UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMATICAS Y FISICAS

CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

PROYECTO DE ALMACENES Y MINERIA DE DATOS

AUTOR

Salazar Alvarado Marlon Omar

ASIGNATURA

Almacenes y Minería de Datos

DOCENTE

ING. OSCAR DARIOLEON GRANIZO

CICLO

II 2026-2027

Contenido

Modelo Crips – DM

1. Comprensión del negocio

La deserción estudiantil constituye uno de los principales problemas que enfrentan las instituciones de educación superior, ya que afecta de manera directa la planificación académica, la eficiencia en el uso de recursos, los indicadores de calidad educativa y al cumplimiento de los objetivos institucionales. La perdida de estudiantes durante los ciclos académicos y profesional de los propios estudiantes. Por esta razón, resulta fundamental contar con mecanismo que permitan anticipar la deserción y actuar de manera preventiva.

En este contexto es importante poder predecir la cantidad de estudiantes que están propensos a desertar en el siguiente ciclo académico, y con esta información crearemos un modelo de clasificación que permita conocer cuántos estudiantes podrían retirarse de la carrera de Ingeniería en inteligencia artificial, con el fin de mejorar la toma de decisiones y optimizar la gestión académica. Contar con esta información de manera anticipada puede permitir tomar acciones ya sea en refuerzo académico, tutorías personalizadas, seguimiento a estudiantes con bajo rendimiento, o acompañamiento institucional.

En conclusión, la comprensión del negocio establece que este proyecto busca poder anticipar la deserción estudiantil mediante el análisis de datos académicos, permitiendo que se estime la cantidad de estudiantes que podrían retirarse en el próximo ciclo y facilitando la implementación de acciones preventivas que ayude a fortalecer la retención estudiantil y el desempeño institucional.

2. Comprensión de los datos

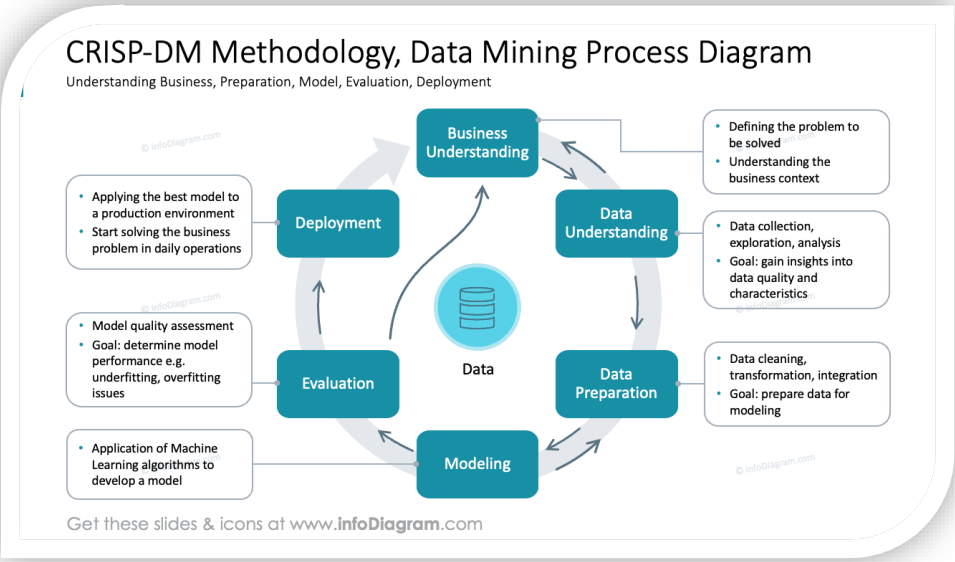
Los datos que se han utilizado en este proyecto provienen de un registro académico institucional anonimizado, almacenado en un archivo Excel y que en este conjunto de datos contiene información relevante de los estudiantes, incluyendo variables académicas, administrativas y de rendimiento.

Entre las principales variables encontramos

1. Facultad
2. Carrera
3. Nivel Académico
4. Promedio Académico
5. Porcentaje de asistencia
6. Estado del estudiante (deserta y no deserta)

En esta fase se realizó un análisis exploratorio (EDA) para poder comprender la estructura del Dataset, identificar el número de registros y las variables, analizamos los tipos de datos, también detectamos los valores nulos y observamos la distribución de la variable objetivo.

El análisis muestra que el dataset contiene información suficiente para poder abordar el problema de clasificación, aunque fue necesario realizar



procesos de limpieza y normalización de los nombres de las columnas para garantizar consistencia en el modelado.

3. Preparación de los datos

En la fase de preparación de los datos se llevaron a cabo las siguientes actividades:

- Limpieza de datos: revisión y tratamiento de valores nulos en las variables relevantes
- Normalización de nombres de columnas conversión a mayúsculas y reemplazo de espacio por guiones bajos
- Selección de variables: se eligieron las variables mas relevantes para el análisis, enfocándose en indicaciones académicos y administrativos
- Transformación de variables categóricas: Las variables como facultad y carrera fueron codificadas mediante técnicas de preprocesamiento de los datos (Por ejemplo, One-Hot Encoding)
- Separación de variables predictoras y variable objetivo: la variable “ESTADO” fue definida como variable objetivo del modelo.

Esta fase permitió poder obtener un conjunto de datos limpio y estructurado, listo para poder ser utilizado en la construcción del modelo.

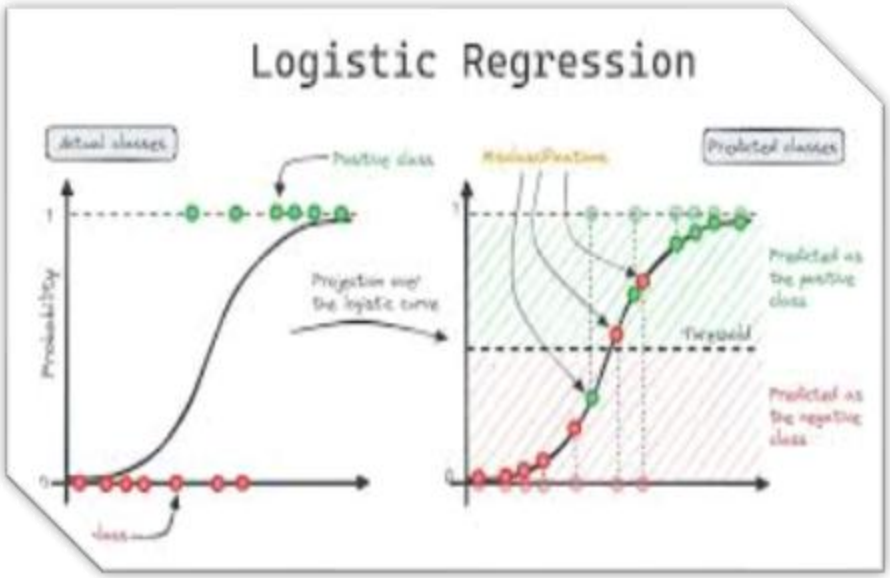
4. Modelado

Para el modelado se utilizado el algoritmo de Regresión Logística, adecuado para problemas de clasificación binaria como la predicción de deserción estudiantil

El proceso de modelado incluyo

- División del dataset en conjuntos de entrenamiento y prueba
- Construcción de un pipeline que integrante el preprocesamiento de datos y el clasificador
- Entrenamiento del modelo utilizando el conjunto de datos de entrenamiento

La regresión Logística fue seleccionada debido a su interpretabilidad, eficiencia y capacidad para analizar la influencia de cada variable en la probabilidad de deserción



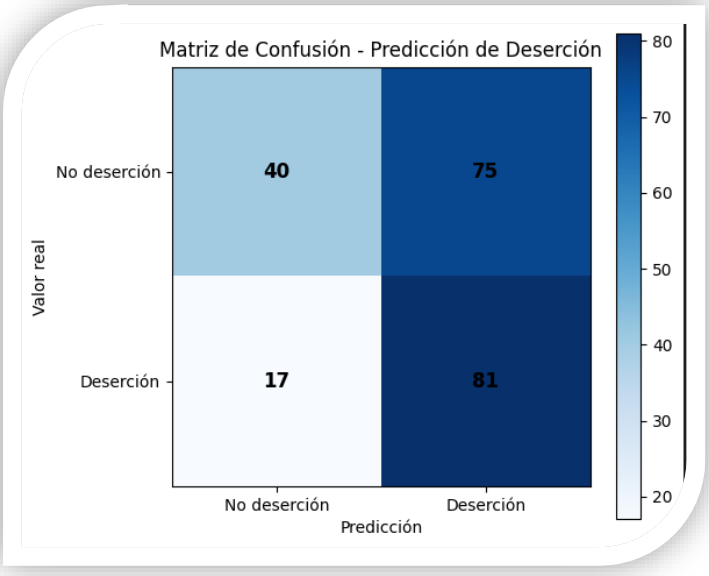
5. Evaluación

La evaluación del modelo se realizó utilizando métricas de clasificación ampliamente aceptadas.

- Accuaracy (Exactitud)
- Precisión (Precisión)
- Recall (Sensibilidad)
- F1 – score

Los resultados obtenidos muestran un desempeño aceptable del modelado, destacando un valor alto de recall, lo cual es especialmente importante en este contexto, ya que permite identificar a la mayoría de los estudiantes en riesgo de deserción.

Adicionalmente, se utilizado una matriz de confusión para visualizar el desempeño del modelo y analizar los errores de clasificación. Esto permitió identificar oportunidades de mejora ajustar del umbral de decisión según las necesidades de la universidad.



6. Despliegue

El modelo desarrollado fue desplegado una aplicación interactiva construida con streamlit. Esta aplicación permite:

- Visualizar información general y estadística del dataset
- Analizar la distribución del estado de los estudiantes
- Evaluar el desempeño del modelo predictivo
- Ingreso datos académicos de un estudiante y obtener una predicción del riesgo de deserción (bajo, medio o alto)
- Visualizar la importancia de las variables en el modelo.

El despliegue del modelo facilita su uso por parte de personal académico y administrativo, permitiendo aplicar el modelo en escenarios reales y apoyar la toma de decisiones estratégicas.

7. Conclusiones

El proyecto demuestra que la aplicación de la metodología CRISP-DM y técnicas de minería de datos permite abordar de manera afectiva el problema de deserción estudiantil. El modelo desarrollado ofrece una herramienta útil para la identificación temprana de los estudiantes en riesgo. Contribuyendo a la implementación de estrategias preventivas.
Como trabajo futuro, se recomienda.

- Incorporar más variables socioeconómicas y psicoeducativas
- Probar otros algoritmos de clasificación
- Mejorar el balance del dataset
- Realizar evaluaciones periódicas del modelo con nuevos datos

En conclusión, este proyecto sienta una base solida para el uso analítica predictiva en la gestión académica institucional.