****

**FACULTAD DE INGENIERIA Y ARQUITECTURA**

**CLOUD COMPUTING**

**Título del Tema:** Predicción de la deserción estudiantil en la UPeU

**Docente:**   
Abel Ángel Sullon Macalupu

**Grupo 2:**

* Orihuela Orozco Luis
* Ponce Galarza Brayan

2025

# PERFIL DEL PROYECTO

## Definir el problema de negocio/social que motiva el análisis

En la Universidad Peruana Unión (UPeU), la deserción estudiantil interrumpe trayectorias académicas y genera impactos académicos, económicos y sociales. Desde la perspectiva del estudiante, abandonar los estudios implica pérdida de oportunidades (título profesional, mejores ingresos futuros) y, muchas veces, está asociado a dificultades académicas, económicas o psicosociales. Desde la institución, la deserción reduce tasas de retención y graduación oportuna, afecta la planificación de cursos y carga docente, y presiona la sostenibilidad financiera (menores matrículas, becas mal focalizadas, capacidad ociosa).

**El problema que abordamos:**

Identificar tempranamente a los estudiantes con mayor riesgo de no continuar sus estudios en el siguiente ciclo, para poder priorizar intervenciones (tutoría académica, apoyo socioeconómico, consejería) que eviten la deserción y mejoren la permanencia.

## Establecer objetivos generales y específicos

**Objetivo general**

Diseñar y documentar un análisis de datos que ayude a detectar a tiempo el riesgo de deserción en estudiantes de la UPeU, para orientar acciones de apoyo académico y socioeconómico.

**Objetivos específicos**

* **Entender la base de datos:** cargarla y ver cuántas filas/columnas tiene y de qué tipo son.
* **Revisar la calidad de los datos:** ubicar faltantes, duplicados y valores raros.
* **Explorar relaciones:** ver con tablas y gráficos qué factores se asocian más con la deserción.

## Redactar criterios de éxito medibles

* **Notebook corre sin errores**

Meta: 100% de celdas ejecutan. Evidencia: notebook corrido en la entrega.

* **Diagnóstico de calidad completo**

Meta: 100% de columnas con % de faltantes, duplicados contados y outliers revisados en ≥80% de variables numéricas. Evidencia: tablas en el notebook.

* **Definir y medir deserción**

Meta: definición operativa escrita y tasa de deserción calculada con n válido. Evidencia: bloque de resultados en el notebook.

* **Visualizaciones mínimas**

Meta: ≥2 gráficos con título, ejes y 2–3 líneas de interpretación. Evidencia: secciones de gráficos en el notebook.

# DATASET CRUDO (análisis descriptivo)

## Cargar Dataset

Usé pandas para leer el archivo Excel data/upue\_students\_data.xlsx.

Como el Excel puede tener varias hojas, el script elige la hoja con más filas (la más probable para el análisis).

También limpié los nombres de columnas (minúsculas, sin tildes ni espacios raros) para evitar errores al trabajar después.

Al final, muestro el tamaño (filas/columnas), el tipo de datos y una vista previa.

## Describir dimensiones, tipos de variables y estadísticos básicos

**Dimensiones.**  
Trabajamos con la hoja “Sheet1” que tiene 3,000 filas y 30 columnas.

**Tipos de variables.**

Hay 29 columnas numéricas (28 enteras y 1 decimal) y 1 columna de texto (nombres\_completos). No hay columnas de fecha. Varios campos “categóricos” vienen codificados con números (por ejemplo: genero, carrera, estado\_civil, formas\_de\_pago, situacion\_laboral, etc.).

**Estadísticos básicos (los más relevantes).**

* **Edad (edad)**: media 27.53, mediana 28, mínimo 16, máximo 39 (IQR 22–33).
* **Promedio ponderado del semestre (promedio\_ponderado\_del\_semestre)**: media 11.97, mediana 11.98, p25 10.03, p75 13.97, rango 2.06–20.00.
* **Asistencia (asistencia, %)**: media 69.75%, mediana 69%, p25 55%, p75 85%, rango 40–100.
* **Desarrollo de tareas (desarrollo\_de\_tareas, %)**: media 70.48%, mediana 71%, p25 55%, p75 86%, rango 40–100.
* **Carga y logro por periodo**:
  + unidades\_curriculares\_matriculadas: media 5.55, mediana 6, rango 3–8 (p25 4, p75 7).
  + unidades\_curriculares\_evaluadas: media 5.02, mediana 5, rango 2–8 (p25 4, p75 7).
  + unidades\_curriculares\_aprobadas: media 2.49, mediana 2, rango 0–8 (p25 1, p75 4).
* **Indicadores binarios (proporción de “1”)**: matricula\_al\_dia 81.27%, deudor 29.93%, becario 15.20%, trasladado\_a 49.33%, necesidades\_educativas\_especiales 5.47%, crianza\_adventista 49.17%.
* **Condición final (condicion\_final)**: categoría 1 = 1,599 casos (53.3%), categoría 2 = 1,401 casos (46.7%).

## Identificar problemas de calidad (faltantes, duplicados, outliers)

1. **Faltantes (huecos)**

* Qué es: celdas vacías.
* Cómo lo voy a ver: sacar el % de huecos por columna.
* Qué haré: si son pocos, relleno (mediana/moda); si son muchos, no uso esa columna o la elimino.

1. **Duplicados (registros repetidos)**

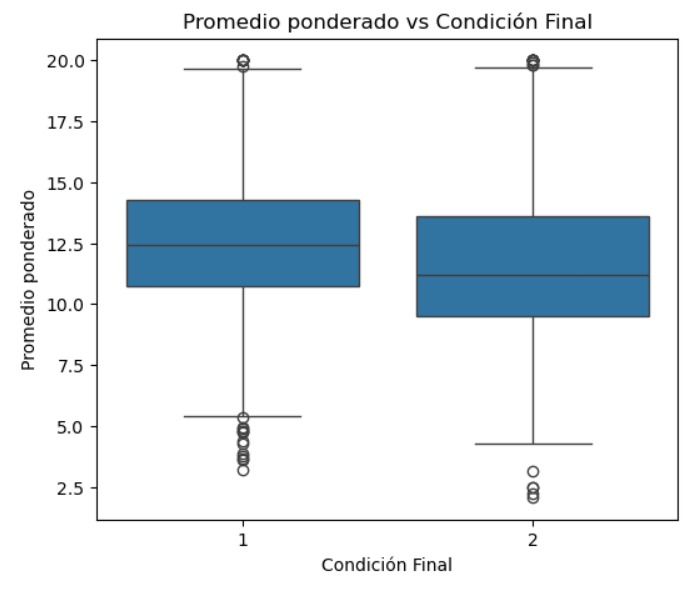
* Qué es: filas iguales o el mismo estudiante repetido (si hay código/DNI/ID).
* Cómo lo voy a ver: contar repetidos totales y, si existe ID, repetidos por ID de estudiante.
* Qué haré: si son copias exactas, borro las extras; si difieren, me quedo con la versión más completa o pido corrección.

1. **Outliers (valores raros)**

* Qué es: números imposibles o muy extremos (ej.: asistencia > 100, nota fuera de 0–20).
* Cómo lo voy a ver: revisar si están muy lejos del resto y validar rangos lógicos.
* Qué haré: si es error, corrijo o dejo vacío; si es extremo pero posible, recorto a límites o uso mediana/IQR en el análisis.

## Generar 2 visualizaciones descriptivas

El gráfico muestra cuántos estudiantes hay en cada edad. Se ve un rango aproximado de 16 a 39 años, con más presencia en los veintitantos y un pequeño “lomito” cerca de los 30–31. La forma es bastante pareja, sin picos raros ni edades imposibles, así que la muestra luce normal para un grupo universitario joven–adulto.



Aquí se compara la distribución de promedios (0–20) entre las dos categorías de “Condición final”. La mediana del grupo “1” está un poco por encima de la del grupo “2”, y en el grupo “2” aparecen más valores bajos y una dispersión mayor hacia abajo.

# DOCUMENTAR EN NOTEBOOK JUPYTER CON DOS SECCIONES

<https://github.com/LuisOrihuela08/project-fase-2/blob/main/project-students.ipynb>

# SUBIR A GITHUB

<https://github.com/LuisOrihuela08/project-fase-2.git>