RESUMEN EJECUTIVO
Sistema Inteligente de Consejería Académica

Introducción

Este documento sintetiza el proyecto de ciencia de datos para transformar la consejería académica universitaria mediante un sistema de agrupamiento jerárquico. El sistema permitirá fundamentar recomendaciones en evidencia empírica derivada de patrones históricos de desempeño estudiantil, abarcando desde la problemática hasta el análisis exploratorio que valida la viabilidad del sistema propuesto.

1 Definición de la Problemática y Entendimiento del Negocio

La consejería académica universitaria enfrenta desafíos por altas cantidades de estudiantes con perfiles diversos. Los procesos actuales se sustentan principalmente en experiencia personal y criterio empírico de los consejeros, sin aprovechar datos históricos para identificar patrones de éxito o riesgo. Esta situación limita la consistencia, objetividad y personalización de las orientaciones, impactando tasas de graduación, deserción estudiantil y promedios académicos.

La solución propuesta incorpora una herramienta de ciencia de datos que agrupa estudiantes según patrones históricos de desempeño (cursos inscritos, promedios, secuencias de materias). Estos agrupamientos permitirán predecir cargas académicas viables y apoyar recomendaciones fundamentadas mediante análisis exploratorio y aprendizaje no supervisado, siguiendo enfoques exitosos implementados en sistemas de apoyo a consejería académica [Gutiérrez et al., 2020].

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 sis-

tema 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 sistema tiene dos usuarios principales: consejeros académicos (usuarios primarios) que actualmente basan recomendaciones en revisión manual de carpetas y experiencia personal; y estudiantes (usuarios indirectos) que recibirán asesoría más confiable y personalizada.

El principal dolor es la naturaleza empírica del proceso actual, generando inconsistencias y uso ineficiente del tiempo. La oportunidad radica en proporcionar fundamentos objetivos basados en datos históricos de perfiles similares.

2.1 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.

El proyecto incluye dos componentes: analítico (identificar patrones mediante clustering) y tecnológico (aplicación web que integre el modelo para clasificar estudiantes y mostrar recomendaciones personalizadas).

2.2 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.

El primer prototipo fue diseñado utilizando Figma y puede verse en el siguiente enlace: https://www.figma.com/community/file/1563043270302634802

3 Implicaciones Éticas

El análisis requiere considerar la Constitución Colombiana (Artículo 15) y la Ley 1581 de 2012 sobre protección de datos personales, exigiendo consentimiento previo, expreso e informado [y a Distancia (UNAD), 2025, ProtecData Latam, 2022]. La universidad cumple con estas obligaciones según su normatividad interna, garantizando la confidencialidad en el tratamiento de información académica [FEVAS, 2023]. Los datos académicos son sensibles y deben ser accesibles exclusivamente por personal autorizado.

Se aplicarán medidas contra reidentificación mediante técnicas de agregación, perturbación estadística, y supresión selectiva de datos [Archivo General de la Nación, 2021], junto con controles de acceso, encriptación y protocolos de seguridad robustos. Los modelos de machine learning pueden funcionar como cajas negras; siguiendo recomendaciones internacionales sobre el uso ético de la inteligencia artificial en educación [UNESCO, 2021], se garantizará transparencia algorítmica, documentación exhaustiva y auditorías periódicas para detectar sesgos algorítmicos [Área eLearning, 2024]. Los estudiantes serán informados proactivamente sobre clasificaciones asignadas, pudiendo cuestionarlas o refutarlas.

4 Enfoque Analítico

Las hipótesis principales establecen que existen segmentos claros de estudiantes con trayectorias similares para predecir riesgo o éxito académico; que las recomendaciones fundamentadas en el grupo de pertenencia pueden disminuir probabilidad de baja académica;

y que la automatización parcial proporcionará orientación con fundamentos más sólidos.

Las preguntas de negocio buscan identificar características discriminantes (promedios, créditos, patrones de retiro, secuencias curriculares) y determinar la proporción de estudiantes correctamente agrupables en segmentos accionables.

Se adoptará el enfoque metodológico propuesto por Clarke [Clarke, 2008] y Ochoa et al. [Ochoa et al., 2016] para manejar estructuras jerárquicas en contextos educativos. Se implementará un modelo multinivel que capture variabilidad entre estudiantes individuales y entre programas académicos, adaptando dinámicamente la granularidad del clustering según densidad y distribución de datos.

Las métricas de evaluación incluyen: ganancia de ajuste multinivel vs. un nivel; homogeneidad interna de grupos; varianza entre niveles jerárquicos; estabilidad mediante bootstrapping; AUC-ROC y AUC-PR para discriminación; Brier score y curvas de calibración para validar probabilidades predichas.

5 Recolección de Datos

El proyecto utiliza ocho datasets anonimizados del observatorio académico institucional con protocolo riguroso de pseudonimización, generalización y supresión selectiva. Los identificadores directos fueron sustituidos por códigos alfanuméricos únicos que mantienen consistencia relacional.

Los datasets principales incluyen: Historial Estados Estudiante (616,085 registros, trayectoria académica completa); Historial Materias Estudiante (4,931,740 registros, cada materia inscrita con créditos y calificaciones); Historial Rendimiento Académico (611,654 registros, indicadores como PGA y distribución de créditos); Información Actual Estudiante (222,407 registros, estado reciente con datos demográficos generalizados); Horarios Curso (444,834 registros, programación académica); Riesgos Estudiante Pregrado (369,370 registros, indicadores binarios de factores de riesgo); Información Financiera Estudiante (793,198 registros, modalidades de financiamiento); Percentiles Académicos Estudiante (158,429 registros, rendimiento relativo por programa).

6 Entendimiento de los Datos

La calidad general es notablemente buena con menos del 1 % de nulos en variables críticas. Excepciones: variables de segundo programa (¿80 % nulos, se descartarán), información pre-universitaria (35-66 % nulos, uso limitado), indicadores de riesgo (nulos por diseño representan ausencia de riesgo). Se detectaron valores atípicos: edad máxima de 1935 años y créditos máximos de un millón, requiriendo limpieza.

Variables críticas identificadas en cuatro dimensiones: Capacidad académica (percentil PGA, PGA con media 3.97, porcentaje créditos aprobados 89.4 %, promedio semestral); Experiencia y progreso (semestre estimado, grupo de semestres, total semestres matriculados media 4.84, créditos acumulados); Indicadores de riesgo (créditos reprobados/retirados, tres o más materias retiradas 5 %, estado académico, materias por tercera vez); Contexto del estudiante (programa y nivel, tipo de matrícula, límite de créditos, estrato socioeconómico).

Del Historial de Materias emergen hallazgos relevantes: media de créditos por materia 2.77 (mediana 3.0); posibilidad de analizar redes de co-inscripción y secuencias curriculares exitosas vs. problemáticas. Del dataset de Riesgos: 15 % presenta tasa de aprobación ¡50 % (riesgo extremo); 52.55 % no cumple requisito de lectura/inglés al sexto semestre. El percentil de PGA por programa emerge como variable estrella al normalizar contextualment el rendimiento.

La integración usará código de estudiante como llave primaria, partiendo de información actual y enriqueciéndola progresivamente con historial de rendimiento, percentiles, riesgos, estados académicos, información financiera, historiales de materias y horarios para validar factibilidad práctica.

Para ahondar más en el entendimiento de los datos es necesario revisar el documento anexo y los cuadernos de Jupyter que se encuentran en el repositorio.

7 Conclusiones y Próximos Pasos

El análisis exploratorio confirma viabilidad técnica y pertinencia estratégica. Los datos demuestran patrones diferenciables y recurrentes que justifican clustering jerárquico. El percentil de PGA por programa es la variable estrella, complementada con indicadores de riesgo y métricas de experiencia.

El preprocesamiento requerirá: descartar variables de segundo programa (¿80 % nulos); limpieza de valores atípicos en edad (¿100 años) y créditos (¿1000); tratamiento de nulos en riesgos como ausencia de riesgo (codificar como cero); segmentación obligatoria por nivel y programa académico antes de clustering.

Las acciones próximas se estructuran en cuatro fases:

Fase 1 - Preparación de datos: Limpieza de atípicos y tratamiento de nulos; integración completa de ocho datasets; ingeniería de características (score compuesto de riesgo, variables de tendencia temporal, ratios informativos).

Fase 2 - Modelamiento: Segmentación primaria por programa y nivel; reducción de dimensionalidad (PCA); implementación de clustering multinivel adaptativo; validación exhaustiva con métricas definidas (comparación multinivel vs. un nivel, coeficiente de silueta, AUC-ROC, AUC-PR, Brier score).

Fase 3 - Desarrollo tecnológico: API robusta para clasificación en tiempo real; interfaz web intuitiva para consejeros con visualización de perfiles y exploración de escenarios; integración con sistemas institucionales.

Fase 4 - Validación y despliegue: Prueba piloto con grupo controlado de consejeros durante un semestre; medición de KPIs (tiempo de consejería, satisfacción, tasas de aprobación); recolección de retroalimentación; despliegue gradual institucional.

Este proyecto transforma el acompañamiento académico desde un enfoque empírico hacia un modelo híbrido que combina juicio humano con evidencia empírica sólida, potenciando la labor del consejero y beneficiando a estudiantes con orientación más personalizada y efectiva.

Referencias

[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.

[Gutiérrez et al., 2020] Gutiérrez, F., Ochoa, X., Chiluiza, K., and De Laet, T. (2020). Lada: A learning analytics dashboard for academic advising. *Computers in Human Behavior*, 107:105826.

[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/.