Informe del Proyecto Final: Identificación de Clientes en Riesgo de Abandono en Gimnasios

Econometría I

Jonathan Amado [14002285**]**

Universidad Galileo

Maestría en Investigación de Operaciones

Introducción

El informe detalla el desarrollo y evaluación de un modelo de clasificación diseñado para identificar a clientes de un gimnasio en riesgo de abandono. Utilizando datos demográficos, fisiológicos y hábitos de ejercicio, se implementaron tres modelos de aprendizaje automático con el objetivo de predecir la frecuencia de asistencia semanal de los clientes. El modelo seleccionado permite proporcionar información valiosa para diseñar estrategias personalizadas de retención y mejorar la experiencia de los usuarios.

Descripción General del Modelo

Se utilizaron tres modelos de clasificación para resolver el problema:

1. **Random Forest Base:** Modelo sin ajustes adicionales, diseñado para capturar relaciones no lineales en los datos.
2. **Random Forest con SMOTE:** Ajustado para balancear las clases subrepresentadas mediante técnicas de sobremuestreo.
3. **Gradient Boosting:** Modelo que mejora de manera incremental la clasificación, enfocándose en corregir errores iterativamente.

El objetivo principal fue clasificar a los clientes en tres categorías de frecuencia de asistencia:

- **Baja:** 2 o menos sesiones por semana.
- **Media:** De 2 a 4 sesiones por semana.
- **Alta:** Más de 4 sesiones por semana.

Los modelos fueron evaluados utilizando métricas como precisión, recall, F1-score y matriz de confusión para determinar su desempeño en cada clase.

Justificación del Modelo

El desarrollo de este modelo responde a la necesidad de los gimnasios de identificar a los clientes que podrían abandonar sus actividades. Al anticipar esta situación, se pueden implementar estrategias de retención como programas de motivación personalizados, lo que impacta directamente en la satisfacción del cliente y en los ingresos del gimnasio.

Los modelos seleccionados (Random Forest y Gradient Boosting) fueron elegidos por su robustez en el manejo de datos desbalanceados y complejos. Ambos permiten capturar relaciones no lineales entre las variables y la variable objetivo, siendo adecuados para el problema planteado.

Valor Generado

Este modelo genera valor al proporcionar un análisis detallado de los patrones de comportamiento de los clientes. Los resultados ayudan a los administradores del gimnasio a:

- Diseñar estrategias que aumenten la retención de clientes.
- Reducir el abandono de miembros mediante intervenciones personalizadas.
- Optimizar el uso de los recursos del gimnasio.

Además, permite personalizar las experiencias de los clientes, mejorando su compromiso y satisfacción general.

Resultados

A continuación, se presenta una tabla comparativa de los resultados obtenidos para cada modelo implementado:

Modelo	Precisión Global	Precisión (Alta)	Recall (Alta)	Precisión (Baja)	Recall (Baja)	Precisión (Media)	Recall (Media)
Random Forest (Base)	70%	61%	45%	51%	29%	74%	87%
Random Forest (SMOTE)	63%	46%	58%	42%	40%	74%	72%
Gradient Boosting	61%	46%	55%	38%	43%	73%	67%

Interpretación

- **Random Forest Base:** Se destacó por su precisión global y excelente recall en la clase predominante (Media), pero tuvo dificultades con las clases menos representadas (Alta y Baja).
- **Random Forest con SMOTE:** Mejoró significativamente en las clases menos representadas a expensas de una ligera disminución en el recall de la clase Media.
- **Gradient Boosting:** Mostró un mejor balance entre todas las clases, aunque su precisión global fue menor que la de Random Forest Base.

Gráficas y Análisis

Distribución de Categorías de Frecuencia

La siguiente gráfica muestra la distribución de las categorías de frecuencia (Baja, Media y Alta) en el dataset. Esta distribución destaca que la mayoría de los clientes pertenecen a la categoría Media, mientras que las categorías Alta y Baja están subrepresentadas.

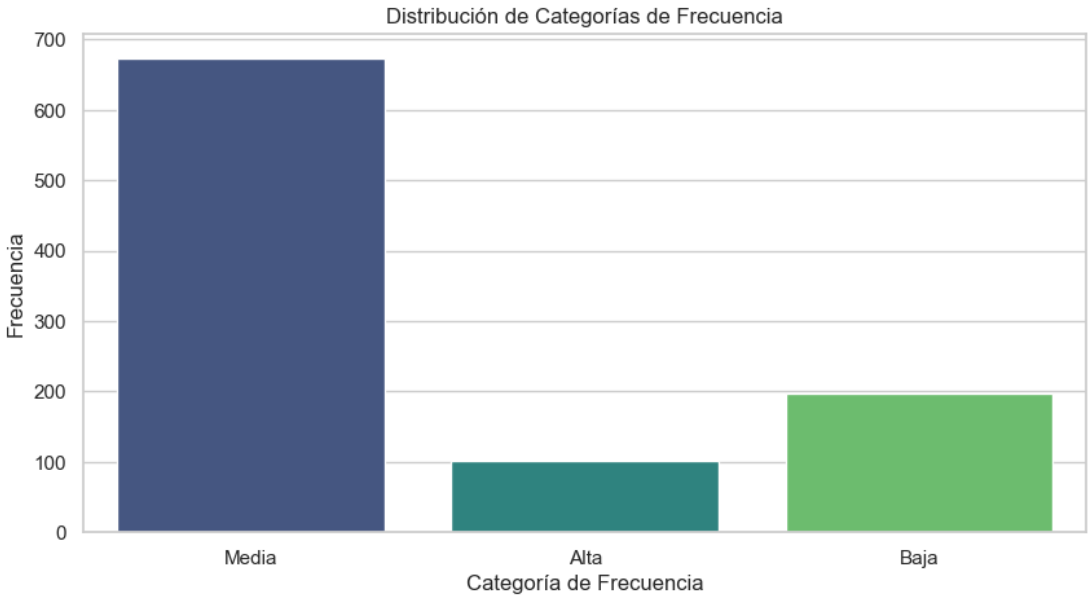

Figura 1: Distribución de Categorías de Frecuencia.

Relación entre Variables Predictoras y Frecuencia de Asistencia

Se analizaron las relaciones entre las variables predictoras seleccionadas (e.g., Age, BMI, Session_Duration) y la variable objetivo (Frequency_Category) mediante gráficas de boxplot. Estas gráficas permiten identificar cómo varían las distribuciones de las variables predictoras según las categorías de frecuencia.

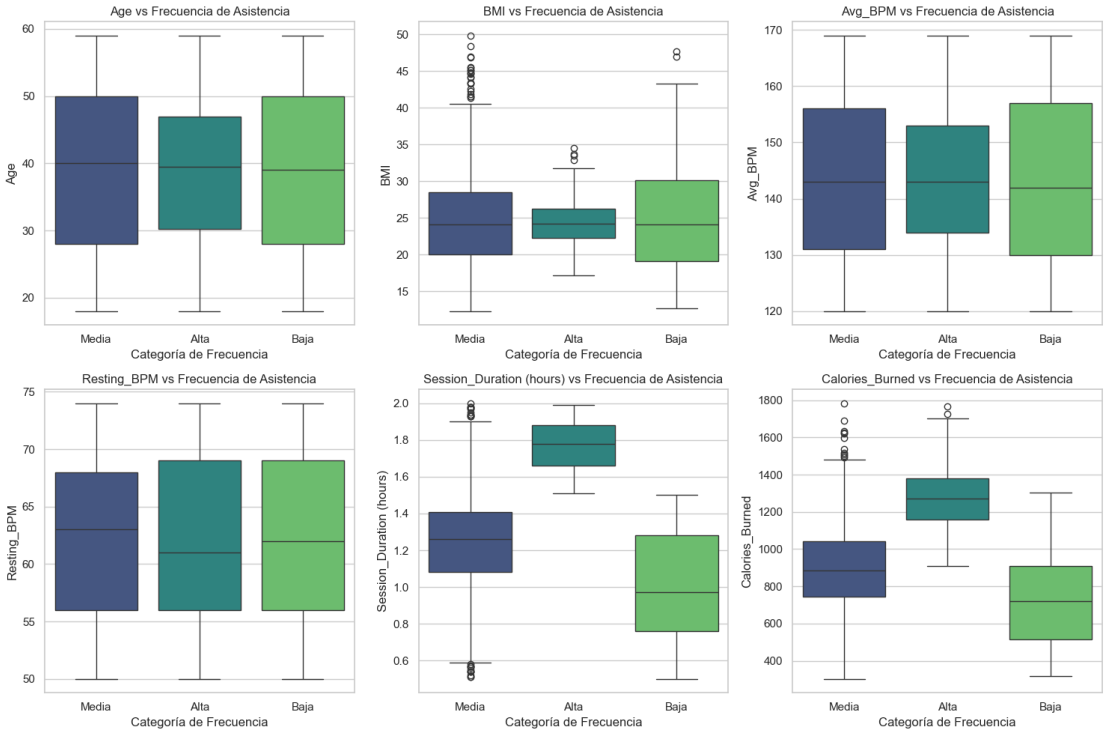

Figura 2: Boxplots de Variables Predictoras vs. Frecuencia de Asistencia.

Matriz de Correlación

La matriz de correlación muestra las relaciones entre las variables numéricas seleccionadas. Ayuda a identificar dependencias lineales que podrían influir en la efectividad del modelo.

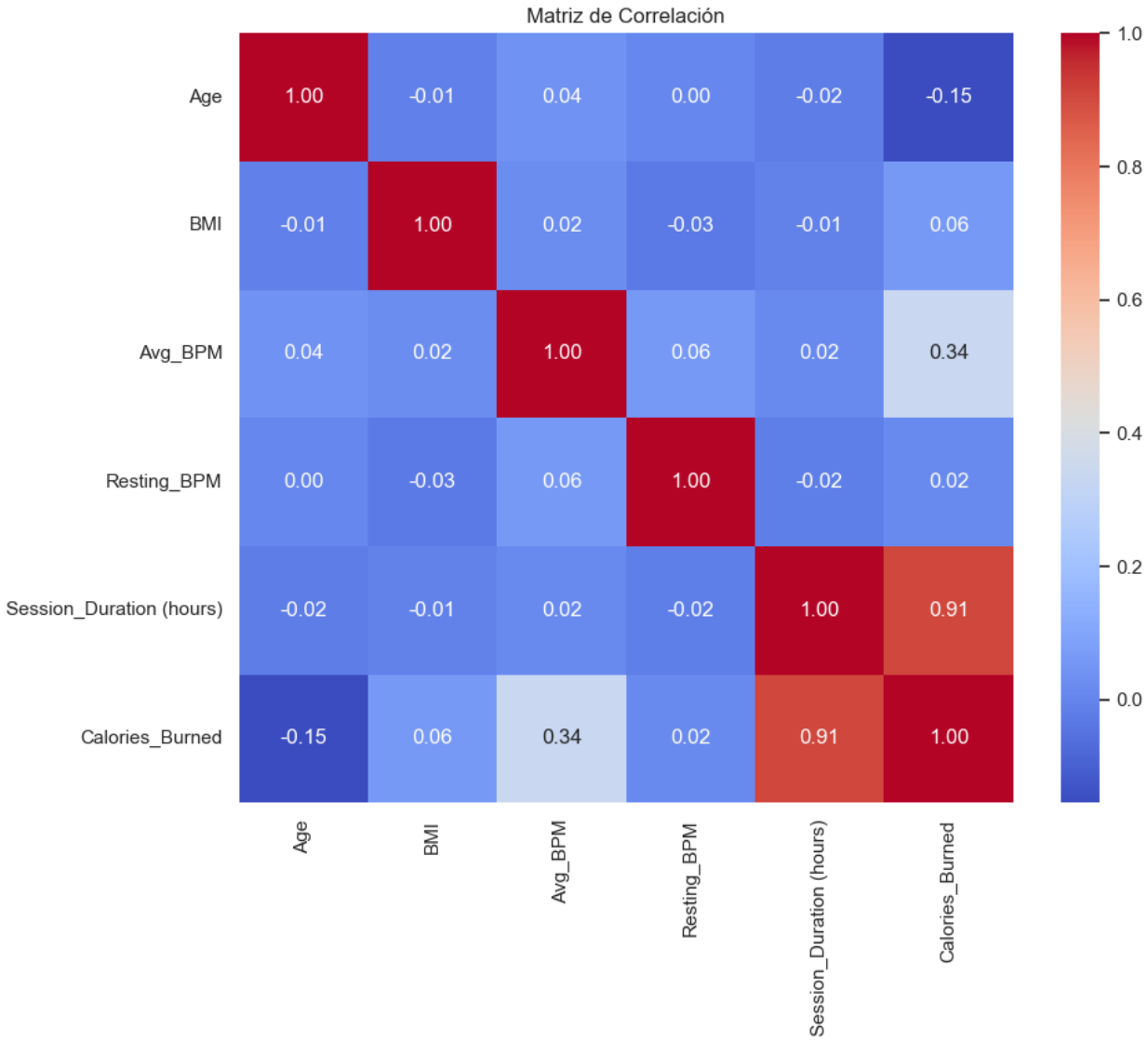

Figura 3: Matriz de Correlación entre Variables Numéricas.

Conclusiones

(I) El modelo **Random Forest Base** mostró el mejor desempeño global, con una precisión del 70% y un excelente recall en la clase predominante (Media). Sin embargo, tuvo dificultades con las clases menos representadas (Alta y Baja).

(II) El modelo **Random Forest ajustado con SMOTE** mejoró significativamente en las clases menos representadas, pero sacrificó algo de precisión global.

(III) **Gradient Boosting** logró un mejor balance entre precisión y recall para todas las clases, aunque con una precisión global ligeramente menor (61%).
(IV) La clase **Baja** sigue siendo un desafío para todos los modelos, destacando la necesidad de recopilar más datos para esta categoría.

Recomendaciones

(I) **Recolección de Datos:** Incrementar el volumen de datos para las clases menos representadas (Alta y Baja) para mejorar su representación en el dataset.
(II) **Implementación del Modelo:** Utilizar **Random Forest Base** si el enfoque principal es la precisión global. Considerar **Gradient Boosting** para un balance entre las clases.
(III) **Optimización:** Continuar ajustando los hiperparámetros de los modelos seleccionados para maximizar su desempeño.
(IV) **Pruebas en Producción:** Implementar el modelo en un entorno real para monitorear su efectividad en datos nuevos.
(V) **Análisis Adicional:** Investigar la importancia de variables adicionales para enriquecer el modelo.