Entendimiento de los Datos

Análisis Exploratorio y Calidad de los Datos

Índice

1. Introducción

Este documento presenta el análisis exploratorio de datos (EDA) y evaluación de calidad para las ocho fuentes de datos del proyecto de clusterización jerárquica para recomendaciones académicas. Se aplicaron técnicas univariadas, multivariadas, gráficas y no gráficas para comprender la estructura, distribución y relaciones de las variables en cada dataset.

El objetivo del proyecto es desarrollar un sistema de recomendaciones de materias y créditos a inscribir por semestre, basado en clusterización jerárquica de estudiantes con perfiles académicos similares.

2. Fuentes de Datos Analizadas

2.1. Historial de Estados Académicos de Estudiantes

2.1.1. Descripción General

El dataset contiene **616,085 registros** y **21 variables**: 6 numéricas (1 int64, 5 float64), 13 categóricas (object) y 2 booleanas. Tamaño en memoria: 576.66 MB.

2.1.2. Valores Nulos

Cuadro 1: Variables con valores nulos - Historial Estados Académicos

Variable	**Nulos**	**%**
PROGRAMA_2	499,274	81.04
INCOMPLETOS_TOTALES_ACUMULADOS	40,856	6.63
SUSPENSIONES_ACUMULADAS	39,352	6.39
PRUEBAS_ACADEMICAS_ACUMULADAS	39,352	6.39
PERIODO_ESTADO_ACADEMICO_INICIAL	10,303	1.67
ESTADO_ACADEMICO_FINAL	158	0.03

Decisión: Variables relacionadas con PROGRAMA_2 serán descartadas (¿81 % nulos). Variables con 6-7 % nulos pueden imputarse o eliminarse según el objetivo del modelo.

2.1.3. Análisis de Variables Clave

PERIODO: Distribución aproximadamente uniforme. Los periodos intersemestrales tienen la mitad de registros que los semestrales debido a menor matrícula. Útil para segmentación temporal en el análisis de clusterización.

CODIGO_ESTUDIANTE: Identificador único, puede usarse como llave primaria para joins con otros datasets. Fundamental para asociar información del estudiante con su historial de materias y recomendaciones personalizadas.

SEMESTRE_SEGUN_CREDITOS: La categoría "0" (recién ingresados) domina la distribución. Variable crítica para recomendaciones académicas: estudiantes en semestres iniciales requieren diferentes patrones de carga crediticia que estudiantes avanzados.

MATRICULADO_PERIODO_ACTUAL: Fuerte desbalance (98 % matriculados vs 1.5 % no matriculados). No usar directamente en modelos. Analizar no matriculados por separado.

CODIGO_PROGRAMA_1 / PROGRAMA_1: Top 3: Ingeniería Industrial (12.6 %), Administración (9.1 %), Derecho (8.9 %). El programa académico es esencial para la clusterización: diferentes carreras tienen estructuras curriculares distintas que afectan las recomendaciones de materias.

NIVEL_PROGRAMA_1: Pregrado 71.4 %, Magíster 18.8 %, Especialización 5.4 %. El nivel del programa impacta directamente el número y tipo de créditos que se recomienda inscribir por semestre.

Variables PROGRAMA_2: 84.9 % sin segundo programa. Todas las variables relacionadas serán descartadas.

DESCRIPCION_ESTADO_RETIRO: 99.5 % "Habilitado para inscribirse". Pocos casos de retiro limitan su utilidad. Complementar con indicadores de desempeño académico.

REINTEGRO_REINGRESO: 95 % N/A. Totalmente desbalanceada, no será utilizada.

ESTADO_ACADEMICO_INICIAL / FINAL: Estado Normal domina (79 % inicial, 95 % final). Variable relevante para recomendaciones: estudiantes en prueba académica o suspensión requieren cargas crediticias más conservadoras. Categorías minoritarias ofrecen oportunidades para analizar paths que conducen a estados críticos.

ENTRA_A_SUSPENSION / ENTRA_A_PRUEBA_ACADEMICA: Solo 1.1 % y 1.3 % respectivamente. A pesar de baja frecuencia, son indicadores críticos de riesgo académico que deben considerarse al recomendar número de créditos a inscribir.

Variables Acumuladas: Las variables de acumulados (suspensiones, pruebas académicas e incompletos totales) presentan fuerte sesgo izquierdo (¿90 % con valor 0). Son señales importantes de trayectoria académica histórica.

2.1.4. Correlaciones

Las variables de periodo de estado académico inicial y final tienen correlación Spearman = 1.0 (esperado por naturaleza temporal). No se identificaron otras correlaciones fuertes entre variables numéricas.

2.1.5. Conclusiones

Variables a descartar: Todas relacionadas con PROGRAMA_2 (¿81 % nulos) y REINTEGRO_REINGRESO (95 % N/A).

Variables clave para clusterización: PROGRAMA_1, SEMESTRE_SEGUN_CREDITOS, ESTADO_ACADEMICO (inicial/final), y las variables acumuladas de riesgo académico son fundamentales para generar clusters significativos de estudiantes con perfiles similares.

Desbalance de clases: La mayoría de variables categóricas presentan sesgos significativos. No pueden usarse directamente en modelos sin tratamiento previo. Definir grupos

focales específicos para análisis.

Estrategia recomendada: (1) Segmentar por nivel de programa y potencialmente por carrera; (2) Analizar trayectorias de estudiantes en estados críticos para identificar patrones de riesgo; (3) Aplicar técnicas de balanceo de clases; (4) Imputar o eliminar valores faltantes según objetivos del proyecto de recomendación.

2.2. Historial de Materias de Estudiantes

2.2.1. Descripción General

El dataset contiene **4,931,740 registros** y **15 variables**: 2 numéricas (1 int64, 1 float64) y 13 categóricas (object). Tamaño en memoria: 3,767.35 MB. Este es el dataset más grande del proyecto y representa el historial completo de inscripciones a materias.

2.2.2. Valores Nulos

Cuadro 2: Variables con valores nulos - Historial Materias

Variable	Nulos	%
ATRIBUTOS	4,041,931	81.96
CALIFICACION_PARCIAL	461,858	9.37
CODIGO_MODO_DE_CALIFICACION	337,762	6.85
CODIGO_CAMPUS	335,994	6.81
SECCION	224,606	4.55
CRN	123,836	2.51
ESTATUS	100,770	2.04

Decisión: ATRIBUTOS tiene ¿81 % nulos, puede descartarse. Las demás variables con nulos moderados pueden manejarse con imputación o análisis de casos completos.

2.2.3. Análisis de Variables Clave

PERIODO: Distribución de periodos académicos desde 200010 hasta 999999. Media en 201496, cubriendo múltiples años de historial académico. Esencial para análisis temporal de patrones de inscripción.

CODIGO_ESTUDIANTE: Identificador único que vincula con otros datasets. Variable clave para integración de datos y generación de recomendaciones personalizadas por estudiante.

NIVEL_ACADEMICO_ESTUDIANTE: Pregrado domina la distribución. Importante para segmentar patrones de inscripción según nivel académico.

CRN: Código de referencia del curso. Alta cardinalidad indicando diversidad de cursos ofertados. Permite identificar materias específicas y sus patrones de co-inscripción.

CODIGO_CURSO: Similar a CRN, identifica el curso académico. Variable fundamental para el sistema de recomendación: permite identificar qué materias se toman juntas frecuentemente.

NUMERO_CREDITOS: Media de 2.77 créditos, mediana de 3.0. La mayoría de materias son de 2-3 créditos. Variable crítica para recomendaciones: permite calcular carga crediticia total al recomendar combinaciones de materias.

SECCION: Identifica secciones específicas de cursos. Útil para análisis de capacidad y disponibilidad.

CODIGO_MODO_DE_CALIFICACION: Indica el sistema de calificación (Numérico, Aprobado/Reprobado, etc.). Puede influir en decisiones de inscripción.

CALIFICACION_PARCIAL / FINAL: Registran el desempeño del estudiante. Variable fundamental para la clusterización: permite identificar estudiantes con patrones de rendimiento similares y recomendar cargas apropiadas.

ESTATUS: Estado del registro (Historia, En Progreso). Distingue entre materias completadas y actuales.

ATRIBUTOS: Alta proporción de nulos (81.96 %). Puede descartarse o usar solo para casos específicos.

CODIGO_CAMPUS: Identifica el campus donde se dicta el curso. Puede ser relevante para recomendaciones considerando ubicación.

RESULTADO: Resultado final de la materia (Aprobado, Reprobado, Retirado). Variable clave para identificar patrones de éxito: ayuda a recomendar materias que históricamente tienen mejor tasa de aprobación cuando se combinan.

ES_HISTORICO_ACTUAL: Indica si es registro histórico o actual. Útil para filtrar datos recientes vs antiguos.

2.2.4. Correlaciones

Con solo 2 variables numéricas (PERIODO y NUMERO_CREDITOS), las correlaciones son limitadas. La correlación Spearman entre estas es baja, indicando que el número de créditos de las materias no ha cambiado significativamente con el tiempo.

2.2.5. Conclusiones

Variables a descartar: ATRIBUTOS por alto porcentaje de nulos (¿81 %).

Variables críticas para recomendación: CODIGO_CURSO, NUMERO_CREDITOS, RESULTADO, y CALIFICACION_FINAL son fundamentales. Permiten identificar: (1) qué materias se toman juntas frecuentemente, (2) cuántos créditos suma cada combinación, (3) qué combinaciones tienen mejores tasas de éxito.

Oportunidades de análisis: Este dataset permite construir redes de co-inscripción de materias, identificar secuencias curriculares comunes, y detectar combinaciones de materias que resultan en sobrecarga académica o bajo rendimiento.

Estrategia recomendada: (1) Crear grafos de co-ocurrencia de materias por programa; (2) Analizar tasas de éxito por combinaciones de materias; (3) Identificar patrones de sobrecarga (alto número de créditos + bajo rendimiento); (4) Usar para validar recomendaciones del modelo de clusterización.

2.3. Historial de Rendimiento Académico de Estudiantes

2.3.1. Descripción General

El dataset contiene **611,654 registros** y **22 variables**: 19 numéricas (1 int64, 3 float32, 15 float64) y 3 categóricas (object). Tamaño en memoria: 198.32 MB. Este dataset captura métricas de rendimiento académico por periodo.

2.3.2. Valores Nulos

Cuadro 3: Variables con valores nulos - Rendimiento Académico

Variable	Nulos	%
TOTAL_SEMESTRES_MATRICULADOS	2,869	0.47
PROMEDIO_SEMESTRAL	91	0.01
CREDITOS_APROBADOS	5	0.00
CREDITOS_REPROBADOS	5	0.00
CREDITOS_INCOMPLETOS	5	0.00

Decisión: Valores nulos muy bajos (¡1 %). Pueden imputarse fácilmente o eliminarse sin pérdida significativa de información.

2.3.3. Análisis de Variables Clave

PERIODO: Media en 201790, cubriendo múltiples periodos académicos. Permite análisis temporal de evolución del rendimiento.

CODIGO_ESTUDIANTE: Identificador único para joins con otros datasets.

PGA (Promedio General Acumulado): Media de 3.97, mediana de 4.03. Variable crítica para clusterización: estudiantes con PGA similar pueden tener capacidades similares para manejar cargas crediticias. Distribución ligeramente asimétrica hacia valores altos.

PROMEDIO_SEMESTRAL: Media de 3.67, más variable que PGA. Refleja desempeño en periodo específico. Útil para identificar estudiantes que mejoran o empeoran con el tiempo.

CREDITOS_PGA: Media de 68.92 créditos. Indica progreso en la carrera. Fundamental para recomendaciones: estudiantes con más créditos acumulados suelen poder manejar cargas mayores.

DESCRIPCION_NIVEL_PROGRAMA: Mayormente Pregrado. Importante para segmentar análisis por nivel académico.

CREDITOS_APROBADOS / REPROBADOS: Media de 72.33 aprobados vs 4.39 reprobados. La mayoría de estudiantes tienen buen rendimiento. Los créditos reprobados son indicador de riesgo para recomendaciones futuras.

CREDITOS_INCOMPLETOS / RETIRADOS: Medias bajas (0.15 y 5.41 respectivamente). Los retiros son más comunes que incompletos. Variable relevante: alta tasa de retiros puede indicar sobrecarga o materias mal seleccionadas.

CREDITOS_PENDIENTES / HOMOLOGADOS: Valores muy bajos en promedio. Los pendientes pueden afectar recomendaciones de prerrequisitos.

PORCENTAJE_CREDITOS_APROBADOS: Media de 89.4 %, mediana de 95.3 %. Variable clave para clusterización: tasa de éxito histórica predice capacidad futura. Estudiantes con ¡50 % requieren recomendaciones más conservadoras.

TOTAL_SEMESTRES_MATRICULADOS: Media de 4.84 semestres. Indica experiencia académica. Estudiantes con más semestres suelen tener mejor comprensión de su capacidad.

MATERIA_POR_TERCERA_VEZ: Categórica SI/NO. Indicador de dificultades académicas específicas. Relevante para evitar sobrecargas en recomendaciones.

Variables por Semestre (CREDITOS_APROBADOS_SEMESTRE, etc.): Reflejan desempeño en el periodo actual. Media de 12.89 créditos aprobados por semestre. Útil para calibrar recomendaciones según desempeño reciente vs histórico.

PORCENTAJE_CREDITOS_APROBADOS_SEMESTRE: Media de 88.5 %. Comparar con porcentaje acumulado permite identificar tendencias de mejora o deterioro.

2.3.4. Correlaciones

Correlaciones fuertes identificadas:

- PGA y PROMEDIO_SEMESTRAL: correlación alta (esperada)
- CREDITOS_PGA y CREDITOS_APROBADOS: correlación muy fuerte (casi perfecta)
- Variables acumuladas vs variables semestrales: correlaciones moderadas
- TOTAL_SEMESTRES_MATRICULADOS con créditos acumulados: correlación fuerte

2.3.5. Conclusiones

Variables mínimas de nulos: Excelente calidad de datos, prácticamente sin valores faltantes.

Variables críticas para clusterización: PGA, PORCENTAJE CREDITOS APROBADOS, CREDITOS_REPROBADOS, CREDITOS_RETIRADOS y TOTAL SEMESTRES MATRICULADOS permiten identificar perfiles de estudiantes: (1) alto rendimiento/mucha experiencia, (2) rendimiento medio/progreso normal, (3) bajo rendimiento/en riesgo.

Indicadores de riesgo: CREDITOS_REPROBADOS, CREDITOS_RETIRADOS, MATERIA_POR_TERCERA_VEZ y PORCENTAJE CREDITOS APROBADOS SEMESTRE bajo son señales para recomendar cargas reducidas.

Estrategia recomendada: (1) Usar PGA y porcentajes de aprobación para clasificar capacidad académica; (2) Considerar CREDITOS_RETIRADOS como indicador de sobrecarga previa; (3) Comparar desempeño semestral vs acumulado para detectar tendencias;

(4) Ajustar recomendaciones según experiencia (TOTAL SEMESTRES MATRICULADOS).

2.4. Horarios de Cursos

2.4.1. Descripción General

El dataset contiene **444,834 registros** y **16 variables**: 8 numéricas (1 int64, 7 int8) y 8 categóricas (object). Tamaño en memoria: 186.67 MB. Representa la programación horaria de cursos ofertados.

2.4.2. Valores Nulos

Cuadro 4: Variables con valores nulos - Horarios

Variable	Nulos	%
CODIGO_SALON	118,713	26.69
HORA_FIN_FRANJA	118,357	26.61
HORA_INICIO_FRANJA	118,357	26.61
FECHA_INICIO_FRANJA	76,457	17.19
FECHA_FIN_FRANJA	76,457	17.19

Decisión: Aproximadamente 26 % de registros carecen de información de horario y salón. Esto puede corresponder a cursos virtuales, dirigidos, o registros incompletos. Importante considerar al generar recomendaciones basadas en disponibilidad horaria.

2.4.3. Análisis de Variables Clave

PERIODO: Media en 201551, cubriendo desde 199019 hasta 202612. Distribución amplia de periodos académicos.

CRN: Código de referencia del curso. Permite vincular con dataset de historial de materias para análisis integrado.

CODIGO_CURSO: Identifica el curso específico. Variable clave para asociar horarios con materias en el sistema de recomendación.

SECCION: Número de sección del curso. Múltiples secciones del mismo curso pueden tener horarios diferentes, importante para recomendaciones considerando conflictos horarios.

FECHA_INICIO_FRANJA / FECHA_FIN_FRANJA: Definen duración del curso. 17 % nulos. Útil para identificar cursos de ciclo completo vs parcial.

HORA_INICIO_FRANJA / HORA_FIN_FRANJA: Horarios de clase. 26.61 % nulos. Variable crítica para detectar conflictos horarios al recomendar combinaciones de materias. Sistema debe evitar recomendar materias que se traslapen temporalmente.

CODIGO_SALON: Identifica el salón físico. 26.69 % nulos probablemente por cursos virtuales o sin asignación aún. Puede usarse para considerar distancias entre clases consecutivas.

Variables de días (LUNES a DOMINGO): Variables binarias (0/1) indicando días de clase. Distribución: Jueves (22.3 %), Martes (21.1 %), Viernes (21.1 %), Miércoles (20.6 %), Lunes (17.9 %), Sábado (7.8 %), Domingo (0.5 %). Para clusterización jerárquica y recomendaciones: importante considerar patrones de distribución semanal para evitar sobrecargas en días específicos.

2.4.4. Correlaciones

Las variables de días de semana no muestran correlaciones fuertes entre sí, sugiriendo que la programación de cursos distribuye las clases de manera relativamente independiente entre días. PERIODO no correlaciona fuertemente con ninguna variable horaria.

2.4.5. Conclusiones

Variables con nulos significativos: 26 % de registros sin información completa de horario/salón. Requiere tratamiento especial: pueden ser cursos virtuales o datos incompletos.

Utilidad para recomendaciones: Este dataset es complementario al sistema de recomendación principal. Permite: (1) validar que materias recomendadas no tengan conflictos horarios, (2) considerar distribución de carga durante la semana, (3) optimizar trayectos físicos entre clases.

Patrones temporales: Jueves es el día más utilizado (22.3 %), Domingo casi no se usa (0.5 %). Clases de sábado son minoritarias (7.8 %), probablemente programas especiales o posgrados.

Estrategia recomendada: (1) Usar como filtro post-clusterización para validar viabilidad de combinaciones recomendadas; (2) Identificar patrones de carga horaria por carrera; (3) Considerar cursos sin horario definido como opciones flexibles; (4) Al recomendar número de créditos, validar que existan horarios compatibles disponibles para las materias sugeridas.

2.5. Información Actual de Estudiantes

2.5.1. Descripción General

El dataset contiene **222,407 registros** y **30 variables**: 4 numéricas (1 int64, 3 float64) y 26 categóricas (object). Tamaño en memoria: 315.06 MB. Representa el estado actual y características demográficas de estudiantes.

2.5.2. Valores Nulos

Cuadro 5: Variables con valores nulos (top 10) - Información Actual

Variable	**Nulos**	**%**
PROGRAMA_2	209,639	94.26
ICFES_CRITERIO_4	157,078	70.63
ICFES_CRITERIO_5	148,246	66.66
ICFES_CRITERIO_2	148,244	66.65
ICFES_CRITERIO_1	148,243	66.65
PUNTAJE_ICFES	148,239	66.65
ULTIMO_PERIODO_MATRICULADO	97,705	43.93
CIUDAD_COLEGIO	78,321	35.22
DEPARTAMENTO_COLEGIO	78,321	35.22
NOMBRE_COLEGIO	78,130	35.13

Decisión: PROGRAMA_2 (94 % nulos) debe descartarse. Variables ICFES (66 % nulos) y datos de colegio (35 % nulos) tienen limitaciones pero pueden usarse para análisis de subpoblaciones específicas.

2.5.3. Análisis de Variables Clave

CODIGO_ESTUDIANTE: Identificador único, llave primaria para integración con otros datasets.

PERIODO: Media en 201523, cubriendo múltiples periodos. Identifica momento de captura de datos.

EDAD: Media de 33.88 años, con valores atípicos (máximo 1935, probablemente error). Mediana de 31 años más confiable. Para recomendaciones: estudiantes de mayor edad pueden tener restricciones de tiempo por compromisos laborales/familiares.

GENERO: Variable categórica con códigos anonimizados (A, B). Puede usarse para análisis de equidad pero no es determinante para recomendaciones académicas.

CODIGO_PROGRAMA_1 / PROGRAMA_1: Identifica el programa académico principal. Variable crítica para clusterización: diferentes programas tienen estructuras curriculares y dificultades distintas. Recomendaciones deben ser específicas por programa.

NIVEL_PROGRAMA_1: Distribución por nivel académico. Fundamental: pregrado, magíster y especialización tienen dinámicas de carga crediticia completamente diferentes.

PROGRAMA_2: 94.26 % nulos. Solo 5.7 % tiene doble programa. Descartar de análisis principal.

ESTADO_ACADEMICO: Estado actual del estudiante (Normal, Prueba, Suspensión, etc.). Variable crucial para recomendaciones: estudiantes en prueba académica necesitan cargas reducidas y materias de alta probabilidad de aprobación.

MATRICULADO_PERIODO_ACTUAL: Indica si está actualmente matriculado. Útil para segmentar población activa vs inactiva.

ULTIMO_PERIODO_MATRICULADO: 43.93 % nulos. Para no nulos, indica continuidad académica. Interrupciones largas pueden afectar recomendaciones.

PERIODO_CATALOGO: Periodo de ingreso del estudiante. Permite calcular tiempo en el programa. Estudiantes con muchos periodos pueden requerir planes de finalización acelerada.

PENALIZADO_EXTRACREDITACION / PENSUM_REFORMADO: Variables binarias/categóricas sobre restricciones académicas. La penalización por extracreditación limita créditos máximos recomendables.

PROMEDIO_MOVIL: Media de 1.46, pero distribución sesgada. Métrica de rendimiento reciente. Complementa PGA para recomendaciones.

CREDITOS_MAXIMOS: Media de 11,222 pero con valor máximo de 1,000,000 (claramente atípico). Mediana de 20 créditos más realista. Variable fundamental: define límite superior de recomendación de créditos.

Variables ICFES: 66 % nulos limita utilidad. Para casos completos, pueden predecir rendimiento inicial pero pierden relevancia con trayectoria académica establecida.

ESTRATO_SOCIOECONOMICO: Variable completa (0 % nulos). Puede correlacionar con disponibilidad de tiempo para estudiar (estudiantes de estratos bajos suelen trabajar). Considerar para recomendaciones contextualizadas.

2.5.4. Correlaciones

EDAD y PERIODO muestran correlación moderada (esperada). CREDITOS_MAXIMOS tiene valores atípicos que distorsionan correlaciones. PROMEDIO_MOVIL debería correlacionar con rendimiento académico al cruzar con otros datasets.

2.5.5. Conclusiones

Variables a descartar: PROGRAMA_2 (94 % nulos), posiblemente variables ICFES por alta proporción de nulos (66 %).

Variables críticas para clusterización: PROGRAMA_1, NIVEL_PROGRAMA_1, ESTADO_ACADEMICO, CREDITOS_MAXIMOS (limpiando atípicos), ESTRATO SOCIOECONOMICO. Estas permiten segmentar estudiantes por perfil académico y contexto socioeconómico.

Calidad de datos: Presencia de valores atípicos extremos en EDAD y CREDITOS_MAXIMOS requiere limpieza previa.

Estrategia recomendada: (1) Limpiar valores atípicos en EDAD y CREDITOS MAXIMOS; (2) Usar PROGRAMA y NIVEL como segmentadores primarios; (3) Aplicar ESTADO ACADEMICO como filtro de restricciones; (4) Considerar ESTRATO para ajustes contextuales; (5) Descartar PROGRAMA_2 y opcionalmente variables ICFES; (6) CREDITOS_MAXIMOS define límite duro para recomendaciones por estudiante.

2.6. Información Financiera de Estudiantes

2.6.1. Descripción General

El dataset contiene **793,198 registros** y **12 variables**: 3 numéricas (1 int64, 1 float64, 1 uint32) y 9 categóricas (object). Tamaño en memoria: 475.94 MB. Representa información sobre financiamiento y matrícula de estudiantes.

2.6.2. Valores Nulos

Cuadro 6: Variables con valores nulos - Información Financiera

Variable	Nulos	%
TIPO_MATRICULA_SEMESTRE	6,615	0.83
DESCRIPCION_NIVEL_PROGRAMA_1	1,659	0.21
TIPO_FINANCIAMIENTO_PRINCIPAL	1,659	0.21

Decisión: Valores nulos muy bajos (¡1 %). Excelente calidad de datos, pueden imputarse fácilmente sin pérdida significativa.

2.6.3. Análisis de Variables Clave

PERIODO: Media en 201903, cubriendo desde 201210 hasta 202610. Refleja información financiera por periodo académico.

CODIGO_ESTUDIANTE: Identificador único para integración con otros datasets.

TIPO_FINANCIAMIENTO_ESPECIFICO: Detalla la fuente específica de financiamiento. Alta granularidad. Útil para análisis de ayudas financieras disponibles.

FECHA_AFECTACION: Fecha del registro financiero. Permite análisis temporal de pagos y financiamientos.

TIPO_SECCION: Categoriza el tipo de financiamiento. Relacionado con TIPO_FINANCIAMIENTO ESPECIFICO.

OP_PRINCIPAL_PARCIAL: Código de operación financiera. Alta cardinalidad, probablemente identificador de transacción.

PORCENTAJE_APORTE_ITEM: Media de 78.25 %, mediana de 100 %. Indica qué porcentaje de un ítem es cubierto por la fuente de financiamiento. Variable relevante: estudiantes con financiamiento completo (100 %) pueden tener menos restricciones de tiempo vs quienes deben trabajar.

RANKING: Media de 1.25, mediana de 1. Parece ser un ranking de importancia o prioridad del financiamiento (cuando hay múltiples fuentes). Bajo valor indica financiamiento principal.

DESCRIPCION_NIVEL_PROGRAMA_1: Nivel académico (PREGRADO mayormente). Variable de segmentación.

TIPO_FINANCIAMIENTO_PRINCIPAL_SEMESTRE: Identifica la fuente principal de financiamiento para el semestre. Para clusterización: puede correlacionar con

disponibilidad de tiempo. Estudiantes con "Recursos Propios" y estratos bajos probablemente trabajan, limitando tiempo de estudio.

TIPO_FINANCIAMIENTO_PRINCIPAL: Similar a la variable anterior pero agregada. Importante para identificar estudiantes con becas vs recursos propios vs créditos.

TIPO_MATRICULA_SEMESTRE: Categoriza si la matrícula es completa, media, de vacaciones, etc. Variable crítica para recomendaciones: estudiantes con matrícula parcial pueden inscribir menos créditos. "COMPLETA_PREGRADO" es la categoría más común.

2.6.4. Correlaciones

RANKING y PORCENTAJE_APORTE_ITEM muestran correlación negativa moderada: rankings bajos (principales) tienden a tener porcentajes altos. PERIODO no muestra correlaciones fuertes con otras variables numéricas.

2.6.5. Conclusiones

Calidad de datos: Excelente, menos de 1 % de nulos en todas las variables.

Relevancia para clusterización: Aunque no es el foco principal del proyecto de recomendación académica, este dataset aporta contexto importante. TIPO FINANCIAMIENTO y TIPO_MATRICULA pueden explicar por qué algunos estudiantes inscriben menos créditos de los recomendados.

Variables clave: TIPO_MATRICULA_SEMESTRE (define capacidad de inscripción), TIPO_FINANCIAMIENTO_PRINCIPAL (puede correlacionar con disponibilidad de tiempo), PORCENTAJE_APORTE_ITEM (estudiantes con bajo porcentaje pueden tener restricciones económicas).

Oportunidades de análisis: Cruzar con rendimiento académico para verificar si tipo de financiamiento afecta desempeño. Estudiantes con becas de alto rendimiento podrían manejar cargas mayores.

Estrategia recomendada: (1) Usar TIPO_MATRICULA_SEMESTRE como restricción: matrícula parcial limita créditos recomendables; (2) Considerar TIPO FINANCIAMIENTO para ajustes contextuales; (3) No usar como variable principal de clusterización, pero sí como feature complementaria; (4) Combinar con ESTRATO_SOCIOECONOMICO del dataset de información actual para mejor contexto socioeconómico.

2.7. Percentiles Académicos de Estudiantes

2.7.1. Descripción General

El dataset contiene **158,429 registros** y **11 variables**: 5 numéricas (1 int64, 1 float32, 3 float64) y 6 categóricas (object). Tamaño en memoria: 71.20 MB. Proporciona información sobre el rendimiento relativo de estudiantes dentro de su programa académico.

2.7.2. Valores Nulos

Cuadro 7: Variables con valores nulos - Percentiles Académicos

Variable	Nulos	%
SEMESTRE_AVANCE	102,062	64.42
NUMERO_PROGRAMA	102,062	64.42

Decisión: 64.42 % de nulos en SEMESTRE_AVANCE y NUMERO_PROGRAMA. Esto puede deberse a estudiantes de posgrado, educación continua, o extensión donde estas métricas no aplican. Usar con precaución, limitado a subpoblaciones específicas.

2.7.3. Análisis de Variables Clave

PERIODO: Media en 201670, cubriendo desde 200020 hasta 202520. Amplio rango temporal.

CODIGO_ESTUDIANTE: Identificador único para joins con otros datasets.

NIVEL_ACADEMICO: Categoriza nivel del programa. Variable de segmentación importante.

CODIGO_PROGRAMA / PROGRAMA: Identifica el programa académico específico. Fundamental para contexto: percentiles solo son significativos dentro del mismo programa.

PGA: Promedio General Acumulado. Media de 4.02, mediana de 4.10. Distribución ligeramente sesgada hacia valores altos. Variable crítica para clusterización: estudiantes con PGA similar tienen capacidades académicas comparables.

PERCENTIL_PGA_PROGRAMA: Media de 49.8 (distribución uniforme esperada). Variable única y valiosa: posiciona al estudiante respecto a sus pares del mismo programa. Útil para identificar estudiantes de alto rendimiento (percentil ¿75) que pueden manejar cargas crediticias mayores, vs estudiantes en riesgo (percentil ¡25) que requieren cargas reducidas. Esta variable normaliza el PGA por programa, permitiendo comparaciones justas entre carreras con diferentes niveles de dificultad.

TIPO_ESTUDIANTE: Clasifica estudiantes (Activo, Graduado, Retirado, etc.). Filtrar por "Activo" para recomendaciones.

SEMESTRE_AVANCE: Media de 3.80 semestres (para no nulos). Indica progreso en la carrera. Estudiantes avanzados (¿6 semestres) suelen tener mejor comprensión de su capacidad vs principiantes. Útil para ajustar recomendaciones según experiencia.

GRUPO_SEMESTRES: Categoriza SEMESTRE_AVANCE en rangos (0-1, 2-3, 4-5, 6, 7+). Variable discreta útil para clusterización jerárquica: permite agrupar estudiantes por etapa de carrera sin depender de valor numérico exacto.

NUMERO_PROGRAMA: Para casos no nulos, media de 1.13. Indica si tiene doble programa. Baja variabilidad (mayoría tiene 1 programa).

2.7.4. Correlaciones

PERCENTIL_PGA_PROGRAMA correlaciona fuertemente con PGA (esperado por construcción). SEMESTRE_AVANCE puede correlacionar moderadamente con PGA (estudiantes que permanecen tienden a mejorar). No se identifican correlaciones inesperadas.

2.7.5. Conclusiones

Variable estrella: PERCENTIL_PGA_PROGRAMA es única en este dataset. Proporciona normalización por programa, permitiendo comparaciones justas. Altamente recomendada para clusterización: separa estudiantes de alto/medio/bajo rendimiento dentro de contexto específico de su carrera.

Variables con nulos significativos: SEMESTRE_AVANCE y NUMERO_PROGRAMA (64 % nulos) limitan su uso. Para pregrado activo, estos datos suelen estar disponibles. Para posgrados/extensión, frecuentemente ausentes.

Segmentación por experiencia: GRUPO_SEMESTRES permite estratificar recomendaciones. Estudiantes en rango 0-1 (principiantes) necesitan orientación diferente que 7+ (veteranos).

Integración con otros datasets: Este dataset es complementario. PERCENTIL_PGA_PROGRAMA enriquece el PGA absoluto de otros datasets con contexto relativo.

Estrategia recomendada: (1) Usar PERCENTIL_PGA_PROGRAMA como variable principal de clusterización para capacidad académica; (2) Combinar con GRUPO_SEMESTRES para segmentar por experiencia; (3) Filtrar TIPO_ESTUDIANTE = “Activo” para población objetivo; (4) Para estudiantes con percentil ¡25, recomendar cargas reducidas y materias con alta tasa de aprobación; (5) Para percentil ¿75, permitir cargas mayores y combinaciones más desafiantes; (6) Considerar que 64 % de casos no tienen SEMESTRE_AVANCE - usar PGA y PERCENTIL como alternativa.

2.8. Riesgos Históricos de Estudiantes de Pregrado

2.8.1. Descripción General

El dataset contiene **369,370 registros** y **12 variables**: 9 numéricas (2 int32, 4 int32 binarias, 3 float64) y 3 categóricas (object). Tamaño en memoria: 62.76 MB. Este dataset es específico para estudiantes de pregrado y contiene indicadores de riesgo académico calculados.

2.8.2. Valores Nulos

Cuadro 8: Variables con valores nulos - Riesgos Históricos

Variable	**Nulos**	**%**
pocos_creditos_inscritos_para_matricula_pagada	358,549	97.07
tres_o_mas_materias_retiradas	352,164	95.34
materias_bloqueantes	343,034	92.87
semestres_perdidos_con_estado_normal	332,683	90.07
tres_o_mas_materias_parcialmente_perdidas	331,711	89.80

Decisión: Altos porcentajes de nulos (89-97 %) indican que estas variables son indicadores de riesgo que solo se activan cuando el problema existe. Valores nulos equivalen a "sin riesgo detectado" en ese indicador. Tratamiento especial: nulos = 0 o "sin riesgo".

2.8.3. Análisis de Variables Clave

CODIGO_ESTUDIANTE: Identificador único. Llave para integración con otros datasets.

PERIODO: Media en 201824, cubriendo periodos recientes. Indica cuándo se calculó el indicador de riesgo.

tres_o_mas_materias_parcialmente_perdidas: 89.8 % nulos (sin riesgo). Para casos no nulos, media de 3.62 materias. Variable crítica de riesgo: perder parcialmente 3+ materias indica dificultades serias. Para clusterización jerárquica: estudiantes con este indicador activo requieren cargas muy reducidas y materias menos demandantes.

tres_o_mas_materias_retiradas: 95.34 % nulos. Media de 4.26 retiros para casos no nulos. Indicador fuerte de sobrecarga: retirar 3+ materias sugiere inscripción excesiva o selección inadecuada. Fundamental para recomendaciones: evitar repetir patrones que llevaron a retiros masivos.

bajo_promedio_segundo_semestre: Variable binaria (0/1). Solo 1.97 % positivos. Identifica estudiantes con mal desempeño temprano. Aunque poco frecuente, es predictor de deserción. Requiere intervención con cargas reducidas y acompañamiento.

bajo_promedio_tercer_semestre_o_mas: 32.76 % positivos. Más común que bajo promedio en segundo semestre. Indica dificultades académicas sostenidas después de la adaptación inicial. Variable importante para clusterización: segmenta estudiantes en riesgo crónico.

bajo_promedio_movil: Solo 1.86 % positivos. Promedio móvil bajo indica problemas recientes. Diferente a PGA bajo: estudiante puede tener buen historial pero rendimiento reciente pobre. Útil para detectar deterioro académico.

pocos_creditos_inscritos_para_matricula_pagada: 97.07 % nulos (variable tipo diccionario en casos no nulos). Indica sub-utilización de matrícula pagada. Puede reflejar restricciones externas (trabajo, familia) o estrategia conservadora. Relevante para ajustar expectativas de carga crediticia.

materias_bloqueantes: 92.87 % nulos. Variable tipo diccionario con detalles de materias que bloquean progreso curricular. Crítica para recomendaciones: identificar y priorizar materias bloqueantes en sugerencias.

semestres_perdidos_con_estado_normal: 90.07 % nulos. Media de 2.52 semestres perdidos para casos no nulos. Perder semestres estando en estado "Normal" (no en prueba/suspensión) indica problemas crónicos no reflejados en estado académico formal. Señal de riesgo oculto.

no_lectura_ingles_sexto_semestre: 52.55 % positivos. Más de la mitad no ha cumplido este requisito para sexto semestre. Variable específica del contexto institucional. Para recomendaciones: considerar como materia pendiente importante.

porcentaje_creditos_aprobados_menor_50: 14.94 % positivos. Aproximadamente 1 de cada 7 estudiantes tiene tasa de aprobación ¡50 %. Indicador crítico de riesgo extremo. Para clusterización: este grupo requiere intervenciones especiales, cargas mínimas (6-9 créditos) y materias con alta tasa de éxito.

2.8.4. Correlaciones

Variables de riesgo (bajo_promedio_segundo_semestre, bajo_promedio_tercer_semestre_o_mas, bajo_promedio_movil, porcentaje_creditos_aprobados_menor_50) deberían correlacionar entre sí, indicando patrones de riesgo acumulativo. Estudiantes con múltiples indicadores activos están en riesgo muy alto.

2.8.5. Conclusiones

Dataset de alertas tempranas: Este dataset es fundamentalmente diferente: no describe características sino señales de riesgo. Alto porcentaje de nulos es esperado y correcto - la mayoría de estudiantes no tiene estos problemas.

Variables críticas para recomendaciones conservadoras: tres_o_mas materias retiradas, tres_o_mas_materias_parcialmente_perdidas, porcentaje_creditos_aprobados_menor_50 son banderas rojas. Presencia de cualquiera indica necesidad de cargas reducidas (máximo 12-15 créditos).

Indicadores de intervención: bajo_promedio_segundo_semestre detecta problemas tempranos. materias_bloqueantes identifica obstáculos curriculares. Ambos requieren atención especial en recomendaciones.

Tratamiento de nulos: IMPORTANTE: nulos en este dataset significan "ausencia de riesgo", no datos faltantes. Imputar como 0 o crear variable binaria "riesgo_presente/ausente".

Estrategia recomendada: (1) Crear score de riesgo sumando indicadores activos; (2) Estudiantes con score alto (3+ indicadores) → cluster de "alto riesgo" con recomendaciones muy conservadoras; (3) Priorizar materias_bloqueantes en recomendaciones para desbloquear progreso curricular; (4) Combinar con PGA y percentil: estudiante puede tener bajo PGA sin estos indicadores (dificultad general) vs bajo PGA con múltiples indicadores (riesgo de deserción); (5) Usar para validación post-clusterización: verificar que recomendaciones no repliquen patrones que activaron indicadores de riesgo.

3. Conclusiones Generales

3.1. Calidad de Datos

La calidad general de los datos es buena, con la mayoría de datasets presentando menos del 1 % de valores nulos en variables críticas. Las excepciones principales son:

- Variables relacionadas con segundo programa (PROGRAMA_2): consistentemente ¿80 % nulos en múltiples datasets - descartar.
- Variables de información pre-universitaria (ICFES, colegio): 35-66 % nulos - usar solo para análisis de subpoblaciones.

- Indicadores de riesgo académico: altos porcentajes de nulos son correctos por diseño (ausencia de riesgo), pero si se quiere usar se va a tener que pensar la forma de transformar a los datos.
- Información de horarios: 26 % sin datos completos, probablemente cursos virtuales.

Se identificaron valores atípicos en EDAD (máximo 1935) y CREDITOS_MAXIMOS (máximo 1,000,000) que requieren limpieza.

3.2. Variables Críticas para Clusterización Jerárquica

Con el propósito de recomendar materias y créditos de forma personalizada, se seleccionaron variables que representan cuatro dimensiones principales. En cuanto a la **capacidad académica**, destacan el percentil del promedio general del programa (`PERCENTIL_PGA_PROGRAMA`), el promedio general acumulado (`PGA`), el porcentaje de créditos aprobados y el promedio semestral, que refleja el rendimiento reciente. La dimensión de **experiencia y progreso** incluye el semestre estimado según créditos (`SEMESTRE_SEGUN_CREDITOS`), el grupo de semestres, el total de semestres matriculados y los créditos acumulados (`CREDITOS_PGA`). En los **indicadores de riesgo** se consideran los créditos reprobados o retirados, las variables del dataset de riesgos (como `tres_o_mas_materias_retiradas`), el estado académico (Normal, Prueba o Suspensión) y la presencia de materias cursadas por tercera vez. Finalmente, el **contexto del estudiante** contempla el programa y nivel académico (`PROGRAMA_1`, `NIVEL_PROGRAMA_1`), el tipo de matrícula del semestre, el límite institucional de créditos y el estrato socioeconómico, que aporta información sobre su entorno personal.

3.3. Estrategia de Integración de Datasets

La integración de los ocho datasets se realiza usando `CODIGO_ESTUDIANTE` como llave principal. El proceso parte del dataset de información actual de los estudiantes, que contiene su estado, programa y posibles restricciones. A partir de allí, se enriquece con el historial de rendimiento académico, incorporando datos como el promedio (PGA), los créditos cursados y el desempeño general. Luego, se añaden los percentiles académicos para representar el rendimiento relativo de cada estudiante, seguidos por los registros de riesgos históricos que actúan como indicadores de alerta temprana. Posteriormente, se incluyen los historiales de estados académicos para reflejar la trayectoria institucional y la información financiera, útil para entender el tipo de matrícula y posibles restricciones. Finalmente, se utilizan los historiales de materias para identificar patrones de coinscripción exitosos y se valida la factibilidad de las recomendaciones mediante los horarios de cursos, garantizando la disponibilidad y evitando conflictos.

3.4. Recomendaciones para Clusterización

Se recomienda realizar la segmentación inicial por `NIVEL_PROGRAMA` (pregrado o posgrado) y `PROGRAMA` antes de aplicar cualquier técnica de clusterización. Las variables sugeridas para este proceso combinan indicadores de capacidad académica (`PERCENTIL_PGA_PROGRAMA`), experiencia estudiantil (`GRUPO_SEMESTRES`) e indicadores de riesgo. En cuanto al tratamiento de datos, se deben descartar los registros asociados a `PROGRAMA_2`, asumir los valores nulos en indicadores de riesgo como ausencia de riesgo (0) y utilizar las

variables del ICFES únicamente cuando estén disponibles, sin imputación. Asimismo, es importante limpiar valores atípicos en `EDAD` (mayores a 100 años) y `CREDITOS_MAXIMOS` (mayores a 1000). Finalmente, las recomendaciones deben validarse respetando los límites de créditos por estudiante, el tipo de matrícula del semestre, el estado académico, la viabilidad horaria según el dataset de horarios y priorizando las materias bloqueantes previamente identificadas.