Taller de Título

“Caracterización del consumo de alcohol y drogas en Chile”

Alumno: Cristian Riquelme Fernández

1- Comprensión de los datos:

a. Problemática o tema de interés

El consumo de alcohol y drogas ilícitas es una problemática social que afecta indiscriminadamente a la población de Chile desde temprana edad y en diversos estratos, provocando conflictos en ámbitos familiares, laborales o educacionales. La facilidad y disponibilidad de estas sustancias son factores que se buscan combatir institucionalmente para aminorar los daños provocados por estas. Además, son muchas las variables que pueden determinar la susceptibilidad o influencia que transforman este fenómeno en adicciones difícil de controlar y estabilizar, por lo tanto, éstas deben ser analizadas y estudiadas constantemente según el avance y cambio de la sociedad.

La pregunta de valor que guiará al presente trabajo, busca dar forma a la problemática social de las adicciones y sus condicionantes a través de los resultados obtenidos de la aplicación de un instrumento oficial (cuestionario de SENDA) sobre la población. Con esto, se elaborará un informe que permita conocer y estar al tanto de las características, condiciones y factores sociales que influyen en las personas y las incentivan al consumo, para que así se genere información clara y detallada que, en alguna situación supuesta, podría servir a los organismos e instituciones correspondientes a orientar políticas públicas o programas de prevención, tratamiento e integración social de manera óptima y acorde a la realidad del país sobre su población objetivo.

.- Objetivo General

- Clasificar las tendencias de consumo de drogas y alcohol en la población.

.- Objetivos Específicos

- Explorar las diferencias estadísticas entre regiones.

- Caracterizar el consumo de drogas ilícitas, alcohol y tabaco en relación a condiciones socioeconómicas.

- Identificar factores que influyen en el consumo de alcohol y drogas.

Se realizará un análisis exploratorio con la base de datos del “Décimo Cuarto Estudio Nacional de Drogas en Población General de Chile”, estudio que se llevó a cabo por el Servicio Nacional para la Prevención y Rehabilitación del Consumo de Drogas y Alcohol (SENDA), el cual se realizó entre los meses de noviembre de 2020 y junio de 2021 y que se encuentra disponible en su página web. Posteriormente, se aplicará un modelo de aprendizaje

supervisado de clasificación para etiquetar una variable categórica dependiente creada, consumidores o no consumidores, que indicará qué características sociales y socioeconómicas influyen en estos hábitos.

Dado el periodo en que se aplicó el cuestionario a la población de muestra, se está considerando la presencia del COVID-19 y las medidas aplicadas por el gobierno y organismos de salud que modificaron y/o influyeron sobre los aspectos habituales que configuran la realidad social del país, y que de cierta manera pudieron afectar sobre el consumo de alcohol y drogas. Esto permitiría realizar comparaciones con estudios anteriores y observar la existencia de diferencias estadísticamente significativas, pero se considerará de acuerdo se avance en el objetivo general elaborado con el estudio actual.

Fuentes: https://radio.uchile.cl/2021/11/27/consumo-de-drogas-en-chile-los-matices-de-una-realidad-silenciosa-y-persistente/

https://www.senda.gob.cl/bases-de-datos/

b. Diccionario de datos

Conjuntos iniciales

Variable	Etiqueta	Tipo de dato	Categoría observada	N
SbjNum	Identificador de encuestados	float64		16662
S01	Sexo encuestados	category	1 = Hombre 2 = Mujer Total=	6878 9784 16662
REGIONNOM	Región de encuestados	object	1 Región de Tarapacá 2 Región de Antofagasta 3 Región de Atacama 4 Región de Coquimbo 5 Región de Valparaíso 6 Región del Libertador Gral. Bernardo O'Higgins 7 Región del Maule 8 Región del Biobío 9 Región de La Araucanía 10 Región de Los Lagos 11 Región de Aysén del Gral. Carlos Ibáñez del Campo 12 Región de	969 689 846 586 1508 966 1133 1264 1237 932 660 742

			Magallanes y de la Antártica Chilena 13 Región Metropolitana de Santiago 14 Región de Los Ríos 15 Región de Arica y Parinacota 16 Región de Ñuble	 2628 835 698 969
S02	Edad encuestados	float64		
ST_1	Estado de salud	float	1 Muy malo 2 Malo 3 Regular 4 Bueno 5 Muy bueno 6 Excelente 88 No sabe 99 No contesta	188 861 4081 8804 1551 1171 4 2
DP_2	¿Cuál es su estado civil actual (legal)?	category	1 Soltero/a 2 Casado/a 3 Divorciado/a 4 Viudo/a 5 Anulado/a 6 Conviviente civil	9305 4612 1412 634 0 699
DP_16	Aproximadamente y considerando un mes normal, ¿a cuánto asciende el ingreso total del hogar al mes?	float64	1 Menos de $100.000 2 Entre $100.001 y 200.000 3 Entre $200.001 y 300.000 4 Entre $300.001 y 400.000 5 Entre $400.001 y 500.000 6 Entre $500.001 y 750.000 7 Entre $750.001 y 1.000.000	64 22 112 2657 323 5722 536

			8 Entre $1.000.001 y 1.500.000 9 Entre $1.500.001 y 2.000.000 10 Más de $2.000.000 88 No sabe 99 No contesta	1407 2136 3307 0 0
DP_12	¿Cuál es el nivel educacional más alto alcanzado o el nivel educacional actual de usted?		1 Nunca asistió 2 Educación Especial (Diferencial) 3 Primaria o preparatoria (Sistema antiguo) 4 Educación Básica 5 Humanidades (Sistema antiguo) 6 Educación Media Científico Humanista 7 Técnica Comercial, Industrial o Normalista (Sistema antiguo) 8 Educación Media Técnico Profesional 9 Técnico de Nivel Superior 10 Profesional 11 Postítulo 12 Magíster 13 Doctorado 88 No sabe	64 22 112 2657 323 5722 536 1407 2136 3307 128 153 64 68
DP_9	sistema previsional de salud	float64	1 FONASA grupo A 2 FONASA grupo B 3 FONASA grupo C 4 FONASA grupo D 5 FONASA no sabe grupo 6 FF. AA. y del Orden 7 ISAPRE 8 Ninguno (particular) 9 Otro sistema 88 No sabe 99 No contesta	

CO_1	La semana pasada, ¿trabajó al menos una hora, sin considerar los quehaceres del hogar? (Respuesta espontánea)	float64	1 Sí 2 No	9550 7112
CO_6	En su trabajo o negocio principal, ¿usted trabaja como?	float64	1 Patrón o empleador/a 2 Trabajador/a por cuenta propia 3 Empleado/a u obrero/a del sector público (gobierno central o municipal) 4 Empleado/a u obrero/a de empresas públicas 5 Empleado/a u obrero/a del sector privado 6 Servicio doméstico puertas adentro 7 Servicio doméstico puertas afuera 8 FF.AA. y del Orden 9 Familiar no remunerado NULOS(quienes no trabajan no respondieron)	549 3482 926 1051 4614 46 270 37 427 5260(31%)
DP_18	Calidad del barrio	float64	1 Tipo 1. Barrio residencial elegante, donde el valor del terreno y el monto de los arriendos es alto. 2 Tipo 2. Barrio residencial, todavía acomodado, de calles amplias, pero sin tanta área verde; casas confortables y bien mantenidas. 3 Tipo 3. Barrios de comercio o calles estrechas, antiguas, sin áreas verdes, de menos	409 5688 8085

			agrado para vivir y de valores medios de arrendamiento. 4 Tipo 4. Barrio obrero o barrio populoso o mal ventilado. El valor de las viviendas está disminuido por la proximidad a talleres, fábricas, estaciones de ferrocarriles, basurales, etc. 5 Tipo 5. Barrio de mejoras y pocilgas desaseadas, sin pavimentación, de mal aspecto y sin condiciones sanitarias.	2398 82
OH_1	¿Ha tomado Ud. alcohol alguna vez en su vida?	float64	1 Sí 2 No 88 No sabe 99 No contesta	12639 4004 16 3
OH_4	¿Cuándo fue la última vez que Ud. consumió alcohol?	float64	1 Durante los últimos 30 días 2 Hace más de un mes, pero menos de un año 3 Hace más de un año 88 No sabe 99 No contesta	6519 2437 3556 99 28
MAR_1	¿Ha probado Ud. marihuana alguna vez en su vida?	float64	1 Sí 2 No 88 No sabe 99 No contesta	5033 11583 38 8
MAR_4	¿Cuándo fue la última vez que Ud. consumió marihuana?	float64	1 Durante los últimos 30 días 2 Hace más de un mes, pero menos de un año 3 Hace más de un año 88 No sabe 99 No contesta	942 464 3566 46 15

COC_1	¿Ha probado Ud. cocaína alguna vez en su vida?	float64	1 Sí 2 No 88 No sabe 99 No contesta	770 15846 34 12
COC_4	¿Cuándo fue la última vez que Ud. consumió cocaína?	float64	1 Durante los últimos 30 días 2 Hace más de un mes, pero menos de un año 3 Hace más de un año 88 No sabe 99 No contesta	35 47 673 12 3

Conjuntos derivados

edad_cat	Edad categorizada por tramos etarios	category	12 - 18 19 - 25 26 - 34 35 - 44 45 - 54 55 - 65	1354 1950 3006 2867 3134 4351
macrozona	Regiones agrupadas según macrozonas del país	object	norte centro centro sur sur austral región Metropolitana	
tram_ingr	Columna con etiquetas de la variable DP_16 que era sólo numérica	object		
sis_sld	Columna con etiquetas de la variable DP_9	object		
area_lab	Columna con etiquetas de la variable CO_6 que era sólo numérica	object		

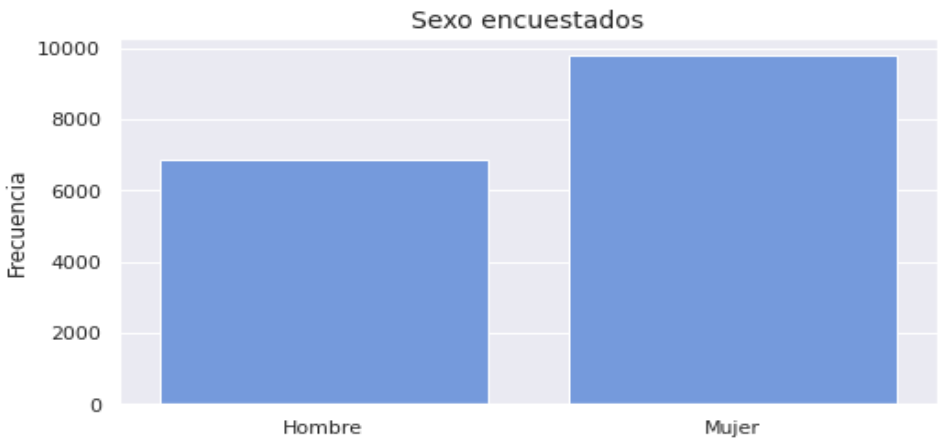
tip_barrio	Columna con etiquetas de la variable DP_18 que era sólo numérica	object		

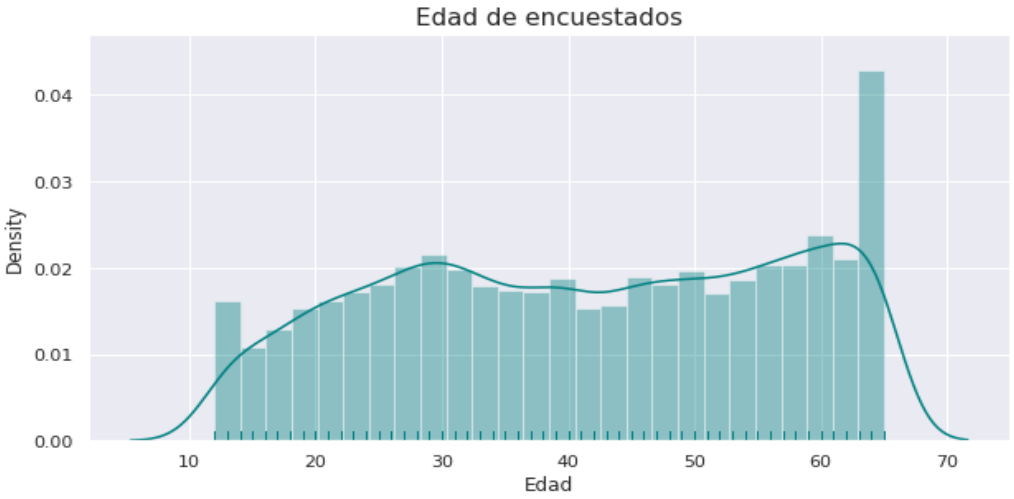
- **Variables dependientes construidas para el modelo de clasificación**

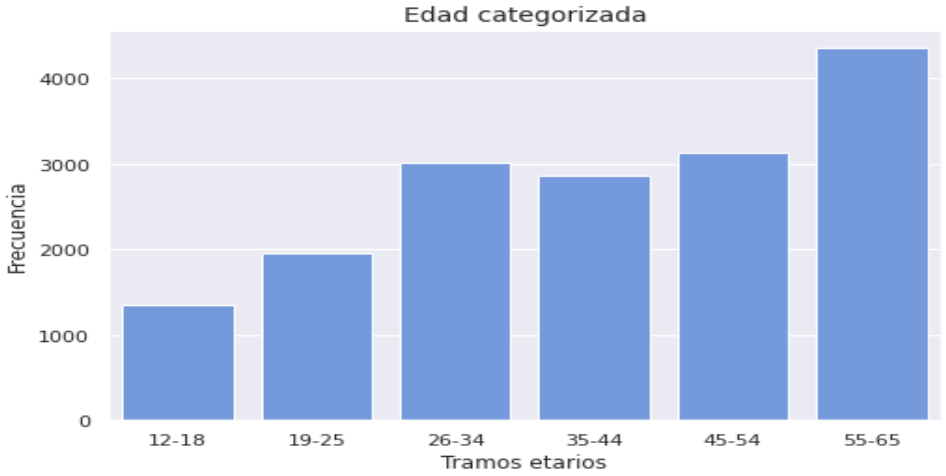
alh_cons	Quienes han consumido alcohol dentro del último mes		No Sí	10143 6519
mar_cons	Quienes han consumido marihuana dentro del mes y hace menos de un año		No Sí	15256 1406
coc_cons	Quienes han consumido cocaína dentro del mes y hace menos de un año		No Sí	16580 82

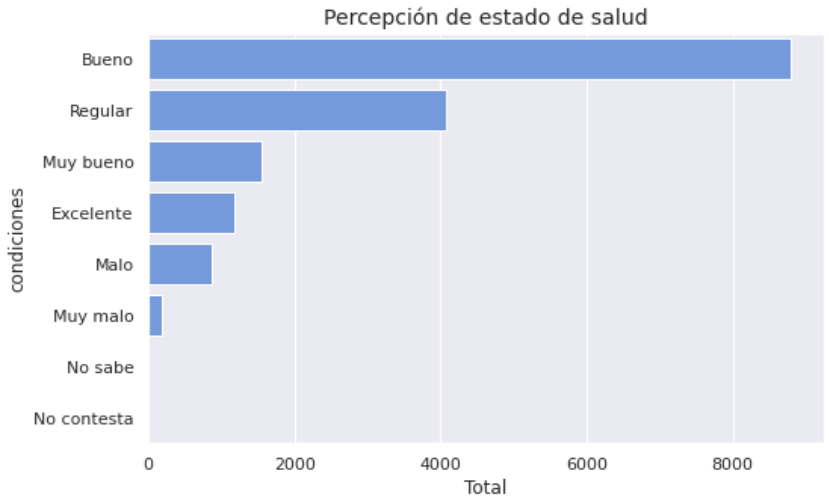
c. Exploración

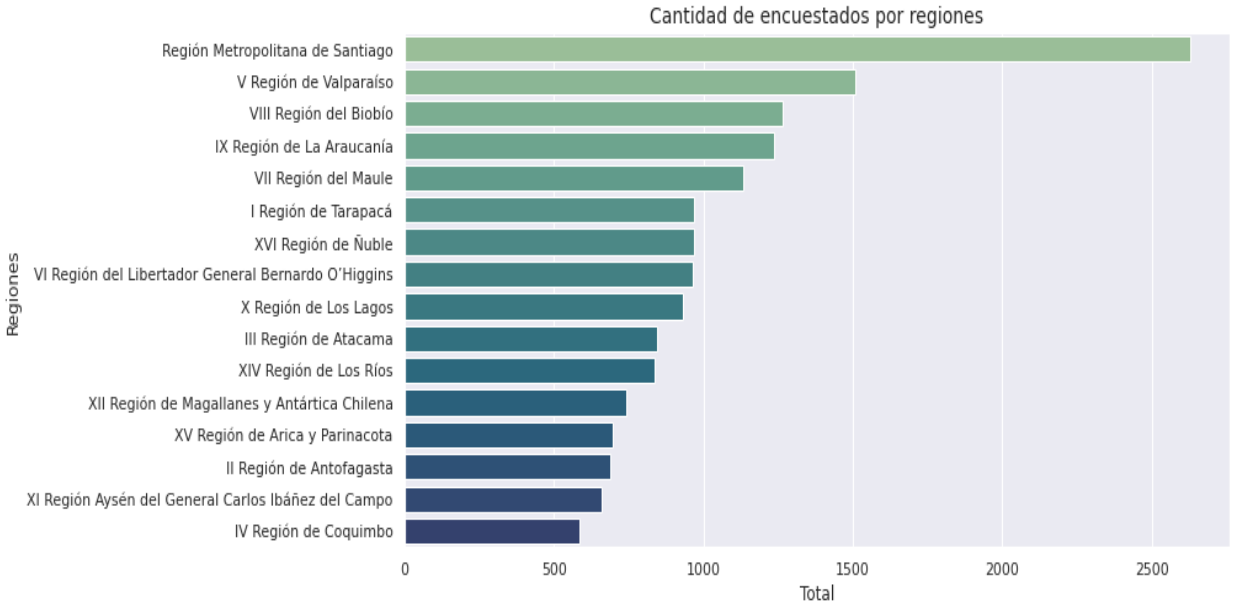
Gráficos descriptivos a nivel univariado y bivariado

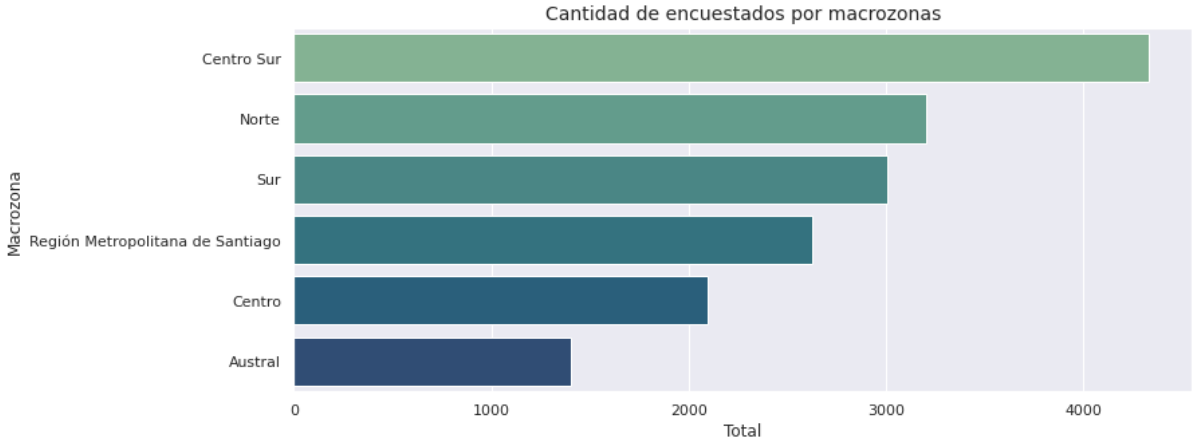


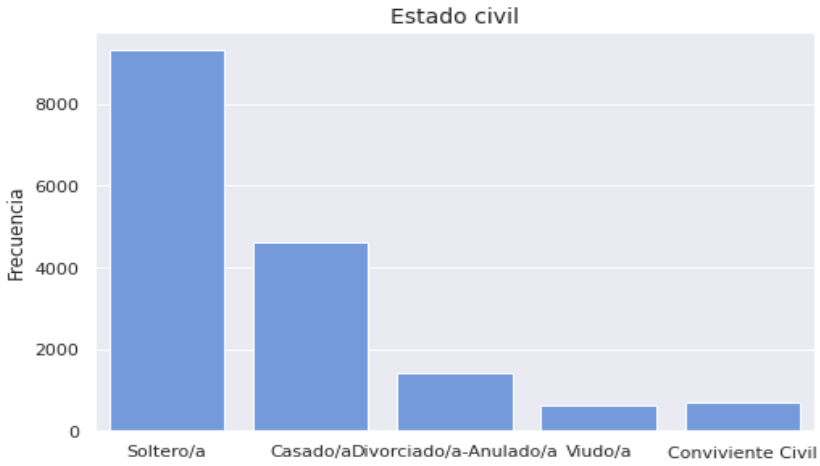

Edad de encuestados
Density
0.00
0.01
0.02
0.03
0.04
10
20
30
40
50
60
70
Edad

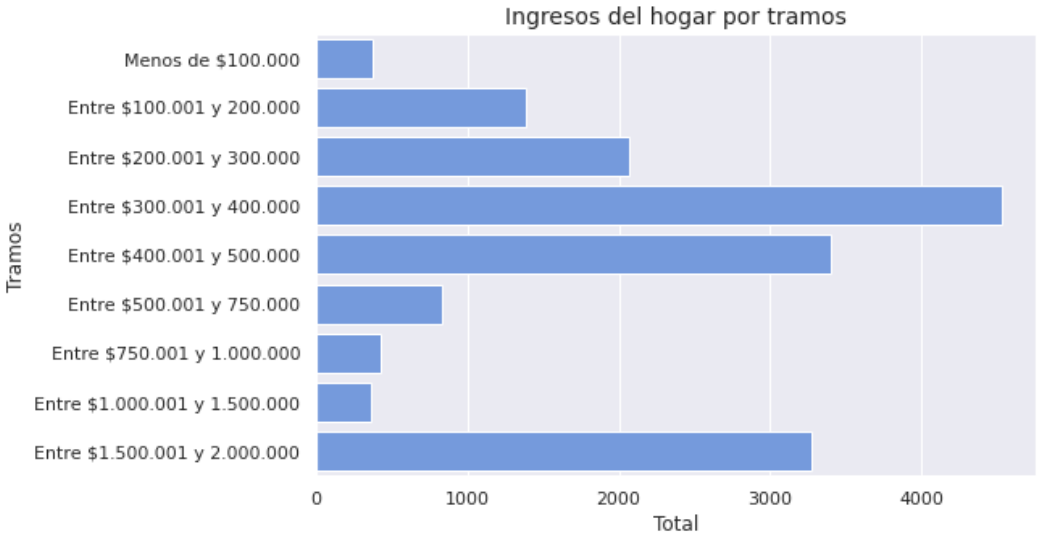

Edad categorizada
Frecuencia
0
1000
2000
3000
4000
12-18
19-25
26-34
35-44
45-54
55-65
Tramos etarios

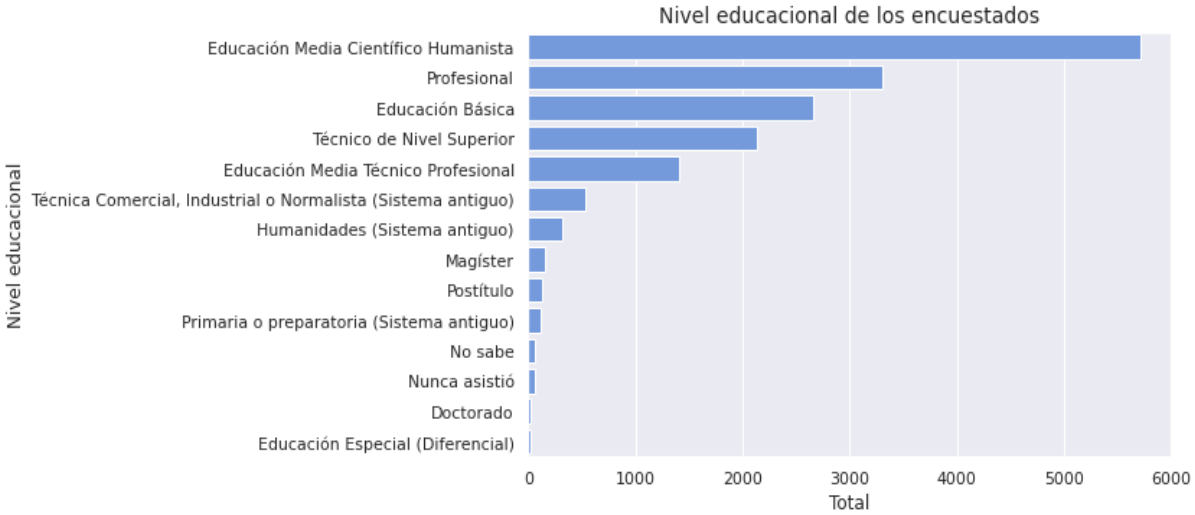

Percepción de estado de salud
condiciones
Bueno
Regular
Muy bueno
Excelente
Malo
Muy malo
No sabe
No contesta
0
2000
4000
6000
8000
Total

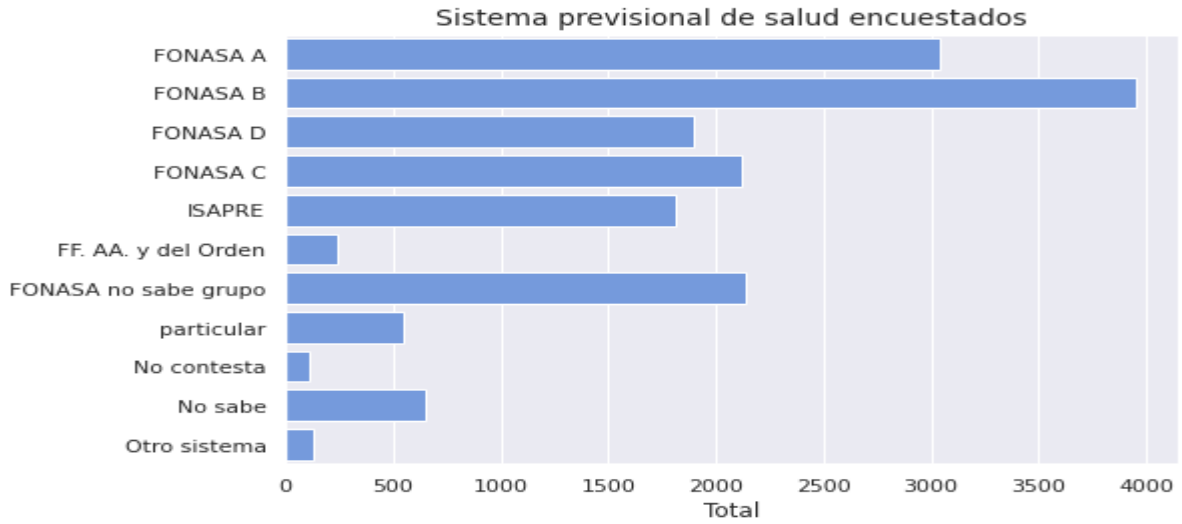

Cantidad de encuestados por regiones
Región Metropolitana de Santiago
V Región de Valparaíso
VIII Región del Biobío
IX Región de La Araucanía
VII Región del Maule
I Región de Tarapacá
XVI Región de Ñuble
VI Región del Libertador General Bernardo O'Higgins
X Región de Los Lagos
III Región de Atacama
XIV Región de Los Ríos
XII Región de Magallanes y Antártica Chilena
XV Región de Arica y Parinacota
II Región de Antofagasta
XI Región Aysén del General Carlos Ibáñez del Campo
IV Región de Coquimbo
Regiones
0
500
1000
1500
2000
2500
Total

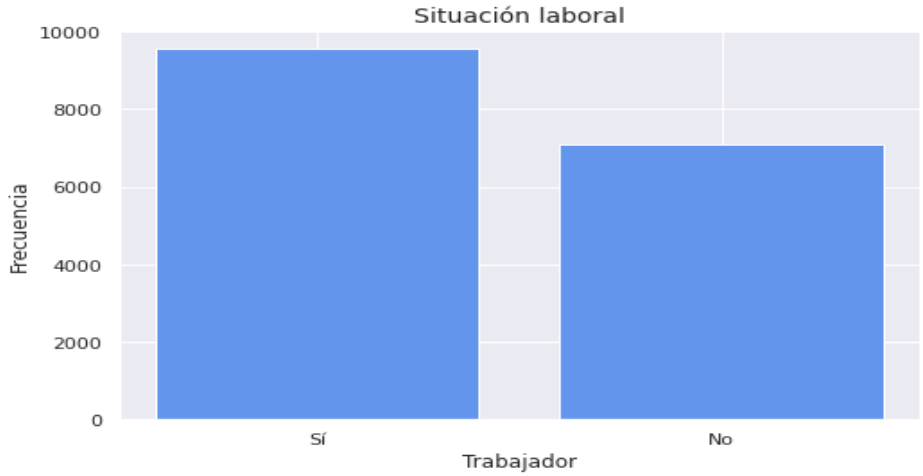

Cantidad de encuestados por macrozonas
Centro Sur
Norte
Sur
Región Metropolitana de Santiago
Centro
Austral
Macrozona
0
1000
2000
3000
4000
Total

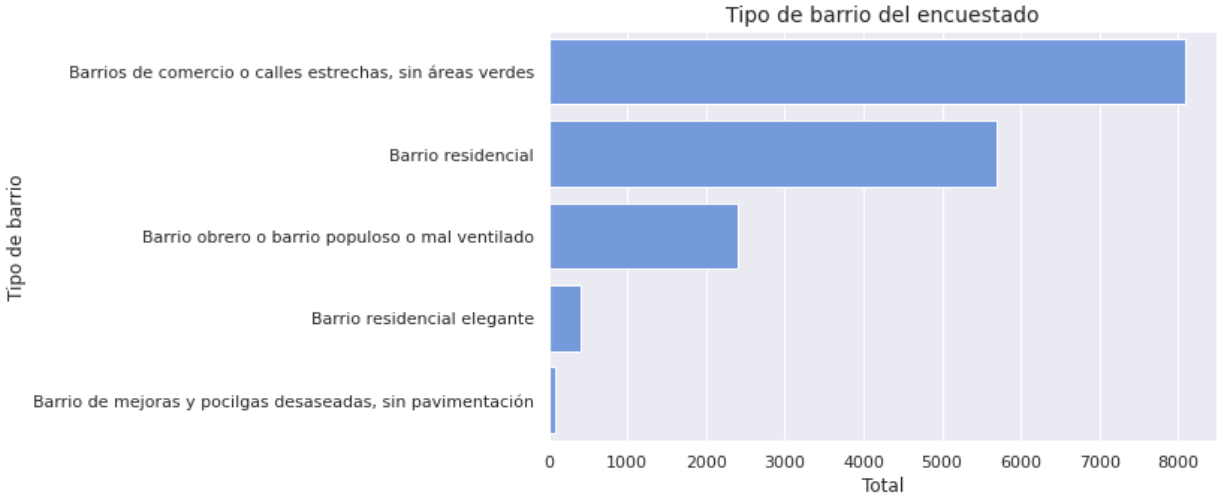

Estado civil
8000
6000
4000
2000
0
Frecuencia
Soltero/a
Casado/a
Divorciado/a
Anulado/a
Viudo/a
Conviviente Civil


Ingresos del hogar por tramos
Tramos
Menos de $100.000
Entre $100.001 y 200.000
Entre $200.001 y 300.000
Entre $300.001 y 400.000
Entre $400.001 y 500.000
Entre $500.001 y 750.000
Entre $750.001 y 1.000.000
Entre $1.000.001 y 1.500.000
Entre $1.500.001 y 2.000.000
0
1000
2000
3000
4000
Total

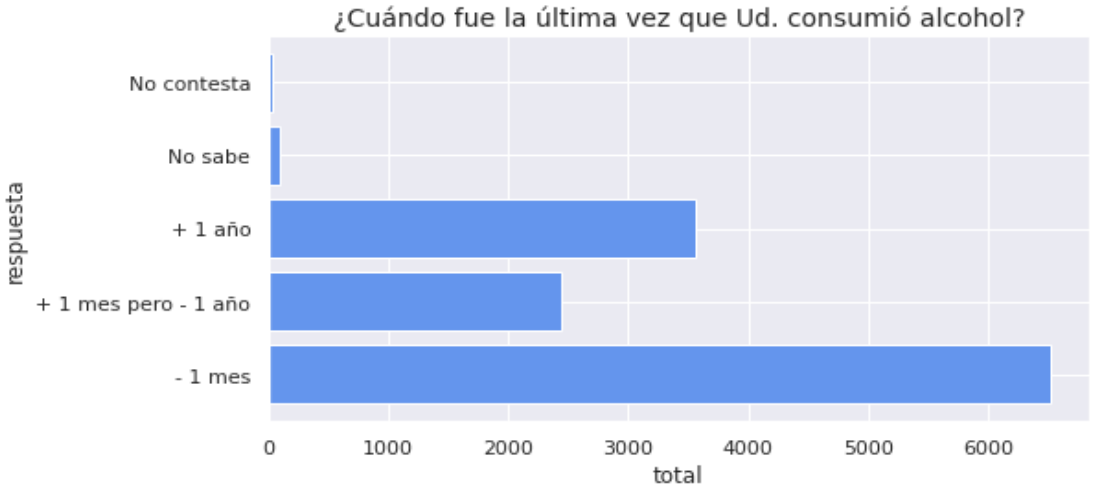

Nivel educacional de los encuestados
Nivel educacional
Educación Media Científico Humanista
Profesional
Educación Básica
Técnico de Nivel Superior
Educación Media Técnico Profesional
Técnica Comercial, Industrial o Normalista (Sistema antiguo)
Humanidades (Sistema antiguo)
Magíster
Postítulo
Primaria o preparatoria (Sistema antiguo)
No sabe
Nunca asistió
Doctorado
Educación Especial (Diferencial)
0
1000
2000
3000
4000
5000
6000
Total

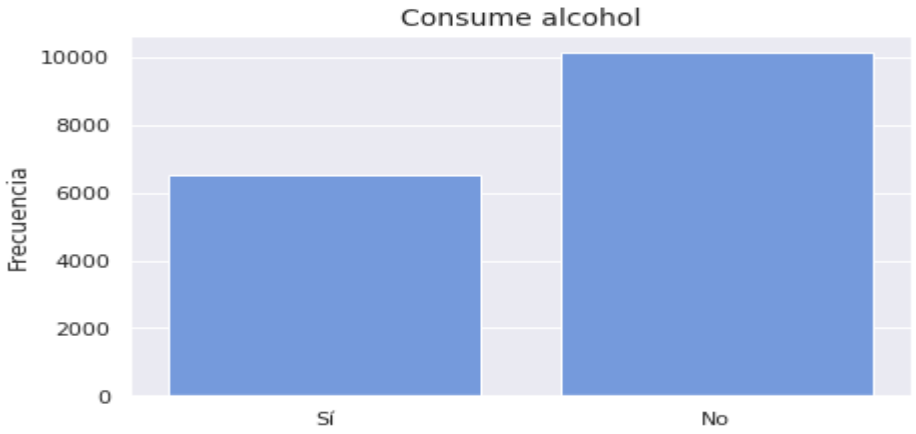

Sistema previsional de salud encuestados
FONASA A
FONASA B
FONASA D
FONASA C
ISAPRE
FF. AA. y del Orden
FONASA no sabe grupo
particular
No contesta
No sabe
Otro sistema
0
500
1000
1500
2000
2500
3000
3500
4000
Total

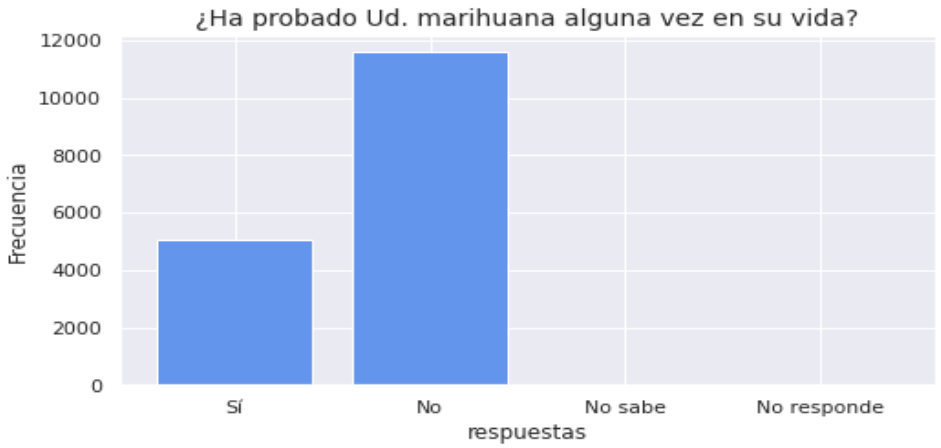


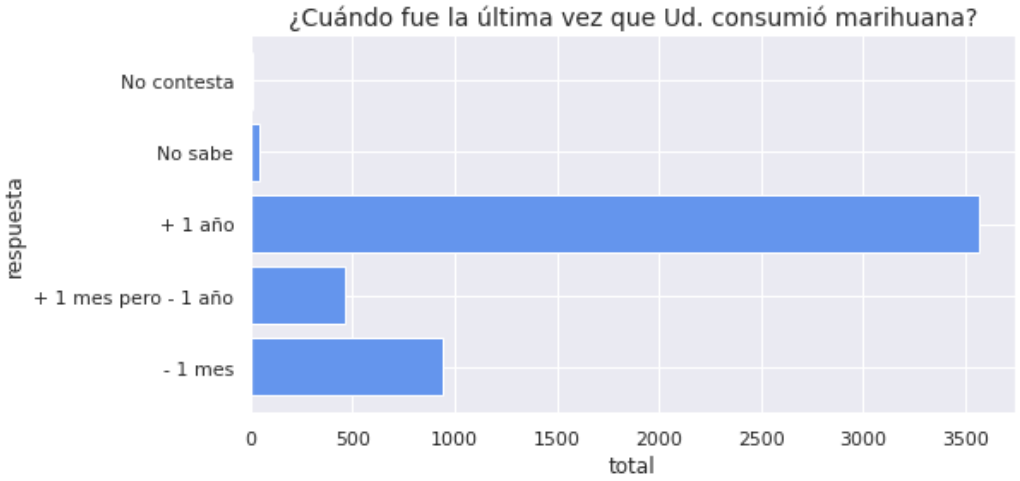


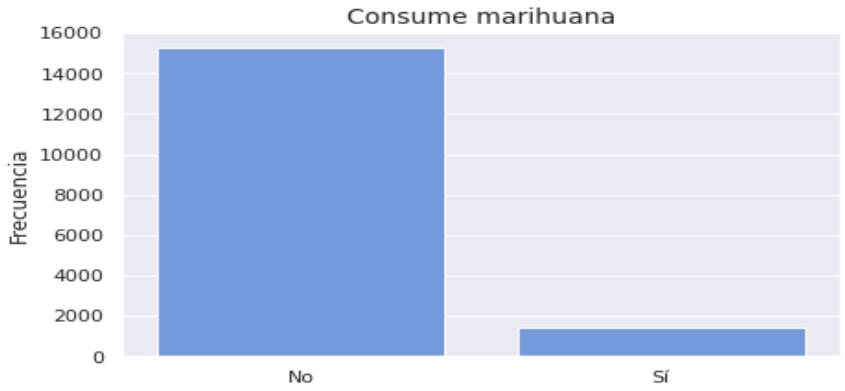
Variables sobre consumo de alcohol y drogas


¿Cuándo fue la última vez que Ud. consumió alcohol?
No contesta
No sabe
+ 1 año
+ 1 mes pero - 1 año
- 1 mes
respuesta
0
1000
2000
3000
4000
5000
6000
total

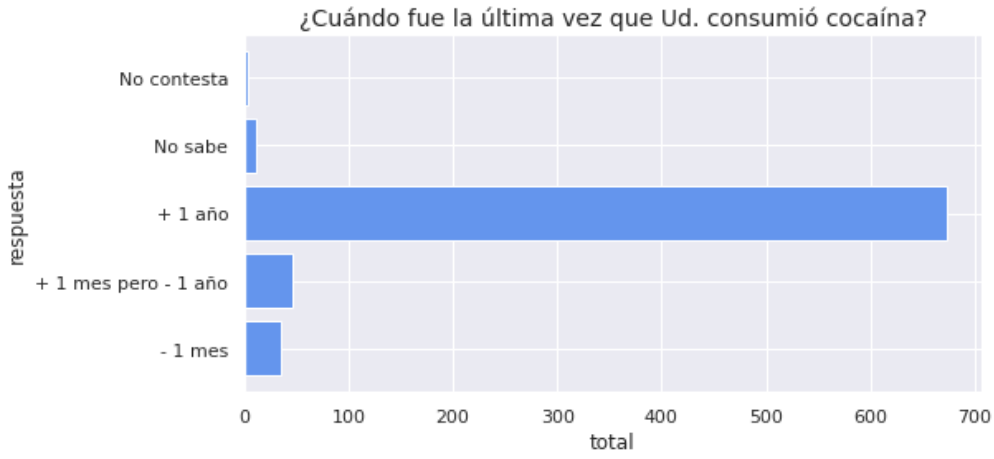

Consume alcohol
10000
8000
6000
4000
2000
0
Frecuencia
Sí
No

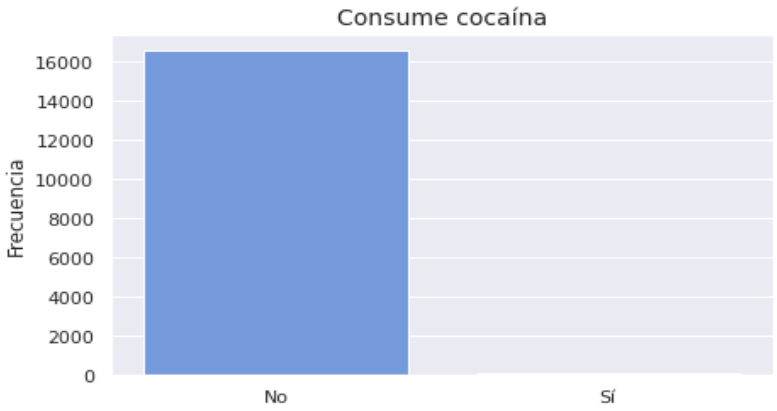

¿Ha probado Ud. marihuana alguna vez en su vida?
12000
10000
8000
6000
4000
2000
0
Frecuencia
Sí
No
No sabe
No responde
respuestas


¿Cuándo fue la última vez que Ud. consumió marihuana?
No contesta
No sabe
+ 1 año
+ 1 mes pero - 1 año
- 1 mes
respuesta
0
500
1000
1500
2000
2500
3000
3500
total

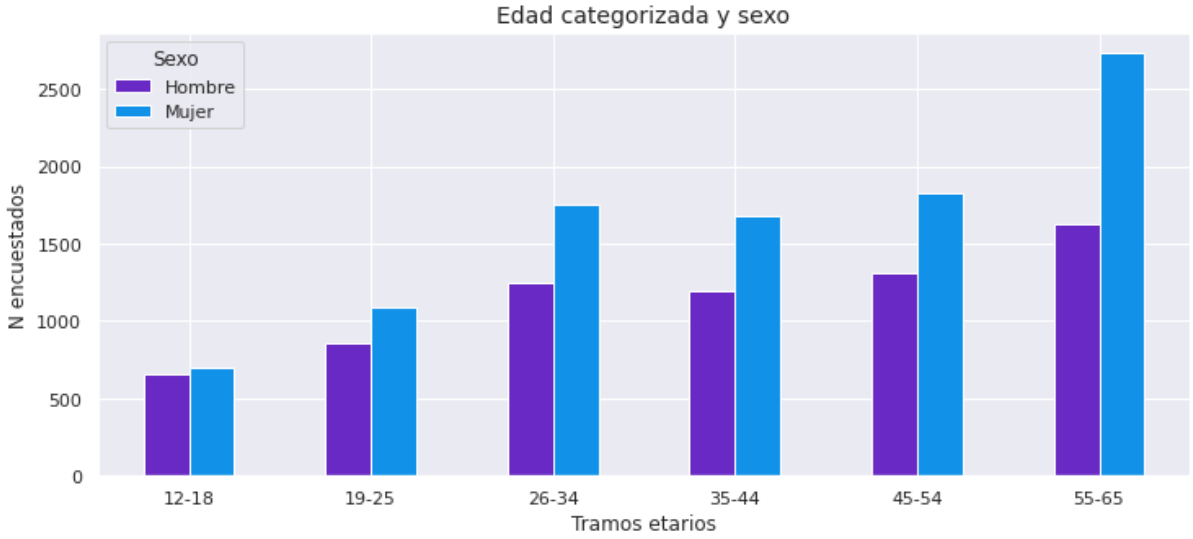

Consume marihuana
16000
14000
12000
10000
8000
6000
4000
2000
0
Frecuencia
No
Sí

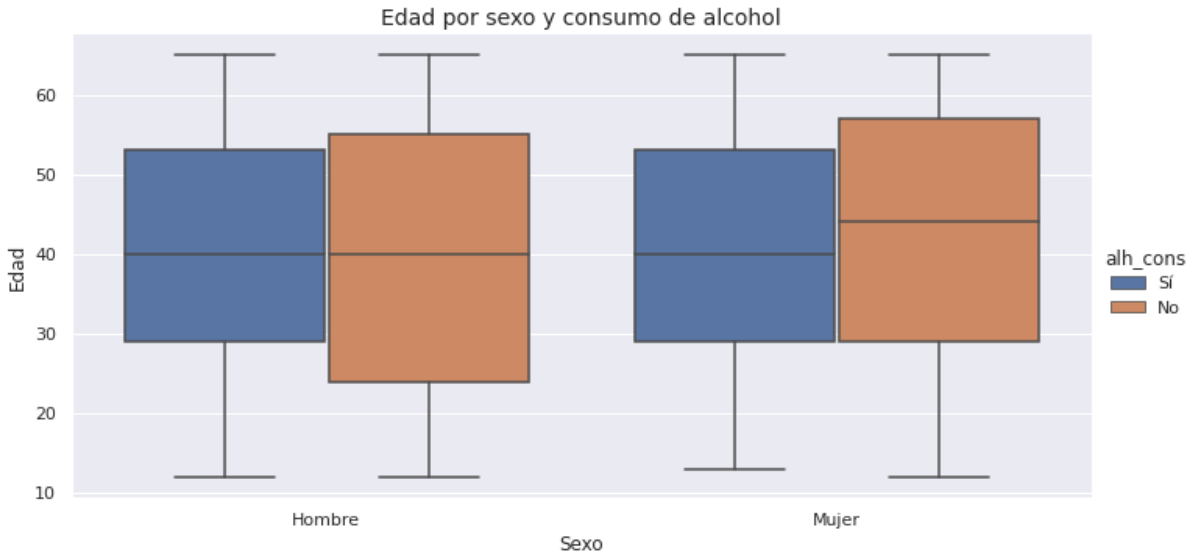

¿Ha probado Ud. cocaína alguna vez en su vida?
16000
14000
12000
10000
8000
6000
4000
2000
0
Frecuencia
Sí
No
No sabe
No responde
respuestas

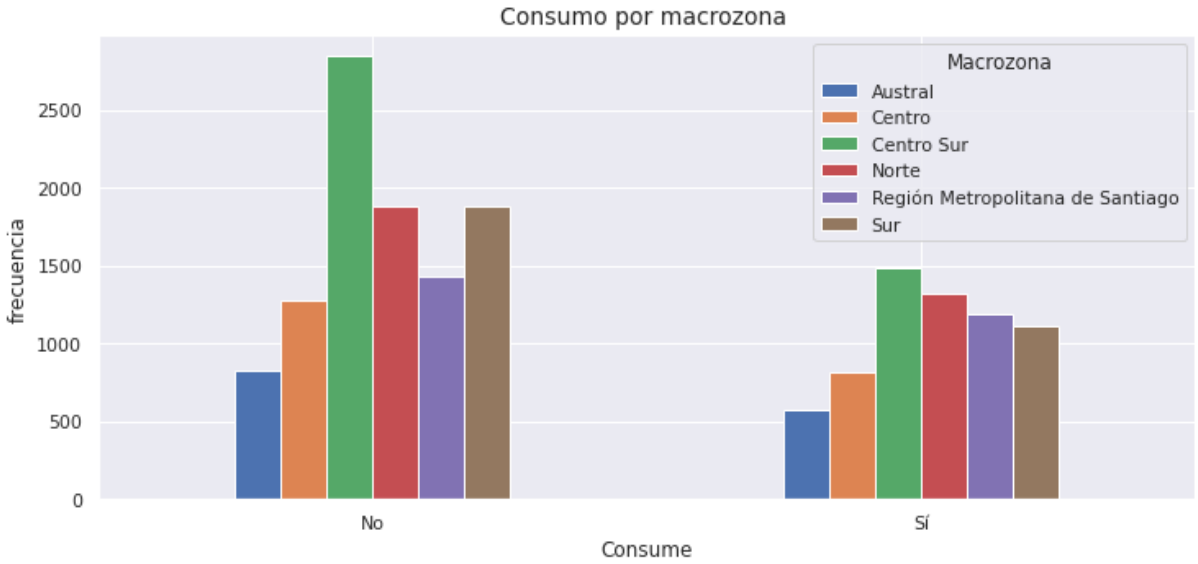


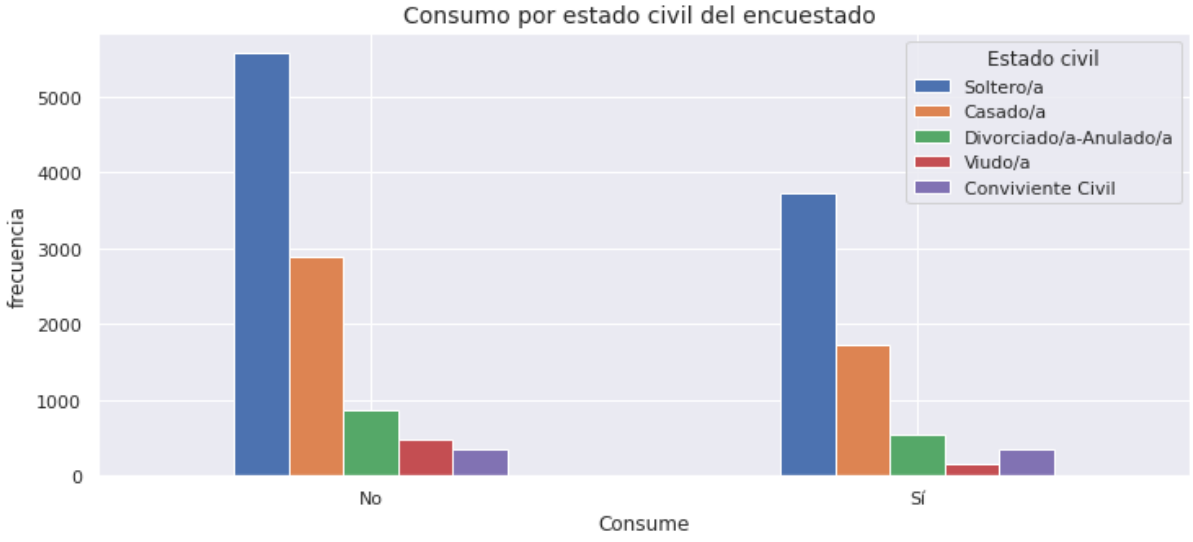


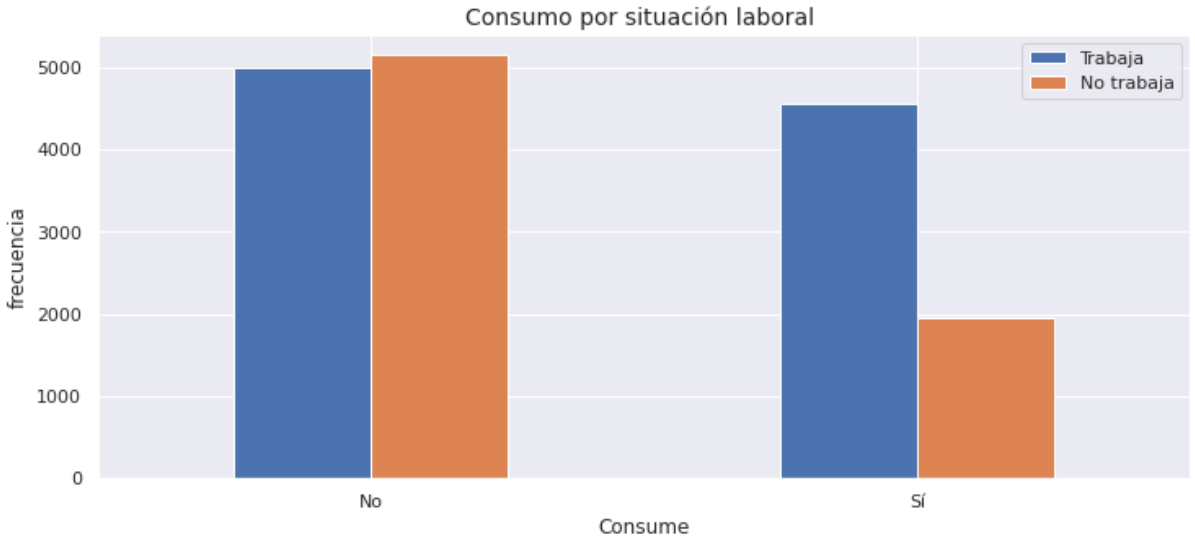
Análisis bivariado según consumo de alcohol

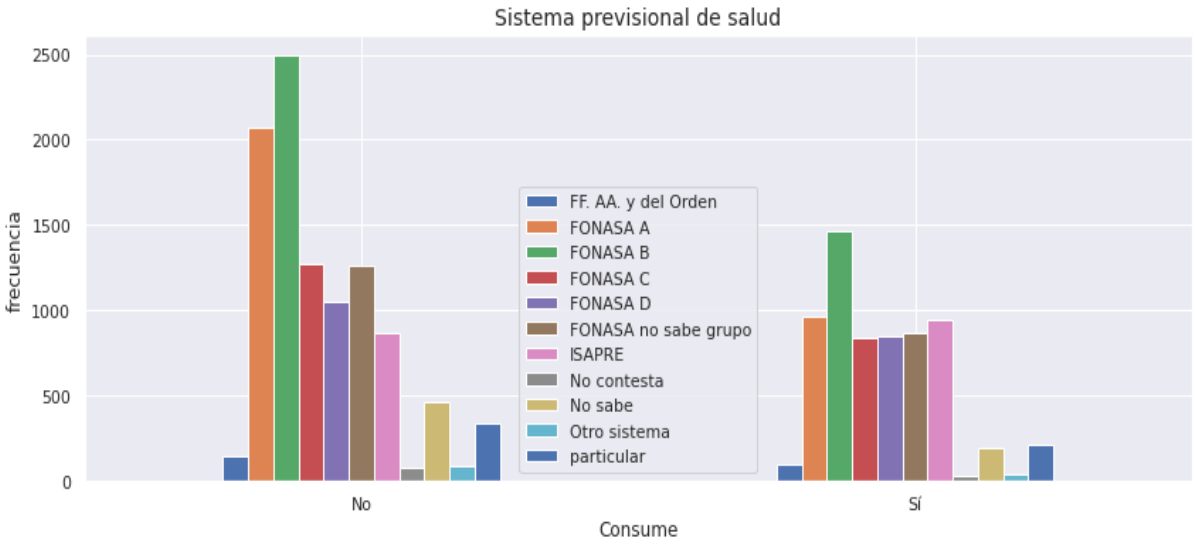


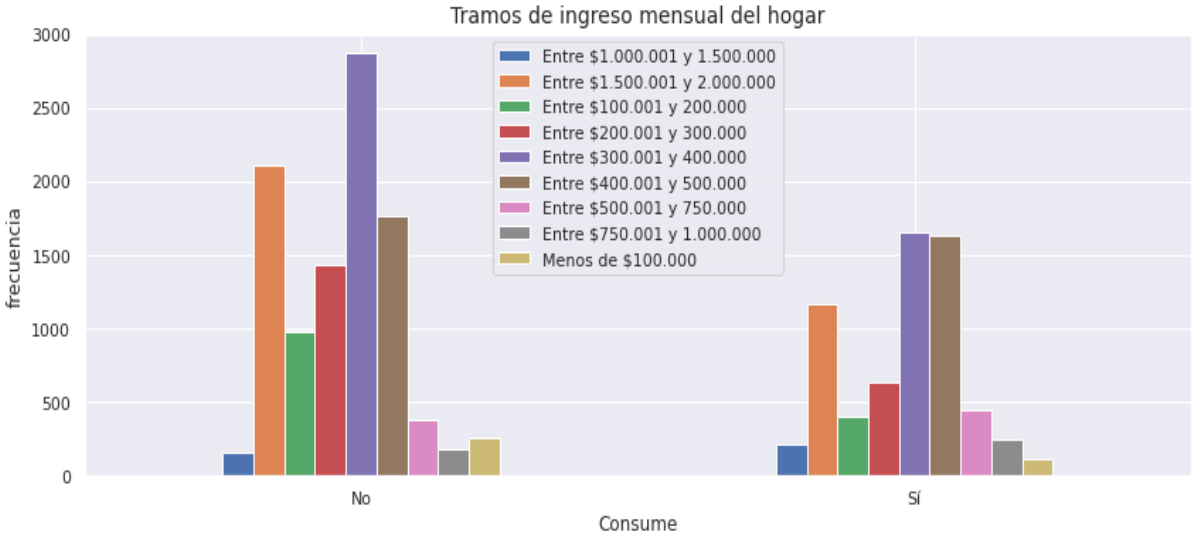

Edad por sexo y consumo de alcohol
Edad
60
50
40
30
20
10
Hombre
Mujer
Sexo
alh_cons
Sí
No

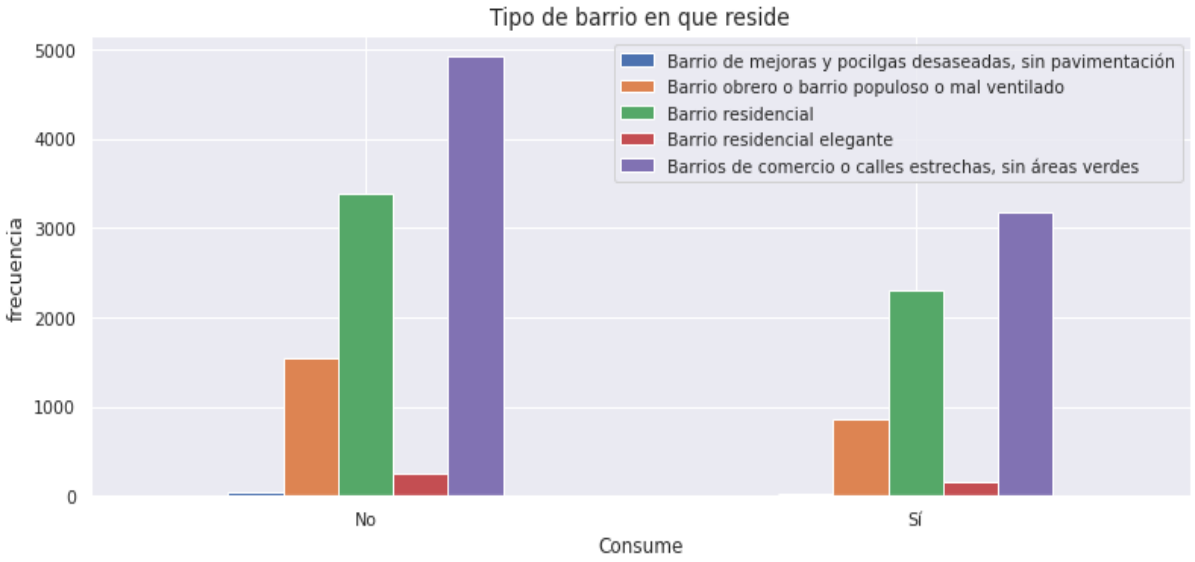

Consumo por macrozona
frecuencia
2500
2000
1500
1000
500
0
No
Sí
Consume
Macrozona
Austral
Centro
Centro Sur
Norte
Región Metropolitana de Santiago
Sur

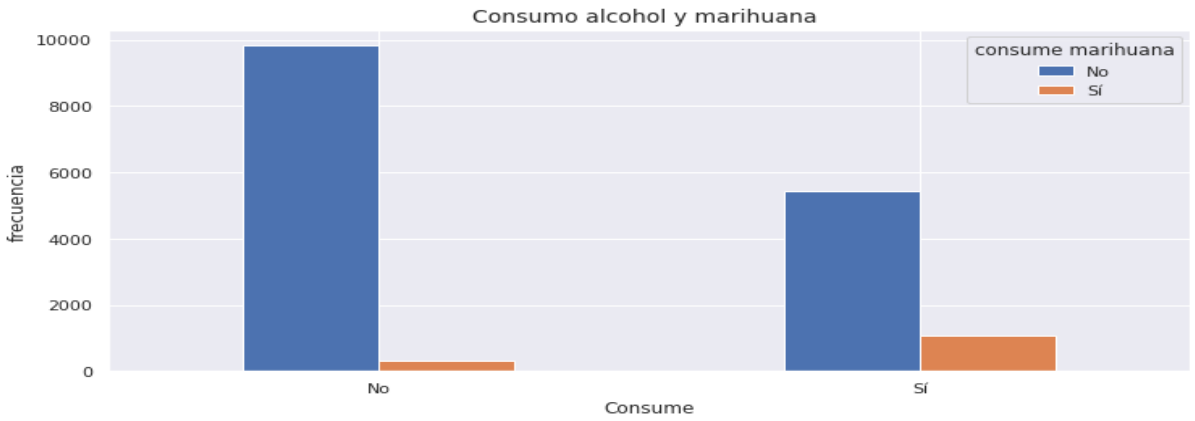

Consumo por estado civil del encuestado
frecuencia
5000
4000
3000
2000
1000
0
No
Sí
Consume
Estado civil
Soltero/a
Casado/a
Divorciado/a-Anulado/a
Viudo/a
Conviviente Civil

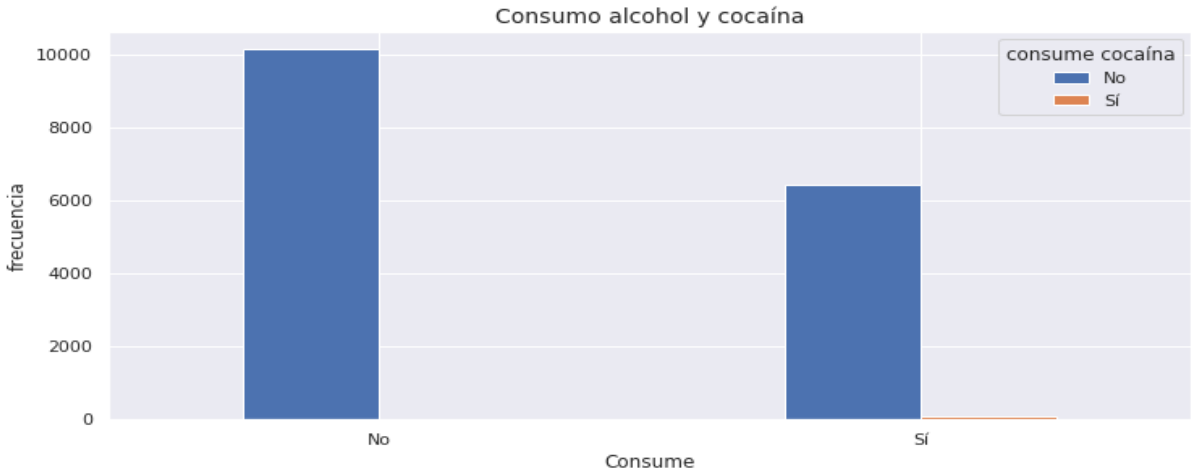

Consumo por situación laboral
Trabaja
No trabaja
frecuencia
5000
4000
3000
2000
1000
0
No
Sí
Consume


Sistema previsional de salud
FF. AA. y del Orden
FONASA A
FONASA B
FONASA C
FONASA D
FONASA no sabe grupo
ISAPRE
No contesta
No sabe
Otro sistema
particular
frecuencia
2500
2000
1500
1000
500
0
No
Sí
Consume


Tramos de ingreso mensual del hogar
Entre $1.000.001 y 1.500.000
Entre $1.500.001 y 2.000.000
Entre $100.001 y 200.000
Entre $200.001 y 300.000
Entre $300.001 y 400.000
Entre $400.001 y 500.000
Entre $500.001 y 750.000
Entre $750.001 y 1.000.000
Menos de $100.000
frecuencia
3000
2500
2000
1500
1000
500
0
No
Sí
Consume


Tipo de barrio en que reside
Barrio de mejoras y pocilgas desaseadas, sin pavimentación
Barrio obrero o barrio populoso o mal ventilado
Barrio residencial
Barrio residencial elegante
Barrios de comercio o calles estrechas, sin áreas verdes
frecuencia
5000
4000
3000
2000
1000
0
No
Sí
Consume


Consumo alcohol y marihuana
consume marihuana
No
Sí
frecuencia
10000
8000
6000
4000
2000
0
No
Sí
Consume


Consumo alcohol y cocaína
consume cocaína
No
Sí
frecuencia
10000
8000
6000
4000
2000
0
No
Sí
Consume

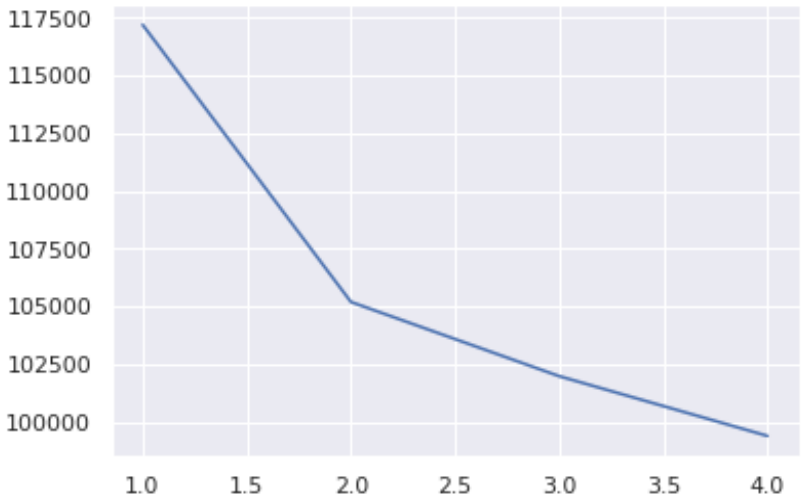
Clustering con el método K - modes

A continuación, se presenta el resultado de la aplicación del algoritmo de aprendizaje no supervisado de clustering por modas para el cálculo de centroides en variables categóricas.

- Primero se construyó una base de datos copia de la principal, luego se utilizó from sklearn import preprocessing para codificar las variables a numéricas con LabelEncoder.

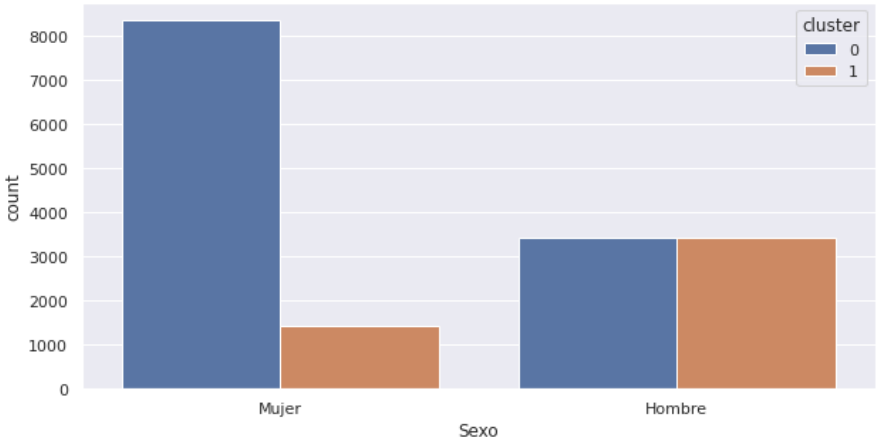
	S01	REGIONNOM	macrozona	edad_cat	est_salud	tram_ingr	educ_niv	sis_sld	area_lab	tip_barrio	alh_cons	mar_cons	coc_cons
0	1	6	1	3	0	3	3	1	5	1	1	0	0
1	1	6	1	4	7	3	3	2	9	1	1	0	0
2	0	6	1	3	0	5	4	1	0	1	1	0	0
3	0	6	1	0	0	4	12	4	5	2	0	0	0
4	1	6	1	3	0	1	3	2	5	1	0	0	0

- Se instala la librería KModes.
- Se busca el número óptimo de grupos y se visualiza con el método de codo, obteniendo la siguiente gráfica que muestra que en los dos grupos en se encuentra el óptimo:

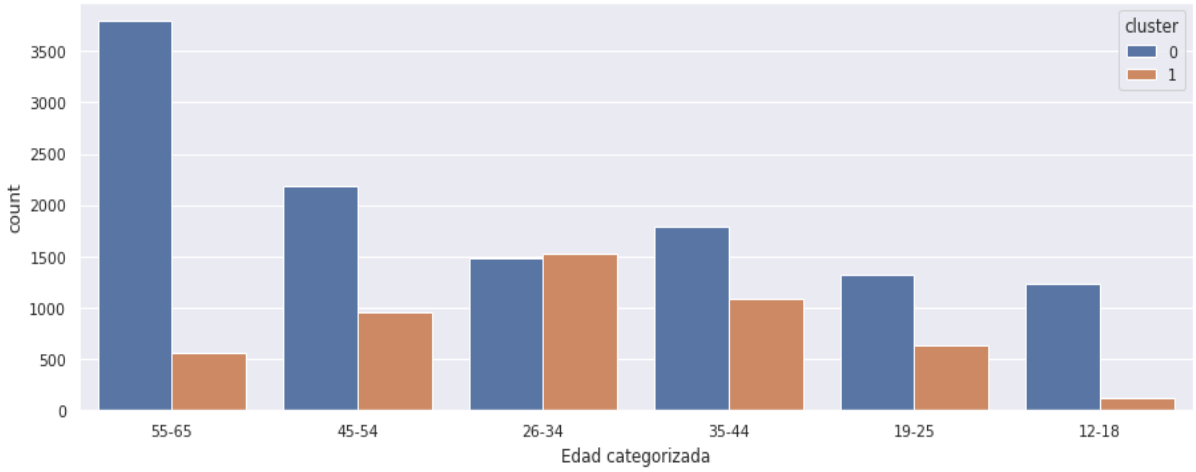


Visualización de agrupamientos obtenidos

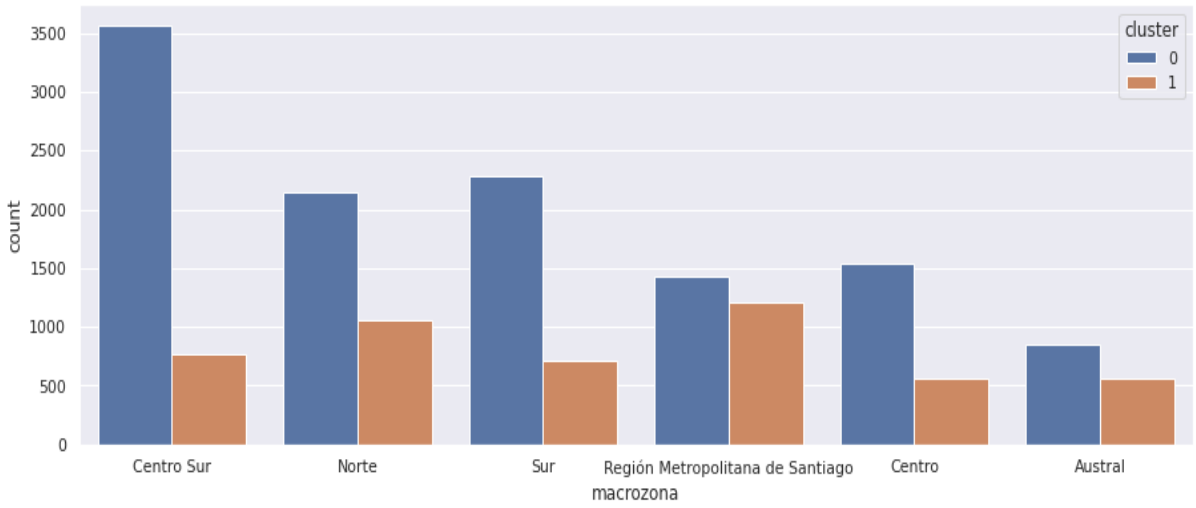
Se presentan a continuación las principales gráficas que resultaron del proceso de agrupamiento y su interpretación.



Se observa que en el primer cluster ‘0’, se agrupó a la mayoría de la mujeres y en el cluster ‘1’ a hombres, igualando en cantidad entre ambos cluster.

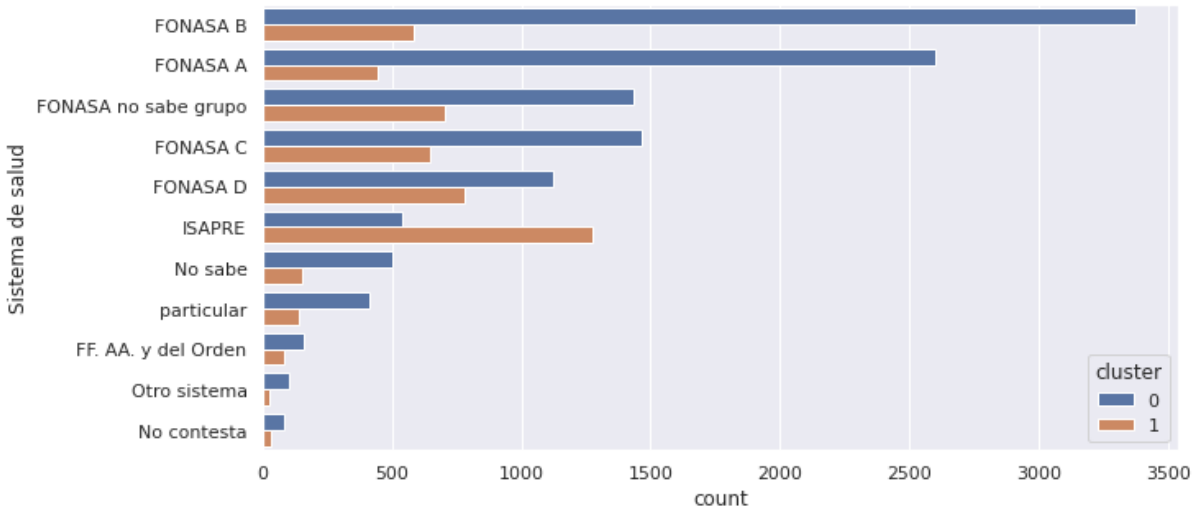


Al observar la edad categorizada y su agrupamiento cluster, se obtiene a una mayor cantidad de personas entre los 55 a 65 años de edad en el cluster ‘0’ y en el segundo cluster, la mayoría de edad categorizada estuvo presente entre los 26 a 34 años. La menor medida del primer cluster estuvo en el rango de 12 a 18, así como también lo fue para el segundo cluster.

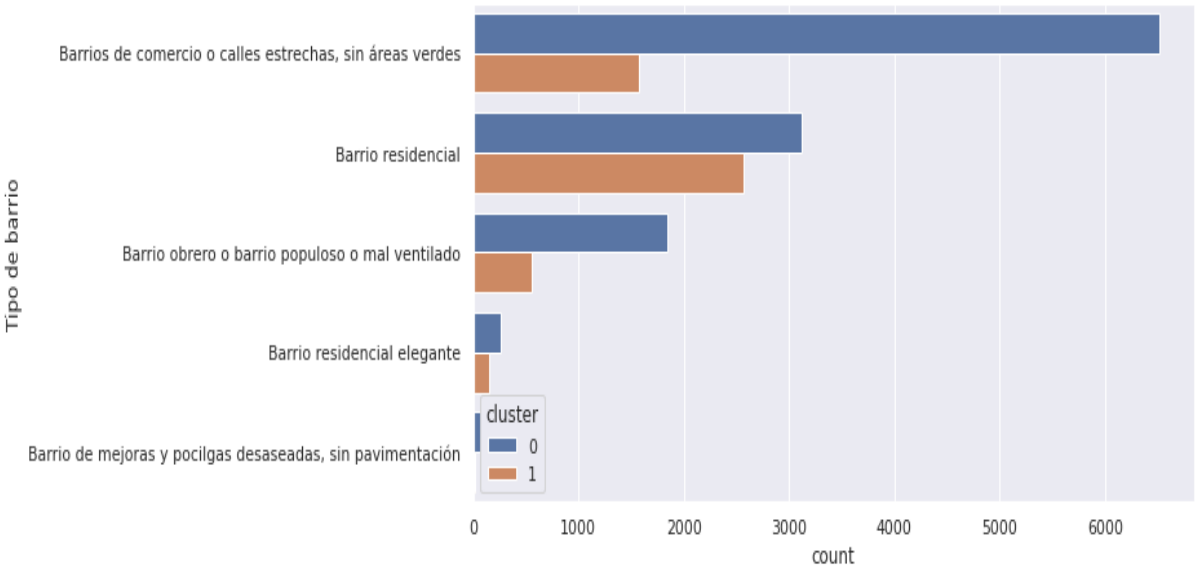


Al agrupar los cluster por macrozona se observa mayor presencia en la zona Centro sur en el caso del primer cluster, y en el segundo cluster una mayor presencia en ala zona de la Región Metropolitana.

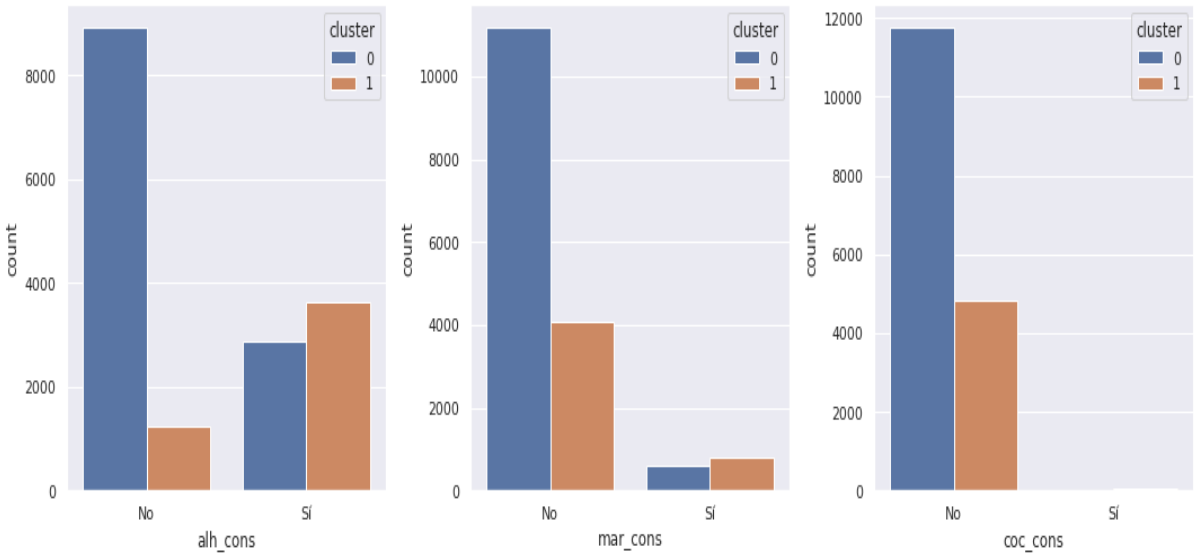
La variables ‘Estado de Salud’ es primordial para este estudio al ser una pregunta que va directamente hacia la percepción personal de los encuestados. Se observa que la agrupación se centró en el estado de buena condición de salud para ambos clusters.



También ligado a la variable anterior se debe analizar el sistema de salud de los encuestados y su relación al momento de agrupar. Se obtiene que el primer cluster pertenece a 'FONASA tramo B' seguido de 'FONASA tramo A', y para el segundo cluster se observa que hay una mayoría en el sistema 'ISAPRE' seguido en 'FONASA tramo D'.



Al agrupar según 'tipo de barrio', item que debía ser rellenado por criterio del encuestador, se obtiene que en el primer cluster la mayoría está en 'barrios de comercios o calles estrechas', y en el segundo cluster la mayoría se agrupó en 'barrio residencial'.



Y para finalizar el análisis de cluster, al tomar las variables respecto al consumo de sustancias se obtiene; para el consumo de alcohol una importante presencia del cluster ‘1’ y en menor medida del cluster ’0’ que sí consume, lo mismo se da en relación al consumo de marihuana. Y Finalmente, el consumo de cocaína no presenta grandes cantidades de encuestados que respondieron de manera afirmativa.

d. Calidad

Luego de realizar un análisis exploratorio con las variables categóricas seleccionadas en un primer momento desde la base de datos de la encuesta del SENDA, se observa que, según cantidad de consumidores, y por razón de balance de datos, la estrategia para continuar con el proyecto de modelamiento debe tomar como variable dependiente la referente al consumo de alcohol. Con esto, se deja de lado a la marihuana y a la cocaína como variables secundarias de momento, por la baja cantidad de consumidores, pero con los cuales se podrían implementar técnicas de balanceo de datos según mejores resultados. Cabe mencionar que, previamente, luego de una lectura general del informe que sintetiza la base que sustenta este proyecto, se omitió la información respecto a las demás sustancias que contenía el cuestionario aplicado, estas eran: tabaco y cigarrillos electrónicos, pasta base, bebidas energéticas, y otras drogas.

La variable sobre el ingreso mensual del hogar no es de tipo numérica, en el cuestionario se presentaba solamente como categorías de alternativas entre rangos que el encuestado debía estimar y seleccionar. Esto le quita posibilidades de manipulación, pero sí sirve para el modelamiento final donde debe integrarse con sus categorías. Podría haber una manera de generar valores random entre los rangos y

según sus frecuencias, esto con el fin de implementar cluster agrupados con la variable numérica de la edad o años de educación. Sobre esta última variable, se señalan los niveles de educación pero no si están completos o incompletos - como sí lo indica la encuesta CASEN por ejemplo- con lo cual igual se podrían estimar años de referencia para emular un procedimiento anterior.

La variable sobre el “tipo de barrio” está incompleta respecto a la función que cumple desde la observación del encuestador. Esta va junta a otra pregunta sobre el tipo de vivienda, ambas son rellenadas por el encuestador sin mencionar nada al encuestado, y a partir de una suma hecha entre las dos se estima el nivel socioeconómico de la persona. Esto se señala en el informe de SENDA de esta manera, pero no muestra el método o criterio de segmentación que realizan y sólo presentan la agrupación según tres niveles socioeconómicos. Al no tener certeza de cómo incluir esta información, sólo se incluyó la caracterización del barrio como imagen general.

2- Preparación de datos:

a. Selección:

-alh_cons = variable que indica si la persona consume o no alcohol
-Edad
-Estado de salud
-Región
-Macrozona
-Ingresos mensuales del hogar
-Educación
-Situación laboral
-Sistema previsional de salud
-Tipo de barrio

b. Limpieza

Respecto a la limpieza que fue necesaria realizar en la base de datos, según el análisis exploratorio, no implicó modificar o imputar grandes cantidades de datos.
Los valores nulos fueron:

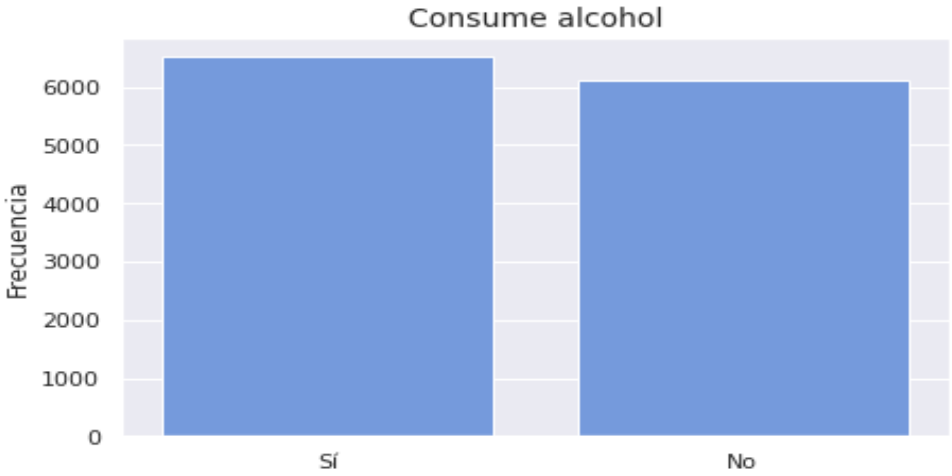
```
Valores porcentuales de nulos
COC_4         95.378706
MAR_4         69.793542
CO_6          31.568839
OH_4          24.144761
mar_cons       0.000000
alh_cons       0.000000
tip_barrio     0.000000
```

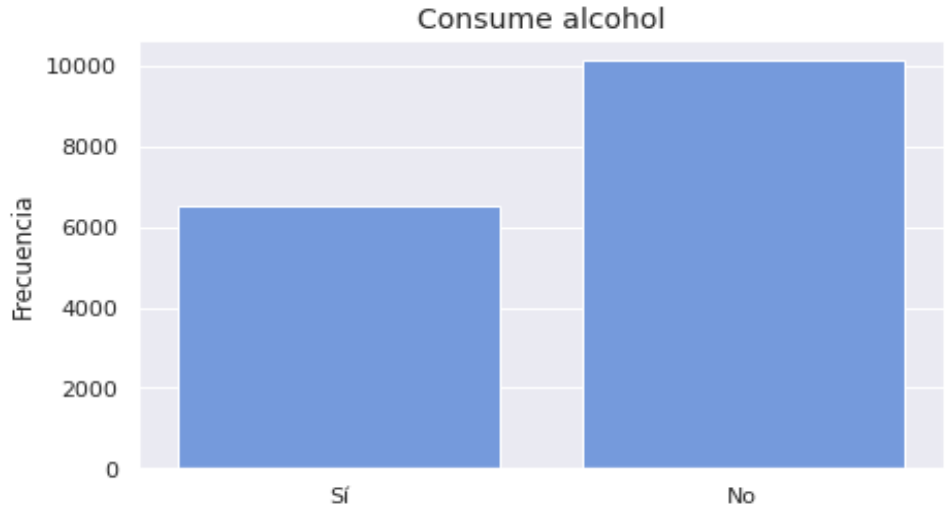
```
area_lab       0.000000
sis_sld        0.000000
educ_niv       0.000000
tram_ingr      0.000000
edad_cat       0.000000
macrozona      0.000000
REGIONNOM      0.000000
DP_9           0.000000
SbjNum         0.000000
COC_1          0.000000
S01            0.000000
MAR_1          0.000000
OH_1           0.000000
DP_18          0.000000
CO_1           0.000000
DP_12          0.000000
DP_16          0.000000
DP_5           0.000000
DP_2           0.000000
S02            0.000000
REGION         0.000000
coc_cons       0.000000
```

Las variables que muestran grandes datos nulos (coc_4 , mar_4 y oh_4) son las que preguntaban acerca del periodo de tiempo que había pasado desde el último consumo realizado, lo que daba 'sin información' debido a la pregunta previa sobre si las consumía o no: coc_1 , mar_1 y oh_1 (cocaína, marihuana y alcohol).

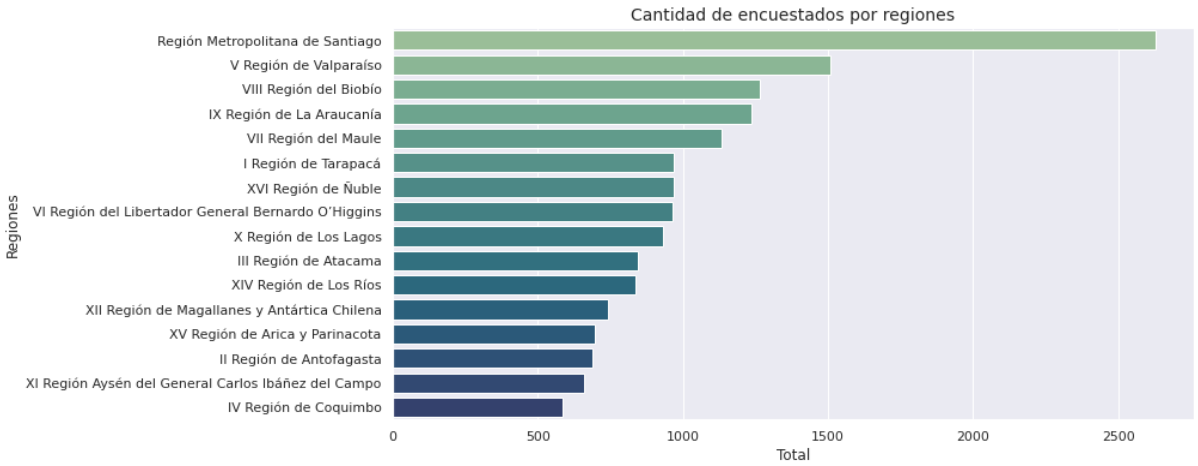
Para ordenar y categorizar de mejor manera el consumo de los encuestados, codifiqué tres columnas nuevas:
-'alh_cons', con los que habían respondido que habían consumido alcohol dentro del último mes (consideré sólo esta opción al ser muy abierta la posibilidad de consumir alcohol en situaciones que no indiquen dependencia más allá del mes)
-'coc_cons', con los que habían respondido que habían consumido cocaína dentro del último mes y la segunda opción que señalaba más de un mes pero menos de un año
-'mar_cons' igualmente con los que habían respondido que habían consumido cocaína dentro del último mes y la segunda opción que señalaba más de un mes pero menos de un año.

La totalidad de estas variables tomaban sólo a los que habían respondido afirmativamente la primera pregunta respecto al consumo, esto dejaba altos porcentajes de valores nulos, y para que no generen problemas en el modelo a implementar en el siguiente informe, se imputaron los valores nulos dándoles el valor de no consumidores, así de esta forma la base de datos queda más clara con la totalidad etiquetada. Entonces las variables quedan compuestas por quienes agrupé como consumidores y quienes no respondieron las siguientes preguntas. Luego de esto sólo escogí a los consumidores de alcohol para el análisis.



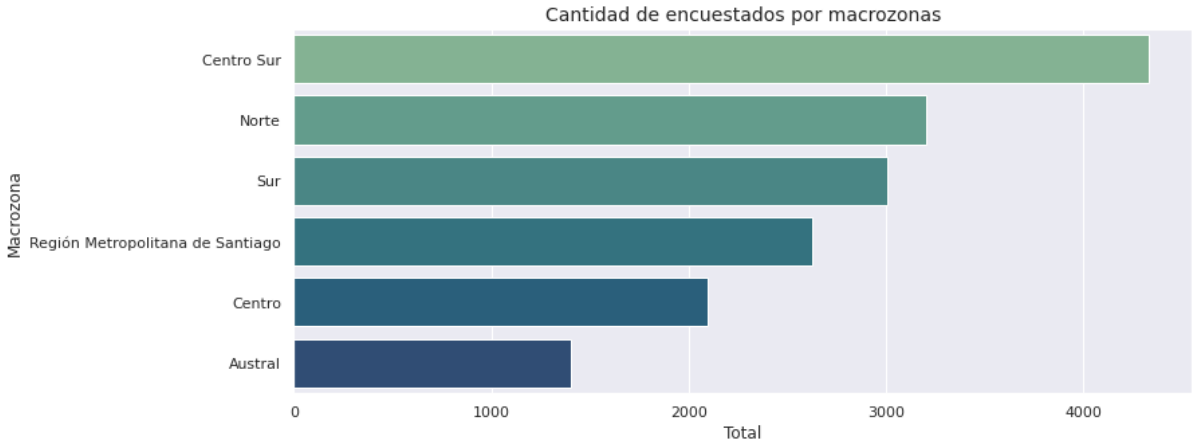


La variable que indica la región de los encuestados también fue categorizada en otra columna. Primero les agregué los nombres a las regiones ya que sólo indicaban el número de estas. Luego las modifiqué por macrozonas.




```
#Categorización de regiones según macrozonas
```

```
df['macrozona'] = df['REGION'].map( {     1: 'Norte',
                                          2: 'Norte',
                                          3: 'Norte',
                                          4: 'Centro',
                                          5: 'Centro',
                                          6: 'Centro Sur',
                                          7: 'Centro Sur',
                                          8: 'Centro Sur',
                                          9: 'Sur',
                                         10: 'Sur',
                                         11: 'Austral',
                                         12: 'Austral',
                                         13:  'Región Metropolitana de
Santiago',
                                    14:'Sur',
                                    15:'Norte',
                                    16: 'Centro Sur'}
                                      ).astype(object)
```



II Parte

Modelamiento

1. Selección:

Para el modelamiento del proyecto se utilizarán Árboles de decisiones y pruebas de Naives Bayes. Ambos se compararán en sus resultados de clasificación para observar cual posee mejor rendimiento y puede demostrar de mejor manera cuales son las características de los consumidores de sustancias que respondieron la encuesta SENDA.

El árbol de decisiones tiene la ventaja de que según las características va realizando el corte por nodos y ramas hasta llegar a la hoja final que determina la clasificación buscada. Para esto se determina el criterio de ganancia de información, que en este caso será la 'entropía', al ser un modelo con variables categóricas, y la profundidad se determinará con algoritmos de 'sklearn.model_selection'.

El modelo Gaussian Naive Bayes trabaja con el supuesto de independencia de las características y la distribución normal de estas, calcula la probabilidad de cada clase dada a partir del conjunto de datos ingresados. Puede ser susceptible a sobreestimar la importancia de ciertas características por lo que se le considera de una interpretabilidad compleja pero de fácil aplicación a nivel de código.

Así, se probarán dos modelos de machine learning utilizados para trabajar con variables categóricas, con diferentes parámetros y consideraciones al momento de evaluar los resultados de sus indicadores de rendimiento.

2. Definir pruebas

Para medir rendimiento de los modelos se utilizará una matriz de confusión con su respectiva gráfica, esto permite señalar la cantidad de errores cometidos en la estimación y observar qué tipo de errores se está cometiendo; si estima demasiados falsos positivos, que serían consumidores cuando no lo son, o muchos falsos negativos, no consumidores cuando sí lo son.Por lo tanto, el objetivo del modelo es calcular el porcentaje total de clases positivas respecto de los reales totales.

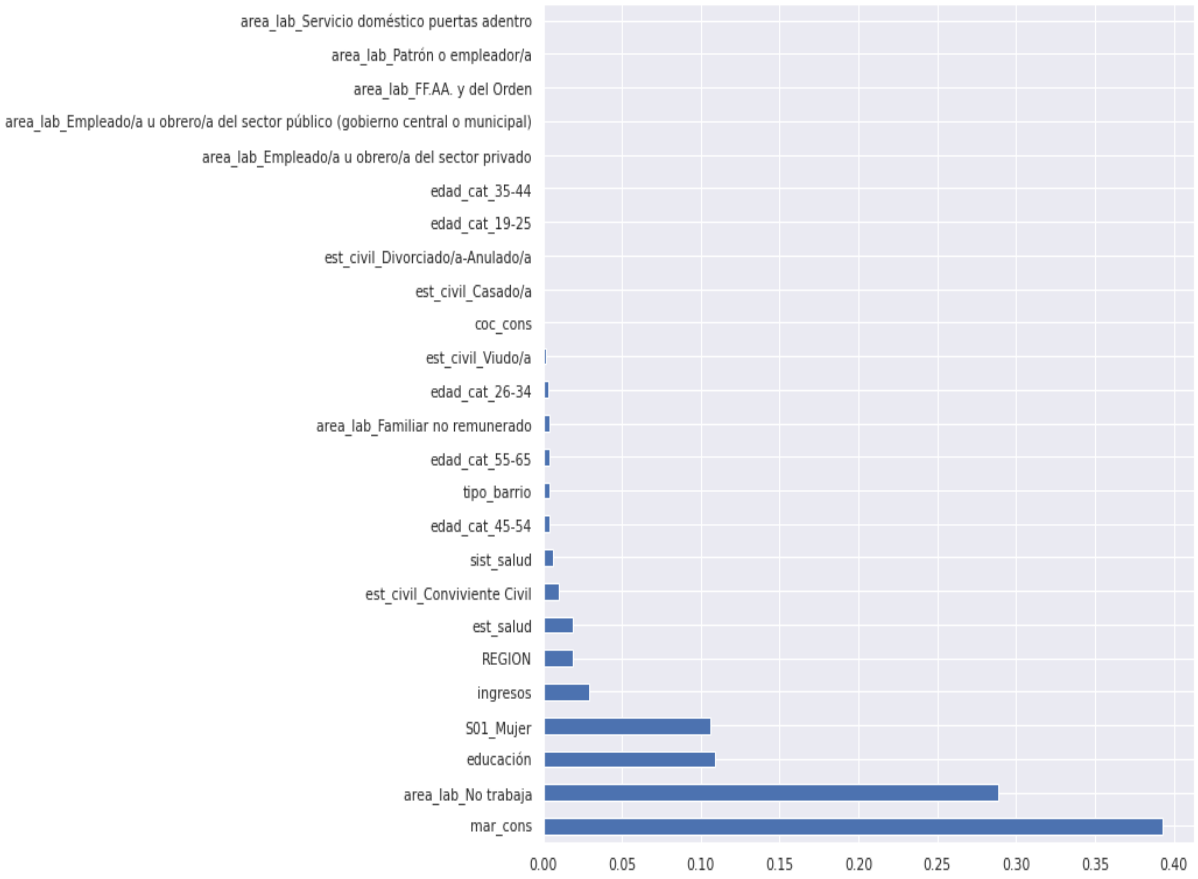
Para lo anterior señalado se considerarán las métricas del indicador de rendimiento 'recall', ya que se busca calificar bien los verdaderos positivos de clase 2, sin sobre calificar con errores a los no consumidores. Se puede considerar además el indicador F1 score al ser una métrica intermedia entre precisión y recall, es decir, la media que puede considerar falsos negativos (pero que calificarían indebidamente a los

encuestados) y la métrica que indica a los verdaderos. No se considera el indicador ‘accuracy’ por ser clases evidentemente desbalanceadas. Con crossvalidation se verificará que las iteraciones del modelo son confiables sin dar medidas dispares o muy alejadas de las dadas por los resultados. Se codificará para que trabaje con el indicador de calidad Recall.

3. Diseño y aplicación

Se utilizará un split de la data de 80/20, se probará con los consumidores de alcohol y los de marihuana por separado, los consumidores de cocaína se omitirán por poseer un cantidad de casos muy reducida en relación al total.

Los parámetros de los ‘Árboles de decisiones’ se obtendrán con el algoritmo Grid Search y Randomizaed Search de ‘sklearn.model_selection’, así, se compararán ambos resultados de estos indicadores de los óptimos niveles de min_split y max_depth. Luego se evaluará la importancia de las variables con ‘tree_class.feature_importances_’, se evaluará y procederá. En criterion se seleccionará el método ‘entropia’ que mide la ganancia de información de cada variable categórica.
Además, para los árboles se consideró evaluar la importancia de las variables. Se consideró que hay muchas con valor cero o muy bajo, se procedió a eliminarlas de la base pero al realizar nuevamente la iteración, los valores incluso bajaron alrededor de .05 décimas de valor de recall, por lo tanto, no se considera quitar las variables con baja importancia. Se presenta a continuación la gráfica obtenida de las variables.



Para el caso de la prueba de Naives Bayes, se utilizará simplemente el modelo Gaussian.
En cuanto al desbalance de los datos, se procederá a aplicar cuatro modos diferentes de balanceo, estos serán:
- Balanceo por penalización, que equilibra la clase minoritaria penalizando a la mayoritaria.
- Balanceo Under Sampling, que elimina datos de la clase mayoritaria.
- Balanceo Over Sampling, que replica datos de la clase minoritaria.
- Balanceo SMOTE, que utiliza muestras sintéticas de la clase minoritaria.
Con estos métodos se probará cuales dan mejores resultados para los objetivos de clasificación del proyecto y se usará crossvalidation para estimar la precisión.

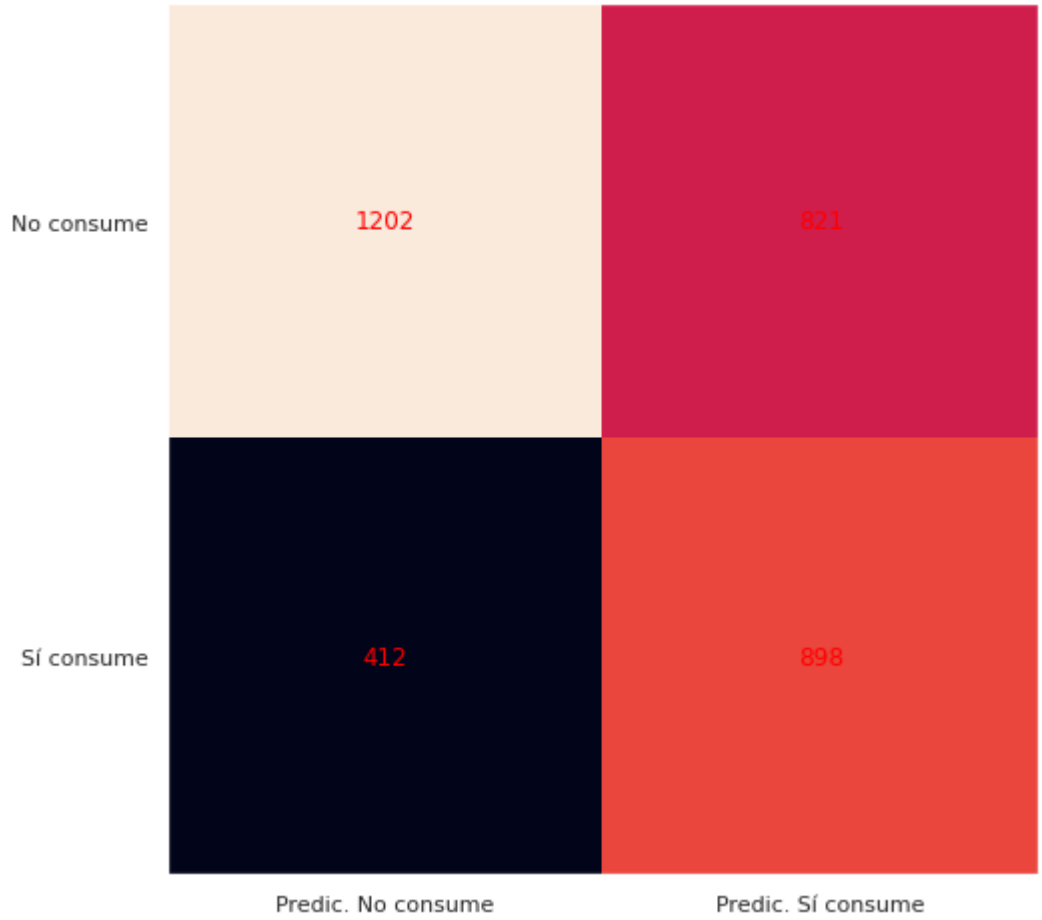
Se aplicará balance de datos frente a los siguientes valores:
-Consumen alcohol
no 10143
sí 6519

-Consumen marihuana
no 15256
sí 1406

4. Evaluación del rendimiento

4.1. Para la iteración en la clasificación para detectar a los ‘consumidores de alcohol’, entre los modelos con mejor rendimiento, están:

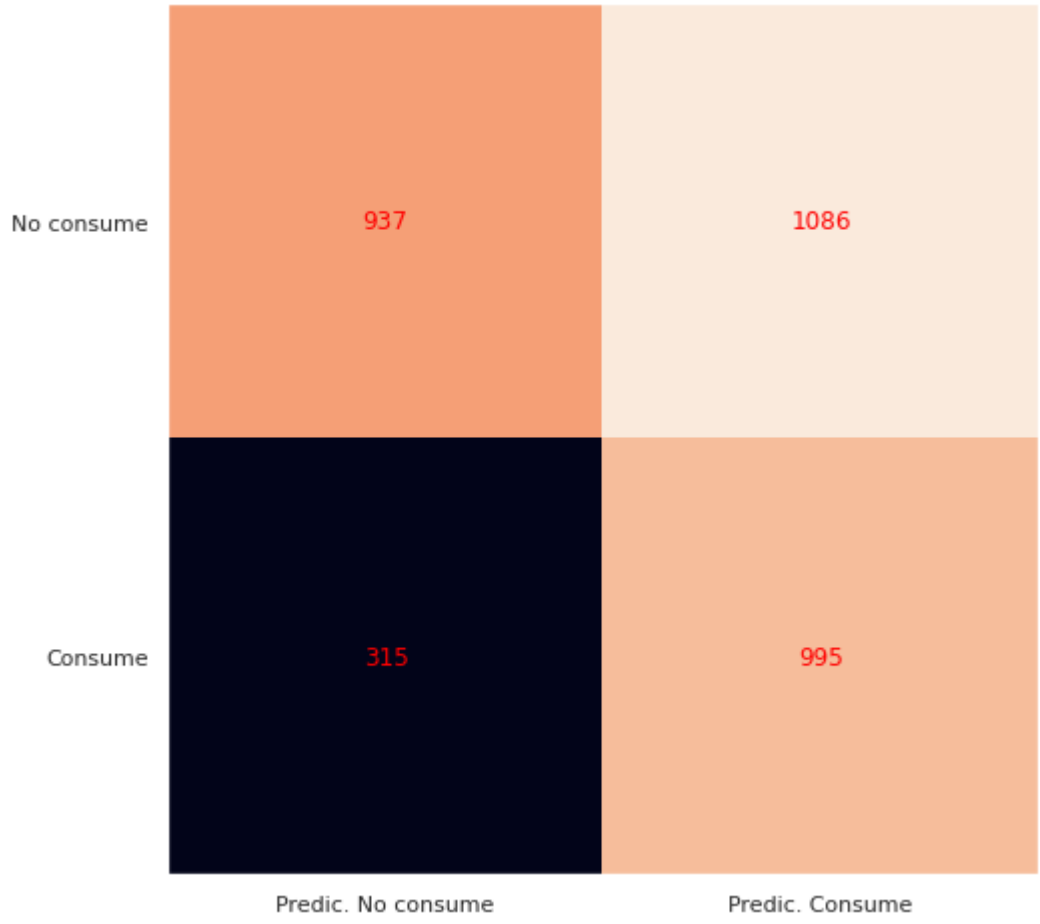
- **Para árbol de decisión la siguiente matriz de confusión**



	precision	recall	f1-score	support
0	0.74	0.59	0.66	2023
1	0.52	0.69	0.59	1310
accuracy			0.63	3333
macro avg	0.63	0.64	0.63	3333
weighted avg	0.66	0.63	0.63	3333

```
[[1202  821]
 [ 412  898]]
recall en el set de Test: 0.69
crossvalidation: 0.6761374544058361
```

- **Para Gaussian Naives Bayes la siguiente matriz de confusión**



	precision	recall	f1-score	support
0	0.75	0.46	0.57	2023
1	0.48	0.76	0.59	1310
accuracy			0.58	3333
macro avg	0.61	0.61	0.58	3333
weighted avg	0.64	0.58	0.58	3333

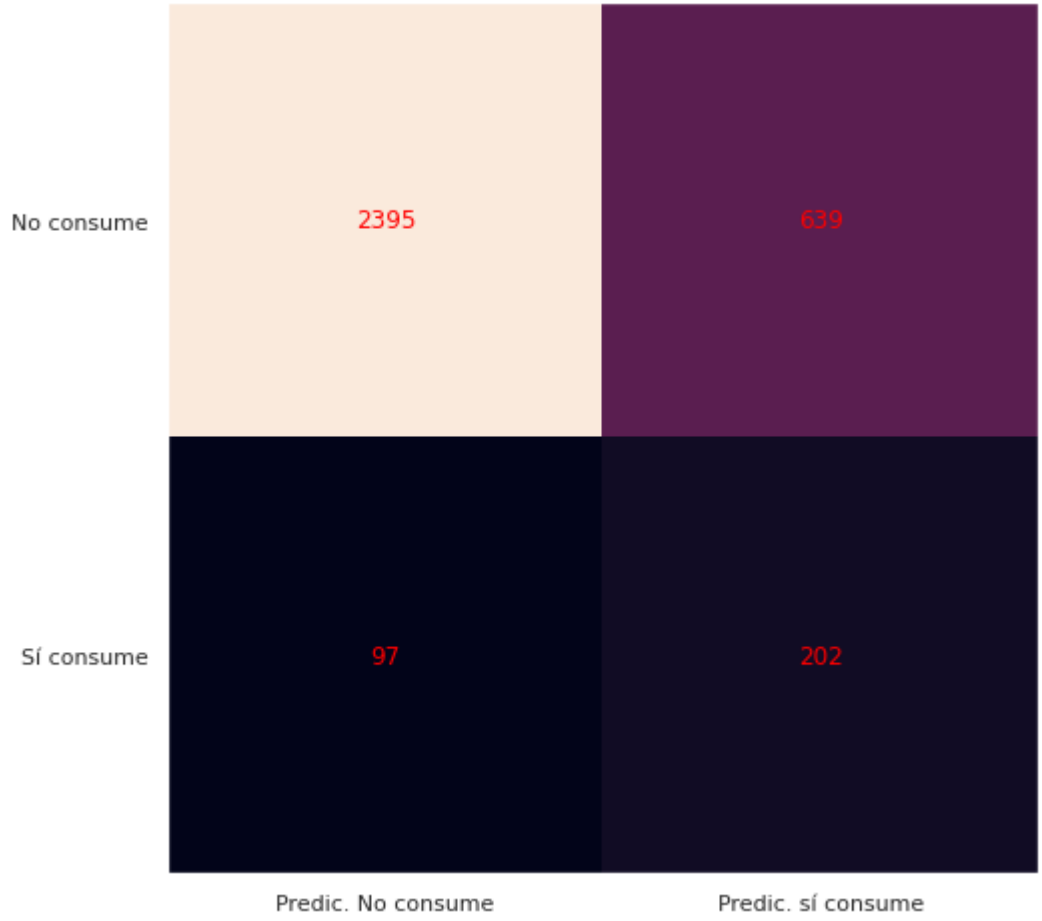
```
[[ 937 1086]
 [ 315  995]]
Recall en el set de Test: 0.76
crossvalidation: 0.6793103448275862
```

- **Conclusiones del modelo clasificador de consumidores de alcohol:**

Se observó que para los Árboles de Decisiones se trabajó de mejor manera el método de balanceo de datos por Under Sampling. Los resultados de recall fueron de 69% con un f1 score de 59%, es decir pudo clasificar a los verdaderos positivos de clase 2 pero dejando a varios afuera y además clasificando a falsos positivos (821 casos). El crossvalidation de la prueba dio 0.67, sin variaciones grandes entre diferentes iteraciones.
En el modelo de Naives Bayes se trabajó de mejor manera con el método de balanceo Smote. Los resultados de recall fueron de 76% con un f1 score de 59%. Clasificó por cien casos mejor que el modelo anterior, pero también significó clasificar más falso positivos (1086 casos), más de la mitad. El crossvalidation de la prueba dio 0.67, es decir una notoria variación entre iteraciones.
En este caso optaría por escoger el árbol de decisiones, puede ser un 7% más bajo pero se muestra más estable en sus diferentes pruebas. Además, clasificó menos falsos positivos en consideración de que se busca obtener un modelo que demuestre las características de los consumidores.

4.2. Para la iteración en la clasificación para detectar a los ‘consumidores de alcohol’, entre los modelos con mejor rendimiento, están:

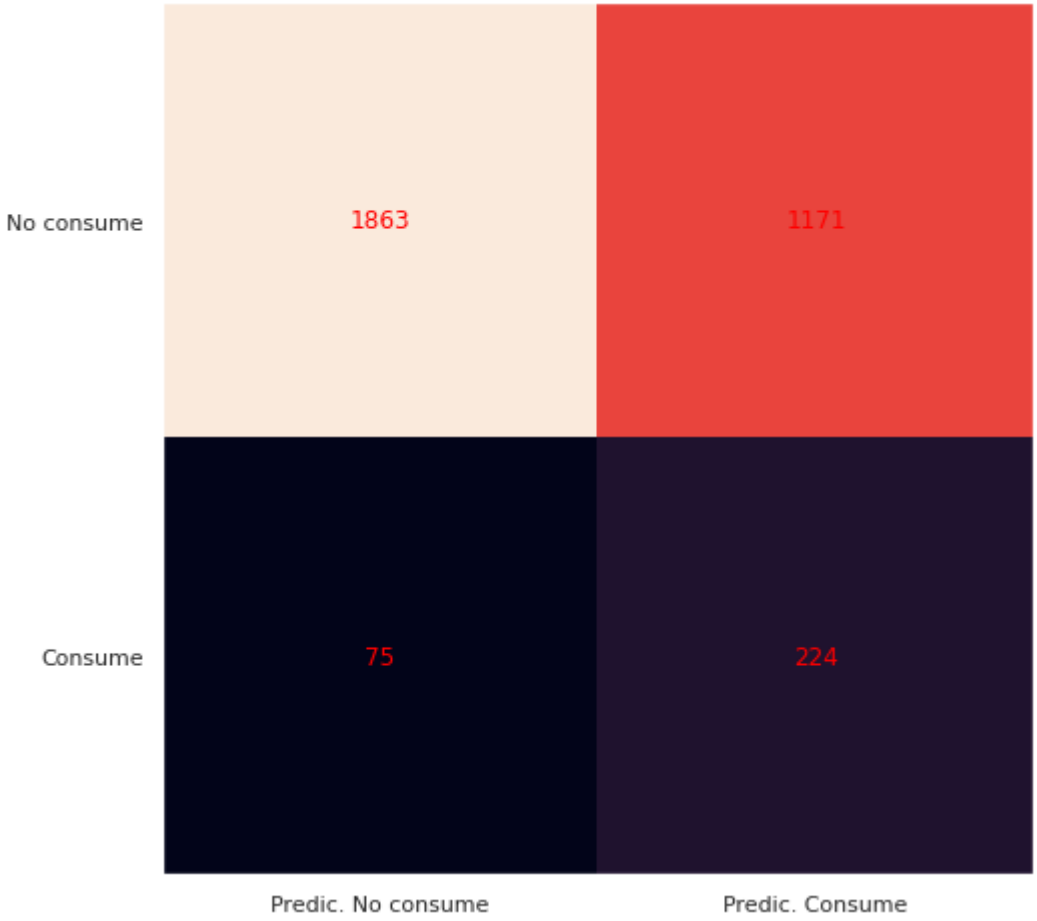
- **Para árbol de decisión la siguiente matriz de confusión**



	precision	recall	f1-score	support
0	0.96	0.79	0.87	3034
1	0.24	0.68	0.35	299
accuracy			0.78	3333
macro avg	0.60	0.73	0.61	3333
weighted avg	0.90	0.78	0.82	3333

```
[[2395  639]
 [  97  202]]
recall en el set de Test: 0.68
crossvalidation: 0.7064137308039747
```

- **Para Gaussian Naives Bayes la siguiente matriz de confusión**



	precision	recall	f1-score	support
0	0.96	0.61	0.75	3034
1	0.16	0.75	0.26	299
accuracy			0.63	3333
macro avg	0.56	0.68	0.51	3333
weighted avg	0.89	0.63	0.71	3333

```
[[1863 1171]
 [  75  224]]
Recall en el set de Test: 0.75
crosvalidation: 0.6793103448275862
```

- **Conclusiones del modelo clasificador de consumidores de marihuana:**

Se observó que para los árboles de decisiones se trabajó de mejor manera el método de balanceo de datos por penalización para compensar. Los resultados de recall fueron de 68% con un f1 score de 35%, es decir pudo clasificar a los verdaderos positivos de clase 2 pero dejando a varios afuera y además clasificando a falsos positivos (639 casos). El crossvalidation de la prueba dio 0.70, sin variaciones grandes entre diferentes iteraciones.
En el modelo de naives bayes se trabajó de mejor manera con el método de balanceo Smote. Los resultados de recall fueron de 75% con un f1 score de 20% clasificó levemente mejor que el modelo anterior, pero también significó clasificar más falso positivos (1171). El crossvalidation de la prueba dio 0.67, es decir una notoria variación entre iteraciones.
Hay que mencionar que para este modelo hubieron pruebas que llegaron al 96% de recall indicando una muy buena clasificación de quienes sí consumen marihuana, pero esto también determinaba que los que no lo hacían fueron mal clasificados, una alta tasa de falsos positivos.
En este caso optaría por escoger el árbol de decisiones, puede ser un 7% más bajo pero se muestra más estable en sus diferentes pruebas. Además, clasificó menos falsos positivos en consideración de que se busca obtener un modelo que demuestre las características de los consumidores.

Evaluación

1. Evaluación de resultados

Los resultados obtenidos con las pruebas no llegaron a niveles óptimos de clasificación, sí pudieron diferenciar e identificar a un porcentaje importante de los dos tipos de consumidores de sustancias que se buscaron, pero según el análisis central del proyecto faltó construir un diseño donde se trabajara más con la ingeniería de datos, o quizá el tipo de preguntas de la encuesta central era muy compleja de modificar.
El árbol de decisiones permite aplicar parámetros para ir modificando el resultado de manera que se esclarezca más la intención de búsqueda, pero en este caso, al hacer la revisión de las ‘variables importantes’ se encontraron muchas de poca relevancia y que al sacarlas de la base no mejoraba los resultados. Por ello, se procedió sólo a identificarlas pero a mantenerlas.
Con Naives bayes se obtienen buenos resultados de clasificación pero se sacrifica mucho error, y al ser este trabajo un intento de clasificar para caracterizar a los consumidores de alcohol y marihuana, es poco ético no mirar hacia los falsos positivos que señalaban a los no consumidores como sí consumidores.

Fueron primordiales los métodos de balanceo de datos para obtener resultados acordes, esto fue dado a la naturaleza de la base de datos y el conjunto de preguntas y que se procede a explicar en el siguiente apartado.

2. Revisión del proceso

Dentro de las dificultades se puede considerar el desbalance de los datos, por eso se procedió sólo a determinar la clasificación entre alcohol y marihuana. La información contenida en la base de datos no daba para extraer mayor conjetura al momento de crear variables que indican a los encuestados como consumidores de tales sustancias. El proceso de codificar a estos consumidores se valió de forma arbitraria por el criterio del investigador, lo que puede además señalar un efecto negativo en la objetividad de las clasificaciones realizadas en la ingeniería de datos, pero esto era necesario para poder identificar las categorías con las cuales se trabajaría.

3. Determinar las próximas etapas

Al pensar en las próximas acciones que vendrían luego de un análisis como el presente, se podría evaluar la implementación de una encuesta que extraiga mayor información y también recurrir a otros medios para tener un panorama más abierto del consumo de sustancias en Chile. Siendo la intención del SENDA prevenir estos hábitos, la ciencia de datos permitiría realizar procesos de enfoque más rápidos al determinar cuáles son los grupos sociales más propensos a caer en adicciones. Este trabajo debe enfocarse en ese fin, prevenir el consumo, y para ello hay que superar los límites que tuvo presente en la manipulación de la información para lograr una caracterización con modelos con valores óptimos, no tan bajos como los alcanzados pero sí considerando el avance que se hizo para implementar metodología de clasificación.