Practica [3] “**Análisis Exploratorio de Datos (EDA) del Dataset OULAD**”

Por

Leandro Agustín Pilar Martínez

Grupo Reflexivo 12: Correlation Pearson

Asignación presentada a la

Escuela de Informática, Facultad de Ciencias, Como Cumplimiento de la Maestría en Ciencias de Datos e Inteligencia Artificial “Ciencia de Datos 1” (INF-7303-C1),

Dr. Silverio del Orbe Abad.

Universidad Autónoma de Santo Domingo (UASD)

Contenido

[**Introducción** ii](#_Toc200832305)

[**Metodología** iii](#_Toc200832306)

[**Resumen** iv](#_Toc200832307)

[**Conclusión** vi](#_Toc200832308)

[Bibliografía vii](#_Toc200832309)

## **Introducción**

El análisis exploratorio de datos (EDA) constituye una de las primeras fases críticas en el desarrollo de proyectos de ciencia de datos. En esta práctica se empleó el dataset OULAD (Open University Learning Analytics Dataset), el cual contiene información sobre estudiantes de educación a distancia. El objetivo fue identificar patrones que permitan comprender las variables asociadas al desempeño académico, a partir de técnicas estadísticas, visualización de datos y procesamiento previo mediante un proceso ETL. Un reto inicial importante fue la exportación del dataset a un gestor de base de datos relacional (SQL Server), ya que fue necesario crear las tablas desde cero, definir las claves primarias y foráneas, y resolver conflictos relacionados con integridad referencial y valores atípicos. Superar estas dificultades permitió estructurar adecuadamente la base de datos, garantizando su integridad y preparando el terreno para un análisis confiable. Este informe recoge los principales hallazgos, organizados bajo una perspectiva técnica pero accesible, con énfasis en su utilidad práctica y pedagógica.

## **Metodología**

Para este análisis se utilizó el lenguaje Python con bibliotecas como pandas, seaborn, matplotlib y scipy. El proceso comenzó con la importación de los archivos CSV originales del dataset OULAD y su transformación en un esquema relacional completo dentro de SQL Server. Este paso implicó retos técnicos significativos, como la necesidad de crear manualmente las tablas, definir correctamente las claves primarias y foráneas, y asegurar la correspondencia entre los tipos de datos. Además, fue necesario validar la integridad de los registros y resolver discrepancias entre los identificadores de claves externas, especialmente al relacionar las tablas studentVle y vle. Una vez consolidada la base de datos, se realizó la limpieza de valores atípicos, la codificación ordinal de variables categóricas y el análisis estadístico, incluyendo medidas de correlación y dispersión.

## **Resumen**

Los principales hallazgos del análisis se detallan a continuación, con base en evidencia empírica:

• Distribución de tipos de evaluación: Se observó una predominancia de las tareas (TMA), seguidas de exámenes (Exam) y evaluaciones continuas (CMA). Este patrón sugiere una estrategia pedagógica centrada en el trabajo continuo, posiblemente para fomentar una evaluación progresiva del aprendizaje.

• Resultados académicos: La variable `final\_result` mostró mayor concentración en las categorías “Pass” y “Distinction”, lo que podría interpretarse como un indicador positivo de desempeño o reflejo de políticas institucionales que favorecen la retención y el apoyo al estudiante.

• Relación entre créditos y desempeño: El análisis de la variable `studied\_credits` evidenció una asociación directa con el resultado final. Los estudiantes que aprobaron cursaron más créditos, lo cual puede ser interpretado como un reflejo de su nivel de compromiso académico o capacidad de carga.

• Tratamiento de outliers: La eliminación de valores atípicos utilizando el rango intercuartílico (IQR) permitió representar mejor las diferencias entre grupos en los boxplots, mejorando la claridad en la interpretación de tendencias generales sin distorsión por casos extremos.

• Correlación entre intentos previos y créditos: Se identificó una correlación moderada entre el número de intentos previos y los créditos estudiados, especialmente en perfiles con resultados de tipo “Fail”. Este hallazgo sugiere la existencia de trayectorias académicas no lineales, posiblemente asociadas a factores como la resiliencia o cambios en estrategias de estudio.

• Distribución de los créditos estudiados: La variable `studied\_credits` mostró una distribución asimétrica a la izquierda, con ligera curtosis positiva. Esto indica que la mayoría de los estudiantes cursa una cantidad elevada de créditos, mientras que pocos estudiantes presentan valores bajos, lo que podría estar relacionado con deserción temprana o una inscripción parcial.

## **Conclusión**

El análisis exploratorio realizado sobre el dataset OULAD permitió identificar relaciones significativas entre características estudiantiles y su desempeño académico. Variables como el número de créditos cursados y los intentos previos mostraron una asociación directa con los resultados finales, revelando patrones de compromiso y rendimiento. La estructuración adecuada de los datos mediante el proceso ETL facilitó la aplicación de técnicas estadísticas y visualizaciones que ofrecieron una comprensión más clara del comportamiento de los estudiantes en entornos de educación a distancia. Estos hallazgos constituyen una base útil para el desarrollo de modelos predictivos enfocados en reducir la deserción y mejorar las decisiones académicas apoyadas en datos.

## Anexo

El código fuente utilizado para el proceso ETL, análisis exploratorio y generación de visualizaciones se encuentra disponible en el siguiente repositorio de GitHub: <https://github.com/Leandroag26/Dataset_OULAD>

# Bibliografía

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset. https://analyse.kmi.open.ac.uk/open\_dataset