Entrega final - Multivariado

Lucca Frachelle, Joaquín Silva, Cecilia Waksman

2025-02-07

Datos

Los datos utilizados en este análisis surgen de un estudio epidemiológico llevado a cabo en la Facultad de Odontología de la Universidad de la República, Uruguay, durante el período 2015-2016. Los mismos son recogidos de 602 individuos que acudieron a consulta.

Contamos con variables binarias de tipo: comportamental, ENT, odnontológicas. Además contamos con las variables: sexo, edad, ingreso, C, P, O y CPO, donde C es la cantidad de dientes con caries, P es la cantidad de dientes perdidos debido a caries y O la cantidad de dientes con caries obturadas, además CPO es la suma de las tres anteriores.

Datos

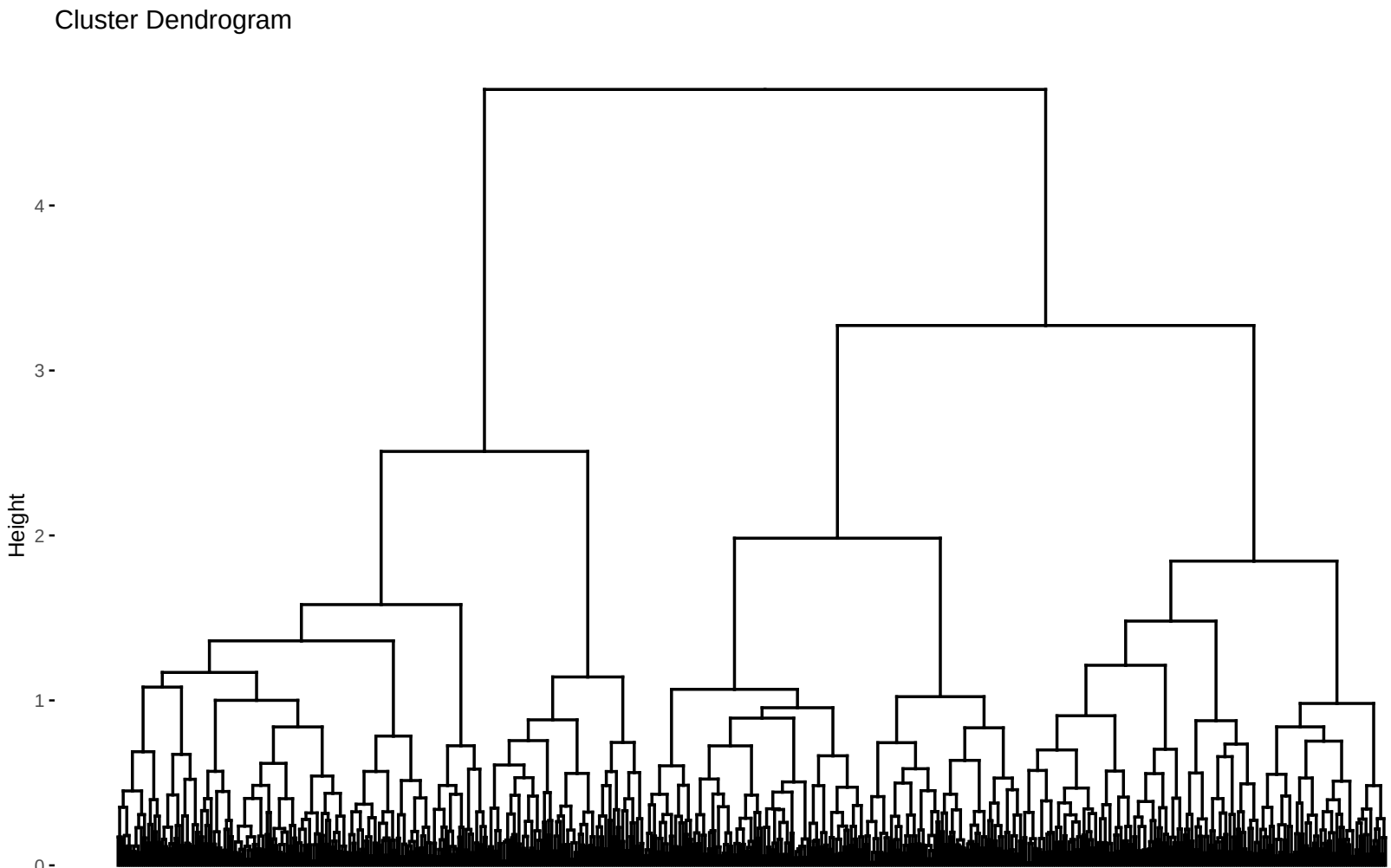
Table 1: Tabla de Variables

Variable	Descripción	Bloque	Tipo
V1	Fuma a diario	1	Comportamental
V2	Consumo nocivo de alcohol	1	Comportamental
V3	Actividad física insuficiente	1	Comportamental
V4	IMC sobrepeso/obesidad	2	ENT
V5	Razón de Cintura Cadera	2	ENT
V6	Hipertensión	2	ENT
V7	Diabetes	2	ENT
V8	Prev. bolsa	3	Odontológicas
V9	Pérdida Dentaria	3	Odontológicas
V10	Prevalencia de Caries	3	Odontológicas
V11	Prevalencia de PIP	3	Odontológicas

Cluster Jerárquico

Dendograma

Se utiliza el **método de Ward** para formar los clusters, los cuales se van agrupando en cada etapa de la forma en que se muestra en el siguiente dendograma.



Cluster Jerárquico

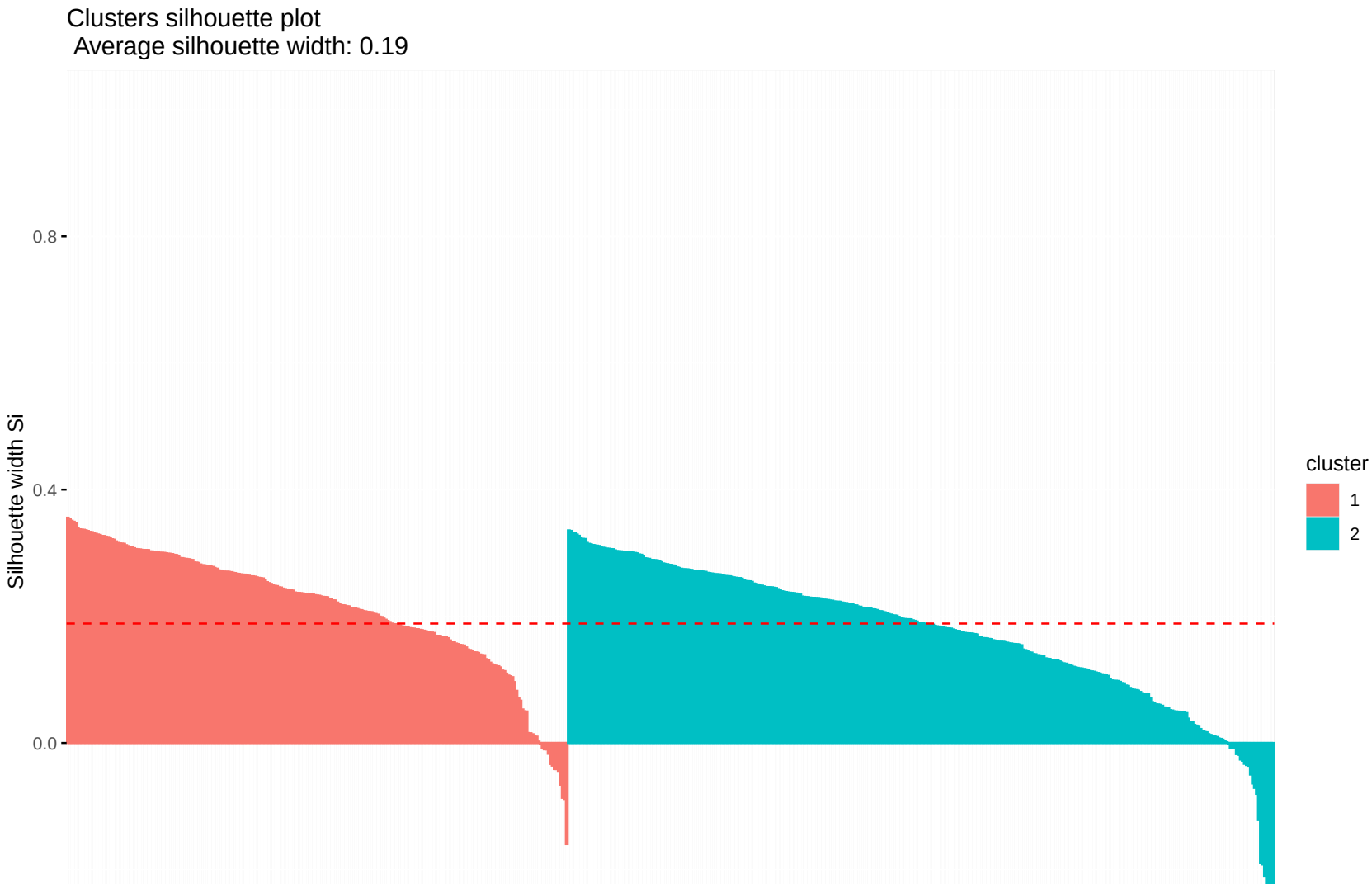
Para elegir la cantidad de clusters, se recurre a los índices: *Cindex*, *Silouhette*, *Mcclain*, *Dunn* y *Frey*.

	Cindex	Silohuette	Mcclain	Dunn	Frey
Ward	15	2	2	2	1

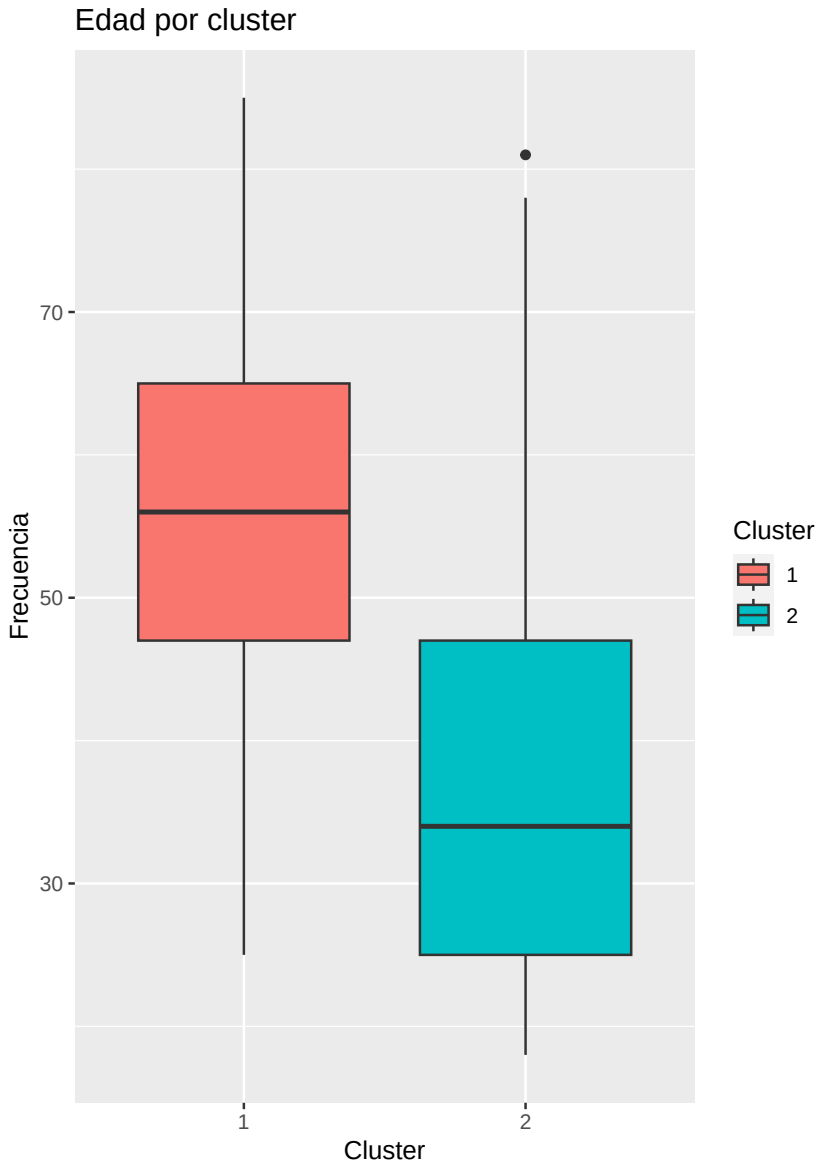
Como la mayoría de ellos plantean 2 como la cantidad aconsejable, se continuará el análisis para clusters jerárquicos en función de esta división.

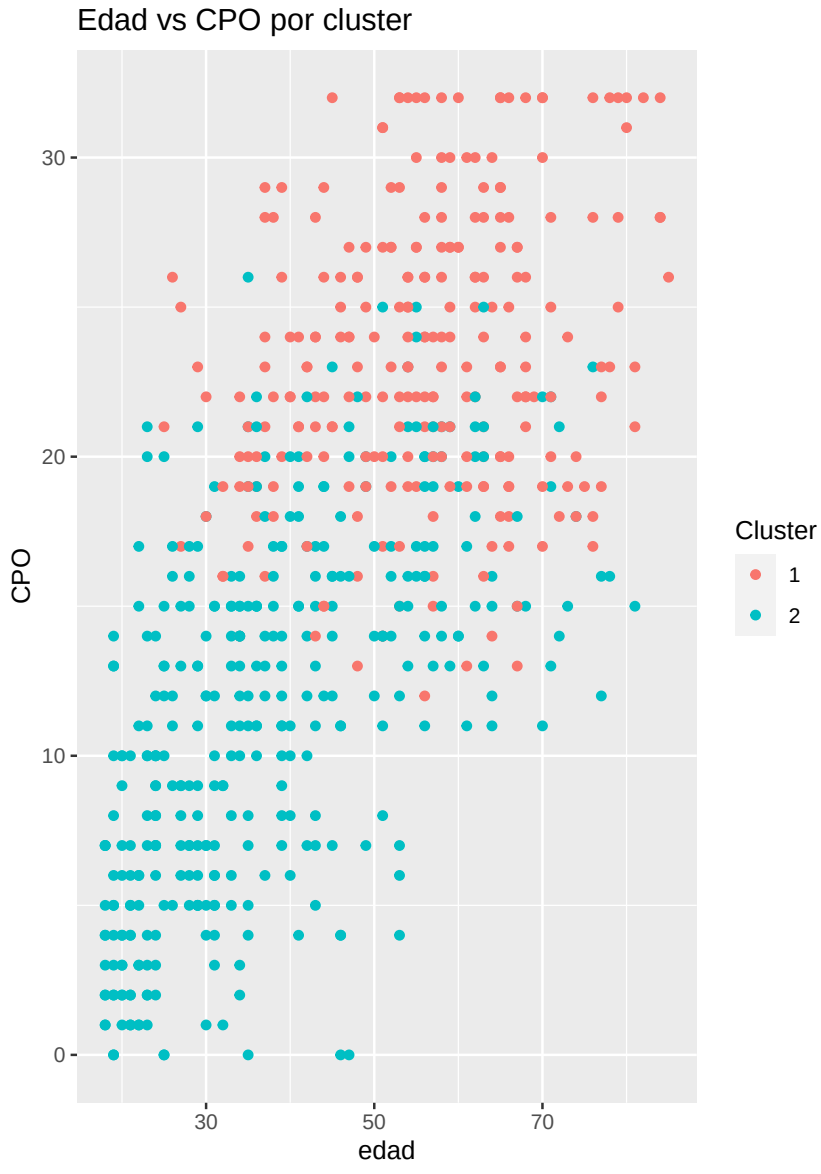
Cluster Jerárquico

	cluster	size	ave.sil.width
1	1	249	0.21
2	2	351	0.17



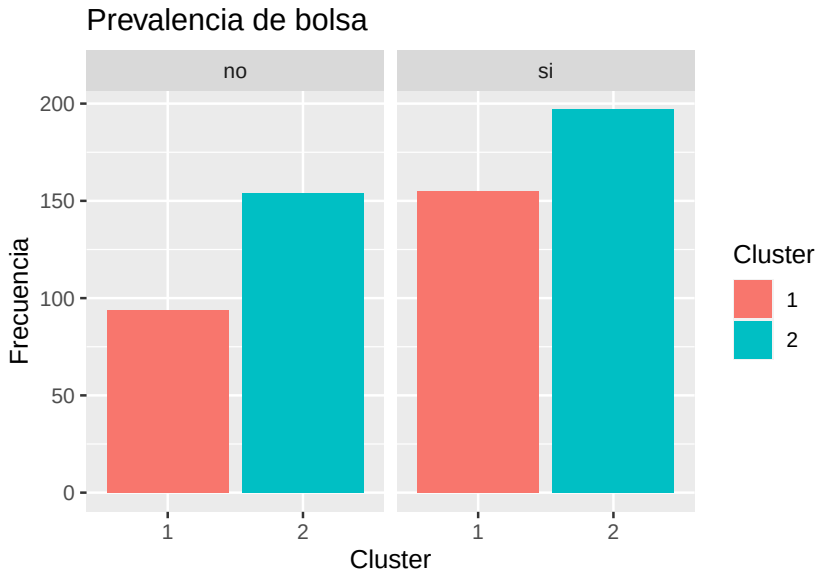
Cluster Jerárquico

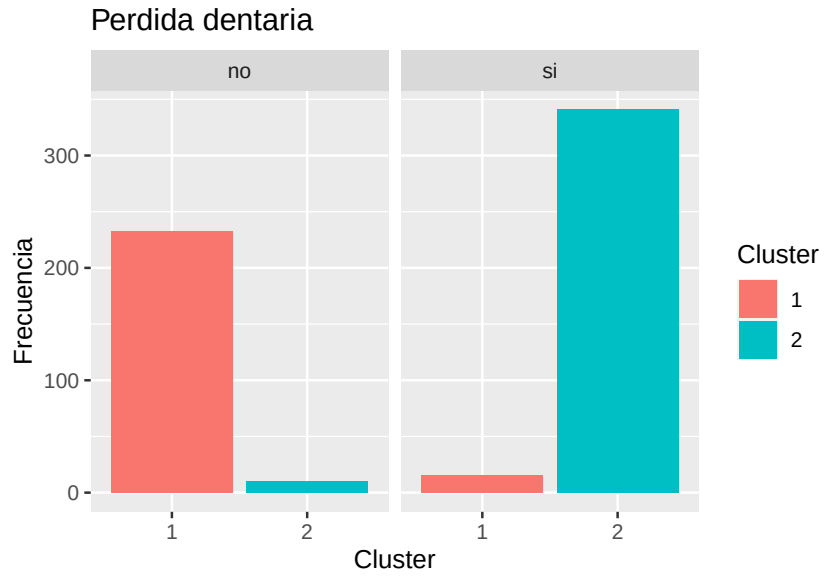


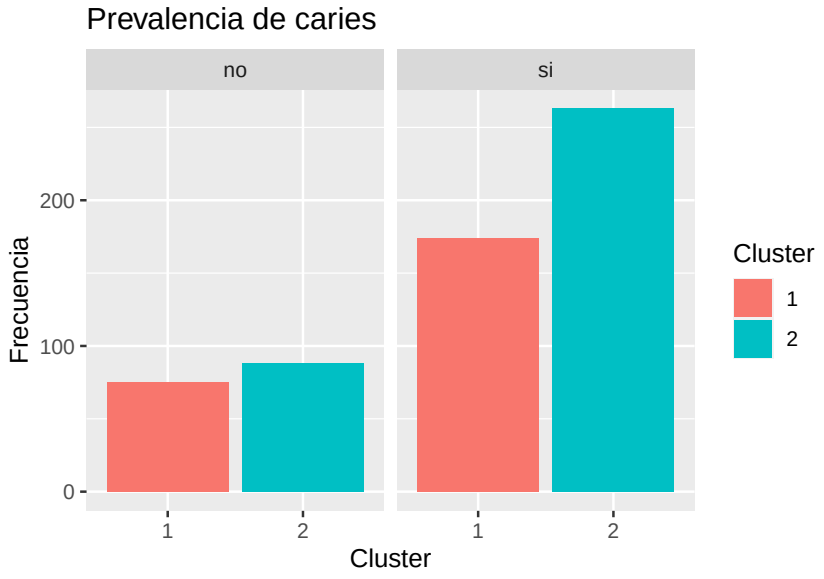


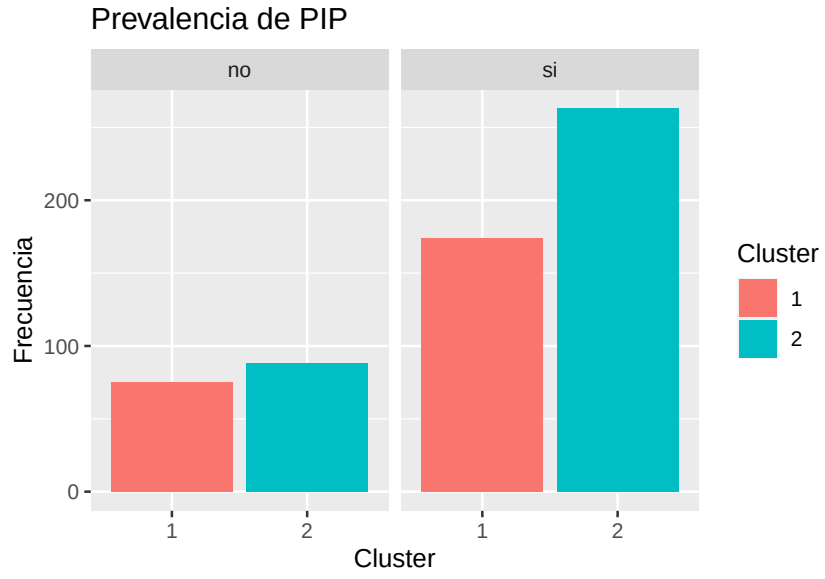
Cluster Jerárquico

Variables odontológicas



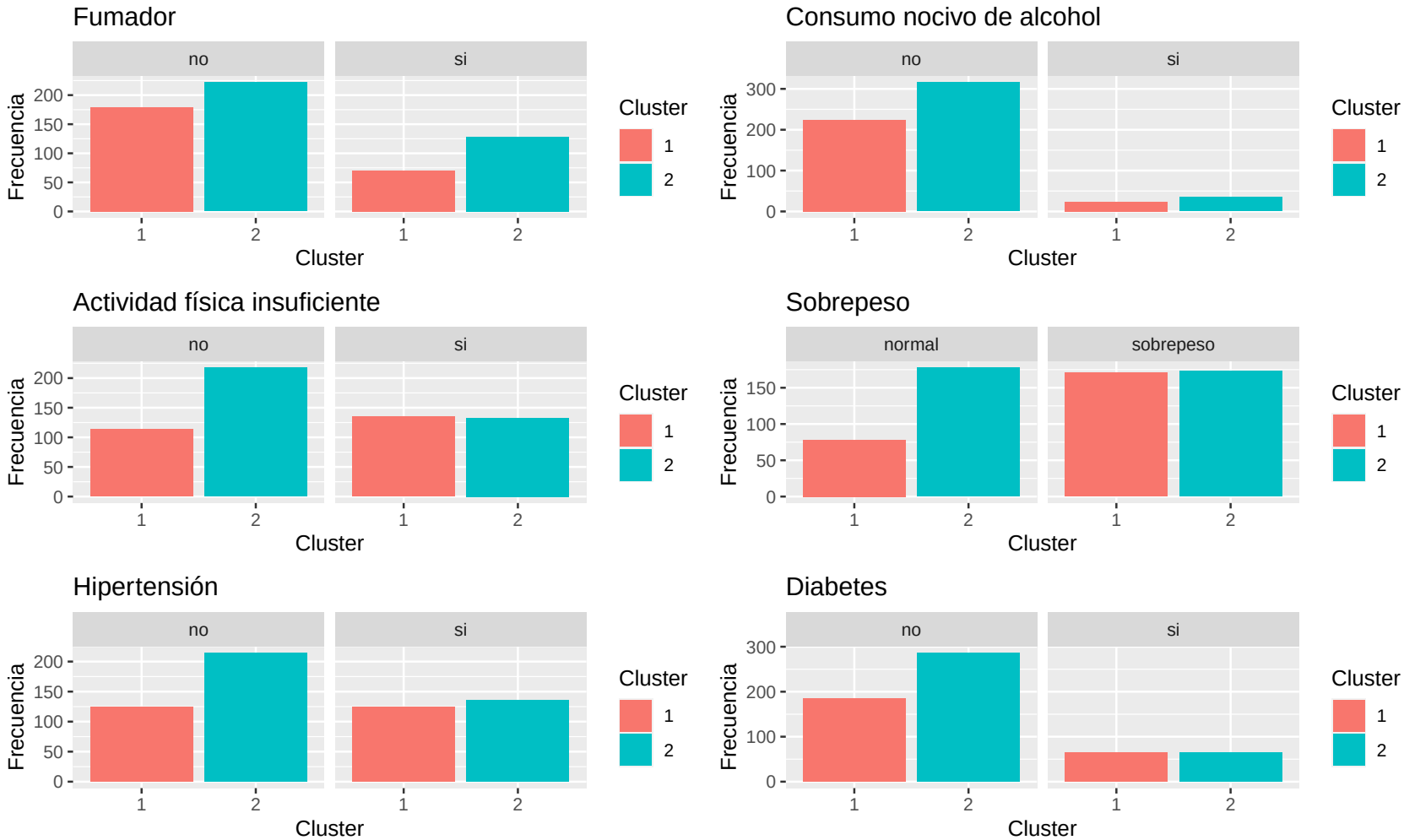






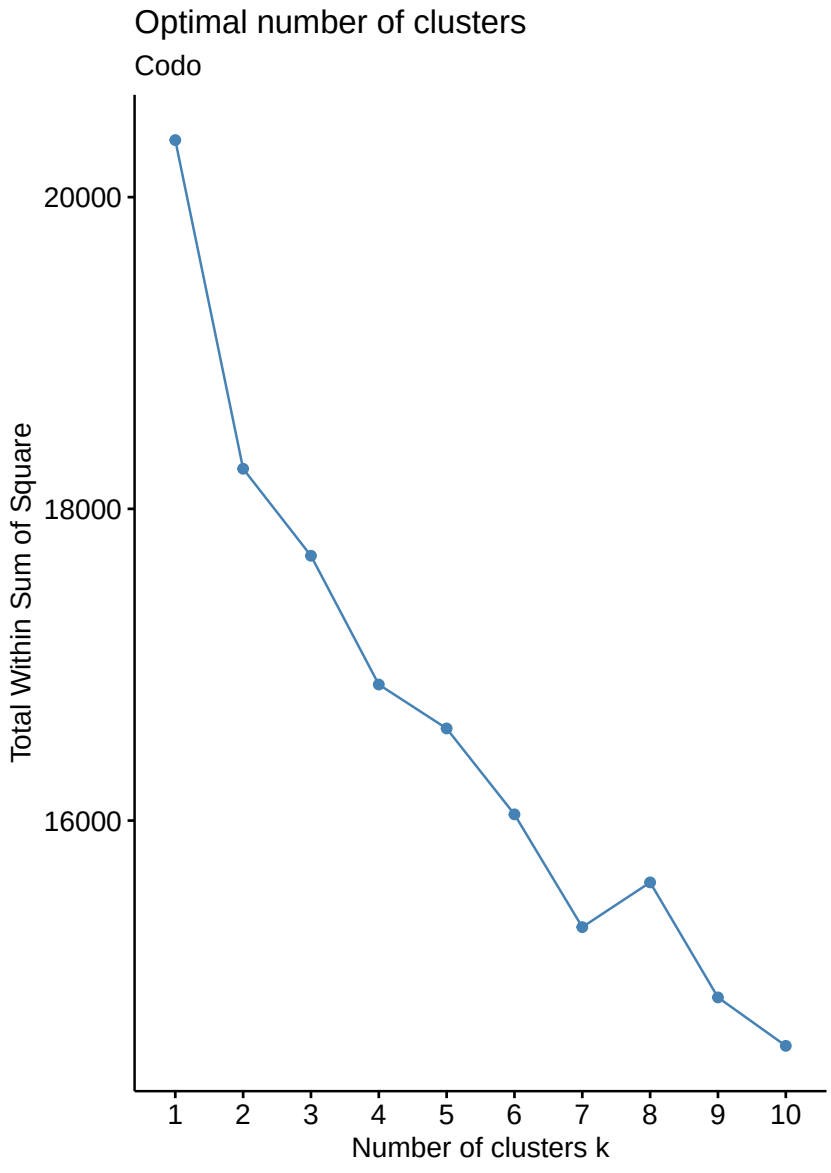
Cluster Jerárquico

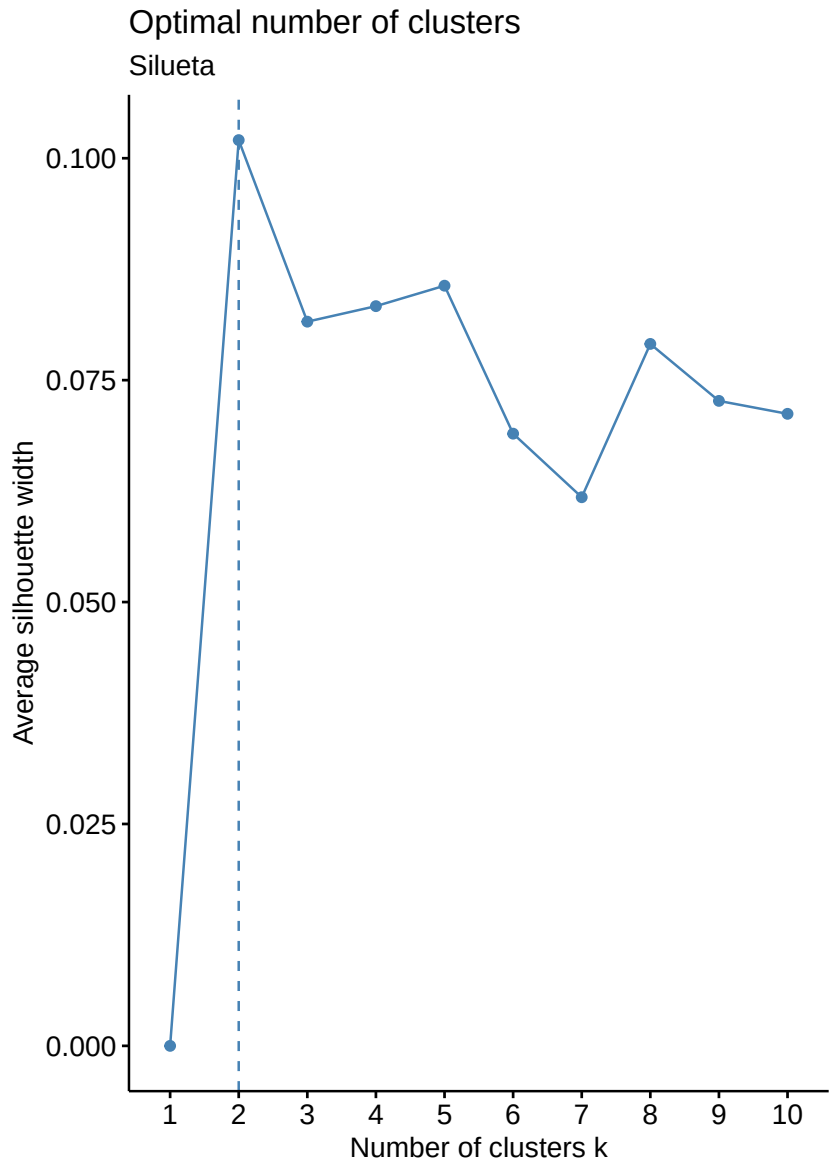
Variables comportamentales y ENT



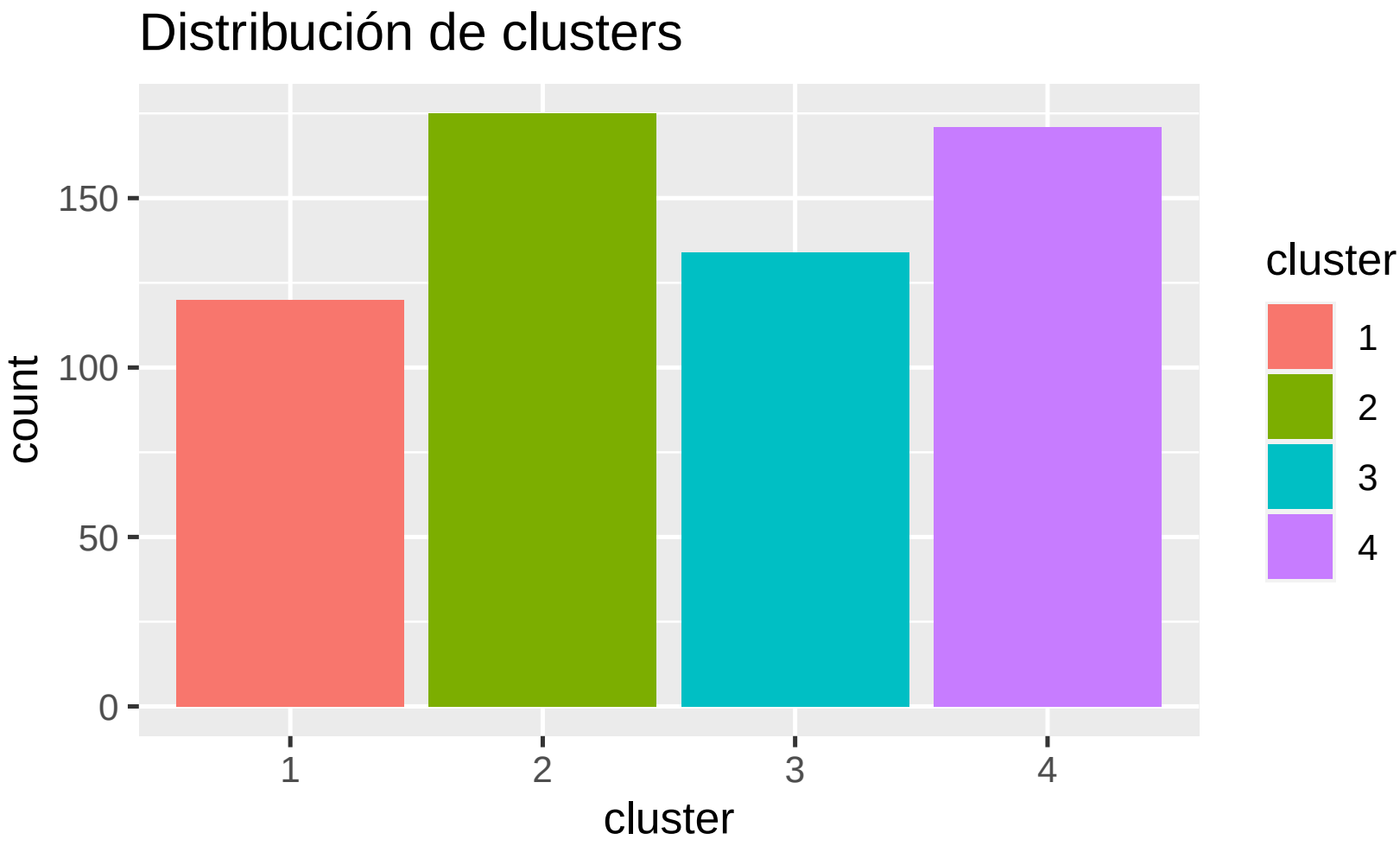
El cluster 2 presenta la mayor proporción en la categoría "no" entre aquellas variables comportamentales y de enfermedades.

Cluster No Jerárquico



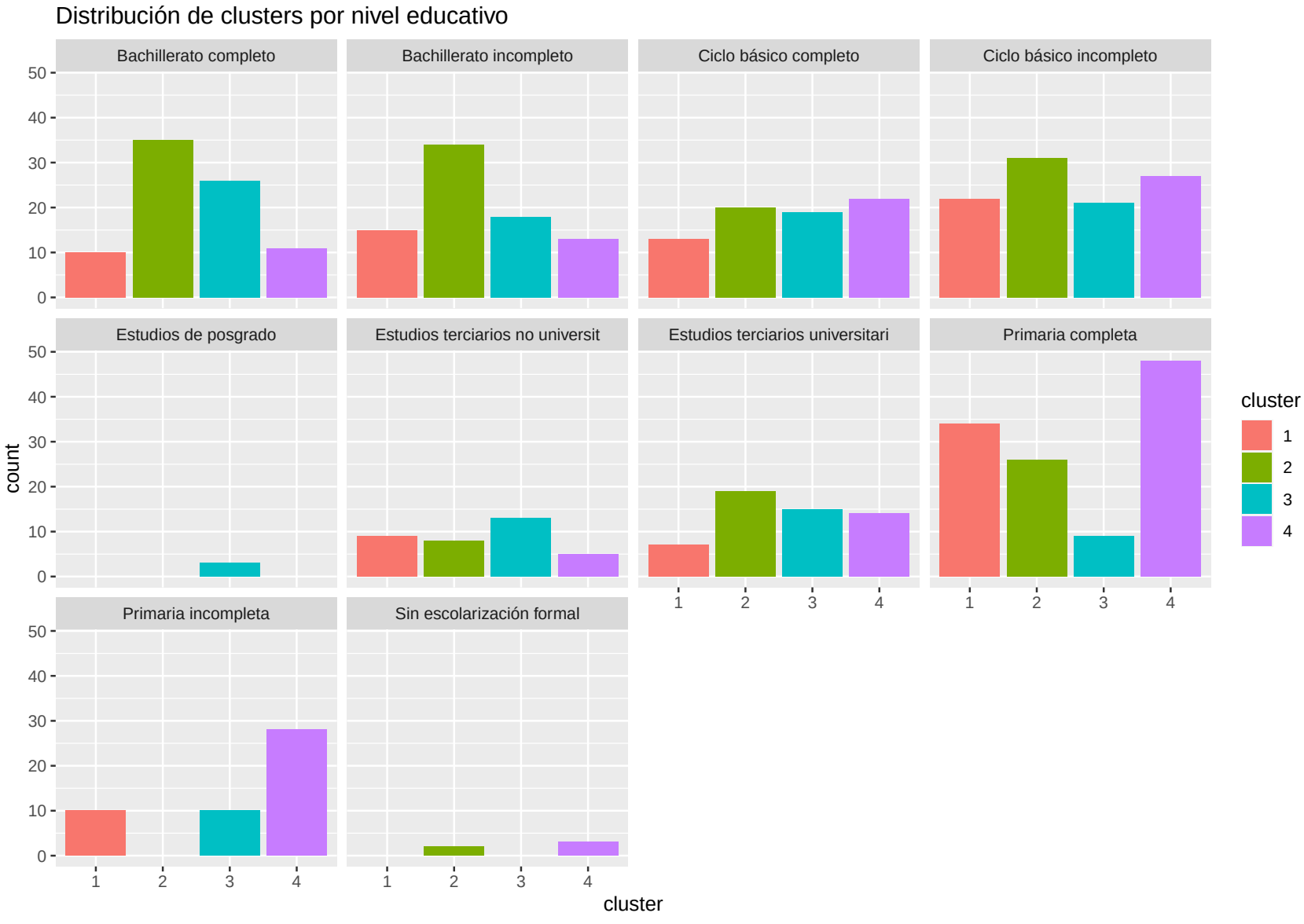


Cluster No Jerárquico

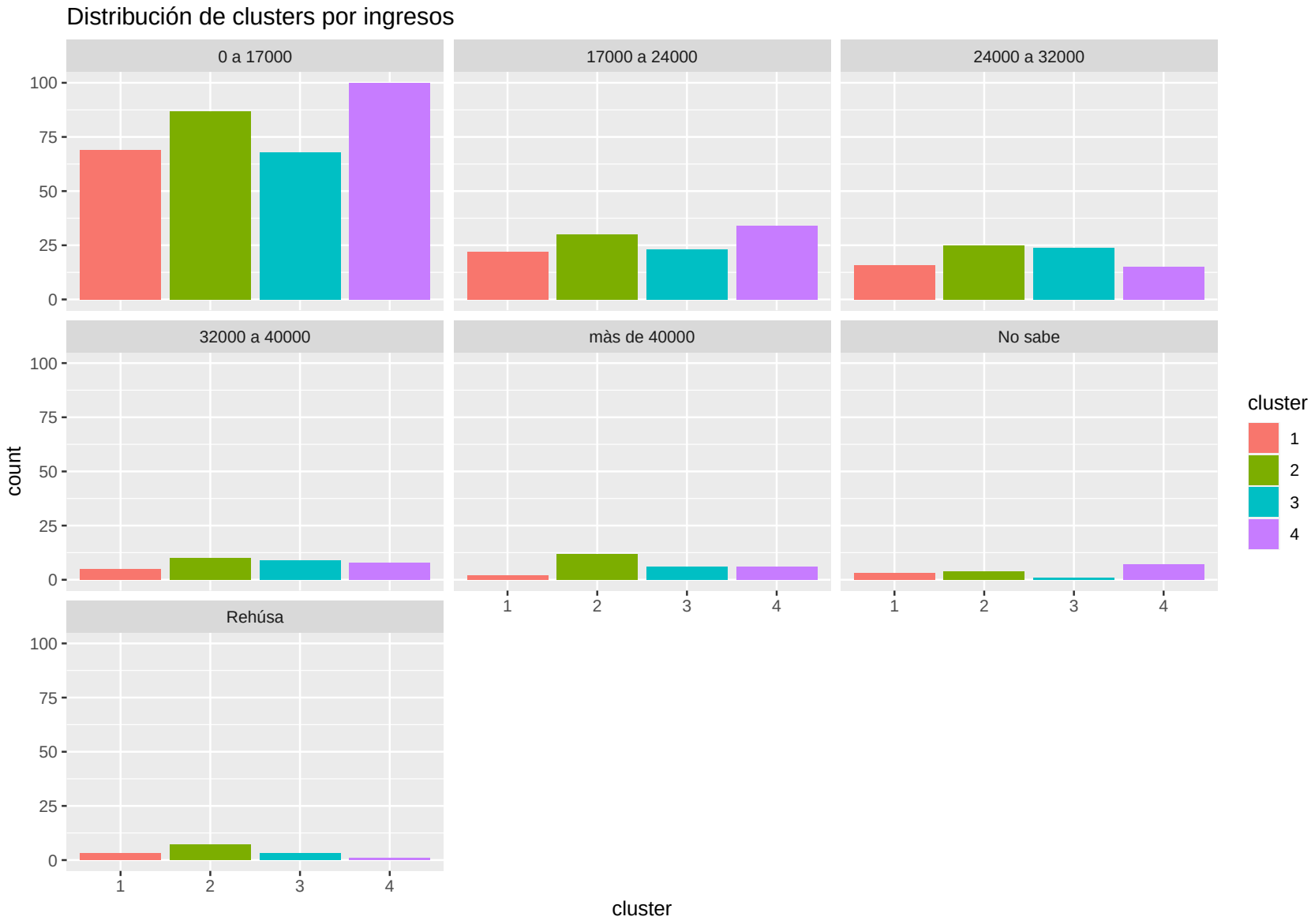


La distribución de personas entre los clusters es bastante homogénea. No hay un cluster que se destaque por tener una cantidad de personas mucho mayor que los otros.

Cluster No Jerárquico

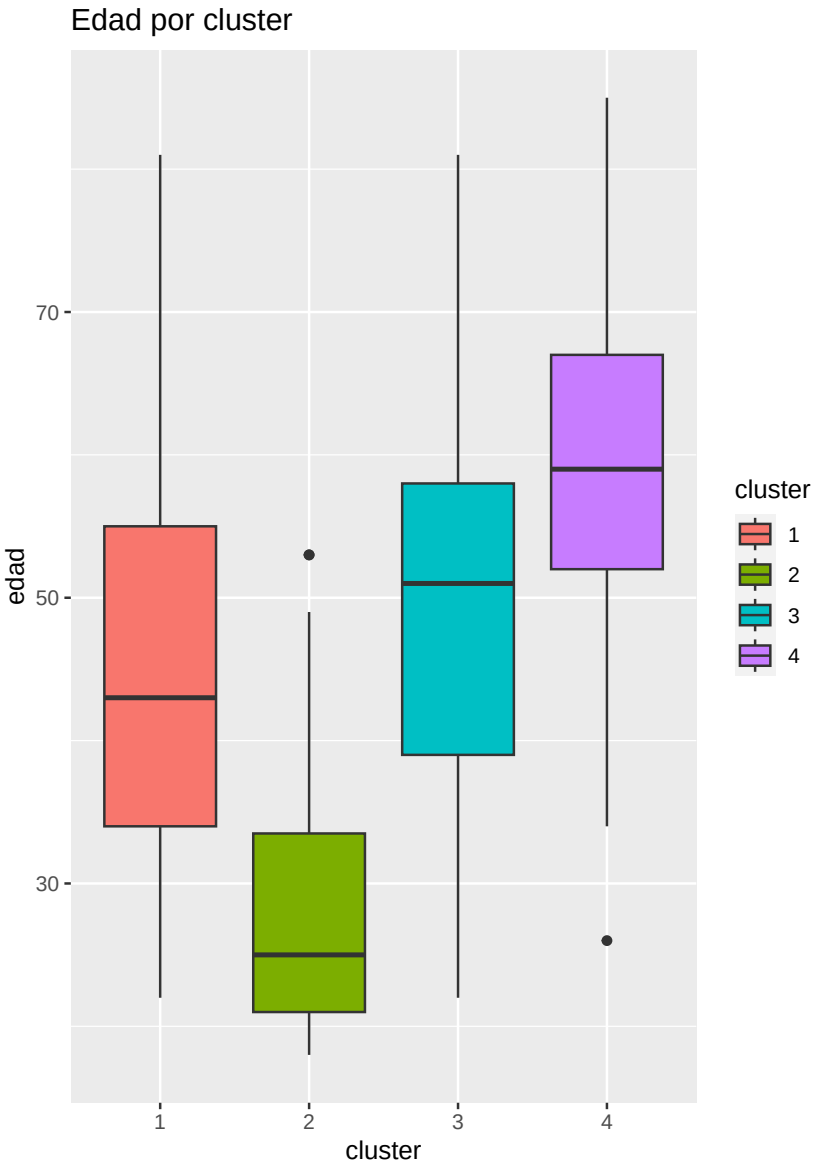


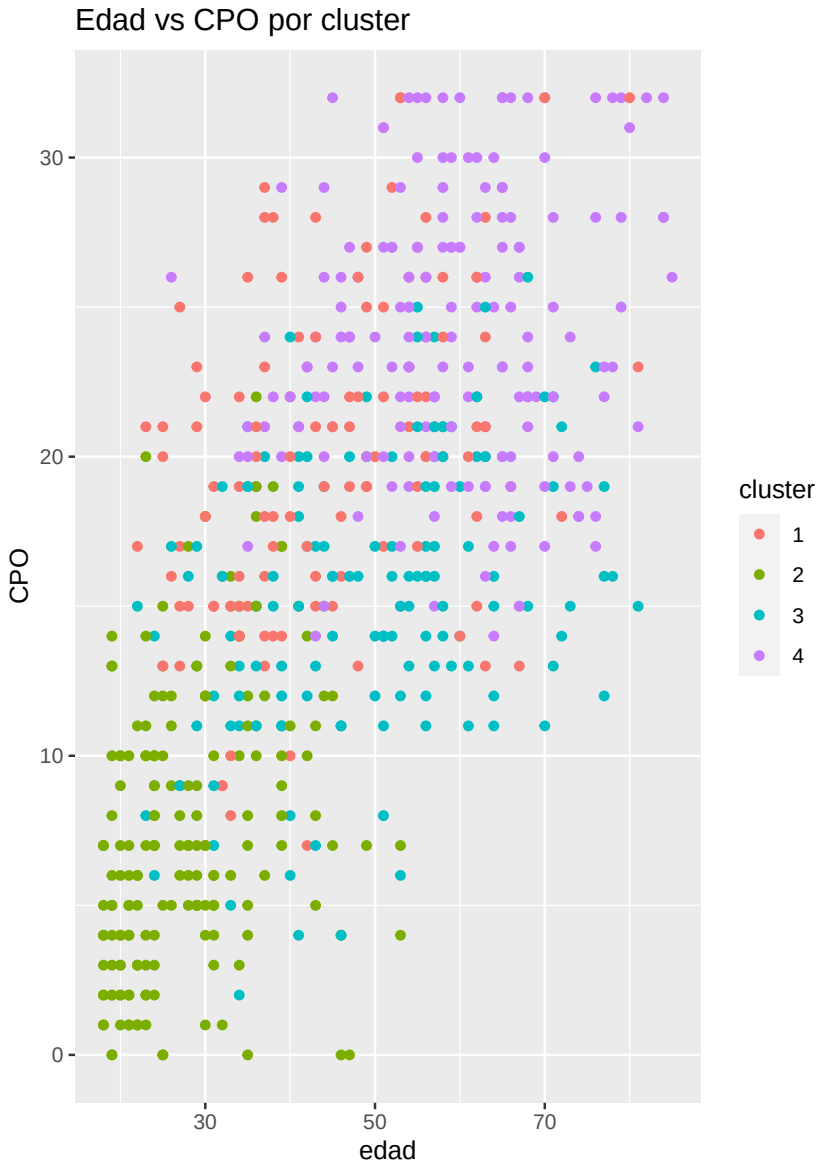
Cluster No Jerárquico



Los ingresos no parecen influir en la variabilidad entre los clusters.

Cluster No Jerárquico

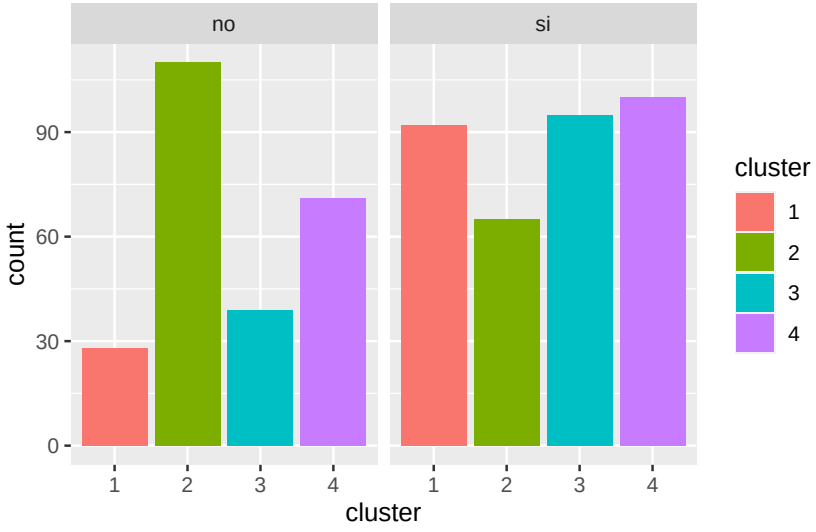




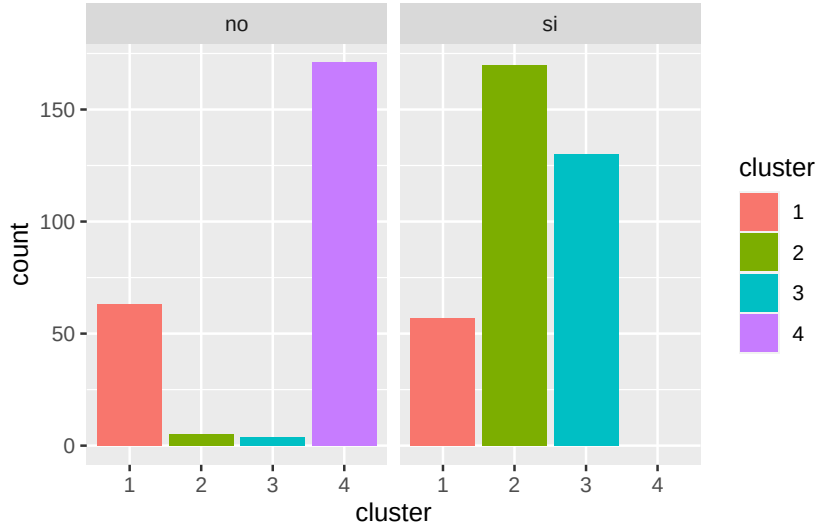
Cluster No Jerárquico

Análisis problemas odontológicos

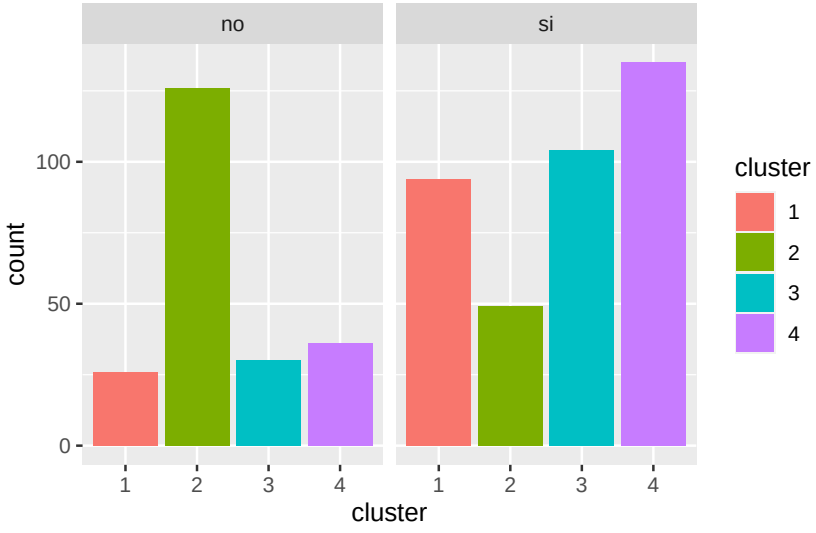
Distribución de clusters por prevalencia de bolsa



Distribución de clusters por perdida dentaria

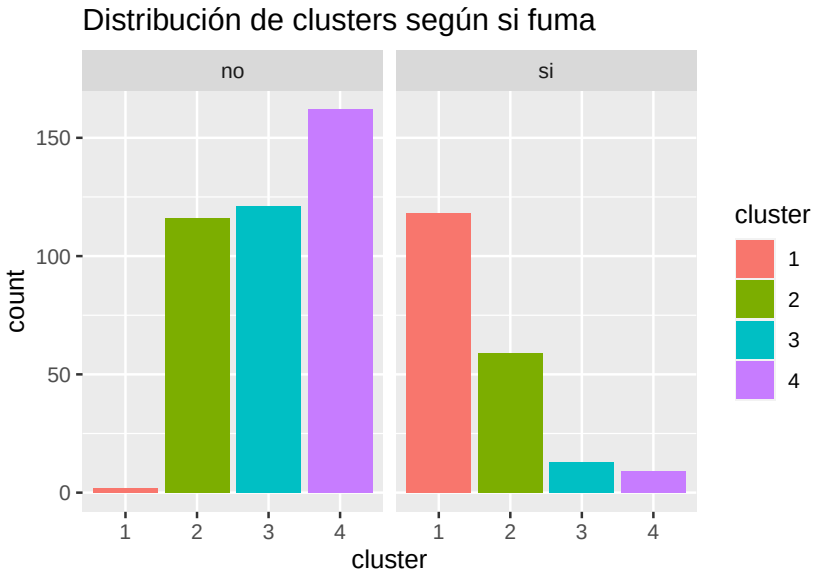


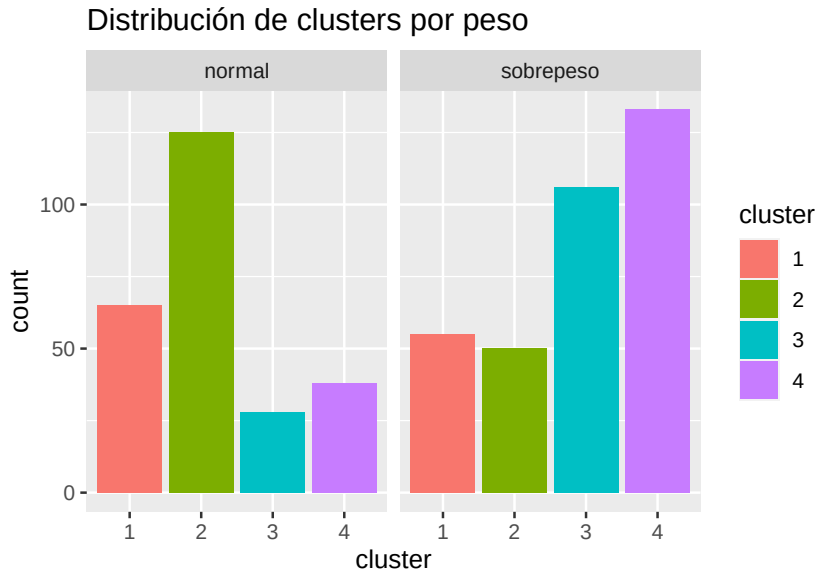
Distribución de clusters por prevalencia de PIP



Cluster No Jerárquico

Análisis según factores de incidencia en la salud bucal





Resumen cluster no jerárquico

Cluster 2

- **Edad:** Las personas más jóvenes.
- **CPO:** Bajo.
- **Pérdida Dentaria:** Presenta pérdida dentaria.
- **Prevalencia de PIP:** Mayormente no tiene prevalencia de PIP.
- **Fumar:** La mayoría no fuma.

Cluster 4

- **Edad:** Las edades más altas.
- **CPO:** Los CPO más altos.
- **Pérdida Dentaria:** No tienen pérdida dentaria.
- **Prevalencia de PIP:** Tienen prevalencia de PIP.
- **Fumar:** La mayoría no fuma.

Resumen cluster no jerárquico

Cluster 1

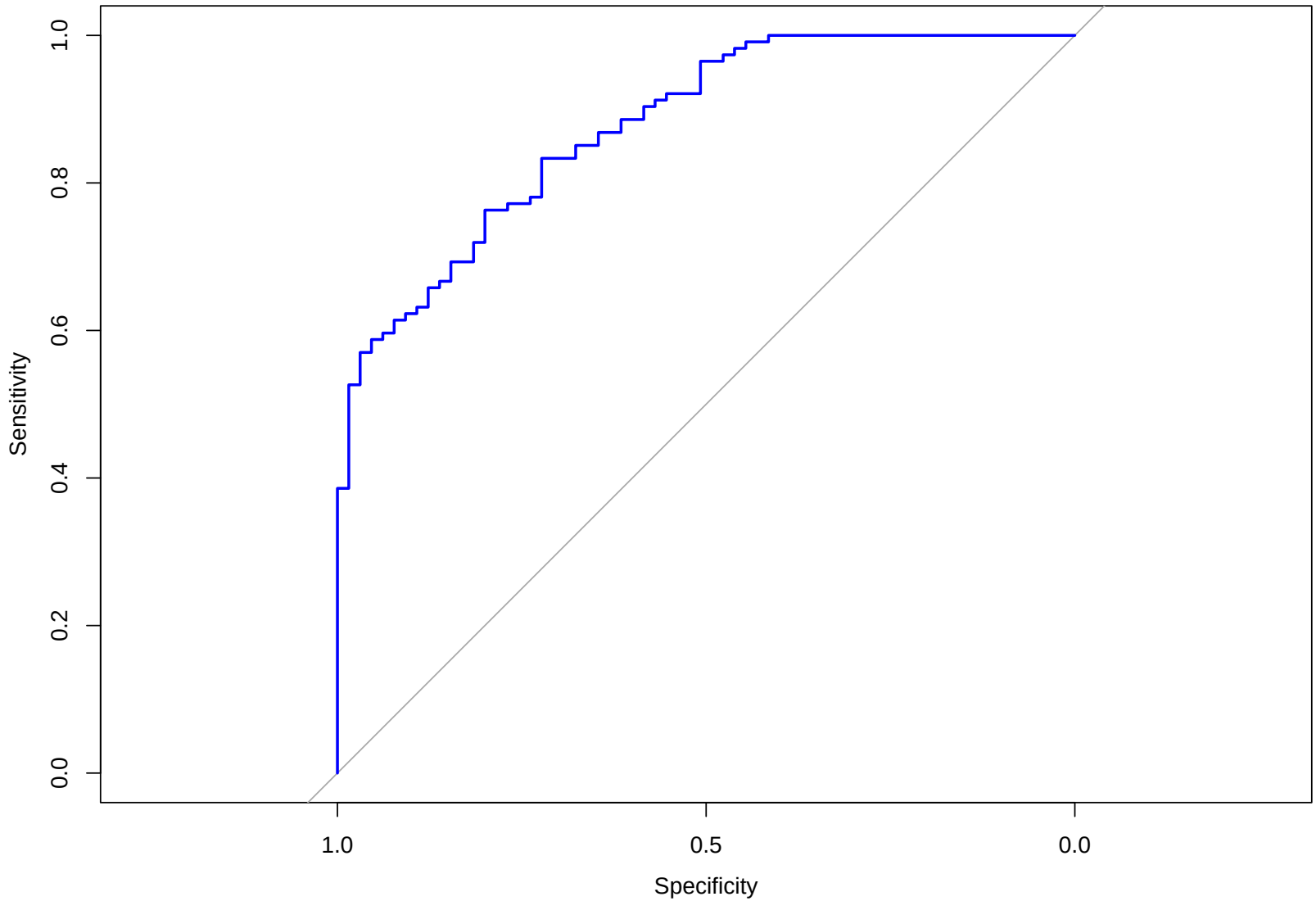
- **Edad:** Similar al cluster 3.
- **CPO:** Un poco mayor que el cluster 3.
- **Pérdida Dentaria:** No está bien explicado por la pérdida dentaria.
- **Prevalencia de PIP:** Prevalencia de PIP un poco alta.
- **Fumar:** La mayoría fuma.

Cluster 3

- **Edad:** Similar al cluster 1.
- **CPO:** Un poco menor que el cluster 1.
- **Pérdida Dentaria:** Presenta pérdida dentaria.
- **Prevalencia de PIP:** Prevalencia de PIP un poco alta.
- **Fumar:** La mayoría no fuma.
- **Peso:** La mayoría tiene sobrepeso.

Análisis de discriminante logistico

Curva ROC modelo GLM con V3 (actividad física insuficiente), V7 (diabetes), V8 (prevalencia de bolsa), sexo, S, C, P y O.



Análisis de discriminante logistico

	no	si
no	52	27
si	13	87

Extra no paramétrico

Como agregado al trabajo se realizó un análisis de clasificación no paramétrico, utilizando el método de K vecinos más cercanos, para intentar predecir la variable V11.

Se van a usar las mismas variables que en el análisis anterior, para poder comparar los resultados. Que son V11 $\sim$ V3 + V8 + sexo + edad + S + C + P + O

Extra no paramétrico

k	Accuracy	Kappa
5	0.76	0.46
7	0.76	0.48
9	0.77	0.49
11	0.77	0.50
13	0.78	0.51
15	0.78	0.51
17	0.78	0.51
19	0.79	0.52
21	0.79	0.52
23	0.78	0.51

Extra no paramétrico

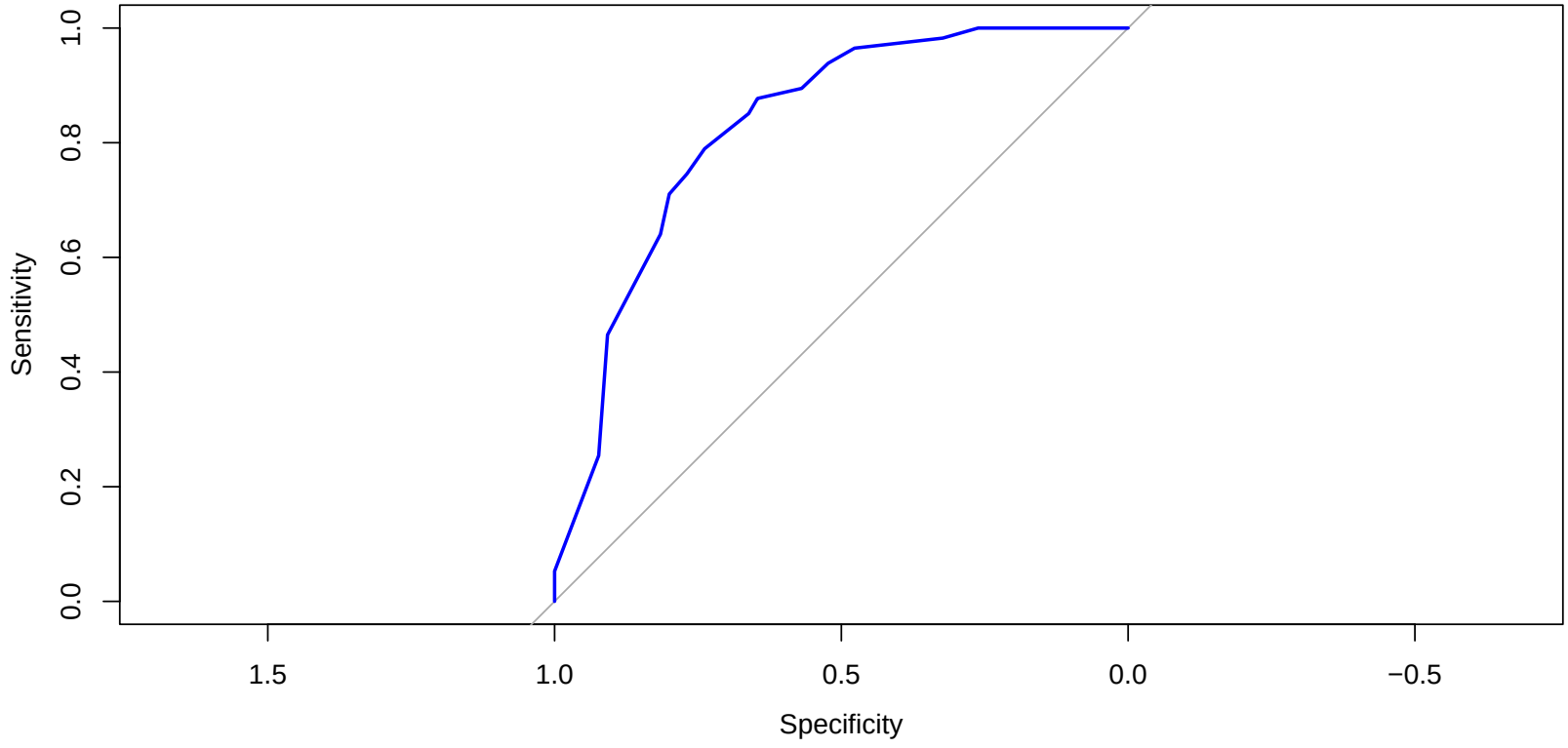
	no	si
no	42	14
si	23	100

Extra no paramétrico

	Métrica
Accuracy	0.793
Kappa	0.539
AccuracyLower	0.727
AccuracyUpper	0.850
AccuracyNull	0.637
AccuracyPValue	0.000
McnemarPValue	0.188

Este modelo en general funciona mejor que el logístico. Predice de mejor forma los casos en los que hay prevalencia de PIP a costa de tener un error mayor en los casos en los que no hay prevalencia de PIP (clasificandolo como si en prevalencia de PIP).

Curva ROC



En este caso el punto de corte óptimo es de 0.605. Que es un poco menor que en el modelo de regresión logística. Lo cual tiene sentido ya que este modelo tiende a clasificar más casos como “si” y la mayoría de los errores provienen de clasificar como “si” a casos que en realidad son “no”.

Conclusiones finales

Clústeres jerárquicos:

- No fueron útiles para separar la población en grupos que se diferencien en las variables.

Clústeres no jerárquicos:

- La edad es un factor importante para separar a la población en grupos
- Se identifican dos grupos que a priori parecían similares; sin embargo, al analizar las variables, se observa que quedan diferenciados por variables de riesgo. Uno de estos clústeres está compuesto por fumadores y el otro por personas con sobrepeso.

Análisis discriminante logístico y K vecinos más cercanos:

- Permitieron predecir la prevalencia de PIP en base a las variables de la muestra, siendo el modelo de K vecinos más cercanos el que mejor se ajusta a los datos, aunque ambos modelos funcionan bien.