Clasificación del Consumo de Drogas Basado en Rasgos de Personalidad y Datos Demográficos

Jeisson A. Barrantes, Natalia A. García, Alexander V. Delgado

***Abstract*—El presente trabajo aborda la predicción del nivel de consumo de seis sustancias psicoactivas utilizando técnicas de Machine Learning supervisado. A partir del conjunto de datos UCI Drug Consumption, se evaluaron cinco modelos de regresión (Ridge, KNN, Redes Neuronales, SVR y Random Forest) para estimar el riesgo en una escala ordinal de 7 niveles. Los resultados indican que el Random Forest Regressor ofrece el mejor desempeño global (RMSE $\approx$ 1.48), superando a los enfoques lineales. Adicionalmente, se demostró mediante PCA que es posible reducir la dimensionalidad del espacio de entrada en un 66% sin pérdida significativa de precisión, validando la redundancia de ciertos rasgos psicométricos.**

***Index Terms*—Machine Learning, Regresión Multi-salida, Consumo de Drogas, Random Forest, PCA, Rasgos de Personalidad.**

I. Descripción del Problema

El presente proyecto busca predecir el riesgo de consumo de 6 drogas distintas basándose en rasgos de personalidad y variables demográficas. Se utilizará el conjunto de datos *Drug Consumption (Quantified)* del repositorio UCI, el cual contiene registros de 1.885 encuestados. Cada registro incluye 12 atributos: variables demográficas (edad, género, educación, país, etnia) y métricas de personalidad (NEO-FFI-R, impulsividad y búsqueda de sensaciones). Las variables objetivo son los reportes de consumo para las 6 drogas, categorizados en siete niveles ordinales (CL0 a CL6) según la frecuencia de uso.

A diferencia de enfoques que simplifican el problema a una clasificación binaria (usuario/no usuario), este proyecto implementará un modelo de regresión *multi-salida (multi-output)*. El objetivo es estimar simultáneamente un índice de riesgo para cada una de las 6 drogas. Para lograrlo, la escala ordinal de consumo (CL0 a CL6) se transformará en un *índice numérico (0-6)*. Este índice, que representa la variable objetivo, cuantifica la frecuencia e intensidad del consumo, donde un valor mayor indica un riesgo más elevado.

El uso de Machine Learning (ML) es fundamental debido a la alta complejidad y no linealidad de las relaciones entre los 12 atributos y los 6 patrones de consumo. Un modelo estadístico tradicional (ej. regresión lineal múltiple) resultaría insuficiente, ya que asume que los rasgos contribuyen de forma simple y aditiva al riesgo. Esta suposición es improbable en este dominio; por ejemplo, el perfil de personalidad que predice el consumo de heroína difiere drásticamente del que predice el consumo de cafeína. Además, es probable que existan *interacciones complejas*, como que una "alta apertura a la experiencia" solo aumente el riesgo de consumo de LSD si se combina con una "baja escrupulosidad".

Por lo tanto, se emplearán algoritmos de ML (como Redes Neuronales, Random Forest o Support Vector Regression) *diseñados específicamente para inferir estos patrones latentes y manejar la alta dimensionalidad del problema*. El resultado será un modelo predictivo unificado que conecte el perfil de entrada con el vector de 6 índices de riesgo, ofreciendo así una herramienta robusta para la estratificación de riesgos y el diseño de intervenciones personalizadas.

II. Variables de Entrada y Variables a Predecir

El conjunto de datos proporciona información detallada sobre 1,885 participantes. Las variables se dividen en características de entrada (predictores) y variables de salida (objetivos).

A. Variables de Entrada (Features)

Las 12 variables de entrada (o *features*) utilizadas para construir el modelo combinan atributos demográficos y métricas de personalidad. El conjunto de datos original ya proporciona estas variables cuantificadas y normalizadas como valores reales, permitiendo su uso directo en los modelos.

1) Variables Demográficas
   a) **Age:** Edad del participante, agrupada en rangos etarios desde 18 hasta más de 65 años.
   b) **Gender:** Género del participante, codificado como masculino o femenino.
   c) **Education:** Nivel educativo alcanzado, desde abandono escolar antes de los 16 años hasta doctorado.
   d) **Country:** País de residencia actual del participante, incluyendo opciones como Reino Unido, EE. UU., Canadá, Australia, entre otros.
   e) **Ethnicity:** Etnia o grupo racial del participante (por ejemplo, blanco, negro, asiático o mixto).
2) Variables de Personalidad
   a) **Nscore (Neuroticism):** Medida del rasgo de neuroticismo, que refleja la tendencia a experimentar emociones negativas como ansiedad o depresión.
   b) **Escore (Extraversion):** Medida de extraversión, asociada con la sociabilidad, la energía y la búsqueda de estimulación externa.
   c) **Oscore (Openness to Experience):** Evalúa la apertura a la experiencia, que refleja curiosidad intelectual, imaginación y preferencia por la novedad.
   d) **Ascore (Agreeableness):** Mide la amabilidad, es decir, la empatía, cooperación y consideración hacia los demás.

e) **Cscore (Conscientiousness):** Indica el nivel de escrupulosidad o responsabilidad, asociado a la autodisciplina y el orden.
f) **Impulsive (Impulsiveness):** Evalúa la impulsividad mediante la escala BIS-11, reflejando la tendencia a actuar sin pensar.
g) **SS (Sensation Seeking):** Mide la búsqueda de sensaciones (ImpSS), es decir, la inclinación a buscar experiencias intensas, variadas y arriesgadas.

B. Variables a Predecir (Targets)

Las variables objetivo son los patrones de consumo de *6 sustancias distintas* (éxtasis, cannabis, heroína, cocaína, benzodiacepinas y LSD).

En el conjunto de datos original, cada una de estas 6 variables se presenta como una categoría ordinal de siete niveles (CL0 a CL6), que describe la frecuencia o recencia del consumo (ej., "nunca usado", "usado hace más de una década", ..., "usado en el último día").

Como se estableció en la descripción del problema (Sección I), este proyecto no simplificará esto a una clasificación binaria. En su lugar, para *cuantificar el nivel de riesgo*, la escala ordinal (CL0-CL6) se transformará en un *índice numérico discreto (0-6)*.

Este índice numérico será la variable objetivo. Por lo tanto, el problema se aborda como una regresión *multi-salida (multi-output)*, donde el modelo debe predecir simultáneamente un *vector de 6 valores* (un índice de 0 a 6 para cada droga), representando el perfil de riesgo de consumo del individuo.

C. Tipo de Configuración

El problema se aborda mediante un paradigma de aprendizaje supervisado, ya que se dispone de un conjunto de datos etiquetados que relacionan variables demográficas y psicológicas con el consumo de diferentes sustancias.

En este contexto, el modelo aprende a partir de ejemplos conocidos para predecir la probabilidad de consumo futuro o potencial, configurándose como un problema de regresión multi-salida.

De acuerdo con *Selvi & Chandrasekaran (2022)*, los métodos supervisados permiten "identificar patrones predictivos a partir de características psicológicas utilizando técnicas como Árboles de Decisión, Random Forest o SVM" (p. 4). Los autores enfatizan que este enfoque resulta adecuado para "analizar la influencia de los rasgos de personalidad sobre el comportamiento de consumo y construir modelos interpretables".

De forma complementaria, *Crumrine, Stivers & Helms (2024)* destacan que los Árboles de Decisión son herramientas especialmente útiles en este tipo de análisis, ya que "proporcionan modelos explicativos que muestran la relación entre variables de personalidad y patrones de uso de drogas" (p. 6). Además, el estudio señala que "la interpretabilidad de los árboles facilita la comprensión de los factores que más contribuyen al consumo", lo que los hace apropiados en contextos de investigación psicológica y social.

Por tanto, se adoptó un *paradigma de aprendizaje supervisado multi-salida* para predecir los índices de riesgo. Se implementaron y compararon cinco modelos distintos para evaluar su capacidad de generalización ante la complejidad de los datos:

1) Regresión Ridge (Modelo Lineal Base): Se estableció ese modelo lineal con regularización L2 como la línea base (*baseline*) del estudio. Su función es predecir el índice de riesgo como una variable continua, sirviendo de punto de referencia para evaluar si la complejidad adicional de los modelos no lineales aporta una mejora significativa en la precisión.

2) K-Vecinos Más Cercanos (kNN Regressor): Se utilizó la variante de regresión de este algoritmo basado en instancias. El modelo estima el riesgo individual calculando el promedio ponderado de los índices de consumo de los k usuarios vecinos con perfiles de personalidad más similares.

3) Redes Neuronales Artificiales (MLP Regressor): Se diseñó un Perceptrón Multicapa con una capa de salida lineal. Este modelo busca aprender un mapeo profundo entre los rasgos de entrada y los índices numéricos de riesgo, minimizando el error cuadrático medio (MSE) global.

4) Support Vector Regression (SVR): Se implementó la versión de regresión de las Máquinas de Soporte Vectorial. El objetivo de este modelo es encontrar un hiperplano que se ajuste a los valores de riesgo dentro de un margen de tolerancia (ϵ) específico, optimizando la capacidad de generalización en espacios de alta dimensión.

5) Random Forest Regressor: Se configuró como el modelo principal de ensamble. Al promediar las predicciones numéricas de múltiples árboles de decisión independientes, este enfoque reduce la varianza y ofrece una estimación robusta de la intensidad del consumo, mitigando el sobreajuste.

III. Estado del Arte

El análisis predictivo del consumo de drogas basado en rasgos de personalidad ha sido abordado desde distintas ópticas en el Machine Learning. **Es relevante destacar que los estudios seleccionados para esta revisión utilizan el mismo conjunto de datos empleado en este proyecto (*Drug Consumption Quantified* del repositorio UCI)**, lo que permite una comparación directa de las metodologías y métricas de desempeño sobre un mismo espacio muestral de 1.885 registros.

A continuación, se describen dos enfoques relevantes: el uso de arquitecturas híbridas para maximizar la precisión y el uso de modelos interpretables para gestionar datos desbalanceados.

A. Enfoque Híbrido de Redes Neuronales y Rough Sets

Una línea de investigación se centra en la optimización del rendimiento predictivo mediante arquitecturas híbridas. *Selvi & Chandrasekaran (2022)* proponen un modelo de aprendizaje supervisado que integra la Teoría de Conjuntos Difusos (*Rough Set Theory* - RST) con Redes Neuronales (NN).

En su trabajo, RST (específicamente el algoritmo QuickReduct) se utiliza como un paso de preprocesamiento para la selección de características. El objetivo es encontrar

un subconjunto mínimo de atributos relevantes, reduciendo la dimensionalidad y simplificando la estructura de la red neuronal posterior. El modelo de clasificación resultante, denominado RS-RPROP-NN, emplea una red Perceptrón Multicapa (MLP) entrenada con el algoritmo de Propagación Hacia Atrás Resiliente (RPROP).

Para la validación, los autores utilizaron una división simple (*split-set*) 60/40 (entrenamiento/prueba). El desempeño del sistema se evaluó priorizando la Entropía Cruzada (CE) —métrica que evalúa la disimilitud entre la distribución de probabilidad predicha y la real— sobre la precisión simple. El modelo RS-RPROP-NN obtuvo resultados notables, alcanzando valores de Entropía Cruzada significativamente bajos (entre 0.0039 y 0.0101 para las drogas estudiadas) y una precisión (*Accuracy*) final del 99.02%, superando a modelos convencionales.

B. Enfoque en Clasificación Sensible al Costo y Modelos de Conjunto

Otro enfoque, presentado por *Crumrine, Stivers, & Helms (2024)*, también utiliza el aprendizaje supervisado sobre la misma base de datos, pero se enfoca en la interpretabilidad de las reglas y la gestión de datos desbalanceados. Este estudio abordó el problema desde dos escenarios: 1) una clasificación binaria (consumo "Legal" vs. "Ilegal") y 2) una predicción de frecuencia de consumo.

Para el primer escenario, implementaron un Clasificador Sensible al Costo (*Cost-Sensitive Classifier*) basado en el algoritmo de reglas JRip. Este modelo asignó una penalización mayor a la clasificación incorrecta de un usuario de drogas ilegales (un Falso Negativo). Para el segundo escenario (predicción de frecuencia), recurrieron al aprendizaje de conjunto (*ensemble learning*), utilizando el algoritmo Vote que combinaba las predicciones de cinco clasificadores distintos (J48, JRip, OneR, entre otros).

Su metodología de validación fue una división simple (*split-set*) con una proporción 90/10. El modelo sensible al costo (JRip) obtuvo una alta precisión del 92.9% para la detección de consumo ilegal. Aunque el modelo de conjunto (Vote) para la frecuencia de cannabis alcanzó una precisión moderada (49.73%), los autores concluyeron que estos resultados apoyaban la hipótesis del cannabis como una "droga de entrada".

C. Benchmarking Exhaustivo y Estructuras de Correlación (Pleiades)

Como punto de partida fundamental, *Fehrman et al. (2017)* —autores responsables de la consolidación del conjunto de datos utilizado en este trabajo— realizaron una evaluación exhaustiva probando ocho algoritmos de clasificación distintos, incluyendo regresión logística, análisis discriminante lineal y k-vecinos más cercanos (kNN). Su objetivo fue identificar el subconjunto de características más eficaz para cada droga individual.

El estudio reveló que no existe un único clasificador universalmente superior; sin embargo, los Árboles de Decisión y los métodos de kernels mostraron un rendimiento consistente. Un hallazgo importante de su investigación fue la identificación de una "estructura modular" en el consumo de sustancias, denominadas "pleiades de correlación". Identificaron tres clústeres principales de consumo correlacionado: la pleiade de heroína (incluyendo crack y metadona), la pleiade de éxtasis (incluyendo anfetaminas y cannabis) y la pleiade de benzodiazepinas. En términos de métricas, lograron una sensibilidad y especificidad superior al 70% para la mayoría de las sustancias utilizando validación cruzada (LOOCV), estableciendo que los rasgos de personalidad del modelo de cinco factores (especialmente Neuroticismo alto y Amabilidad baja) son predictores robustos del riesgo de consumo.

D. Comparativa entre Aprendizaje Clásico y Redes Neuronales

En un estudio más reciente enfocado específicamente en el riesgo de consumo de marihuana, *Zoboroski, Wagner & Langhals (2021)* compararon el rendimiento de modelos clásicos (Regresión Logística y Árboles de Decisión) frente a Redes Neuronales (NN) profundas. A diferencia de los enfoques anteriores, este trabajo implementó una búsqueda iterativa multidimensional de hiperparámetros para optimizar la arquitectura de la red neuronal.

Sus resultados demostraron la superioridad de los modelos de aprendizaje profundo sobre los algoritmos clásicos en este dominio. El modelo de Red Neuronal optimizado alcanzó una sensibilidad de 0.87 y una exactitud (Accuracy) de 0.86, superando las métricas reportadas en la literatura previa (como la sensibilidad de 0.80 de Fehrman et al.). No obstante, los autores destacaron el valor de la Regresión Logística para la inferencia explicativa, confirmando mediante razones de probabilidad (odds ratios) que los individuos más jóvenes, con menor nivel educativo y altos puntajes en "búsqueda de sensaciones" y "apertura a la experiencia" presentan un riesgo significativamente mayor de consumo de THC.

IV. Entrenamiento y Evaluación de los Modelos

En esta sección se describe el diseño experimental implementado para el desarrollo de modelos predictivos orientados a estimar el nivel de consumo de seis tipos de drogas: *Cannabis, Cocaína, Heroína, Éxtasis, Benzodiacepinas y LSD*.

A. Preprocesamiento y Configuración de Datos

El conjunto de datos original (1,885 registros) se sometió a un proceso riguroso de partición y balanceo para asegurar la validez de los resultados:

1) División y Estratificación: Se realizó una división *Hold-Out* estratificada (80% entrenamiento, 20% prueba) utilizando como referencia la variable de consumo de *Heroína*. Esto garantiza que las clases minoritarias se mantengan representadas proporcionalmente en ambos subconjuntos, evitando sesgos en la evaluación.

2) Balanceo de Clases: Dado el desbalanceo severo en los niveles de consumo (0-6), como se evidencia en la Fig. 1, se aplicó la técnica de *Random OverSampling (ROS)* exclusivamente sobre el conjunto de entrenamiento. Esto incrementó el tamaño muestral de 1,500 a aproximadamente 9,000 instancias, permitiendo que los modelos aprendan patrones de las clases menos frecuentes sin contaminar el conjunto de prueba.

TABLE I
ESPACIO DE BÚSQUEDA DE HIPERPARÁMETROS POR TIPO DE MODELO

Tipo de Modelo	Algoritmo	Hiperparámetros / Valores Explorados
Paramétrico	Ridge Regression (Multi-Output)	α: [0.001, 0.01, 0.1, 1, 10, 50, 100] `solver`: [`auto`, `svd`, `cholesky`]
No Paramétrico	K-Nearest Neighbors (kNN)	`n_neighbors`: [3, 5, 7, 11, 15, 21] `weights`: [`uniform`, `distance`] `metric`: [`euclidean`, `manhattan`]
Ensamble	Random Forest Regressor	`n_estimators`: 50–300 (via Optuna) `max_depth`: 5–30 `min_samples_split`: 2–15
Red Neuronal	Multi-Task NN (PyTorch)	`learning_rate`: [0.0001, 0.001, 0.01] `hidden_dim`: [32, 64, 128, 256] `dropout`: [0.2, 0.3, 0.4, 0.5] `weight_decay`: [1e-5, 1e-4]
Vectores de Soporte	Support Vector Regression (SVR)	C (Penalización): 0.1–100 ϵ (Margen): 0.01–1.0 `gamma`: [`scale`, `auto`]

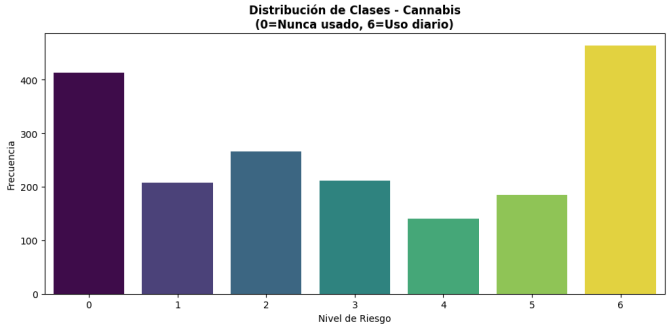

Fig. 1. Distribución de clases para la variable Cannabis. El marcado desbalance hacia las clases extremas (0 y 6) justifica la necesidad de técnicas de re-muestreo.

3) Validación Cruzada: Para la selección de modelos se utilizó un esquema de *k-fold cross-validation* (3 a 5 pliegues) optimizando el `neg_root_mean_squared_error`.

B. Espacio de Búsqueda de Hiperparámetros

Se evaluaron cinco familias de algoritmos bajo una estrategia híbrida: búsqueda en rejilla (*Grid Search*) para parámetros discretos y optimización bayesiana (*Optuna*) para espacios continuos. La Tabla I detalla la configuración final explorada.

C. Métricas de Evaluación

Para evaluar el desempeño se seleccionaron tres métricas complementarias que penalizan tanto el error continuo como el error de clasificación tras el redondeo de las predicciones:

1) RMSE (Root Mean Squared Error): Seleccionada como función de pérdida principal por su capacidad de penalizar cuadráticamente los errores grandes, crítico en la predicción de riesgo.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

2) MAE (Mean Absolute Error): Proporciona una interpretación directa del error promedio en unidades de la escala de riesgo (0-6).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

D. Resultados del Entrenamiento de Modelos

1) Regresión Ridge: Tras la optimización de hiperparámetros, se identificó un $\alpha = 100$ como el valor óptimo, superando al modelo base. Este modelo optimizado obtuvo resultados variables entre sustancias. Como se detalla en la Tabla II, el Cannabis mantiene un error moderado (RMSE 1.83), mientras que en sustancias con fuerte desbalance como la Heroína, el entrenamiento balanceado penaliza el error cuadrático (RMSE 2.19), sugiriendo una tendencia a sobreestimar el riesgo en no-consumidores para intentar capturar a los pocos usuarios reales.

TABLE II
DESEMPEÑO DEL MODELO RIDGE OPTIMIZADO ($\alpha = 100$)

Sustancia	RMSE
Cannabis	1.83
Cocaína	1.69
Éxtasis	1.55
Benzodiacepinas	1.98
Heroína	2.19
LSD	1.36

Las limitaciones de la linealidad se visualizan explícitamente en las superficies de regresión generadas por el notebook. Esta rigidez impide capturar patrones de consumo complejos o límites abruptos entre usuarios y no usuarios, lo que explica el alto error residual.

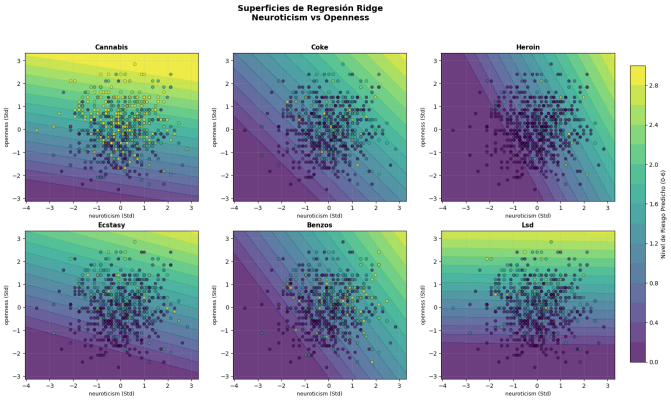

Fig. 2. Superficies de regresión Ridge (Neuroticism vs Openness). El modelo impone planos inclinados como predicción, limitando su capacidad para modelar comportamientos no lineales.

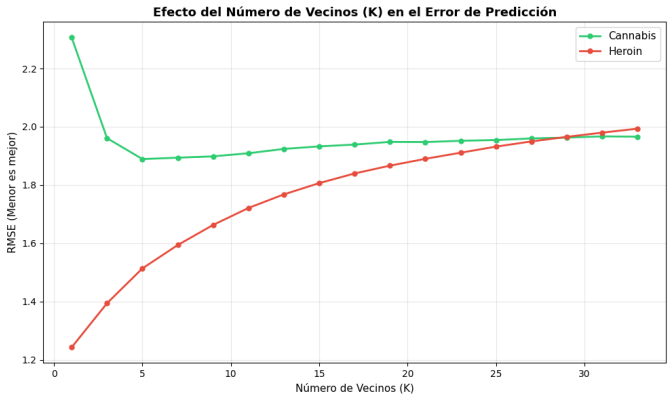

Fig. 3. Curva de sensibilidad a k (Cannabis vs Heroína). Se observa cómo la predicción de Heroína (rojo) se degrada al aumentar k, mientras que Cannabis (verde) mejora, validando la necesidad de hiperparámetros específicos por sustancia.

2) K-Nearest Neighbors (KNN): La optimización de hiperparámetros confirmó la sensibilidad del modelo al tamaño del vecindario (k). Mientras que para sustancias con mayor varianza de consumo como el Cannabis o la Cocaína el modelo generalizó mejor con vecindarios amplios ($k = 21$), para la Heroína se requirió un enfoque extremadamente local ($k = 3$) para maximizar la detección de casos positivos. A diferencia del modelo base, la búsqueda de cuadrícula seleccionó consistentemente la ponderación por distancia (*weights='distance'*) para todas las sustancias.

TABLE III
DESEMPEÑO DEL MODELO KNN OPTIMIZADO

Sustancia	**RMSE**	**Mejor k**
Cannabis	1.93	21
Cocaína	1.64	21
Éxtasis	1.65	21
Benzodiacepinas	2.04	15
Heroína	1.32	3
LSD	1.53	21

El KNN superó a la regresión Ridge en la identificación de patrones minoritarios críticos. Como se muestra en la Tabla III, el modelo alcanzó un RMSE de **1.32** para la Heroína, su mejor desempeño relativo. Esto sugiere que los usuarios de drogas duras tienden a agruparse en nichos muy específicos del espacio de características, los cuales se diluyen si se promedian con demasiados vecinos.

La Fig. 3 ilustra este comportamiento divergente: la curva de aprendizaje para la Heroína alcanza su mínimo error rápidamente con un k bajo y se deteriora al aumentar el vecindario, mientras que el Cannabis se beneficia de una mayor suavización.

3) Red Neuronal Artificial (Multi-Task): La arquitectura implementada consistió en un *backbone* compartido de 128 neuronas y cabezas específicas para cada sustancia. El modelo alcanzó un RMSE promedio global de **1.50** tras 60 épocas de entrenamiento. El uso de *Dropout* ($p = 0.4$) y regularización L2 fue necesario para controlar la complejidad del modelo, dado que el número de parámetros superaba la cantidad de muestras disponibles en las clases minoritarias.

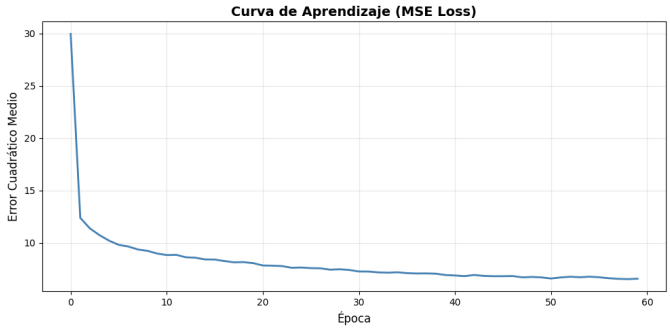

Fig. 4. Curva de aprendizaje (Training MSE). El modelo reduce el error de entrenamiento consistentemente, demostrando una convergencia estable sin oscilaciones severas.

TABLE IV
DESEMPEÑO DE LA RED NEURONAL MULTI-TASK

Sustancia	**RMSE**	**MAE**
Cannabis	1.85	1.56
Cocaína	1.48	1.25
Éxtasis	1.41	1.19
Benzodiacepinas	1.67	1.41
Heroína	1.31	1.05
LSD	1.25	1.00

La Fig. 4 muestra la convergencia del modelo durante el entrenamiento. Se observa una disminución monótona y suave del Error Cuadrático Medio (MSE), estabilizándose hacia la época 60. Esta estabilidad confirma que el optimizador (Adam, $lr = 0.001$) logró minimizar efectivamente el error sobre los datos de entrenamiento, aunque el desempeño final en prueba (Tabla IV) sugiere que la generalización sigue limitada por el tamaño del dataset.

4) Random Forest Regressor: Este modelo demostró ser el más robusto durante la fase experimental. La optimización bayesiana (Optuna) convergió hacia una arquitectura de **150**

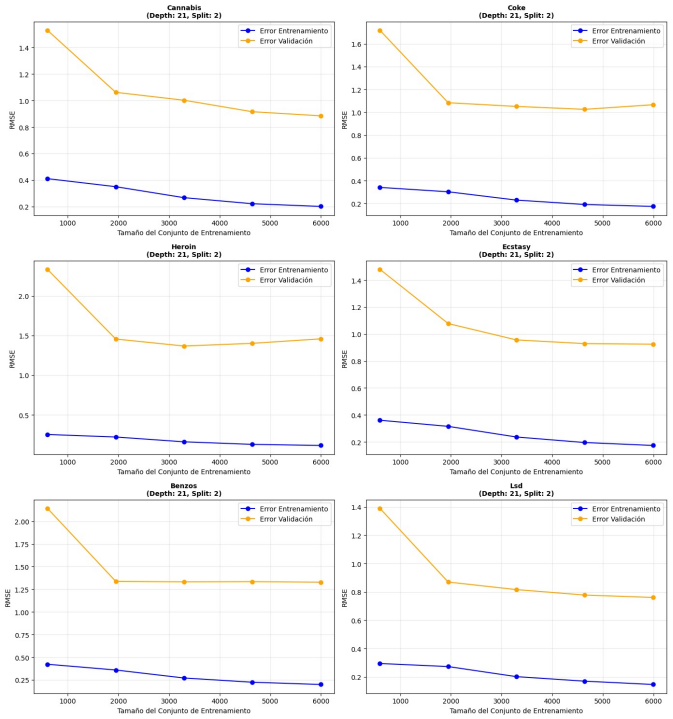

Fig. 5. Curvas de Aprendizaje por Droga. Se observa la convergencia del modelo, donde el error de validación se estabiliza rápidamente, confirmando la robustez de los hiperparámetros seleccionados ($N = 150$).

estimadores y una profundidad máxima de **21 niveles**. La capacidad del ensamble para manejar interacciones no lineales permitió reducir significativamente el error en las clases minoritarias, superando el desempeño de los modelos lineales previos.

TABLE V
DESEMPEÑO DEL MODELO RANDOM FOREST OPTIMIZADO

Sustancia	RMSE
Cannabis	2.07
Cocaína	1.47
Éxtasis	1.41
Benzodiacepinas	1.68
Heroína	1.10
LSD	1.26

Como muestra la Tabla V, Random Forest obtuvo el mejor RMSE global para Heroína (1.10) y LSD (1.26). La Fig. 5 presenta las curvas de aprendizaje generadas por el notebook, donde se evidencia una brecha estable entre el error de entrenamiento y validación. Esto indica que, aunque el modelo tiene capacidad suficiente para aprender los datos de entrenamiento (línea azul descendente), la generalización (línea naranja) está acotada por la complejidad intrínseca de los patrones de consumo, evitando el sobreajuste extremo observado en la red neuronal.

5) Máquina de Vectores de Soporte (SVR): La optimización bayesiana para el problema completo de 7 clases arrojó una configuración de alta regularización, con un coeficiente de penalización $C \approx 97.29$ y un margen de insensibilidad $\epsilon \approx 0.29$. A pesar del uso de un kernel RBF para capturar no linealidades, el modelo SVR tuvo dificultades significativas para modelar la granularidad fina de la escala original, resultando en errores cuadráticos elevados comparados con otros modelos.

TABLE VI
DESEMPEÑO DEL MODELO SVR OPTIMIZADO

Sustancia	RMSE
Cannabis	2.04
Cocaína	1.67
Heroína	1.30
Éxtasis	1.80
Benzodiacepinas	1.92
LSD	1.51

Como se detalla en la Tabla VI, el SVR no logró reducir el RMSE por debajo de 1.3 en ninguna sustancia. Su mejor desempeño relativo fue en la Heroína (RMSE 1.30), impulsado principalmente por la correcta identificación de la clase mayoritaria (no usuarios).

La Figura 6 muestra la matriz de confusión para la Heroína generada por el modelo. Se observa que, si bien el SVR clasifica correctamente a los no consumidores (clase 0), falla al distinguir entre los niveles intermedios de consumo (clases 1-6), dispersando las predicciones a lo largo de la diagonal. Esto confirma que la regresión de vectores de soporte es sensible al ruido en escalas ordinales con muchas categorías y desbalance severo.

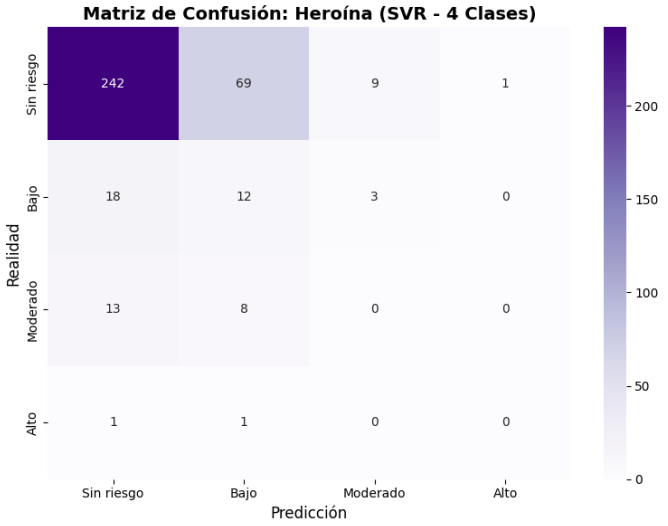

Fig. 6. Matriz de confusión para Heroína (SVR). El modelo muestra una alta eficacia en la clase 0 pero pierde precisión en los niveles de consumo positivo, dispersando las predicciones.

V. REDUCCIÓN DE DIMENSIÓN

A. Análisis Individual de Variables

Antes de aplicar técnicas de compresión global, se realizó una evaluación de la capacidad discriminativa individual de las

TABLE VII
IMPACTO DE PCA (11 COMPONENTES) EN EL ERROR LOGARÍTMICO

Modelo	RMSLE Original	RMSLE con PCA	Variación
Random Forest	0.620	0.626	+0.9% (Estable)
Red Neuronal	0.667	0.654	-1.9% (Mejora)

33 variables de entrada. Utilizando la importancia de características del Random Forest (Gini importance) y el coeficiente de Información Mutua, se detectó una fuerte jerarquía en la relevancia de los predictores.

Como se observa en el ranking de variables más influyentes (Fig. 7), existe un subconjunto crítico de rasgos psicológicos que dominan la predicción:

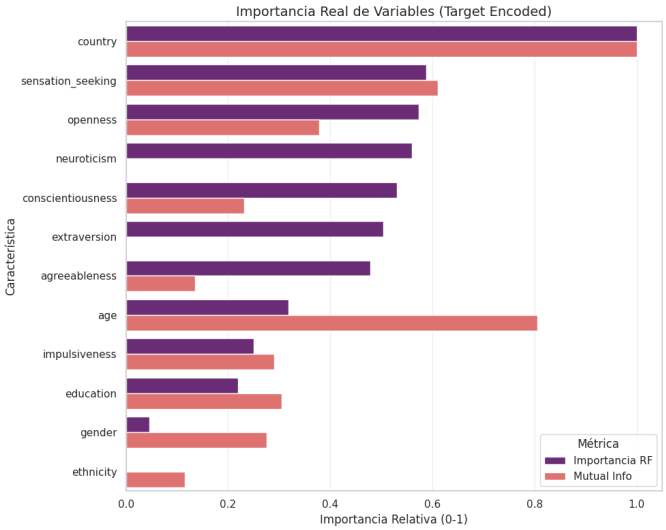

Fig. 7. Importancia de características (Random Forest). La búsqueda de sensaciones (ss) es el factor determinante, superando significativamente al resto de variables psicométricas y demográficas.

- **Variables Dominantes:** *Sensation Seeking* (Importancia RF: 0.58) y *Openness* (0.57) mostraron consistentemente la mayor correlación con el consumo de sustancias ilícitas.
- **Variables Redundantes:** Rasgos como *Agreeableness* o *Gender* presentaron contribuciones marginales (Importancia < 0.05). Además, la alta colinealidad entre los "Big Five" sugiere que el espacio de 33 dimensiones está inflado artificialmente, justificando la reducción.

B. Extracción de Características Lineal (PCA)

Se aplicó el Análisis de Componentes Principales (PCA) sobre los datos estandarizados. El criterio de selección fue conservar el **90% de la varianza explicada**, lo cual se logró reteniendo las primeras **11 componentes principales**. Esto implica una compresión del 66.7% respecto al espacio original de 33 variables.

La efectividad de esta transformación se evaluó re-entrenando los dos mejores modelos (Random Forest y Red Neuronal) en este nuevo subespacio latente. Los resultados en la Tabla VII demuestran una estabilidad notable.

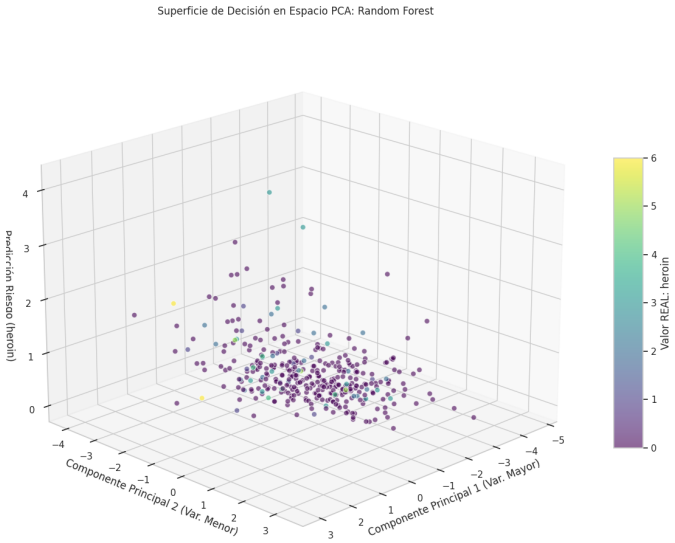

Fig. 8. Superficie de decisión del Random Forest (Heroína) en el espacio PCA. El modelo logra construir un gradiente de riesgo suave y coherente (eje Z) utilizando solo las proyecciones lineales, lo que confirma que la estructura predictiva principal se conserva intacta tras la reducción.

El hecho de que la Red Neuronal incluso mejorara ligeramente su rendimiento (RMSLE bajó de 0.667 a 0.654) sugiere que PCA actuó como un filtro de ruido efectivo, eliminando la varianza no informativa de las variables demográficas irrelevantes.

Para validar visualmente la calidad de este espacio reducido, se generó la superficie de decisión del Random Forest proyectada sobre los tres primeros componentes principales (Fig. 8).

C. Extracción de Características No Lineal (UMAP)

Buscando capturar relaciones más complejas, se aplicó UMAP (*Uniform Manifold Approximation and Projection*) configurado con $n\_neighbors = 15$ y $min\_dist = 0.1$, reduciendo el espacio a **10 dimensiones**.

A diferencia de PCA, esta transformación no lineal resultó perjudicial para la tarea de regresión. Como se observa en la Tabla VIII, el error del Random Forest se disparó un 22%, pasando de 0.620 a 0.762.

La Figura 9 ayuda a explicar este fenómeno. UMAP agrupa exitosamente a los individuos en clusters compactos (preservando la estructura local), lo cual es excelente para clasificación categórica o visualización. Sin embargo, al distorsionar las distancias globales para mantener esa cohesión local, rompe la linealidad métrica que los modelos de regresión necesitan para estimar la *magnitud* del consumo (e.g., distinguir entre nivel 3 y nivel 4).

TABLE VIII
IMPACTO DE UMAP (10 COMPONENTES) EN EL ERROR LOGARÍTMICO

Modelo	RMSLE Original	RMSLE con UMAP	Variación
Random Forest	0.620	0.762	+22.0% (Degradación)
Red Neuronal	0.667	0.684	+2.5% (Degradación)

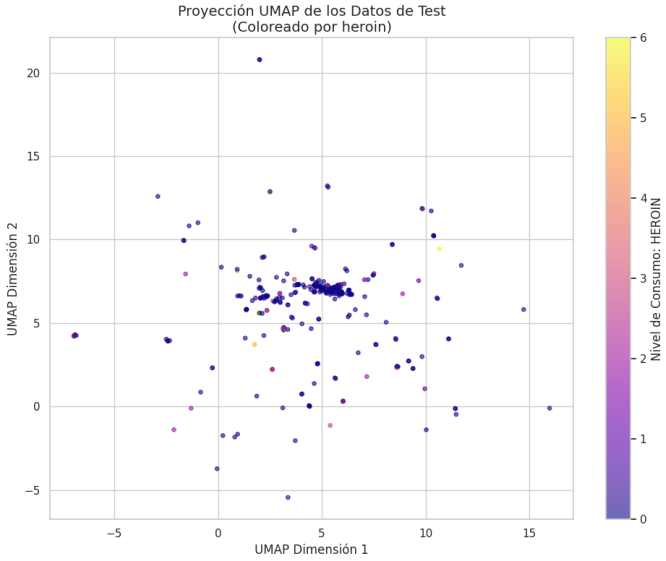

Fig. 9. Proyección UMAP 2D coloreada por nivel de consumo de Heroína. Aunque separa bien a los no-usuarios (puntos morados), la topología resultante es fragmentada, dificultando que el regresor trace una función continua de riesgo.

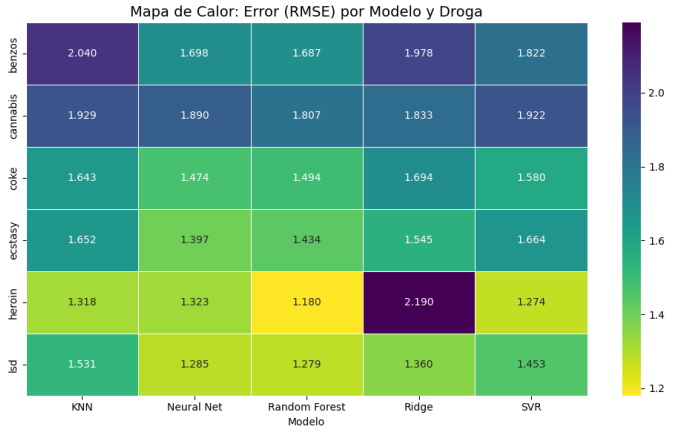

Fig. 10. Mapa de Calor de RMSE por Modelo y Sustancia. La columna del Random Forest muestra la menor intensidad de error acumulado. Nótese la dificultad transversal de todos los modelos para predecir Cannabis y Benzos (colores claros).

VI. DISCUSIÓN Y CONCLUSIONES

El presente estudio evaluó exhaustivamente la capacidad de predecir el nivel de consumo de 7 sustancias psicoactivas utilizando un enfoque de regresión sobre 7 clases ordinales.

A. Comparación de Modelos y Complejidad

El **Random Forest Regressor** se consolidó como el modelo más robusto, alcanzando un RMSE promedio global de **1.48** en la escala de 0-6. Superó sistemáticamente a modelos lineales como Ridge (RMSE 1.77) y SVR (RMSE 1.62), así como a la Red Neuronal Multi-Task (RMSE 1.51).

La Figura 10 resume el desempeño comparativo. Se evidencia que las drogas "duras" (Heroína, LSD) presentan menores errores de predicción (zonas oscuras), lo que indica que sus consumidores poseen perfiles de personalidad más distintivos y fáciles de modelar que los consumidores de sustancias de uso más extendido y social como el Cannabis o el Alcohol.

B. Conclusiones Finales

1) **Límite Predictivo del Problema de 7 Clases:** A pesar de la optimización de hiperparámetros, el error residual (RMSE $\approx$ 1.5) indica que predecir la frecuencia exacta de consumo es una tarea con mucho ruido aleatorio. La granularidad de 7 niveles excede la resolución informativa de los tests psicométricos estándar.
2) **Eficacia de la Reducción Lineal:** PCA demostró ser una herramienta valiosa para producción. Es posible sustituir el cuestionario extenso de 33 preguntas por una versión reducida que capture los 11 componentes principales, ahorrando tiempo de recolección de datos sin sacrificar precisión.
3) **Inviabilidad de Manifold Learning para Regresión:** Aunque UMAP es potente para visualización exploratoria, su uso como pre-procesador para regresión numérica está contraindicado en este dominio, ya que la distorsión de la métrica global introduce más error del que resuelve.

REFERENCES

[1] C. Crumrine, C. Stivers, and C. Helms, "Unraveling connections: Personality traits, drug usage frequency", Furman Univ., Greenville, SC, USA, CSC-272 Final Project Rep., Apr. 2024.
[2] S. Selvi and M. Chandrasekaran, "Detection of drug abuse using rough set and neural network-based elevated mathematical predictive modelling", *Neural Process. Lett.*, vol. 55, no. 3, pp. 2633–2660, Nov. 2022, DOI: 10.1007/s11063-022-11086-z.
[3] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The five factor model of personality and evaluation of drug consumption risk", *arXiv preprint arXiv:1506.06297*, Jun. 2015.
[4] L. Zoboroski, T. Wagner, and B. Langhals, "Classical and neural network machine learning to determine the risk of marijuana use", *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, Art. no. 7466, Jul. 2021, DOI: 10.3390/ijerph18147466.