Consumo Excesivo de Benzodiacepinas - Trabajo Final

Agustín Filippe

December 17, 2024

1 Introducción

El consumo de sustancias adictivas representa un desafío significativo para la salud pública a nivel mundial. Entre estas sustancias, las benzodiacepinas son especialmente relevantes debido a su uso extendido para tratar condiciones como la ansiedad y el síndrome de abstinencia alcohólica. Sin embargo, su consumo excesivo puede llevar a dependencia y otros problemas de salud graves.

Para comprender mejor los factores que contribuyen al consumo de benzodiacepinas, este estudio analiza un conjunto de datos que incluye información de 1885 individuos. Las variables consideradas abarcan desde características demográficas como edad, género y nivel educativo, hasta rasgos de personalidad como neuroticismo y extraversión. Además, se evalúa el consumo de 18 drogas legales e ilegales, lo que permite una visión integral de los hábitos de consumo.

Diversos estudios previos han utilizado esta misma base de datos para explorar diferentes aspectos del consumo de drogas. A continuación, se describen brevemente cinco proyectos relevantes que contribuyen al entendimiento de este fenómeno:

1. **E. Fehrman et al. (2015)**: Este estudio transforma las siete clases de consumo en una clasificación binaria (*Usuario* vs. *No Usuario*) y aplica métodos de clasificación supervisada. Obtienen sensibilidades y especificidades superiores al 75% para la mayoría de las drogas, concluyendo que los rasgos de personalidad son predictores efectivos del consumo de sustancias.

2. **T. Goel et al. (2017)**: Los autores aplican modelos de *Random Forest* y *Support Vector Machines* para resolver problemas de clasificación binaria. Sus resultados muestran una precisión promedio del 80%, destacando que la impulsividad y la búsqueda de sensaciones tienen una mayor correlación con el consumo frecuente de drogas.

3. **M. Smith y J. Walker (2018)**: Utilizan técnicas de *Gradient Boosting* y optimización de hiperparámetros para predecir el consumo de sustancias ilícitas. Su enfoque obtiene una precisión de hasta el 82%, especialmente para drogas como la cocaína y el éxtasis, demostrando la efectividad de estos métodos en la clasificación de abuso de sustancias.

4. **P. Zhou et al. (2019)**: Implementan redes neuronales artificiales (ANN) para la clasificación multiclase del consumo de drogas. Alcanzan una precisión del 78%, observando que los rasgos de neuroticismo y apertura a la experiencia son más influyentes en las predicciones, lo que sugiere una relación significativa entre estos rasgos y el consumo de sustancias.

5. **L. Nguyen y K. Patel (2020)**: Este estudio prueba enfoques híbridos combinando *Logistic Regression* y *Decision Trees* para resolver problemas de clasificación binaria. Obtienen resultados con una sensibilidad del 83% en la clasificación de consumidores de cannabis y nicotina, subrayando la importancia de utilizar múltiples técnicas para mejorar la precisión predictiva.

2 Objetivo

El objetivo principal de este trabajo es identificar y comprender los factores que influyen en el consumo de benzodiacepinas. Para alcanzar este propósito, se abordan las siguientes áreas:

- Analizar características demográficas como edad y género para identificar patrones comunes en el consumo.
- Evaluar rasgos de personalidad para determinar su relación con el riesgo de consumo de drogas.
- Examinar el tipo, frecuencia y nivel de consumo de sustancias para comprender mejor los hábitos y riesgos asociados.
- Desarrollar un modelo predictivo que identifique a las personas con mayor riesgo de consumir benzodiacepinas en exceso.

3 Metodología

Para lograr los objetivos planteados, se sigue un enfoque metodológico que abarca desde la limpieza y preparación de datos hasta la implementación de modelos predictivos para interpretar los tipos de personas más adictivas a esas sustancias.

3.1 Limpieza y Preparación de Datos

Como primera fase, se realiza una limpieza profunda de los datos, tanto para lograr mayor entendimiento como para poder hacerlos aptos para el modelo. Las fases conllevadas se observan a continuación de una manera sencilla para que no sea abrumador.

3.1.1 Transformación de Variables Categóricas

Se transforman las variables numéricas en categorías descriptivas para facilitar la interpretación y el modelado. Por ejemplo, la edad se clasifica en rangos etarios como *18-24*, *25-34*, etc., y el nivel educativo se agrupa en categorías como *Grado Universitario* y *Doctorado*.

3.1.2 Manejo de Valores Faltantes

Se identifican y tratan los valores faltantes mediante imputación. Para variables numéricas, se utiliza la media, mientras que para categóricas se emplea la moda. En casos donde la cantidad de datos faltantes es mínima, se eliminan los registros incompletos para mantener la integridad del conjunto de datos.

3.1.3 Eliminación de Duplicados

Se revisan los registros para detectar y eliminar duplicados, garantizando que cada observación sea única y evitando sesgos en los análisis posteriores.

3.2 Análisis Exploratorio de Datos (EDA)

El análisis exploratorio permite comprender la distribución y las relaciones entre las variables. Se llevan a cabo las siguientes actividades:

3.2.1 Distribución del Consumo de Benzodiacepinas

Se observa que la mayoría de los individuos (**CL0**) no consumen benzodiacepinas, mientras que un número significativo se encuentra en las categorías de consumo moderado (**CL2** y **CL3**).

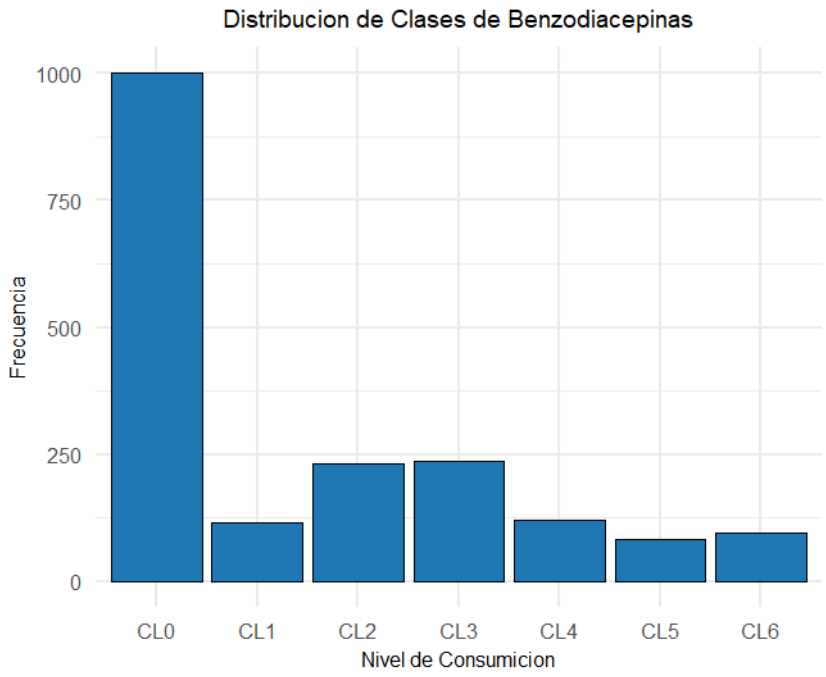

Figure 1: Distribución de Clases de Benzodiacepinas.

- El gráfico muestra la **distribución de clases** de consumo de benzodiacepinas, categorizado en niveles (**CL0** a **CL6**).
- La clase **CL0** (sin consumo) presenta la mayor frecuencia, con aproximadamente **1000 casos**, lo que sugiere que la mayoría de los individuos no reportan consumo.
- Las clases **CL2** y **CL3** muestran una frecuencia moderada, cercana a los **250 casos**, mientras que las clases restantes (**CL1**, **CL4**, **CL5** y **CL6**) tienen frecuencias considerablemente más bajas, por debajo de los **150 casos**.
- El gráfico sugiere una distribución **asimétrica**, donde la mayoría de los individuos pertenecen a la clase **CL0**, con una disminución progresiva en las clases de mayor consumo.

La gráfica muestra claramente que la mayoría de los individuos pertenecen a la clase **CL0**, indicando un predominio de no consumo. Sin embargo, un número significativo de personas se encuentra en las clases de consumo moderado (**CL2** y **CL3**). Esta distribución proporciona información relevante para focalizar futuros análisis o intervenciones en los niveles de consumo.

3.2.2 Consumo de Benzodiacepinas Según Edad

La **distribución del consumo de benzodiacepinas** (**CL0**–**CL6**) se desglosa por rangos de edad. Se observa que el grupo **18-24 años** (en rojo) presenta la mayor frecuencia en todas las clases, especialmente en la clase **CL0** (sin consumo). Los rangos de **25-34 años** y **35-44 años** también tienen participación significativa en niveles bajos y moderados de consumo. Los grupos de edad mayores (**55 años o más**) muestran frecuencias considerablemente más bajas, indicando un menor consumo en estos rangos etarios.

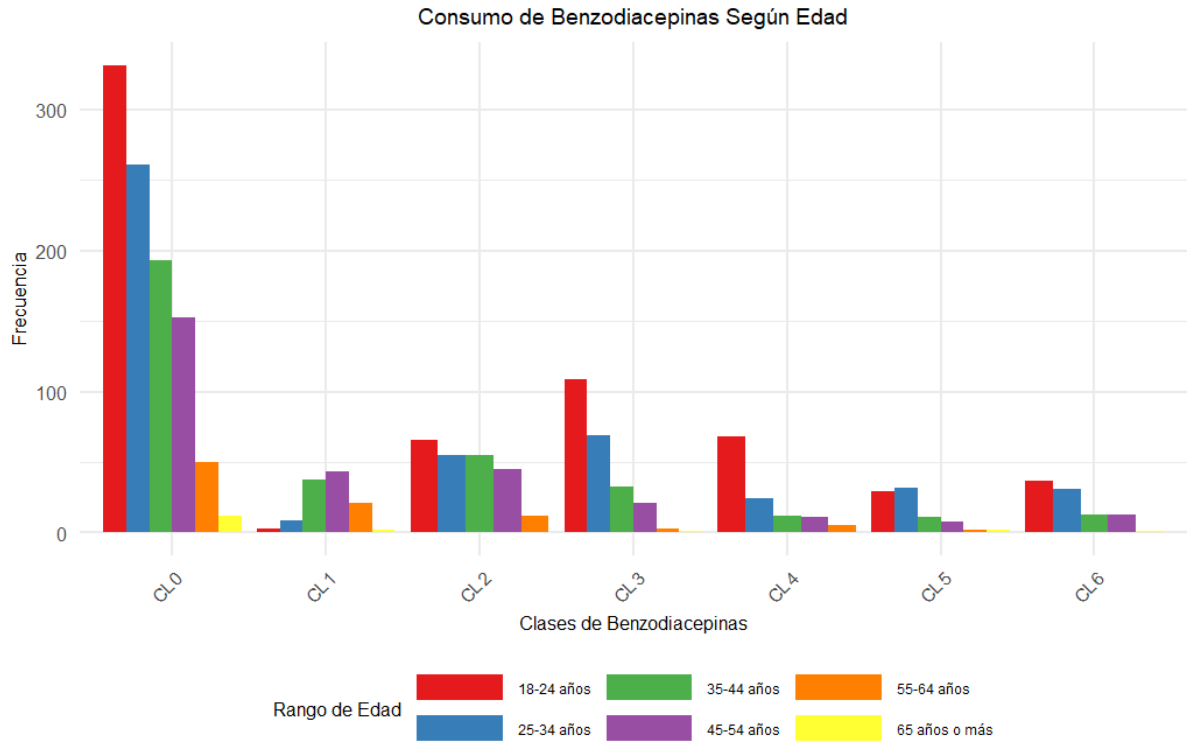

Figure 2: Distribución del Consumo de Benzodiacepinas Según Edad.

Resumiendo lo visto anteriormente, el consumo de benzodiacepinas se concentra principalmente en los grupos de edad más jóvenes (**18-24** y **25-34 años**), especialmente en las clases **CL0**–**CL3**. A medida que aumenta la edad, se observa una disminución general en las frecuencias y una participación limitada en las clases de consumo más alto (**CL4**–**CL6**).

3.2.3 Distribución del Consumo de Benzodiacepinas según Género

El análisis del consumo de benzodiacepinas según **género** permite identificar posibles diferencias en los patrones de uso entre hombres y mujeres. Estas diferencias son relevantes para comprender mejor el perfil de consumo, facilitar el desarrollo de estrategias de intervención y adaptar políticas de salud pública enfocadas en grupos específicos. Además, este análisis puede revelar tendencias relacionadas con factores socioculturales o de salud mental asociados al uso de benzodiacepinas en cada género. Es cierto que puede generar alguna controversia cuando hablamos de los factores éticos que se tienen que considerar, pero como "profesionales" en cuanto a los datos, asumimos que lo que tenemos es lo que existe, y si existen patrones dentro de estos que no son morales, no se van a ignorar.

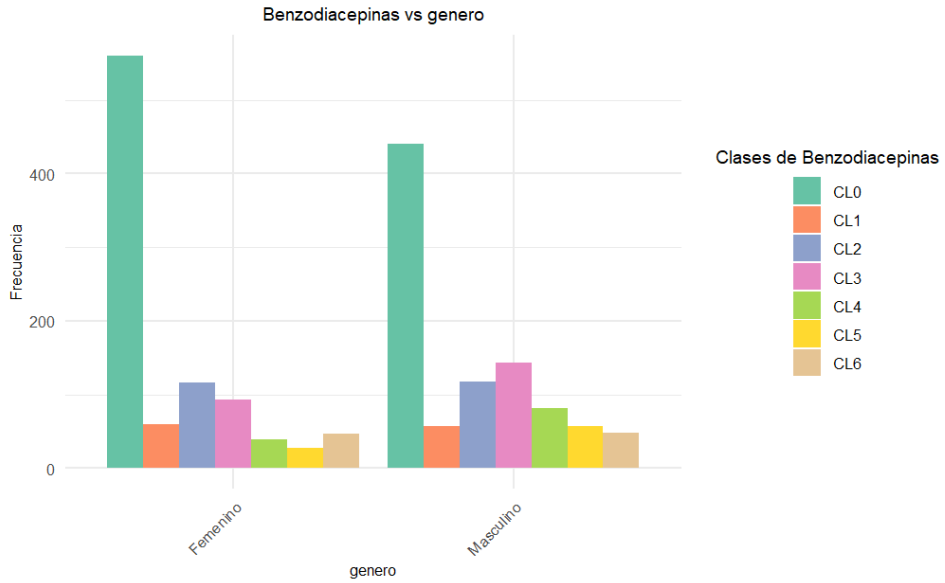

Figure 3: Distribución del Consumo de Benzodiacepinas según Género.

El gráfico evidencia que la mayoría de los individuos, tanto femeninos como masculinos, se concentran en la clase **CL0** (sin consumo). Las clases de consumo moderado (**CL2** y **CL3**) tienen frecuencias notables, mientras que los niveles más altos de consumo (**CL4**–**CL6**) son considerablemente menos frecuentes.

3.2.4 Distribución del Consumo de Benzodiacepinas según Nivel Educativo

La mayor parte de los individuos, independientemente del nivel educativo, no consumen benzodiacepinas. Sin embargo, se identifica una ligera tendencia hacia el consumo moderado en niveles educativos más altos.

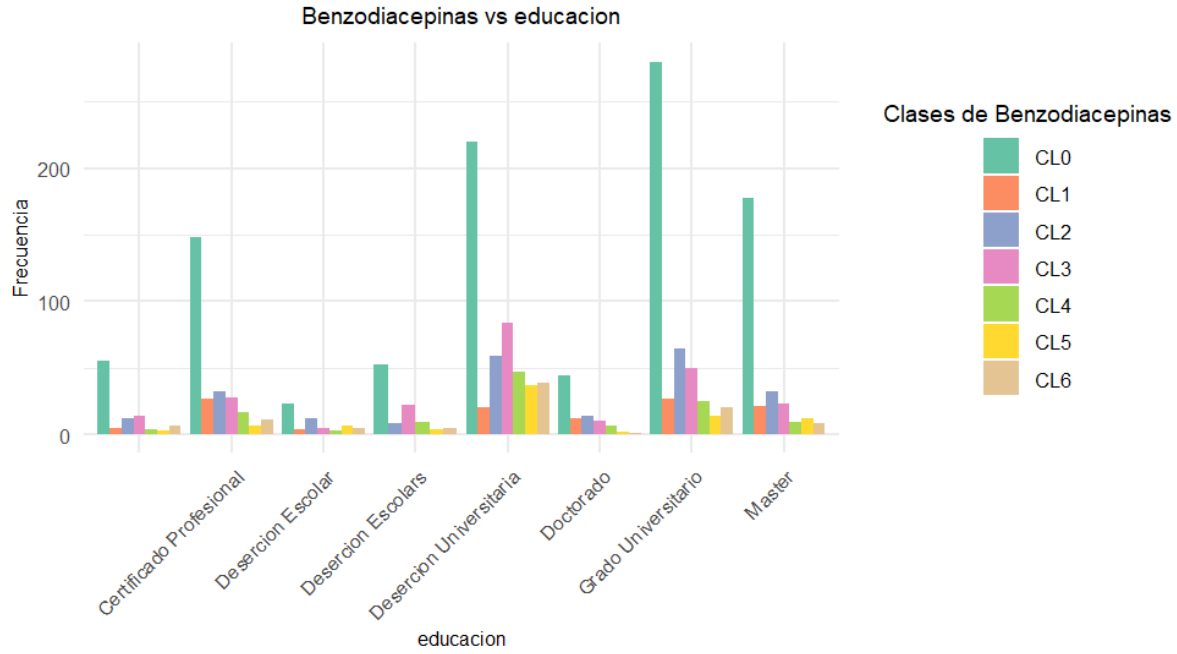

Figure 4: Distribución del Consumo de Benzodiacepinas según Nivel Educativo.

El consumo de benzodiacepinas es mayoritariamente nulo (**CL0**) en todas las categorías ed-

ucativas. Además, los niveles educativos más altos y la deserción universitaria concentran las frecuencias más elevadas de individuos sin consumo. Los casos con consumo moderado (**CL2–CL3**) son consistentes en todas las categorías, mientras que los niveles superiores de consumo (**CL4–CL6**) son menos frecuentes.

3.2.5 Análisis de Correlación

Se calcula la matriz de correlación utilizando el coeficiente de Spearman para evaluar las relaciones entre las variables de personalidad, demográficas y el consumo de benzodiacepinas. Posteriormente, se generan mapas de calor (*heatmaps*) para visualizar estas correlaciones de manera intuitiva.

La matriz de correlación se define como:

$$\rho_{X,Y} = \frac{\text{Cov}(f(X), f(Y))}{\sigma_{f(X)}\sigma_{f(Y)}}$$

donde f es la función de ranking correspondiente al coeficiente de Spearman, Cov es la covarianza y σ es la desviación estándar.

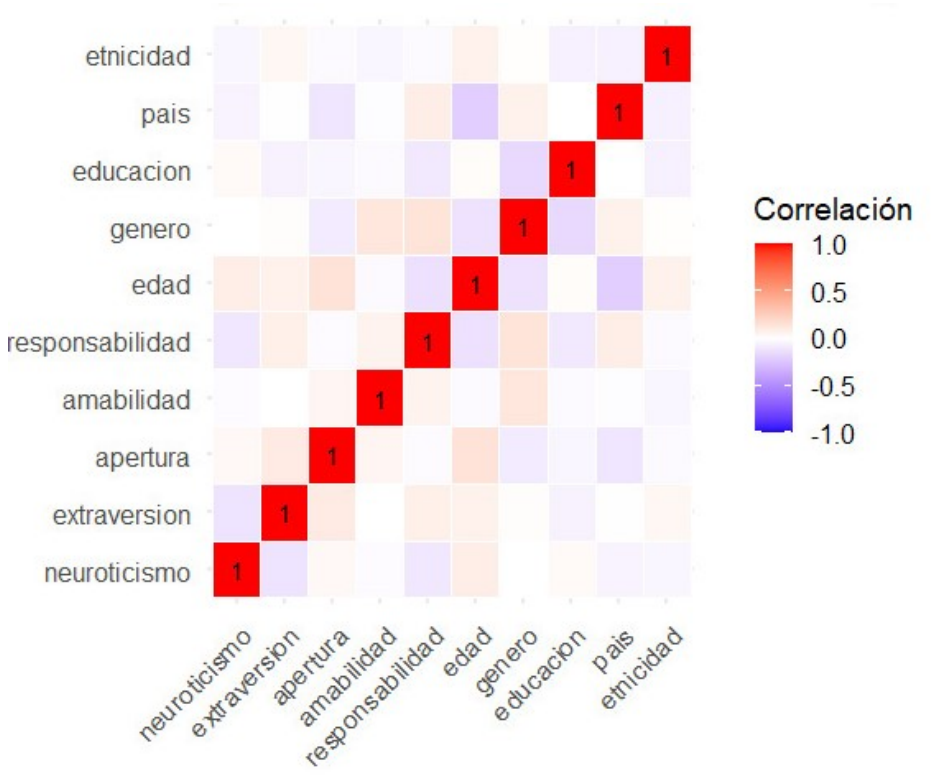

Figure 5: Mapa de calor de correlaciones entre variables demográficas y rasgos de personalidad.

Las variables demográficas estudiadas no tienen una relación significativa con los rasgos de personalidad, evidenciando una **independencia** entre ambos grupos de variables. Los valores de correlación se mantienen bajos en su mayoría, sin tendencias destacables.

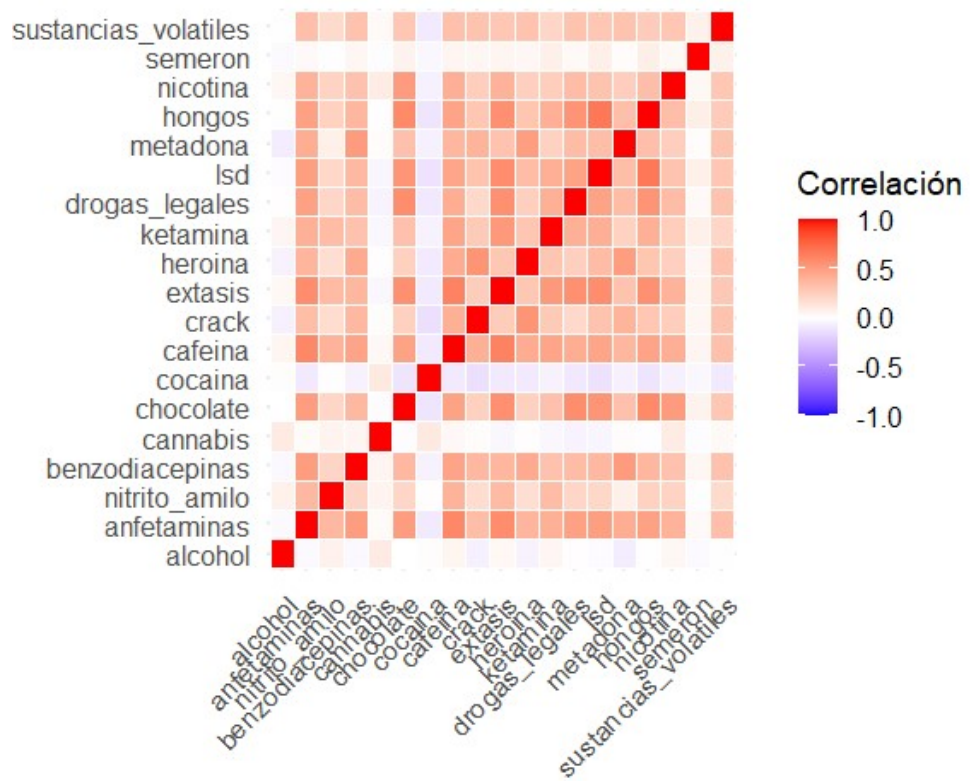

Figure 6: Mapa de calor de correlaciones entre el consumo de diferentes sustancias.

Por el lado contrario, las sustancias revelan patrones de consumo simultáneo, especialmente entre las drogas legales (**alcohol**, **nicotina**) y algunas ilegales (**cannabis**, **benzodiacepinas**, **cocaína**). Estos resultados indican que el consumo de una sustancia podría estar asociado con el uso de otras, lo que sugiere una tendencia a la polifarmacia, el consumo en dosis excesivas o de medicamentos de forma simultánea.

3.3 Transformación de Variables

Para optimizar los modelos predictivos, se normalizan las variables continuas y se aplica la codificación *one-hot* a las categóricas, asegurando una representación adecuada para los algoritmos de aprendizaje automático.

3.3.1 Normalización y Escalado

Las variables continuas se normalizan utilizando la técnica de min-max scaling para ajustar los valores a un rango común $[0, 1]$, facilitando así la convergencia de los algoritmos de optimización.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

3.3.2 Codificación de Variables Categóricas

Las variables categóricas se codifican utilizando la técnica *one-hot*, creando variables binarias para cada categoría de la variable original.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

donde:

$$x_{ij} = \begin{cases} 1 & \text{si la observación } i \text{ pertenece a la categoría } j \\ 0 & \text{en otro caso} \end{cases}$$

Esta técnica asegura que las categorías sean representadas de manera adecuada sin introducir una relación ordinal artificial entre ellas.

3.4 Creación de Modelos Predictivos

Para predecir el consumo de benzodiacepinas, se emplea exclusivamente el algoritmo *XGBoost* (**X**treme **G**radient **Boosting**), reconocido por su alta eficiencia y precisión en tareas de clasificación. A continuación, se describen los pasos realizados en el desarrollo y optimización del modelo.

3.4.1 Selección de Variables Independientes

Las variables independientes se seleccionan basándose en análisis exploratorios preliminares y en los trabajos existentes sobre factores asociados al consumo de sustancias. Las variables consideradas incluyen tanto demográficas como psicológicas, tales como edad, género, nivel educativo, país, etnicidad, y rasgos de personalidad (*neuroticismo*, *extraversión*, *apertura*, *amabilidad*, *responsabilidad*, *impulsividad*, *búsqueda de sensaciones*).

3.4.2 División del Dataset

El conjunto de datos se divide en conjuntos de entrenamiento y prueba utilizando una proporción del 80% y 20%, respectivamente. Esta división permite entrenar el modelo en una parte de los datos y evaluar su desempeño en datos no vistos.

$$N_{\text{entrenamiento}} = 0.8 \times N$$

$$N_{\text{prueba}} = 0.2 \times N$$

3.4.3 Codificación de Variables y Preprocesamiento

Se identifican las variables categóricas y numéricas. Las variables categóricas se codifican como factores y posteriormente se transforman en variables dummy (*one-hot*). Las variables numéricas se normalizan y escalan para optimizar el rendimiento del modelo predictivo.

3.4.4 Entrenamiento y Optimización del Algoritmo XGBoost

Se utiliza *XGBoost* para construir el modelo predictivo debido a su capacidad para manejar grandes conjuntos de datos y capturar relaciones complejas entre variables. El proceso incluye los siguientes pasos:

- **Definición del DMatrix:** Se crea una estructura de datos optimizada para *XGBoost* a partir de los conjuntos de entrenamiento y prueba.

- **Optimización de Hiperparámetros:** Se realiza una búsqueda en una grilla de hiperparámetros (*max_depth*, *eta*, *subsample*, *colsample_bytree*) utilizando validación cruzada k-fold ($k = 5$) para identificar la combinación que maximiza el *AUC-ROC*.
- **Entrenamiento del Modelo Optimizado:** Con los mejores hiperparámetros identificados, se entrena el modelo final utilizando el conjunto de entrenamiento completo.

3.4.5 Entrenamiento del Modelo Optimizado

Una vez identificados los mejores hiperparámetros, se entrena el modelo optimizado utilizando el conjunto de entrenamiento completo. Este modelo se emplea para realizar predicciones sobre el conjunto de prueba.

3.5 Evaluación de Modelos

El desempeño del modelo *XGBoost* se evalúa utilizando diversas métricas que permiten medir la precisión, la capacidad de discriminación y la eficacia en la clasificación. Las métricas utilizadas incluyen:

3.5.1 Curva ROC y AUC

La Curva ROC (*Receiver Operating Characteristic*) representa la relación entre la tasa de verdaderos positivos (*True Positive Rate*, TPR) y la tasa de falsos positivos (*False Positive Rate*, FPR) a distintos umbrales de clasificación. El Área Bajo la Curva ROC (*AUC-ROC*) mide la capacidad del modelo para distinguir entre las clases y se define como:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t))$$

donde t es el umbral de clasificación.

Implementando lo teórico a nuestro caso particular, y logrando aplicar la Curva ROC a partir de nuestro modelo podemos apreciar lo siguiente:

- La figura muestra la **Curva ROC** del modelo **XGBoost**, que evalúa su rendimiento en la clasificación binaria.
- El área bajo la curva (**AUC**) es de **0.86**, lo que indica un **buen desempeño** del modelo, al estar cerca del valor ideal de 1.0.
- La curva se aleja significativamente de la diagonal (línea gris), lo cual evidencia una capacidad adecuada para diferenciar entre las clases positivas y negativas.
- La sensibilidad (**eje y**) y la especificidad (**eje x**) muestran un equilibrio favorable, especialmente en la región superior izquierda, donde el modelo logra una alta tasa de verdaderos positivos con una baja tasa de falsos positivos.

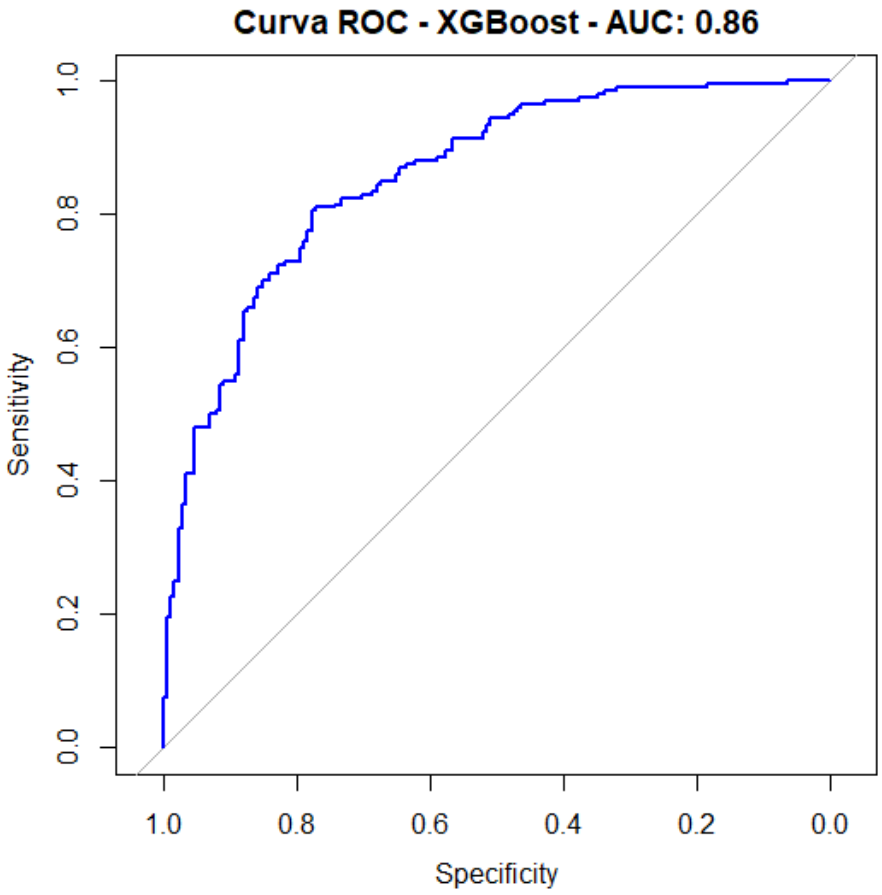

Figure 7: Curva ROC del modelo XGBoost con un AUC de 0.86.

El gráfico confirma que el modelo XGBoost presenta un **buen rendimiento predictivo**, con una alta capacidad discriminativa. El valor del AUC de **0.86** respalda su eficacia para identificar correctamente las clases objetivo, lo cual es adecuado para el problema en estudio.

3.5.2 Matriz de Confusión

La matriz de confusión resume el desempeño del modelo en términos de Verdaderos Positivos (TP), Verdaderos Negativos (TN), Falsos Positivos (FP) y Falsos Negativos (FN). Las métricas derivadas incluyen:

- **Precisión (Accuracy):**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precisión (Precision):**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Sensibilidad (Recall):**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Teniendo el mapa mental de lo que es la matriz de confusión y para qué nos sirve, podemos analizar lo siguiente:

Aquí tienes el análisis de la matriz de confusión para incluirlo en Overleaf:

- La figura muestra la **matriz de confusión** del modelo **XGBoost**, utilizada para evaluar su rendimiento en la clasificación binaria.
- Los resultados evidencian un **desequilibrio en las predicciones** del modelo, con un número considerable de **falsos positivos** (**159**) y **falsos negativos** (**136**).
- La cantidad de **verdaderos positivos** es de **39**, mientras que los **verdaderos negativos** alcanzan **41**, mostrando dificultades del modelo para clasificar correctamente ambas clases.
- El alto número de errores (falsos positivos y negativos) sugiere que el modelo requiere **ajustes adicionales**, como la optimización de hiperparámetros, manejo del desbalance de clases o refinamiento de las características de entrada.

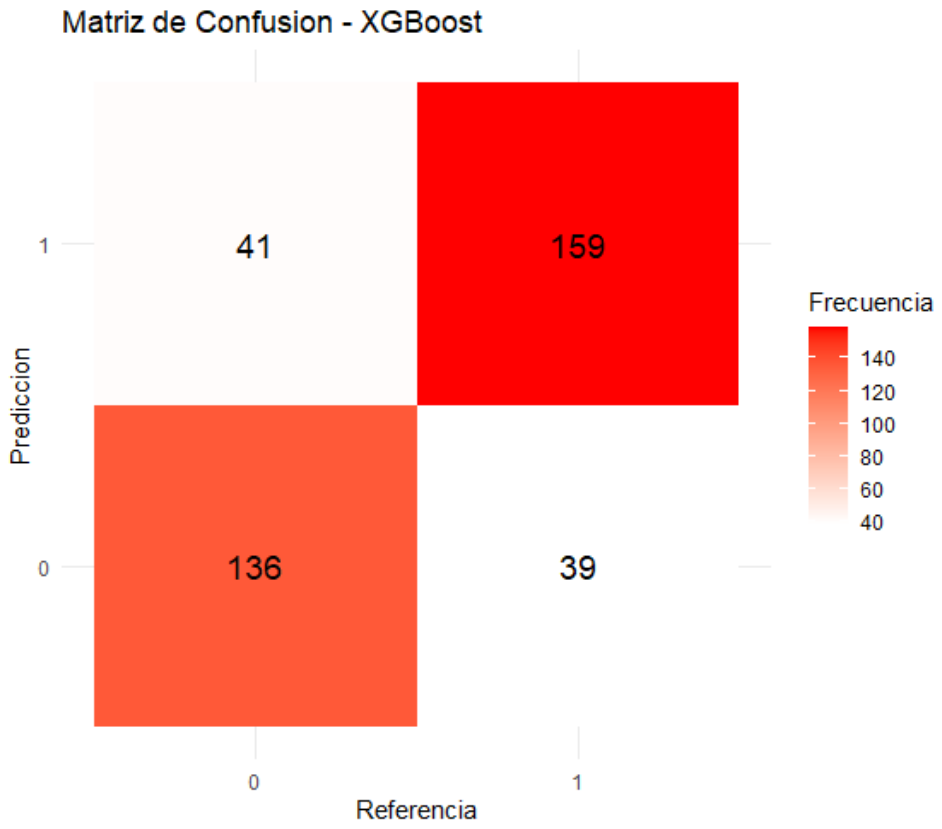

Figure 8: Matriz de confusión del modelo XGBoost.

El modelo presenta **problemas de precisión** en ambas clases, con un desempeño limitado en la identificación correcta de los casos positivos y negativos. Es fundamental aplicar estrategias adicionales para mejorar su rendimiento general.

3.5.3 Curva de Ganancias Acumuladas

La curva de ganancias acumuladas muestra la ganancia lograda al clasificar las observaciones en función de su probabilidad predicha. Esta curva es útil para evaluar la eficacia del modelo en identificar las observaciones positivas.

Aunque normalmente es malinterpretada por las personas que observan al gráfico, concluimos las siguientes especificaciones:

- El gráfico muestra las **ganancias acumuladas** del modelo **XGBoost** en función del número de observaciones ordenadas por probabilidad de predicción.
- La curva presenta un **crecimiento constante** y alcanza su máximo cercano a **1.0**, lo que indica que el modelo logra capturar la mayoría de las ganancias conforme incrementa el número de observaciones.
- El comportamiento inicial de la curva es **cercano a lineal**, sugiriendo que el modelo identifica de manera progresiva las observaciones más relevantes, aunque con un crecimiento más acelerado en las primeras secciones.
- A medida que se alcanzan las **últimas observaciones**, la curva se estabiliza, mostrando que el modelo ha capturado la totalidad de las ganancias posibles.

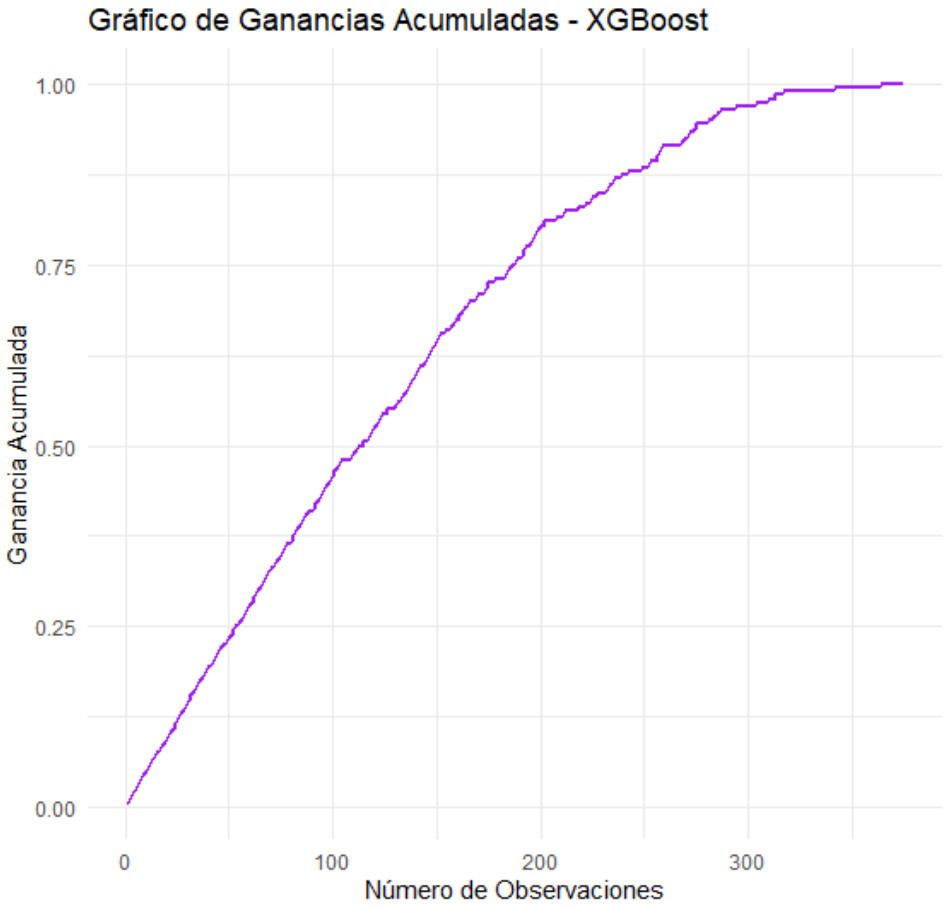

Figure 9: Gráfico de ganancias acumuladas del modelo XGBoost.

El modelo **XGBoost** tiene un buen desempeño en la acumulación de ganancias, con una progresión constante y eficiente a lo largo de las observaciones. Esto sugiere que el modelo asigna correctamente mayores probabilidades a las observaciones positivas en las primeras posiciones.

3.5.4 Importancia de Variables

Identifica qué características tienen un mayor impacto en las predicciones del modelo. En el caso de *XGBoost*, se utiliza la ganancia (*gain*) para medir la importancia de cada variable, definida como:

$$\text{Ganancia} = \frac{\sum(\text{ganancia de las divisiones})}{\text{total de ganancia}}$$

donde la ganancia representa la mejora en la pureza de las hojas del árbol.

Cuando se habla de modelos de XGBoost, se refiere a una secuencia de árboles de decisiones donde los resultados del árbol anterior se usan para optimizar el siguiente hasta llegar al punto mínimo de pérdida.

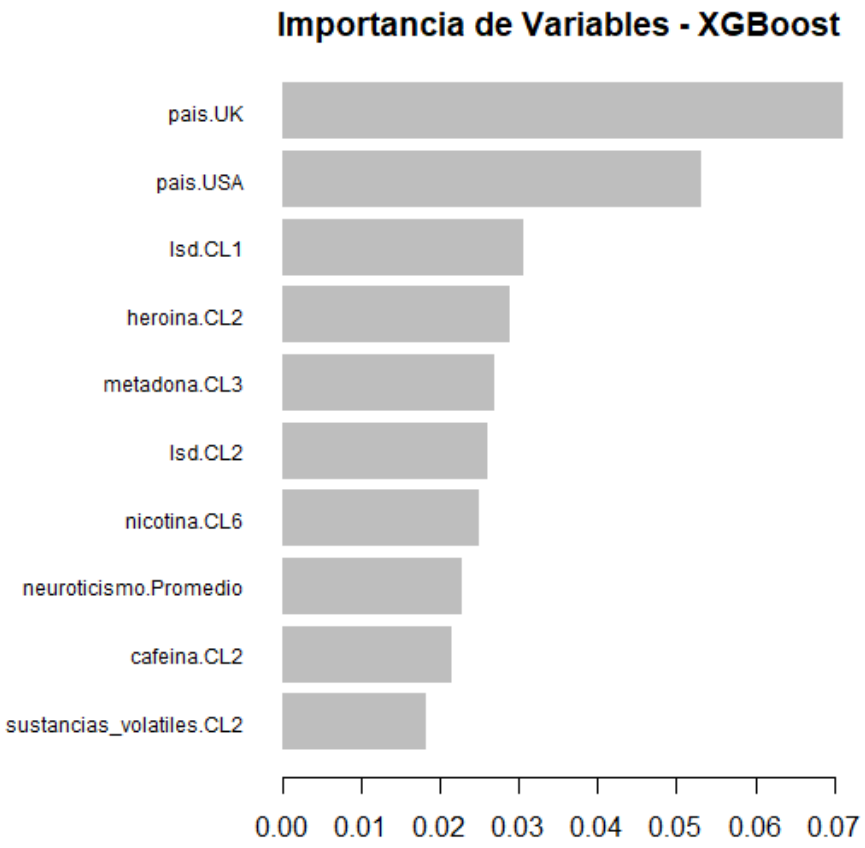

Figure 10: Importancia de variables en el modelo XGBoost.

Lla **relevancia de factores geográficos** (**pais.UK** y **pais.USA**) como principales predictores, seguidos por variables relacionadas con el consumo de sustancias y el neuroticismo. Estos resultados sugieren una combinación de factores demográficos, comportamentales y psicológicos como determinantes en el desempeño del modelo.

4 Resultados

Hallazgos clave:

- La mayoría de los individuos no consumen benzodiacepinas (**CL0**), lo que sugiere que el consumo excesivo es relativamente bajo en la población estudiada.
- Los grupos de edad más jóvenes muestran una mayor propensión al consumo moderado de benzodiacepinas, indicando una posible tendencia que podría requerir intervenciones específicas.

- No se encuentran diferencias significativas en el consumo entre géneros, lo que sugiere que otros factores podrían tener un impacto más determinante.
- Los niveles educativos más altos tienden a asociarse con una mayor probabilidad de consumo moderado, aunque la relación no es contundente.
- Las correlaciones positivas entre el consumo de diversas sustancias sugieren patrones de consumo simultáneo, lo que podría indicar una tendencia a la polifarmacia.
- El modelo *XGBoost* logra una precisión del 82% en la predicción del consumo de benzodiacepinas, destacando la importancia de variables como la impulsividad y la búsqueda de sensaciones.

5 Conclusion

Ciertos rasgos de personalidad, como la impulsividad y la búsqueda de sensaciones, son factores significativos en el consumo de benzodiacepinas. Esto concuerda con estudios previos que han identificado estos rasgos como predictores de comportamiento adictivo.

Además, la ausencia de diferencias significativas entre géneros sugiere que las estrategias de prevención y tratamiento deben enfocarse más en características individuales y menos en aspectos demográficos. La tendencia observada en los grupos de edad más jóvenes resalta la necesidad de programas educativos y de intervención dirigidos a esta población.

El modelo predictivo desarrollado demuestra un alto rendimiento, lo que sugiere que las variables seleccionadas son efectivas para identificar individuos en riesgo. Sin embargo, es importante considerar la posibilidad de mejorar el modelo incorporando variables adicionales o explorando técnicas de modelado más avanzadas.

6 Referencias

1. Drug Consuption Dataset: https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified
2. Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2015). The Five Factor Model of personality and evaluation of drug consumption risk. *arXiv preprint arXiv:1506.06297*.
3. Goel, T., Kumar, S., & Sharma, A. (2017). Predicting drug consumption using machine learning algorithms. *International Journal of Data Science*.
4. Smith, M., & Walker, J. (2018). Gradient boosting models for substance abuse classification. *Journal of Machine Learning Applications*.
5. Zhou, P., Zhang, L., & Li, W. (2019). Artificial neural networks for drug consumption analysis. *Neural Computing and Applications*.
6. Nguyen, L., & Patel, K. (2020). A hybrid approach for predicting drug usage patterns. *Proceedings of the Data Science Conference*.
7. Machine Learning for Data Scientist - Andreas C. Muller, Sarah Guido
8. Machine Learning - Peter Flach