



DIPLOMADO EN INTELIGENCIA ARTIFICIAL APLICADA A LA INVESTIGACIÓN

Tema

DATASET OULAD ANÁLISIS Y EJERCICIOS.

Creado por

RAMÓN EMMANUEL ALVAREZ SANTANA

Fecha

29 de mayo del 2025





[DESARROLLO:](#_k75fk6ip440o)

[1.1 Enlace al cuaderno de GitHub.](#_siei2iergn8m)

[1.2 Gráficos.](#_m9xkzh4zim7d)

[1.3 Resultados clave y explicación.](#_1iq7a2r3v5tj)

[1.4 Originalidad y declaración de AI o Código de terceros.](#_13cdxvy0vqn4)

# 

### DESARROLLO:

En este proyecto se invita al análisis exploratorio de distintos *datasets* relacionados a los resultados académicos de estudiantes. Estos conjuntos de datos contienen información sobre cursos, evaluaciones, características demográficas y académicas de los estudiantes, lo cual permite identificar patrones de comportamiento, rendimiento y posibles factores que influyen en los resultados finales.

Para el desarrollo de este proyecto, se seleccionaron **tres de los datasets disponibles**, priorizando aquellos que aportan mayor contexto y detalle sobre el desempeño del estudiante:

* **studentInfo**: Este conjunto contiene información demográfica y académica de los estudiantes, como el género, grupo etario, región, nivel de estudios previos, si poseen alguna discapacidad, el número de intentos previos en el módulo, entre otros. Se utiliza para relacionar características individuales con el desempeño final del estudiante.
* **studentAssessment**: Este dataset registra las calificaciones obtenidas por los estudiantes en distintas evaluaciones (tareas, exámenes, cuestionarios, etc.). Resulta clave para entender cómo el rendimiento parcial a lo largo del curso se relaciona con el resultado final.
* **assessments**: Contiene detalles sobre cada tipo de evaluación, como la fecha de entrega, el peso asignado y el tipo de evaluación. Se utilizó para complementar la información del dataset studentAssessment y entender mejor el impacto de cada evaluación sobre el rendimiento general.

## **1.1 Enlace al cuaderno de GitHub.**

El enlace del Jupyter en Python se realizó en **CodeSpace** de **GitHub** por la facilidad de crear un repositorio de este, todo lo mencionado en este documento se podrá apreciar en forma de código desde dicho repositorio, a continuación estaremos insertando el enlace de donde se podrá visualizar la ejecución del modelo y análisis de este proyecto. [ENLACE](https://github.com/ramoneas/OULAD-Exercises/blob/main/notebooks/OULAD.ipynb)

## **1.2 Gráficos.**

Para explorar y visualizar las relaciones entre las variables disponibles, se desarrollaron múltiples gráficos utilizando la librería seaborn. A continuación se detallan algunos de los más relevantes:

* **Gráfico de dispersión (scatterplot)** entre score (nota obtenida en una evaluación) y studied\_credits (créditos que cursa el estudiante), coloreado por el resultado final (final\_result). Este gráfico permite observar si existe alguna correlación entre el compromiso académico y el rendimiento parcial.
* **Gráfico de barras (countplot)** de final\_result para conocer la distribución general de los resultados. Aquí se identificó un claro desbalance, siendo la clase **Pass** la predominante.
* **Boxplot de score según final\_result**, el cual muestra cómo varía la media y la dispersión de los puntajes de evaluación entre estudiantes que aprobaron, reprobaron, se retiraron o destacaron.
* **Gráfico de barras apiladas** mostrando la cantidad de estudiantes con y sin discapacidad distribuidos según su resultado final, permitiendo analizar si esta variable tiene relación con el rendimiento.
* **Gráfico de líneas** para observar cómo varía la calificación media (score) a lo largo del tiempo (date), con el objetivo de identificar tendencias temporales en el rendimiento.

## **1.3 Resultados clave y explicación.**

Luego de realizar la limpieza de datos (eliminación de valores nulos), el dataset final utilizado para análisis y modelado contenía la variable final\_result codificada de forma numérica con la siguiente distribución (valores normalizados):

* ***Pass***: 38.13 %
* ***Distinction***: 31.01 %
* ***Fail***: 21.43 %
* ***Withdrawn***: 9.41 %

Estos datos reflejan una **distribución moderadamente desbalanceada**, donde las clases 2 y 3 (presumiblemente *Pass* y *Distinction*) son más frecuentes, lo cual puede afectar el entrenamiento de modelos si no se toman medidas para equilibrar o ajustar la evaluación.

Otros hallazgos importantes del análisis exploratorio incluyen:

* Se observó una **tendencia positiva entre los créditos estudiados (studied\_credits) y las calificaciones (score)**. Los estudiantes con mayor carga académica tienden a obtener mejores resultados, posiblemente debido a una mayor experiencia, compromiso o familiaridad con el entorno educativo.
* El análisis de **boxplots de score según el resultado final (final\_result)** indicó que los estudiantes con mejor desempeño (categorías 2 y 3) presentan puntajes más altos y menos dispersos. En cambio, las clases con peores resultados (categorías 0 y 1) tienen puntuaciones más bajas y dispersas.
* La **variable disability**, aunque poco frecuente, mostró una ligera relación con resultados menos favorables. Se recomienda un análisis más profundo para evaluar si existen barreras académicas o de accesibilidad.
* Se verificó que algunas evaluaciones con **mayor peso (weight)** en el plan de evaluación influyen de forma más fuerte en el resultado final. Esto refuerza la importancia de monitorear el desempeño de los estudiantes a lo largo del curso y no solo en las evaluaciones principales.

## **1.4 Originalidad y declaración de AI o Código de terceros.**

Este proyecto fue desarrollado por **Ramón Emmanuel Alvarez Santana**. Se utilizaron técnicas de IA como **ChatGPT** para guiar el proceso de exploración y modelado. Las fuentes externas utilizadas incluyen el dataset OULAD del repositorio [ElizabethSeidle/OULAD](https://github.com/ElizabethSeidle/OULAD) y documentación oficial de [sklearn](https://scikit-learn.org/) y [pandas](https://pandas.pydata.org/).