UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
T U N J A
VIGILADA MINEDUCACIÓN - SNIES 1732

ANÁLISIS ESTADÍSTICO: Drug Consumption Dataset[1]

Nicolas Javier Carreño Perea
Emmanuel Mosquera Casas

Universidad Santo Tomás

***Abstract*— Este proyecto evidencia la práctica e implementación de una serie de algoritmos de *machine learning* para determinar la mejor clasificación binaria de clases para cada atributo. Así mismo, se podrá entrenar el modelo para que realice una evaluación del riesgo de ser consumidor de narcóticos.**
***Index Terms*—Aprendizaje automático, clasificación multivariable, consumo de narcóticos.**

I. INTRODUCCIÓN

La creciente tasa de consumo de drogas es un problema grave a nivel mundial, y el riesgo de recaer en ellas está presente en cualquier sociedad y cualquier rango de edad. Sin embargo, las razones de este aumento no se investigan a fondo. Por lo tanto, es muy importante predecir el riesgo de consumir que podría tener un paciente específico y diseñar un plan de tratamiento adecuado en consecuencia. Este dataset contiene registros de 1885 encuestados, sobre sus atributos, que incluyen los rasgos del modelo de cinco factores, la impulsividad, la búsqueda de sensaciones y otras características demográficas, y el historial de consumo de drogas psicoactivas del sistema nervioso central.

Objetivo:

a. General

Analizar la información y determinar un modelo de machine learning lo suficientemente preciso para clasificar el riesgo de un consumidor de ser consumidor de narcóticos.

b. Específicos:

- Identificar las variables de mayor correlación del dataset.
- Normalizar la información del dataset.
- Aplicar distintos modelos de *machine learning* clásico y determinar cuál se acopla mejor al dataset dado.

II. DESCRIPCIÓN DEL DATASET

Este dataset contiene registros de 1885 encuestados pertenecientes a países de la Commonwealth, Irlanda y Estados Unidos. Para cada encuestado se conocen 12 atributos: Medidas de personalidad que incluyen el NEO-FFI-R (Inventario de Personalidad Neo Revisado) que examina los cinco rasgos de personalidad de una persona como lo son: el neuroticismo, extraversión, apertura a la experiencia, agradabilidad y la concienciación, el BIS-11 (impulsividad) y el ImpSS (búsqueda de sensaciones), el nivel de estudios, la edad, el sexo, el país de residencia y la etnia. Todos los atributos de entrada son originalmente categóricos y se cuantifican. Tras la cuantificación, los valores de todas las características de entrada pueden considerarse de valor real. Además, se preguntó a los participantes sobre su consumo de 18 drogas legales e ilegales (alcohol, anfetaminas, nitrito de amilo, benzodiacepina, cannabis, chocolate, cocaína, cafeína, crack, éxtasis, heroína, ketamina, euforizantes legales, LSD, metadona, setas, nicotina y abuso de sustancias volátiles, así como una droga ficticia (Semeron) que se introdujo para identificar a los consumidores excesivos. Para cada droga tienen que seleccionar una de las respuestas: nunca ha consumido la droga, la ha consumido hace más de una década o en la última década, año, mes, semana o día. [1]

[1]Proyecto correspondiente a la asignatura de MACHINE LEARNING, presentado a Martha Susana. Fecha: 28 de noviembre de 2021.

Este estudio tiene limitaciones, ya que la muestra recogida estaba sesgada con respecto a la población general, pero seguía siendo útil para la evaluación del riesgo. Por lo tanto, parece que el objetivo razonable de esta modelización ML podría ser probar si es posible predecir el consumo de drogas e identificar los atributos más informativos. El dataset contiene 18 problemas de clasificación. Cada una de las variables de etiqueta independientes contiene siete clases: "Nunca usó", "Usó hace más de una década", "Usó en la última década", "Usó en el último año", "Usó en el último mes", "Usó en la última semana" y "Usó en el último día". El problema puede transformarse en una clasificación binaria mediante la unión de parte de las clases en una nueva clase. Por ejemplo, "Nunca usado", "Usado hace más de una década" forman la clase "No usuario" y todas las demás clases forman la clase "Usuario".

a) Inventario de personalidad Neo inventado:

Históricamente, el desarrollo del NEO PI-R revisado comenzó en 1978 con la publicación de un inventario de personalidad por parte de Costa y McCrae. El Inventario de Personalidad NEO Revisado (NEO PI-R) es un inventario de personalidad que examina los cinco grandes rasgos de personalidad de una persona (apertura a la experiencia, concienciación, extraversión, amabilidad y neuroticismo). Además, el NEO PI-R también informa sobre seis subcategorías de cada uno de los Cinco Grandes rasgos de la personalidad (llamadas facetas).

Neuroticismo	Extraversión	Apertura a la experiencia	Agradabilidad	Concienciación
Ansiedad	Calidez / amabilidad	Fantasía / Imaginación	Confianza (en otros)	Competencia / Autoeficacia
Hostilidad/Ira	Gregarismo	Estética / Interés artístico	Sencillez / Moralidad	Orden / Organización
Depresión	Asertividad	Sentimientos / Emocionalidad	Altruismo	Obligación / Sentido del deber / Obligación
Autoconciencia	Actividad (temperamento muy animado)	Acciones / Aventura / Exploración	Cumplimiento / Cooperación	Esfuerzo de logros
Impulsividad / inmoderación	búsqueda de emoción	Ideas / interés intelectual [curiosidad]	Modestia	Autodisciplina / Fuerza de voluntad
Vulnerabilidad al estrés / Miedo [indefensión aprendida]	Emoción positiva / Alegría / Vivacidad	Valores / Liberalismo psicológico / Tolerancia a la ambigüedad	Sensibilidad / simpatía	Deliberación / Cautela

b) Descripción de las variables:

1. ID es el número de registro en la base de datos original. No puede estar relacionado con el participante. Solo puede utilizarse como referencia.
2. Age (Real) es el grupo del participante y está agrupada en intervalos de edad.
3. Gender (Real) es el género del participante.
4. Education (Real) es el nivel de educación del participante y tiene uno de los valores.
5. Country (Real) es el país de residencia actual del participante.

6. Ethnicity (Real) es la etnia del participante.
7. Neuroticismo (Real) es el valor de esta variable en el NEO-FFI-R.
8. Extraversión (Real) es el valor de esta variable en el NEO-FFI-R.
9. Apertura a la experiencia (Real) es el valor de esta variable en el NEO-FFI-R.
10. Agradabilidad (Real) es el valor de esta variable en el NEO-FFI-R.
11. Concienciación (Real) es el valor de esta variable en el NEO-FFI-R.
12. Impulsividad (Real) es el valor de esta variable en el NEO-FFI-R.
13. Búsqueda de sensaciones (Real) es el valor de esta variable en el NEO-FFI-R.
14. Alcohol es una variable categórica, es decir el participante consume o no alcohol en que frecuencia.
15. Anfetamina es una variable categórica, es decir el participante consume o no Anfetamina en que frecuencia.
16. Nitrito de amilo es una variable categórica, es decir el participante consume o no Nitrito de amilo en que frecuencia.
17. Benzodiazepina es una variable categórica, es decir el participante consume o no Benzodiazepina en que frecuencia.
18. Cannabis es una variable categórica, es decir el participante consume o no Cannabis en que frecuencia.
19. Cocaína es una variable categórica, es decir el participante consume o no Cocaína en que frecuencia.
20. Crack es una variable categórica, es decir el participante consume o no Crack en que frecuencia.
21. Éxtasis es una variable categórica, es decir el participante consume o no Éxtasis en que frecuencia.
22. Heroína es una variable categórica, es decir el participante consume o no Heroína en que frecuencia.
23. Ketamina es una variable categórica, es decir el participante consume o no Ketamina en que frecuencia.
24. Drogas legales es una variable categórica, es decir el participante consume o no Drogas legales en que frecuencia.
25. LSD es una variable categórica, es decir el participante consume o no LSD en que frecuencia.
26. Metadona es una variable categórica, es decir el participante consume o no Metadona en que frecuencia.
27. Hongos es una variable categórica, es decir el participante consume o no Hongos en que frecuencia.
28. Semeron es una variable categórica, es decir el participante consume o no Semeron en que frecuencia.
29. VSA es una variable categórica, es decir el participante consume o no de sustancias volátiles en que frecuencia.

(*Nota: Aquí ya se descartan las no consideradas drogas: Nicotina, Cafeína, etc.*)

III. PROCEDIMIENTO Y ANÁLISIS

Se obtienen los datos del dataset y nos damos cuenta de que los datos de edad, género, educación, país y etnia están codificados con un número real:

Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Impulsiveness	Sensation_s
0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1
-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0
0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0
-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1
0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0

Estas columnas quedan de esta manera (*Nota: El proceso se detalla en la sección Tratamiento y normalización de los datos del notebook.*):

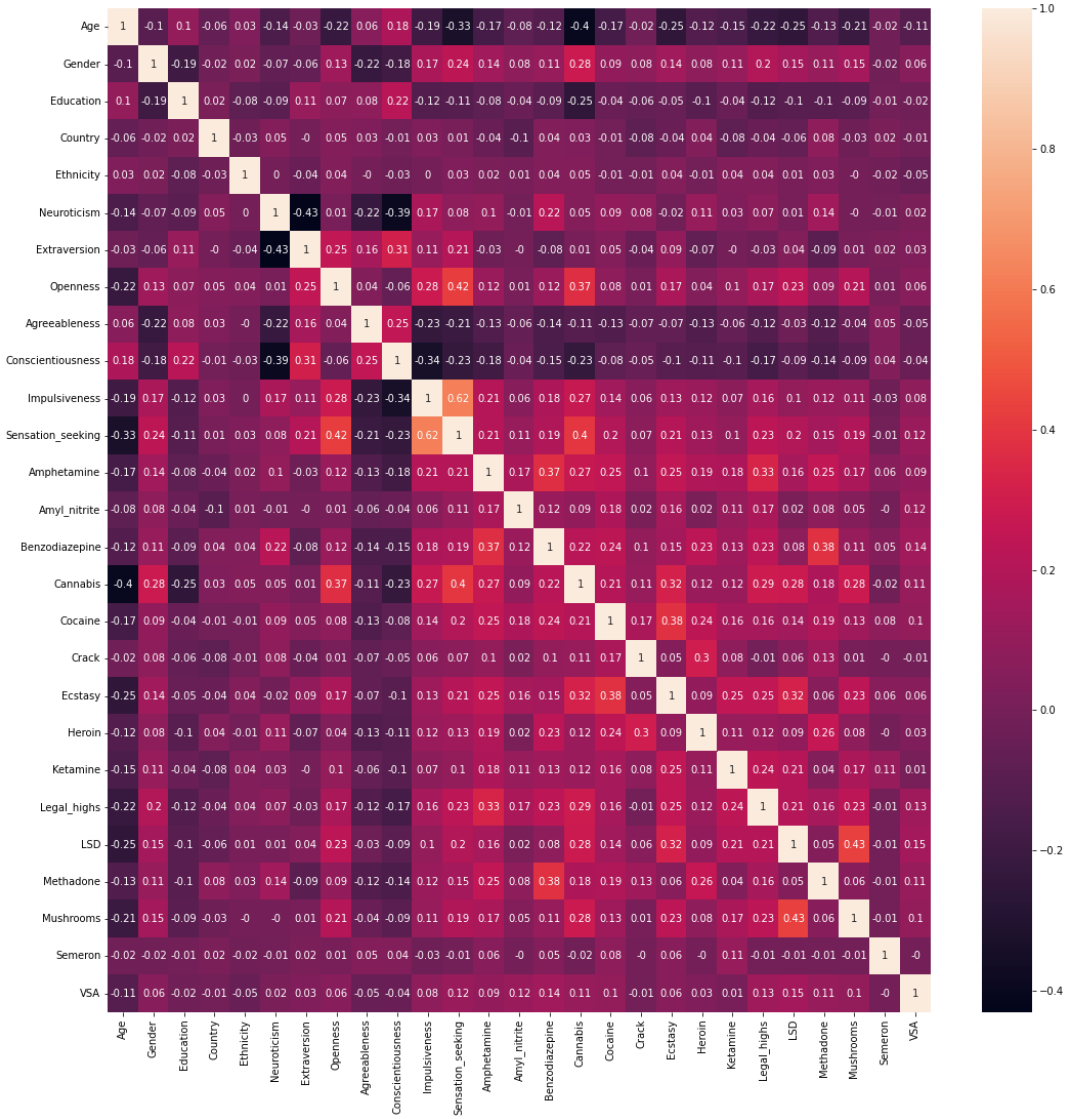
Age	Gender	Education	Country	Ethnicity
2	0	5	5	3
1	1	8	5	6
2	1	5	5	6
0	0	7	5	6
2	0	8	5	6

Se realiza una binarizacion de los datos, para pasar de 7 clases de frecuencia de consumo a solo dos (*Nota: El proceso se detalla en la sección Tratamiento y normalización de los datos del notebook.*):

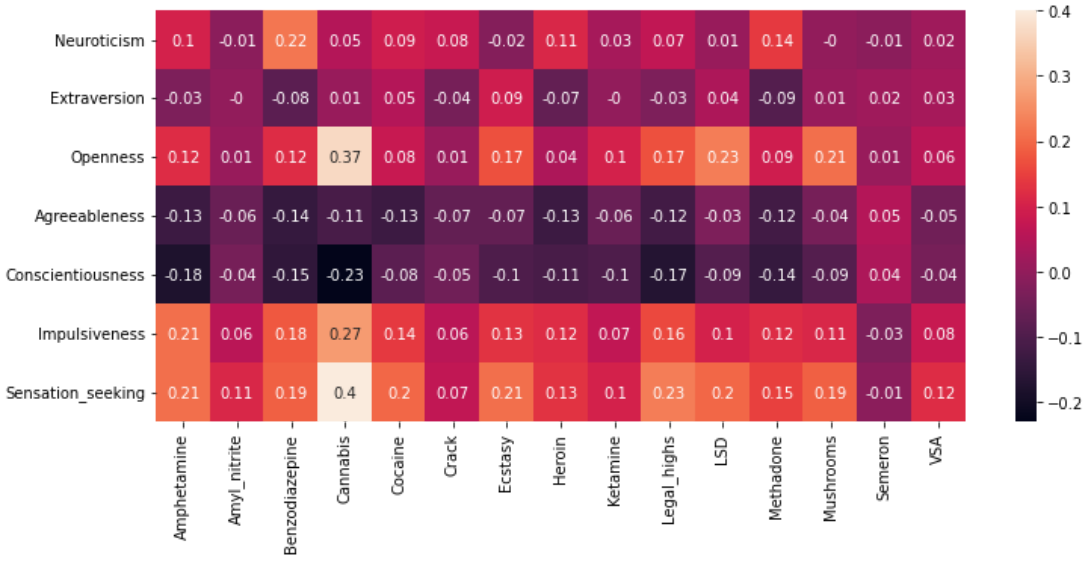
Amphetamine	Amyl_nitrite	Benzodiazepine	Caffeine	Cannabis	Chocolate	Cocaine	Crack	Ecstasy	Heroin	Ketamine	Legal_highs	LSD	Methadone
CL2	CL0	CL2	CL6	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4	CL0	CL2	CL0	CL2	CL3
CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0
CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0	CL0	CL2	CL0	CL0	CL0
CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1	CL0	CL0	CL1	CL0	CL0

Amphetamine	Amyl_nitrite	Benzodiazepine	Cannabis	Cocaine	Crack	Ecstasy	Heroin	Ketamine	Legal_highs	LSD	Methadone	Mushrooms	Semeron
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

A continuación, se proyecta la gráfica de la matriz de correlación principal:



Sin embargo, esta matriz se puede simplificar para enfocarnos en el inventario de personalidad NEO-FFI-R y su correlación con cada una de las drogas, para decidir cuáles parámetros vamos a escoger para los modelos de predicción y además seleccionar las drogas que presentan más correlación con estas características de la personalidad.

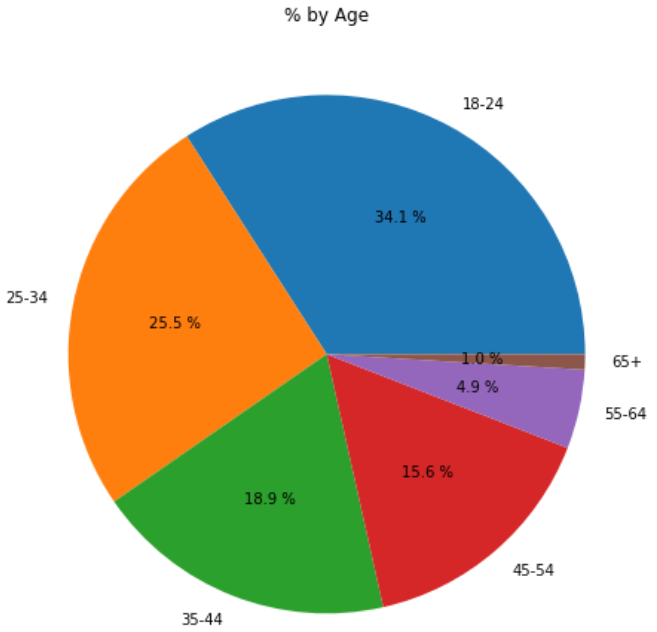


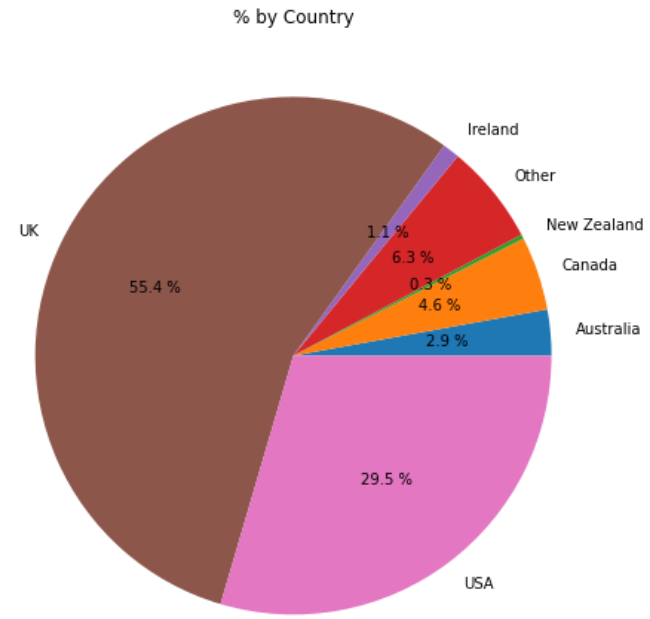
Como se observa, según las correlaciones entre las variables de estudio, los efectos que mayor conllevan a las personas de estos países a consumir son:

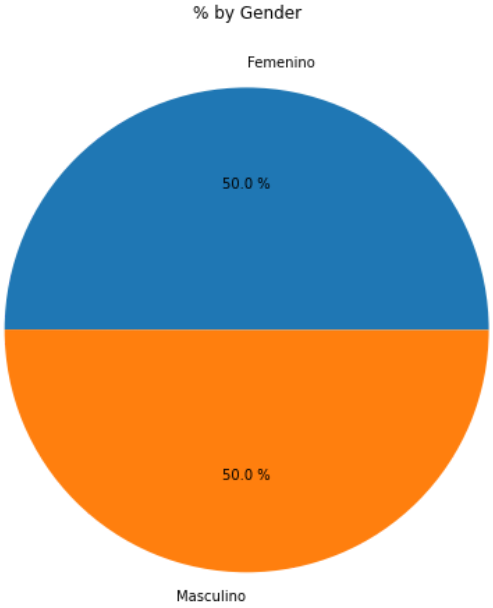
- Búsqueda de sensaciones (*Sensation seeking*)
- Impulsividad (*Impulsiveness*)
- Apertura a la experiencia (*Openness*)

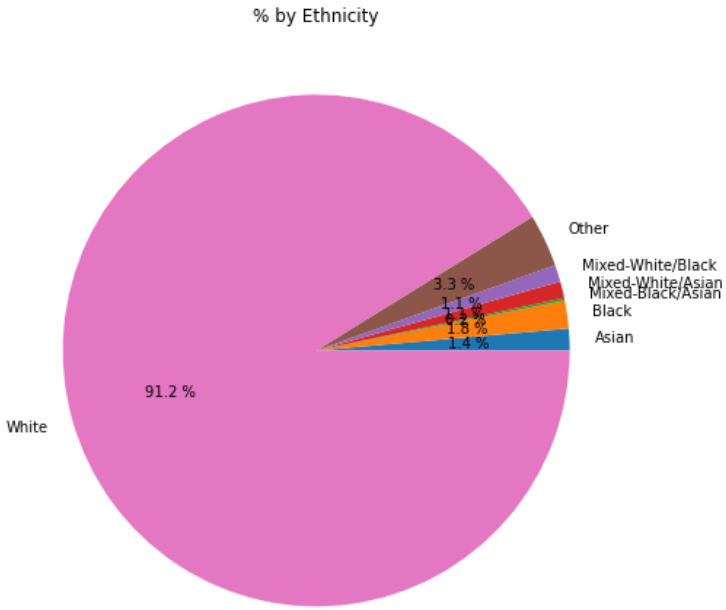
A partir de esto, las sustancias que mayor influencia tienen sobre las personas consumidoras son el Cannabis, el LSD, las *Legal highs* o drogas legales, la Cocaína y el éxtasis. Por lo tanto, teniendo en cuenta estos resultados, se usarán estos como parámetros para la creación de los modelos de *machine learning* de predicción.

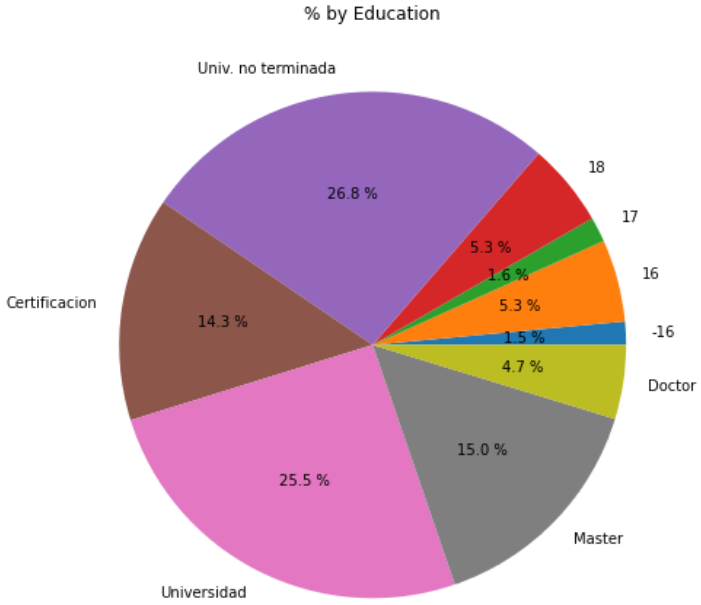
Visualizamos más gráficos para mostrar la información del dataset y contextualizar el problema:







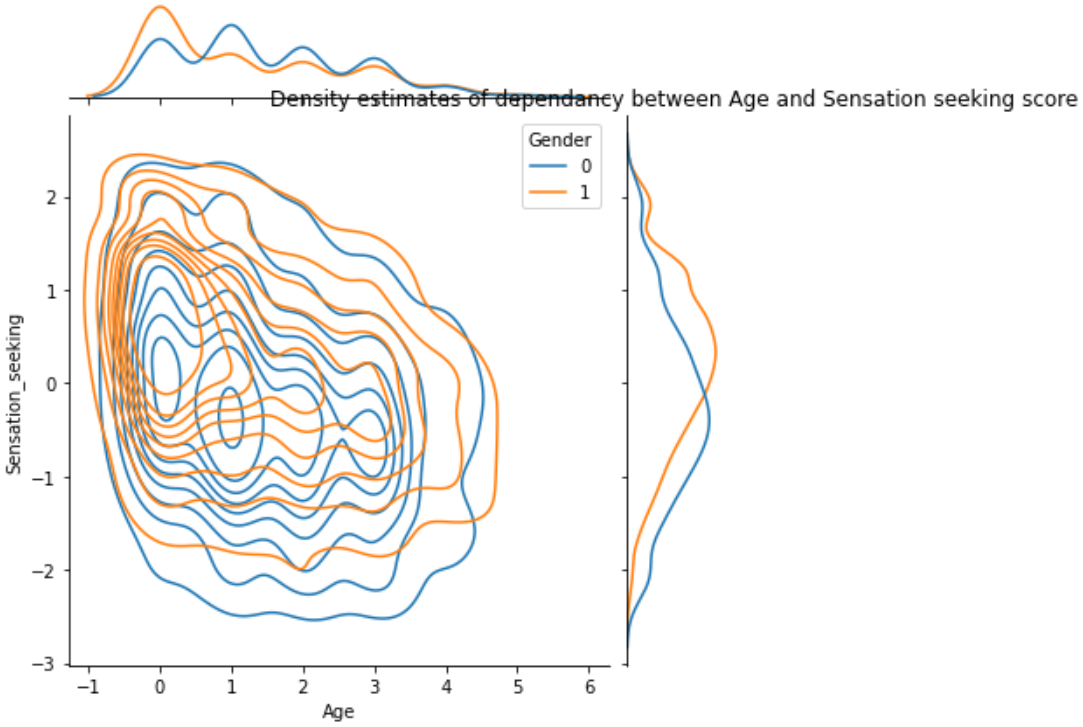




Como se observa en los datos, se pueden encontrar las siguientes características:

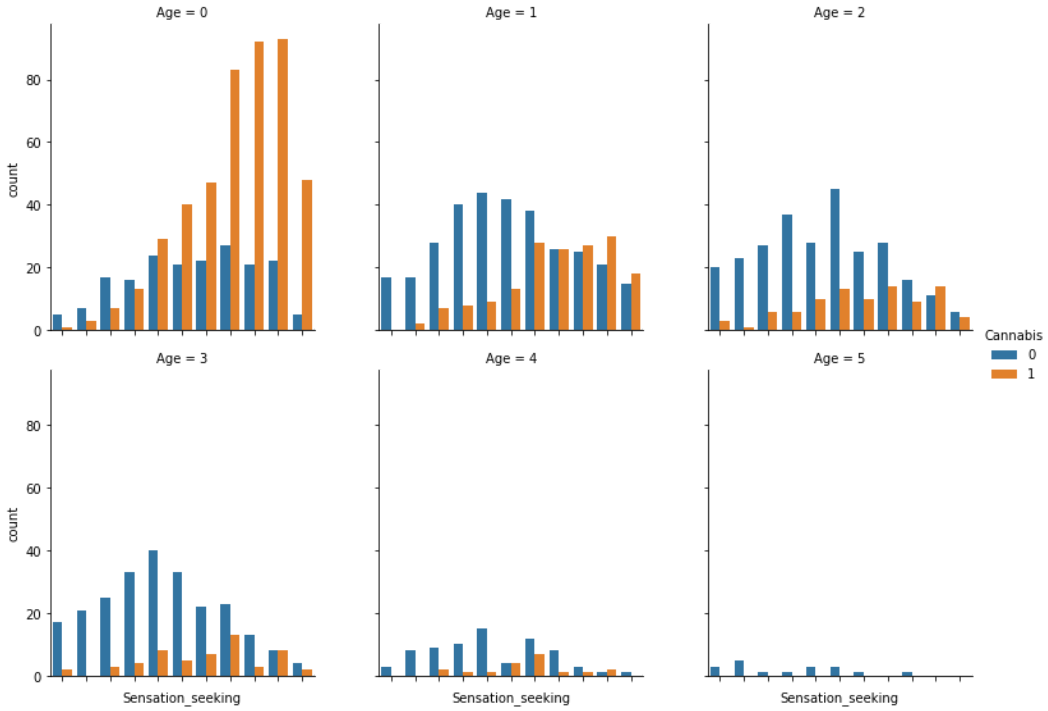
- Este grupo muestral está distribuido uniformemente entre hombres y mujeres.
- Más del 50% de la población es menor de 34 años y se considera joven.
- Menos del 10% de los grupos étnicos de esta población son de otra etnia que no sea blanca
- El 55% de los encuestados viven en el Reino Unido (Inglaterra, Escocia, Gales e Irlanda del Norte).

En cuanto a la distribución de población de acuerdo a la característica de personalidad más determinante encontrada, la cual fue búsqueda de sensaciones, se visualiza en una gráfica multivariada para reconocer los grupos sociales de acuerdo a su edad y género:



De acuerdo a la gráfica de densidad, encontramos que la mayoría de gente en búsqueda de drogas que aumenten su búsqueda de sensaciones se encuentra en los hombres jóvenes entre 18-24. Sin embargo, la paridad con las mujeres es ligeramente inferior, por lo que aquí encontramos que la edad y el género es una característica bastante proporcional.

A continuación, se puede visualizar la distribución de personas de acuerdo a la búsqueda de sensaciones de acuerdo a las 3 drogas más consumidas:



Esto reafirma las deducciones anteriores. A su vez observamos que la mayoría de personas consume Cannabis, esto puede deberse a su facilidad de acceso al consumo y a que puede resultar menos adictiva y costosa que otras drogas como la Cocaína o el LSD. Así mismo, al ser una droga de objetivo relajante y no psicótico es más utilizado.

Nota: El proceso se detalla en la sección Visualización de datos del notebook.

IV. RESULTADOS

Con el desarrollo del proyecto no se presentaron mayores inconvenientes. Para cada una de las drogas seleccionadas se aplicaron cuatro modelos de *machine learning*, y estos fueron los resultados de cada uno de los entrenamientos y pruebas:

Cocaína

	Model	Score	Accuracy
3	KNN	0.9193	0.9174
1	Decision Tree	0.9151	0.9237
2	Support Vector Machine	0.9122	0.9258
0	Logistic Regression	0.9115	0.9216

Cannabis

	Model	Score	Accuracy
3	KNN	0.8075	0.7669
2	Support Vector Machine	0.8011	0.7945
0	Logistic Regression	0.7912	0.7945
1	Decision Tree	0.7771	0.7521

LSD

	Model	Score	Accuracy
1	Decision Tree	0.9250	0.9195
3	KNN	0.9108	0.9237
0	Logistic Regression	0.9094	0.9195
2	Support Vector Machine	0.9080	0.9237

Éxtasis

	Model	Score	Accuracy
3	KNN	0.8740	0.8729
1	Decision Tree	0.8627	0.9025
2	Support Vector Machine	0.8627	0.9025
0	Logistic Regression	0.8613	0.9004

Drogas legales

	Model	Score	Accuracy
3	KNN	0.8896	0.8750
1	Decision Tree	0.8719	0.8771
2	Support Vector Machine	0.8705	0.8771
0	Logistic Regression	0.8655	0.8814

Los resultados de los modelos son bastante buenos, por lo que podrían llegar a predecir si una persona es riesgosamente sensible a consumir determinada droga. No obstante, el Cannabis obtuvo un puntaje muy bajo, esto puede deberse a que esta droga en particular dependa de otras variables categóricas extra, que no afectan al resto.

V. CONCLUSIONES

Aunque las diferencias individuales (orden de rango) tienden a ser relativamente estables en la edad adulta, hay cambios madurativos en la personalidad que son comunes a la mayoría de las personas (cambios de nivel medio). La mayoría de los estudios transversales y longitudinales sugieren que el neuroticismo, la extraversión y la apertura tienden a disminuir, mientras que la amabilidad y la conciencia tienden a aumentar durante la edad adulta.[3]

Los datos están demasiado enfocados a los países angloparlantes de la Commonwealth, Irlanda y los Estados Unidos. Y además, están desequilibradamente distribuidos, ya que la mayoría de ellos está ubicada en el Reino Unido, por lo tanto, este estudio no puede generalizarse o extrapolarse a la situación de consumo de drogas mundial.

Sin embargo, es muy probable que las características no demográficas, sean aplicables al resto de poblaciones del mundo. Ya que existen patrones de consumo de acuerdo a la personalidad bastante notables. A su vez harían falta más parámetros, como las condiciones económicas en las que vive cada persona. Así se podría mejorar y tener un modelo más realista.

VI. REFERENCIAS

[1] 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.", 2015. Disponible: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#

[2] Paul T. Costa, Jr. & Robert R. McCrae. Wikipedia. Revised NEO Personality Inventory. Disponible: https://en.wikipedia.org/wiki/Revised_NEO_Personality_Inventory

[3] Paul T. Costa, Jr. & Robert R. McCrae (2006). "Age Changes in Personality and Their Origins: Comment on Roberts, Walton, and Viechtbauer (2006)" (PDF). Psychological Bulletin. University of Toronto.