Alcohol Consumption

Pablo Sáez Morales
Antonio Ruiz Jiménez
Candela Esparrica Torrecilla

July 7, 2025

Contents

1 Introducción

Descripción del problema

El consumo de alcohol es un fenómeno social ampliamente extendido que afecta a personas de todas las edades y niveles educativos. Comprender qué factores influyen con mayor fuerza en este comportamiento resulta fundamental para el diseño de estrategias preventivas y políticas públicas más eficaces. Aunque existen numerosos estudios sobre el tema, aún no está del todo claro cuáles son las variables demográficas que más contribuyen al consumo de alcohol.

Motivación del problema y su relevancia

Los problemas relacionados con el alcohol, como la adicción, las complicaciones de salud y los efectos sociales negativos, siguen siendo una preocupación relevante a nivel mundial. Identificar qué características personales —como la edad, el sexo o el nivel de estudios— están más asociadas al consumo puede ayudar a dirigir mejor los esfuerzos de prevención y concienciación. Este proyecto nace de la necesidad de aportar una respuesta basada en datos a esa cuestión.

Audiencia interesada en el problema

Los resultados de este estudio pueden resultar útiles para profesionales del ámbito sanitario, responsables de políticas públicas, educadores y trabajadores sociales. Todos ellos están involucrados en la creación de campañas de concienciación o en programas de intervención, y conocer qué factores tienen mayor peso en el consumo de alcohol les permite enfocar mejor sus recursos.

Beneficios de una solución propuesta

Contar con una visión clara sobre los principales factores que inciden en el consumo de alcohol puede facilitar la elaboración de campañas preventivas más eficaces y dirigidas a grupos de riesgo concretos. A medio plazo, esto podría traducirse en una reducción de los daños asociados al alcohol y en una sociedad más informada y saludable.

Solución propuesta

Para abordar este problema, desarrollamos un modelo predictivo basado en técnicas de aprendizaje automático. El objetivo es determinar cuál de las variables disponibles —sexo, edad o nivel de estudios— tiene una mayor influencia en el consumo de alcohol. Para ello, entrenamos un clasificador Random Forest y analizamos su comportamiento utilizando valores SHAP, lo que nos permite obtener predicciones precisas y al mismo tiempo comprensibles.

Desafíos computacionales enfrentados

El principal reto consistió en interpretar el modelo de manera que fuera estadísticamente válida y fácil de entender. Además, fue necesario transformar adecuadamente las variables categóricas y gestionar posibles desequilibrios en las clases objetivo. Otro aspecto

importante fue garantizar la reproducibilidad y la imparcialidad en el proceso de evaluación.

Distribución de tareas dentro del grupo

Todos los integrantes del grupo colaboraron en el desarrollo del proyecto. Candela Esparrica se centró en la transformación de datos y en la elaboración del predictor RandomForest. Pablo Sáez y Antonio Ruiz crearon el notebook,e implementaron el algoritmo SHAP. La elaboración del informe se hizo de forma colavorativa.

Resumen de los resultados obtenidos

El análisis permitió identificar que una de las variables tenía una influencia notablemente mayor en la predicción del consumo de alcohol. Gracias al uso de los valores SHAP, pudimos visualizar y cuantificar estos efectos, lo que ofreció una perspectiva no solo sobre qué variable es la más determinante, sino también sobre cómo afecta a las predicciones individuales. Estos resultados servirán de base para la discusión y las conclusiones que se presentan más adelante.

2 Método Propuesto

Elección de la solución

Para abordar el problema de identificar qué variable tiene mayor influencia en el consumo de alcohol, consideramos distintas opciones dentro del ámbito del aprendizaje automático. Se evaluaron modelos como regresión logística, árboles de decisión simples y redes neuronales. Sin embargo, optamos por utilizar un modelo Random Forest debido a su buen equilibrio entre precisión, interpretabilidad y robustez ante datos categóricos.

Este enfoque no solo permite obtener predicciones fiables, sino también evaluar la importancia relativa de cada variable de entrada. Además, es un modelo que no requiere una gran cantidad de ajustes y funciona bien incluso con datos moderadamente complejos, como es nuestro caso.

Justificación del enfoque elegido

El modelo Random Forest fue elegido porque combina múltiples árboles de decisión y permite reducir el sobreajuste que puede darse en modelos más simples. Esta técnica también proporciona una medida interna de la importancia de las variables, lo cual es especialmente útil para responder a nuestro objetivo principal: identificar qué factor influye más en el consumo de alcohol.

A esto se suma el uso de valores SHAP (SHapley Additive exPlanations), una técnica moderna que nos permitió interpretar el modelo en profundidad. SHAP no solo muestra qué variables son más importantes en promedio, sino también cómo afectan cada una de las predicciones individuales, lo que añade transparencia al proceso.

Metodología para medir el rendimiento

Para evaluar el rendimiento del modelo, dividimos los datos en conjuntos de entrenamiento y prueba. Utilizamos métricas estándar como la precisión (accuracy), la precisión positiva (precision) y la sensibilidad (recall) para cuantificar qué tan bien estaba funcionando el modelo.

También analizamos gráficamente la importancia de cada variable tanto a través del propio Random Forest como mediante los valores SHAP. Esto nos permitió comparar los enfoques tradicionales con técnicas más avanzadas de interpretación de modelos.

3 Resultados Experimentales

Demostración y tecnologías

El proyecto se desarrolló en un entorno basado en Python, utilizando un cuaderno Jupyter para organizar y ejecutar el código. Para entrenar el modelo y analizar los resultados, se emplearon las siguientes tecnologías:

- **Python 3.10**: Lenguaje principal de programación.
- **pandas 2.0**: Para la carga y manipulación de datos.
- **scikit-learn 1.3**: Para la implementación del modelo Random Forest y el proceso de entrenamiento y evaluación.
- **shap 0.43**: Para el análisis interpretativo del modelo mediante valores SHAP.
- **matplotlib y seaborn**: Para la visualización de resultados.

El código está organizado para ser reproducible, con semillas aleatorias fijadas para garantizar consistencia en los resultados.

Resultados de la mejor configuración

El modelo Random Forest fue entrenado con 100 árboles (`n_estimators=100`) y una profundidad máxima de 5 niveles, usando la estrategia `max_features='log2'` para seleccionar características en cada división.

Con esta configuración, se obtuvieron los siguientes resultados sobre el conjunto de prueba:

- **Precisión (accuracy)**: 0.687
- **Precisión ponderada (weighted precision)**: 0.672
- **Sensibilidad ponderada (weighted recall)**: 0.687

Estos valores muestran que el modelo ofrece un rendimiento aceptable para un problema de clasificación básica con variables demográficas. Aunque no se alcanzaron niveles muy altos de exactitud, los resultados son consistentes y permiten extraer conclusiones útiles a través del análisis interpretativo.

Estudio de ablación: comparación entre configuraciones

Se experimentó con distintas configuraciones del modelo para observar cómo influían en el rendimiento:

- Un Random Forest sin límite de profundidad tendía a sobreajustar los datos, mostrando una alta precisión en entrenamiento pero bajo rendimiento en prueba.
- Reducir el número de árboles (`n_estimators=10`) provocaba una pérdida de precisión y mayor variabilidad en los resultados.
- Modelos más simples, como un único árbol de decisión, no captaban bien las relaciones entre las variables y mostraban un rendimiento inferior.

También se comparó la importancia de las variables obtenida directamente del modelo con la proporcionada por los valores SHAP. Ambos enfoques coincidieron en resaltar una variable como la más influyente, aunque SHAP ofreció una explicación más detallada, mostrando cómo cada característica afecta a cada predicción individual de forma específica.

4 Discusión y Conclusiones

Discusión de los resultados

Los resultados obtenidos muestran que el modelo Random Forest fue capaz de detectar patrones significativos en los datos, alcanzando una precisión de aproximadamente 68.7%. Aunque este valor no indica un rendimiento especialmente alto, sí es suficiente para realizar un análisis interpretativo fiable.

El uso de valores SHAP permitió observar con mayor detalle cómo influye cada variable en la predicción. A diferencia de las métricas de importancia del modelo, que ofrecen una visión global, SHAP muestra cómo cada variable afecta individualmente a cada ejemplo. Este enfoque hizo posible identificar no solo qué variable es la más influyente en promedio, sino también en qué dirección y con qué intensidad afecta a cada predicción.

Validez del método

La combinación del modelo Random Forest con el análisis SHAP demostró ser adecuada para los objetivos del proyecto. El modelo proporcionó resultados razonablemente estables, y los valores SHAP permitieron interpretar sus decisiones de forma transparente. Esta interpretación resulta fundamental cuando se trabaja con temas sociales, ya que aporta explicaciones comprensibles a las predicciones realizadas por el sistema.

Limitaciones y grado de madurez

Una de las principales limitaciones del modelo es la simplicidad del conjunto de datos utilizado. Solo se consideraron tres variables: sexo, edad y nivel de estudios. Esto limita la capacidad del modelo para captar matices más complejos que podrían influir en el consumo de alcohol. Además, la codificación de las variables categóricas puede introducir sesgos si no se gestiona con cuidado.

Desde el punto de vista técnico, el modelo es funcional y reproducible, pero aún no está listo para ser utilizado en aplicaciones reales. Para ello, sería necesario validarlo con un conjunto de datos más amplio y representativo, así como incorporar nuevas variables que reflejen mejor el contexto social y personal de los individuos.

Trabajos futuros

Como continuación de este proyecto, se proponen varias líneas de trabajo:

- Ampliar el conjunto de datos incluyendo más variables relevantes, como factores familiares, económicos o de salud.
- Aplicar técnicas de balanceo de clases para mejorar el rendimiento del modelo en casos de desbalance.
- Comparar el modelo Random Forest con otros enfoques interpretables, como la regresión logística o árboles de decisión simples.
- Evaluar la capacidad de generalización del modelo en diferentes poblaciones o contextos geográficos.

Estas mejoras permitirían avanzar hacia una solución más robusta, precisa y aplicable en la práctica, manteniendo siempre la interpretabilidad como un eje central.