Segmentación de pacientes con consumo abusivo de sustancias psicoactivas mediante algoritmos de aprendizaje no supervisado para apoyar el diseño de estrategias y campañas públicas preventivas en la ciudad de Bogotá

Integrantes

Juan Diego Solórzano Gómez
David Nieto Cardona
Vanessa Valencia Giraldo
Jorge Eduardo Velásquez Sánchez

Ejecutores

Natalia Betancur Herrera
Frank Yesid Zapata Castaño
Margarita Maria Orozco
Andrés Sanchez

Universidad de Antioquia, Universidad de Caldas
Talento TECH
BOOTCAMP Inteligencia Artificial

Mayo 2025

1. Introducción

El consumo abusivo de sustancias psicoactivas representa un desafío estructural para la salud pública y la formulación de políticas sociales en Bogotá. Según datos del Ministerio de Salud, el uso problemático de drogas ha mostrado un crecimiento sostenido, especialmente entre jóvenes en contextos urbanos marginados de la capital. Este fenómeno no solo afecta la salud individual, sino que también incide negativamente en la productividad laboral, la seguridad ciudadana y la cohesión social.

Si bien existen datos relevantes sobre las características de las personas que consumen estas sustancias —como edad, género, nivel educativo, zona de residencia y patrones de uso—, estos no han sido suficientemente explotados para generar conocimiento útil que permita anticipar riesgos y orientar decisiones estratégicas.

En este contexto, el presente proyecto propone el uso de técnicas de inteligencia artificial, específicamente aprendizaje no supervisado, para identificar perfiles de riesgo asociados al consumo abusivo de sustancias. La identificación de estos perfiles permitirá a las autoridades de salud diseñar estrategias y campañas preventivas e intervenciones públicas más eficaces, basadas en evidencia y ajustadas al contexto real de la ciudad.

2. Justificación

Este proyecto se justifica en la necesidad de aprovechar la información disponible sobre la población catalogada como consumidora abusiva de sustancias psicoactivas mediante técnicas de inteligencia artificial, particularmente el aprendizaje no supervisado, para identificar perfiles de riesgo que no son evidentes a simple vista. La segmentación basada en datos permitirá reconocer patrones emergentes y construir una base objetiva para diseñar intervenciones preventivas más eficaces, focalizadas y sensibles al contexto social y territorial de Bogotá.

Además, al tratarse de una prueba de concepto en el marco del Bootcamp de Inteligencia Artificial de Talento Tech, este proyecto tiene un valor formativo y exploratorio. Sirve como ejemplo del potencial de la ciencia de datos aplicada a problemas sociales complejos, generando insumos que pueden escalarse o replicarse en futuras investigaciones o iniciativas institucionales.

En síntesis, este proyecto contribuye a cerrar la brecha entre la disponibilidad de datos y su uso estratégico para la formulación de estrategias y planes de prevención basados en evidencia.

3.Objetivo general

Identificar perfiles de riesgo asociados al consumo abusivo de sustancias psicoactivas en Bogotá mediante técnicas de aprendizaje no supervisado, con el fin de apoyar el diseño de estrategias y campañas preventivas basadas en evidencia, orientadas a mejorar la salud pública y la cohesión social.

4.Objetivos específicos

- Recolectar y depurar datos relevantes sobre consumo de sustancias psicoactivas en Bogotá, incluyendo variables sociodemográficas, geográficas y conductuales.
- Aplicar técnicas de aprendizaje no supervisado, como análisis de clusters, para identificar grupos o perfiles con patrones similares de consumo problemático.
- Caracterizar los perfiles de riesgo identificados, con el fin de identificar zonas o poblaciones prioritarias para la intervención.

5.Alcance del proyecto

El proyecto se desarrolla como una prueba de concepto en el marco académico del Bootcamp de Inteligencia Artificial de Talento Tech, programa de formación tecnológica impulsado por el Ministerio de las TIC y ejecutado por la Universidad de Caldas y la Universidad de Antioquia. No se contempla implementación en ambientes productivos.

6. Metodología: CRISP-DM

6.1. Comprensión del Proyecto

Problema de salud pública: La detección tardía de patrones de consumo y la estigmatización dificultan la prevención efectiva. El objetivo es identificar patrones sin sesgos, que permitan intervención oportuna.

Objetivos del proyecto:

- Detectar patrones no evidentes de consumo abusivo mediante clustering.

- Segmentar la población en perfiles homogéneos.
- Generar insumos para el diseño de estrategias y campañas preventivas basadas en evidencia.

Criterios de éxito:

- Identificación clara de perfiles diferenciados.
- Capacidad de traducir los clústeres en estrategias concretas de intervención.
- Documentación accesible para audiencias no técnicas.

6.2. Comprensión de los Datos

Fuente: Secretaría Distrital de Salud - Alcaldía Mayor de Bogotá

- https://datosabiertos.bogota.gov.co/dataset/consumo-abusivo-o-problematico-de-sustancias-psicoactivas

Estructura:

- **Variables categóricas:** Year - Género - Localidad - Tipo Seguro - Nivel Educativo - Etapa Vida - Inició Tratamiento - UPZ - Estado Civil
- **Variables enteras:** Mes Notificación - Consume Vivienda Consume_Parque Consume _ Institución Educativa - Consume Bares - Consume Vía Pública Consume Casa Amigos - Núm Casos - Núm Casos Rango

Problemas detectados

- **Datos faltantes**: Se detectaron 4 valores nulos en la variable género y 20 en la variable localidad. Adicionalmente, en varias de las variables se identifican categorías con valores tales como "Sin dato", "Sin información", “N.A.”.
- **Errores**: Se detectan los siguientes errores en la base
    - Error de formato en cadenas: Las variables categóricas con valores que incluían tilde presentaban un error de formato, visualizándose de forma incorrecta. **Ejemplo:** Para la localidad de Fontibón, la variable se visualiza como FontibÃ³n
- **Duplicados**: No se observan errores relacionados a registros duplicados. Tener en cuenta que los datos de los pacientes son únicos y se encuentran anonimizados.

6.3. Preparación de los Datos

- Limpieza de datos

- Se retiran los registros con campos vacíos o etiquetados como "N.A.", "Sin dato" o "Sin informacion", etc.

- Se retiran los registros correspondientes al año 2025 por su incompletitud. Solo se encontraban reportados para el primer trimestre del año.

Tras aplicar estos dos criterios, el número de pacientes pasó de 88.042 a 66.812

- Transformaciones realizadas (normalización, codificación, etc.)
  - Aplicamos el principio de Pareto para reducir las categorías con muchos valores de baja frecuencia, incluyendo la localidad, el tipo de seguro, nivel educativo, el inicio del tratamiento y estado civil.
  - Se convierten las variables categóricas en numéricas aplicando el método One-Hot Encoding.

- Selección de variables

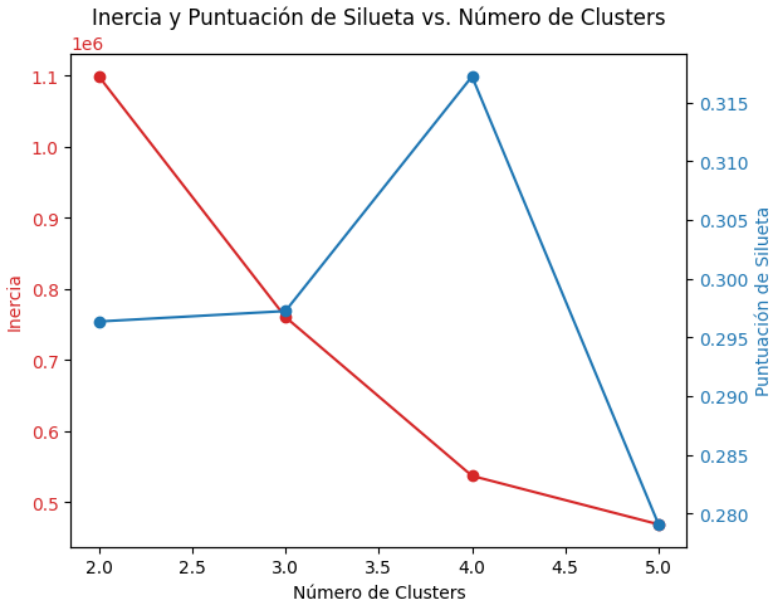
  - Se decide omitir la variable categórica Unidad de Planeamiento Zonal (UPZ), dado que cuenta con 116 valores únicos, con participaciones porcentuales que no difieren significativamente entre ellas.

6.4. Modelado

Algoritmo seleccionado: Se eligió el algoritmo de aprendizaje no supervisado K-means, por su eficiencia y facilidad de interpretación.

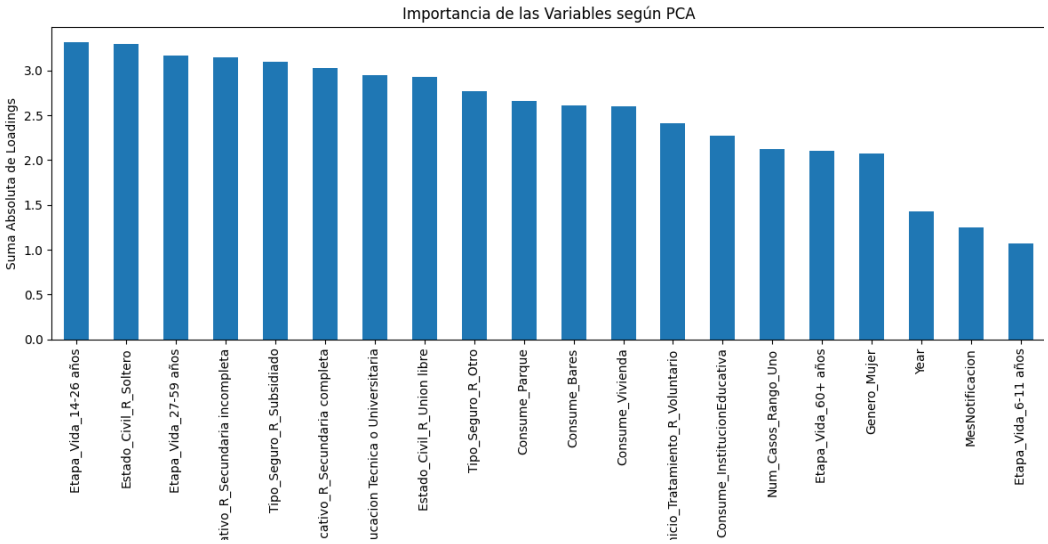
Entrenamiento:

- **Primera iteración del modelo**: En la primera iteración, el uso de K-Means arrojó que el número óptimo de clusters, apoyado en el análisis del codo, era 4 (k=4).

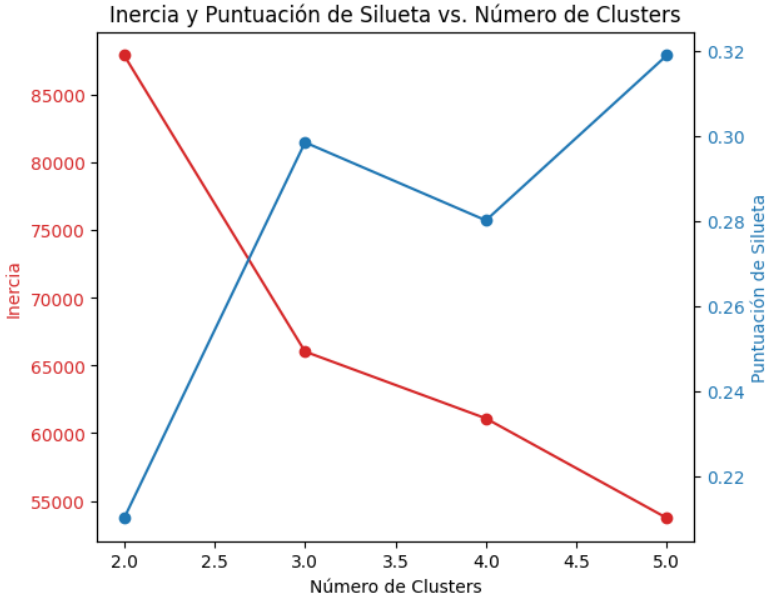


El gráfico de líneas muestra cómo la inercia y el score de silueta se cruzan cuando k es igual a 3, con la silueta alcanzando su punto máximo con 4 clústers.

- **Segunda iteración del modelo**: Para la segunda iteración, se optimizó el modelo inicial haciendo uso del análisis de componentes principales (PCA), el cual calculó la "importancia" de cada característica como la suma del valor absoluto de sus loadings a través de todas las componentes principales, tal como se muestra en el siguiente gráfico.

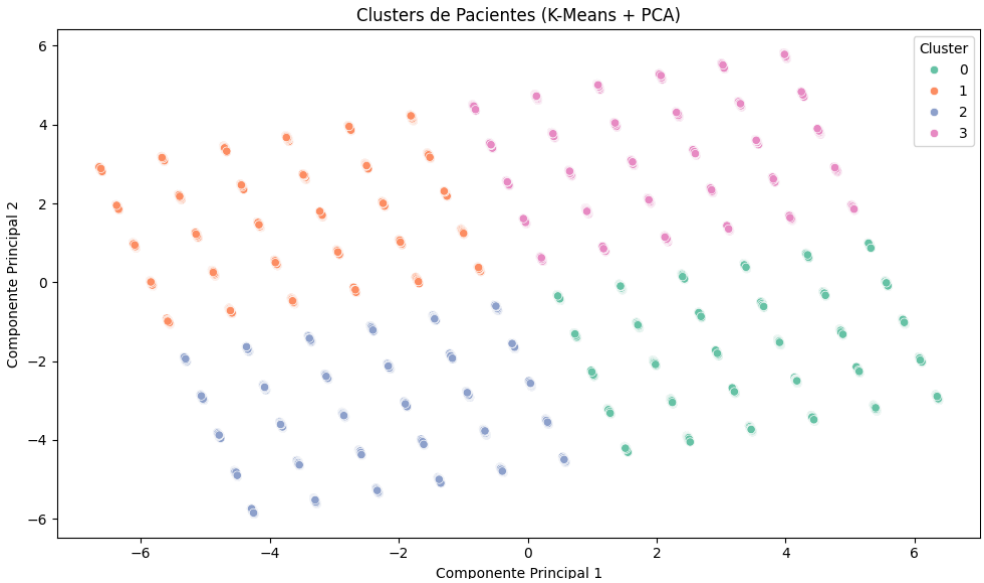


Considerando los resultados obtenidos, se redujeron las variables en el modelo a 8 de las 20 inicialmente incluidas, con lo cual el número de clusters óptimos estimados por el modelo pasaron a ser 3 (k=3).

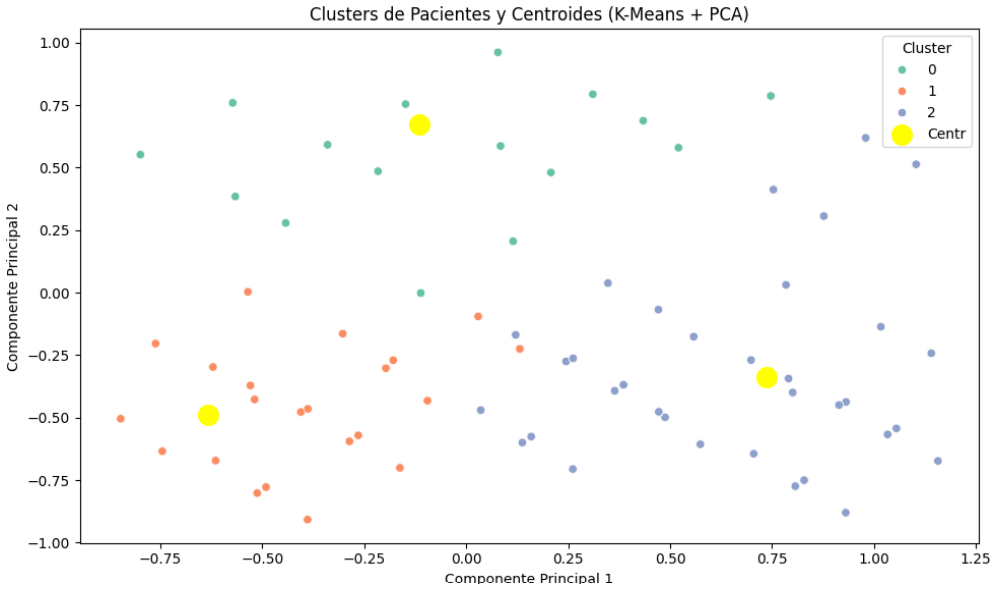


Evaluación:

Para la evaluación, el principal método empleado fue el análisis de la visualización de los clústeres, agrupando las dimensiones en dos componentes principales. La primera iteración del modelo mostró patrones de linealidad entre los clusters, con puntos alejados significativamente del centroide.



En la segunda iteración, la distribución por cluster presenta unas agrupaciones con menos patrones de linealidad, con centroides suficientemente alejados entre sí y una menor cantidad de outliers alejados de dichos centroides.



- **Inercia intra-clúster**: La inercia final promedio fue de 66.025, una reducción del 87% versus la inercia de 529.806 obtenida con el modelo inicial, lo que indica que los 3 clústers finales son mucho más compactos.

  Además, al no ser inercia demasiado baja, el modelo no muestra tendencia al sobreajuste, la cual se presentaría, por ejemplo, si hubiesen muchos clústeres pequeños.

- **Índice de silueta**: La puntuación de silueta para 3 clústeres fue 0.2985, un valor que muestra que el agrupamiento no es óptimo, pero tampoco es completamente arbitrario.

  Con esta estructura de clústeres débil o solapada, en la que los grupos no están claramente separados, queda implícita la necesidad adicional de validar si los clústers tienen forma no esférica.

6.5. Evaluación

Perfiles de los pacientes por clúster

- **Clúster 0 – Adolescentes en contextos vulnerables**

  Descripción del perfil: Este grupo está compuesto principalmente por hombres adolescentes con nivel educativo incompleto, vinculados al régimen subsidiado de salud. Predominan los casos voluntarios y el estado civil soltero.

Patrones destacados:

- Alta frecuencia de consumo en vía pública (67%) y parques (57%).
- Nivel bajo de escolaridad (secundaria incompleta).
- Perfil típico de consumo temprano en contextos de riesgo social.
- Representa un grupo prioritario para intervenciones comunitarias y escolares.

- **Clúster 1 – Jóvenes adultos con consumo social**

  Descripción del perfil: Integrado principalmente por hombres jóvenes, con secundaria completa y afiliados al régimen contributivo. Son casos voluntarios, mayoritariamente solteros.

  Patrones destacados:

  - Consumo elevado en bares (44%) y vía pública (61%), asociado a entornos recreativos.
  - Acceso a educación media, lo que sugiere un perfil funcional con riesgo de consumo social/recreativo.
  - Representa un grupo clave para campañas de concienciación en entornos urbanos y universitarios.

- **Clúster 2 – Adultos en situación de consumo persistente**

  Descripción del perfil: Grupo conformado por hombres adultos (27-59 años) con secundaria completa, atendidos mayoritariamente bajo el régimen subsidiado y en estado soltero.

  Patrones destacados:

  - Consumo frecuente en vía pública (63%), bares (38%) y parques (40%).
  - Nivel educativo medio, pero en contextos posiblemente laborales o familiares complejos.
  - Riesgo asociado a consumo crónico o reincidente, con implicaciones en salud mental y productividad.

Lecciones aprendidas:

- **Importancia del preprocesamiento**: El tratamiento previo de los datos permitió construir el modelo con datos de mejor calidad. Gracias a ello, se redujo el ruido y se encontraron agrupaciones más representativas.
- **Reducción de dimensionalidad:** El uso de técnicas como PCA (Análisis de Componentes Principales) permiten reducir las dimensiones conservando la mayor varianza informativa, lo que facilita el agrupamiento de los datos.
- **El clustering no predice, pero permite generar hipótesis**: La división de los datos en clusters fue el punto de partida para caracterizar a tres grupos poblacionales, sobre los cuales se podrán diseñar estrategias específicas, acordes a su entorno y necesidades particulares.
- **Limitaciones:** Las métricas de rendimiento del modelo permiten inferir que aún hay un margen para su mejora. Esto podría lograrse con la recolección de información de más pacientes o la ampliación de las características de la población ya identificada.

6.6. Implementación

Plan de despliegue (Estos salarios no están incluidos en el presupuesto)

Líder de Proyecto / Project Manager

- Coordinar los equipos, asegurar cumplimiento de plazos, gestionar recursos y servir de enlace con instituciones aliadas (alcaldías, secretarías, universidades).

Científico/a de Datos (Data Scientist)

- Diseñar y entrenar los modelos, interpretar los resultados y traducir los hallazgos a decisiones accionables.

Ingeniero/a de Datos (Data Engineer)

- Preparar, depurar y mantener las bases de datos, asegurar escalabilidad y automatización del flujo de datos.

Epidemiólogo/a o Profesional en Salud Pública

- Validar los perfiles desde una perspectiva clínica/social y orientar la aplicación de los hallazgos a intervenciones comunitarias.

Psicólogo/a o Trabajador/a Social

- Interpretar el contexto de los datos, diseñar campañas preventivas focalizadas y asesorar en ética del manejo de la información.

Enlace Institucional / Gestor de Alianzas

- Coordinar acuerdos con entidades públicas (Secretarías, Ministerio de Salud, observatorios), conseguir acceso a nuevas fuentes de datos y fomentar la adopción del modelo.

Líderes comunitarios o gestores sociales (colaboradores externos)

- Participar en la co-creación de estrategias preventivas y validar la pertinencia de las intervenciones propuestas.

Plan propuesto

1. Piloto local con instituciones educativas y EPS
2. Validación con profesionales de salud mental
3. Desarrollo de dashboard con Streamlit o Power BI

Herramientas utilizadas:

- Python (Colab), Pandas, Scikit-learn, Matplotlib, Seaborn
- PCA para visualización
- K-means para agrupamiento

Mantenimiento:

- **Documentación replicable**: Una documentación detallada permite que otros investigadores, profesionales de la salud o entidades públicas puedan entender, validar y reproducir los pasos del análisis. Esto fortalece la transparencia del proceso, facilita la mejora continua del modelo y garantiza que las decisiones derivadas del análisis se basen en procedimientos verificables.
- **Monitoreo de la calidad de los datos**: Es clave para asegurar que los patrones descubiertos por el modelo reflejen la realidad y no errores o inconsistencias. Un monitoreo riguroso permite detectar y corregir problemas a tiempo, mejorando la fiabilidad de los resultados y la efectividad de las estrategias implementadas con base en ellos.
- **Gobernanza ética**: La gobernanza ética es un componente crítico en el manejo de datos de salud, aún cuando estos hayan sido anonimizados. Es fundamental que el diseño del proyecto contemple políticas claras de uso responsable, consentimiento informado (cuando aplique), y mecanismos para proteger los derechos de las poblaciones vulnerables.

7. Presupuesto

8. Apéndices / Anexos

Anexo A: Diccionario de datos

Concepto	Cantidad	Costo unitario estimado	Subtotal
1. Licencias y herramientas			
Power BI Pro (2 usuarios, 1 mes)	2	45.000	90.000
Servidores en la nube (Colab Pro/VMs)	1	150.000	150.000
Repositorio GitHub privado / hosting	1	$0 (usando versiones gratuitas)	$0
Apache spark	1	$0	$0
Flask/Django	1	$0	$0
React.js	1	$0	$0
Azure	1 mes	$1.512.000	$1.512.000
2. Recursos humanos			
Analista de datos (1 mes)	1	$3.000.000	$3.000.000
Científico de datos junior (1 mes)	1	$2.500.000	$2.500.000
Mentoría técnica (20 h)	1	$100.000/h	$2.000.000
3. Visualización y entrega			

Anexo B: Código fuente (puedes incluir fragmentos o referencias a GitHub)

Diseño de dashboard (Power BI)	1	500.000	500.000
Preparación de informes y documentación	1	500.000	500.000
4. Operativos y logísticos			
Taller validación con expertos (1 sesión)	1	1.000.000	1.000.000
Gastos operativos (conectividad, logística virtual)	1	300.000	300.000
Total estimado			11.552.000

El código fuente así como la base de datos empleada para entrenar el modelo de aprendizaje no supervisado puede encontrarse en el siguiente repositorio de Github:

https://github.com/jdsolorzanog/Bootcamp_TalentoTech

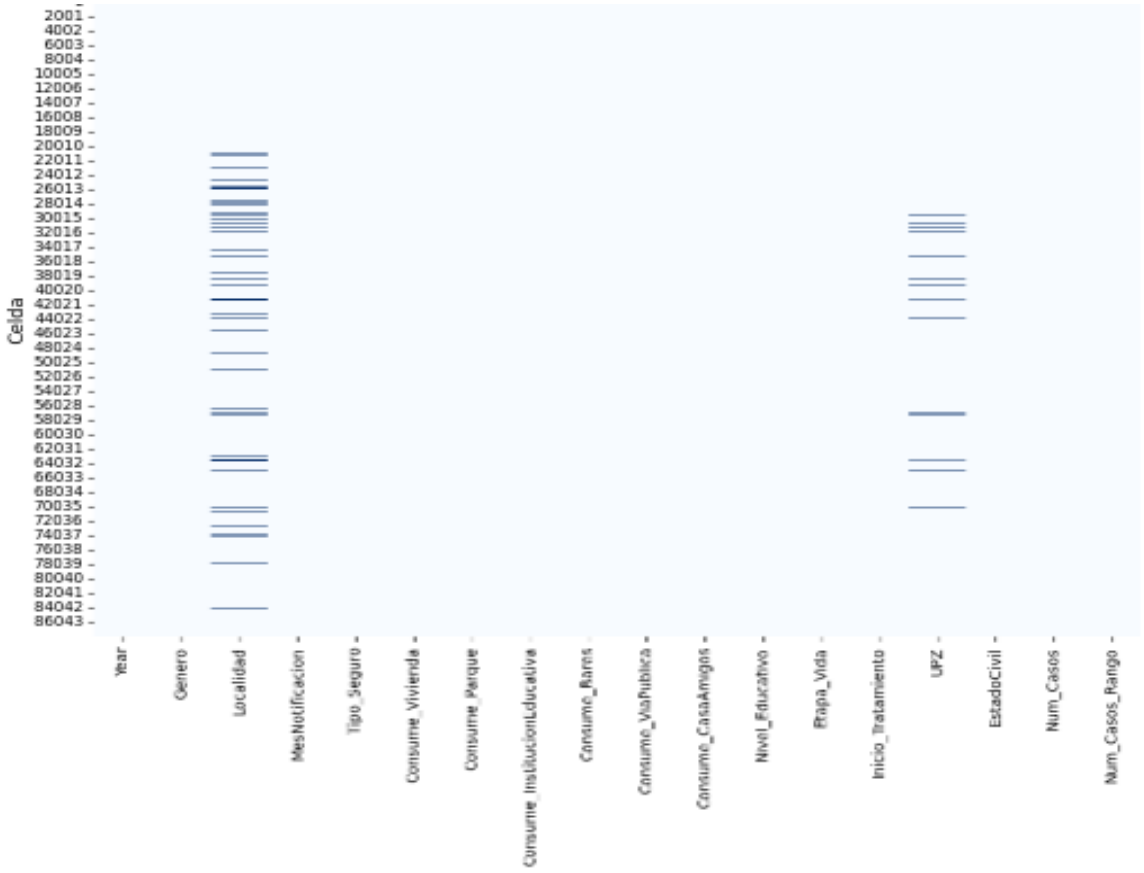
Anexo C: Visualizaciones principales

A continuación se presentan las principales visualizaciones empleadas para el análisis exploratorio de los datos y el análisis del modelo

- **Análisis de valores nulos o vacíos**

Nombre de la columna	Tipo de dato	Descripción	Cantidad de datos	Valores únicos (si aplica)
Year	Entero	Año de notificación del caso.	88042	11
Género	Texto	Género del paciente (Hombre o Mujer).	88038	2
Localidad	Texto	Localidad de residencia del paciente en Bogotá.	88022	22
Mes Notificación	Entero	Mes en el que se notificó el caso (1 a 12).	88042	12
Tipo Seguro	Texto	Tipo de aseguramiento en salud (Contributivo, Subsidiado, etc.).	88042	8
Consume Vivienda	Entero (0/1)	Indica si el consumo ocurre en la vivienda.	88042	2
Consume_Parque	Entero (0/1)	Indica si el consumo ocurre en parques.	88042	2
Consume _ Institución Educativa	Entero (0/1)	Indica si el consumo ocurre en instituciones educativas.	88042	2
Consume Bares	Entero (0/1)	Indica si el consumo ocurre en bares.	88042	2
Consume Vía Pública	Entero (0/1)	Indica si el consumo ocurre en la vía pública.	88042	2
Consume Casa Amigos	Entero (0/1)	Indica si el consumo ocurre en casa de amigos.	88042	2
Nivel Educativo	Texto	Nivel educativo alcanzado por el paciente.	88042	13
Etapa Vida	Texto	Etapa de vida del paciente (Adolescencia, Juventud, Adultez, etc.).	88042	5
Inicio Tratamiento	Texto	Cómo inició el tratamiento (voluntario, llevado por amigos, etc.).	88042	8
UPZ	Texto	Unidad de Planeamiento Zonal de Residencia.	88042	116

Estado Civil	Texto	Estado civil del paciente.	88042	7
Núm Casos	Entero	Número de casos registrados para ese paciente.	88042	35
Núm Casos Rango	Texto	Rango de número de casos. Valores posibles: Uno, dos o más	88042	2



Se realiza limpieza de registros, donde se normalizan y se reemplazan, los datos nulos o etiquetados como "N.A.", "Sin dato", "Sin información", etc. A su vez, se identifica que el año 2025 (último año de muestra), se encuentra con datos incompletos; por lo cual, decide eliminarse.

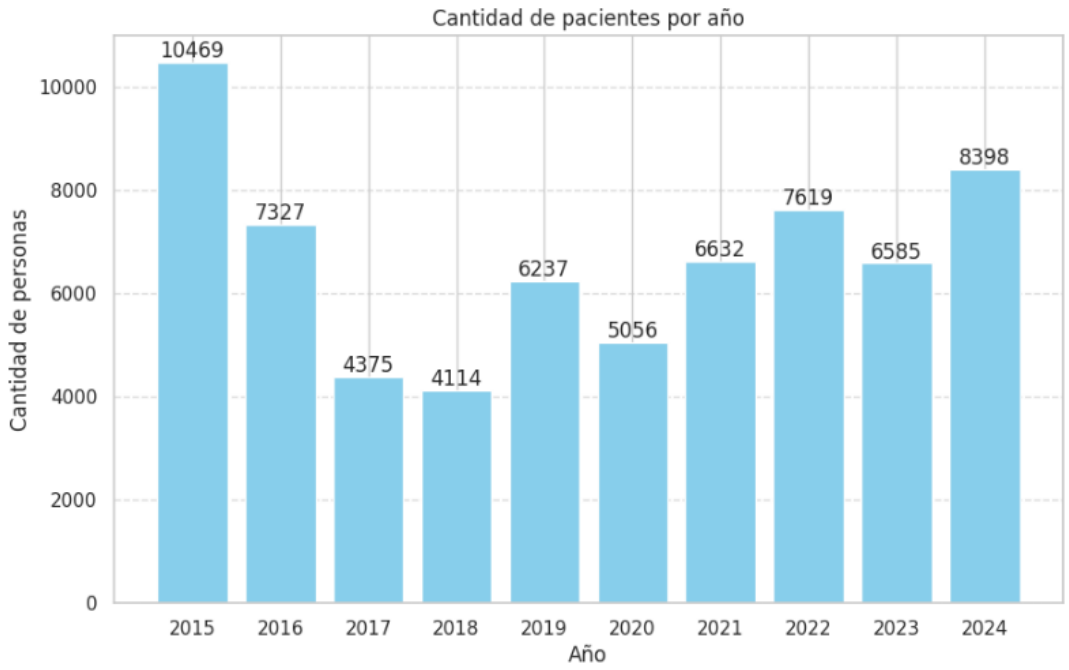
Antes

Year	
2015	11625
2016	8165
2017	4943
2018	4772
2019	9674
2020	6778
2021	9827
2022	9807
2023	8987
2024	10470
2025	2994

Después

Year	
2015	10469
2016	7327
2017	4375
2018	4114
2019	6237
2020	5056
2021	6632
2022	7619
2023	6585
2024	8398

- **Cantidad de pacientes por año**



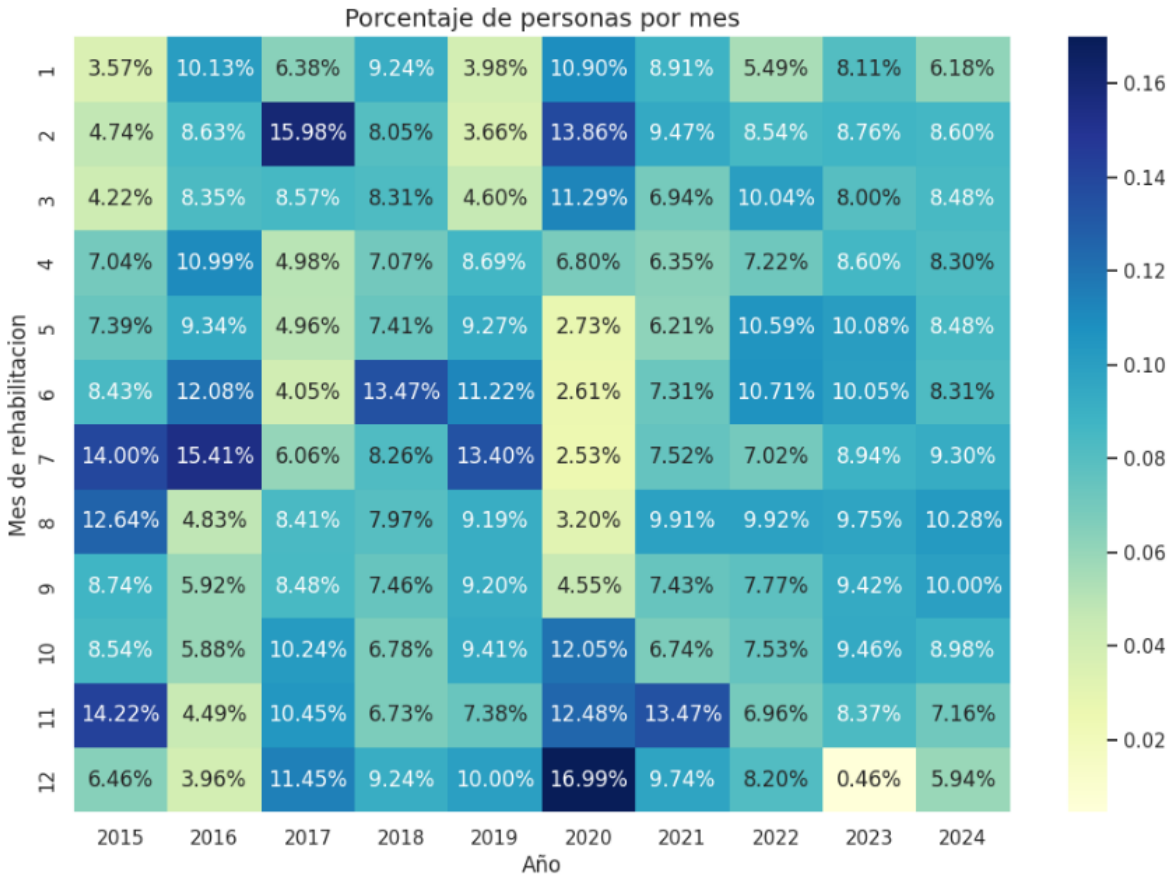
Se identifica que los años ‘2015’ y ‘2024’, tienen una cantidad de pacientes significativamente más alta que los demás años. Esto podría explicarse por los siguientes motivos

- En 2015, se registró un aumento del desempleo significativo. Además, las políticas

de prevención y tratamiento de la ciudad aún se encontraban en desarrollo.

- En el 2024, aumentó el consumo de sustancias psicoactivas entre jóvenes y mujeres. Es posible que más personas hayan accedido a “buscar ayuda”, por lo cual aumentó el número de pacientes registrados.

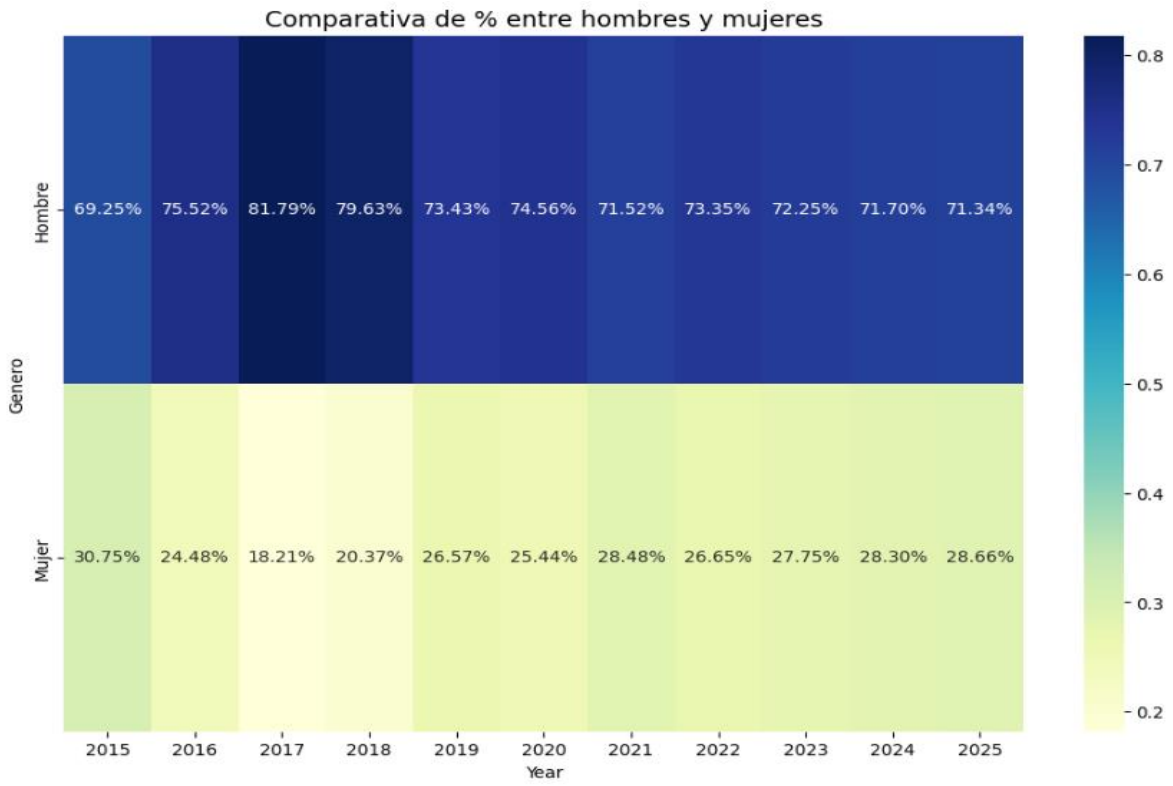
- **Porcentaje de pacientes, por año y por mes**



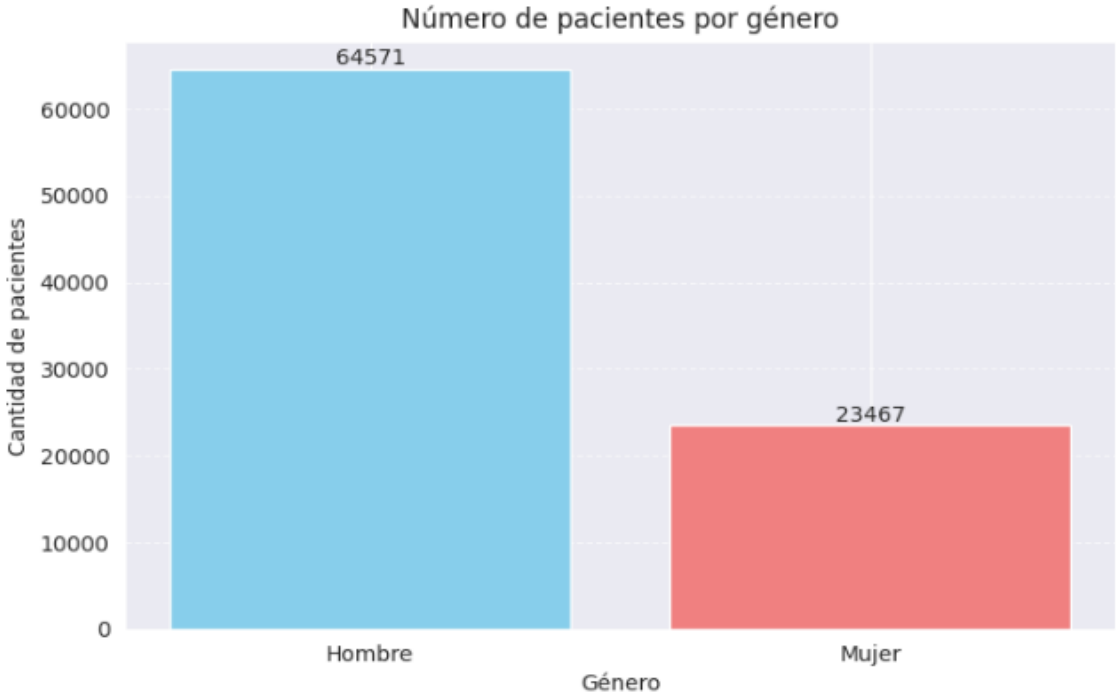
Como se puede evidenciar, los porcentajes varían entre 2% y 16%, donde se observan picos en meses específicos como febrero de 2017 con 15,98%, julio de 2016 con 15,41% y diciembre de 2020 con 16,99%. Colombia enfrentó en 2016 y 2017 crisis económicas importantes, acompañadas de paros en diversos sectores de la economía, lo que pudo haber influido en el aumento de consumo.

Además, los aumentos de consumo en el último trimestre de 2020 pudieron ser el resultado de la crisis sanitaria y el confinamiento en los meses previos. Durante estos últimos meses, con las medidas de seguridad mucho menos estrictas, las personas pudieron retornar al consumo de sustancias, incluso en espacios públicos.

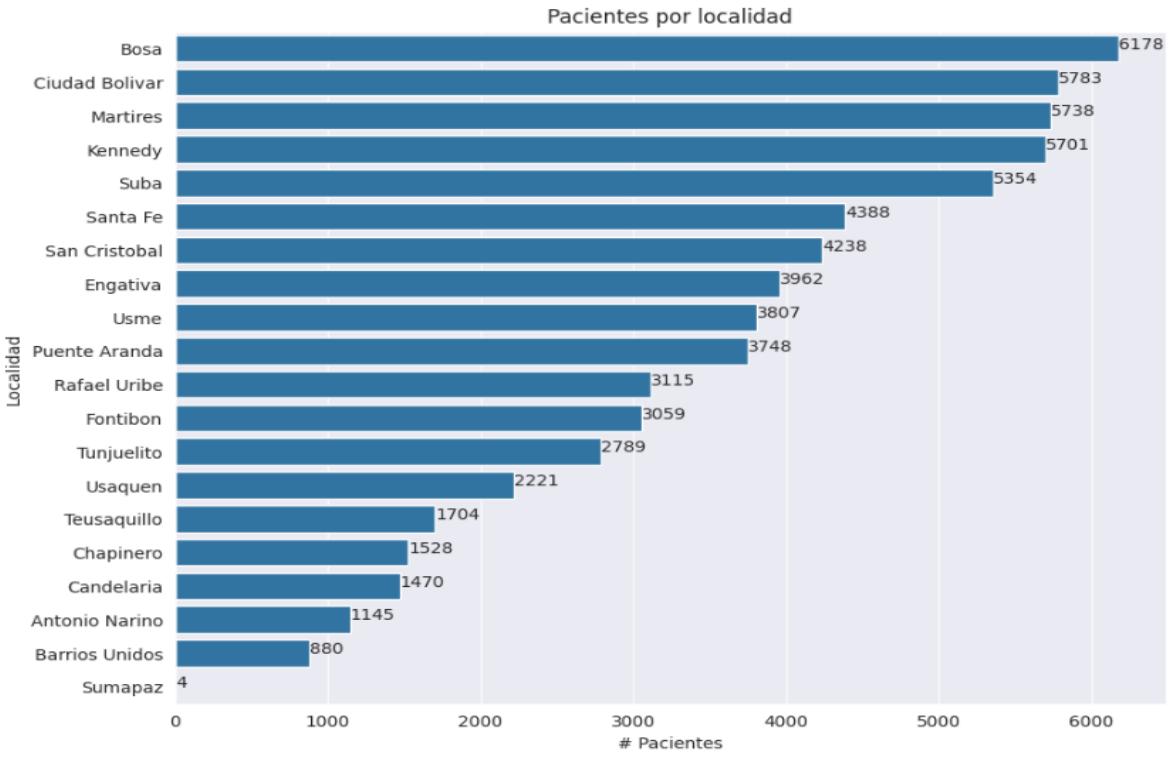
- **Pacientes por género**



La figura permite observar que la proporción de hombres consumidores siempre ha rondado entre el 70 y el 80%, mostrando que el consumo recae principalmente en este género.



- **Pacientes por localidad**



Localidades con mayor número de pacientes

1. **Bosa (6.178 pacientes)**: Es la localidad con más casos registrados. Esto puede deberse a su alta densidad poblacional y factores socioeconómicos.
2. **Ciudad Bolívar (5.783 pacientes)**: Históricamente, esta zona ha tenido altos índices de consumo de sustancias psicoactivas, posiblemente por condiciones de vulnerabilidad.
3. **Los Mártires (5.738 pacientes)**: Es una zona con presencia de habitantes en situación de calle, lo que puede influir en el número de pacientes.

Localidades con menor número de pacientes

1. **Sumapaz (4 pacientes)**: Al ser una zona rural con baja densidad poblacional, el número de casos es significativamente menor.
2. Los barrios **Antonio Nariño (1.145 pacientes) y Barrios Unidos (880 pacientes)**, son localidades con el menor número de pacientes. Esto por diversas causas como lo son cantidad de habitantes; ya que el barrio Antonio Nariño, tiene menos de 120,000 mientras que Suba y Bosa, tienen más de 700,000 cada una. También podría ser, que estos barrios tienen mejores indicadores socioeconómicos por lo que aumenta el índice de calidad de vida, cobertura educativa, acceso a la salud, etc.

Factores que pueden influir en la distribución

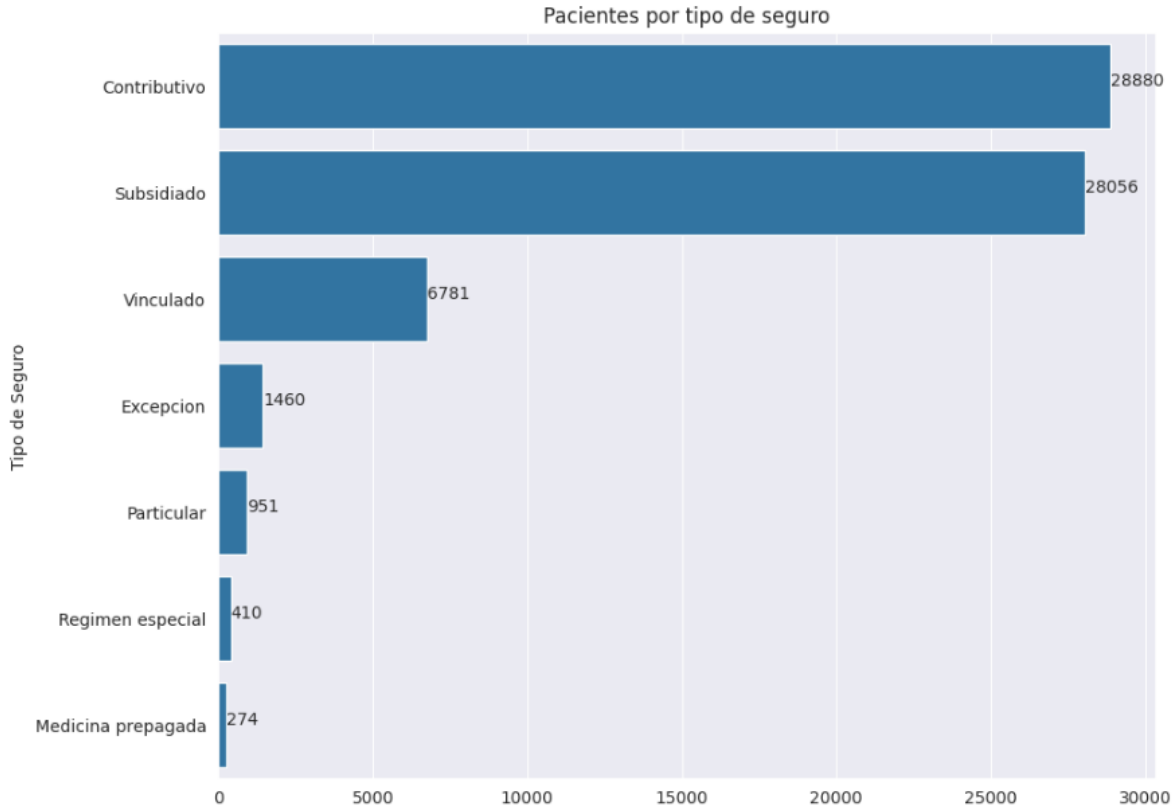
- **Condiciones socioeconómicas**: Las localidades con mayor población suelen tener mayores índices de pobreza y desempleo.
- **Acceso a servicios de salud**: Algunas zonas pueden tener más centros de atención, lo que facilita el registro de pacientes.
- **Presencia de puntos de consumo**: Según estudios recientes, localidades como **Chapinero, Teusaquillo y Usaquén** tienen altos índices de consumo de sustancias ilícitas.

- **Pacientes por tipo de seguro (régimen)**

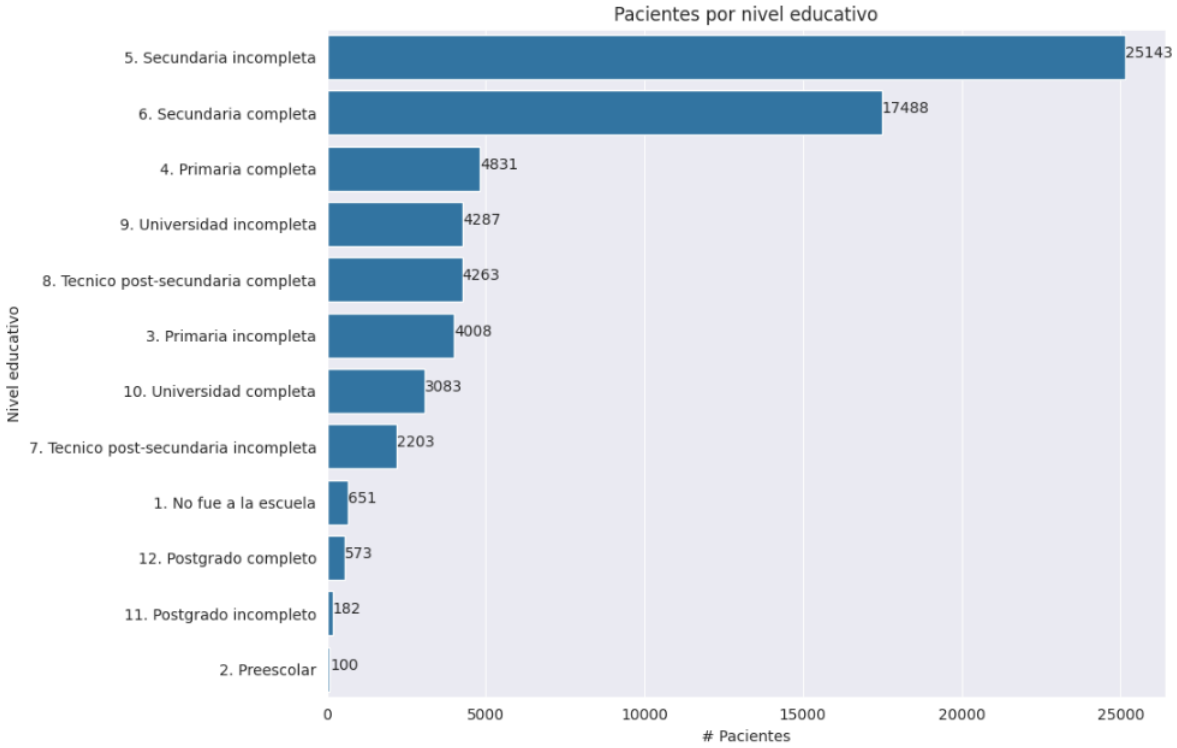
El régimen contributivo y el régimen subsidiado encabezan la lista de el tipo de seguro con el que cuentan los pacientes registrados, lo que da para inferir:

- Las personas de régimen subsidiado, pueden enfrentar mayores riesgos de consumo, debido a factores de pobreza y desempleo.

- Los pacientes que cuentan con medicina prepagada o régimen especial son menos, lo que nos puede indicar que estos sistemas no priorizan tratamientos de adicciones.

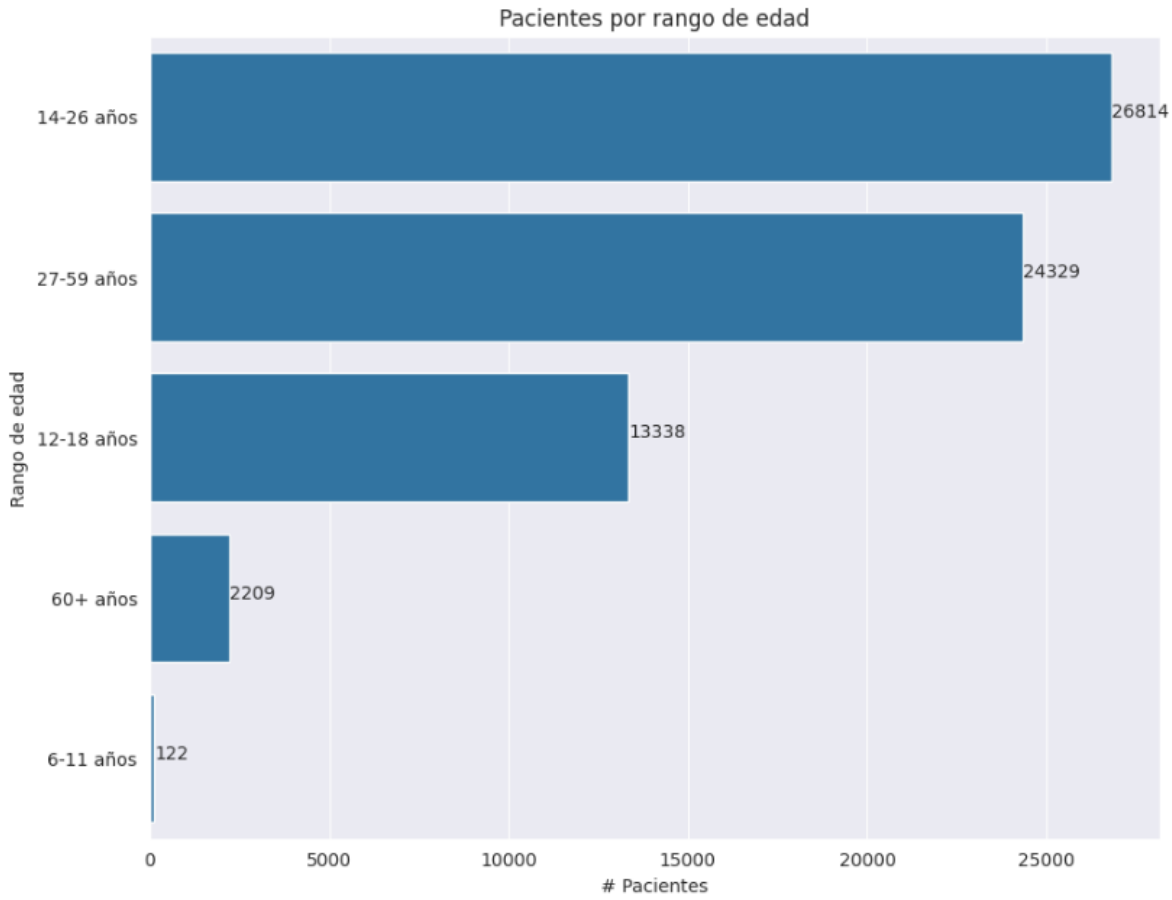


- **Pacientes por nivel educativo**



La categoría más relevante para esta variables es la de “Secundaria Incompleta”; lo que no puede decir que el grado de escolaridad tiene una incidencia relevante en el consumo. No obstante, la segunda categoría es la de “secundaria completa”; con lo cual se infiere que el consumo no solo se limita a individuos que abandonaron la escuela.

- **Distribución de pacientes por edades**



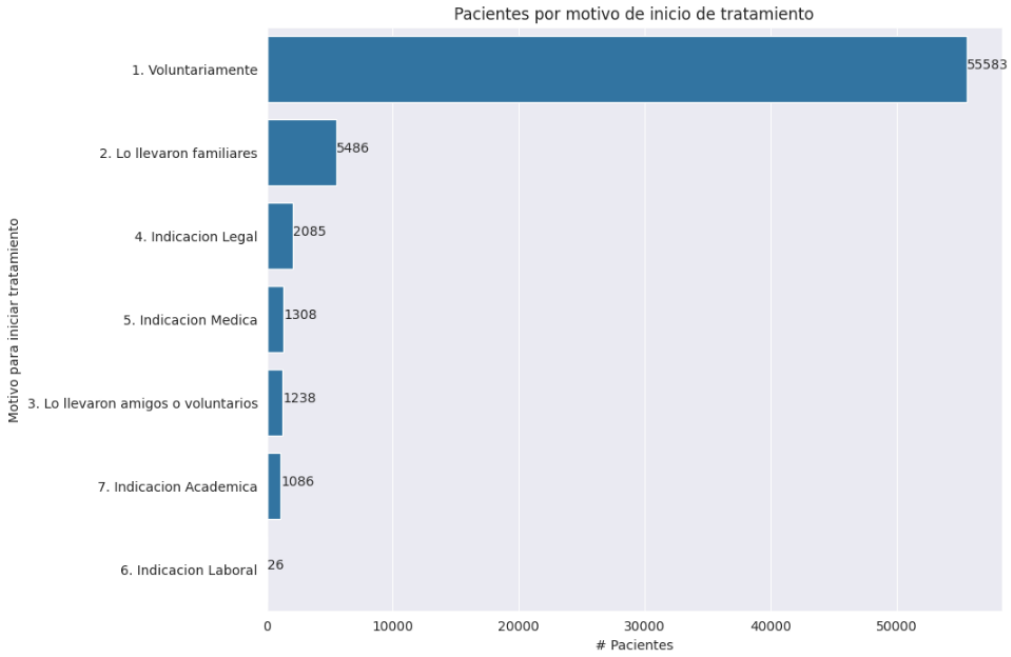
Se evidencia que el grupo más afectado de pacientes pertenece a población joven (de 14 a 26 años), seguidos por personas en edad adulta (27 a 59 años).

Según estudios de la **Secretaría Distrital de Salud**, el consumo en Bogotá ha crecido en poblaciones jóvenes debido a la facilidad de acceso y normalización en entornos sociales.

Los problemas económicos, tienden al aumento de consumo en la población adulta. Después de la pandemia en 2020, se ha observado un aumento de problemáticas sociales ligadas al estrés, la ansiedad y depresión, lo que pudo haber influido en el incremento.

Como un punto a destacar; el consumo de sustancias en personas mayores de 60 años, a pesar de ser baja en comparación, puede deberse a problemas mentales desarrollados por la edad y mal manejo de sus medicamentos.

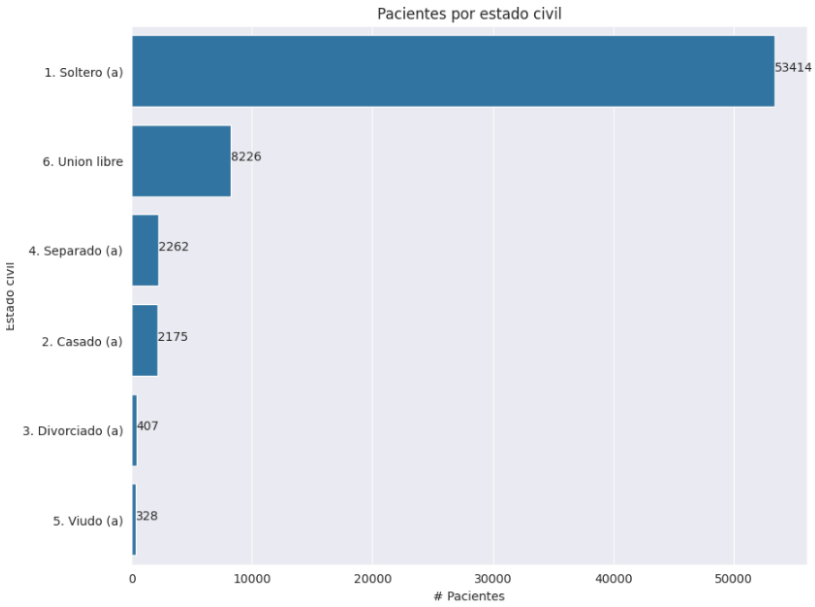
- **Motivo de inicio de tratamiento**



Como se evidencia, los pacientes “voluntarios” abarcan casi la totalidad de registros, lo que puede indicar que se está adquiriendo una conciencia sobre la adicción y la disposición de recibir ayuda externa.

La intervención por familiares y amigos también nos dice que las personas con adicción muchas veces no pueden reconocer la necesidad de tratamiento y requieren intervención de su entorno cercano.

- **Pacientes por estado civil**

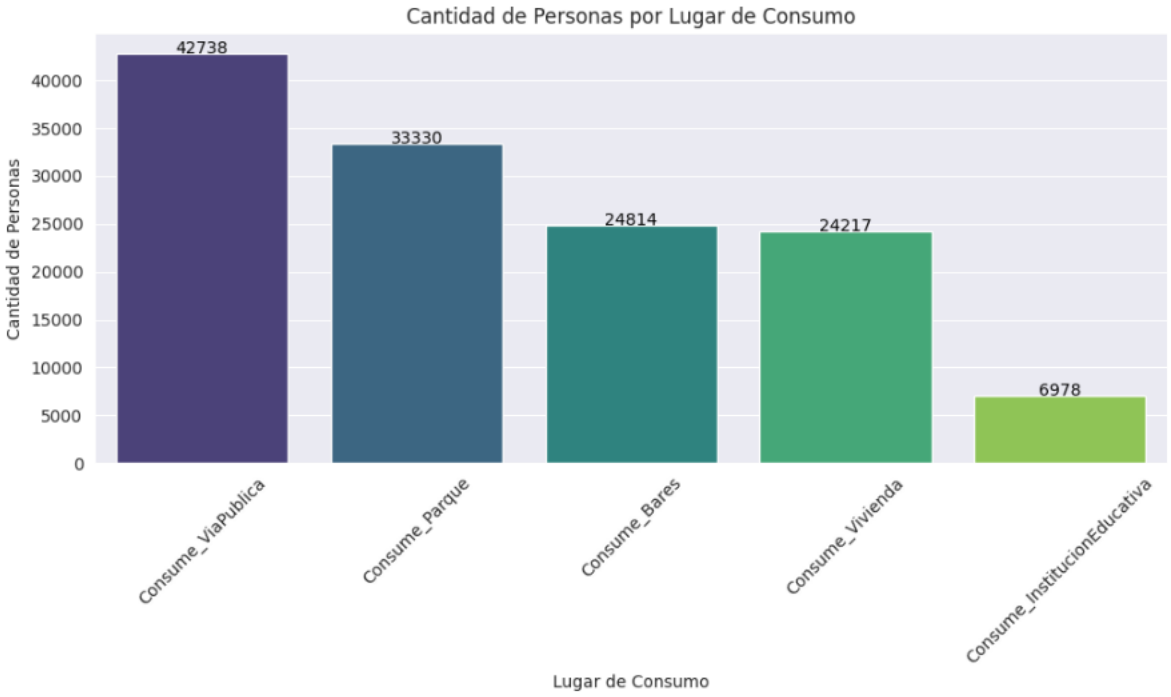


Al igual que en el caso anterior, hay una categoría que abarca casi la totalidad de la variable estado civil, siendo estos los solteros, lo cual podría explicarse por contar con menos redes de apoyo que las personas en unión libre o casados, quienes tendrían menores necesidades de tratamiento.

Estudios han mostrado que el consumo de sustancias suele ser más frecuente en jóvenes y adultos solteros, corroborando el comportamiento exhibido por la base.

- **Lugares más comunes de consumo**

  El mayor número de casos, se encuentran categorizados como “Consume en vía pública” y “Consume en parque”, lo que puede indicar la facilidad de acceso, anonimato y la normalización de consumo en estos espacios. Puede estar también relacionado con el microtráfico, considerando que en estos lugares las sustancias son más accesibles.



Las categorías de “Consumo Bares” (24.814 personas) y “Consumo Vivienda” (24.217 personas)” tienen cifras similares. En bares, el consumo puede estar ligado al consumo recreativo y social, mientras que en viviendas puede reflejar consumo privado o dependencia dentro del hogar. Además, es posible que el consumo en casa se relacione con patrones de consumo a largo plazo y problemas familiares.

La última de las categorías corresponde a “Consumo en instituciones educativas”, lo que podría explicarse por la ejecución de campañas de sensibilización que han reducido el consumo en estos espacios, aunque aún hay casos registrados.

Anexo D: Hiperparámetros: n_clusters, random_state

- Reducción de dimensiones: PCA
- Validación: índice de silueta, inercia.

Anexo E: Plan de implementación técnica o de escalabilidad

El plan de implementación técnica o de escalabilidad estará estructurado de la siguiente manera:

a. **Recopilación y Procesamiento de Datos**
   - **Fuentes de datos:** Se utilizarán bases de datos gubernamentales, informes de salud pública y encuestas poblacionales para obtener registros históricos (2015-2024).
   - **Limpieza de datos:** Se aplicarán técnicas de preprocesamiento en Python (pandas y numpy) para eliminar valores atípicos y asegurar la calidad de la información.
   - **Normalización y almacenamiento:** La información se almacenará en bases de datos como (Colab Pro/VMs), dependiendo de si se manejan datos estructurados o no.

b. **Visualización y Análisis**
   - Mapas de calor y gráficos de tendencias para identificar patrones de consumo en distintos sectores de Bogotá.
   - Modelos estadísticos y de machine learning para predecir incrementos en el consumo en función de factores socioeconómicos y geográficos.
   - Implementación de dashboards interactivos en herramientas como Power BI para facilitar la interpretación de datos.

c. **Desarrollo de Plataforma o Sistema de Monitoreo**
   - Se propone desarrollar una aplicación web o sistema que permita visualizar datos en tiempo real, utilizando Flask/Django para el backend y React.js para el frontend.
   - Integración con APIs gubernamentales para actualizar registros y mantener una base de datos dinámica.

d. Infraestructura Tecnológica

- Uso de servidores en la nube: Implementar almacenamiento escalable mediante Azure para manejar grandes volúmenes de datos.
- Optimización de consultas: Aplicar índices en bases de datos para mejorar la eficiencia en la recuperación de información.
- Procesamiento distribuido: Si el volumen de datos crece, se puede emplear Apache Spark para procesamiento en paralelo.

e. Expansión Geográfica y Social

- Piloto inicial en Bogotá, con posibilidad de replicar el modelo en otras ciudades del país.
- Colaboraciones con instituciones de salud y organizaciones gubernamentales para mejorar el acceso a información y tratamientos.
- Desarrollo de estrategias de prevención basadas en los datos recolectados, enfocadas en zonas críticas.

f. Interoperabilidad y Conectividad

- Integración con plataformas de salud para cruzar información sobre pacientes y tendencias de consumo.
- Automatización de reportes para facilitar la toma de decisiones por parte de entidades gubernamentales y ONGs.
- Uso de Big Data y técnicas de IA para identificar patrones predictivos y generar alertas tempranas.

9. Conclusiones y Recomendaciones

Conclusiones clave:

- El modelo de clustering permite segmentar perfiles de consumo con características diferenciadas.
- Estas agrupaciones pueden ser útiles para diseñar campañas y estrategias preventivas focalizadas y más eficientes.
- La inteligencia artificial aplicada a problemas sociales representa una herramienta estratégica para la intervención temprana.

Recomendaciones:

- Profundizar en la calidad y volumen de los datos, incluyendo variables contextuales externas.
- Vincular instituciones públicas o académicas para validar los hallazgos y explorar aplicaciones reales.
- Repetir el análisis en otras poblaciones para evaluar la generalización de los perfiles.

Consideraciones éticas:

- Todos los datos deben manejarse de forma anonimizada.
- Evitar que los perfiles sean utilizados para estigmatización o discriminación.
- Incluir a la población objetivo en futuras etapas de validación participativa.