CDDEIA-ELNO-5-2

Ciencias de Datos e inteligencia Artificial

Estudiante:
Roberto Alvarez

Proyecto final:
Predicción de Deserción Estudiantil mediante Minería de Datos

Materia:
Almacenes y minería de datos

Docente:
Msc. Oscar Dario Leon Granizo

Año:
2026

Informe Técnico: Predicción de Deserción Estudiantil mediante Minería de Datos

1. Comprensión del Negocio (Business Understanding) 3
Contexto del Problema 3
Objetivos del Proyecto 3
2. Comprensión de los Datos (Data Understanding) 3
Fuente de Datos 3
Análisis de Variables 3
Justificación de la Variable Objetivo (Target) 4
3. Preparación de los Datos (Data Preparation) 4
4. Modelado (Modeling) 4
Selección del Algoritmo 4
Configuración del Modelo 4
5. Evaluación (Evaluation) 5
6. Despliegue (Deployment) 5
Conclusión 8
Anexos 8
LINK AL REPOSITORIO EN GITHUB: 8

Informe Técnico: Predicción de Deserción Estudiantil mediante Minería de Datos

Metodología: CRISP-DM (Cross-Industry Standard Process for Data Mining)

1. Comprensión del Negocio (Business Understanding)

Contexto del Problema

La deserción estudiantil representa un desafío crítico para las instituciones de educación superior. La pérdida de estudiantes no solo afecta los indicadores institucionales, sino que impacta el desarrollo profesional de los jóvenes.

Objetivos del Proyecto

- Identificar patrones: Determinar qué variables académicas correlacionan con el abandono.

- Modelo Predictivo: Desarrollar un algoritmo capaz de clasificar a los estudiantes según su riesgo de deserción real.

- Acción Temprana: Proveer una herramienta a los tutores para intervenir antes de que el estudiante abandone la carrera.

2. Comprensión de los Datos (Data Understanding)

Fuente de Datos

Se utiliza el archivo REPORTE_RECORD_ESTUDIANTIL_ANONIMIZADO.xlsx, el cual contiene el historial académico detallado por estudiante, materia y periodo.

Análisis de Variables

- Académicas: Promedio General, Peor Promedio obtenido en el ciclo, Tasa de Reprobación.

- Conductuales: Asistencia Promedio (filtrada para evitar sesgos de materias convalidadas).

- Históricas: Máximo de intentos por materia (indicador de rezago).

Justificación de la Variable Objetivo (Target)

Se implementó una lógica de Deserción Real: un estudiante es marcado como desertor (Target = 1) si, habiendo cursado un periodo ordinario, no registra actividad en los periodos siguientes del dataset.

3. Preparación de los Datos (Data Preparation)

En esta fase se transformaron los datos crudos en un dataset apto para Machine Learning mediante el script entrenar.py:

- Limpieza y Filtrado: Se eliminaron registros de periodos no ordinarios (Cursos de Verano/Inglés) y materias de "Movilidad" o "Prácticas".

- Ingeniería de Características: Creación de un sistema de puntos para asignar Categorías de Riesgo (BAJO, MEDIO, ALTO, DESERTOR).

- Tratamiento de Nulos: Conversión de formatos numéricos y manejo de valores vacíos.

4. Modelado (Modeling)

Selección del Algoritmo

Se seleccionó Random Forest Classifier por su capacidad para manejar relaciones no lineales y su robustez frente a datos desequilibrados.

Configuración del Modelo

- Bosque: 100 árboles de decisión.

- Balanceo: class_weight="balanced" para manejar el desequilibrio de clases.

- Validación: División de datos en 80% entrenamiento y 20% prueba.

5. Evaluación (Evaluation)

El modelo es evaluado mediante cuatro métricas fundamentales:

Métrica	Propósito
Accuracy	Porcentaje total de predicciones correctas.
Precisión	Capacidad de no marcar como desertor a un estudiante que va a continuar.
Recall	Capacidad del modelo para detectar a la mayor cantidad de desertores reales.
F1-Score	Balance entre precisión y recall.

Justificación: Se priorizó el Recall en la evaluación, ya que es preferible monitorear a un falso positivo que ignorar a un desertor real.

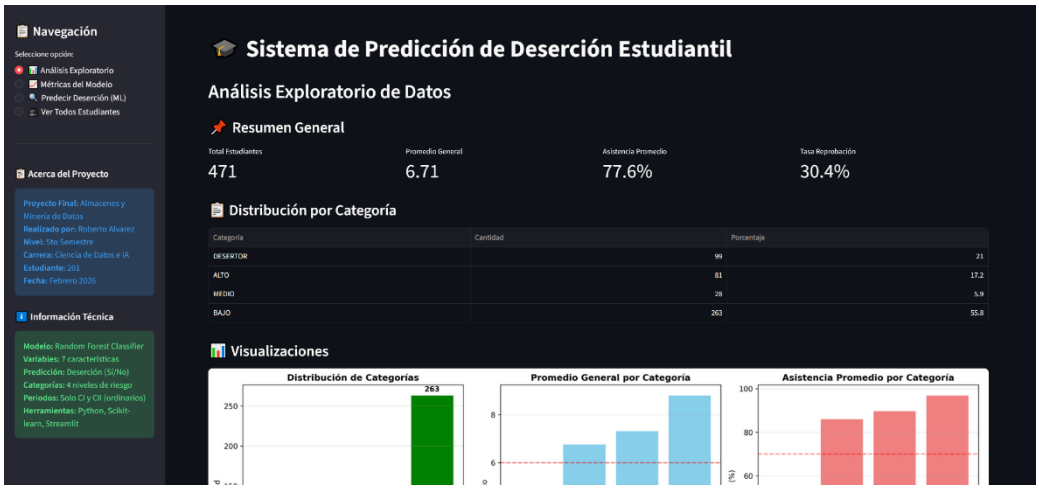
6. Despliegue (Deployment)

El proyecto se despliega a través de una aplicación interactiva en Streamlit (app.py):

- Dashboard Exploratorio: Visualización institucional de la salud académica.

- Módulo de Predicción Individual: Formulario para ingresar datos de un estudiante y obtener probabilidad de riesgo.

- Gestión de Reportes: Exportación de listas a Excel/CSV para tutores.

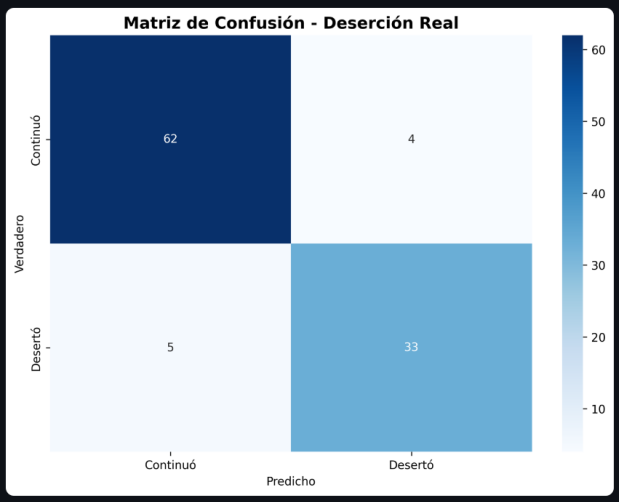


Evaluación del Modelo de Machine Learning

🎯 Métricas de Rendimiento

Accuracy	Precisión	Recall	F1-Score
0.9135	0.8919	0.8684	0.8800

Matriz de Confusión

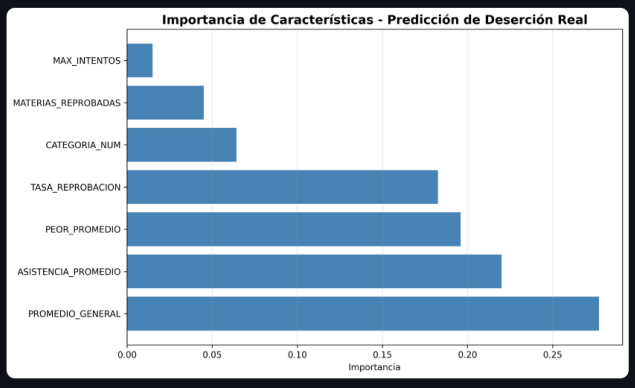


Matriz de Confusión del Modelo

Interpretación:

- **Verdaderos Positivos:** Desertores correctamente identificados
- **Falsos Positivos:** No desertores identificados como desertores
- **Falsos Negativos:** Desertores no detectados
- **Verdaderos Negativos:** No desertores correctamente identificados

📊 Importancia de Características



Importancia de Variables para la Predicción

Características ordenadas por importancia:

1. **PROMEDIO_GENERAL:** 0.2771
2. **ASISTENCIA_PROMEDIO:** 0.2200
3. **PEOR_PROMEDIO:** 0.1958
4. **TASA_REPROBACION:** 0.1827
5. **CATEGORIA_NUM:** 0.0643
6. **MATERIAS_REPROBADAS:** 0.0451
7. **MAX_INTENTOS:** 0.0150

⚙ Detalles del Modelo

Algoritmo: Random Forest Classifier

Método de validación: Train/Test Split (80%/20%)

Número de árboles: 100

Características utilizadas: 7

Ponderación de clases: Balanceada

Periodos analizados: Solo CI y CII (ordinarios)

Target: Deserción Real (0=Continuó, 1=Desertó)

	ESTUDIANTE	PERIODO	CATEGORIA	PROMEDIO_GENERAL	ASISTENCIA_PROMEDIO	TASA_REPROBACION	MAX_INTENTOS
113	Estudiante 114	2024 - 2025 CII	DESERTOR	0.15	6.3%	100.0%	1
121	Estudiante 122	2024 - 2025 CII	DESERTOR	0.81	10.7%	100.0%	1
125	Estudiante 126	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	1
135	Estudiante 136	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	1
13	Estudiante 14	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	1
153	Estudiante 154	2025 - 2026 CI	DESERTOR	0	0.7%	100.0%	2
165	Estudiante 166	2024 - 2025 CII	DESERTOR	2.61	62.0%	100.0%	1
167	Estudiante 168	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	1
174	Estudiante 175	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	2
178	Estudiante 179	2024 - 2025 CII	DESERTOR	0	0.0%	100.0%	1

Conclusión

Este sistema nos sirve para cerrar la brecha entre el almacenamiento de datos académicos y la toma de decisiones, transformando registros históricos en una herramienta preventiva activa para la Universidad de Guayaquil.

Anexos

Link al repositorio en GitHub:
https://github.com/Robe1o/prediccion-desercion-estudiantil