



FACULTAD DE CIENCIAS MATEMÁTICAS Y  
FÍSICAS

Carrera: Ciencia de Datos e Inteligencia Artificial

## **Proyecto Final**

**ALMACENES DE DATOS Y MINERÍA DE DATOS**

**CDDEIA-ELNO-5-2**

**Integrantes:**

Sancán Chóez Javier Fernando

**Docente:**

**Ing. León Granizo Oscar Dario**

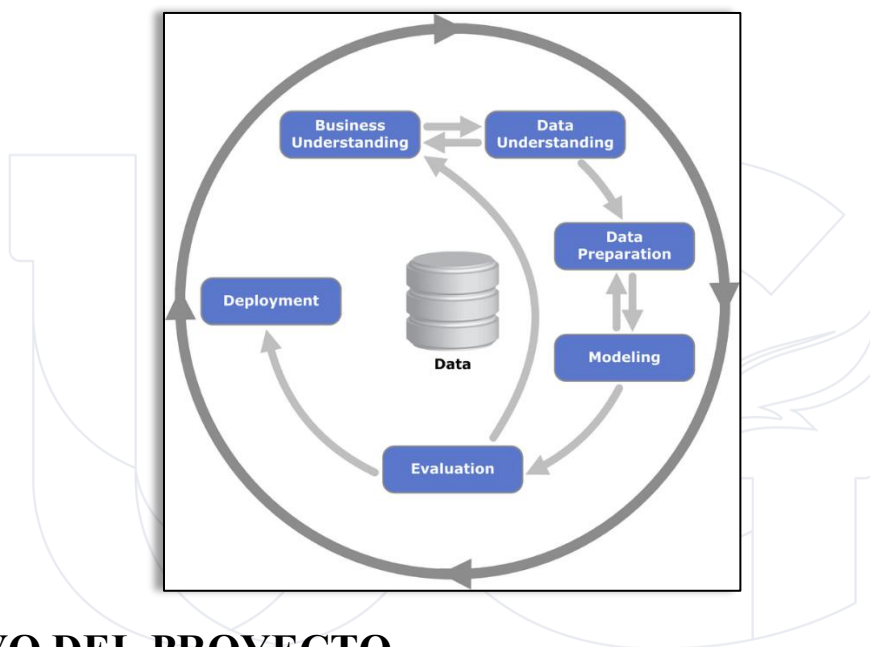
**Año Lectivo:**

2025 – 2026 CII

# PROYECTO DE SISTEMA DE PREDICCIÓN DE DESERCIÓN ESTUDIANTIL

## INTRODUCCIÓN

Este informe técnico documenta el desarrollo del sistema de predicción de deserción estudiantil, estructurado bajo la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) conforme a los requerimientos del proyecto.



## OBJETIVO DEL PROYECTO

### Problema a Resolver

La deserción estudiantil representa un desafío crítico en educación superior. Identificar tempranamente a estudiantes en riesgo permite implementar intervenciones oportunas.

### Objetivo General

Desarrollar un modelo predictivo de Machine Learning que identifique estudiantes con alta probabilidad de deserción, utilizando datos históricos académicos.

## RESUMEN DEL PROYECTO

## Datos Disponibles

- **Total de registros:** 4,448 registros académicos
- **Estudiantes únicos:** 488 estudiantes
- **Períodos académicos:** 9 períodos (2023-2026)
- **Facultad:** Ciencias Matemáticas Y Físicas
- **Carrera:** Ciencia De Datos E Inteligencia Artificial

PERIODO	FACULTAD	CARRERA	ESTUDIANTE	COD_MATERIA	MATERIA	COD_GRUPO	GRUPO/PARALELO	ESTADO	PROMEDIO	ASISTENCIA	NO. VEZ
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 1	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	REPROBADA	0,00	0	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 2	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	REPROBADA	0,00	0	2
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 3	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,07	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 4	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,48	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 5	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	REPROBADA	6,87	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 6	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,15	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 7	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	7,97	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 8	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	7,75	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 9	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	REPROBADA	0,00	0	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 10	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,60	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 11	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	APROBADA	8,98	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 12	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,01	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 13	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,56	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 14	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	REPROBADA	0,00	0	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 15	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,90	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 16	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	7,79	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 17	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,24	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 18	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,32	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 19	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,98	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 20	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	REPROBADA	0,00	0	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 21	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,98	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 22	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	APROBADA	9,25	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 23	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,66	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 24	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	8,98	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 25	318010	CÁLCULO I	3180011	CDDEIA-ELNO-1-11	VAN20	APROBADA	9,07	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 26	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	REPROBADA	4,13	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 27	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	9,24	100	1
2024 - 2025 CI	CIENCIAS MATEMATICAS Y FISICAS	CIENCIA DE DATOS E INTELIGENCIA A: Estudiante 28	318010	CÁLCULO I	3180001	CDDEIA-ELMA-1-1	VAN20	APROBADA	7,11	100	1

## Variables del DataSet

Variables	Tipo	Descripción	Utilidad
PERIODO	Categorica	Período académico	Media
ESTUDIANTE	ID	Identificador anónimo del estudiante	Alta
COD_MATERIA	Numérica	Código de la materia	Baja
MATERIA	Texto	Nombre de la materia	Baja
ESTADO	Categorica	APROBADA / REPROBADA	Muy Alta
PROMEDIO	Numérica*	Calificación (0-10)	Muy Alta
ASISTENCIA	Numérica	Porcentaje de asistencia (0-100)	Muy Alta
NO. VEZ	Numérica	Veces cursando la materia	Muy Alta
NIVEL	Numérica	Semestre del estudiante (1-4)	Media

## LENGUAJE Y HERRAMIENTAS A UTILIZAR

- Python
- Google Colab
- GitHub
- Streamlit

## FASES METODOLOGÍA CRISP-DM (CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING)

### 1. Comprensión del Negocio

La deserción estudiantil es uno de los desafíos más críticos para las instituciones de educación superior. El objetivo primordial es identificar de manera temprana a los estudiantes con mayor riesgo de abandono para implementar estrategias de intervención oportunas.

Objetivo del Modelo: Construir un clasificador que prediga la probabilidad de que un estudiante se retire basándose en su rendimiento académico histórico.

### 2. Comprensión de los Datos

Se utilizó un conjunto de datos anonimizado con el historial académico. Las variables clave identificadas fueron:

- PROMEDIO: Calificación media del estudiante.
- ASISTENCIA: Porcentaje de presencia en clases.
- NO. VEZ: Número de veces que se ha cursado una materia (reincidencia).
- ESTADO: Si la materia fue aprobada o reprobada.
- NIVEL: Semestre o nivel actual del estudiante.

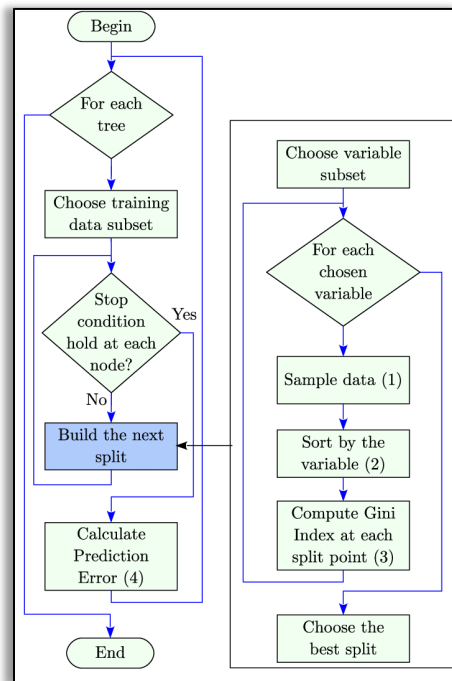
### 3. Preparación de los Datos



Para transformar los datos crudos en un formato apto para la minería de datos, se realizaron las siguientes tareas:

1. Limpieza: Se corrigió el formato de la columna PROMEDIO (reemplazo de ',' por '.') para convertirla a tipo numérico decimal.
2. Agregación: Se consolidó la información por cada ESTUDIANTE único, calculando promedios generales y máximos niveles alcanzados.
3. Ingeniería de Características: Se creó la variable REPROBADAS contando las ocurrencias de materias con estado "REPROBADA".
4. Definición de Variable Objetivo: Se definió la Deserción (1) como aquellos estudiantes que no registran actividad en el último periodo académico disponible en el dataset, y Persistencia (0) para los que sí aparecen en el periodo vigente.

## 4. Modelado



Se seleccionó la técnica de modelo **Random Forest Classifier** (Bosques Aleatorios).

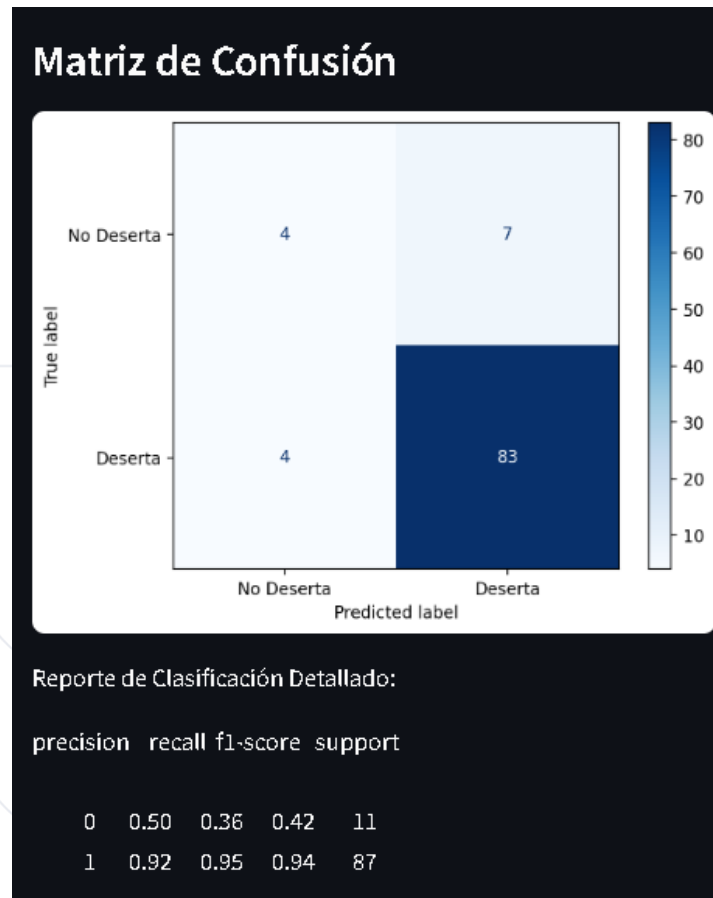
- Justificación: Este modelo es robusto frente a valores atípicos y permite identificar qué variables (como la asistencia o las materias reprobadas) tienen más peso en la decisión final.
- Entrenamiento: Los datos se dividieron en un 80% para entrenamiento y un 20% para pruebas para garantizar la capacidad de generalización del modelo.

## 5. Evaluación

El modelo se evaluó utilizando las métricas estándar requeridas:

- Accuracy: Precisión global del sistema.
- Matriz de Confusión: Para visualizar cuántos estudiantes en riesgo fueron detectados correctamente vs. falsas alarmas.

- Reporte de Clasificación: Incluye Precisión, Recall y F1-Score para ambas clases (Desertor/No Desertor).
- Importancia de Variables: El análisis reveló que factores como el promedio y el número de materias reprobadas son los predictores más fuertes del riesgo académico.



## 6. Despliegue

Se desarrolló una interfaz gráfica interactiva utilizando Streamlit. Las funcionalidades implementadas incluyen:

- Resumen: Introducción y objetivos del proyecto.
- Análisis de Datos (EDA): Gráficos de distribución de deserción y diagramas de caja para comparar la asistencia.
- Evaluación: Presentación visual de las métricas de rendimiento y la matriz de confusión.

- Predicción Individual: Formulario dinámico donde el usuario puede ingresar datos de un estudiante específico y obtener su nivel de riesgo en tiempo real.

