Entrega Final - Multivariado

Lucca Frachelle, Joaquín Silva, Cecilia Waksman

Resumen Ejecutivo

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. Este análisis se centra mayormente en el uso de variables binarias para identificar la presencia de enfermedades no transmisibles (ENT) y otros factores de riesgo asociados con la salud bucal, tratados aquí como comorbilidades. Los atributos estudiados se agrupan en tres bloques principales: comportamentales, ENT generales, y patologías odontológicas específicas, además se cuenta con variables demográficas diversas (edad, ingreso, sexo, nivel educativo). Entre los comportamentales se incluyen el consumo diario de tabaco, el consumo nocivo de alcohol y la insuficiencia de actividad física. Los ENT se refieren a condiciones como sobrepeso/obesidad, razón de cintura-cadera, hipertensión y diabetes. Finalmente, las condiciones odontológicas incluyen la prevalencia de bolsas periodontales, pérdida dentaria, caries y prevalencia de patología periapical infecciosa (PIP).

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

Además contamos con las variables 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.

El estudio empleó técnicas de muestreo sistemático para seleccionar una muestra representativa, ajustada para medir prevalencias de hasta un 25% con un margen de error de 0.05 y un nivel de confianza del 95%. Se aplicaron cuestionarios sociodemográficos y exámenes bucodentales completos, evaluando la salud dental y de la mucosa, además de medidas antropométricas y análisis de presión arterial y glucemia.

Introducción

En el presente trabajo, aprovechamos un conjunto de datos derivados de individuos que solicitaron atención en la Facultad de Odontología de la Universidad de la República, Uruguay, entre 2015 y 2016, para realizar un estudio exploratorio que indaga cómo los factores de riesgo comportamentales y las enfermedades no transmisibles (ENT) influyen en las patologías odontológicas. A través del uso de métodos de clasificación tanto supervisada, Análisis de Discriminantes, como no supervisada, Clustering, buscamos comprender las interacciones y posibles correlaciones entre estos distintos aspectos de la salud.

Este trabajo se centra en analizar las asociaciones entre variables, con el fin de identificar patrones que sugieran cómo los hábitos de vida y las condiciones de salud impactan en la salud bucal.

Así, nuestro análisis no solo ayuda a esclarecer la estructura de los datos y la relación entre diversas condiciones de salud, sino que también facilita el entendimiento de los factores de riesgo que afectan en la salud bucal.

Marco metodológico

Clustering

Este es un método de clasificación no supervisada que, aplicado a una base de datos, consiste en dividir la misma en subgrupo/clusters de observaciones mediante un orden determinado. Dicho orden toma en cuenta cierta distancia (a definir, depende del tipo de clustering a usar y de la estructura de los datos) entre subgrupo y/o observaciones, según los valores que tomen las variables de dicha base.

Clusters jerárquicos:

Los clusters de tipo jerárquico se entienden como particiones encajadas que deben seguir cierta jerarquía basada para la aglomeración o separación de los subconjuntos. Gracias a ello

los mismos permiten seguir claramente la historia de construcción de cada grupo. Existen dos tipos de clusters jerárquicos:

- **Divisivos**: Comienza tomando como grupo a todo el conjunto de individuos, los cuales irá separando en subgrupo en las siguientes etapas hasta obtener tantos clusters como observaciones.
- **Agregativo**: Cada individuo comienza siendo un cluster unitario que se une con otros en las etapas siguientes hasta formarse un único cluster que cuenta con todos los individuos.

En nuestro análisis usaremos únicamente clusters de tipo agregativo.

En estos, la distancia a utilizar es seleccionada según las características de las variables en cuestión, así como también la relación entre las mismas. En nuestro caso, como contamos con variables tanto numéricas como categóricas, es necesario usar la similaridad de Gower, a través de la cual se calcula una distancia comparable, independientemente del tipo de variable.

La **similaridad de Gower** para dos individuos i, j se calcula como:

$$\delta_{ij} = \frac{1}{T} \sum_{t=1}^{T} \delta_{ijt}$$

donde δ_{ijt} es la similaridad entre ambos individuos en la variable t y T es la cantidad de variables.

El método de clustering jerárquico que se usará será el **método de Ward**. El cual se basa en minimizar la varianza intra-grupos resultante de la agregación de dos grupos. Dicha varianza se calcula en función de la distancia de los individuos al centro de gravedad de su grupo en cada etapa.

Clusters no jerárquicos:

Para el análisis usaremos unicamente los clusters no jerárquicos por el método de **K-means**. Este consiste en seleccionas K centros provisorios, los cuales inducen una partición del conjunto de individuos en k subgrupo. Luego, se vuelven a seleccionar nuevos centros de gravedad, en base a las distancias de los individuos del grupo a los centros. Esta nueva partición cuenta con una reducción de la variabilidad intra-grupo. Este proceso se repite siempre y cuando las particiones en etapas sucesivas sean diferentes o hasta que se alcanza un número de iteraciones determinado.

Análisis Discriminante

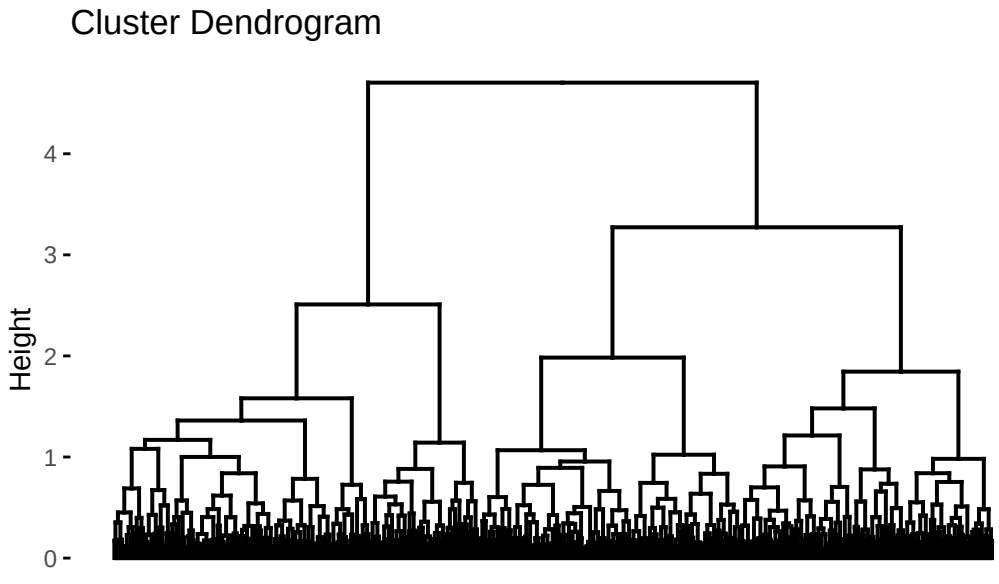
Dado que buscamos predecir la variable V11, la cual corresponde a prevalencia de PIP, la cual es binaria debemos de utilizar un discriminante logístico, este consiste en realizar una regresión logística, tomando como variable a predecir V11, ya que esta nos dará predicciones en el intervalo $[0, 1]$. Luego se busca el punto de corte óptimo, es decir tomamos un valor tal

que toda predicción sobre este se toma como 1 y por debajo como 0, tal que minimice el error de predicción.

Clasificasión no supervisada

Cluster Jerárquico

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.

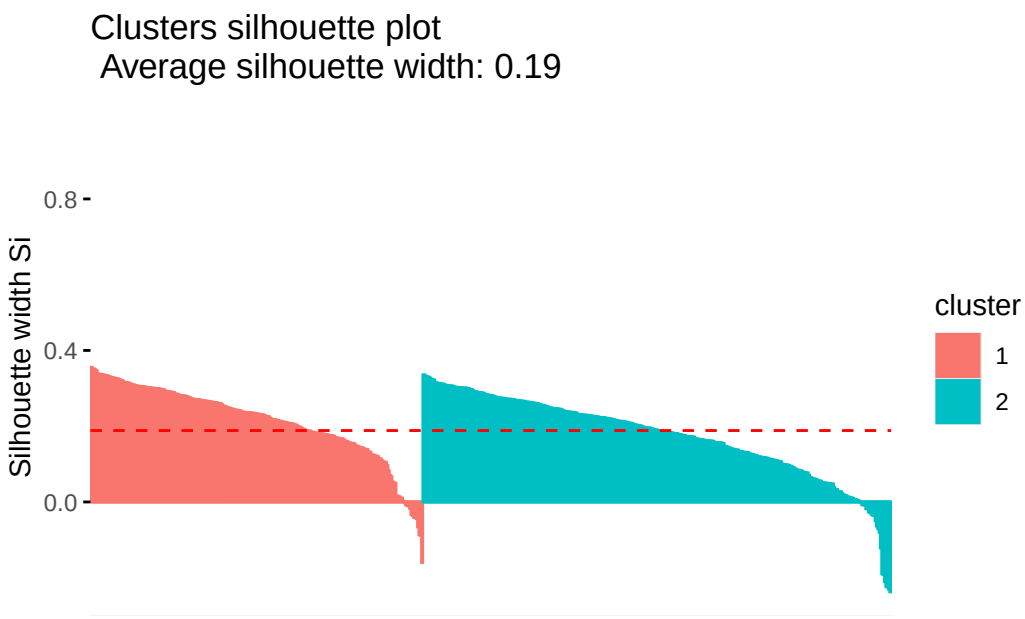


Para elegir la cantidad de clusters, se recurre a los índices: *Cindex*, *Silouhette*, *Mcclain*, *Dunn* y *Frey*.

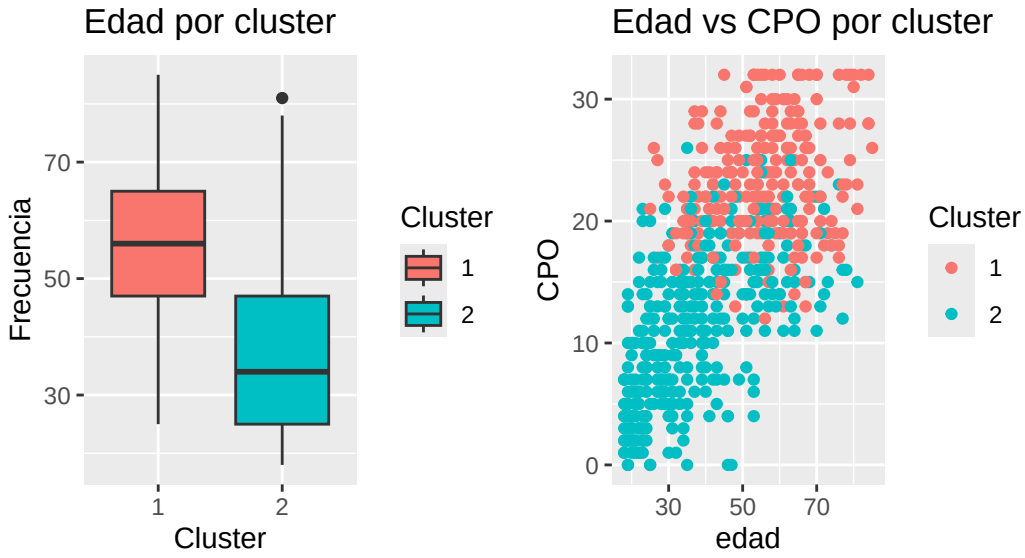
	Cindex	Silohuette	Mcclain	Dunn	Frey
Ward	15	2	2	2	1

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

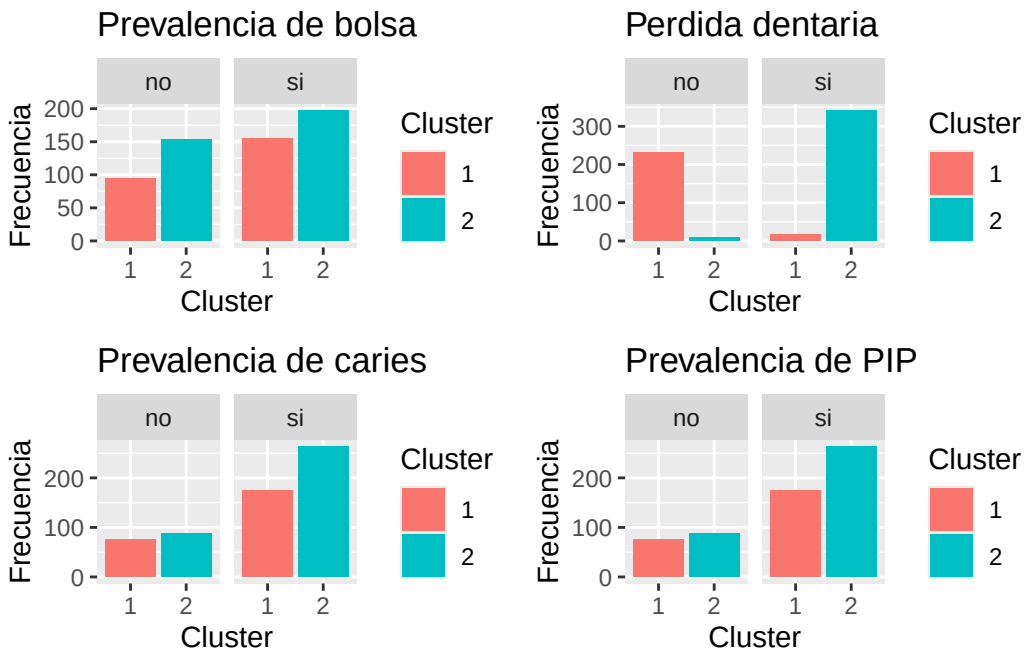
```
  cluster size ave.sil.width
1       1  249          0.21
2       2  351          0.17
```



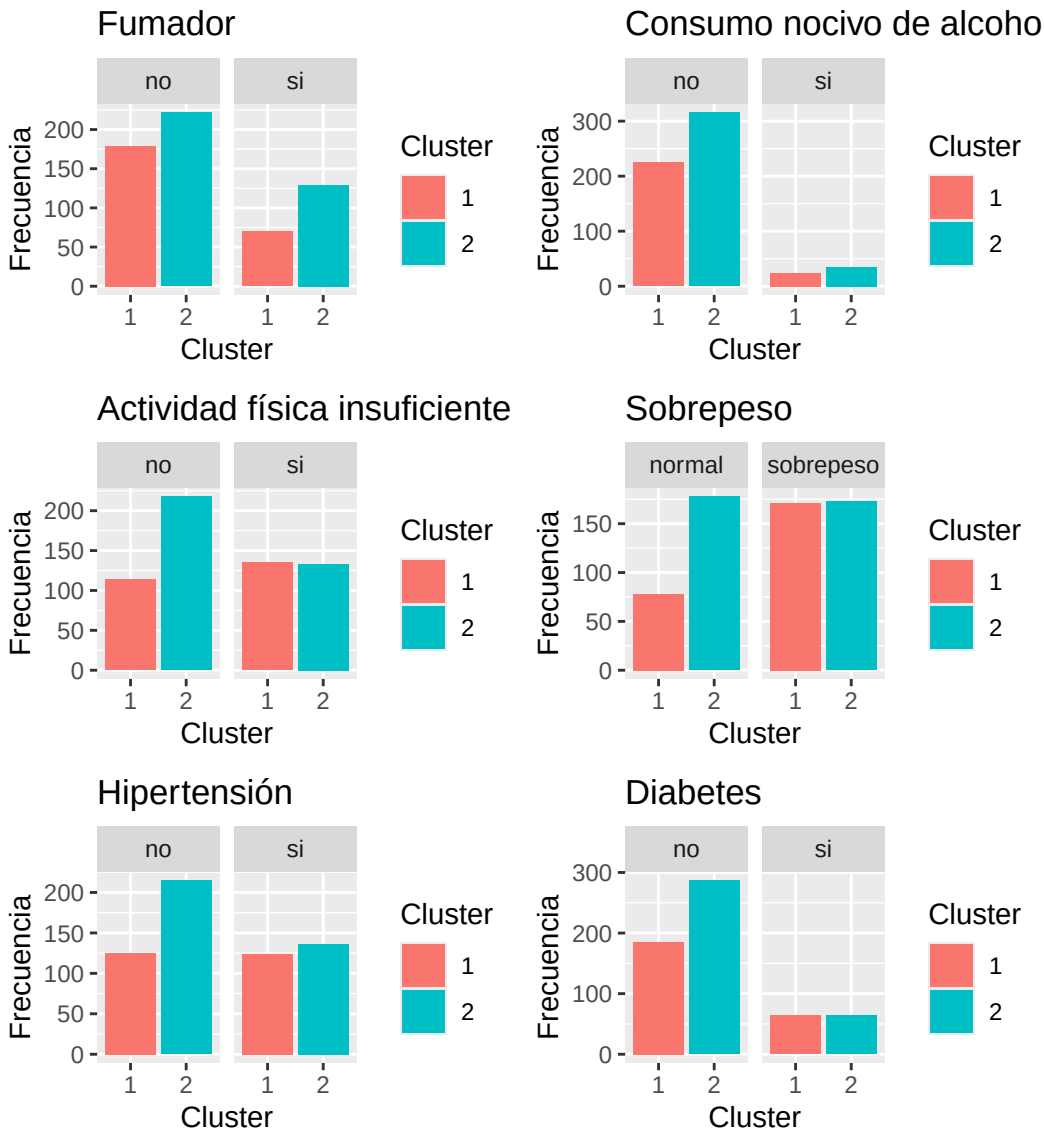
El ancho de silueta es bajo, por lo que podría haber una mejor forma de agrupar las observaciones. Las observaciones que se encuentran por debajo del 0 posiblemente se encuentran mal clasificadas.



Los clusters parecen estar divididos entre quienes son menores y mayores a 47 años. También se observa que aquéllos mayores a dicha edad, son los que presentan un CPO en general más alto.

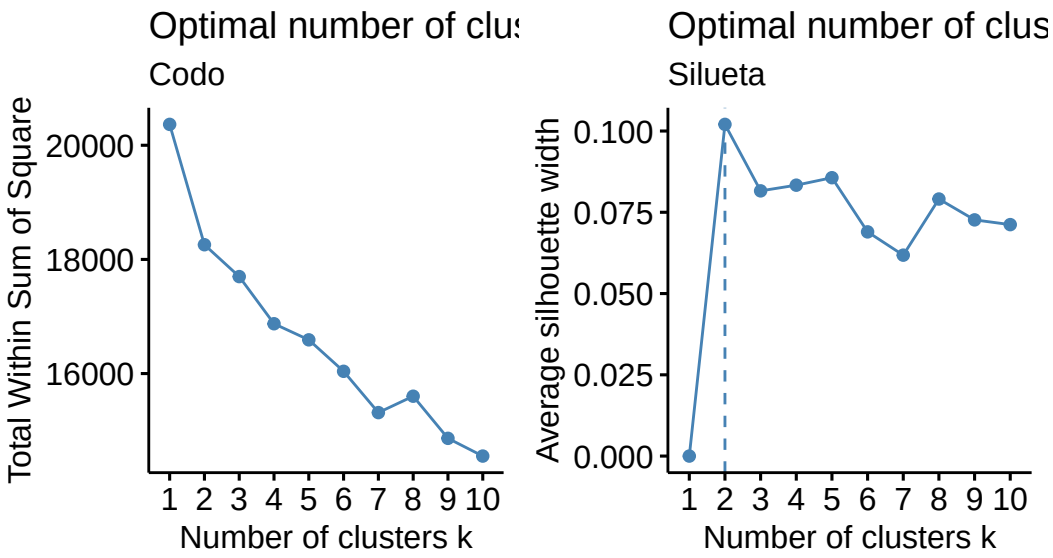


La diferencia más notoria se observa en la pérdida dentaria, donde el primer grupo, aquellas personas más mayores, en su mayoría no presenta pérdida dentaria, mientras que el segundo grupo, aquéllos más jóvenes, cuentan con una frecuencia alta de la misma variable. Las demás variables no parecen mostrar diferencias significativas entre las categorías según el grupo.



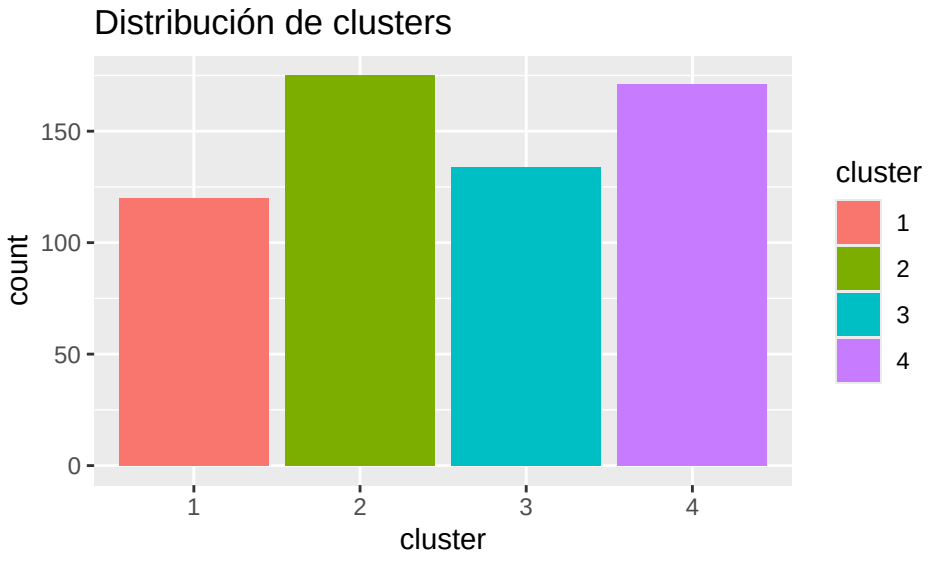
El cluster 2 es el que presenta la mayor proporción en la categoría "no" entre aquellas variables comportamentales y de enfermedades.

Cluster No Jerárquico

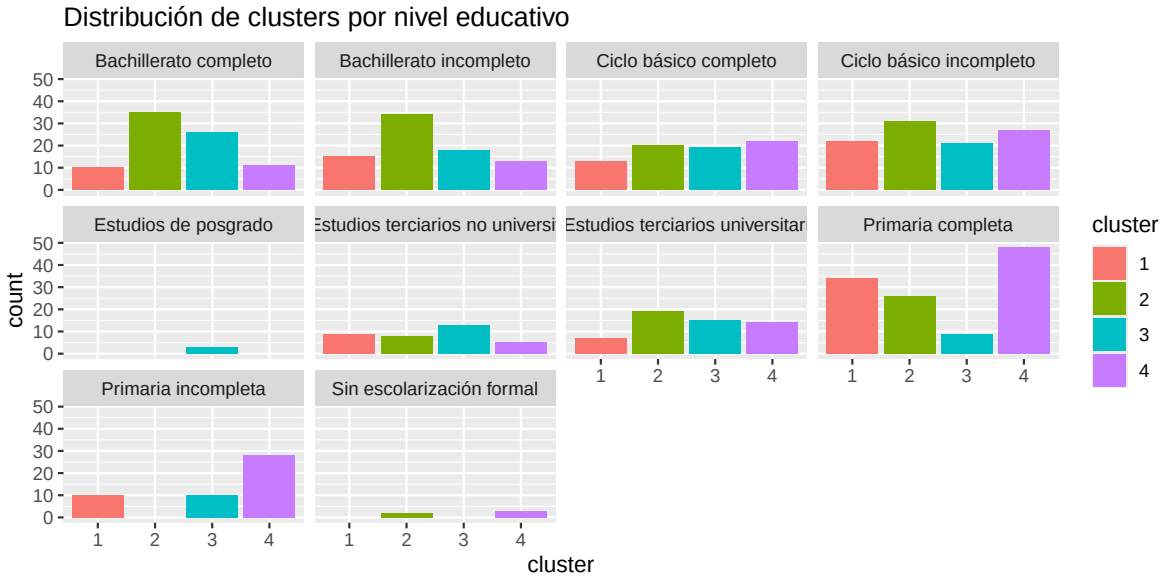


Dada la información obtenida de los gráficos de codo y silueta, se procede a realizar un análisis de cluster con 4 grupos. Ya que el gráfico de codo no presenta un punto de inflexión claro se opta por un punto medio entre silueta y codo.

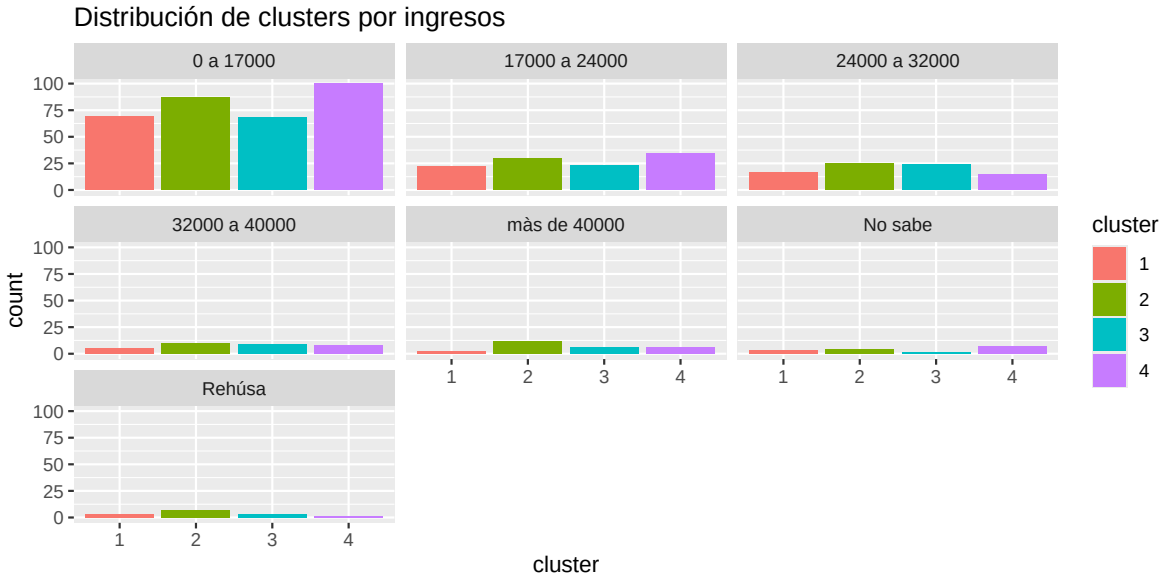
Análisis entre clusters



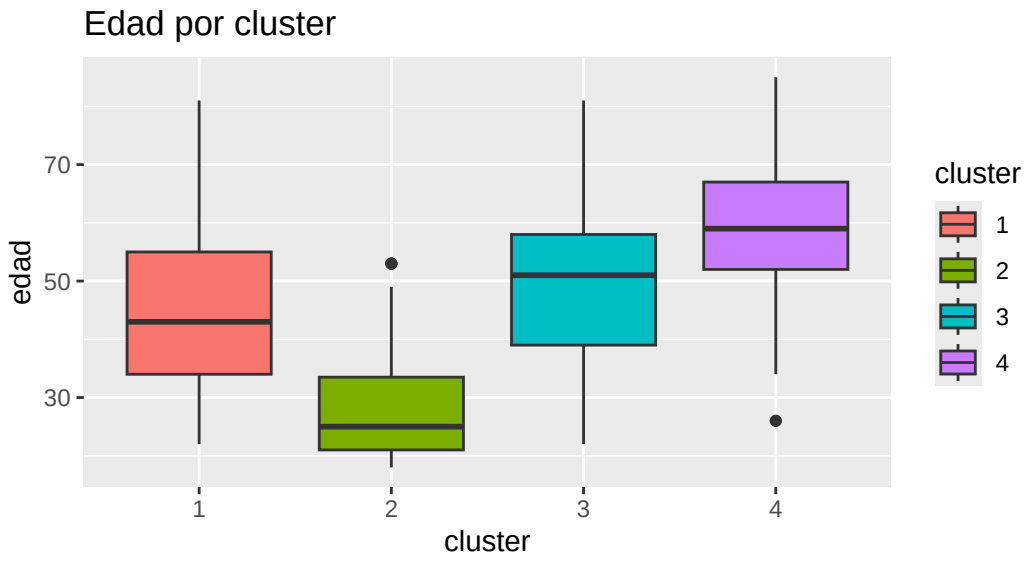
La distribución de personas entre los clusters es bastante homogénea. No hay un cluster que se destaque por tener una cantidad de personas mucho mayor que los otros.



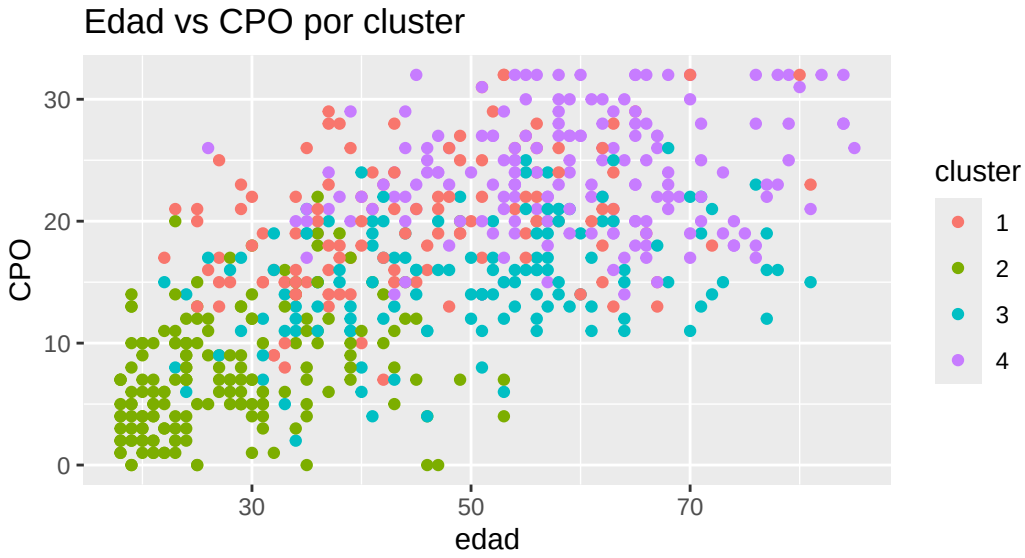
El cluster 4 es el que tiene la mayor cantidad de personas con nivel educativo bajo. La mayoría de las personas en este cluster tienen un nivel educativo que varía entre ciclo básico y primaria incompleta. Los demás clusters tienen una distribución más homogénea entre los distintos niveles educativos.



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

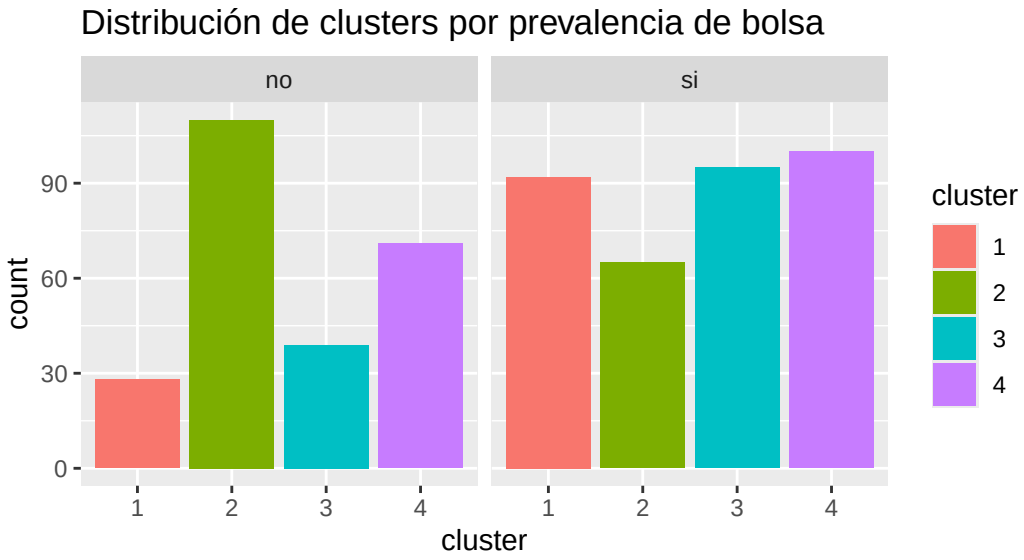


Se identifican dos clusters claramente diferenciados por edad: el cluster 2, con una mediana de 25 años, y el cluster 4, con una mediana de 60 años. En contraste, los clusters 1 y 3 tienen edades similares, con una mediana de alrededor de 50 años, y presentan la mayor variabilidad en sus distribuciones de edad

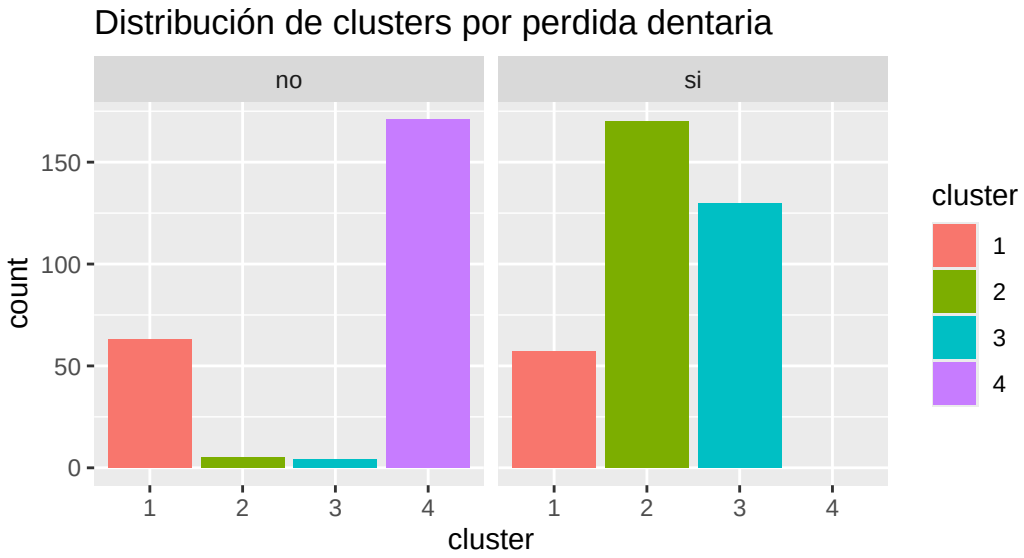


Se observa una correlación positivia entre la edad y el CPO. El cluster 4 en general tiene tanto el CPO más alto como la mayor edad. Por otro lado, se nota una diferencia entre el cluster 1 y el cluster 3 respecto al CPO. Aunque tienen edades similares, con una mediana de alrededor de 50 años, el cluster 3 presenta un CPO mayor. El cluster 1, en cambio, está por debajo tanto en CPO como en edad.

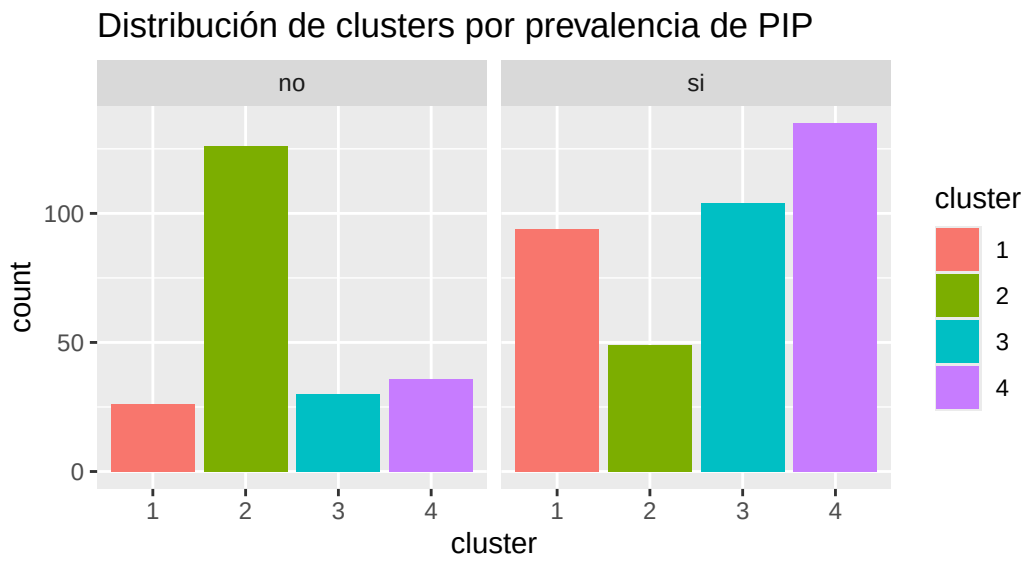
Análisis problemas odontológicos



No hay una gran diferencia entre los clusters en cuanto a la prevalencia de bolsa. Solo el cluster 2 pareciera contar con una mayor fecuencia en no prevalencia de bolsa, en comparación al resto.

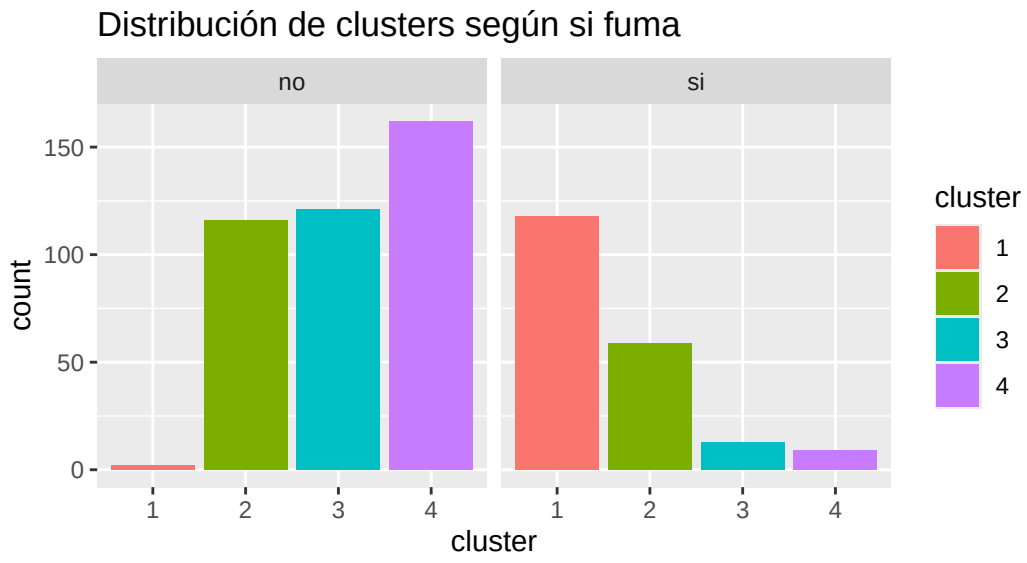


Los grupos parecen diferenciarse claramente según su frecuencia de pérdida dentaria. Por un lado, el cluster 4 está compuesto completamente por personas que no poseen pérdida dentaria, mientras que los clusters 2 y 3 pertenecen en su gran mayoría a personas que sí tienen pérdida dentaria. El cluster 1, sin embargo, no parece estar tan diferenciado según esta variable.

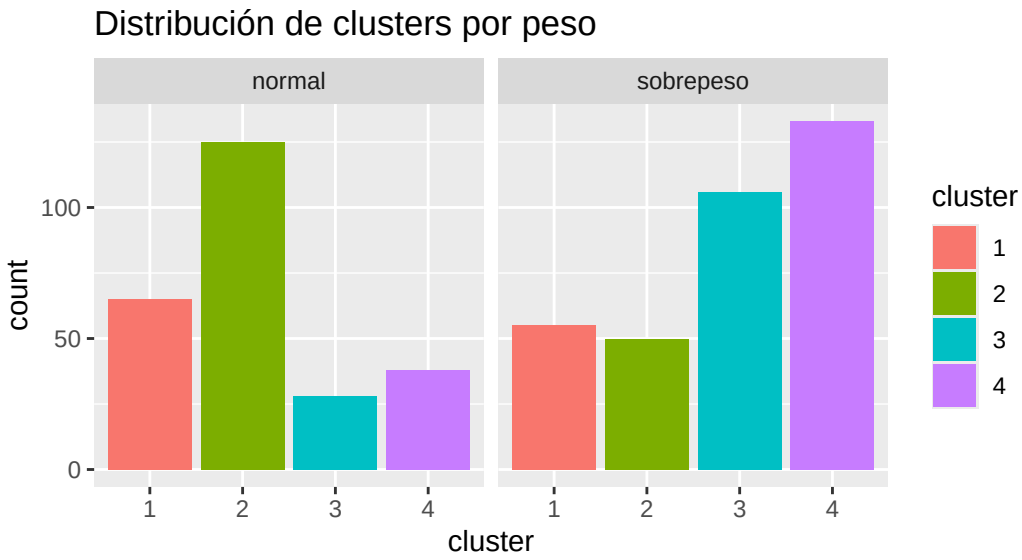


Se observa que el cluster más joven es el cluster 2, el cual tiene la menor prevalencia de PIP. En contraste, el cluster 4 tiene la mayor edad y la mayor prevalencia de PIP. Además, los clusters 1 y 3 tienen una prevalencia de PIP similar.

Análisis según factores de incidencia en la salud bucal



Aquí se observa una diferencia clara entre los cluster en los gráficos anteriores eran mas parecidos: por un lado, el cluster 1 consiste en su mayoría de fumadores y por otro lado, el cluster 3 está compuesto mayormente de no fumadores. El cluster 4, por su parte, también cuenta con una gran proporción de no fumadores.



En este caso se tiene que el cluster 3 y 4 son mayoritariamente personas con sobrepeso, mientras que el cluster 2 cuenta en su mayoría con personas con peso normal. El cluster 1 no se diferencia según el peso.

En resumen:

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.

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.

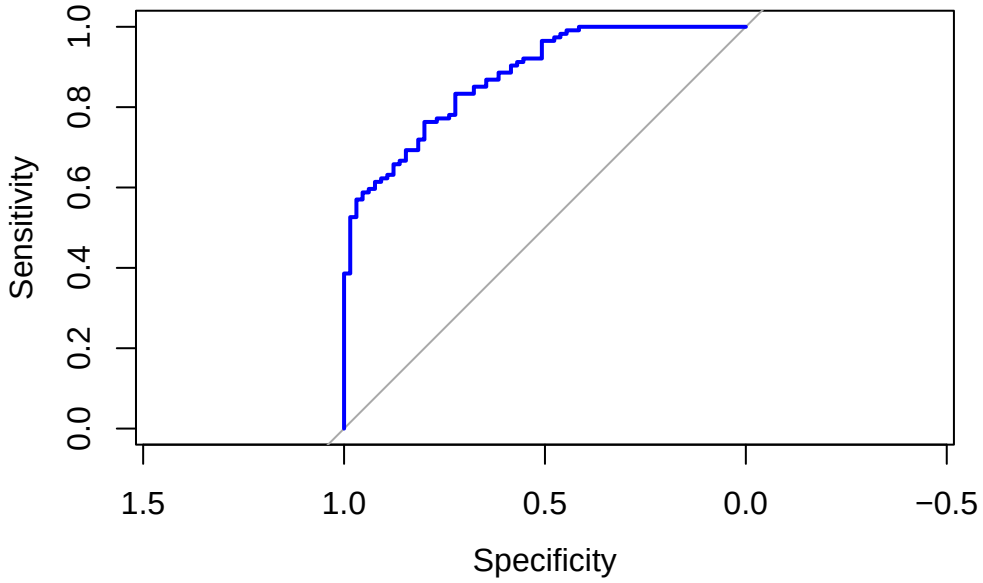
Análisis supervisado

Análisis de discriminante logistico

Teniendo en cuenta que la prevalencia de pip (V11) es un factor importante de riesgo odnologico se procede a realizar un análisis de discriminante logistico. Para intentar predecir la prevalencia de pip en base a las otras variables.

Utilizando el método stepwise ajustamos un modelo GLM que prediciera el valor de la variable V11, este se quedó con las variables V3 (actividad física insuficiente), V7 (diabetes) y V8 (prevalencia de bolsa) de las variables ENT, siendo la última la más importante con un coeficiente de casi 3 . Luego también mantuvo las variables sexo, edad y S, C, P y O.

Observando ahora la curva ROC del modelo obtenemos lo siguiente.



De la cual podemos obtener que el punto de corte con menor error en la parte de entrenamiento de la muestra es 0.647 , el cual es considerablemente mayor de $0,5$, esto nos puede dar a intuir que el modelo está dando valores en general cercanos a 1.

Si observamos la matriz de confusión para la muestra de testeo tenemos lo siguiente

	no	si
no	52	27
si	13	87

Podemos ver que en general el modelo tiende a sobreestimar la cantidad de individuos con prevalencia de PIP, teniendo un error de aproximadamente un 50% cuando estos no cuentan con PIP.

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.

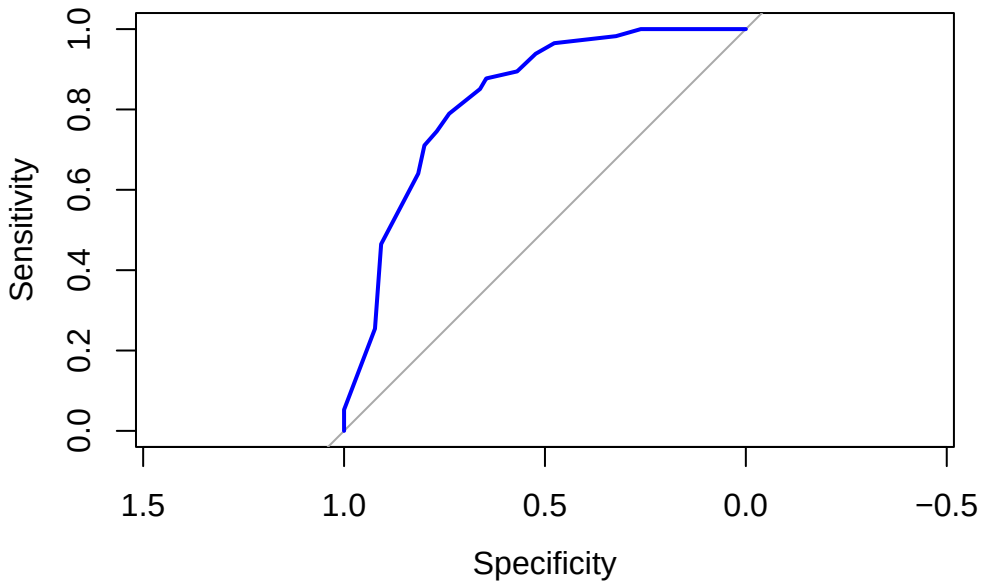
Se van a usar las mismas variables que en el análisis anterior, para poder comparar los resultados. Que son V11 ~ V3 + V8 + sexo + edad + S + C + P + O

	no	si
no	42	14
si	23	100

	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

Se observa que los clústeres jerárquicos no fueron útiles para separar la población en grupos que se diferencien en las variables.

En los clústeres no jerárquicos se reveló que la edad es un factor importante para separar a la población en grupos, siendo los más jóvenes los que presentan pérdida dentaria y prevalencia de PIP más baja. Mientras que los más viejos presentan un CPO más alto y una prevalencia de PIP más alta. Por otro lado, se identificaron 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. A su vez, se pudo validar alguna de las conclusiones de la primera entrega, como la relación entre el CPO y la edad.

Por otro lado, el análisis discriminante logístico y el de 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.