



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.

LENGUAJE Y HERRAMIENTAS A UTILIZAR

- Python
- Google Colab
- GitHub
- Streamlit

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

Variables del DataSet

Variables	Tipo	Descripción	Utilidad
PERIODO	Categórica	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
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
ESTADO	Categórica	APROBADA / REPROBADA	Muy Alta
NIVEL	Numérica	Semestre del estudiante (1-4)	Media

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.