UNIVERSIDAD DE GUAYAQUIL

Universidad De Guayaquil
Facultad De Ciencias Matemáticas y Físicas
Carrera de Ciencias de Datos e Inteligencia Artificial

ALMACENES DE DATOS Y MINERIA DE DATOS

Proyecto Final

Estudiante:

Álvarez Sánchez José Alejandro

CURSO:

CDDEIA-ELNO-5-2

DOCENTE:

LEON GRANIZO OSCAR DARIO

CICLO II

2025 – 2026

Introducción

La deserción estudiantil es un fenómeno que afecta de manera significativa a las instituciones de educación superior, ya que implica la interrupción del proceso formativo de los estudiantes y repercute en los indicadores académicos y de retención institucional. Identificar de forma temprana a los estudiantes con riesgo de abandonar sus estudios es un aspecto clave para implementar acciones de apoyo oportunas.

En este contexto, el presente proyecto aplica técnicas de minería de datos para analizar información académica histórica y desarrollar un modelo predictivo de deserción estudiantil. El análisis se basa en datos académicos anonimados de estudiantes de la carrera de Ciencia de datos e inteligencia artificial, considerando variables relacionadas con el rendimiento académico, la asistencia y la repitencia.

El desarrollo del proyecto sigue la metodología CRISP-DM, la cual proporciona un marco estructurado para abordar problemas de minería de datos, desde la comprensión del problema hasta el despliegue de los resultados. Como parte final del proyecto, se implementa una aplicación interactiva utilizando Streamlit, que permite visualizar el análisis realizado y obtener predicciones individuales sobre el riesgo de deserción.

Comprensión del negocio

La deserción estudiantil es un problema relevante en las instituciones de educación superior, ya que afecta tanto a la continuidad académica de los estudiantes como a los indicadores institucionales de rendimiento y retención.

La identificación temprana de estudiantes con riesgo de abandono permite implementar estrategias de acompañamiento académico oportunas, con el objetivo de mejorar la permanencia estudiantil.

En este proyecto se aplican técnicas de minería de datos para analizar información académica histórica y construir un modelo predictivo que permita anticipar la deserción estudiantil.

Definición del problema

La institución dispone de un conjunto de datos académicos anonimizado que contiene información histórica de estudiantes de la carrera de Ciencia de Datos e Inteligencia Artificial, registrada a lo largo de varios períodos académicos.

El problema consiste en predecir si un estudiante abandonará sus estudios en el siguiente período académico, utilizando únicamente información académica disponible hasta el período actual.

Objetivo general

Desarrollar un modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos, y presentar los resultados mediante una aplicación interactiva desarrollada en Streamlit.

Objetivos específicos

- Analizar el conjunto de datos académicos proporcionado.
- Identificar variables relevantes para la predicción de deserción.
- Definir una variable objetivo que represente el fenómeno de deserción.

Alcance

El proyecto se limita al análisis de datos académicos de una única carrera universitaria. No se consideran variables personales, socioeconómicas ni demográficas, y el modelo tiene fines académicos y de apoyo a la toma de decisiones.

Criterio de éxito

El proyecto se considera exitoso si el modelo logra identificar adecuadamente a estudiantes en riesgo de deserción y la aplicación desarrollada permite visualizar los resultados y realizar predicciones de forma clara e intuitiva.

Comprensión de los datos

Se describe el conjunto de datos proporcionado y se realiza un análisis exploratorio inicial con el objetivo de comprender su estructura, contenido y principales características, antes de proceder a la preparación y modelado de los datos.

El análisis se basa en un conjunto de datos académicos anonimizado correspondiente a estudiantes de la carrera de ciencia de datos e inteligencia artificial, registrados a lo largo de varios períodos académicos.

Descripción del conjunto de datos

El dataset original contiene información académica a nivel de asignatura. Cada registro representa una materia cursada por un estudiante en un período académico específico. Entre las variables disponibles se incluyen identificadores de estudiante, período académico, carrera, nivel, calificaciones finales, asistencia, estado de la asignatura (aprobada o reprobada) y el número de veces que una asignatura ha sido cursada.

Debido a la naturaleza del problema de deserción, este nivel de detalle fue posteriormente transformado a un formato agregado por estudiante y período académico, permitiendo analizar el comportamiento académico global del estudiante en cada período.

Estructura y organización de los datos

Tras el proceso de agregación, el conjunto de datos final está compuesto por 519 registros, que corresponden a 352 estudiantes únicos, distribuidos en tres períodos académicos regulares. Cada fila del dataset procesado representa el desempeño académico de un estudiante en un período determinado.

La estructura final del dataset incluye variables académicas agregadas, una variable temporal que identifica el período académico y una clave numérica (PERIODO_KEY) que permite ordenar cronológicamente los registros.

Variables disponibles

Las principales variables consideradas en el conjunto de datos procesado son:

- **nivel**: nivel académico en el que se encuentra el estudiante.
- **materias_cursadas**: número de asignaturas cursadas durante el período.
- **promedio_periodo**: promedio de calificaciones obtenidas en el período.
- **asistencia_prom**: porcentaje promedio de asistencia del estudiante.
- **reprobadas** y **aprobadas**: cantidad de asignaturas reprobadas y aprobadas.
- **max_no_vez**: máximo número de veces que el estudiante ha cursado una asignatura.
- **repitencia_prom**: promedio del número de veces que las asignaturas han sido cursadas.
- **tasa_reprobacion**: proporción de asignaturas reprobadas respecto al total cursado.

Estas variables permiten caracterizar el rendimiento académico y el historial de repitencia del estudiante, factores comúnmente asociados al riesgo de deserción.

Variable objetivo

La variable objetivo-definida para el proyecto es DESERCION_NEXT, la cual indica si el estudiante no se matricula en el período académico siguiente**.**

Valor 1: el estudiante no se reinscribe en el siguiente período (deserción).

Valor 0**:** el estudiante continúa sus estudios en el período siguiente.

Esta definición permite abordar la deserción desde una perspectiva de detección temprana**,** ya que el modelo utiliza únicamente información disponible hasta el período actual para realizar la predicción.

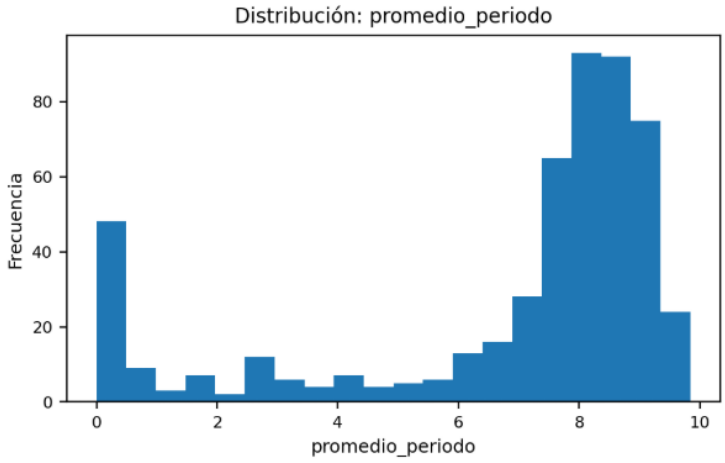
Análisis exploratorio inicial

El análisis exploratorio de los datos permitió identificar las siguientes características relevantes:

- La variable objetivo presenta una distribución aproximada de 62% de estudiantes que continúan **y** 38% que desertan, lo que indica un desbalance moderado entre clases.
- Las variables numéricas muestran distribuciones coherentes con el contexto académico, sin la presencia de valores extremos anómalos.
- La mayoría de los estudiantes presenta bajas tasas de repitencia y reprobación, lo cual es consistente con un desempeño académico regular.
- No se detectaron valores faltantes en las variables seleccionadas para el modelado.

Estos resultados indican que el conjunto de datos cuenta con una calidad adecuada para la construcción de un modelo de clasificación supervisado.

Durante el análisis exploratorio se evaluó la distribución de las principales variables académicas. En particular, la Figura X muestra la distribución del promedio de calificaciones por período académico, la cual presenta un comportamiento coherente con el contexto educativo.



Conclusión de comprensión de datos

A partir del análisis realizado, se concluye que el conjunto de datos proporciona información suficiente y relevante para abordar el problema de predicción de deserción estudiantil. Las variables académicas disponibles permiten describir el desempeño del estudiante en cada período y constituyen una base sólida para las etapas posteriores de preparación de datos y modelado.

Preparación de los datos

En esta sección se describen las transformaciones realizadas al conjunto de datos con el objetivo de obtener una versión limpia y adecuada para el modelado predictivo. Dado que los datos originales se encuentran a nivel de asignatura (múltiples filas por estudiante y

período), fue necesario aplicar procesos de limpieza, estandarización, agregación y construcción de la variable objetivo.

Limpieza y estandarización de variables

Durante la revisión inicial se identificó que la variable de calificación (PROMEDIO) se encontraba almacenada como texto debido al uso de coma como separador decimal. Para permitir su análisis numérico, se convirtió a formato numérico generando una nueva variable:

PROMEDIO_NUM: conversión de la calificación a valor numérico (reemplazo de coma por punto y conversión a tipo float).

Adicionalmente, se verificó la consistencia de variables numéricas como asistencia, nivel académico y número de veces que se cursa una asignatura, asegurando que los datos estén en rangos razonables para el contexto académico.

Organización temporal de períodos académicos

Para modelar deserción es necesario respetar el orden cronológico de los períodos académicos. Por ello, se construyó una clave numérica:

PERIODO_KEY: variable que permite ordenar los períodos académicos de forma temporal.

Esta clave evita problemas de ordenamiento al trabajar con períodos almacenados como texto y permite aplicar un esquema de evaluación más realista (por ejemplo, usando el último período como conjunto de prueba).

Selección de períodos académicos

Se trabajó únicamente con períodos regulares (por ejemplo, “CI” y “CII”), debido a que períodos especiales pueden corresponder a casos particulares (nivelación, módulos específicos u otras modalidades) y podrían introducir ruido en la definición de continuidad académica.

Esta decisión busca que la definición de deserción sea consistente bajo un mismo criterio académico (reinscripción en el siguiente período regular).

Agregación por estudiante y período

Como el dataset original contiene información por asignatura, se transformó a un nivel de análisis por estudiante y período académico**,** generando un dataset donde cada fila representa el desempeño global del estudiante durante un período. Se calcularon variables agregadas que resumen el comportamiento académico del estudiante, incluyendo:

- **materias_cursadas**: total de asignaturas cursadas en el período.
- **promedio_periodo**: promedio de calificaciones del período.

- **asistencia_prom**: asistencia promedio del período.
- **aprobadas** y **reprobadas**: conteo de asignaturas aprobadas y reprobadas.
- **tasa_reprobacion**: reprobadas / materias_cursadas.
- **max_no_vez** y **repitencia_prom**: indicadores de repitencia en el período.

Estas variables fueron seleccionadas porque reflejan factores académicos relacionados con el riesgo de deserción, como bajo rendimiento, alta reprobación o repitencia.

Construcción de la variable objetivo (deserción)

Dado que el dataset no incluye un campo explícito de "desertor", se definió la deserción de forma operativa a partir de la continuidad del estudiante en los registros académicos.

Se definió la variable objetivo:

DESERCION_NEXT: indica si el estudiante no aparece matriculado en el siguiente período académico regular.

Criterio:

DESERCION_NEXT = 1 si el estudiante no tiene registros en el período siguiente.

DESERCION_NEXT = 0 si el estudiante sí aparece en el período siguiente.

Esta definición permite construir un modelo con enfoque de alerta temprana, es decir, identificar riesgo de deserción usando información disponible hasta el período actual.

Resultado final de los datos preparados

Como resultado del proceso de preparación, se obtuvo un conjunto de datos procesado y listo para modelado, almacenado en formato CSV. Este dataset contiene variables agregadas por estudiante y período, y la variable objetivo definida, manteniendo consistencia temporal y evitando el uso de información futura.

Modelado

En esta sección se describe el proceso de construcción del modelo predictivo de deserción estudiantil, a partir del conjunto de datos previamente preparado. El objetivo del modelado es aprender patrones a partir del desempeño académico histórico que permitan identificar estudiantes con riesgo de deserción en el siguiente período académico.

Definición del problema de modelado

El problema abordado corresponde a un problema de clasificación binaria, donde la variable objetivo DESERCION_NEXT toma los siguientes valores:

1: el estudiante no se reinscribe en el siguiente período académico (deserción).

0: el estudiante continúa sus estudios en el período siguiente.

El modelo utiliza únicamente información disponible hasta el período actual, respetando la naturaleza temporal del problema.

Selección de variables de entrada

Para el modelado se seleccionaron exclusivamente **variables académicas agregadas**, ya que son las únicas disponibles y directamente relacionadas con el rendimiento del estudiante. Las variables utilizadas incluyen:

- nivel
- materias_cursadas
- promedio_periodo
- asistencia_prom
- tasa_reprobacion
- max_no_vez
- repitencia_prom

Adicionalmente, se incluyó la variable categórica **carrera**, la cual fue codificada mediante técnicas de transformación adecuadas dentro del pipeline de preprocesamiento.

No se utilizaron identificadores como el código del estudiante ni variables temporales como el período, con el fin de evitar sesgos y fuga de información.

Selección del modelo

Para la construcción del modelo predictivo se seleccionó el algoritmo de regresión logística, dado que es una técnica ampliamente utilizada en problemas de clasificación binaria y resulta adecuada para el objetivo del proyecto. Este modelo permite estimar la probabilidad de que un estudiante abandone sus estudios en el siguiente período académico, a partir de variables explicativas relacionadas con su desempeño académico. Además, la regresión logística ofrece un alto nivel de interpretabilidad, lo que facilita el análisis del impacto de cada variable sobre el riesgo de deserción, aspecto especialmente relevante en un contexto educativo donde se requiere justificar las decisiones tomadas a partir del modelo. Su simplicidad, eficiencia computacional y buen desempeño en conjuntos de datos de tamaño moderado justifican su elección frente a modelos más complejos.

Pipeline de preprocesamiento y entrenamiento

El proceso de modelado se implementó mediante un pipeline, integrando en una sola estructura las etapas de preprocesamiento de datos y entrenamiento del modelo. Este enfoque garantiza consistencia entre las transformaciones aplicadas durante el entrenamiento y aquellas utilizadas en la fase de predicción. Dentro del pipeline se incluyó

la selección de variables numéricas y categóricas, así como la transformación de las variables categóricas mediante técnicas de codificación apropiadas. Posteriormente, se entrenó el modelo de regresión logística utilizando el conjunto de datos preparado. Considerando que la variable objetivo presenta un desbalance moderado entre las clases, se configuró el modelo para utilizar pesos balanceados, con el fin de penalizar de manera adecuada los errores asociados a la clase de deserción. El uso del pipeline facilita la reproducibilidad del proceso y permite su integración directa en la etapa de despliegue del sistema.

Consideraciones finales del modelado

El modelo entrenado genera una probabilidad de deserción para cada estudiante, la cual puede ser interpretada como un nivel de riesgo. Esta probabilidad permite ajustar umbrales de decisión según el contexto institucional y las necesidades de intervención temprana.

El modelo final entrenado fue almacenado para su posterior uso en la aplicación de despliegue desarrollada en Streamlit.

Evaluación

Se evalúa el desempeño del modelo predictivo de deserción estudiantil, con el objetivo de analizar su capacidad para identificar correctamente a los estudiantes en riesgo de abandonar sus estudios en el siguiente período académico.

Estrategia de evaluación

La evaluación del modelo se realizó utilizando un esquema de validación temporal, con el fin de simular un escenario real de predicción. Para ello, se utilizó el último período académico disponible como conjunto de prueba, mientras que los períodos anteriores se emplearon para el entrenamiento del modelo.

Este enfoque permite evitar la fuga de información futura y evaluar el modelo bajo condiciones similares a las que se presentarían en un entorno de aplicación real, donde solo se dispone de información histórica para realizar predicciones.

Métricas de evaluación

Para evaluar el desempeño del modelo se utilizaron métricas estándar de clasificación, incluyendo accuracy, precision, recall, F1-score y la matriz de confusión. Dado el objetivo del proyecto, se puso especial énfasis en el recall de la clase de deserción, ya que resulta prioritario identificar a la mayor cantidad posible de estudiantes en riesgo, incluso a costa de aceptar algunos falsos positivos.

Resultados obtenidos

El modelo de regresión logística obtuvo un desempeño satisfactorio en el conjunto de prueba, alcanzando un valor de accuracy cercano al 93%. En particular, el recall para la clase de deserción fue superior al 94%, lo que indica que el modelo logra identificar correctamente la gran mayoría de los estudiantes que efectivamente abandonan sus estudios en el período siguiente.

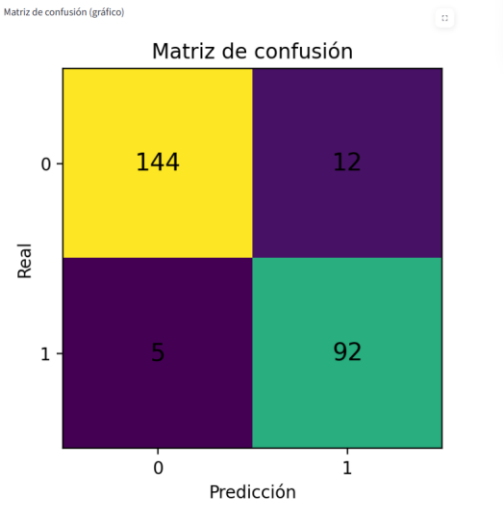
La matriz de confusión muestra un número reducido de falsos negativos, es decir, casos en los que el modelo no identifica a un estudiante que posteriormente deserta. Este resultado es especialmente relevante desde el punto de vista institucional, ya que minimiza el riesgo de no detectar estudiantes que podrían beneficiarse de acciones de acompañamiento académico.

Interpretación de los resultados

Los resultados obtenidos indican que el modelo presenta un buen equilibrio entre capacidad predictiva y estabilidad. La alta tasa de recall en la clase de deserción sugiere que el modelo es adecuado como herramienta de detección temprana, permitiendo identificar estudiantes con alto riesgo de abandono antes de que este ocurra.

Si bien el modelo puede generar algunos falsos positivos, estos casos pueden ser gestionados mediante intervenciones preventivas, lo cual es preferible a no detectar estudiantes en riesgo real. En este sentido, el desempeño del modelo resulta apropiado para su uso como sistema de apoyo a la toma de decisiones académicas.

La matriz de confusión permite analizar en detalle los aciertos y errores del modelo. Como se observa, el número de falsos negativos es reducido, lo que indica que el modelo logra identificar correctamente a la mayoría de los estudiantes que efectivamente desertan, minimizando el riesgo de no detectar casos críticos.



Conclusión de la evaluación

A partir de la evaluación realizada, se concluye que el modelo desarrollado cumple con los objetivos planteados en el proyecto y presenta un desempeño adecuado para la predicción de deserción estudiantil. Los resultados obtenidos validan la elección del modelo y las variables utilizadas, y respaldan su integración en una aplicación de despliegue para su uso práctico.

Despliegue

Como etapa final del proyecto, el modelo predictivo de deserción estudiantil fue desplegado mediante una aplicación web interactiva desarrollada con la herramienta Streamlit. El objetivo del despliegue es facilitar la visualización de los resultados del análisis y permitir la interacción con el modelo de forma sencilla y accesible.

La aplicación integra el modelo entrenado y el conjunto de datos procesado, permitiendo ejecutar inferencias sin necesidad de conocimientos técnicos avanzados. De esta manera, el sistema puede ser utilizado como una herramienta de apoyo para la toma de decisiones académicas.

Funcionalidades de la aplicación

La aplicación Streamlit se organiza en diferentes secciones que permiten explorar los datos y los resultados del modelo:

- **Análisis exploratorio (EDA)**: visualización de la estructura del conjunto de datos, estadísticas descriptivas y distribuciones de las principales variables académicas.
- **Métricas del modelo**: presentación de las métricas de evaluación obtenidas sobre el conjunto de prueba, incluyendo accuracy, precision, recall, F1-score y la matriz de confusión.
- **Predicción individual**: funcionalidad que permite ingresar manualmente los datos académicos de un estudiante y obtener una predicción sobre su riesgo de deserción en el siguiente período académico.
- **Importancia de variables**: visualización de la influencia relativa de las variables académicas en la predicción del modelo, facilitando la interpretación de los resultados.

Uso del sistema

El sistema fue diseñado para ser ejecutado localmente a través de un entorno de desarrollo en Python. La aplicación se inicia mediante un comando simple, lo que permite desplegar la interfaz web en un navegador sin configuraciones adicionales complejas.

El modelo utilizado en la aplicación corresponde al modelo entrenado y validado durante la etapa de modelado y evaluación, garantizando consistencia entre los resultados presentados y los obtenidos durante el análisis.

Beneficios del despliegue

El despliegue del modelo mediante una aplicación interactiva permite transformar los resultados del proyecto en una herramienta práctica, orientada a la detección temprana del riesgo de deserción. La interfaz facilita la interpretación de los resultados y contribuye a que el modelo pueda ser utilizado como apoyo para la identificación de estudiantes que podrían beneficiarse de acciones de acompañamiento académico.

Conclusión final

En el presente proyecto se desarrolló un sistema de predicción de deserción estudiantil utilizando técnicas de minería de datos, siguiendo la metodología CRISP-DM. A partir de información académica histórica, se logró construir un modelo de clasificación capaz de identificar estudiantes con riesgo de abandonar sus estudios en el siguiente período académico.

El modelo desarrollado, basado en regresión logística, mostró un desempeño adecuado en términos de métricas de evaluación, destacando especialmente un alto nivel de recall para la clase de deserción, lo cual resulta fundamental en escenarios de detección temprana. Las variables académicas relacionadas con el rendimiento, la carga académica y la repitencia demostraron ser relevantes para la predicción del fenómeno de deserción.

Finalmente, el despliegue del modelo mediante una aplicación interactiva desarrollada en Streamlit permitió transformar el análisis realizado en una herramienta práctica, facilitando la visualización de resultados y la realización de predicciones individuales. El sistema propuesto puede servir como apoyo para la toma de decisiones académicas orientadas a la prevención de la deserción estudiantil.

Como trabajo futuro, se plantea la posibilidad de incorporar nuevas variables, como información socioeconómica o demográfica, así como evaluar modelos adicionales que permitan comparar y mejorar el desempeño predictivo obtenido.