

# Clasificación de consumo de cannabis mediante variables psicológicas y demográficas

Laboratorio de Aprendizaje Estadístico

- Guillermo Casillas Copca
- Francisco J. Ramírez Ávila
- Juan A. Morales R.



# Contexto y origen de los datos

Datos extraídos de "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk" (Fehrman et al., 2015), disponible en UCI Machine Learning Repository (Drug Consumption, Quantified). El objetivo: aprovechar variables psicológicas y demográficas para predecir consumo de cannabis.

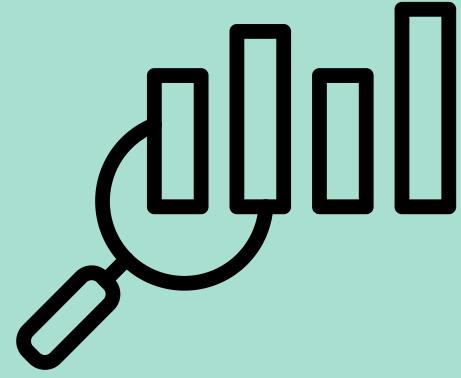


## Descripción

### Visión general

- Total: 1,885 observaciones.
- 32 variables originales; para este estudio se usan 13 predictoras numéricas.
- Variables: demográficas y psicológicas.
- Target: Cannabis binario (1 = usuario, 0 = no usuario).
- Desbalanceo de clases: Doble de proporción en usuarios respecto a no usuarios

# EDA - Variables Utilizadas



## Demográficas

- ID: identificador único de cada persona.
- Age: edad de la persona.
- Gender: género de la persona (por ejemplo: masculino, femenino, otro).
- Education: nivel de educación alcanzado (por ejemplo: secundaria, universidad, posgrado).
- Country: país de residencia de la persona.
- Ethnicity: grupo étnico al que pertenece la persona.

## Psicológicas

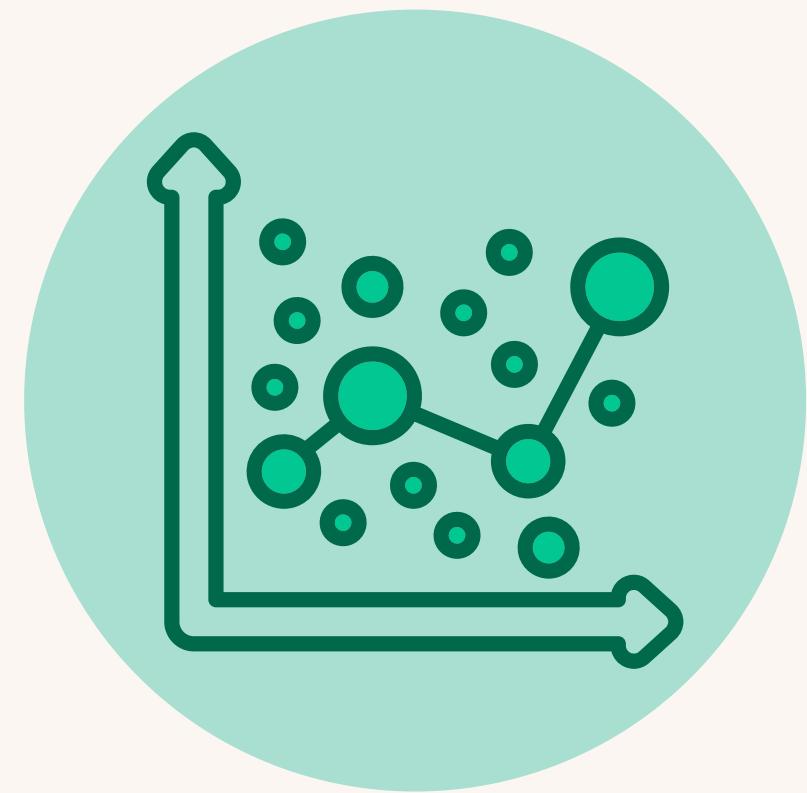
- Nscore (Neuroticism): mide la estabilidad emocional.
- Escore (Extraversion): mide la sociabilidad y energía.
- Oscore (Openness): mide la apertura a nuevas experiencias.
- Ascore (Agreeableness): mide la amabilidad y cooperación.
- Cscore (Conscientiousness): mide la responsabilidad y organización.
- Impulsive: mide el nivel de impulsividad (qué tanto actúa sin pensar).
- SS (Sensation Seeking): mide la búsqueda de sensaciones nuevas o intensas.

# Modelo 1: Regresión Logística

Se usó regresión logística con regularización para evitar sobreajuste y obtener coeficientes interpretables.

Pipeline: StandardScaler + validación cruzada 10-Fold + BayesSearchCV

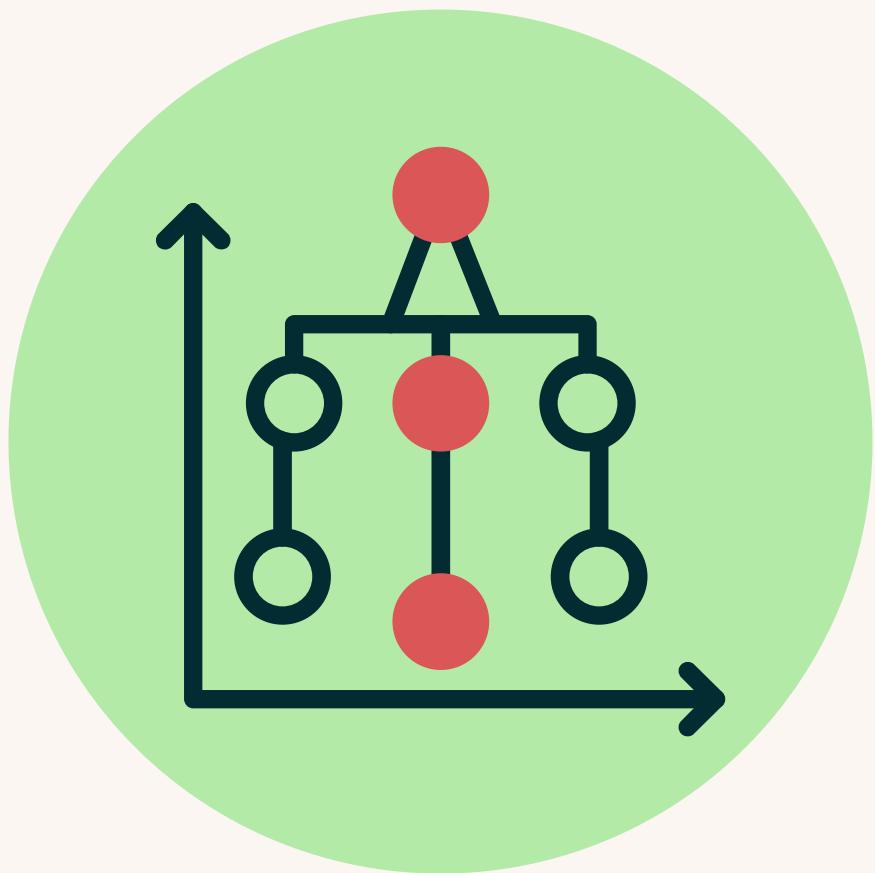
- Busqueda de hiperparámetros: C (0.001, 100), penalty (l1, l2) y solver (liblinear, saga)
- Mejor configuración: C = 0.1639, penalty = L1, solver = saga.
- AUC (mejor): 0.8834 — AUC 10-Fold promedio: 0.8834 (std 0.0190).



## Modelo 2 – SVM con kernel RBF

Máquina de Vectores de Soporte con kernel RBF para capturar relaciones no lineales. Preprocesado idéntico y optimización bayesiana de hiperparámetros (C, gamma, class\_weight).

- Busqueda de hiperparámetros: C(0.001 a 1000) y gamma(0.001 a 10).
- Mejores hiperparámetros: C = 0.4863, gamma = 0.0178, class\_weight = balanced.
- AUC promedio (10-Fold): 0.8887 (std 0.0327).
- AUC total en dataset: 0.8914 — rendimiento más alto entre los modelos.



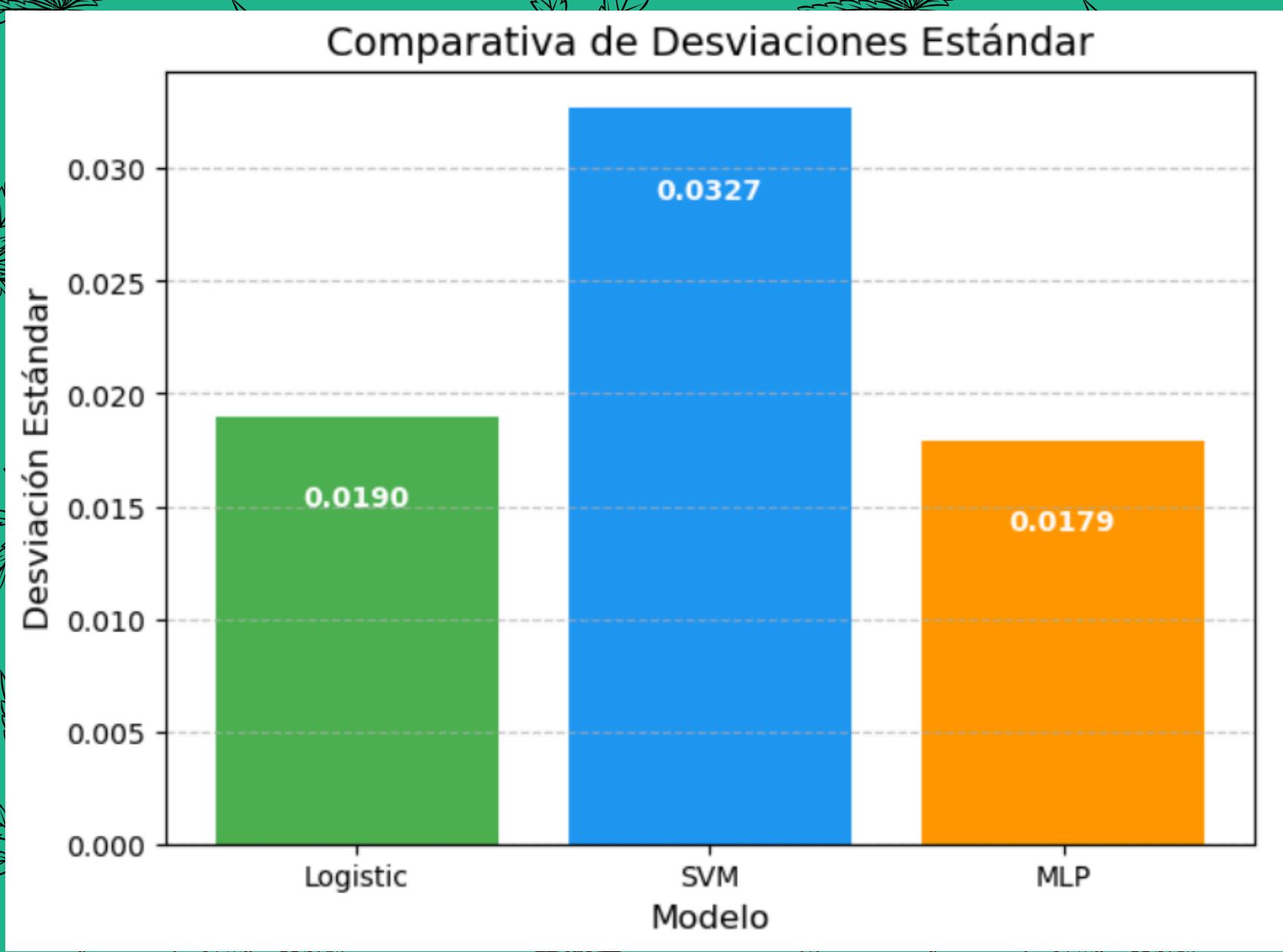
# Modelo 3 – Red Neuronal Multicapa

Perceptrón Multicapa para modelar interacciones complejas. Optimización bayesiana del tamaño de capas; pipeline con escalado y evaluación por 10-Fold CV.

- Busqueda de Hiperparametros: 2 Capas ocultas  
Capa 1:[10,100], Capa 2:[5,50]
- Mejor arquitectura: [10, 5] (dos capas ocultas).
- Mejor AUC: 0.8792 — AUC promedio: 0.8340 (std 0.0179).
- Resultado competitivo, pero menor que SVM para este conjunto de datos.

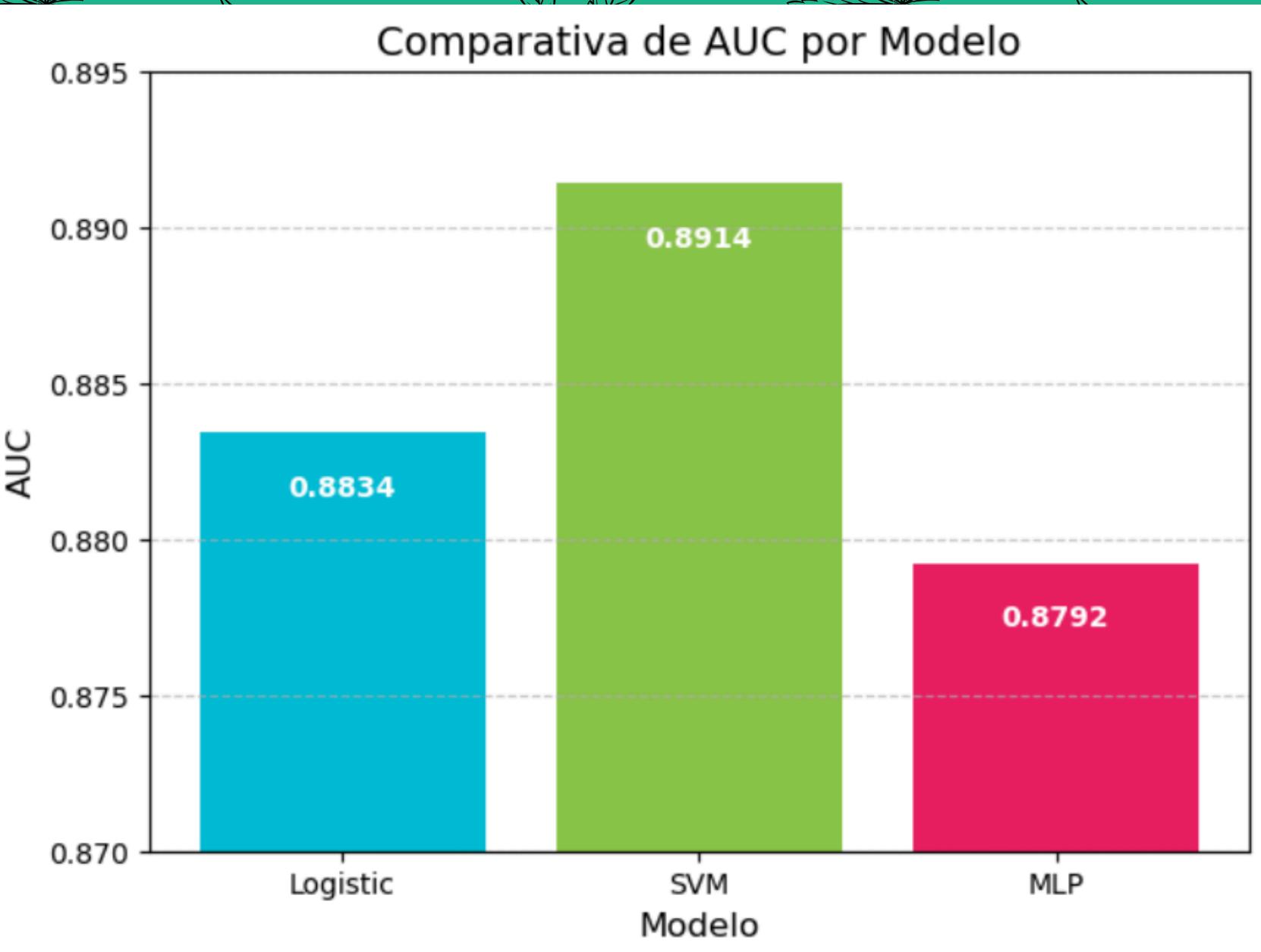


# Comparativa Desviación Estándar



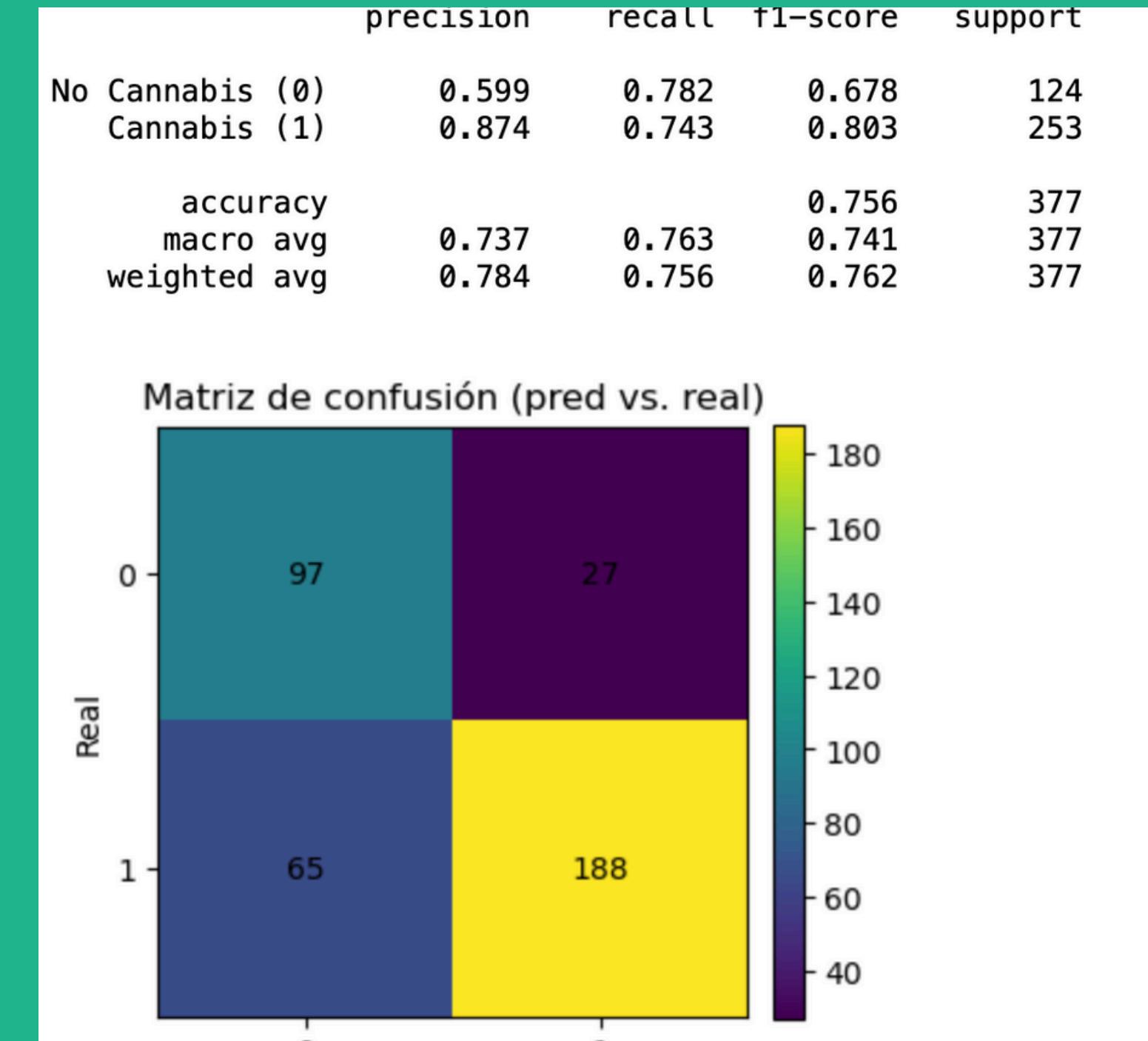
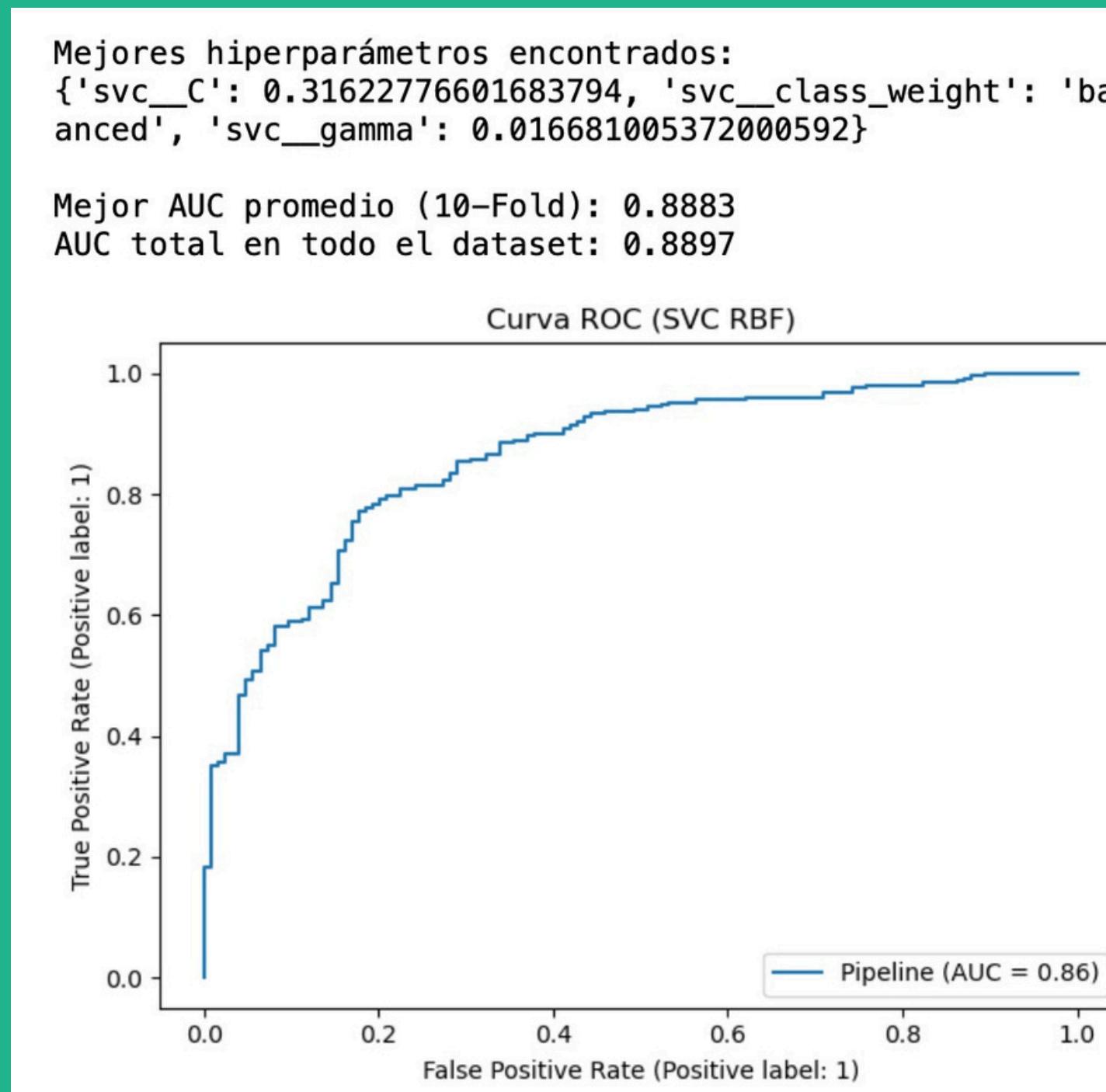
**SVM (RBF) tiene la desviación estándar más alta**

# Comparativa AUC



SVM (RBF) obtiene la AUC más alta

# Mejor modelo:

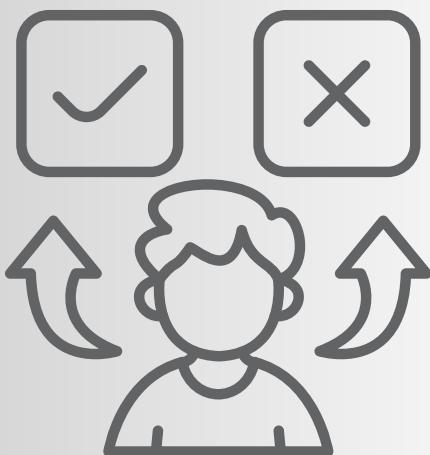


# Conclusiones

Nuestro mejor modelo fue el SVM con kernel RBF, con un AUC de 0.8914, lo que representa un excelente indicador de desempeño al distinguir entre consumidores y no consumidores de cannabis.

Las variables que más aportaron a la predicción fueron principalmente los rasgos de personalidad relacionados con la impulsividad y la búsqueda de sensaciones, ya que reflejan comportamientos asociados a la toma de riesgos y la curiosidad por nuevas experiencias.

En cambio, las variables menos relevantes fueron algunas de tipo demográfico, como el país o la etnicidad, lo que puede deberse a que estos factores no influyen directamente en las decisiones individuales de consumo, o su efecto queda opacado por las características psicológicas.



# Link del repositorio:

[MemoCasillasC/Proyecto\\_02: Notebook de Jupyter sobre el Proyecto 2](#)