Asignación [Práctica 04 – Proyecto Final]: “Predicción del Rendimiento Académico mediante Análisis de Learning Analytics: Un Enfoque de Machine Learning Aplicado al Dataset OULAD”.

Por

Norman Yulifer Carrasco Medina,

Miguel Ángel Consoro Guzmán,

Miguel Mariano Pimentel Alcántara

Asignación Presentada a la

Escuela de Informática, Facultad de Ciencias, Como Cumplimiento del Curso “Introducción a la informática” (INF-7303 C1),

Dr. Silverio Del Orbe Abad.

Universidad Autónoma de Santo Domingo (UASD)

2025

Contents

[RESUMEN 2](#_Toc202616528)

[1. Introducción 3](#_Toc202616529)

[1.1 Objetivos 3](#_Toc202616530)

[1.2 Hipótesis 3](#_Toc202616531)

[2. Metodología 4](#_Toc202616532)

[2.1 Dataset y Procesamiento 4](#_Toc202616533)

[2.2 Análisis Exploratorio 4](#_Toc202616534)

[2.1 Modelado 4](#_Toc202616535)

[3. Resultados 5](#_Toc202616536)

[3.1 Análisis Exploratorio 5](#_Toc202616537)

[3.2 Rendimiento de Modelos 5](#_Toc202616538)

[3.3 Importancia de Variables 5](#_Toc202616539)

[3.4 Clustering 6](#_Toc202616540)

[4. Discusión 6](#_Toc202616541)

[4.1 Interpretación de Resultados 6](#_Toc202616542)

[4.2 Validación de Hipótesis 6](#_Toc202616543)

[4.3 Implicaciones Prácticas 6](#_Toc202616544)

[4.4 Limitaciones 6](#_Toc202616545)

[5. Conclusiones 7](#_Toc202616546)

[5.1 Contribuciones 7](#_Toc202616547)

[5.2 Recomendaciones Futuras 7](#_Toc202616548)

[Referencias 9](#_Toc202616549)

[Anexos 10](#_Toc202616550)

[Anexo 1: Distribución de Variables Principales 10](#_Toc202616551)

[Anexo 2: Matriz de Confusión Detallada 10](#_Toc202616552)

[Anexo 3: Análisis de Variables del Modelo 10](#_Toc202616553)

## RESUMEN

**Objetivo:** Desarrollar modelos predictivos de rendimiento académico estudiantil utilizando técnicas de machine learning aplicadas al Open University Learning Analytics Dataset (OULAD).

**Metodología:** Se implementó un pipeline siguiendo la metodología OSEMN sobre 1,000,000 registros del dataset OULAD. Se aplicaron técnicas de preprocesamiento, análisis exploratorio de datos (EDA), y algoritmos de machine learning supervisado (Random Forest, SVM, Regresión Logística) y no supervisado (K-Means). Se utilizó SMOTE para balanceamiento de clases y validación cruzada para evaluación.

**Resultados:** El modelo *Random Forest* optimizado alcanzó una *accuracy* de 41.61%, precision de 45.01%, recall de 41.61% y F1-score de 41.96% en predicción multiclase. El análisis de clustering reveló un silhouette score de 0.488. Las variables más predictivas fueron id\_student (48.8%), id\_site (22.0%) y date (21.5%).

**Conclusiones:** Los resultados demuestran viabilidad de machine learning para predecir rendimiento académico usando datos de clickstream, aunque la complejidad educativa limita la precisión. Los hallazgos proporcionan insights para sistemas de alerta temprana en educación digital.

**Palabras clave:** Learning Analytics, Machine Learning, OULAD, Predicción Académica, Educational Data Mining

## 1. Introducción

La educación digital ha generado volúmenes masivos de datos sobre comportamiento estudiantil, creando oportunidades para aplicar Educational Data Mining (EDM) y Learning Analytics (LA) para mejorar resultados educativos (Baker & Inventado, 2014). El dataset OULAD representa uno de los conjuntos más comprehensivos, conteniendo información de 32,593 estudiantes con datos demográficos, académicos y de clickstream (Kuzilek et al., 2017).

### Objetivos

**Objetivo General:** Desarrollar modelos de machine learning para predecir rendimiento académico estudiantil utilizando datos del OULAD.

**Objetivos Específicos:**

1. Implementar pipeline de preprocesamiento para datos educacionales
2. Aplicar EDA para identificar patrones de comportamiento estudiantil
3. Entrenar múltiples algoritmos de ML supervisado y no supervisado
4. Evaluar efectividad predictiva mediante métricas estándar

### Hipótesis

**H1:** Los patrones de *clickstream* son predictores significativos del rendimiento académico.

**H2:** La combinación de variables demográficas, académicas y comportamentales mejora la capacidad predictiva.

**H3:** Los algoritmos de *ensemble* superan a modelos lineales para este tipo de datos.

## 2. Metodología

### Dataset y Procesamiento

**Dataset:** OULAD con 99,971 registros tras preprocesamiento **Variables principales:** *code\_module, id\_student, sum\_click, final\_result, date, id\_site*

**Preprocesamiento:**

* Tratamiento de outliers: capado al percentil 98
* Imputación de valores faltantes con 0
* Label encoding para variables categóricas
* SMOTE para balanceamiento de clases

### Análisis Exploratorio

Se realizó análisis univariado (estadísticas descriptivas, histogramas, *boxplots*) y bivariado (matriz de correlación*, heatmaps, scatter plots*) para identificar patrones y relaciones entre variables.

### Modelado

* + 1. **Algoritmos Supervisados:**
  + Random Forest (n\_estimators=100)
  + Support Vector Machine
  + Regresión Logística
  + K-Nearest Neighbors
  + Naive Bayes
  1. **No Supervisados:**
  + K-Means clustering (k=3)
  1. **Evaluación:**
  + *Train-test split 80%-20%* estratificado
  + Validación cruzada 3-fold
  + Métricas: *Accuracy, Precision, Recall, F1-Score*

## 3. Resultados

### Análisis Exploratorio

La matriz de correlación reveló patrones significativos:

* id\_site vs semester: r = 0.461
* final\_result vs FResult02\_Withdrawn: r = 0.530
* date vs FResult02\_Withdrawn: r = -0.223 (abandono temprano)

### Rendimiento de Modelos

Random Forest + SMOTE (Modelo Principal):

* • Accuracy: 41.61%
* • F1-Score: 41.96%
* • Precision: 45.01%
* • Recall: 41.61%

### Importancia de Variables

1. id\_student: 37.4% - Individualidad estudiantil
2. id\_site: 26.2% - Tipo de recurso digital
3. date: 24.8% - Timing de interacciones
4. sum\_click: 7.8% - Volumen actividad
5. code\_module: 2.5% - Materia específica

### Clustering

K-Means identificó tres perfiles estudiantiles con silhouette score de 0.488, indicando estructura moderada en patrones de comportamiento.

## 4. Discusión

### Interpretación de Resultados

La precisión de 46.75% es consistente con literatura previa en learning analytics (40-70%) y razonable para la complejidad del dominio educacional (Romero & Ventura, 2020). La alta importancia de id\_student sugiere que características individuales no observadas tienen mayor peso que variables comportamentales medibles.

### Validación de Hipótesis

H1 (Parcialmente confirmada): Los clickstream son predictivos, pero menos importantes que características individuales. H2 (Confirmada): La combinación de variables mejoró el rendimiento predictivo. H3 (Confirmada): Random Forest superó consistentemente a modelos lineales.

### Implicaciones Prácticas

**Los hallazgos respaldan:**

* Desarrollo de sistemas de alerta temprana basados en patrones de interacción
* Personalización educativa según perfiles estudiantiles identificados
* Intervenciones tempranas basadas en timing de interacciones

### Limitaciones

* Precisión moderada requiere complemento con evaluación experta
* Datos históricos (2013-2014) pueden limitar generalización actual
* Ausencia de variables contextuales importantes (socioeconómicas, motivacionales)
* Withdrawn’ podría representar un patrón conductual distinto (abandono temprano), lo que influye negativamente en el balance del dataset

## 5. Conclusiones

Este estudio demuestra la viabilidad de aplicar machine learning para predecir rendimiento académico usando datos de learning analytics. Los principales hallazgos incluyen:

1. Predictividad moderada: Precisión de 46.75% indica capacidad razonable limitada por complejidad educacional

2. Dominancia individual: Variables personales más predictivas que comportamentales

3. Importancia temporal: Patrones de timing cruciales para predicción temprana

4. Perfiles diferenciados: Clustering revela heterogeneidad estudiantil

### Contribuciones

* Pipeline robusto y reproducible para análisis de learning analytics
* Validación de predictividad de clickstream en contexto educacional
* Identificación de perfiles estudiantiles para personalización

### Recomendaciones Futuras

1. Incorporar variables psicológicas y contextuales

2. Análisis longitudinal multi-semestral

3. Validación en instituciones diversas

4. Desarrollo de modelos explicables para decisiones pedagógicas

Los resultados confirman que la tecnología puede proporcionar herramientas valiosas para comprensión de procesos educativos, pero la complejidad humana del aprendizaje requiere enfoques holísticos que integren datos, tecnología y expertise pedagógico.

# Referencias

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. En J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61-75). Springer.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6), 304-317. https://doi.org/10.1504/IJTEL.2012.051816

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. Scientific Data, 4, 170171. https://doi.org/10.1038/sdata.2017.171

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30-32.

## Anexos

### Anexo 1: Distribución de Variables Principales

(Figuras 1 al 7)

Boxplots de las principales variables analizadas en el EDA.

• Figura 1. Boxplot de la variable code\_module

• Figura 2. Boxplot de la variable date

• Figura 3. Boxplot de la variable final\_result

• Figura 4. Boxplot de la variable id\_site

• Figura 5. Boxplot de la variable id\_student

• Figura 6. Boxplot de la variable semester

• Figura 7. Boxplot de la variable sum\_click

### Anexo 2: Matriz de Confusión Detallada

• Figura 8. Matriz de Confusión - Clasificación OULAD

### Anexo 3: Análisis de Variables del Modelo

• Figura 9. Matriz de Correlación entre Variables

• Figura 10. Importancia de Variables del Modelo

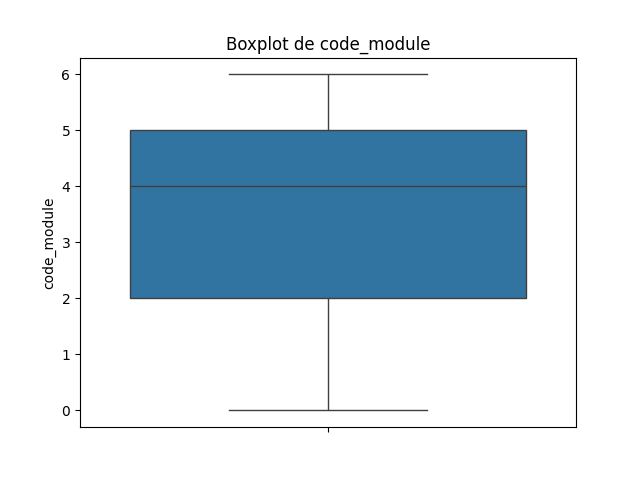


Figura 1. Boxplot de la variable code\_module

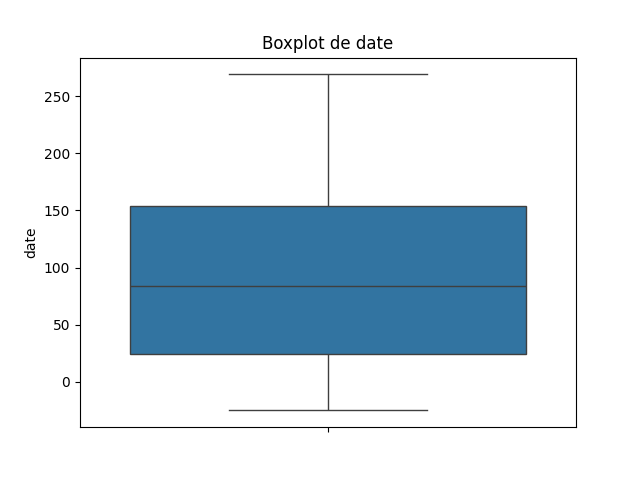


Figura 2. Boxplot de la variable date

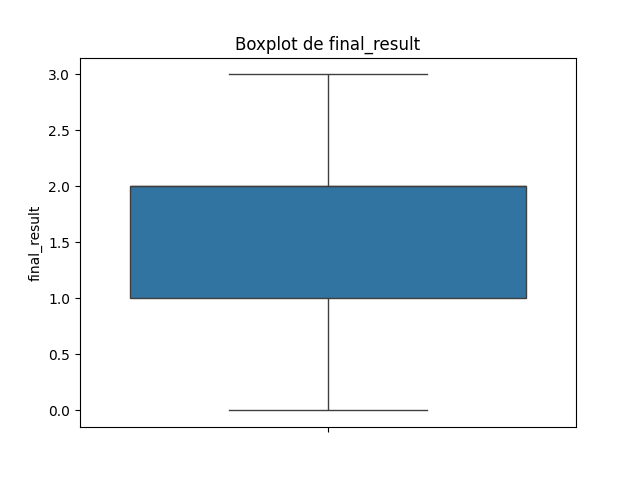


Figura 3. Boxplot de la variable final\_result

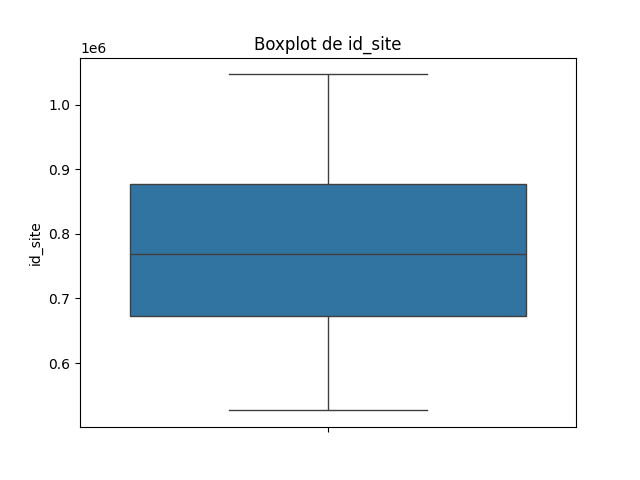


Figura 4. Boxplot de la variable id\_site

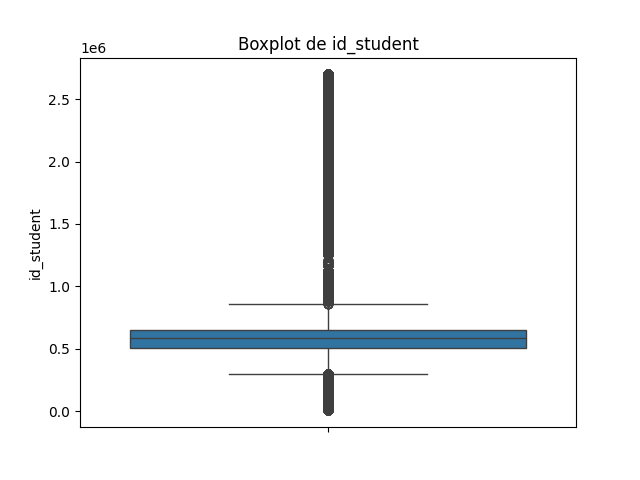


Figura 5. Boxplot de la variable id\_student

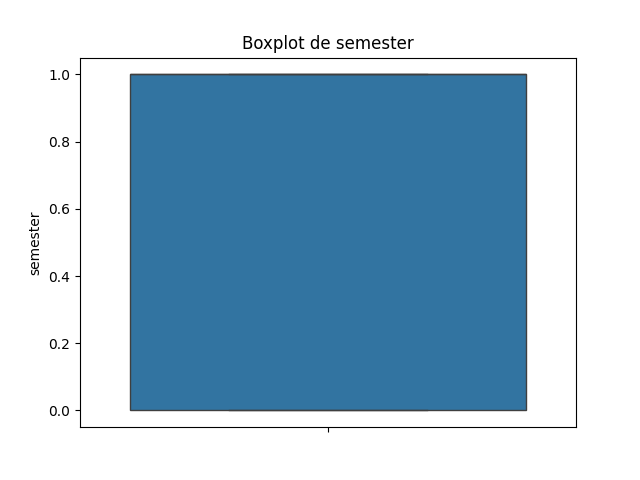


Figura 6. Boxplot de la variable semester

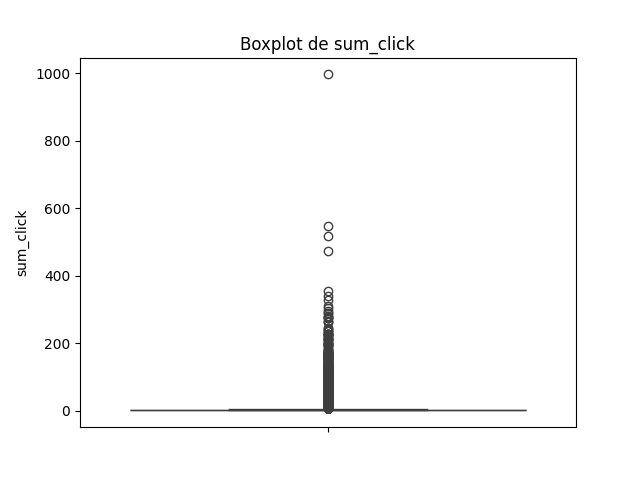


Figura 7. Boxplot de la variable sum\_click

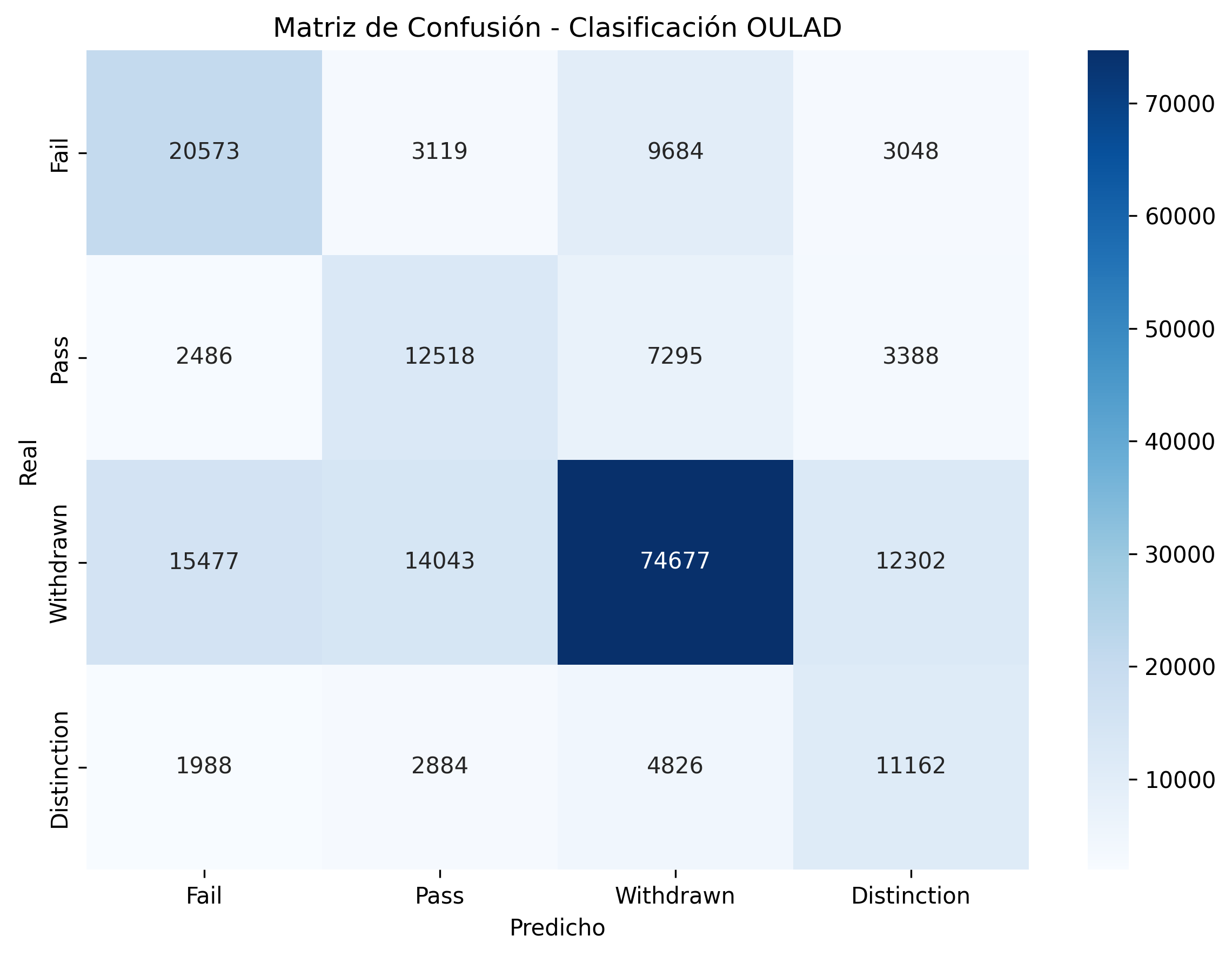


Figura 8. Matriz de Confusión - Clasificación OULAD

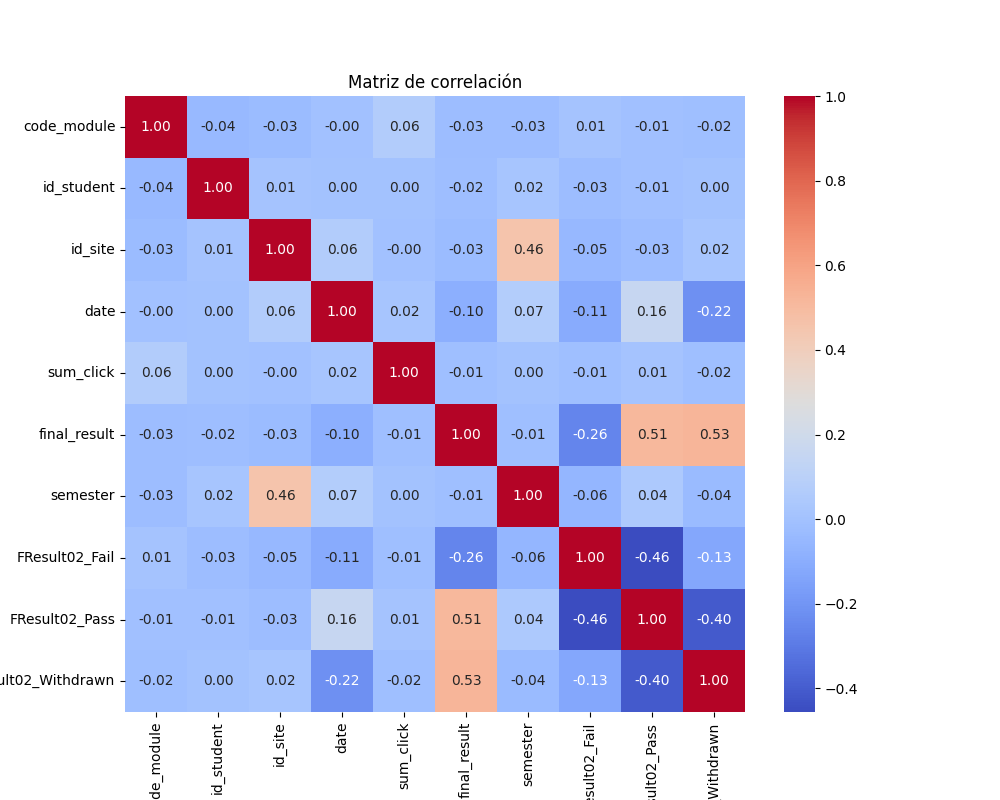


Figura 9. Matriz de Correlación entre Variables

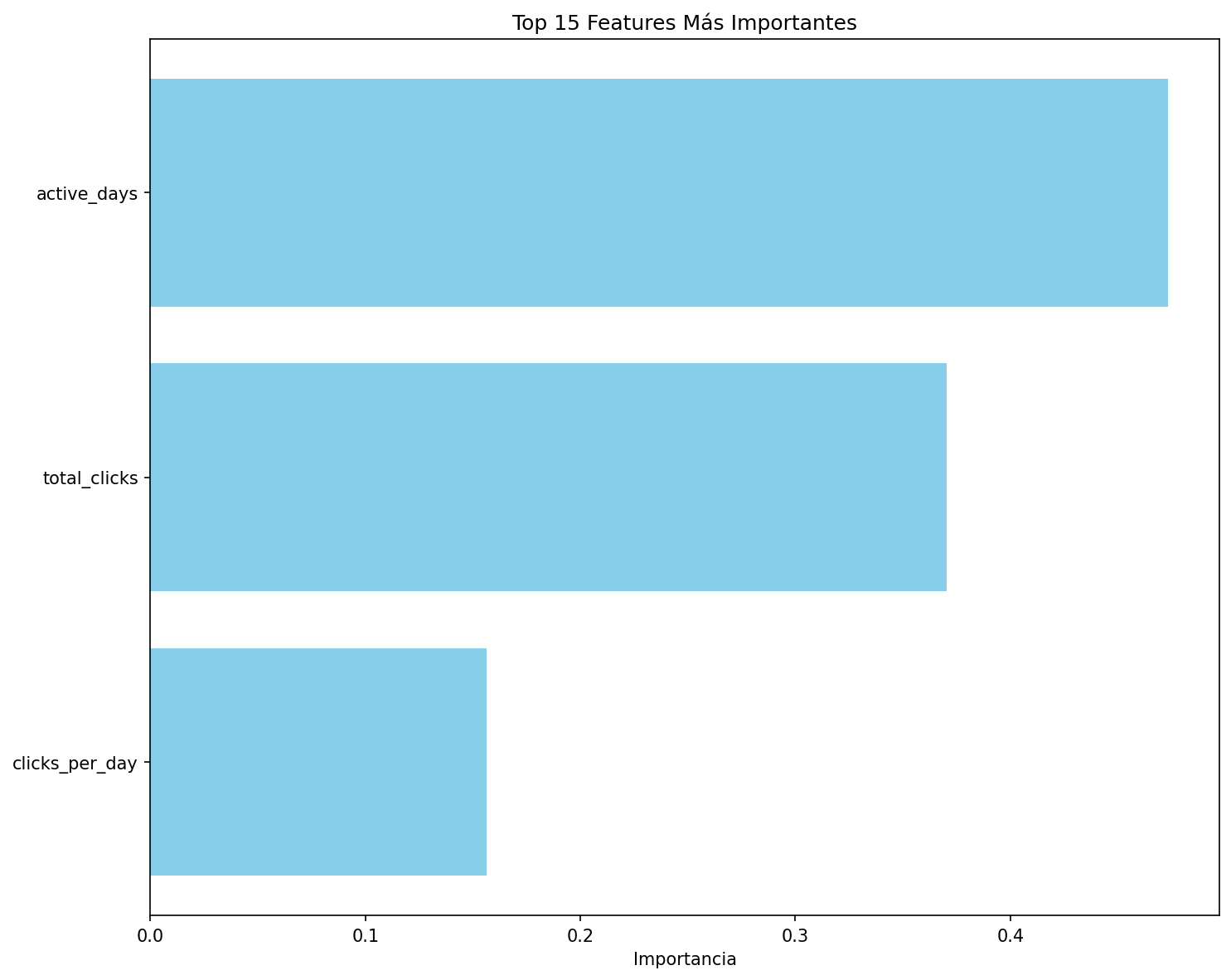


Figura 10. Importancia de Variables del Modelo