectoria académica, toma de decisiones sobre carga de cursos e identificación de estrategias para mejorar su rendimiento y bienestar académico. Hoy en día, muchas universidades enfrentan desafíos sobre las altas cantidades de estudiantes y sus diferentes y diversos perfiles académicos, lo que exige procesos de orientación más personalizados, ágiles y basados en evidencia.

En este contexto, la oficina de consejería académica de cada deartamento de la universidad tiene un rol fundamental dentro del ecosistema educativo, con acceso a información valiosa sobre historiales académicos de estudiantes. Sin embargo, estos datos rara vez se usan de manera sistemática y analítica para apoyar la toma de decisiones o mejorar los procesos de orientación estudiantil.

1.1 Problemática

Como ya se mencionó anteriormente, los procesos actuales de consejería académica suelen estar sustentados principalmente por experiencias personales y criterio empírico de los consejeros. Estos, basan sus recomendaciones en casos precios o en percepciones individuales, en muchos casos solo conociendo una parte de toda la información disponible del estudiante.

Es cierto que esta práctica ha funcionado históricamente, pero no garantiza consistencia ni objetividad en las orientaciones. Cada semestre, los consejeros deben atender a muchos estudiantes con diversos contextos y trayectorias, lo que dificulta las recomendaciones personalizadas y basadas en evidencia.

La falta de un sistema analítico que aproveche los mencionados datos históricos estudiantiles limita la capacidad de identificar patrones de éxito o riesgo para diferentes casos, y predecir posibles resultados que un estudiante podría alcanzar según sus condiciones.

La problemática tiene un alto impacto sobre la efectividad de la orientación y la planeación académica. Esto, además, puede tener por su parte efecto en aumentar las tasas de graduación semestrales, reducir la deserción, mejorar promedios académicos y optimizar el tiempo y cantidad de consultas privadas entre estudiante y consejero. Por último, dado el contexto de educación superior, el acompañamiento personalizado usando datos puede representar una ventaja competitiva y un fortalecimiento de la experiencia estudiantil.

Dada esta problemática, se propone incorporar una herramienta de ciencia de datos que permita transformar la asesoría académica en un proceso más objetivo, informado y sustentado por datos reales visibles.

1.2 Objetivos y métricas de evaluación

El proyecto propone el desarrollo de un sistema de agrupamiento construido a partir de datos históricos de desempeño estudiantil. Este sistema buscará identificar patrones comunes entre estudiantes con trayectorias similares (cursos

inscritos, en qué semestre, con qué promedio, secuencias de materias, etc.). Estos agrupamientos se usarán luego para predecir posibles cargas académicas futuras y apoyar a los consejeros con las recomendaciones.

Aunque luego se va a explicar a mayor profundidad, el enfoque analítico inicial contempla la aplicación de técnicas de análisis exploratorio de datos y modelos de aprendizaje no supervisado, que permitan segmentar a los estudiantes en grupos de comportamientos académicos comparables.

Las métricas de desempeño (KPIs) que se van a usar para evaluar la efectividad del sistema se dividen en dos: a corto y largo plazo.

A corto plazo:

- **Tiempo promedio de consejería:** reducción del tiempo necesario para cada sesión gracias al acceso a información analítica fundamentada.
- **Nivel de satisfacción del estudiante:** percepción de los estudfiantes respecto a la utilidad, claridad y personalización de esta nueva implementación.

Y, a largo plazo:

- **Promedio ponderado acumulado:** cambio del promedio en el rendimiento académico de los estudiantes luego de esta nueva implementación.
- **Tasa de aprobación de cursos:** proporción de estudiantes que aprueban materias con alta tasa de reprobación dentro de cada grupo identificado.

2 Diseño del producto de datos

El producto de datos propuesto busca transformar el proceso de consejería académica universitaria en uno más fundamentado, eficiente y personalizado.

2.1 Potenciales usuarios y procesos actuales

Hay dos principales usuarios del producto de datos. El primero son los **consejeros académicos** de organizaciones estudiantiles de educación superior. Estos desempeñan un rol importante en el acompañamiento y orientación de los estudiantes durante su proceso estudiantil. Como ya se mencionó antes, actualmente una sesión de consejería se basa en la revisión manual de la carpeta del estudiante y las materias disponibles según el plan de estudios de la carrera en cuestión. Con esta información, el consejero hace recomendaciones que dependen en gran medida de su conocimientos personal del programa académico y su experiencia previa con otros estudiantes y casos similares. Luego, están los **estudiantes**. Estos son usuarios indirectos del sistema, ya que son quienes realmente se benefician de las recomendaciones fundamentadas y justificadas por datos. Para ellos, el producto promete una asesoría más confiable y personalizada que reduzca la duda sobre qué materias ver o cómo planificar el semestre.

El principal dolor identificado en el proceso actual es la naturaleza empírica y poco sistematizada del proceso de recomendación. Las decisiones se sustentan mayoritariamente en la intuición o experiencia individual del consejero, lo que puede derivar en inconsistencias entre diferentes consejeros y un uso ineficiente del tiempo durante las sesiones. Además, la falta de una base de datos analítica dificulta la detección de patrones de éxito o de riesgo académico, así como cualquier tipo de validación sobre el éxito del mismo. La oportunidad de mejora radica en integrar un sistema que proporcione fundamentos objetivos basados en datos históricos, permitiendo así justificar las recomendaciones con evidencia empírica y reducir significativamente el tiempo requerido para el análisis de cada caso.

Por supuesto, este producto no busca reemplazar a un humano consejero. Su objetivo se cumple, también, cuando se complementa con conocimiento humano. Es necesario proveer a los consejeros académicos con herramientas que les permiten tomar decisiones mucho más fundamentadas y con mayor seguridad.

2.2 Requerimientos

El producto debe ser capaz de recuperar automáticamente el historial académico completo de un estudiante mediante su código de identificación y, a partir de una carga académica propuesta para el próximo semestre, posicionarlo dentro de un grupo de referencia mediante un algoritmo de agrupamiento multinivel (multilayer clustering). El sistema proporcionará una predicción de desempeño esperado junto con un índice de confianza que refleja la certeza de la clasificación realizada, permitiendo al consejero tomar decisiones informadas sobre la viabilidad del plan académico propuesto.

Desde un punto de vista funcional, el sistema deberá cumplir con los siguientes requerimientos:

- El consejero ingresará el código del estudiante y el sistema cargará automáticamente su historial académico completo desde la base de datos institucional.
- El consejero definirá la carga académica propuesta para el próximo semestre y el modelo determinará el clúster de pertenencia más probable.
- El sistema calculará y presentará un índice de confianza que cuantifica la certeza de la agrupación, junto con predicciones de rendimiento esperado basadas en el comportamiento histórico del clúster identificado.
- Una aplicación web permitirá visualizar el perfil del estudiante, su clúster proyectado, el índice de confianza y las predicciones, además de explorar escenarios alternativos modificando la carga académica propuesta.
- El sistema almacenará las clasificaciones generadas y los resultados reales posteriores para validación y mejora continua del modelo.

2.3 Componentes analíticos y tecnológicos

El proyecto se compone de dos partes principales. En el **componente analítico**, se emplearán datos históricos de expedientes estudiantiles para identificar patrones mediante técnicas de análisis exploratorio y algoritmos de agrupamiento (clustering), con el fin de crear un modelo capaz de clasificar estudiantes en grupos de desempeño similares. En el **componente tecnológico**, este modelo se integrará en una aplicación web que permitirá ingresar los datos de un estudiante, procesarlos a través del modelo entrenado y mostrar su grupo de pertenencia junto con recomendaciones académicas personalizadas.

3 Implicaciones Éticas

Aunque contamos con datos anonimizados, el análisis de datos académicos con técnicas de *machine learning* requiere tener en cuenta principios éticos y legales rigurosos. En particular, considerando nuestro contexto local en Colombia, la Constitución (Art. 15) y la Ley 1581 de 2012 —*Régimen General de Protección de Datos Personales*— protegen la privacidad e intimidad de los estudiantes. Como explica [y a Distancia (UNAD), 2025], esta ley exige que todo tratamiento de datos personales se realice con consentimiento previo, expreso e informado del titular y bajo principios rectores como legalidad, confidencialidad y acceso restringido.

En la práctica, la universidad ya cumplió con estos pasos y debió haber informado a los estudiantes para qué fines se usarán sus datos académicos y financieros, y sólo tratar la información necesaria. Esto puede comprobarse en la *Normatividad Interna de la Universidad de los Andes respecto al uso de datos personales* (`https://usodedatospersonales.uniandes.edu.co/es/normatividad/normatividad-interna`, `https://usodedatospersonales.uniandes.edu.co/es/`).

La Ley 1581 establece que el objetivo principal del tratamiento de datos es proteger los derechos fundamentales de las personas, especialmente la privacidad y la intimidad [y a Distancia (UNAD), 2025]. En este contexto, los datos académicos (calificaciones, materias cursadas, puntajes, historial) se consideran información sensible en el sentido de que reflejan el desempeño personal. Por ello, cualquier análisis —como *clustering* o modelos de aprendizaje automático— debe garantizar que los datos estudiantiles sean accesibles sólo por el personal autorizado y que su uso corresponda a los fines informados.

Como sugiere el *Archivo General de la Nación* en la *Guía de Anonimización de Datos Estructurados* (agn2021), incluso si los datos están anonimizados, se recomienda aplicar medidas adicionales contra la reidentificación (por ejemplo, agregación de datos o perturbación estadística) para reforzar la privacidad [Archivo General de la Nación, 2021].

Los registros académicos de los estudiantes (matrícula, calificaciones, etc.) son considerados datos privados. Según [ProtecData Latam, 2022], su divulgación indebida podría afectar la dignidad o el buen nombre de los alumnos. Por ello, tanto ética como legalmente, se impone el deber de confidencialidad: "no divulgar ni permitir que se divulgue lo que otras personas nos han confiado" ([FEVAS, 2023]). En la práctica, esto implica emplear controles de acceso adecuados, encriptación y protocolos de seguridad (claves, VPN, etc.) para proteger la base de datos. Sin embargo, dado que sólo contamos con una parte de la base de datos de la universidad, anonimizada y sin toda la información de cada estudiante, el riesgo es menor comparado con la base completa institucional.

Por otro lado, los modelos de *machine learning* (clustering, predicción de riesgo) pueden funcionar como cajas negras para los estudiantes y los consejeros. Como subraya la [UNESCO, 2021], es fundamental garantizar transparencia en cómo estos algoritmos toman decisiones. Se recomienda que los procesos automatizados sean comprensibles para estudiantes y consejeros. La falta de transparencia algorítmica puede generar sesgos en decisiones automatizadas [Área eLearning, 2024].

En consecuencia, es éticamente aconsejable documentar y explicar el funcionamiento del modelo (por ejemplo, qué factores influyen en la predicción de riesgo académico), e incluso considerar auditorías internas de los algoritmos. Esto incluye informar a los estudiantes sobre por qué un cierto perfil o grupo de riesgo fue asignado, de modo que puedan cuestionarlo o refutarlo si es necesario.

En conjunto, estas consideraciones éticas y normativas buscan que el uso de datos y técnicas analíticas en la consejería académica sea seguro, justo y beneficioso para los estudiantes. Todas las prácticas —desde la anonimización de datos hasta la explicación de los resultados— deben regirse por principios de protección de la privacidad, transparencia y equidad.

4 Enfoque Analítico

4.1 Hipótesis y preguntas de negocio

En primer lugar, se plantean las **hipótesis y preguntas de negocio** que guiarán la experimentación. Algunas hipótesis son:

- Existen segmentos claros de estudiantes con trayectorias académicas similares que permiten predecir riesgo de éxito o fracaso con suficiente antelación.
- Las recomendaciones basadas en el grupo al que pertenece un estudiante (por ejemplo, reducir carga, cambiar secuencia de materias o solicitar tutoría) disminuyen la probabilidad de baja académica en el siguiente

semestre.

- La automatización de recomendaciones permitirá dar recomendaciones con mayor fundamentos más allá de la experiencia y aprendizaje empírico del consejero.

A partir de estas hipótesis surgen las siguientes preguntas de negocio:

- ¿Qué características (promedios por materia, número de créditos, etc.) son más discriminantes para identificar qué tan probable o improbable es que un estudiante curse exitosamente su carga académica?
- ¿Qué proporción de estudiantes queda correctamente agrupada en segmentos accionables para la consejería académica?

4.2 Pipeline analítico propuesto

Para responder dichas preguntas se propone un **pipeline analítico** que combine análisis exploratorio de datos, técnicas de reducción de dimensionalidad, algoritmos de agrupamiento y modelos supervisados complementarios.

En la etapa de **análisis exploratorio de datos (EDA)**, se incluirán análisis univariados y bivariados, así como evaluación de la calidad de los datos y análisis y descubrimiento de variables claves.

4.3 Técnicas estadísticas y de aprendizaje automático

Para el desarrollo del modelo analítico se adoptará un enfoque fundamentado en los trabajos de [Clarke, 2008] y [Ochoa et al., 2016], que proponen estrategias para manejar estructuras jerárquicas y heterogeneidad entre grupos en contextos educativos. En el estudio de Clarke (2008), se discute de manera detallada cuándo la estructura de agrupamiento de los datos (por ejemplo, estudiantes anidados dentro de programas académicos) puede o no ser ignorada sin introducir sesgos significativos. Esta distinción es muy importante en entornos con datos escasos o distribuciones desbalanceadas entre grupos, donde un modelo de un solo nivel puede subestimar la varianza y sobreestimar los efectos individuales.

El proyecto considerará la naturaleza jerárquica de los datos académicos, integrando un modelo multinivel que capture la variabilidad tanto entre estudiantes como entre programas. Este enfoque permite representar de manera más realista la estructura de los datos y evitar errores de inferencia derivados de ignorar la dependencia entre observaciones. Además, siguiendo la propuesta de Ochoa (2016), se implementará un modelo de agrupamiento multinivel adaptativo que ajusta dinámicamente el nivel de granularidad del agrupamiento según la densidad y distribución de los datos. Este modelo resulta especialmente útil para predecir riesgo académico en contextos donde coexisten grupos de estudiantes con trayectorias homogéneas y otros con comportamientos atípicos o poco frecuentes.

‘

4.4 Métricas de evaluación

Las métricas de evaluación del modelo se derivan directamente de las consideraciones metodológicas propuestas por Clarke (2008) y Ochoa (2016). Desde la perspectiva de Clarke, es esencial analizar el impacto de ignorar el nivel de agrupamiento, ya que un modelo de un solo nivel puede producir estimaciones sesgadas o sobreconfianza en los resultados, especialmente en contextos donde existen pocos estudiantes por grupo. Por tanto, se evaluará explícitamente la ganancia de ajuste y la reducción del error estándar al comparar un modelo multinivel frente a un modelo de un solo nivel. Esta comparación permitirá determinar la magnitud de la varianza explicada por el nivel grupal y cuantificar si el agrupamiento contribuye significativamente a la precisión de las predicciones.

Por otro lado, siguiendo la propuesta de Ochoa (2016), el modelo adaptativo será evaluado tanto por su capacidad para discriminar niveles de riesgo académico como por su estabilidad frente a conjuntos de datos con distinta densidad o composición. En este sentido, se analizará la homogeneidad interna de los grupos, la varianza entre niveles y la estabilidad de los agrupamientos mediante técnicas de bootstrapping o validación cruzada. En los componentes predictivos, se emplearán métricas clásicas de discriminación como el área bajo la curva ROC (AUC-ROC) y el área bajo la curva de precisión-recobrado (AUC-PR), así como medidas de calibración como el Brier score y las curvas de calibración.

5 Recolección de Datos

5.1 Descripción General de las Fuentes de Datos

El presente proyecto utiliza ocho datasets anonimizados provenientes del observatorio académico institucional. Todos los datos han sido procesados siguiendo un protocolo riguroso de anonimización que garantiza la protección de la identidad de los estudiantes mediante pseudonimización, generalización y supresión selectiva de información identificadora. Los identificadores directos han sido sustituidos por códigos alfanuméricos únicos que mantienen la consistencia relacional entre datasets.

5.1.1 Historial Estados Estudiante

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante y periodo académico
Alcance: Todos los periodos académicos registrados desde el ingreso del estudiante

Este dataset constituye la columna vertebral del análisis longitudinal, registrando la trayectoria académica completa de cada estudiante. Contiene 21 columnas que documentan el estado académico, programas matriculados, cambios de estado (suspensiones, pruebas académicas, reintegros), y métricas acumuladas de alertas académicas en cada periodo.

Características principales:

- Permite seguimiento temporal del estado académico de los estudiantes
- Incluye información de hasta dos programas simultáneos por estudiante
- Registra transiciones entre estados académicos con sus periodos correspondientes
- Documenta eventos críticos como suspensiones y pruebas académicas acumuladas

Utilidad para el proyecto: Esencial para poder entender el contexto del estudiante en la universidad hasta el momento.

5.1.2 Historial Materias Estudiante

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante, curso, periodo académico y nivel académico
Alcance: Registro completo de todas las materias cursadas

Dataset de máxima granularidad que registra cada materia inscrita por cada estudiante en cada periodo. Contiene 15 columnas con información sobre créditos, calificaciones, estado de la materia (aprobada, reprobada, retirada), y atributos específicos del curso. Los identificadores de estudiante y curso han sido pseudonimizados de

forma consistente.

Características principales:

- Nivel de detalle más fino disponible en el sistema
- Incluye calificaciones parciales y finales
- Registra diferentes modalidades de calificación y estados de materia
- Documenta sección, campus y atributos especiales de cada curso

Utilidad para el proyecto: En el contexto de las consejerías académicas es importantísimo poder entender en que materias el estudiante y en general los estudiantes de cierto perfil suelen tener problemas y que combinaciones de las mismas llevan a mejores resultados académicos (balanceo de la carga académica).

5.1.3 Historial Rendimiento Académico

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante, periodo académico y nivel académico
Alcance: Métricas agregadas semestrales desde el ingreso

Dataset que consolida indicadores cuantitativos de desempeño académico por periodo. Contiene 22 columnas con métricas como PGA (Promedio General Acumulado), promedios semestrales, distribución de créditos por resultado (aprobados, reprobados, retirados, incompletos), y alertas específicas.

Características principales:

- Métricas acumuladas y por semestre para análisis comparativo
- Incluye indicadores de progreso curricular (porcentaje de créditos aprobados)
- Registra materias cursadas por tercera vez (alerta crítica)
- Documenta créditos homologados que afectan el avance curricular

Utilidad para el proyecto: Vista académica general de los estudiantes, esto permite entender mucho mejor el desempeño general del estudiante en un determinado semestre.

5.1.4 Información Actual Estudiante

Naturaleza temporal: Snapshot actual (último periodo disponible)
Granularidad: Por estudiante
Alcance: Estado más reciente de cada estudiante activo

Dataset transversal que captura el estado actual de los estudiantes. Contiene 30 columnas con información demográfica generalizada (edad, género codificado, estrato socioeconómico codificado), académica (programas, estado, penalizaciones), y de admisión (información de colegio pseudonimizada, puntajes ICFES descontextualizados).

Características principales:

- Variables demográficas protegidas mediante generalización

- Género Estrato socioeconómico codificado
- Puntajes ICFES específicos renombrados como ICFES_CRITERIO_1 a ICFES_CRITERIO_5
- Información de colegio con identificadores pseudonimizados (SCH_XXXXXXXX)

Utilidad para el proyecto: Si bien es importante entender a profundidad la dimensión académica del estudiante, también es necesario poder determinar que tanto afecta la dimensión social y económica del estudiante.

5.1.5 Horarios Curso

Naturaleza temporal: Por periodo académico
Granularidad: Por periodo, curso (CRN) y franja horaria
Alcance: Programación académica completa por periodo

Dataset que documenta la programación de cada curso ofertado. Contiene 16 columnas con información de franjas horarias (inicio/fin de fecha y hora), días de la semana, ubicación (código de salón), y sección. Los identificadores CRN y CODIGO_CURSO están pseudonimizados de forma consistente con el dataset de Historial Materias.

Utilidad para el proyecto: Puede llegar a ser importante y relevante en alguno de los de niveles de clusterización.

5.1.6 Riesgos Estudiante Pregrado

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante y periodo académico
Alcance: Indicadores binarios de riesgo por periodo

Dataset especializado para sistemas de alerta temprana. Contiene 12 columnas con indicadores binarios (0/1) que señalan la presencia de factores de riesgo académico predefinidos. Cada indicador representa una condición específica que la literatura identifica como predictor de deserción o bajo rendimiento.

Características principales:

- 10 indicadores de riesgo independientes calculados algorítmicamente
- Valores binarios facilitan modelado de machine learning
- Incluye riesgos de corto plazo (semestre actual) y acumulativos
- Cubre múltiples dimensiones: rendimiento, retiros, progreso curricular

Indicadores incluidos:

- Tres o más materias parcialmente perdidas
- Tres o más materias retiradas
- Bajo promedio en segundo semestre
- Bajo promedio desde tercer semestre en adelante
- Bajo promedio móvil

- Pocos créditos inscritos respecto a matrícula pagada
- Materias bloqueantes no aprobadas
- Semestres perdidos con estado académico normal
- No aprobación de lectura/inglés al sexto semestre
- Porcentaje de créditos aprobados menor al 50%

Utilidad para el proyecto: En las consejerías académicas es fundamental poder conocer los riesgos que tuvo el estudiante en el momento de tomar cierta carga académica.

5.1.7 Información Financiera Estudiante

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante, periodo y tipo de financiamiento
Alcance: Registro de todos los tipos de financiamiento aplicados

Dataset que documenta las modalidades de financiamiento de la matrícula de cada estudiante. Contiene 12 columnas con información sobre tipos de financiamiento (becas, créditos, descuentos), fechas de aplicación, porcentajes de aporte, y clasificación principal del financiamiento por semestre.

Características principales:

- Múltiples registros por estudiante-periodo (diferentes tipos de financiamiento)
- Variable de ranking para identificar financiamiento principal
- Diferenciación entre financiamiento principal y parcial
- Clasificación jerárquica en tipos específicos, principales y generales

Utilidad para el proyecto: Fundamental para análisis de equidad financiera y su relación con permanencia y éxito académico.

5.1.8 Percentiles Académicos Estudiante

Naturaleza temporal: Histórico completo
Granularidad: Por estudiante, periodo y programa académico
Alcance: Ubicación percentil dentro de cada programa

Dataset que proporciona contexto de rendimiento relativo. Contiene 11 columnas que ubican a cada estudiante dentro de la distribución de rendimiento de su programa académico, incluyendo percentil de PGA, clasificación por tipo de estudiante, y agrupación por avance semestral.

Características principales:

- Percentiles calculados por programa específico (comparación intra-programa)
- Clasificación de estudiantes por características (Estudiante activo o graduado)
- Agrupación por semestre de avance para comparaciones justas

- Permite análisis longitudinal de posición relativa

Utilidad para el proyecto: Esencial para análisis comparativo que considera el contexto específico de cada programa. Permite identificar estudiantes en riesgo relativo a su grupo de pares, independientemente de las diferencias de dificultad entre programas.

5.2 Referencias al Diccionario de Datos

Las descripciones detalladas de cada columna, sus tipos de datos, y valores únicos (donde aplique) se encuentran documentadas en el **Anexo A: Diccionario de Datos Anonimizados**. El diccionario incluye:

- Nombre de la columna
- Tipo de dato
- Descripción funcional

6 Entendimiento de los datos

Debido a la cantidad de fuentes de datos, así como el alto número de columnas, se decidió generar un informe únicamente orientado al entendimiento de los datos. El cual se adjunta dentro de la entrega.

7 Conclusiones

Se confirma que la orientación académica basada exclusivamente en la experiencia personal de los consejeros resulta insuficiente frente a la complejidad y diversidad de trayectorias estudiantiles actuales. Los datos explorados evidencian la existencia de patrones recurrentes en el desempeño y la progresión de los estudiantes, lo que valida la hipótesis de que es posible segmentar poblaciones y anticipar riesgos mediante técnicas de agrupamiento multinivel. Además, el diseño metodológico basado en los trabajos de Clarke (2008) y Ochoa et al. (2016) permite integrar un enfoque estadísticamente robusto que considera la jerarquía natural de los datos académicos y la heterogeneidad entre programas. Estos resultados preliminares refuerzan la pertinencia de avanzar hacia la implementación completa del modelo propuesto, con la construcción del pipeline analítico, la integración de los datasets anonimizados y el desarrollo de la aplicación web para consejeros y estudiantes.

En el futuro se podría realizar una serie de pruebas piloto con estudiantes y consejeros para evaluar el impacto del sistema en métricas institucionales ya mencionadas al comienzo de este documento. Finalmente, este proyecto sienta las bases para una transformación gradual del proceso de acompañamiento académico, combinando la experiencia humana con la evidencia empírica proveniente del análisis de datos.

References

[Archivo General de la Nación, 2021] Archivo General de la Nación (2021). Guía de anonimización de datos estructurados. `https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicacionees/Guia_de_Anonimizacion-min.pdf`. Ministerio de Cultura.

[Clarke, 2008] Clarke, P. (2008). When can group level clustering be ignored? multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, 62(8):752–758.

[FEVAS, 2023] FEVAS (2023). Guía: La confidencialidad en el ámbito educativo. https://fevas.org/wp-content/uploads/2023/04/GUIA-LA-CONFIDENCIALIAD-EN-EL-AMBITO-EDUCATIVO.pdf.

[Ochoa et al., 2016] Ochoa, X., Méndez, G., Chiluiza, K., and Luzardo, G. (2016). Adaptive multilevel clustering model for the prediction of academic risk. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge (LAK '16)*, pages 237–241, Edinburgh, United Kingdom. ACM.

[ProtecData Latam, 2022] ProtecData Latam (2022). Concepto sector educación sobre tratamiento de datos personales. https://protecdatalatam.com/wp-content/uploads/2022/03/CONCEPTO-SECTOR-EDUCACION.pdf.

[UNESCO, 2021] UNESCO (2021). El uso ético de la inteligencia artificial en la educación. https://unac.edu.mx/blog-2/el-uso-etico-de-la-inteligencia-artificial-en-la-educacion/. Universidad de las Américas y el Caribe.

[y a Distancia (UNAD), 2025] y a Distancia (UNAD), U. N. A. (2025). Lo que debes saber sobre el tratamiento y protección de datos personales. https://noticias.unad.edu.co/index.php/2025/7277-lo-que-debes-saber-sobre-el-tratamiento-y-proteccion-de-datos-personales. Noticias UNAD.

[Área eLearning, 2024] Área eLearning (2024). Ética y sesgos en ia educativa: retos y soluciones. https://areaelearning.com/etica-y-sesgos-en-ia-educativa-retos-y-soluciones/.

A Diccionario de Datos Anonimizados

Este anexo presenta las descripciones detalladas de todas las columnas presentes en los datasets anonimizados utilizados en el proyecto. Para cada variable se indica:

- **Nombre de la columna**: Nombre final después de la anonimización
- **Tipo de dato**: Tipo de dato original en el sistema fuente
- **Descripción**: Explicación funcional de la variable

A.1 Dataset 1: Historial Estados Estudiante

Table 1: Variables del Dataset: Historial Estados Estudiante

Columna	Tipo	Descripción
CODIGO_ ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP: año + periodo).
SEMESTRE_SEGUN_ CREDITOS	Decimal	Semestre según créditos aprobados dividido entre 17 créditos esperados.
MATRICULADO_ PERIODO_ACTUAL	String	Indica si un estudiante se encuentra matriculado en cada periodo académico. Valores posibles: NO, SI.
ESTADO_ ACADEMICO	String	Estado académico del estudiante periodo a periodo. Existen 26 estados posibles.
CODIGO_ PROGRAMA_1	String	Código del primer programa del estudiante.
PROGRAMA_1	String	Nombre del primer programa del estudiante.
NIVEL_ PROGRAMA_1	String	Nivel académico del primer programa. Valores posibles: Pregrado, Magíster, etc.
CODIGO_ PROGRAMA_2	String	Código del segundo programa (si aplica).
PROGRAMA_2	String	Nombre del segundo programa (si aplica).
NIVEL_ PROGRAMA_2	String	Nivel académico del segundo programa. Valores posibles: Pregrado, Magíster, etc.
DESCRIPCION_ ESTADO_RETIRO	String	Descripción del tipo de retiro si aplica. Valores posibles: Retiro voluntario, Abandono, Grado.
REINTEGRO_ REINGRESO	String	Indica si el estudiante se reintegró o reingresó. Valores posibles: SI, NO.
ESTADO_ ACADEMICO_ INICIAL	String	Estado académico al inicio del periodo.

Continúa en la siguiente página

Table 1 – *Continuación de la página anterior*

Columna	Tipo	Descripción
PERIODO_ESTADO_ ACADEMICO_ INICIAL	Int64	Periodo del estado académico inicial.
ESTADO_ ACADEMICO_FINAL	String	Estado académico al final del periodo.
PERIODO_ESTADO_ ACADEMICO_FINAL	Int64	Periodo del estado académico final.
ENTRA_A_ SUSPENSION	String	Indica si entra a suspensión en el periodo. Valores posibles: SI, NO.
ENTRA_A_PRUEBA_ ACADEMICA	String	Indica si entra a prueba académica en el periodo. Valores posibles: SI, NO.
SUSPENSIONES_ ACUMULADAS	Int64	Número acumulado de suspensiones.
PRUEBAS_ ACADEMICAS_ ACUMULADAS	Int64	Número acumulado de pruebas académicas.
INCOMPLETOS_ TOTALES_ ACUMULADOS	Int64	Número acumulado de incompletos.

A.2 Dataset 2: Historial Materias Estudiante

Table 2: Variables del Dataset: Historial Materias Estudiante

Columna	Tipo	Descripción
CODIGO_ ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
CODIGO_CURSO	String	Código del curso. Todos los valores en esta columna tienen el formato: CRS_XXXXXXXX.
CRN	String	Course Reference Number, identificador único de sección. Todos los valores en esta columna tienen el formato: CRN_ XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
NIVEL_ ACADEMICO_ ESTUDIANTE	String	Nivel académico del estudiante al cursar la materia. Valores posibles: Pregrado, Posgrado.
NUMERO_ CREDITOS	Int64	Número de créditos de la materia.
SECCION	String	Sección específica del curso.
CODIGO_MODO_DE_ CALIFICACION	String	Código del sistema de calificación aplicado. Valores posibles: N, AB, etc.

Continúa en la siguiente página

Table 2 – *Continuación de la página anterior*

Columna	Tipo	Descripción
CALIFICACION_ PARCIAL	Decimal	Calificación parcial obtenida en el curso.
CALIFICACION_ FINAL	Decimal	Calificación final obtenida en el curso.
ESTATUS	String	Estado de la materia.
ATRIBUTOS	String	Atributos especiales del curso.
CODIGO_CAMPUS	String	Código del campus donde se dictó el curso.
RESULTADO	String	Resultado final de la materia. Valores posibles: Aprobado, Reprobado, Retirado.
ES_HISTORICO_ ACTUAL	String	Indica si es parte del historial actual o antiguo. Valores posibles: SI, NO.

A.3 Dataset 3: Historial Rendimiento Académico

Table 3: Variables del Dataset: Historial Rendimiento Académico

Columna	Tipo	Descripción
CODIGO_ ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
PGA	Decimal	Promedio General Acumulado del estudiante.
PROMEDIO_ SEMESTRAL	Decimal	Promedio del semestre actual.
CREDITOS_PGA	Int64	Créditos que cuentan para el cálculo del PGA.
DESCRIPCION_ NIVEL_PROGRAMA	String	Nivel académico del programa. Valores posibles: Pregrado, Magíster, Doctorado.
CREDITOS_ APROBADOS	Int64	Total acumulado de créditos aprobados.
CREDITOS_ REPROBADOS	Int64	Total acumulado de créditos reprobados.
CREDITOS_ INCOMPLETOS	Int64	Total acumulado de créditos incompletos.
CREDITOS_ RETIRADOS	Int64	Total acumulado de créditos retirados.
CREDITOS_ PENDIENTES	Int64	Total acumulado de créditos pendientes.
CREDITOS_ HOMOLOGADOS	Int64	Total acumulado de créditos homologados.
PORCENTAJE_ CREDITOS_ APROBADOS	Decimal	Porcentaje de créditos aprobados sobre el total.

Continúa en la siguiente página

Table 3 – *Continuación de la página anterior*

Columna	Tipo	Descripción
TOTAL_ SEMESTRES_ MATRICULADOS	Int64	Número de semestres matriculados.
MATERIA_POR_ TERCERA_VEZ	String	Indica si cursó alguna materia por tercera vez. Valores posibles: SI, NO.
CREDITOS_ APROBADOS_ SEMESTRE	Int64	Créditos aprobados en el semestre actual.
CREDITOS_ REPROBADOS_ SEMESTRE	Int64	Créditos reprobados en el semestre actual.
CREDITOS_ INCOMPLETOS_ SEMESTRE	Int64	Créditos incompletos en el semestre actual.
CREDITOS_ RETIRADOS_ SEMESTRE	Int64	Créditos retirados en el semestre actual.
CREDITOS_ PENDIENTES_ SEMESTRE	Int64	Créditos pendientes en el semestre actual.
CREDITOS_ HOMOLOGADOS_ SEMESTRE	Int64	Créditos homologados en el semestre actual.
PORCENTAJE_ CREDITOS_ APROBADOS_ SEMESTRE	Decimal	Porcentaje de créditos aprobados en el semestre.

A.4 Dataset 4: Información Actual Estudiante

Table 4: Variables del Dataset: Información Actual Estudiante

Columna	Tipo	Descripción
CODIGO_ ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
PERIODO	Int64	Periodo académico más reciente.
EDAD	Int64	Edad del estudiante en años.
GENERO	String	Género del estudiante. Todos los valores en esta columna tienen el formato: A, B u OTRO.
CODIGO_FACULTAD_ PROGRAMA_1	String	Código de la facultad del primer programa.

Continúa en la siguiente página

Table 4 – *Continuación de la página anterior*

Columna	Tipo	Descripción
CODIGO_ PROGRAMA_1	String	Código del primer programa.
PROGRAMA_1	String	Nombre del primer programa.
NIVEL_PROGRAMA_1	String	Nivel académico del primer programa. Valores posibles: Pregrado, Magíster, Doctorado.
CODIGO_FACULTAD_ PROGRAMA_2	String	Código de la facultad del segundo programa (si aplica).
CODIGO_ PROGRAMA_2	String	Código del segundo programa (si aplica).
PROGRAMA_2	String	Nombre del segundo programa (si aplica).
NIVEL_PROGRAMA_2	String	Nivel académico del segundo programa. Valores posibles: Pregrado, Magíster, Doctorado.
ESTADO_ ACADEMICO	String	Estado académico actual del estudiante.
MATRICULADO_ PERIODO_ACTUAL	String	Indica si está matriculado en el periodo actual. Valores posibles: SI, NO.
ULTIMO_PERIODO_ MATRICULADO	Int64	Último periodo en el que estuvo matriculado.
PERIODO_ CATALOGO	Int64	Periodo de catálogo académico aplicable.
PENALIZADO_ EXTRACREDITACION	String	Indica si tiene penalización por extracreditación. Valores posibles: SI, NO.
PENSUM_ REFORMADO	String	Indica si está bajo pensum reformado. Valores posibles: SI, NO.
PROMEDIO_MOVIL	Decimal	Promedio móvil del estudiante.
CREDITOS_ MAXIMOS	Int64	Número máximo de créditos que puede inscribir.
NOMBRE_COLEGIO	String	Nombre del colegio de origen. Todos los valores en esta columna tienen el formato: SCH_XXXXXXXX.
CIUDAD_COLEGIO	String	Ciudad del colegio de origen.
DEPARTAMENTO_ COLEGIO	String	Departamento del colegio de origen.
PUNTAJE_ICFES	Decimal	Puntaje total ICFES/Saber 11.
ICFES_CRITERIO_1	Decimal	Puntaje en primer criterio de evaluación .
ICFES_CRITERIO_2	Decimal	Puntaje en segundo criterio de evaluación .
ICFES_CRITERIO_3	Decimal	Puntaje en tercer criterio de evaluación .
ICFES_CRITERIO_4	Decimal	Puntaje en cuarto criterio de evaluación .
ICFES_CRITERIO_5	Decimal	Puntaje en quinto criterio de evaluación.
ESTRATO_ SOCIOECONOMICO	String	Estrato socioeconómico del estudiante. Todos los valores en esta columna tienen valores entre U y Z

A.5 Dataset 5: Horarios Curso

Table 5: Variables del Dataset: Horarios Curso

Columna	Tipo	Descripción
CRN	String	Course Reference Number, identificador único de sección. Todos los valores en esta columna tienen el formato: CRN_ XXXXXXXX.
CODIGO_CURSO	String	Código del curso. Todos los valores en esta columna tienen el formato: CRS_XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
SECCION	String	Sección específica del curso.
FECHA_INICIO_ FRANJA	Date	Fecha de inicio de la franja horaria.
FECHA_FIN_ FRANJA	Date	Fecha de finalización de la franja horaria.
HORA_INICIO_ FRANJA	Time	Hora de inicio de la franja horaria.
HORA_FIN_FRANJA	Time	Hora de finalización de la franja horaria.
CODIGO_SALON	String	Código del salón donde se dicta el curso.
LUNES	Int64	Indica si hay clase los lunes. Valores posibles: 0, 1.
MARTES	Int64	Indica si hay clase los martes. Valores posibles: 0, 1.
MIERCOLES	Int64	Indica si hay clase los miércoles. Valores posibles: 0, 1.
JUEVES	Int64	Indica si hay clase los jueves. Valores posibles: 0, 1.
VIERNES	Int64	Indica si hay clase los viernes. Valores posibles: 0, 1.
SABADO	Int64	Indica si hay clase los sábados. Valores posibles: 0, 1.
DOMINGO	Int64	Indica si hay clase los domingos. Valores posibles: 0, 1.

A.6 Dataset 6: Riesgos Estudiante Pregrado

Table 6: Variables del Dataset: Riesgos Estudiante Pregrado

Columna	Tipo	Descripción
CODIGO_ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_ XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
tres_o_mas_materias_parcialmente_ perdidas	Int64	Indicador de riesgo: tres o más materias con calificación parcial reprobatoria. Valores posibles: 0, 1.
tres_o_mas_materias_retiradas	Int64	Indicador de riesgo: tres o más materias retiradas en el semestre. Valores posibles: 0, 1.
bajo_promedio_segundo_semestre	Int64	Indicador de riesgo: promedio bajo en segundo semestre de carrera. Valores posibles: 0, 1.

Continúa en la siguiente página

Table 6 – *Continuación de la página anterior*

Columna	Tipo	Descripción
bajo_promedio_tercer_semestre_o_mas	Int64	Indicador de riesgo: promedio bajo desde tercer semestre en adelante. Valores posibles: 0, 1.
bajo_promedio_movil	Int64	Indicador de riesgo: promedio móvil por debajo del umbral. Valores posibles: 0, 1.
pocos_creditos_inscritos_para_matricula_pagada	Int64	Indicador de riesgo: créditos inscritos significativamente menores a lo esperado. Valores posibles: 0, 1.
materias_bloqueantes	Int64	Indicador de riesgo: tiene materias bloqueantes sin aprobar. Valores posibles: 0, 1.
semestres_perdidos_con_estado_normal	Int64	Indicador de riesgo: semestres reprobados con estado académico normal. Valores posibles: 0, 1.
no_lectura_ingles_sexto_semestre	Int64	Indicador de riesgo: no ha aprobado requisito de lectura/inglés al sexto semestre. Valores posibles: 0, 1.
porcentaje_creditos_aprobados_menor_50	Int64	Indicador de riesgo: menos del 50% de créditos inscritos aprobados. Valores posibles: 0, 1.

A.7 Dataset 7: Información Financiera Estudiante

Table 7: Variables del Dataset: Información Financiera Estudiante

Columna	Tipo	Descripción
CODIGO_ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
TIPO_FINANCIAMIENTO_ESPECIFICO	String	Tipo detallado del financiamiento aplicado. Ejemplos: Beca ICETEX, Descuentos, etc.
FECHA_AFECTACION	Date	Fecha en que se aplicó el financiamiento.
TIPO_SECCION	String	Tipo de sección financiera.
OP_PRINCIPAL_PARCIAL	String	Indica si es operación financiera principal o parcial. Valores posibles: Principal, Parcial.
PORCENTAJE_APORTE_ITEM	Decimal	Porcentaje de aporte del ítem financiero.
RANKING	Int64	Ranking del financiamiento (1=principal).
DESCRIPCION_NIVEL_PROGRAMA_1	String	Nivel académico del programa asociado. Valores posibles: Pregrado, Magíster, etc.
TIPO_FINANCIAMIENTO_PRINCIPAL_SEMESTRE	String	Tipo de financiamiento principal del semestre.
TIPO_FINANCIAMIENTO_PRINCIPAL	String	Categoría general de financiamiento principal. Ejemplos: Beca, Crédito, Propio.
TIPO_MATRICULA_SEMESTRE	String	Tipo de matrícula del semestre.

A.8 Dataset 8: Percentiles Académicos Estudiante

Table 8: Variables del Dataset: Percentiles Académicos Estudiante

Columna	**Tipo**	**Descripción**
CODIGO_ ESTUDIANTE	Int64	Código estudiantil del estudiante. Todos los valores en esta columna tienen el formato: EST_XXXXXXXX.
PERIODO	Int64	Periodo académico como número de 6 dígitos (YYYYPP).
NIVEL_ACADEMICO	String	Nivel académico del programa. Valores posibles: Pregrado, Magíster, Doctorado.
CODIGO_ PROGRAMA	String	Código del programa académico.
PROGRAMA	String	Nombre del programa académico.
PGA	Decimal	Promedio General Acumulado del estudiante.
PERCENTIL_PGA_ PROGRAMA	Decimal	Percentil del PGA dentro del programa específico (0-100).
TIPO_ESTUDIANTE	String	Clasificación del tipo de estudiante. Valores posibles: Regular, Transferencia, etc.
SEMESTRE_ AVANCE	Int64	Semestre de avance curricular del estudiante.
GRUPO_ SEMESTRES	String	Agrupación de estudiantes por rango de semestres. Ejemplos: 1-2, 3-4, 5-6.
NUMERO_ PROGRAMA	Int64	Número de programa al que pertenece. Valores posibles: 1, 2.