UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS DE LA DOCUMENTACIÓN

UNIVERSIDAD COMPLUTENSE DE MADRID

TITULACIÓN: Máster Data Science, Big Data & Business Analytics 2024-2025 (clase 2)

Predicción del abandono universitario a través de modelos de Machine Learning

TRABAJO DE FIN DE MÁSTER

AUTOR: Alejandro Carbonero Linares

CURSO:2024/2025

ÍNDICE

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1 Introducción del problema

El abandono académico es uno de los principales retos que enfrentan las instituciones de educación superior a nivel global. Diversos estudios muestran que entre un 20% y un 40% de los estudiantes en programas universitarios interrumpen sus estudios antes de finalizarlos, lo que supone no solo un impacto negativo para el estudiante en términos de empleabilidad y desarrollo personal, sino también para la institución, que ve afectada su tasa de retención, sus ingresos y su reputación (Research.com, 2025).

El presente trabajo tiene como objetivo desarrollar un modelo predictivo capaz de identificar desde los primeros semestres, a los estudiantes con mayor riesgo de abandono académico. De este modo, la universidad podría diseñar e implementar estrategias preventivas personalizadas (tutorías, apoyo financiero, programas de mentoría), orientadas a mejorar la permanencia estudiantil.

Para este propósito se ha utilizado un conjunto de datos proveniente de una institución de educación superior, que integra información de diferentes fuentes (académica, demográfica y socioeconómica) de estudiantes matriculados en carreras como agronomía, diseño, educación, enfermería, periodismo, gestión, trabajo social y tecnologías. El dataset recoge tanto información disponible al momento de la matrícula (edad, género, historial previo) como el rendimiento académico en los dos primeros semestres.

El problema se plantea como una clasificación multiclase con tres posibles resultados:

- Abandono (Dropout): el estudiante interrumpe sus estudios.
- Éxito (Graduate): el estudiante completa su titulación.
- Continúa (Enrolled): el estudiante sigue en el programa en el momento del registro.

Con el objetivo de simplificar el problema y centrar el análisis en la detección de abandono académico, se decidió transformar la variable objetivo original. El dataset planteaba una clasificación en tres categorías (*Dropout*, *Enrolled* y *Graduate*). Para este trabajo, se optó por eliminar la clase intermedia *Enrolled* y redefinir el problema como una tarea de clasificación binaria, distinguiendo únicamente entre estudiantes que abandonan (*Dropout*) y aquellos que logran completar sus estudios (*Graduate*).

Dada la importancia del abandono como fenómeno, el presente trabajo no se limita únicamente a entrenar un modelo predictivo, sino que también incorpora un dashboard interactivo en Streamlit, que permite tanto a académicos como a gestores explorar el perfil de los estudiantes y obtener predicciones individuales en tiempo real.

En resumen, este proyecto busca demostrar cómo la analítica de datos y el aprendizaje automático pueden convertirse en aliados estratégicos para las universidades, ayudando a anticipar riesgos, optimizar recursos y aumentar la tasa de éxito estudiantil.

1.2. Introducción técnica del Dataset

El conjunto de datos utilizado en este trabajo está compuesto por **4.424 registros de estudiantes**, cada uno de ellos descrito mediante **35 variables** de naturaleza diversa. Estas variables abarcan información **sociodemográfica**, **académica** y **económica**, así como indicadores de rendimiento durante el primer y segundo semestre.

- **Número de observaciones (filas):** 4.424 estudiantes.
- **Número de variables (columnas):** 35.
- **Tipos de variables:**
    - **Categóricas:** 16 (género, estado civil, tipo de beca, curso…).
    - **Numéricas enteras:** 11 (edad de matriculación, número de asignaturas inscritas/aprobadas…).
    - **Numéricas continuas:** 5 (nota promedio, tasas macroeconómicas…).
    - **Variable objetivo (target):** 1 (situación final del estudiante).

La variable objetivo ("**Target**") clasifica a los estudiantes en tres categorías:

- **Graduate** (2.209 estudiantes, 49.9%).
- **Dropout** (1.421 estudiantes, 32.1%).
- **Still enrolled** (794 estudiantes, 18.0%).

No obstante, por motivos de diseño metodológico, como hemos mencionado anteriormente, en este proyecto se ha optado por **simplificar la tarea en una clasificación binaria**, eliminando la categoría *Still enrolled*. De esta forma, el modelo se centra en distinguir entre estudiantes con riesgo de **abandono académico (Dropout)** y aquellos que logran la **graduación (Graduate)**.

1. FASES DEL PROYECTO

2.1. Recopilación y Preparación de Datos

El dataset original estaba compuesto principalmente por variables numéricas. Para poder comprender mejor las variables durante el análisis exploratorio y aportar mayor valor cualitativo, transformamos varias de estas variables en categóricas mediante diccionarios de codificación. Esta transformación permitió analizar patrones y relaciones de manera más intuitiva para un público no técnico.

No se detectaron valores nulos ni datos atípicos, por lo que no fue necesario aplicar técnicas de imputación o limpieza adicional.

El dataset resultante mantiene una estructura limpia y homogénea, lo que permitió centrar los esfuerzos en el análisis y modelado sin necesidad de corregir inconsistencias. Cabe destacar que existe un ligero desbalance de clases, con más estudiantes graduados que abandonos, lo cual se tuvo en cuenta al seleccionar y evaluar los modelos predictivos.

2.2. Análisis Exploratorio (EDA)

El análisis exploratorio buscamos comprender la estructura del dataset, identificar patrones, detectar desequilibrios y posibles variables predictivas, y, sobretodo,establecer un contexto para la construcción del modelo de predicción de abandono o graduación de los estudiantes.

2.2.1. Variable Objetivo: Target

En el dataset contamos con 2.209 estudiantes graduados frente a 1.421 que abandonaron, lo que indica que la mayoría completa la carrera, pero hay un número relevante de casos de abandono.

Aunque el grupo de graduados es mayoritario, la proporción de estudiantes que abandonan no es despreciable. Esto significa que cualquier herramienta predictiva debe ser capaz de identificar correctamente a estos estudiantes en riesgo, ya que detectar a los que podrían abandonar es el objetivo principal.

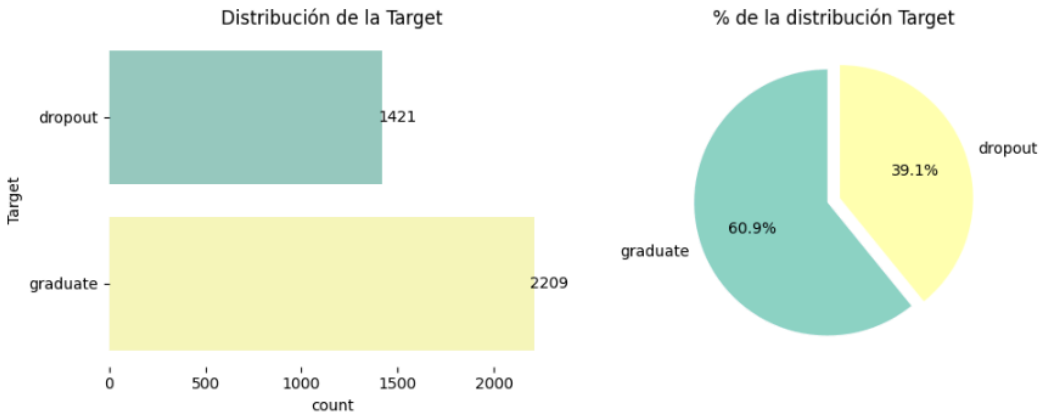

Figura 1: Distribución de la Target

2.2.2. Preprocesamiento y ajuste de variables categóricas

En este punto, para mejorar el análisis y la interpretación decidimos tomar una serie de acciones, que son las siguientes:

- **Problemas de cardinalidad:** Variables con demasiadas categorías se redujeron o agruparon cuando era necesario (ej. Application mode, Previous qualification, Mother's y Father's qualification y occupation).

- **Eliminaciones:** Variables con baja variabilidad y poco valor predictivo, como "**Nationality**", "**Educational special needs**" e "**International**", se descartaron.

- **Transformación:** Todas las variables originalmente numéricas que representaban categorías se transformaron en variables categóricas mediante diccionarios, permitiendo interpretar mejor los patrones de comportamiento.

2.2.3. Insights Variables

2.2.3.1. Variables Categóricas

Dentro del Análisis Exploratorio de Datos (EDA) nuestro objetivo no es únicamente describir las variables, sino identificar aquellas que aportan información relevante para explicar la variable objetivo (Target). En esta fase se analizan distribuciones, sesgos y posibles correlaciones con el riesgo de abandono, lo que permite anticipar qué variables tendrán mayor peso en el modelo predictivo.

Para mantener un enfoque claro y práctico, en este documento presentamos principalmente **los insights más significativos y las variables con mayor capacidad explicativa**. El análisis detallado de todas las variables (incluyendo aquellas descartadas por baja representatividad o falta de poder predictivo) está completamente documentado en el notebook, donde se puede consultar de forma exhaustiva.

En las siguientes secciones se profundiza en las variables clave, justificando las decisiones de preprocesamiento y destacando patrones relevantes que podrían influir directamente en la permanencia o abandono estudiantil.

Marital Status (Estado civil): El 72% de los estudiantes son solteros, 7% casados, y el resto categorías residuales (<2%).

- El estado civil no parece tener un peso directo en el abandono, pero refleja un perfil: los solteros suelen estar en edad universitaria “tradicional”, mientras que los casados/divorciados tienden a ser estudiantes no convencionales
- **Relación con la target:** La mayoría de los abandonos ocurre entre los solteros, pero esto se explica por su volumen. Los casados y divorciados, aunque menos numerosos, tienen una **proporción más alta de abandono**, lo que sugiere mayores dificultades de adaptación.
- **Posible acción:** Los estudiantes con perfiles no tradicionales podrían beneficiarse de apoyos específicos: flexibilidad horaria, conciliación laboral-familiar y seguimiento más cercano para mejorar su permanencia.

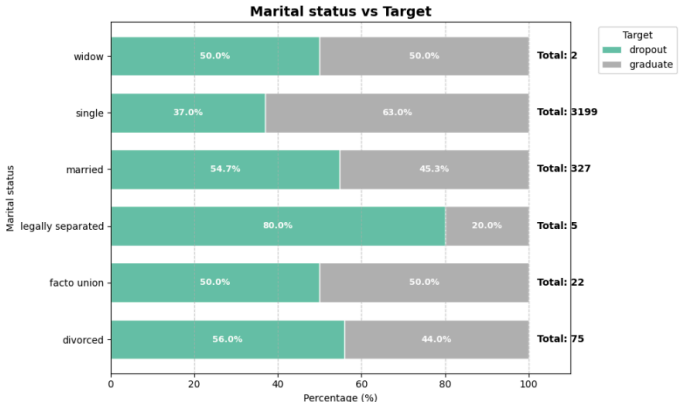

Figura 2: Graduación y abandono por estado civil

Application order: El 67% de los estudiantes están en su primera opción, y a partir de ahí las cifras caen drásticamente.

- Estudiar la primera elección podemos asociarlo a una mayor claridad de objetivos, motivación y compromiso. Los estudiantes que seleccionan opciones posteriores presentan mayor frustración y un riesgo ligeramente superior de abandono. Por ejemplo, en la primera elección hubo 1,408 graduados frente a 1,053 abandonos, mientras cuando es la 4–6 la proporción de abandonos es relativamente más alta.
- **Relación con el target:** La prioridad de aplicación refleja, de manera indirecta, la probabilidad de graduación. Elegir la primera opción aumenta la probabilidad de éxito académico, aunque no es un predictor tan fuerte como las variables económicas o de desempeño.
- **Posible acción:** Detectar desde el inicio a los estudiantes que no están en su primera elección permite ofrecer orientación académica, reencauzar intereses y diseñar intervenciones que aumenten su motivación y compromiso.

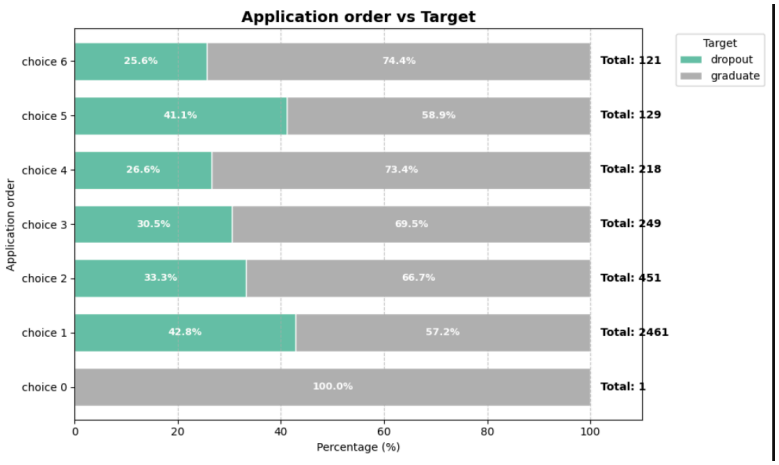

Figura 3: Graduación y abandono por orden de aplicación

Course: Destacan Nursing (18%), Social Service y Management (7–9%), mientras que ingenierías como Informática tienen <3%.

- La universidad tiene un sesgo hacia áreas sociales y de salud. El riesgo de abandono puede variar mucho entre carreras (ej. ingenierías suelen tener más deserción). Probablemente no será de gran utilidad para el modelo debido a la gran segmentación que existe y por tanto falta de información. Hemos probado a agrupar los cursos en facultades pero el resultado acabó siendo muy similar.
- **Posible acción:** Diseñar planes de retención específicos por facultad: refuerzo académico en ingenierías, apoyo vocacional en sociales, etc.

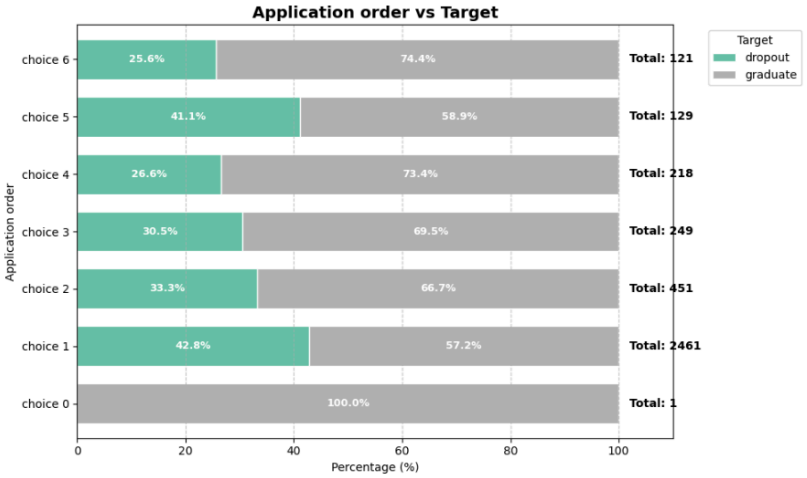

Figura 4: Graduación y abandono por orden de aplicación

Daytime/evening attendance: La gran mayoría de los estudiantes (89%) cursa en horario diurno, mientras que solo un 11% asiste en la modalidad de tarde/noche. Esto refleja un perfil predominante de estudiantes de dedicación completa. También es posible que la información esté desequilibrada en cuanto a información y esto se traslade al modelo final.

- Los estudiantes diurnos tienen una proporción de graduación significativamente mayor que de abandono, mientras que en el horario vespertino la proporción de graduados y dropouts es más equilibrada. Esto sugiere que cursar por la tarde puede asociarse a mayores desafíos, posiblemente debido a obligaciones laborales o personales.
- **Relación con la target:** La asistencia diurna se correlaciona con un mayor éxito académico y menor riesgo de abandono, mientras que la asistencia vespertina indica un grupo potencialmente más vulnerable.
- **Posible acción:** Identificar a los estudiantes vespertinos permite diseñar apoyos específicos, como tutorías flexibles, mentorías o recursos en línea, para ayudarles a superar las barreras de tiempo y aumentar sus probabilidades de graduación.

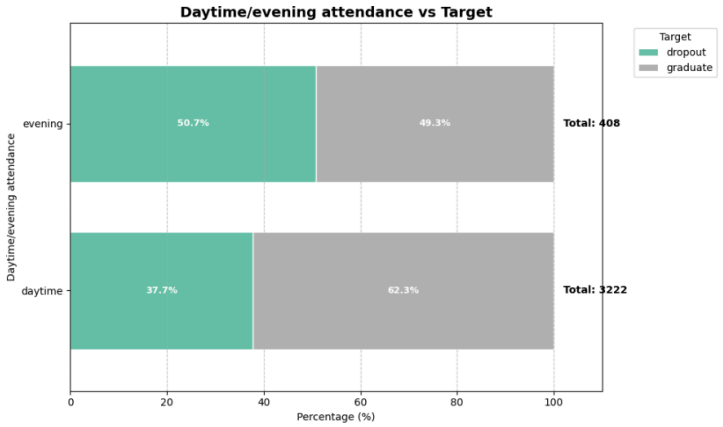

Figura 5: Graduación y abandono por momento de asistencia

Debtor: Aproximadamente el 11% de los estudiantes tiene deudas pendientes. La gran mayoría (89%) no presenta deudas, lo que indica estabilidad financiera relativa para la mayoría.

- La relación con el éxito académico es muy marcada: entre los estudiantes sin deudas, 2,108 se graduaron frente a 1,109 que abandonaron. En cambio, los estudiantes con deudas presentan un patrón inverso preocupante: 312 abandonos frente a solo 101 graduaciones.
- **Relación con la target:** Ser "deudor" es un claro indicador de vulnerabilidad: aumenta significativamente la probabilidad de abandono
- **Posible acción:** Esta variable puede ser utilizada para identificar estudiantes en riesgo desde el inicio del curso. Intervenciones como becas, planes de pago flexibles o asesoramiento financiero podrían ayudar a retener a los estudiantes

Figura 6: Graduación y abandono por tipos de deudores

<u>Tuition fees up to date:</u> El 87% de los estudiantes mantiene las cuotas al día, reflejando estabilidad económica. Solo un 13% tiene pagos pendientes.

- Existe una relación muy fuerte con el éxito académico. Los estudiantes que mantienen sus pagos al día muestran una alta tasa de graduación, mientras que los incumplimientos coinciden casi exclusivamente con el abandono.
- **Relación con la target:** Mantener las cuotas al día es un potente predictor de graduación; los estudiantes con pagos pendientes tienen un riesgo significativamente mayor de abandono.
- **Posible acción:** Estrategias como seguimiento financiero, recordatorios de pago, facilidades de pago o programas de becas pueden ayudar a mejorar la retención.

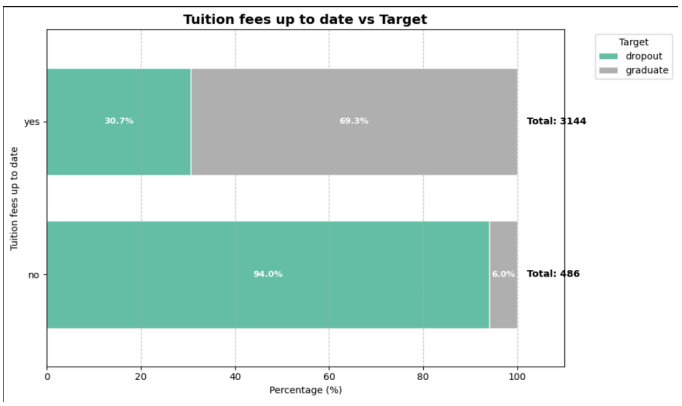

Figura 7: Graduación y abandono por momento de pago

Scholarship holder: Aproximadamente el 27% de los estudiantes recibe una beca, mientras que la mayoría (73%) no tiene apoyo económico directo.

- Los estudiantes becados presentan una proporción de éxito claramente superior: 835 graduados frente a solo 134 dropouts. En contraste, entre quienes no reciben beca, los abandonos (1,287) se acercan a los graduados (1,374), mostrando un riesgo más elevado de deserción. Esto indica que el apoyo financiero no solo facilita la continuidad, sino que también actúa como un factor protector frente al abandono.
- **Relación con la target:** Recibir una beca está fuertemente asociado con la graduación. La diferencia entre graduados y dropouts es significativa, lo que convierte esta variable en un buen predictor de éxito académico.
- **Posible acción:** Identificar a estudiantes que podrían beneficiarse de becas o apoyos financieros puede ser una estrategia clave para mejorar la retención. Además, este dato sugiere que las políticas de becas no solo alivian cargas económicas, sino que también fomentan el compromiso académico y reducen la deserción.

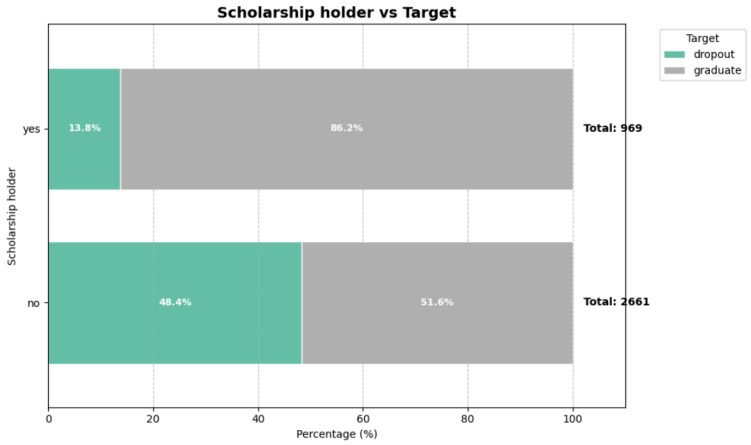

Figura 8: Graduación y abandono por estudiantes becados

2.2.3.1. Variables Numéricas

Edad al ingreso: La edad de los estudiantes muestra una distribución positivamente sesgada, con la mayoría ingresando alrededor de los 20 años, pero existiendo un grupo de estudiantes significativamente mayores (hasta 70 años).

- Los estudiantes más mayores podrían enfrentarse a responsabilidades adicionales que podrían afectar su rendimiento académico y aumentar la probabilidad de abandono. En contraste, los estudiantes más jóvenes tienden a adaptarse más fácilmente al entorno universitario.
- **Relación con el target:** La media de edad de los dropouts es 26 años, mientras que la de los graduados es 22 años. La correlación con el target es negativa (-0.267), indicando que a mayor edad al ingreso, mayor riesgo de abandono. Su correlación con la Target es de -0.267, lo que implica que a mayor edad mayor riesgo de abandono.
- **Posible acción:** La edad puede ser un indicador útil para priorizar seguimiento o soporte académico. Estudiantes mayores podrían beneficiarse de estrategias de

apoyo específicas, como tutorías flexibles, orientación sobre conciliación laboral-académica o seguimiento más cercano de su progreso.

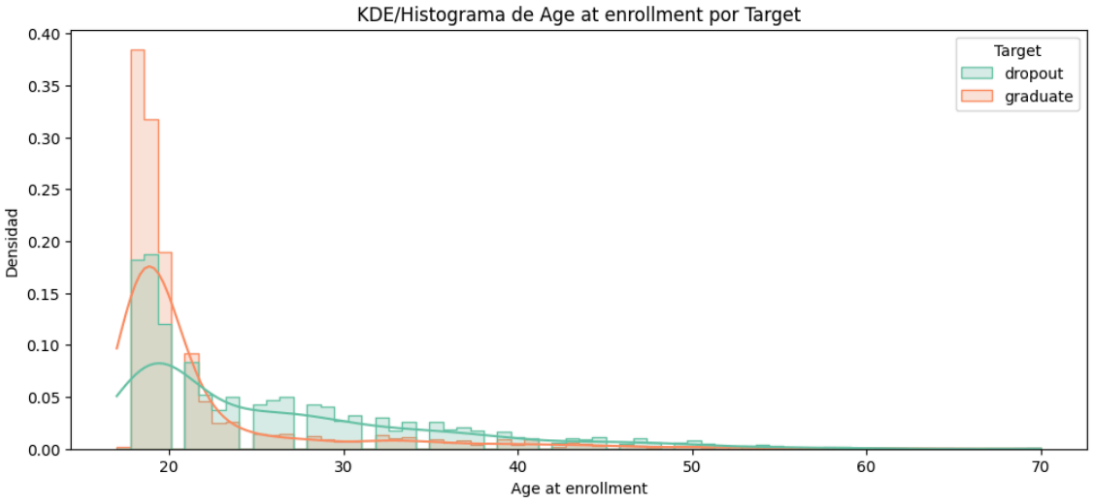

Figura 9: Graduación y abandono por edad

A continuación, se presentan los principales hallazgos relacionados con las variables de **créditos del dataset**. Para facilitar la comprensión y evitar repeticiones, se han unificado los resultados de primer y segundo semestre en un mismo bloque de conclusiones. De este modo, se obtiene una visión integrada del comportamiento de los estudiantes a lo largo del primer año académico, etapa clave para anticipar riesgos de abandono o consolidar trayectorias de éxito.

Curricular units (credited) - Asignaturas convalidadas

- Las asignaturas convalidadas muestran valores muy bajos en ambos grupos y semestres. Tanto en el primero como en el segundo semestre, la mediana es 0, lo que confirma que la mayoría de los estudiantes no tiene créditos reconocidos
- En promedio, los graduados presentan una ligera ventaja (0,85 en el primer semestre y 0,67 en el segundo) frente a los que abandonan (0,61 y 0,45 respectivamente). Sin embargo, la dispersión es alta y la diferencia pequeña
- La correlación con el éxito académico es muy baja (0,047 en 1º y 0,052 en 2º), lo que indica que este factor apenas aporta capacidad predictiva.

Insight conclusivo: las convalidaciones no parecen jugar un papel decisivo en la permanencia o graduación. Aunque quienes logran graduarse tienden a tener ligeramente más créditos reconocidos, la influencia de esta variable es marginal comparada con otras como las asignaturas aprobadas o las sin evaluación.

Curricular units (enrolled) – Asignaturas matriculadas

El número de asignaturas en las que se matriculan los estudiantes es muy similar entre semestres, y refleja una ligera pero consistente diferencia entre grupos:

- En el **primer semestre**, los graduados se matriculan en promedio en **6,67 asignaturas** (mediana 6), mientras que los que abandonan lo hacen en **5,82**.

- En el **segundo semestre**, la brecha se mantiene: graduados **≈6,63**, abandonos **≈5,78**.
  La correlación con el éxito académico es **moderada y positiva** (0,161 en 1º y 0,183 en 2º), lo que indica que matricularse en un mayor número de asignaturas se asocia con mayor probabilidad de graduación.

Insight conclusivo: El número de asignaturas matriculadas funciona como un indicador temprano de compromiso académico. Los estudiantes que se inscriben en más asignaturas desde el inicio muestran mayor probabilidad de graduarse, mientras que quienes se matriculan en menos materias presentan un riesgo más alto de abandono, posiblemente reflejando limitaciones de tiempo, recursos o menor intención de continuidad.

Curricular units (evaluations) – Evaluaciones realizadas

Distribución general:

- En el **primer semestre**, la distribución de evaluaciones está moderadamente sesgada a la derecha (1.15), lo que indica que aunque la mayoría de estudiantes tiene una carga media, existe un grupo que afronta muchas más evaluaciones (hasta 31).
- En el **segundo semestre**, la distribución es más simétrica (sesgo 0.38), con una carga más homogénea, aunque algunos alcanzan valores extremos (hasta 33 evaluaciones).

Relación con el éxito académico:

- En ambos semestres, los graduados tienden a presentar **ligeramente más evaluaciones** en promedio (8,3 en 1º y 8,1 en 2º) frente a los dropouts (7,8 en 1º y 7,2 en 2º).
- La correlación con el Target es **positiva pero débil** (0.06 en 1º y 0.12 en 2º).

Insight conclusivo: La carga de evaluaciones refleja la intensidad académica. Aunque no es un predictor fuerte por sí sola, se observa que los estudiantes que superan más evaluaciones tienden a progresar mejor. Más que el volumen absoluto, lo importante puede ser la **capacidad de afrontar esa carga**: quienes gestionan bien el segundo semestre, cuando la exigencia se estabiliza, tienen más opciones de graduarse.

Curricular units (approved) – Asignaturas aprobadas

Distribución general:

- En el primer semestre, la distribución tiene un sesgo positivo moderado (≈0.75), con la mayoría de estudiantes aprobando alrededor de 5 asignaturas.
  En el segundo semestre, la distribución es más equilibrada (sesgo ≈0.27), con un volumen de creditos aprobados algo más uniforme, aunque igualmente concentrado en torno a valores medios.

Relación con el éxito académico:
La diferencia entre graduados y dropouts es muy marcada:

- **En el primer semestre,** los graduados aprueban en promedio 6,2 asignaturas (mediana 6), mientras que los dropouts apenas alcanzan 2,6 (mediana 2).
- **En el segundo semestre,** la brecha se amplía todavía más: los graduados aprueban unas 6,2 asignaturas (mediana 6), frente a solo 1,9 en los dropouts (mediana 0)**.**
    - La correlación con el Target es muy fuerte: 0.555 en 1º semestre y 0.654 en 2º semestre, lo que convierte a estas variables en de las más predictivas del éxito académico.

Insight conclusivo: El rendimiento medido en créditos aprobados es probablemente el indicador más claro de permanencia. Ya en el primer semestre, un bajo número de aprobados anticipa riesgo de abandono, y en el segundo semestre la diferencia se convierte en abismal. En la práctica, los estudiantes que no logran aprobar al menos la mitad de las asignaturas en los dos primeros semestres quedan en una situación crítica de deserción.

Curricular units (grade) – Calificaciones obtenidas

- **Distribución general:**
  Primer semestre: sesgo negativo fuerte (≈ -1.45), con la mayoría de estudiantes obteniendo notas altas, aunque existen casos de calificaciones muy bajas**.**
- **Segundo semestre:** sesgo negativo moderado (≈ -1.17), reflejando que la mayoría sigue manteniendo un buen desempeño.

Relación con el éxito académico:

- **En el primer semestre,** los graduados tienen un promedio de 12,64 puntos (mediana 13), mientras que los dropouts apenas alcanzan 7,26 (mediana 10,93).
- **En el segundo semestre,** la diferencia se mantiene e incluso se amplía ligeramente: graduados 12,7 (mediana 13) vs dropouts 5,9 (mediana 0).

La correlación con el Target es alta: 0.52 en 1º semestre y 0.605 en 2º semestre, lo que indica que las calificaciones son un predictor sólido de graduación frente a abandono.

Insight conclusivo: Las calificaciones reflejan de manera clara la capacidad de los estudiantes para mantenerse en el programa. Los estudiantes con notas consistentemente bajas en cualquiera de los dos primeros semestres tienen un alto riesgo de abandono, mientras que quienes logran notas altas tienden a completar sus estudios. Este patrón refuerza la importancia de monitorear el desempeño académico desde el inicio.

Curricular units (without evaluations) – Asignaturas sin evaluación

Distribución general:

- **Primer semestre:** sesgo extremadamente positivo (≈ 8.72) y kurtosis muy alta (≈ 101.7), indicando que casi todos los estudiantes tienen 0 unidades sin evaluaciones, con unos pocos casos extremos.
- **Segundo semestre:** sesgo positivo alto (≈ 7.62) y kurtosis elevada (≈ 73.26), manteniendo el patrón de que la mayoría completa todas las evaluaciones.

Relación con el éxito académico:

- **En el primer semestre,** los graduados presentan un promedio de 0.088 unidades sin evaluaciones frente a 0.192 en dropouts.
- **En el segundo semestre,** los graduados tienen 0.081 y los dropouts 0.238**.**

La correlación con el Target es negativa pero baja: -0.075 en 1º semestre y -0.103 en 2º semestre, lo que indica que tener unidades sin evaluaciones puede aumentar ligeramente el riesgo de abandono, pero el efecto es marginal.

Insight conclusivo: La gran mayoría de los estudiantes cumple con sus evaluaciones a tiempo, pero los pocos que no lo hacen tienden a estar en riesgo de abandono. Aunque no es un factor decisivo por sí solo, podría utilizarse para detectar casos aislados que requieran seguimiento académico adicional o apoyo en gestión del tiempo.

2.2.4 Drops

Las siguientes variables se descartan en este punto.

Nationality - Nacionalidad

- Aproximadamente el 97,5% de los estudiantes son portugueses, mientras que el resto de nacionalidades representan menos del 3%, muchas con solo uno o ningún registro.

Educational special needs - Especiales necesidades educativas

- Solo 40 estudiantes presentan necesidades educativas especiales frente a 3,590 que no las tienen.
- Aunque es un factor relevante en términos de apoyo académico, la baja representación hace que sea difícil que el modelo aprenda patrones consistentes.

International - Internacionales

- Solo 86 estudiantes son internacionales.
- La variable aporta muy poca información debido a su baja frecuencia y, en este conjunto de datos, no parece tener impacto relevante sobre el Target.

2.3. Construcción del Modelo Predictivo

2.3.1 Transformación de variables categóricas

Todas las variables categóricas se transformaron finalmente a dummies (one-hot encoding), por varias razones:

Los algoritmos de machine learning requieren variables numéricas; la codificación one-hot permite representar cada categoría sin introducir supuestos de orden. De esta manera cada categoría conserva su propia representación, evitando que categorías nominales se interpreten como ordinales.

Se probaron alternativas, como mapas manuales para variables binarias y ordinales, pero no mejoraron el score. Esto indica que la codificación one-hot permite al modelo capturar patrones más complejos y relevantes entre las categorías.

En consecuencia, se decidió mantener dummies para todas las variables categóricas, junto con la codificación binaria para variables naturalmente binarias, lo que asegura compatibilidad, interpretabilidad y mejor desempeño.

2.3.2 Correlaciones

El análisis de correlación indica que el rendimiento académico durante los dos primeros semestres es el factor más determinante para graduarse. Los créditos aprobados del segundo semestre muestran la correlación más alta con la graduación (0,65), seguidas de las calificaciones del segundo semestre (0,61) y las del primer semestre (0,52–0,55).

La situación financiera también influye significativamente: estar al día con los pagos tiene una correlación de 0,44 y recibir beca de 0,31 con la graduación. Otros factores, como el número de asignaturas matriculadas (0,16–0,18), la elección de curso (por ejemplo, Nursing 0,21), el estado civil o el desplazamiento geográfico, aportan información adicional, aunque su impacto es menor (0,12–0,13).

En conjunto, estos resultados muestran que el éxito académico temprano y la estabilidad económica son los principales determinantes de permanencia, mientras que los factores sociodemográficos y de elección de carrera son marginales. Esto sugiere que enfocar esfuerzos en estudiantes con bajo rendimiento y dificultades financieras podría reducir de forma significativa el abandon

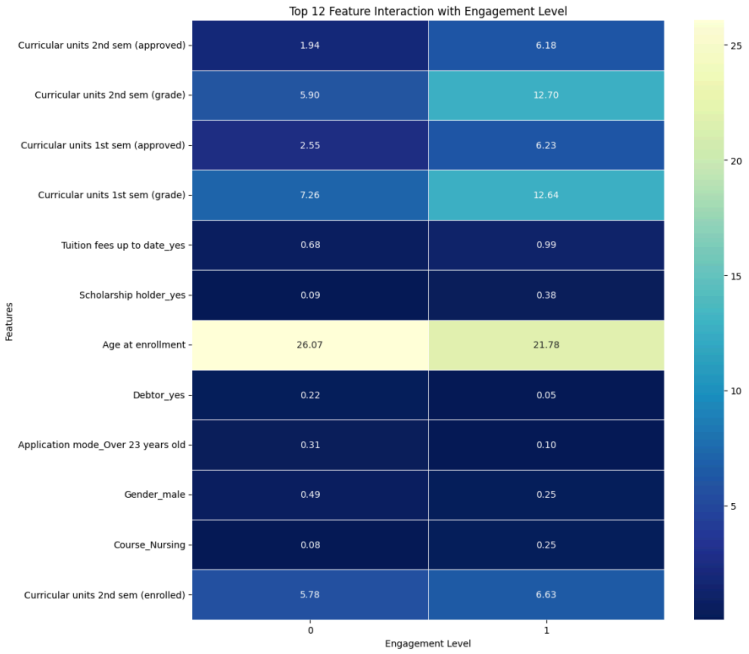

Figura 10: Top 12 correlaciones respecto a la target

2.3.3 Selección de Modelo

2.3.3.1 Modelos descartados (Primera fase)

En la primera fase de experimentación, antes de la depuración de variables, se evaluaron cuatro algoritmos: **Random Forest, Logistic Regression, Gradient Boosting y KNN**. Tras una primera revisión de resultados, se decidió descartar Gradient Boosting y KNN por las siguientes razones:

1. **KNN (K-Nearest Neighbors)**:

Presentó el rendimiento más bajo entre los cuatro modelos (accuracy ≈ 87%) y un recall especialmente reducido en la clase dropout (0.74), lo que significa que dejó de identificar una proporción significativa de abandonos reales. Además, KNN tiende a escalar mal con grandes volúmenes de datos y carece de interpretabilidad en comparación con los modelos restantes..

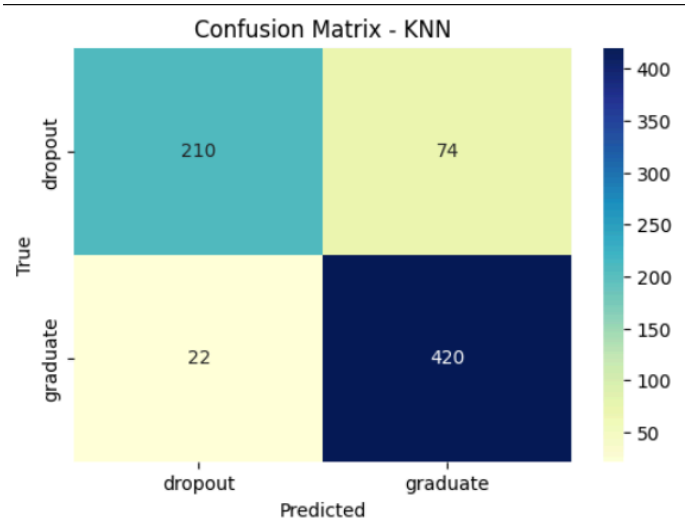

Figura 11: Matriz de confusión - Modelo KNN

2. **Gradient Boosting**:

Aunque obtuvo un rendimiento aceptable (accuracy ≈ 90%), mostró un recall más bajo para la clase dropout (0.82), lo cual es crítico dado que el objetivo principal es identificar estudiantes en riesgo de abandono. Además, el modelo es más complejo de interpretar y ajustar respecto a Random Forest y Logistic Regression.

Y, aunque es cierto que mejora sobre KNN, pero todavía queda por debajo de **Logistic Regression** y **Random Forest** en recall de dropouts (0.82) y F1-score.

Conclusión: Aunque se trata de un modelo sólido, se descarta por tener una **mayor complejidad y menor interpretabilidad**,

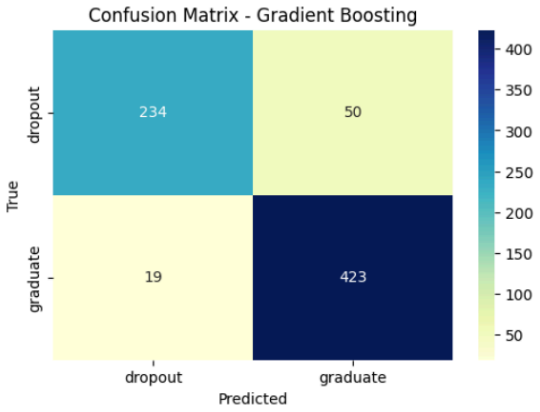

Figura 12: Matriz de confusión - Modelo Gradient Boosting

2.3.3.2 Depuración de variables

En este punto, donde podíamos comprobar un primer vistazo a los modelos, utilizamos los modelos de **Random Forest** y **Logistic Regression** para hacer una depuración de aquellas variables que solo estuviesen proporcionando ruido al modelo, y que pudiésemos descartar obteniendo los mismos resultados o incluso mejores.

1. Primera tanda de drops

Las variables eliminadas en esta primera fase se descartaron principalmente por **baja relevancia en la predicción del modelo**:

- La categoría de **Marital status** mostraba coeficientes cercanos a cero en Logistic Regression y mínima importancia en Random Forest, indicando una contribución prácticamente nula.

- **Application order_choice** y **Previous qualification** con baja representación o muy poca correlación no aportaban información significativa.

- La variable de **Cursos** y la variable **Daytime/evening attendance** también presentaban importancia residual.

Esta limpieza inicial permitió reducir el número de variables irrelevantes, simplificar el dataset y facilitar la interpretabilidad, manteniendo el rendimiento del modelo.

2. Segunda y definitiva tanda de drops

En esta segunda fase se aplicó un filtrado más exhaustivo y sistemático, basado en la **importancia de variables en Random Forest (<0.01)** y en **coeficientes casi nulos en Logistic Regression**.

- Se eliminaron **todos los cursos con relevancia marginal**, así como las categorías de **Application order_choice** y **Application mode** que no tenían peso predictivo.

- También se descartó **Marital status** y varias relacionadas con las **ocupaciones o estudios de los padres**, que mostraban impacto residual.

- Además, se retiraron variables marginales como **Displaced_yes**, **Daytime/evening attendance_evening** y los indicadores **macroeconómicos** (Unemployment rate, Inflation rate, GDP), dado que no aportaban valor al rendimiento.

Con esta depuración definitiva se obtuvo un dataset más compacto y manejable, que mantiene la performance del modelo prácticamente intacta, pero mejora la **interpretabilidad** y facilita la **explicación de resultados a responsables académicos** así como la **productivización en una aplicación real**.

2.3.3.3 Elección de Modelo

Tras la depuración final de variables, se repitió la comparación entre los modelos candidatos (Random Forest y Logistic Regression). Ambos mantuvieron un rendimiento muy similar en términos globales (accuracy 92%), pero se decidió descartar Random Forest y continuar únicamente con Logistic Regression por las siguientes razones:

- **Precisión global:** Logistic Regression mantuvo una accuracy del 92%, muy similar a Random Forest con un 91%, pero con métricas de equilibrio ligeramente más consistentes entre ambas clases
- **Balance entre clases:** Logistic Regression alcanzó un recall superior en la clase dropout (0.85 vs 0.83). Dado que el objetivo principal del trabajo es identificar estudiantes en riesgo de abandono, esta métrica tiene un peso clave en la decisión.
- **Interpretabilidad:** Logistic Regression permite interpretar directamente los coeficientes de las variables y cuantificar su impacto en la probabilidad de abandono o graduación. Random Forest, aunque robusto, requiere técnicas adicionales (importancias de variables, SHAP values) que dificultan la comunicación de los resultados.

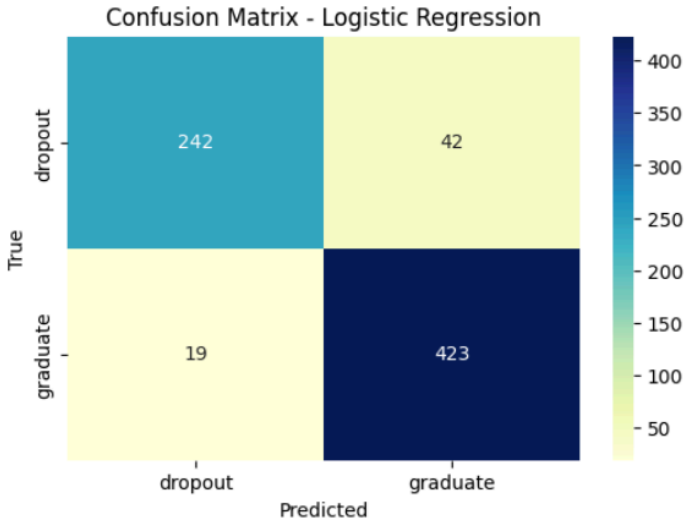

Figura 13: Matriz de confusión - Modelo Logistic Regression

3. Resultados Obtenidos

Después de obtener los resultados vistos anteriormente del modelo, podemos obtener insights más profundos y reveladores gracias a lo que el modelo nos informa sobre la importancia de cada variable. Por ello, los principales y más importantes hallazgos son:

	Feature	Coefficient	AbsCoefficient	OddsRatio
14	Tuition fees up to date_yes	2.2984	2.2984	9.9580
10	Curricular units 2nd sem (approved)	1.0538	1.0538	2.8686
13	Debtor_yes	-1.0477	1.0477	0.3507
8	Curricular units 2nd sem (enrolled)	-0.8936	0.8936	0.4092
16	Scholarship holder_yes	0.7120	0.7120	2.0380
4	Curricular units 1st sem (approved)	0.6139	0.6139	1.8475
15	Gender_male	-0.3659	0.3659	0.6936
2	Curricular units 1st sem (enrolled)	-0.2279	0.2279	0.7962
7	Curricular units 2nd sem (credited)	-0.1894	0.1894	0.8275
1	Curricular units 1st sem (credited)	-0.1569	0.1569	0.8548
12	Curricular units 2nd sem (without evaluations)	0.1541	0.1541	1.1667
6	Curricular units 1st sem (without evaluations)	0.1236	0.1236	1.1316
5	Curricular units 1st sem (grade)	-0.1117	0.1117	0.8943
11	Curricular units 2nd sem (grade)	0.0819	0.0819	1.0854
3	Curricular units 1st sem (evaluations)	0.0208	0.0208	1.0210
9	Curricular units 2nd sem (evaluations)	-0.0185	0.0185	0.9816
0	Age at enrollment	-0.0149	0.0149	0.9852

Figura 14: Tabla de importancia de variables

1. La situación financiera es el factor más determinante

Los resultados muestran que la salud financiera del estudiante es un predictor crítico de éxito académico. Mantener los pagos al día multiplica casi diez veces la probabilidad de graduación, mientras que acumular deudas reduce drásticamente las posibilidades de completar el programa. Este resultado coincide con estudios internacionales que demuestran que las ayudas financieras reducen significativamente las tasas de abandono, como se evidenció en Dinamarca tras una reforma de becas públicas (Arendt, 2012) y en Estados Unidos, donde un mayor acceso a ayudas basadas en la necesidad aumenta las probabilidades para graduarse (Goldrick-Rab, Harris, Kelchen, & Benson, 2011).

- **Pagos al día (OR = 9,96):** los estudiantes que mantienen sus tasas al corriente tienen casi diez veces más probabilidades de graduarse.
- **Deudas activas (OR = 0,35):** acumular deudas reduce la probabilidad de éxito académico a apenas un tercio respecto a quienes no tienen deudas.

Soluciones: los mecanismos de alivio financiero (planes flexibles de pago, programas de condonación parcial de deuda o becas extraordinarias) podrían tener un impacto inmediato y directo en la reducción del abandono.

2. El rendimiento en los primeros semestres define el futuro académico

Aprobar asignaturas desde el primer semestre aumenta significativamente las probabilidades de graduación, especialmente en el segundo semestre. Por el contrario, inscribirse en muchas asignaturas sin aprobarlas refleja esfuerzo no consolidado y se asocia a un alto riesgo de abandono. (González-Morales, 2025).

- **Asignaturas aprobadas (1º sem OR = 1,85 | 2º sem OR = 2,87):** aprobar asignaturas desde el inicio multiplica las opciones de graduación; el efecto es especialmente fuerte en el segundo semestre.
- **Matrículas sin aprobar (1º sem OR = 0,79 | 2º sem OR = 0,41):** inscribirse en muchas asignaturas sin aprobarlas se asocia con mayor probabilidad de abandono, reflejando un esfuerzo no consolidado.

Soluciones: el verdadero punto de control está en el primer año. Detectar desde el semestre inicial a quienes se matriculan pero no logran aprobar puede permitir activar intervenciones tempranas (tutorías, refuerzo académico, reducción de carga).

3. Las becas son un factor protector

Ser beneficiario de una beca no solo proporciona apoyo económico, sino que duplica las probabilidades de completar los estudios. Este hallazgo está alineado con investigaciones que muestran que las becas y ayudas financieras no solo fomentan la equidad, sino que también actúan como mecanismos efectivos de retención estudiantil, reduciendo las brechas de abandono por nivel socioeconómico (Chen & DesJardins, 2008).

Ser becado (OR = 2,04): recibir una beca duplica la probabilidad de terminar los estudios.

Soluciones: Las becas no solo son un mecanismo de apoyo financiero, sino también una palanca de retención institucional. Ampliar su cobertura puede ser una de las políticas más costo-efectivas para mejorar las tasas de graduación.

4. Perfil de mayor riesgo

El análisis indica que los hombres y los estudiantes que ingresan a mayor edad presentan tasas de graduación más bajas. Esto sugiere la necesidad de diseñar planes de acompañamiento específicos para estos segmentos, con estrategias adaptadas que puedan incluir flexibilidad académica, seguimiento personalizado y programas de integración, reduciendo así el riesgo de deserción.

- **Género masculino (OR = 0,69):** los hombres tienen un 30% menos de probabilidades de graduarse frente a las mujeres.un patrón que coincide con hallazgos de estudios previos, los cuales reportan consistentemente mayores tasas de abandono en varones frente a mujeres en distintos contextos universitarios (Frontiers in Education, 2024; González-Morales, 2025).

- **Edad de ingreso (OR = 0,98 por año):** a mayor edad de matrícula, ligeramente menor es la probabilidad de completar los estudios.

Soluciones: estos perfiles requieren planes de acompañamiento diferenciados. Para estudiantes mayores, mayor flexibilidad horaria; para hombres, programas de integración y seguimiento que reduzcan la deserción.

5. Las notas y evaluaciones importan menos de lo esperado

El simple hecho de presentarse a exámenes o la nota promedio no es tan decisivo como lograr aprobaciones concretas. La capacidad de avanzar en créditos y aprobar asignaturas de manera sostenida es el verdadero indicador de éxito académico. Esto implica que la institución debe enfocarse más en asegurar la progresión académica que en evaluar solo el desempeño puntual en pruebas.

- **Calificaciones promedio (1º sem OR = 0,89 | 2º sem OR = 1,08):** las notas influyen, pero en menor medida que el hecho de aprobar o no aprobar asignaturas.
- **Número de evaluaciones (coef = 0):** presentarse a más exámenes por sí solo no predice la permanencia.

Soluciones: más que medir desempeño puntual en pruebas, lo relevante es asegurar que los estudiantes acumulen créditos aprobados. La métrica de "progresión académica" (créditos logrados vs. intentados) es mucho más poderosa que la nota media como predictor de éxito.

3.1 Conclusión

El modelo demuestra que la permanencia estudiantil depende de una combinación de **factores financieros (pagos, becas), académicos tempranos (aprobaciones en los dos primeros semestres)** y **características de perfil (género, edad)**. La intervención más eficaz pasa por atacar esos tres frentes de manera coordinada: apoyo económico, acompañamiento académico inicial y políticas de retención focalizadas en grupos de riesgo.

4. Limitaciones del dataset

1. **Posibles sesgos:** Algunas variables están muy desbalanceadas (ej. nacionalidad, necesidades educativas especiales). Esto puede provocar que el modelo aprenda patrones poco representativos y generalice mal en grupos minoritarios. Dicho esto, la gran cantidad de variables hace que el modelo no se quede para nada escueto respecto a las posibilidades que brinda.

2. **Escenarios de cambio:** El modelo fue entrenado con datos históricos. Si en el futuro cambian políticas académicas (ej. criterios de becas, tasas de matrícula) o el perfil de los estudiantes, la capacidad predictiva puede deteriorarse.

5. Mejoras Futuras

1. **Enriquecer las variables disponibles:** Incorporar más información socioeconómica (nivel de ingresos familiar, empleo del estudiante) o de comportamiento (uso de plataformas online, asistencia a clases) podría mejorar la detección temprana de riesgo.
2. **Explorar modelos más complejos:** Aunque la regresión logística es interpretable y efectiva, algoritmos más avanzados como **árboles de decisión, XGBoost o redes**

neuronales podrían capturar interacciones no lineales entre variables y aumentar la precisión.

3. **Ampliar el tamaño de la muestra:** Algunas variables tuvieron que ser descartadas porque estaban demasiado segmentadas o tenían muy pocos registros en ciertas categorías (ej. nacionalidad, estado civil, modos de aplicación).

6. Productivización del modelo en Streamlit

Con el objetivo de acercar el valor del modelo a los profesionales que trabajan directamente con los estudiantes, se ha desarrollado una **aplicación en Streamlit** que permite su uso de forma sencilla e intuitiva.

- **Predicciones personalizadas**: el usuario puede introducir los datos de un estudiante (edad de ingreso, rendimiento en los primeros semestres, situación económica, etc.) y obtener en tiempo real una predicción de riesgo de abandono o probabilidad de graduación.

- **Dashboard interactivo**: además de la parte predictiva, la aplicación incluye un panel con gráficos exploratorios que permiten **comprender mejor la naturaleza del dataset** y los factores que más influyen en el éxito o fracaso académico.

- **Facilidad de uso**: la interfaz está diseñada para ser utilizada por tutores, orientadores o responsables académicos sin necesidad de conocimientos técnicos.

- **Escalabilidad**: esta primera versión sienta las bases para una futura integración en el sistema académico institucional, de manera que el modelo pueda funcionar en tiempo real y convertirse en una herramienta de apoyo en la toma de decisiones.

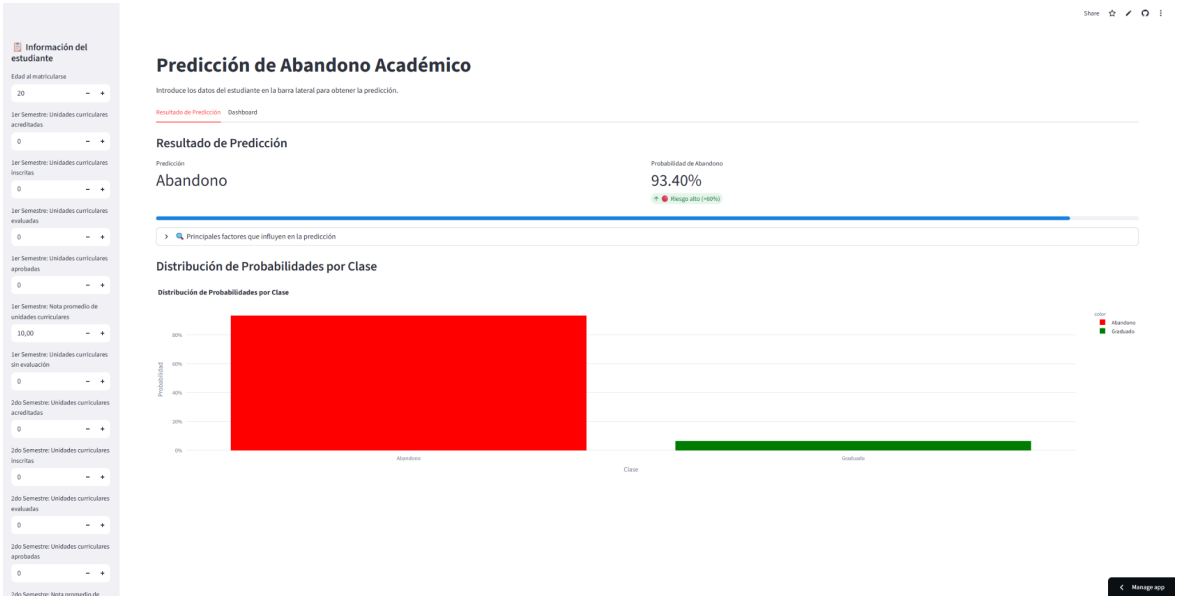

Figura 15: Interfaz Streamlit

7. BIBLIOGRAFÍA

Estudios y referencias académicas:

Estudios y referencias académicas:

- Arendt, J. N. – 2012 – *The effect of public financial aid on dropout from and completion of university education: Evidence from a student grant reform*. Springer. https://link.springer.com/article/10.1007/s00181-012-0638-5

- Goldrick-Rab, S.; Harris, D. N.; Kelchen, R.; Benson, J. – 2011 – *Need-based financial aid and college persistence: Experimental evidence from Wisconsin*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1887826

- Chen, R.; DesJardins, S. L. – 2008 – *Exploring the effects of financial aid on the gap in student dropout risks by income level*. Springer. https://link.springer.com/article/10.1007/s11162-007-9060-9

- González-Morales, M. O. – 2025 – *Dropping out of higher education: Analysis of variables that influence student dropout*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S000169182400547X

- Research.com – 2025 – *College Dropout Rates: 2025 Statistics by Race, Gender & Income*. https://research.com/universities-colleges/college-dropout-rates

8. ANEXO

El código entero junto con todas las anotaciones técnicas y mas personalizadas de todas las variables que aquí han sido descartadas por el bien de la comprensión de lo realmente importante se encuentran en la carpeta subida y/o en este link de github:

LINK DATASET KAGGLE:
https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success

LINK VIDEO EN YOUTUBE: https://www.youtube.com/watch?v=VSgb4b8eCwM

LINK STREAMLIT: (Nota: Si aparece como "Dormida", simplemente haz clic en Despertar app y en unos segundos estará disponible.)
https://tfm-predicci-n-abandono-universitario-kmdsukt8tbqmtwsrf8mappb.streamlit.app/

LINK REPOSITORIO GITHUB:
https://github.com/AlejandroCarbonero/TFM-Predicci-n-Abandono-Universitario

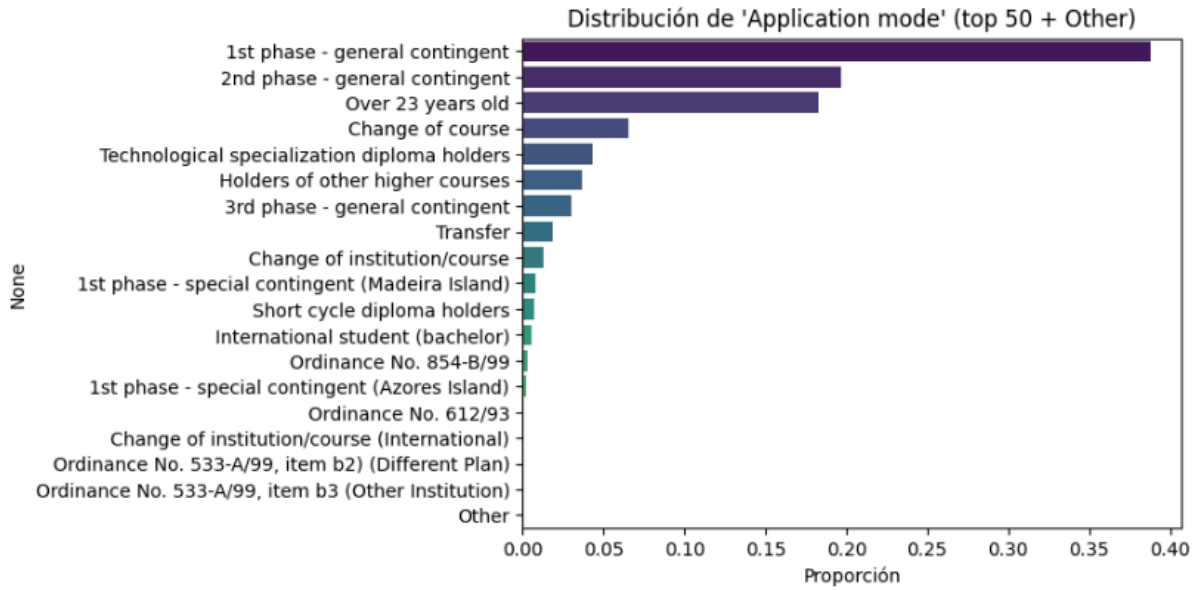

Figura 1: Distribución “Application mode” pre 1ª transformación

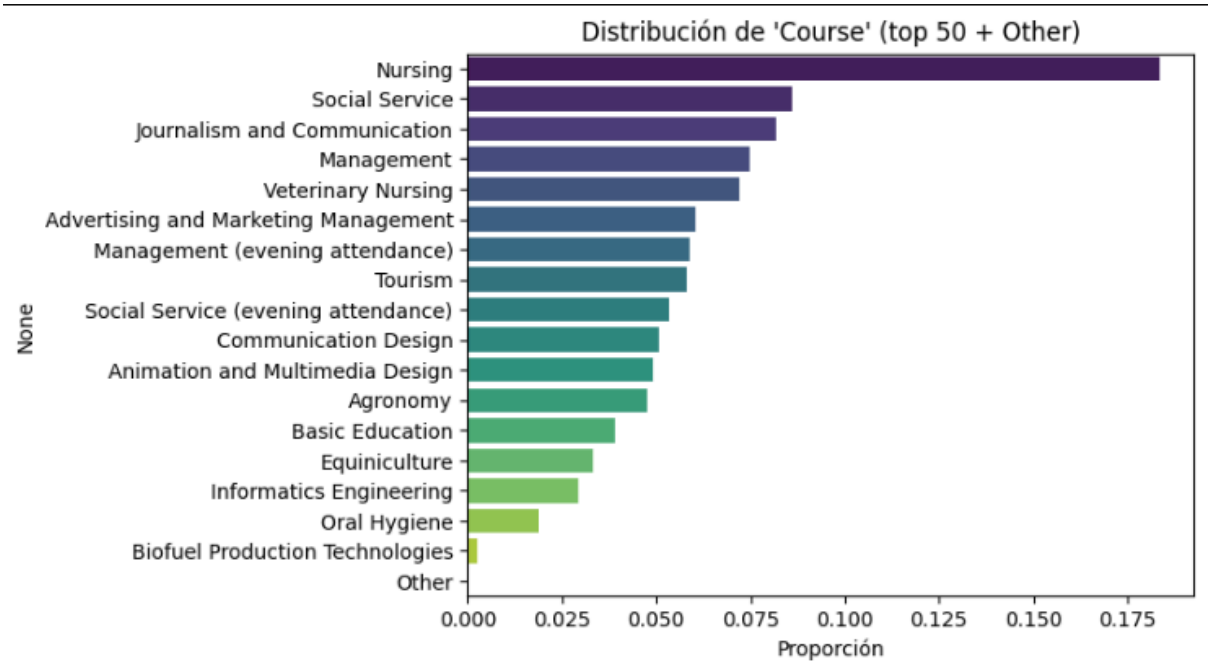

Figura 2: Distribución “Course” pre 1ª transformación

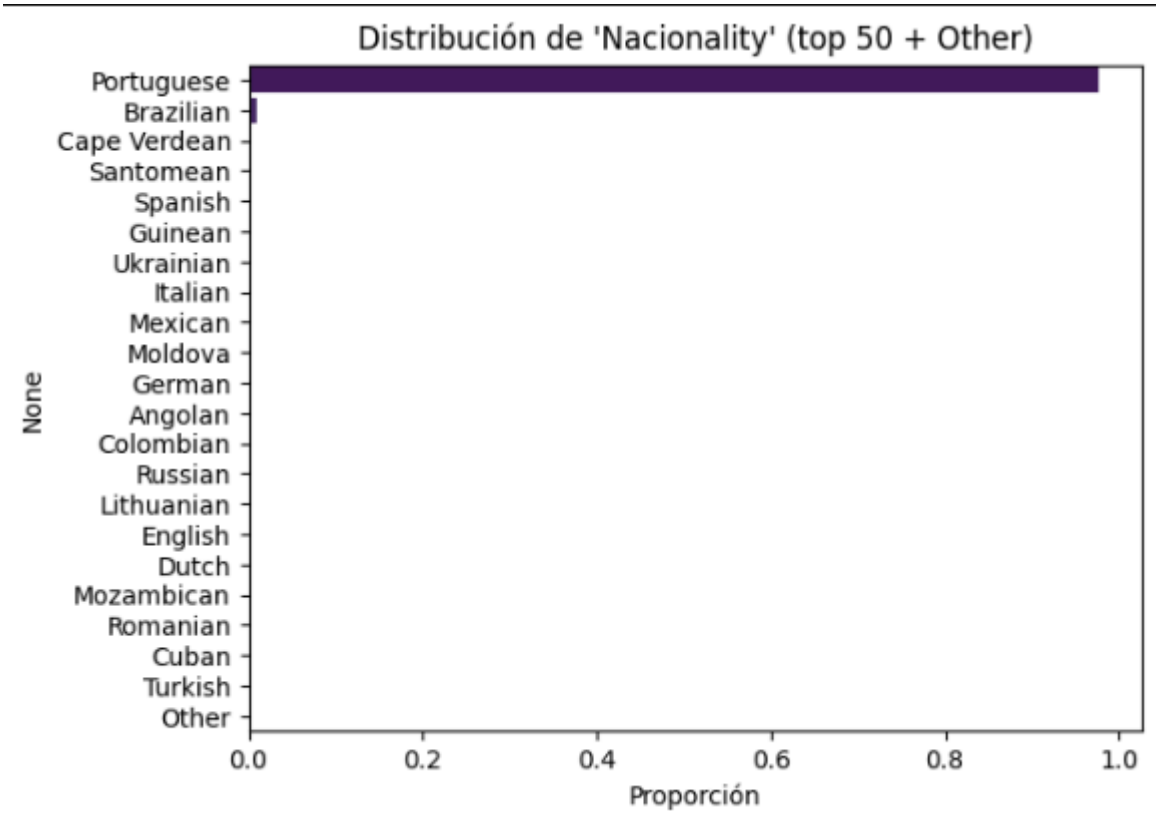

Figura 3: Distribución “Nacionality” pre eliminación

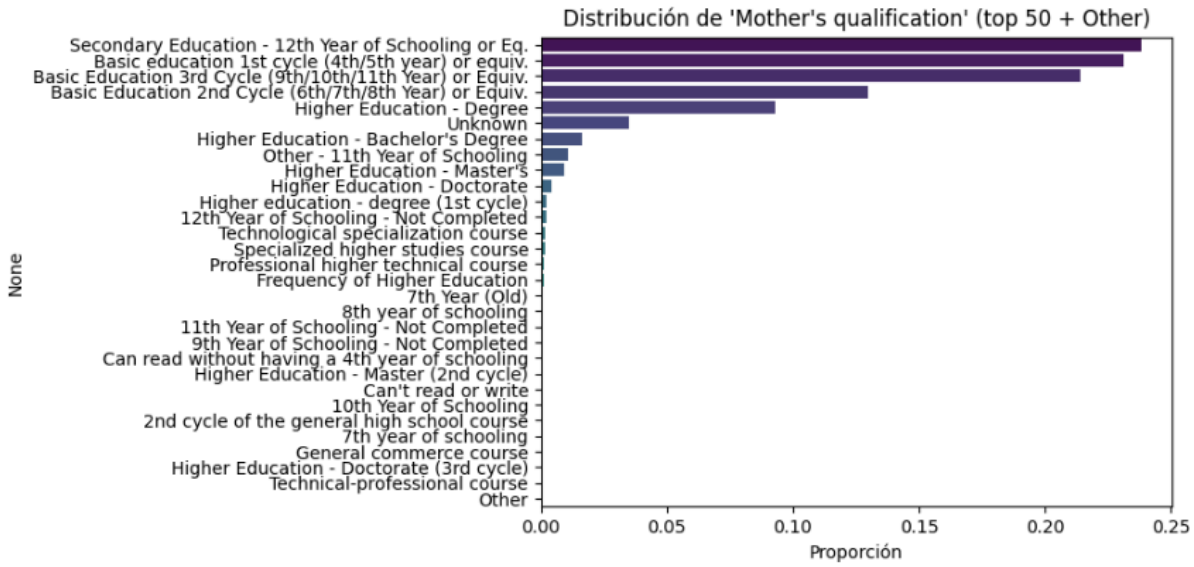

Figura 4: Distribución “Mothers qualification” pre 1ª transformación

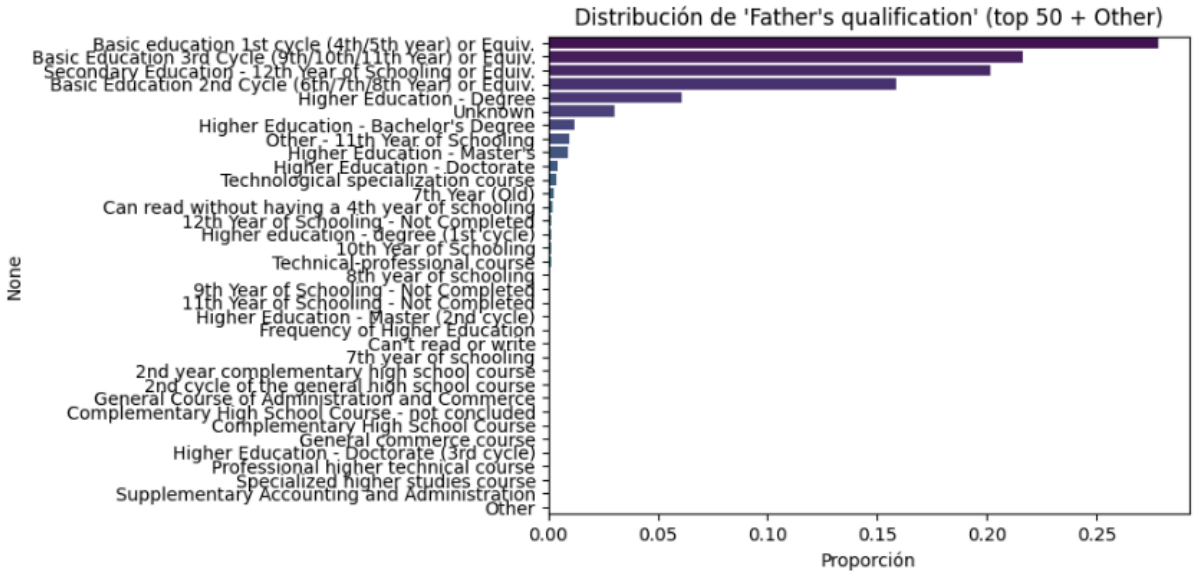

Figura 5: Distribución “Fathers qualification” pre 1ª transformación

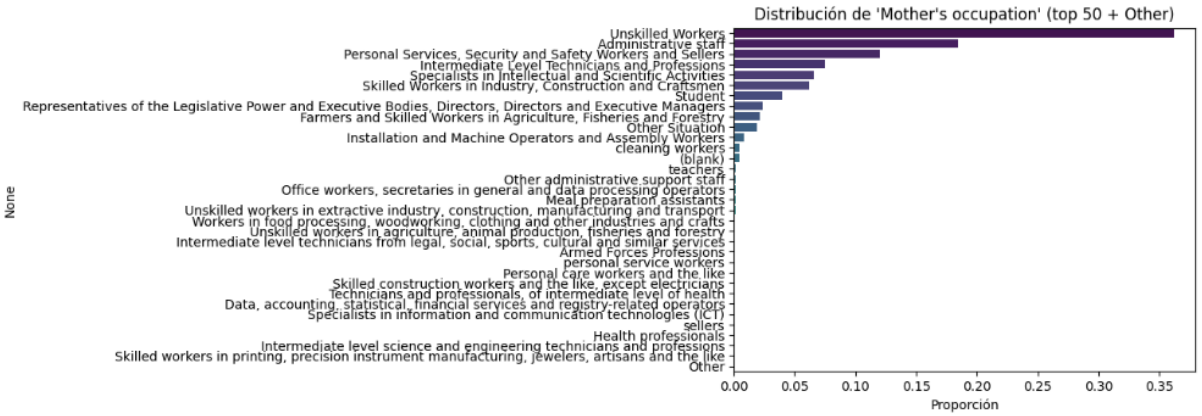

Figura 6: Distribución “Mothers occupation” pre 1ª transformación

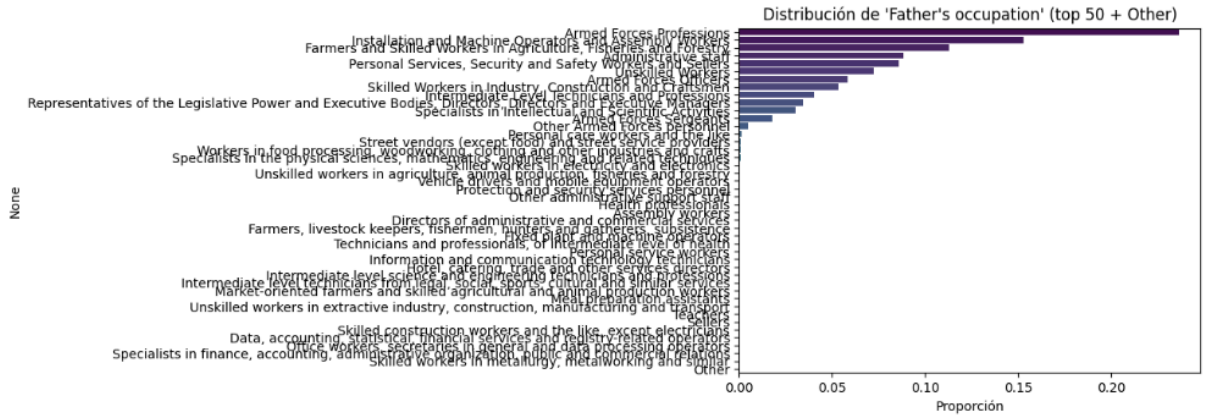

Figura 7: Distribución “Fathers occupation” pre 1ª transformación

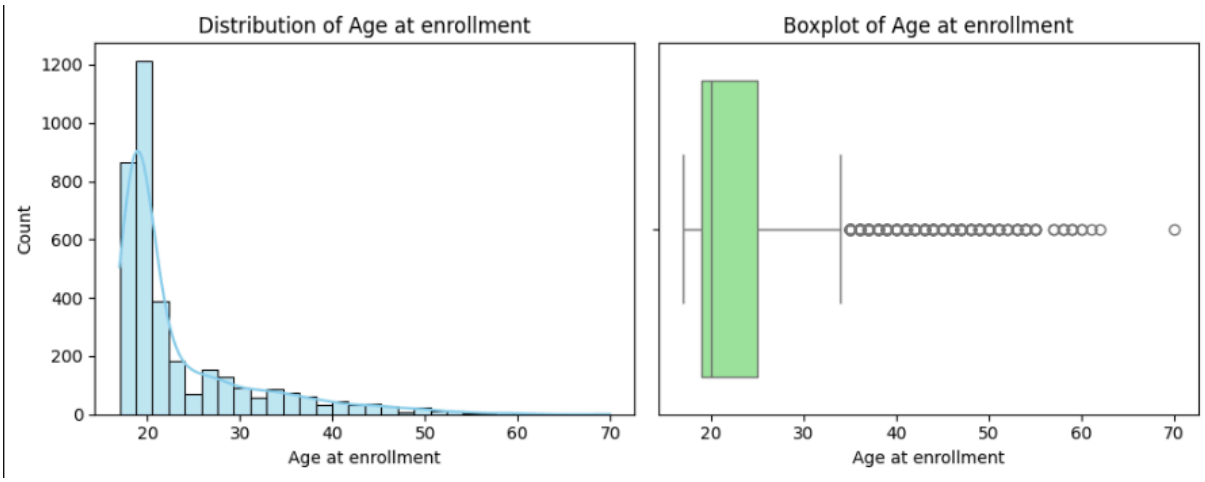

Figura 8: Distribución “Age at enrollment”

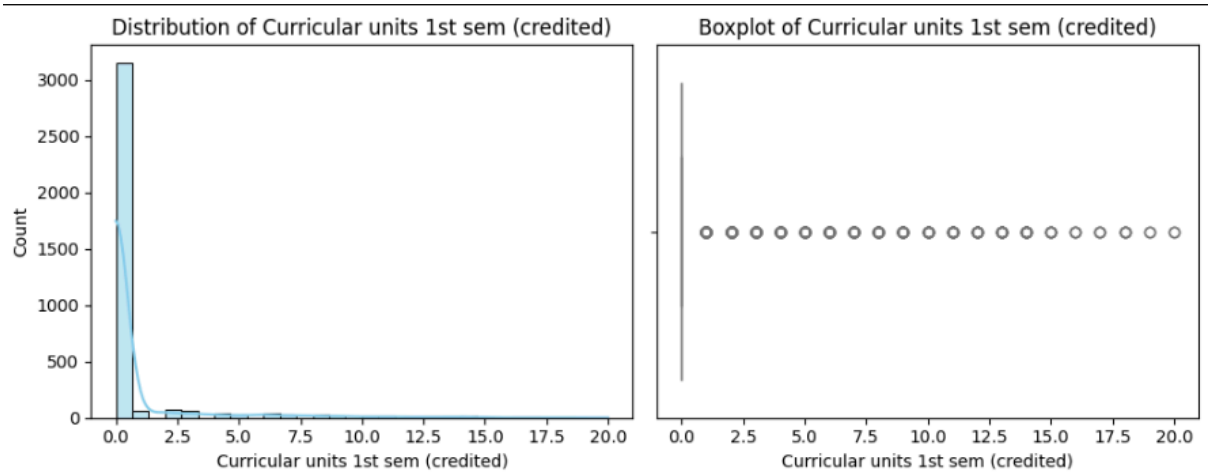

Figura 9: Distribución “Curricular units 1st sem (credited)”

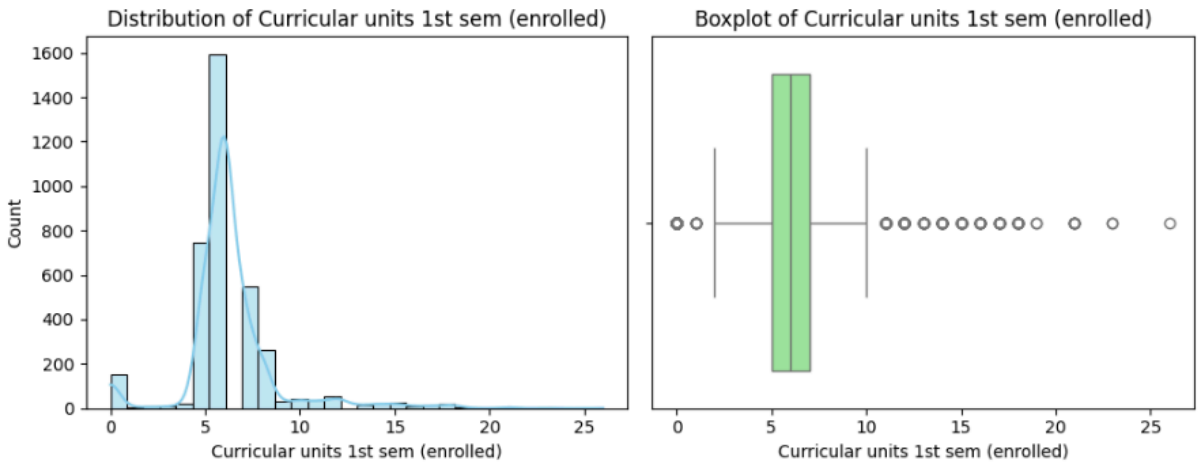

Figura 10: Distribución “Curricular units 1st sem (enrolled)”

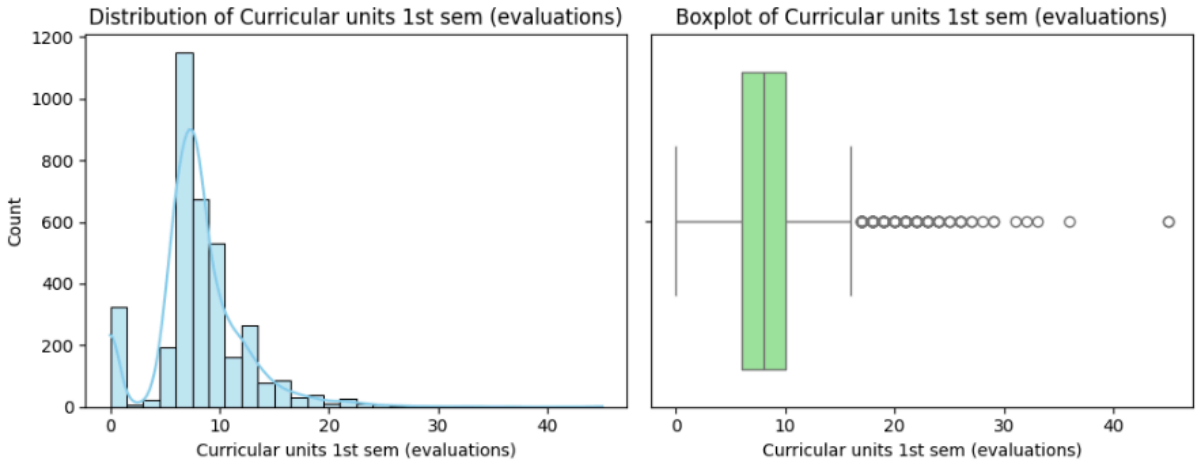

Figura 11: Distribución “Curricular units 1st sem (evaluations)”

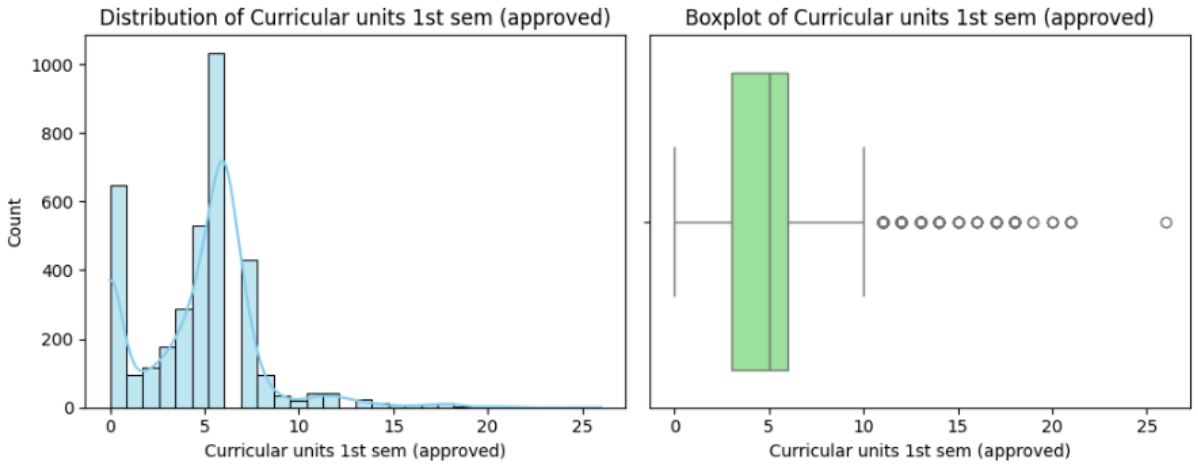

Figura 12: Distribución “Curricular units 1st sem (approved)”

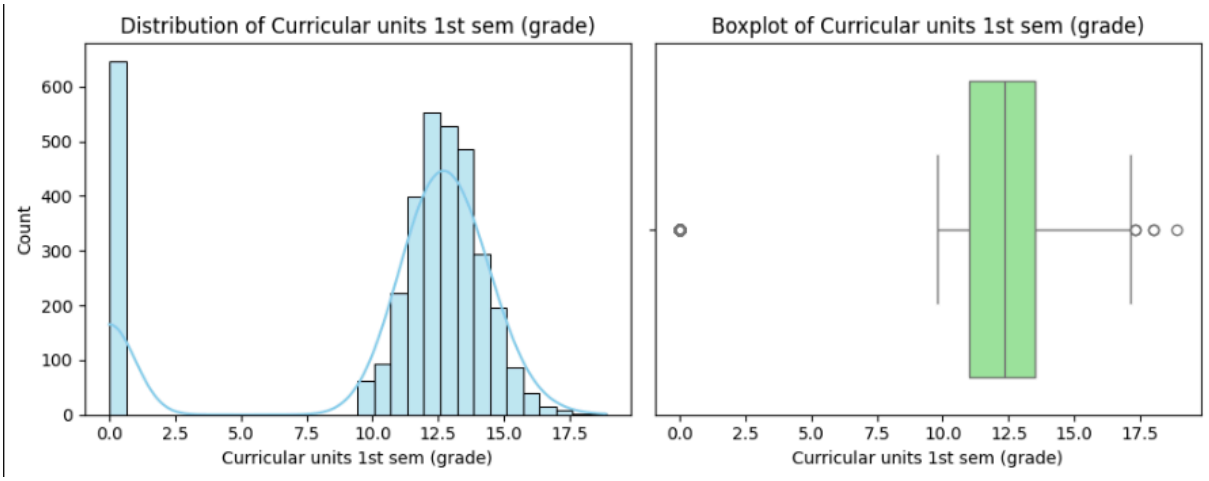

Figura 13: Distribución “Curricular units 1st sem (grade)”

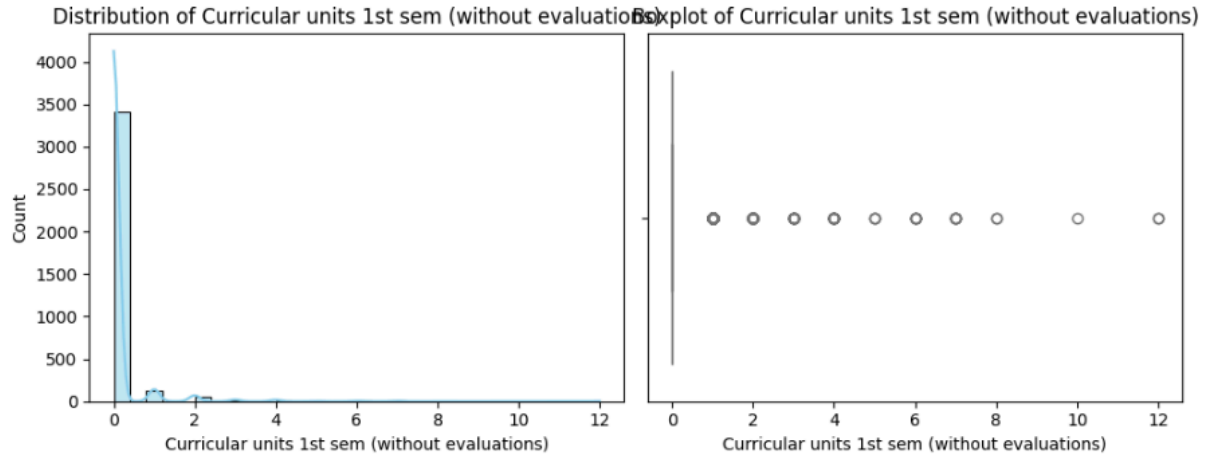

Figura 14: Distribución “Curricular units 1st sem (without evaluation)”

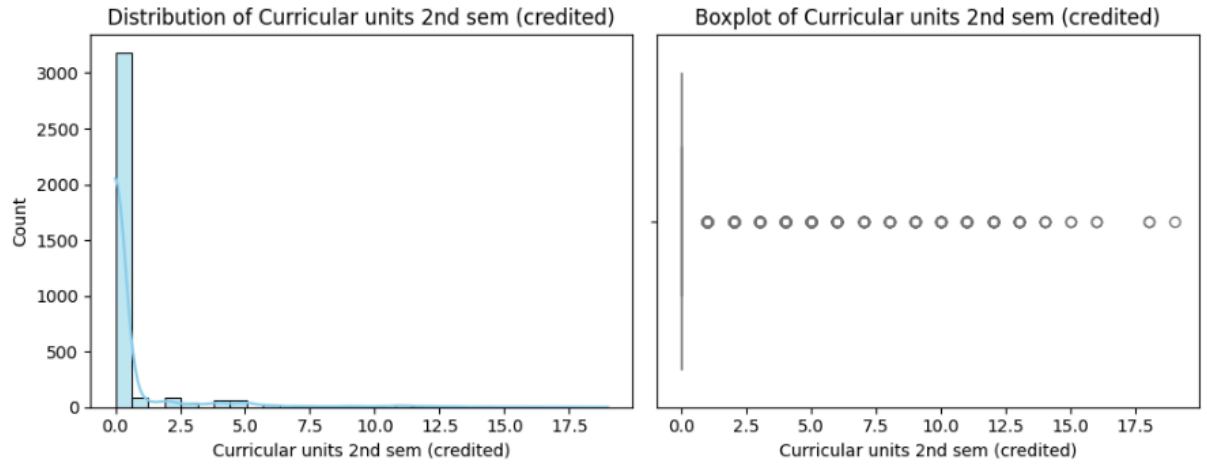

Figura 15: Distribución “Curricular units 2st sem (credited)”

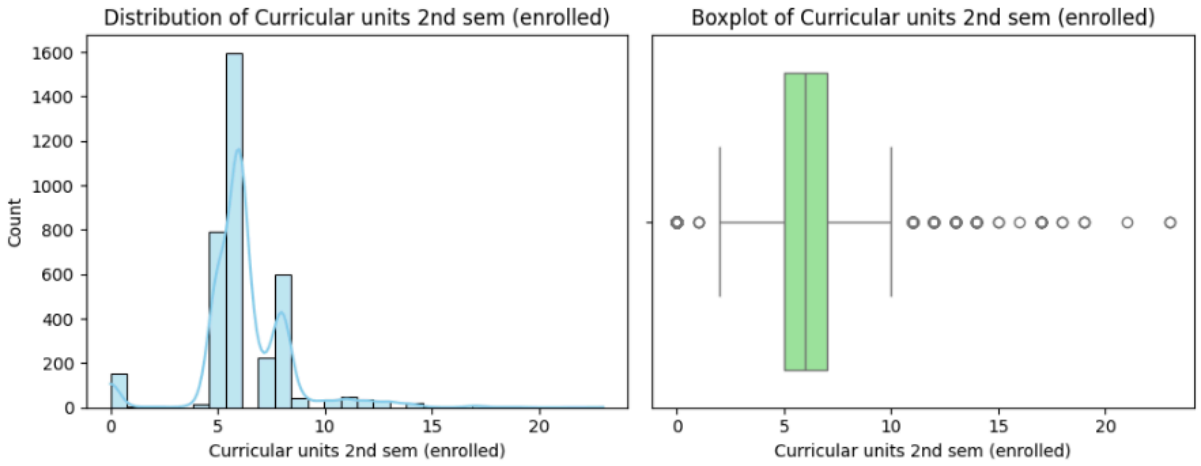

Figura 16: Distribución “Curricular units 2st sem (enrolled)”

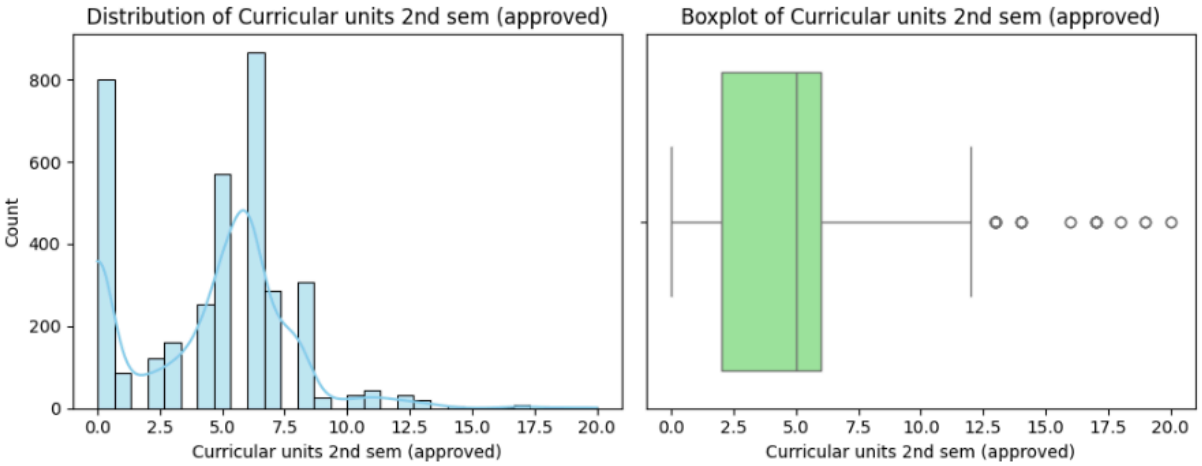

Figura 17: Distribución “Curricular units 2st sem (approved)”

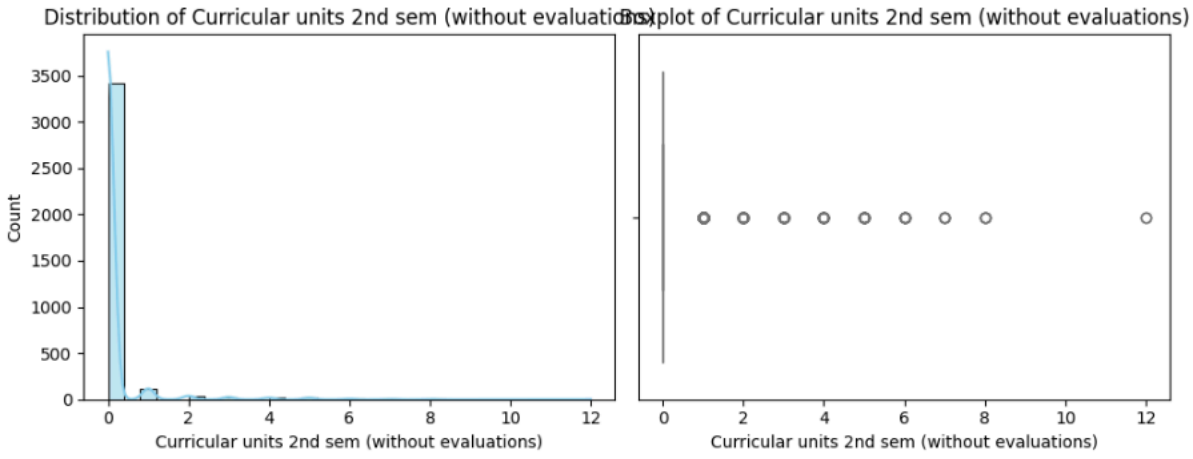

Figura 18: Distribución “Curricular units 2st sem (without evaluations)”

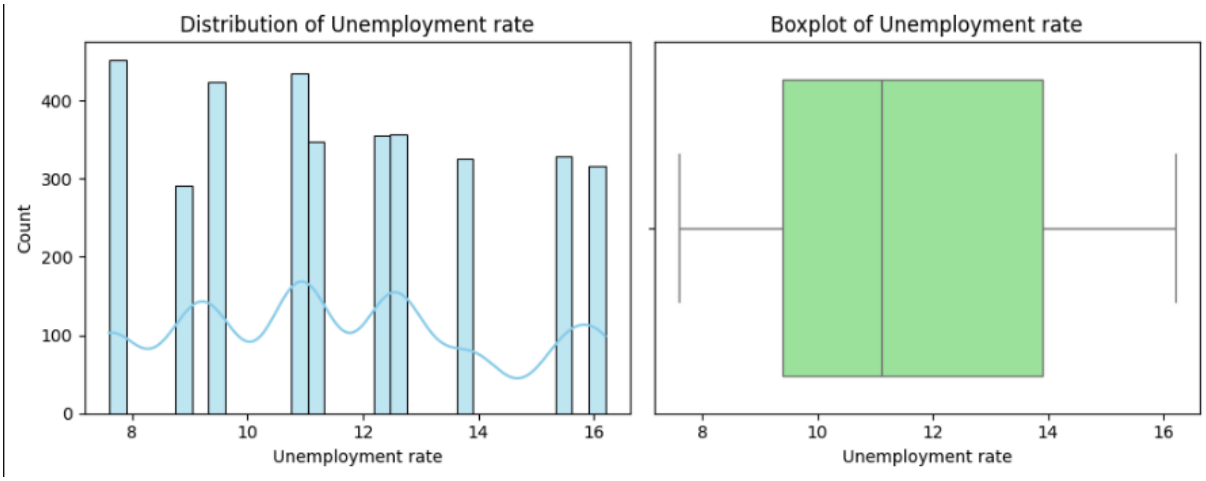

Figura 19: Distribución “Unemployment rate”

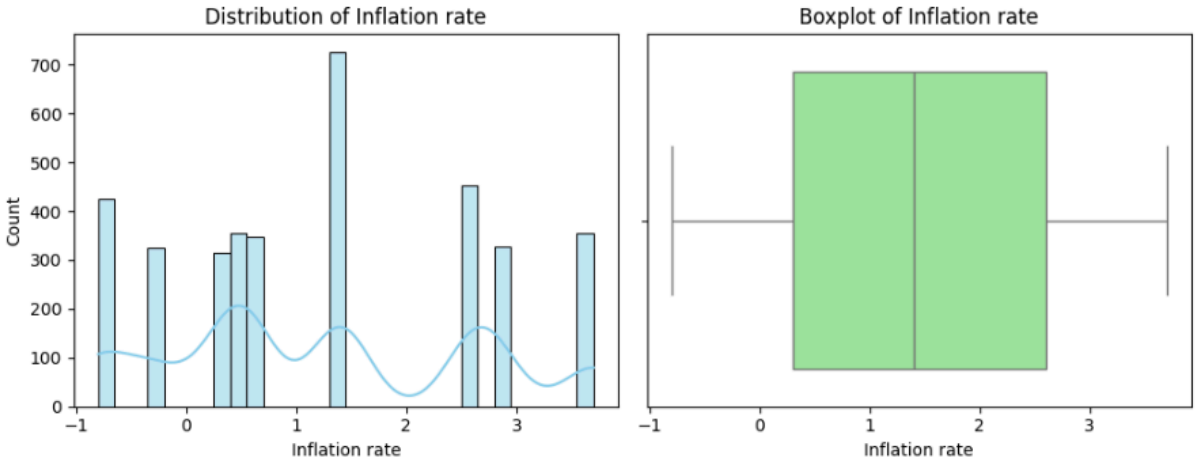

Figura 20: Distribución “Inflation rate”

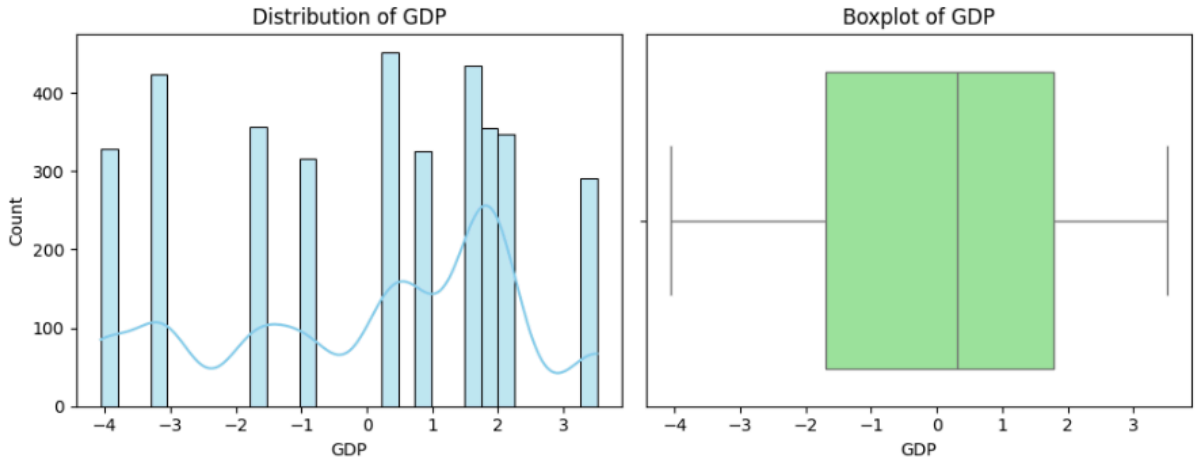

Figura 21: Distribución “GDP”