FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2021/2022

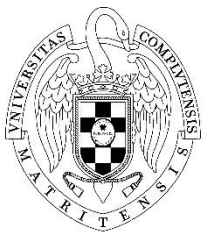
Trabajo de Fin de Máster

TITULO: Identificación de los fármacos asociados a Xerostomía e Hiposalivación, y su predicción con técnicas de Machine Learning

***Alumno*: Yuang Huang**

***Directores*: Antonio Sarasa Cabezuelo y Rosa María López-Pintor Muñoz**

Septiembre de 2022



UNIVERSIDAD COMPLUTENSE
MADRID

Resumen

Algunos fármacos que se utilizan para tratar diferentes enfermedades pueden provocan efectos secundarios en la cavidad oral. Entre dichos efectos adversos se encuentra la xerostomía y la hiposalivación. Existen estudios que han observado como ciertos fármacos se asocian a dichas condiciones. Pero no está claro si la combinación de diferentes fármacos tendría un efecto sinérgico. La finalidad de este trabajo es estudiar los fármacos que influyen en estas dos patologías en un grupo de pacientes polimedicados. Este trabajo utiliza los enfoques de aprendizaje automático para construir dos modelos de regresión logística y de árbol de decisión, para identificar los fármacos que afectan a estos dos problemas diferentes por separado, así como los fármacos que actúan sobre ambas afecciones. También, se proponen dos modelos avanzados de predicción, SVM y XGBoost para cuantificar la posibilidad de que un individuo sufra xerostomía e hiposalivación en función del medicamento que está tomando. Finalmente, regresión logística y árbol de decisión concluyen al menos un factor significativo que causa a la xerostomía e hiposalivación, y al menos dos fármacos protectores, en cada nivel de fármacos. Además, se encuentra que en los niveles dos a cuatro, XGBoost hace mejores predicciones para estimar el riesgo de xerostomía e hiposalivación en virtud de su alta AUC, precisión y sensibilidad. En el nivel cinco, SVM es la mejor opción para ambas patologías. El uso del clustering jerárquico complementa algunos de los fármacos no detectados por regresión logística y árbol de decisión y sus efectos.

Palabras clave: Xerostomía, Hiposalivación, machine learning, fármaco, predicción

ÍNDICE

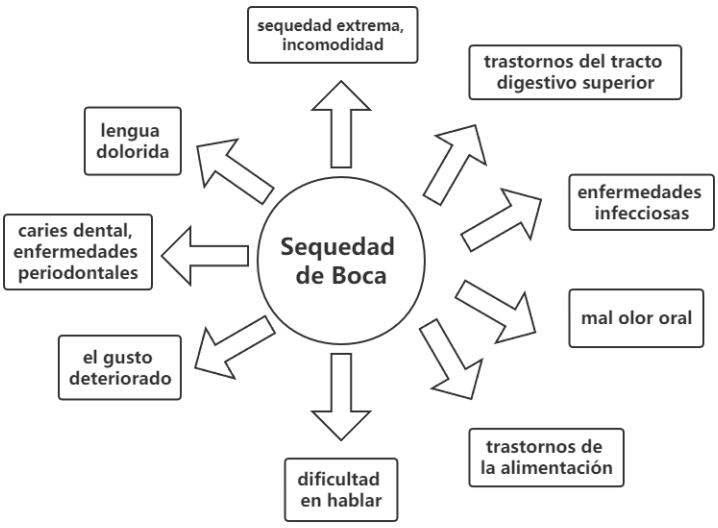
CAPÍTULO 1. Introducción y objetivos

1.1. Introducción

La saliva es el fluido que hidrata la cavidad oral. La saliva tiene importantes funciones: facilita la digestión, lubrica la mucosa oral y los dientes, remineraliza el esmalte reduciendo el riesgo de caries, tiene actividad antifúngica y antibacteriana y tiene un efecto tampón (Carramolino-Cuéllar et al., 2018; Villa et al., 2014). Cuando disminuye el flujo salival aumenta el riesgo de diferentes enfermedades orales (Figura 1) entre las que se encuentran la caries, la candidiasis y la aparición de úlceras orales. También la disminución de saliva dificulta el hablar correctamente, masticar y tragar los alimentos (Saleh et al., 2015; Matsumoto et al., 2020).

Se denomina xerostomía a la sensación subjetiva de sequedad oral e hiposalivación o hiposialia a la reducción objetiva del flujo salival (Aliko et al., 2015; Villa et al., 2014). La prueba más habitual para evaluar la reducción del flujo salival es la sialometría. Se considera que un paciente tiene hiposalivación cuando el flujo salival es menor de 0,1mL/min en reposo o menor de 0,7mL/min cuando es estimulado (Villa et al., 2014). La xerostomía es una enfermedad relativamente frecuente en pacientes mayores. Su prevalencia varía entre el 17 y el 40% en pacientes mayores no institucionalizados (Liu et al., 2012).

Figura 1, Problemas causados por la xerostomía



Fuente: Matsumoto, N., Ushikoshi-Nakayama, R., Yamazaki, T. et al. (2020).

Existen diferentes factores de riesgo que contribuyen a la aparición de xerostomía e hiposalivación y deben ser tenidos en cuenta (López-Pintor et al., 2022). En primer lugar, existen estudios que han observado como las mujeres sufren más estas patologías. Una explicación comúnmente aceptada para este fenómeno es que las mujeres tienen glándulas salivales más pequeñas. En segundo lugar, existen enfermedades que también afectan a las glándulas salivales. Entre dichas enfermedades se encuentran ciertas enfermedades reumatológicas (como el síndrome de Sjögren), la diabetes y enfermedades del sistema nervioso. En tercer lugar, la edad es otro factor que favorece la aparición de estas patologías, existiendo un porcentaje elevado de pacientes de edad avanzada que sufren estos problemas. Por otro lado, la radioterapia de cabeza y cuello y la quimioterapia también inducen sequedad bucal por afectación de las glándulas salivales.

Además de los factores mencionados anteriormente, la toma de medicamentos también puede hacer que una persona desarrolle xerostomía o hiposalivación. Hay explicaciones acerca de la implicación de ciertos fármacos en la reducción del flujo salival. Pero no es posible establecer una explicación para todos los fármacos implicados. El uso de los fármacos en la industria médica actual es diverso, según la lista de medicamentos esenciales de la OMS (Enrique, 2004), existen aproximadamente 250 principios activos, y esta lista se sigue actualizando continuamente. En el conjunto de datos utilizado en este trabajo se han registrado 256 fármacos de uso común y es importante señalar que los principios activos de los medicamentos se han clasificado según la clasificación ATC de la OMS: según su órgano o sistema de acción (A), su mecanismo terapéutico y farmacológico (T) y sus propiedades químicas (C) (ATC de la OMS , 2021). Con referencia a este punto de la influencia de fármacos, en la actualidad se han publicado numerosos trabajos de investigación sobre los efectos del consumo de medicamentos en la xerostomía e hiposalivación. Existen estudios que han mostrado la asociación de estas alteraciones con el uso de antihipertensivos (Ostermann et al., 1988) o con el uso de analgésicos opioides (Looström et al., 2011). En el estudio de Ostermann que comparó el uso de rilmenidine y placebo en el tratamiento de pacientes hipertensos, se produjo un aumento significativo de la incidencia de pacientes con sensación de sequedad de boca cuando se utilizaron 2 mg/día de rilmenidine. En otro ensayo sobre el tratamiento de la boca seca, Looström y cols. (2011) utilizaron Tramadol para inducir sequedad oral en los sujetos y luego utilizaron pilocarpina para tratar dicha sequedad, encontrando finalmente que la pilocarpina aumentaba el flujo salival. En el trabajo de López-Pintor y cols. (2022), se elaboró una tabla más actualizada con todos los fármacos que actualmente se han encontrado asociados a la xerostomía y la hiposalivación y las explicaciones específicas acerca de su posible causa.

Sin embargo, existe controversia acerca de la asociación de diferentes fármacos con la xerostomía e hiposialia. Algunos fármacos producen xerostomía, pero no siempre provocan una reducción del flujo de saliva (Liu et al., 2012). En segundo lugar, no se conoce si la asociación de diferentes fármacos con posible efecto xerostomizante podría aumentar el riesgo de sufrir xerostomía e hiposalivación. Los fármacos tienen diferentes mecanismos de acción para producir xerostomía e hiposialia y otros podrían tener un efecto protector sobre estas patologías. Por lo tanto, existe necesidad de estudiar la asociación entre la ingesta de fármacos y las dos enfermedades. Además de esto, también es necesario estimar los cambios en el riesgo de enfermedad después de tomar estos fármacos específicos.

Para dar resolución a estos dichos problemas, este trabajo trata de proponer algunos enfoques de aprendizaje automático para evaluar cómo afecta cada fármaco a la xerostomía y la hiposalivación. Para ello es necesario clasificar los fármacos en sus cinco niveles (según la OMS): anatómico, terapéutico, farmacológico, químico y compuesto, pudiendo proporcionar un conocimiento a priori para la prevención de la enfermedad. Además de los modelos que responden a la asociación entre las variables y los resultados, se utilizará el modelo de caja negra con una fuerte capacidad predictiva, que es un modelo basado únicamente en las relaciones de entrada y salida, proporcionando predicciones precisas que dan a los médicos una referencia brillante, aunque no se permite captar el proceso interno de toma de decisiones. Una aplicación de clustering jerárquico adicional puede informar sobre las distribuciones del consumo de drogas de los diferentes grupos.

1.2. Objetivos

El objetivo principal del presente estudio es establecer un par de modelos de clasificación y hacer una evaluación de su rendimiento. Con la ayuda de modelos explicativos de clasificación se podrá evaluar la asociación de varios fármacos a la xerostomía y la hiposalivación en un grupo amplio de pacientes polimedicados. A continuación, se intentarán realizar predicciones a partir de las condiciones identificadas. Los objetivos principales pueden subdividirse en los tres puntos siguientes.

1. Identificar en cada uno de los cinco niveles de fármacos de la Clasificación ATC los grupos de fármacos que tienen un efecto sobre la xerostomía e hiposalivación y analizar la dirección de su acción.
2. Hacer predicciones basadas en los datos recogidos sobre los fármacos y determinar si el paciente está en riesgo de padecer xerostomía e hiposalivación.

3. Evaluar la capacidad de predicción del modelo y seleccionar el modelo ganador en función de las métricas predeterminadas en cada nivel de fármacos, en último paso comprobar la capacidad tras su aplicación a una nueva muestra desconocida.
4. Realizar un modelo de agrupación para averiguar si existe un grupo de pacientes que sufren dichas alteraciones y estudiar sus características.

1.3. Estructura de la memoria

La estructura de la memoria es la siguiente:

- Capítulo 1. Introducción: En este capítulo se distingue entre las definiciones de xerostomía e hiposalivación, se explica la motivación del trabajo y se presentan los objetivos previstos y la estructura de la memoria.
- Capítulo 2. Estado del arte: En este capítulo se describe el estado actual de la investigación sobre la relación entre los medicamentos y la xerostomía y la hiposalivación.
- Capítulo 3. Materiales: En este capítulo se presenta el conjunto de datos utilizado para este trabajo y se desarrolla un análisis descriptivo.
- Capítulo 4. Metodología: En este capítulo consta de metodología SEMMA, algoritmos de aprendizaje máquina y métricas de evaluación de modelos, donde SEMMA es una guía completa de métodos para la construcción de modelos y proporciona criterios para la resolución de problemas.
- Capítulo 5. Resultados: En este capítulo se resume el proceso de tuneado de los parámetros del modelo y los resultados óptimos, que son la adquisición inicial del modelo. Además, contiene un análisis de los efectos de los medicamentos propuestos por el objetivo.
- Capítulo 6. Comparación de modelos: En este capítulo se presenta una comparación de los modelos finales obtenidos con validación cruzada, se evalúa el rendimiento para los modelos de xerostomía e hiposalivación, respectivamente, e incluyendo también un análisis del clustering jerárquico.
- Capítulo 7. Discusión: En este capítulo se propone comentarios detallados acerca de los fármacos identificados y de modelos seleccionados, desde un punto de vista médico y técnico respectivamente.

- Capítulo 8. Conclusiones y líneas de trabajo futuro: Las conclusiones consisten en un análisis resumido y de las limitaciones encontradas, y la línea de trabajo futura expone brevemente las opciones de mejora futura en los aspectos técnicos.

CAPÍTULO 2. Estado del arte

2.1. La influencia de fármacos en alteraciones salivales

Existen más de mil fármacos que se han asociado con la xerostomía. Entre los fármacos más frecuentemente asociados a este problema se encuentran los antagonistas de los receptores muscarínicos, los antidepresivos tricíclicos, los opioides, las benzodiacepinas, los antipsicóticos, los antihipertensivos y los antihistamínicos (Aliko et al., 2015; Villa et al., 2014).

Existen estudios que han evaluado el efecto de los fármacos en las alteraciones salivales, pero hoy en día no se ha esclarecido si estos fármacos producen un descenso en el flujo salival o simplemente xerostomía (Ramírez Martínez-Acitores et al., 2021; Ramírez Martínez-Acitores et al., 2020; Ivanovski et al., 2015; Prasanthi et al., 2014; Nonzee et al., 2012; Tahrir y Aldelaimi, 2006; Nederfors et al., 2004). Según ciertos estudios un 52% de los hombres y un 65% de las mujeres de edad avanzada toman al menos un medicamento (Sreebny et al., 1997). Además, muchos de los pacientes toman varios fármacos para tratar las diferentes enfermedades que sufren, por lo que estos fármacos podrían tener un efecto sumativo sobre la xerostomía e hiposalivación (López-Pintor et al., 2022; Dalodom et al., 2016; Pereira et al., 2016; Abdullah, 2015; Thomson et al., 2006). De hecho, hay trabajos que han observado como la prevalencia de xerostomía aumenta con el número de fármacos que recibe un paciente (Han et al., 2015). La mayoría de los estudios previos acerca de este tema son estudios observacionales que analizan si ciertos fármacos se asocian a la xerostomía y la hiposalivación. En el presente trabajo, por primera vez, se utilizarán técnicas de aprendizaje automático para intentar analizar los efectos de los medicamentos en la xerostomía e hiposalivación y predecir de ese modo el riesgo de sufrir estas alteraciones salivales a la hora de recibir dichos fármacos.

2.2. Uso de aprendizaje máquina para analizar las relaciones fármacos-alteraciones salivales

Hay algunos estudios que han logrado buenas predicciones de la xerostomía utilizando métodos de aprendizaje máquina (Chao et al., 2022; Men et al., 2019). En sus estudios, se confirmaron una relación entre los pacientes sometidos a radioterapia de cabeza y cuello para el tratamiento del cáncer y la susceptibilidad a la xerostomía. Sin embargo, los factores de estos estudios solo se centran en el tratamiento de radioterapia de cabeza y cuello y en concreto de su dosis de radiación. Pero, actualmente se echa en falta investigación que utilice el aprendizaje máquina para analizar el efecto de los fármacos sobre la xerostomía e hiposalivación. Cuando se trata de los usos de fármacos, la mayoría de los estudios previos acerca de este tema son estudios

observacionales que analizan si ciertos fármacos se asocian a la xerostomía y la hiposalivación. En el presente trabajo, por primera vez, se utilizarán técnicas de aprendizaje máquina para intentar analizar los efectos de los medicamentos en la xerostomía e hiposalivación y predecir de ese modo el riesgo de sufrir estas alteraciones salivales a la hora de recibir dichos fármacos.

CAPÍTULO 3. Materiales

En este capítulo, se presenta la composición del conjunto de datos aplicado y una sencilla visualización para explorarlos.

3.1. Descripción de los datos

Este conjunto de datos procede de un grupo de pacientes que acudían a dos centros de Salud de atención primaria de la Ciudad de Madrid: el centro de Salud Adelfas y Canal de Panamá. Los datos incluyen información básica como la edad, el sexo y los antecedentes de tabaquismo, además del uso de medicamentos y enfermedades del paciente. También dichos datos incluyen si el paciente tenía sensación de boca seca (xerostomía) y si presentaban hiposalivación (valor de saliva <0,1 ml/min al recoger su flujo salival no estimulado durante 15 minutos entre las 8-11 de la mañana). Las patologías orales a estudiar son dos: la xerostomía (sensación de boca seca experimentada por el paciente, para ello se le realizó un cuestionario) e hiposalivación (o flujo salival reducido al realizar la sialometría, definido siguiendo los parámetros previamente indicados).

3.2. Análisis descriptivo de los datos

A continuación, se muestra un estudio visual de los datos considerados para el modelo. En el estudio se puede observar cómo se están usando los fármacos y en términos generales los efectos probables de una droga en relación con xerostomía e hiposalivación.

3.2.1. Xerostomía

En la figura A27 de anexo se muestra unos diagramas de barras con los diversos fármacos tomados y los estados de los pacientes que padecen xerostomía o no.

Nivel 1

Se puede observar que las drogas de categoría A y N se consumen ampliamente, seguidas de las de categoría B, mientras que las de categoría L son consumidas por un número muy reducido de personas. Analizado desde el punto de vista de la morbilidad, puede haber un efecto xerostomizante para los fármacos de las categorías G y N, ya que el número de pacientes con patología que los toman supera al de los que no sufren la enfermedad.

Nivel 2

Hay un total de 16 tipos de fármacos en nivel secundario, que consisten principalmente en los que comienzan con A, B y C. Aquí destacan A02, A10, C03, N05 y N06 por su efecto xerostomizante, mientras que C08 y C09 muestran un efecto protector.

Nivel 3

Esta es una representación detallada de la clase de fármacos de nivel terciario y es coherente con el análisis anterior. En el sentido de que la mayoría de los fármacos C09, como C09A, C09B y C09C, mostraron protección, mientras que el C09D, que también pertenece a la familia C09 no lo hizo. Además, más de dos tercios del total de personas que tomaron N06A sufrían xerostomía.

Nivel 4

El número de personas que toman C10AA es extremadamente alto. Con este fármaco más del 50% de las personas que lo toman desarrollan xerostomía. Es por ello que este fármaco puede estar asociado a la xerostomía.

Nivel 5

La especificidad de esta clase de códigos de drogas corresponde exactamente a los medicamentos que se toman en la vida real. Observar que surgen algunas incoherencias con el análisis anterior, ya que los medicamentos de la serie C09 son teóricamente protectores, pero el C09AA02 tiene un claro efecto xerostomizante.

3.2.2. Hiposalivación

La figura A28 de anexo incluye información sobre si el paciente tiene hiposalivación y las ingestas de medicamentos.

Nivel 1

En contraste con los resultados de la sección de xerostomía, los fármacos de las series A y N no se asociaban a hiposalivación, ya que más de la mitad de las personas que los tomaron, no sufrían hiposalivación.

Nivel 2

Con el C10 se puede presumir que tiene un efecto protector importante. También algunos fármacos tienen el resultado contrario. Observar que, en este caso, el fármaco A02 que tiene un efecto protector.

Nivel 3

Como se discutió en el nivel 2, el número de pacientes con producción reducida de saliva como resultado de tomar C10A era mucho menor que el número de personas con hiposalivación que lo tomaban, y era el único que aparecía en la serie C10. Esto demuestra el fuerte efecto protector de esa serie.

Nivel 4

La segmentación de los medicamentos hace que la mayoría de los fármacos de esta clase tengan una muestra reducida, aunque sólo una minoría de personas toma el C07AB y más del 60% de las personas sufrían de hiposalivación.

Nivel 5

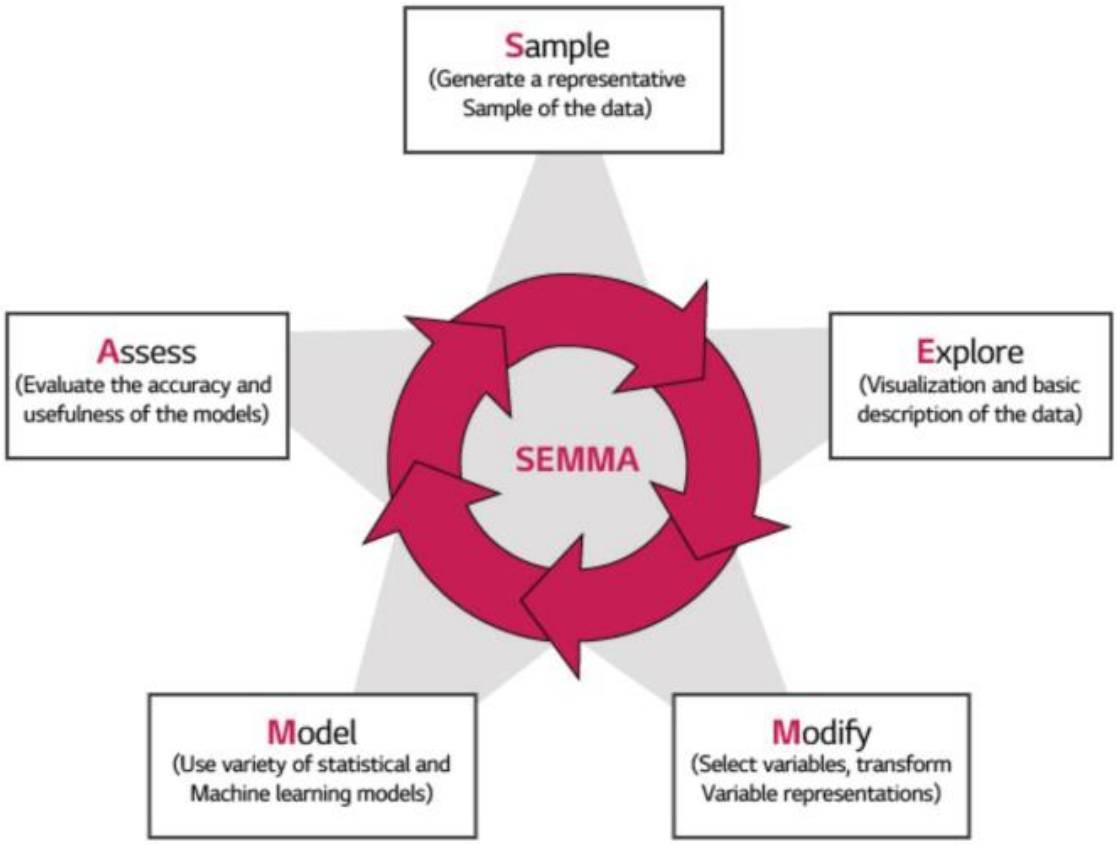
El único fármaco de interés general aquí es el A02BC01, que se asocia a una reducción en la producción de saliva si observamos la proporción de enfermos que lo toman respecto a los que no.

CAPÍTULO 4. Metodología

4.1. Metodología SEMMA

Para resolver los problemas presentados en este trabajo, se utilizará un importante y eficaz esquema de trabajo llamado SEMMA (SAS Institute Inc, 2017). SEMMA está diseñado para formar una especificación completa para la implementación de la minería de datos y está desarrollado principalmente por la empresa de software SAS, y CRISP-DM (Cross Industry Standard Process for Data Mining). Este marco SEMMA tiene nombre completo que debe incluir los siguientes cinco pasos, como se muestra en la Figura 2: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modalización), Assess (Evaluación). Cada uno de estos cinco elementos es crucial y orienta e influye directamente en la precisión de los resultados finales obtenidos. La implementación de cada paso se explica a continuación:

Figura 2. Elementos de Metodología SEMMA



Fuente: Calviño, A. (2021). Técnicas y Metodología de la Minería de Datos (SEMMA)

- **Muestreo**. La selección de la muestra es un requisito importante; un tamaño excesivo de muestras puede ser una dificultad para la fase de modelización y llevar más tiempo para calcular el modelo. Por supuesto, un tamaño de muestra demasiado pequeño también empeora los resultados, ya que la máquina no puede aprender la información de una muestra insuficiente.
- **Exploración**. En este paso se realizan observaciones sobre muestras conocidas para comprobar la distribución de los datos. Si hay valores que faltan y valores atípicos, hay que hacer una eliminación o imputación para garantizar la pureza de los datos. La definición común de un valor atípico (Zheng, 1995) es una medición que se desvían de la media en más del doble de la desviación estándar.
- **Modificación**. Las modificaciones de los datos pueden mejorar el rendimiento de la modelización, lo que incluye la eliminación y la sustitución. La eliminación de las variables que tienen una distribución dispersa puede mejorar la velocidad de cálculo del modelo, y la selección adecuada de las variables puede reducir el error del modelo. La sustitución toma datos que representan un gran número de categorías y los engloba en un pequeño número de categorías según reglas artificiales, lo que implica reducir la complejidad del modelo y aumentar la interpretabilidad.
- **Modelización**. Hallar los modelos es el paso clave, en primer lugar, hay que determinar la selección de la clase de modelo en función del objetivo. En este trabajo requiere utilizar un modelo de predicción de clasificación. Una vez que se han determinado la clase de modelo, también es necesario saber qué modelos específicos son apropiados para utilizar, por ejemplo, en condiciones en las que el tamaño de la muestra es pequeño y el número de características es grande, los modelos de redes neuronales no suelen ser una opción preferida. Además, los objetivos exigen que se utilice al menos un modelo de clasificación interpretable.
- **Evaluación**. La comprobación de la calidad de los modelos y la comparación de las diferencias entre ellos es una parte central de la sección de evaluación. Normalmente, de forma interna, se utilizan una serie de métricas prediseñadas para evaluar la capacidad predictiva de los modelos y si sus resultados son realistas. A continuación, se realiza una comparación externa y se selecciona el modelo con mejores resultados como modelo ganador.

4.2. Preparación de los datos y selección de variables

Gracias al cuidado puesto por los odontólogos involucrados en la recogida de datos, los datos son exhaustivos y no hay missing ni outliers. Por lo tanto, no es necesario realizar ninguna operación de limpieza o imputación para este conjunto de datos.

4.2.1. Remuestreo de los datos

Sin embargo, el conjunto de datos presenta un problema con respecto al tamaño, dado que sólo contiene 220 observaciones. Es por ello que se decidió realizar un remuestreo con reemplazo para aumentar el tamaño de la muestra a 420. Una vez finalizado el remuestreo, hay que realizar una partición para el conjunto de entrenamiento y el conjunto de prueba cumplían una proporción del 75% frente al 25%, la condición de partición debería ser muestreo aleatorio sin reemplazo. Así, el resto de la manipulación de los datos sólo se implementará sobre el conjunto de entrenamiento, mientras que el conjunto de prueba se mantendrá inalterado considerándose una muestra desconocida.

Teniendo en cuenta la especificidad de los datos de la historia clínica, había más de 400 variables que representan el consumo de fármacos en diferentes niveles de clasificación. Es por ello, que fue necesario dividir el conjunto de datos en cinco subconjuntos más pequeños usando el sistema ATC:

- Nivel 1: Órgano o sistema en el cual actúa el fármaco. Existen 14 grupos en total.
- Nivel 2: Subgrupo terapéutico, identificado por un número de dos cifras.
- Nivel 3: Subgrupo terapéutico o farmacológico, identificado por una letra del alfabeto.
- Nivel 4: Subgrupo terapéutico, farmacológico o químico, identificado por una letra del alfabeto.
- Nivel 5: Nombre del principio activo o de la asociación farmacológica, identificado por un número de dos cifras.

Fuente: ATC de la OMS (2021)

Así, para lograr el objetivo planteado, se necesitan cuatro modelos de clasificación en cada nivel que garanticen que los datos de cada nivel puedan interpretarse y predecirse. Por tanto, será necesario comparar 40 modelos: 4(modelos) * 5(niveles) * 2 (patologías: xerostomía o hiposalivación).

Sin embargo, aun dividiendo el conjunto de datos, la dimensionalidad en cada conjunto sigue siendo enorme, de manera que, si se utilizan todas las variables parar obtener los modelos de aprendizaje, esto supondría una enorme carga computacional para el cálculo del modelo y un riesgo de sobreajuste. Es por ello, que se decidió realizar una selección de variables.

4.2.2. Selección de variables

La selección de variables se realizó de la siguiente manera:

- Se revisó la estructura y contenido de las variables, y se apreció que en las diferentes clases cada variable es independiente y representa fármacos completamente diferentes en el ámbito médico.
- La mayoría de los fármacos son consumidos por menos de un 8% de toda la muestra de entrenamiento. Por lo que, su uso en un análisis estadístico no sería representativo Para este tipo de casos, hay dos soluciones: a) Usar data-smothing, o b) Eliminar los datos y utilizar sólo aquellos que tengan una distribución razonable. Aunque se pierde algo de información al utilizar el segundo enfoque, se decidió que la eliminación era la opción más apropiada para esta muestra de entrenamiento.

Finalmente, las variables seleccionadas para su uso en los modelos para investigar el problema de la xerostomía e hiposalivación fueron las siguientes:

Tabla 1. Variables utilizadas de cada nivel en los modelos

Nivel de fármacos	Nombre de variables seleccionadas
Nivel 1: Órgano o sistema	SistemaA, SistemaB, SistemaD, SistemaG, SistemaL, SistemaM, SistemaN, SistemaR, SistemaS
Nivel 2: Terapéutico	A02, A10, A11, A12, B01, C03, C07, C08, C09, C10, G04, N02, N05, N06, R03, S01
Nivel 3: Farmacológico	A02B, A10B, A12A, B01A, C07A, C08C, C09A, C09B, C09C, C09D, C10A, G04C, N02A, N02B, N05B, N06A
Nivel 4: Químico	A02BC, A10BA, A10BD, B01AC, C07AB, C08CA, C09AA, C09BA, C09CA, C09DA, C09DB, C10AA, G04CA, N02BB, N02BE, N05BA, N06AB

Nivel 5: Principio activo	A02BC01, A10BA02, A12AXP1, B01AA07, B01AC06, C08CA01, C09AA02, C09AA03, C09BA03, C09DB01, C09DB02, C10AA01, C10AA05, D09DB06, G04CA02, N02BB02, N02BE01, N05BA06, N05BA08

4.2.3. Balanceado de la muestra

La muestra no requiere ser balanceada dado que la proporción de pacientes con xerostomía respecto a los no pacientes es de 1:1 y la proporción de pacientes con hiposalivación respecto a los que no padecen dicha patología es de 4:6.

4.3. Algoritmos de aprendizaje máquina

En este trabajo, se busca un modelo que relacione la presencia de xerostomía e hiposalivación con la ingesta de determinados medicamentos. A parte de este objetivo, se necesitan clasificadores más avanzados para investigar la probabilidad de que estos individuos sufran xerostomía e hiposalivación basándose de la información de tomar los medicamentos.

Para ello se han utilizado cuatro algoritmos supervisados: Stepwise Regresión Logística, Árboles de Decisión, Support Vector Machine y Extreme Gradient Boosting, y un algoritmo no supervisado: Clustering jerárquico. Concretamente:

- El algoritmo Stepwise Regresion Logística se utiliza para seleccionar las variables valiosas y exponer la influencia de aquellas variables sobre la variable de respuesta. Normalmente, se requiere una transformación matemática (exponencial) para demostrar la relación lineal entre las variables predictoras y la repuesta, la relación lineal incluye la positiva o la negativa y el grado de influencia, esto dará ayuda a analizar los factores que influyen en la xerostomía y la hiposalivación.
- Los árboles de decisión se utilizan para obtener resultados predictivos a través de un proceso de juicio múltiple. Con la visualización de los árboles de decisión se permite obtener información sobre los factores que influyen en la xerostomía y la hiposalivación.
- SVM y XGBOOST se utilizan para estimar el riesgo de un individuo de desarrollar la enfermedad en función del consumo de medicamentos. al ser modelos más complejos, suelen dar lugar a un mejor rendimiento predictivo, sin embargo, adolecen de

dificultades en la interpretación de las variables, por lo que estos dos modelos se analizarán principalmente por su optimización de parámetros y efectos predictivos.

- El algoritmo del Clustering Jerárquico se utiliza para agrupar a todas estas personas de las que se ha recogido información en diferentes perfiles, a continuación, hacer comparación de las diferencias entre ellos. Por ejemplo, en un perfil si ellos tienen características patológicas comunes, cual medicación acostumbran a tomar las personas de este grupo.

En teoría, todos estos son estudios significativos que pueden aportar pistas a profundizar a este tema.

4.3.1. Stepwise Regresión Logística

Se trata de un algoritmo basado en la regresión logística con la adición de la regresión por pasos como función de filtrado de las variables. Es especialmente útil cuando hay un gran número de variables y se necesita reducir la dimensionalidad y la multicolinealidad del modelo. Aunque esta técnica no elimina completamente la multicolinealidad del modelo, sin embargo sí mejora la situación inicial.

El modelo presenta tres variantes de ejecución (Gareth et al. 2013).

- Forward selection. En esta variante, principio el modelo no tiene ninguna variable introducida. A continuación, añade iterativamente predictores que contribuyen al modelo uno por uno hasta que no se pueden introducir nuevos predictores, de acuerdo con el resultado significativo de la Prueba F de Fisher.
- Backward elimination. En esta variante, se ajusta la ecuación de regresión con todas las variables independientes, y los predictores se prueban de menor a mayor en términos de su contribución a la variable dependiente. Aquellos que no son estadísticamente significativos se eliminan sucesivamente hasta que no se puede eliminar ningún predictor.
- Bidirectional elimination, Se trata de una combinación de los dos primeros métodos. Se empieza con una selección hacia delante, añadiendo gradualmente predictores y a continuación, se realiza una eliminación hacia atrás para comprobar si deben ser retenidos o excluidos.

La parte principal del modelo es la regresión logística, por lo que sigue teniendo los inconvenientes de ésta. Si hay muchas entradas muy correlacionadas o si los datos son muy

dispersos (por ejemplo hay muchos ceros en los datos de entrada), el logaritmo no podrá converger bien y el modelo perderá su capacidad de predicción. Una de las ventajas de la regresión logística es la obtención directa del coeficiente del predictor, mediante el cual se puede interpretar la relación entre una variable y las posibilidades de que se produzca un suceso y las posibilidades de que no se produzca. En este caso se correspondería con el problema de clasificación binaria de si una persona tiene probabilidades de estar enferma cuando está tomando un fármaco.

4.3.2. Árbol de Decisión

Un árbol de decisión es un modelo predictivo que representa una relación de mapeo entre los atributos y los valores de los objetos. Cada nodo del árbol representa un objeto, cada camino bifurcado representa un posible valor de atributo, y cada nodo hoja corresponde al valor del objeto representado por el camino desde el nodo raíz hasta ese nodo hoja. Por lo tanto, un árbol de decisión generaliza un conjunto de reglas de clasificación a partir de un conjunto de datos de entrenamiento.

El tamaño del árbol y la precisión del modelo se controlan principalmente ajustando dos parámetros, minsplit y minbucket. El parámetro minsplit representa el número mínimo de nodos de la rama. Así, cuando es mayor que el valor preestablecido, el nodo seguirá dividiéndose, y en caso contrario se detiene. El parámetro minbucket es el número mínimo de muestras contenidas en los nodos hoja del árbol. De manera que si su valor es menor, mayor será el tamaño del árbol de decisión resultante, se ve más grande la forma del árbol.

Una razón para elegirlo es que el árbol de decisión puede ser lógicamente bien explicado en comparación con otros modelos de clasificación de caja negra, lo que puede ayudar a responder al primer pequeño objetivo de este trabajo. Además, el árbol de decisión tiene la ventaja de poder hacer poda para mejorar la generalización, no es un modelo sólido y la estructura del árbol puede cambiar incluso si los datos originales se alteran ligeramente. Por lo tanto, la introducción de un modelo que integre modelos múltiples de árbol de decisión es una buena solución, es decir, la idea boosting.

4.3.3. Support Vector Machine (SVM) con Kernel Radial

La idea en la que se basa el SVM es que las variables del conjunto de datos se pueden interpretar como dimensiones de un plano geométrico, de modo que las variables múltiples formarán un plano hipergeométrico y la información de cada fila del conjunto de datos se mapeará como

puntos en el plano hipergeométrico. De manera que el modelo SVM busca encontrar un hiperplano no lineal que separe todas las categorías binarias.

Los parámetros que controlan el SVM radial son C y gamma. C es el factor de penalización e indica la tolerancia al error. Si el valor de C es mayor, la clasificación del modelo será más ajustada y la clasificación será bastante buena en el conjunto de pruebas. Sin embargo, si C es extremamente grande y estricto, en el conjunto de pruebas no se puede conseguir ese buen efecto y se produce el sobreajuste. Por el contrario, cuanto menor sea el valor de C, el número de clasificaciones erróneas del modelo aumentará y las predicciones serán peores en las nuevas muestras. En cualquier caso, un valor inadecuado hará que el modelo tenga menos capacidad de generalización. El parámetro gamma determina implícitamente la distribución de los datos cuando se mapean en el nuevo espacio de características. Cuanto mayor sea el valor de gamma, menor será el número de vectores de soporte, y cuanto menor sea el valor de la gamma, mayor será el número de vectores de soporte. El número de vectores de soporte afecta a la velocidad del entrenamiento. En la implementación de este modelo en R, se utiliza un parámetro adicional sigma que se relaciona con gamma de acuerdo con la siguiente ecuación:

$$k(x,z) = \exp\left(-\frac{d(x,z)^2}{2*\sigma^2}\right) = \exp\left(-gamma \cdot d(x,z)^2\right) \Rightarrow gamma = \frac{1}{2\cdot\sigma^2}$$

En esta función, x y z representa dos muestras distintas, la función d denota la distancia euclidiana al cuadrado entre los dos vectores de características. σ es un parámetro libre.

Por lo tanto, sigma y gamma muestran una relación opuesta, ya que una sigma mayor indica que se utilizan más vectores de soporte en el modelo.

La SVM radial es una técnica eficaz para resolver problema de características de alta dimensión, ya que sólo utiliza vectores de soporte (puntos cercanos a la superficie de decisión) para tomar decisiones sobre el corte del hiperplano. Es decir, no depende de los datos completos. Sin embargo, no es adecuado para situaciones con tamaños de muestra muy grandes, en las que el mapeo de la función de kernel es de muy alta dimensión dado que impone una enorme carga computacional al ordenador. Por lo tanto, en teoría es un algoritmo muy interesante para la resolución de problemas médicos, ya que normalmente no se disponen de grandes cantidades de datos en la investigación médica.

4.3.4. Extreme Gradient Boosting (XGBoost)

XGBoost es un algoritmo de refuerzo integrado que proporciona una alta eficiencia de entrenamiento y bueno resultados de predicción. Se basa en el Boosting, se trata esencialmente de un proceso continuo de construcción de un árbol, haciendo otra predicción basada en los residuos del árbol de predicción última, reduciendo continuamente el residuo de predicción hasta que no se pueda reducir más o se cumpla la condición del número de iteraciones. Se considera como un algoritmo que mejora los clasificadores débiles, en este caso, XGBoost integra un par de árboles de decisión para formar un clasificador fuerte que garantice resultados de clasificación más fiables.

XGBoost tiene un gran número de parámetros generales, siendo los tres siguientes los más importantes, learning rate, min child weight y nrounds. El parámetro learning rate indica la longitud del paso de cada iteración. Si se configura un learning rate demasiado grande, no funcionará con precisión, y si es demasiado pequeño, funcionará lentamente. El parámetro de nrounds explica el número de iteraciones del algoritmo, es decir, cuántos árboles se generarán (el número de árboles determina la complejidad del modelo). Por último, el parámetro min child weight (What is minimum child weight in XGBoost, 2019), es muy similar a minbucket en los árboles de decisión, sumando mínima de peso de instancia (hessiana) necesaria en un hijo nodo. Si el paso de partición del árbol da como resultado un nodo hoja con la suma de peso de instancia menor que min child weight, el proceso de construcción renunciará a seguir partiendo. Entonces, este parámetro puede describir el número de muestras en los nodos hoja de un árbol, y su ajuste afectará al riesgo de sobreajuste o subajuste.

4.3.5. Clustering Jerárquico: Average Linkage

El método Average Linkage tiene como objetivo distinguir aquellas muestras que no comparten características comunes y clasificarlas en diferentes grupos. Para ello, la lógica es calcular la distancia entre dos clusters como la distancia media entre todos los elementos y, mediante un cálculo iterativo, va fusionando los elementos que tienen la distancia más cercana hasta que no se puedan agregar más al grupo existente. Este modelo utiliza las distancias euclidianas para obtener la matriz de distancias, que se define mediante la siguiente fórmula:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

Fuente: Yusniyanti et al. (2021). Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators

Donde (UV) y W representan dos clústeres diferentes, d_{ik} representa la distancia entre los elementos i de cluster (UV) y k de clúster W, y N indica el número de elementos en este clúster.

4.4. Métricas para la evaluación de modelos

Para evaluar el rendimiento de los modelos de aprendizaje se van a utilizar las siguientes métricas.

4.4.1. Exactitud (Accuracy)

La accuracy es una métrica que se define como la parte de los resultados de predicción del modelo que son bien clasificados, la fórmula para calcular la exactitud se encuentra en la siguiente ecuación:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Los cuatro elementos de la ecuación representan diferentes clases de muestras clasificadas.

True positive (TP). Clase verdadera positiva, el modelo identifica la muestra positiva original correctamente.

False Negative (FN). Clase falsa negativa. La clase verdadera de la muestra es positiva, pero el modelo la identifica como negativa.

False Positive (FP). Clase falsa positiva. La clase verdadera de la muestra es negativa, pero el modelo la identifica como positiva.

True Negative (TN). Clase verdadera negativa, el modelo identifica la muestra negativa original correctamente.

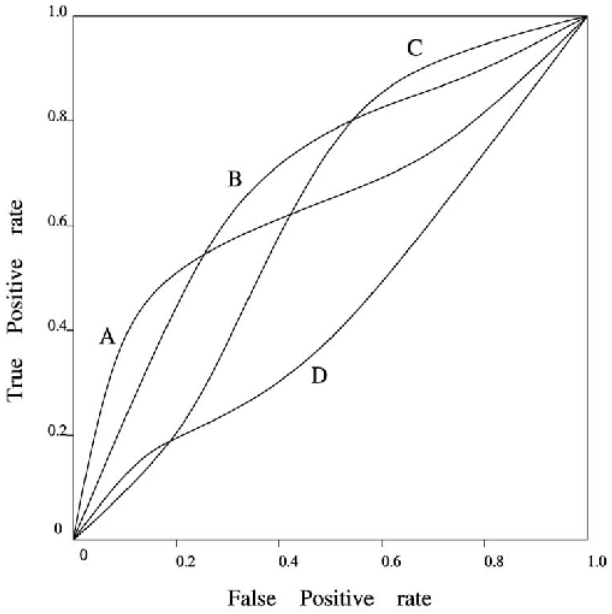
Por lo tanto, la métrica accuracy es el porcentaje de todas muestras bien clasificadas por el modelo predictivo.

4.4.2. AUC – The Area Under the Curve

ROC (The Receiver Operator Characteristic) es una curva de probabilidad que traza TPR contra FPR en diferentes umbrales, TPR es una ratio de bien predecir una muestra positiva y FPR está contrario, indica la ratio de predecir una muestra negativa como una muestra positiva, se encuentran los dos conceptos correspondientes en sección 4.4.3 (sensibilidad) y 4.4.4

(especificidad), respectivamente. Así, el área bajo la curva (AUC) es una medida de la capacidad del clasificador para distinguir entre categorías y se utiliza como un resumen de la curva ROC.

Figura 3. Un ejemplo de cuatro curvas ROC.



Fuente: Huang, J. et al (2005)

Observando este ejemplo con cuatro curvas, una curva ROC completa indica que tiene un valor AUC más alto y que el modelo tiene una mejor capacidad para distinguir entre casos positivos y negativos.

4.4.3. Sensibilidad

La sensibilidad igual a la TPR (True Positive Rate), o se llama recall. Esta métrica representa el porcentaje de casos que son en realidad positivos y están correctamente clasificados como enfermos según los criterios de clasificación del modelo. En el presente trabajo refleja la capacidad de una prueba diagnóstica para detectar pacientes que van a sufrir xerostomía o hiposalivación.

$$\text{Sensitivity } = TPR = \frac{TP}{TP + FN}$$

4.4.4. Especificidad

La especificidad es una métrica opuesta a la sensibilidad, es el TNR (True Negative Rate). La especificidad dice cuál es la proporción de clasificación correcta para los casos negativos,

reflejando la capacidad de una prueba diagnóstica para identificar a los no pacientes. Si la especificidad de una prueba diagnóstica es baja, habrá muchos casos falsos positivos. Esto hace que se desperdicien recursos médicos y provoca miedo y ansiedad injustificados en los pacientes.

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FP}$$

Existe un orden de prioridad para estas métricas dichas cuando se utilizan para evaluar el modelo, con el AUC en primer lugar y la accuracy seguida. Porque el AUC se calcula basándose en todos los umbrales, mientras que la accuracy utiliza un único umbral (normalmente es 0,5) y cuando se cambia el umbral, el resultado de la accuracy cambia con él, lo que hace que el resultado del AUC sea más estable y fiable.

La sensibilidad y la especificidad se utilizan como segunda referencia, dando su capacidad para clasificar correctamente las muestras positivas y negativas, respectivamente, después de determinar los modelos ganadores según el criterio del AUC y la accuracy.

CAPÍTULO 5. Resultados

En este capítulo se describen los resultados obtenidos al genera cada modelo de aprendizaje.

5.1. Modelo de Stepwise Regresión Logística

En este modelo único que se puede ajustar es la dirección del stepwize. Después de varios intentos las direcciones utilizadas para cada nivel de datos se tabulan a continuación:

Tabla 2. Stepwize dirección para regresión logística

Xerostomía	**Stepwize dirección**	**Hiposalivación**	**Stepwize dirección**
Nivel 1	Forward	Nivel 1	Backward
Nivel 2	Bidirection	Nivel 2	Bidirection
Nivel 3	Backward	Nivel 3	Backward
Nivel 4	Backward	Nivel 4	Backward
Nivel 5	Forward	Nivel 5	Forward

Con este método se puede conseguir un número reducido de predictores dado que sólo considerarán las variables significativas en el modelo:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

donde β es el coeficiente de la función de regresión. Este puede interpretarse como el cambio en la variable dependiente resultante de un cambio unitario en la variable independiente.

Hay un nuevo concepto aquí que se llama odds, representa la proporción de evento que ocurre sobre evento que no ocurre, se describe en una función a continuación:

$$\text{odds} = \frac{p}{(1-p)}$$

El conjunto de datos utilizado determina que el valor de x sólo podía ser cero o uno, siendo cero que representaba que no se tomaba la droga y uno lo contrario. Para entender el significado de los coeficientes en este estudio, en primer lugar, es necesario aclarar un requisito previo con un ejemplo de fármaco X_1, hay que suponer toda la condición demás constante menos X_1, es decir, el individuo únicamente toma este fármaco X_1, ahora el coeficiente correspondiente al fármaco

X_1 es β_1, β_1 puede explicarse como el riesgo previsto de xerostomía o hiposalivación con o sin el fármaco X_1 de acuerdo a la siguiente relación de funciones:

$$odds_{tomar\ X1} = e^{\beta_0+\beta_1*1+\cdots}$$

$$odds_{no\ tomar\ X1} = e^{\beta_0+\beta_1*0+\cdots}$$

$$odds\ ratio\ (OR) = \frac{odds_{tomar\ X1}}{odds_{no\ tomar\ X1}} = e^{\beta_1}$$

Según la definición de odds ratios (OR), el valor de referencia de la OR debe ser uno, lo que corresponde a que β_1 sea cero, esta OR representa que no hay cambios en la aparición de la enfermedad de un sujeto que toma dicho fármaco respecto a que no lo toma. El coeficiente β_1 negativo significa que existe menos posibilidad de que un individuo sufra xerostomía o hiposalivación después de tomar el fármaco X_1 respecto a no tomarlo, lo que indica que el fármaco tenía originalmente efecto xerostomizante o podía reducir el flujo de salivación. A la inversa, siempre que β_1 sea mayor que 0, existe un efecto protector del fármaco.

Los coeficientes dados por el modelo y las interpretaciones correspondientes se dan según el análisis presentado anteriormente, todos los coeficientes de este modelo se pueden encontrar en la Figura A1 - A10 del anexo. Así mismo, en la tabla 3 se resumen todas las variables y coeficientes utilizados en el modelo final. y se compararán con los resultados del modelo de árbol de decisión.

5.2. Árbol de Decisión

La poda es la operación principal de los árboles de decisión, y su objetivo es evitar el sobreajuste del modelo. Esto se debe a que el algoritmo del árbol de decisión divide los nodos durante el proceso de aprendizaje para clasificar las muestras de entrenamiento de la forma más correcta posible, lo que conduce a demasiadas ramas en el árbol y, por tanto, a un sobreajuste.

A pesar de haber realizado una prepoda, el modelo es demasiado complejo y da lugar a predicciones que no cumplen las expectativas en el conjunto de prueba. Por lo tanto, para cada modelo ha sido necesario experimentar individualmente con su parámetro cp, también conocido como parámetro de complejidad, que actúa como umbral. Si se eleva, la complejidad del modelo se simplificará, se fusionarán algunos nodos y se reducirá el número final de nodos hoja, lo que también facilita el análisis del proceso de decisión del modelo.

Tras el proceso de tunear, los mejores parámetros para el árbol de decisión se incluyen en la tabla 3.

Tabla 3. Parámetros óptimos para árbol de decisión

Xerostomía	**Minsplit**	**Minbucket**	**CP**	**Hiposalivación**	**Minsplit**	**Minbucket**	**CP**
Nivel 1	12	10	0.01	Nivel 1	20	60	0.01
Nivel 2	4	60	0.01	Nivel 2	20	80	0.02111
Nivel 3	16	80	0.01	Nivel 3	24	20	0.02222
Nivel 4	4	80	0.025	Nivel 4	20	60	0.02111
Nivel 5	24	50	0.0135	Nivel 5	16	70	0.025

A partir de estos parámetros se puede construir un modelo de árbol de decisión final y mostrar el proceso de decisión (véase las figuras A11 - A20 del anexo). En la tabla 4 se resumen los resultados obtenidos y se comparan con el modelo de regresión logística.

Tabla 4. Asociación entre xerostomía y fármacos derivados de los modelos de regresión logística y árbol de decisión.

Nivel 1 ATC						
ATC código	**Fármaco**	**Regresión Logística**		**Árbol Decisión**		**Resultado combinado de los modelos**
		Asociación con la Xerostomía	**P-valor**	**Asociación con la Xerostomía**	**Importancia de Feature**	
D	**Dermatológicos**	Protector	Significativo	Protector	No Importante	Protector
R	**Sistema respiratorio**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
L	**Antineoplásicos e inmunomoduladores**	Protector	No Significativo			No Concluyente
N	**Sistema nervioso**	Xerostomizante	No Significativo	Xerostomizante	Importante	Xerostomizante
G	**Sistema genitourinario y hormonas sexuales**			Xerostomizante	Importante	No Concluyente
Nivel 2 ATC						
A02	**Agentes para el tratamiento de alteraciones causadas por ácidos**			Xerostomizante	Importante	No Concluyente
A10	**Drogas usadas en diabetes**	Xerostomizante	No Significativo			No Concluyente
A12	**Suplementos minerales**					No Concluyente
B01	**Agentes antitrombóticos**	Protector	Significativo	Protector	Importante	Protector
C03	**Diuréticos**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
C07	**Agentes betabloqueantes**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
C08	**Bloqueantes de canales de calcio**	Protector	Significativo	Protector	Importante	Protector
C10	**Agentes modificadores de los lípidos**			Xerostomizante	No Importante	No Concluyente
N05	**Psicolépticos**			Xerostomizante	Importante	No Concluyente

N06	**Psicoanalépticos**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
R03	**Agentes contra padecimientos obstructivos de las vías respiratorias**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
S01	**Oftalmológicos**			Protector	No Importante	No Concluyente
Nivel 3 ATC						
A12A	**Calcio**	Protector	Significativo			No Concluyente
B01A	**Agentes antitrombóticos**	Protector	No Significativo			No Concluyente
C07A	**Agentes betabloqueantes**	Protector	No Significativo			No Concluyente
C08A	**Bloqueantes selectivos de canales de calcio con efectos principalmente vasculares**	Protector	Significativo	Protector	Importante	Protector
C09A	**Inhibidores de la ECA, monoterapia**	Protector	Significativo	Protector	No Importante	Protector
C09B	**Inhibidores de la ECA, asociaciones**			Protector	Importante	No Concluyente
C09C	**Bloqueantes de receptores de angiotensina II (BRA), monofármacos**	Protector	Significativo	Protector	Importante	Protector
N02A	**Opioides**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
N06A	**Antidepresivos**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
G04C	**Drogas usadas en la hipertrofia prostática benigna**			Xerostomizante	Importante	No Concluyente
Nivel 4 ATC						
B01AC	**Inhibidores de la agregación plaquetaria, excluido heparina**	Protector	Significativo	Protector	Importante	Protector

C07AB	**Agentes betabloqueantes selectivos**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
C08CA	**Derivados de la dihidropiridina**	Protector	Significativo	Protector	Importante	Protector
C09AA	**Inhibidores de la ECA, monodrogas**	Protector	Significativo	Protector	Importante	Protector
C09CA	**Bloqueantes de receptores de angiotensina II (BRA), monofármacos**	Protector	Significativo			No Concluyente
C09DB	**Antagonistas de angiotensina II y bloqueantes de canales de calcio**			Xerostomizante	No Importante	No Concluyente
N02BB	**Pirazolonas**	Protector	No Significativo			No Concluyente
N05BA	**Derivados de la benzodiazepina**			Xerostomizante	Importante	No Concluyente
N06AB	**Inhibidores Selectivos De La Recaptación De Serotonina**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
Nivel 5 ATC						
B01AC06	**Ácido Acetilsalicílico**	Protector	Significativo	Protector	Importante	Protector
B01AA07	**Acenocumarol**	Xerostomizante	Significativo	Xerostomizante	Importante	Xerostomizante
C09AA02	**Enalapril**	Xerostomizante	Significativo			No Concluyente
C08CA01	**Amlodipino**	Protector	Significativo	Protector	Importante	Protector
A12AXP1	**Calcio Carbonato/Colecalciferol**	Protector	Significativo	Protector	Importante	Protector
N02BB02	**Metamizol Sódico**	Protector	Significativo			No Concluyente
N02BE01	**Paracetamol**			Protector	Importante	No Concluyente

Tabla 5. Asociación entre la hiposalivación y los fármacos derivados de los modelos de regresión logística y árbol de decisión.

Nivel 1 ATC						
ATC código	**Fármaco**	**Regresión Logística**		**Árbol Decisión**		**Resultado combinado de los modelos**
		Asociación con la Hiposalivación	**P-valor**	**Asociación con la Hiposalivación**	**Importancia de Feature**	
G	**Sistema genitourinario y hormonas sexuales**	Hiposalivación	Significativo	Hiposalivación	Importante	Hiposalivación
R	**Sistema respiratorio**	Protector	Significativo	Protector	Importante	Protector
D	**Dermatológicos**			Hiposalivación	Importante	No Concluyente
N	**Sistema nervioso**			Hiposalivación	Importante	No Concluyente
Nivel 2 ATC						
A02	**Agentes para el tratamiento de alteraciones causadas por ácidos**	Hiposalivación	No Significativo	Hiposalivación	Importante	Hiposalivación
A10	**Drogas usadas en diabetes**	Hiposalivación	Significativo	Hiposalivación	Importante	Hiposalivación
C03	**Diuréticos**	Hiposalivación	Significativo			No Concluyente
C07	**Agentes betabloqueantes**	Hiposalivación	Significativo	Hiposalivación	Importante	Hiposalivación
C08	**Bloqueantes de canales de calcio**	Protector	Significativo	Protector	Importante	Protector
C10	**Agentes modificadores de los lípidos**	Protector	Significativo	Protector	Importante	Protector
N05	**Psicolépticos**	Hiposalivación	Significativo	Protector	Importante	No Concluyente

N06	**Psicoanalépticos**			Protector	Importante	No Concluyente
S01	**Oftalmológicos**			Hiposalivación	No Importante	No Concluyente
Nivel 3 ATC						
C07A	**Agentes betabloqueantes**	Hiposalivación	Significativo	Hiposalivación	No Importante	Hiposalivación
C08C	**Bloqueantes selectivos de canales de calcio con efectos principalmente vasculares**	Protector	No Significativo			No Concluyente
C09A	**Inhibidores de la ECA, monoterapia**	Protector	Significativo	Protector	Importante	Protector
C09B	**Inhibidores de la ECA, combinaciones**	Protector	Significativo			No Concluyente
C09D	**Antagonistas de angiotensina II, combinaciones**	Protector	Significativo	Protector	Importante	Protector
C10A	**Agentes modificadores de los lípidos, monoterapia**	Protector	Significativo			No Concluyente
N02B	**Otros analgésicos y antipiréticos**	Protector	Significativo			No Concluyente
N05B	**Ansiolíticos**	Hiposalivación	Significativo	Hiposalivación	Importante	Hiposalivación
N06A	**Antidepresivos**			Hiposalivación	No Importante	No Concluyente
Nivel 4 ATC						
A02BC	**Inhibidores de la bomba de protones**	Hiposalivación	No Significativo	Hiposalivación	No Importante	Hiposalivación

B01AC	**Inhibidores de la agregación plaquetaria, excluido heparina**	Protector	No Significativo			No Concluyente
C07AB	**Agentes betabloqueantes Selectivos**	Hiposalivación	Significativo			No Concluyente
C08CA	**Derivados de la dihidropiridina**	Protector	No Significativo	Protector	Importante	Protector
C09AA	**Inhibidores de la ECA, monodrogas**	Protector	Significativo			No Concluyente
C09BA	**Inhibidores de la ECA Y Diuréticos**	Protector	Significativo			No Concluyente
C09CA	**Bloqueantes de receptores de angiotensina II (Bra), monofármacos**			Hiposalivación	Importante	No Concluyente
C09DB	**Antagonistas de angiotensina II y bloqueadores de canales de calcio**	Protector	No Significativo			No Concluyente
C10AA	**Inhibidores de la HMG-COA reductasa**	Protector	Significativo	Protector	Importante	Protector
N02BE	**Anilidas**	Protector	Significativo	Protector	Importante	Protector
N05BA	**Derivados de la benzodiazepina**			Hiposalivación	Importante	No Concluyente
N06AB	**Inhibidores selectivos de la recaptación de serotonina**	Hiposalivación	No Significativo			No Concluyente

Nivel 5 ATC						
N02BE01	**Paracetamol**	Protector	Significativo	Protector	No Importante	Protector
N02BB02	**Metamizol sódico**			Hiposalivación	No Importante	No Concluyente
N05BA08	**Bromazepam**			Hiposalivación	Importante	No Concluyente
A02BC01	**Omeprazol**	Hiposalivación	No Significativo	Hiposalivación	Importante	Hiposalivación
A10BA02	**Metformina**			Hiposalivación	Importante	No Concluyente
C09AA02	**Enalapril**			Hiposalivación	Importante	No Concluyente
B01AA07	**Acenocumarol**	Hiposalivación	Significativo			No Concluyente
C08CA01	**Amlodipino**	Protector	Significativo	Protector	Importante	Protector
C10AA01	**Simvastatina**	Protector	No Significativo	Protector	No Importante	Protector
C10AA05	**Atorvastatina**	Protector	No Significativo			No Concluyente

Los códigos de las drogas se pueden encontrar en detalle en página web[1] de la clasificación de fármacos ATC/DDD Index 2022 de la OMS.

[1] https://www.whocc.no/atc_ddd_index/

En la tabla anterior se introducen dos conceptos, el p-valor y la importancia de feature. El P-valor es un parámetro utilizado para determinar el resultado de una prueba de hipótesis, y en la regresión logística determina si el coeficiente es cero o no, representando la existencia del coeficiente de la variable. La importancia de feature tiene una función similar al P-valor, que puede determinar si una variable es necesaria en el proceso de decisión del modelo de árbol de decisión. En la regresión logística, las variables que están presentes en el modelo son todas significativas con un p-valor inferior a 0,1, sin embargo en la tabla 4 y 5 se han elegido una condición más estricta, utilizando 0,05 como umbral para distinguir su significación. Así, sólo las variables con un p-valor menor o igual a 0,05 se consideran significativas.

En cuanto a la importancia de las características, se trata de una medida específica de la familia de los árboles de decisión que consiste en asignar puntuaciones a las características en función de su importancia en la predicción de la variable objetivo, calculándose las puntuaciones a partir de un índice de Gini ponderado. El índice Gini se calcula del siguiente modo:

$$\text{Gini}\ (D) = \sum_{i=1}^{n} p(x_i) * \left(1 - p(x_i)\right)$$
$$= 1 - \sum_{i=1}^{n} p(x_i)^2$$

Donde $p(x_i)$ es la probabilidad de ocurrencia de la clasificación x_i y n es el número de clasificaciones.

En este sentido, Gini (D) refleja la probabilidad de que dos muestras seleccionadas al azar del conjunto de datos D tengan marcadores de categoría inconsistentes. Por tanto, usando esta fórmula se sabe que el índice describe la pureza de los nodos. Así, cuanto menor sea el coeficiente de Gini, mayor será la pureza y mejor será la característica elegida. Por lo tanto, cuanto mayor sea la importancia de la característica, más importante será ésta en el proceso de toma de decisiones. Por el contrario, las características que estén por debajo de un determinado umbral deberían eliminarse del modelo. La importancia de las características se utiliza para la construcción del modelo final.

Por último, observar que se ha añadido una nueva columna que se llama resultado combinado de los modelos, que representa la intersección de la regresión logística y los árboles de decisión. De esta forma, se verifica la validez de los efectos de los fármacos sobre las patologías estudiadas, y reduce la gama de fármacos de interés.

5.3. Modelos de SVM

En este modelo se utiliza el kernel RBF, que es adecuado para el caso linealmente inseparable, con 2 parámetros opcionales y resultados de clasificación muy dependientes de los parámetros introducidos.

Este conjunto de datos tiene un gran volumen de características y contiene demasiada información dispersa para que el problema sea linealmente inseparable. Asimismo, si el número de características es pequeño y el número de muestras es medio Andrew Ng (2017) recomienda utilizar el kernel RBF si se quiere conseguir el objetivo de una predicción precisa, y hay que probar muchas combinaciones diferentes de parámetros.

En las figuras A21 y A22 del anexo se muestran hasta 36 combinaciones los parámetros finales en cada conjunto de datos de los cinco niveles mediante imágenes bidimensionales. Así, se puede observar que la precisión del modelo no tiende a aumentar después de un valor C de 5, lo que es particularmente importante para la elección de los vectores de soporte. Los parámetros óptimos que se utilizarán en modelos se resumen en la siguiente tabla 6.

Tabla 6. Parámetros óptimos para Support Vector Machine

Xerostomía	**C**	**Sigma**	**Hiposalivación**	**C**	**Sigma**
Nivel 1	0.0016	25	Nivel 1	1	0.04
Nivel 2	25	0.2	Nivel 2	10	0.2
Nivel 3	5	0.2	Nivel 3	1	25
Nivel 4	1	0.2	Nivel 4	10	0.2
Nivel 5	5	0.2	Nivel 5	0.008	0.2

De la tabla se desprende que los modelos del segundo nivel de medicamentos son muy estrictos, con un mayor valor de C, independientemente de la condición para la que se utilizan.

5.4. Modelos de XGBoost

XGBoost, como extensión del modelo de árbol de decisión, incorpora la esencia del boosting y establece vínculos entre los árboles, de modo que los bosques ya no existen como árboles independientes entre sí. El primer término de su función objetivo es la función de pérdida, que mide la diferencia entre los valores predichos y los verdaderos. Además, la presencia del término de regularización, el segundo término, se añade a este modelo. Así, cada iteración se regulariza a nivel matemático, controlando con éxito la complejidad del modelo y reduciendo aún más el riesgo de sobreajuste. La función de XGBoost se muestra a continuación:

$$L(\phi) = \sum_{i}^{n} l(\hat{y}_i, y_i) + \sum_{k}^{m} \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2$$

El uso de la técnica de early stopping en XGBoost permite reducir eficazmente el número de cálculos. En concreto, detiene el entrenamiento cuando el lossing en el conjunto de entrenamiento no disminuye (es decir, cuando disminuye menos que un determinado umbral), con ese puede evitar muchos cálculos no requeridos.

Tras probar el número de iteraciones necesarias con parámetros demás fijos , se comprobó que casi todas las muestras requerían sólo 1000 iteraciones como el máximo, excluyendo los datos del quinto nivel bajo el objetivo de xerostomía que requerían 2000 iteraciones. Por otra parte, al comparar las tendencias de precisión con los parámetros usados, los mejores parámetros varían de un modelo a otro. Por último, hay que señalar que los resultados (véase las figuras A23y A24 del anexo) muestran que un total de nueve modelos eligieron 1000 o menos iteraciones, cinco de las cuales fueron inferiores a 1000, lo que demuestra que la técnica de parada temprana es una buena opción.

En tabla 7 están todos parámetros seleccionados al aplicar en el modelo correspondiente.

Tabla 7. Parámetros óptimos para XGBoost

Xerostomía	**Learning Rate**	**Min child weight**	**Nrounds**	**Hiposalivación**	**Learning Rate**	**Min child weight**	**Nrounds**
Nivel 1	0.1	0.5	1000	Nivel 1	0.001	0.5	1000
Nivel 2	0.1	1	300	Nivel 2	0.1	1	500
Nivel 3	0.05	0.5	1000	Nivel 3	0.03	0.5	500
Nivel 4	0.1	0.5	500	Nivel 4	0.1	0.5	500
Nivel 5	0.1	0.5	2000	Nivel 5	0.1	2	1000

Los resultados muestran que un total de nueve modelos eligieron 1000 o menos iteraciones, cinco de las cuales fueron inferiores a 1000, lo que demuestra que la técnica de early stopping tiene sentido.

CAPÍTULO 6. Comparación de modelos

En este capítulo se explica la comparación de los modelos y la selección del modelo ganador en las dos patologías y los cinco niveles correspondientes, mediante una regla predeterminada de que el AUC tiene prioridad, seguido de la accuracy. Este no es un criterio absoluto y también se hace referencia a la variabilidad del AUC, donde un mayor grado de volatilidad indica que el AUC del modelo no es un resultado estable. Una función diseñada por Javier Portera (2021) puede ser de gran utilidad, siendo capaz de genera diagrama de caja, lo que facilita la comparación de modelos. En este caso, la métrica tasa de fallos es equivalente a 1- accuracy, entonces, cuando el valor de tasa de fallos sea menor, tiene más precisión este modelo.

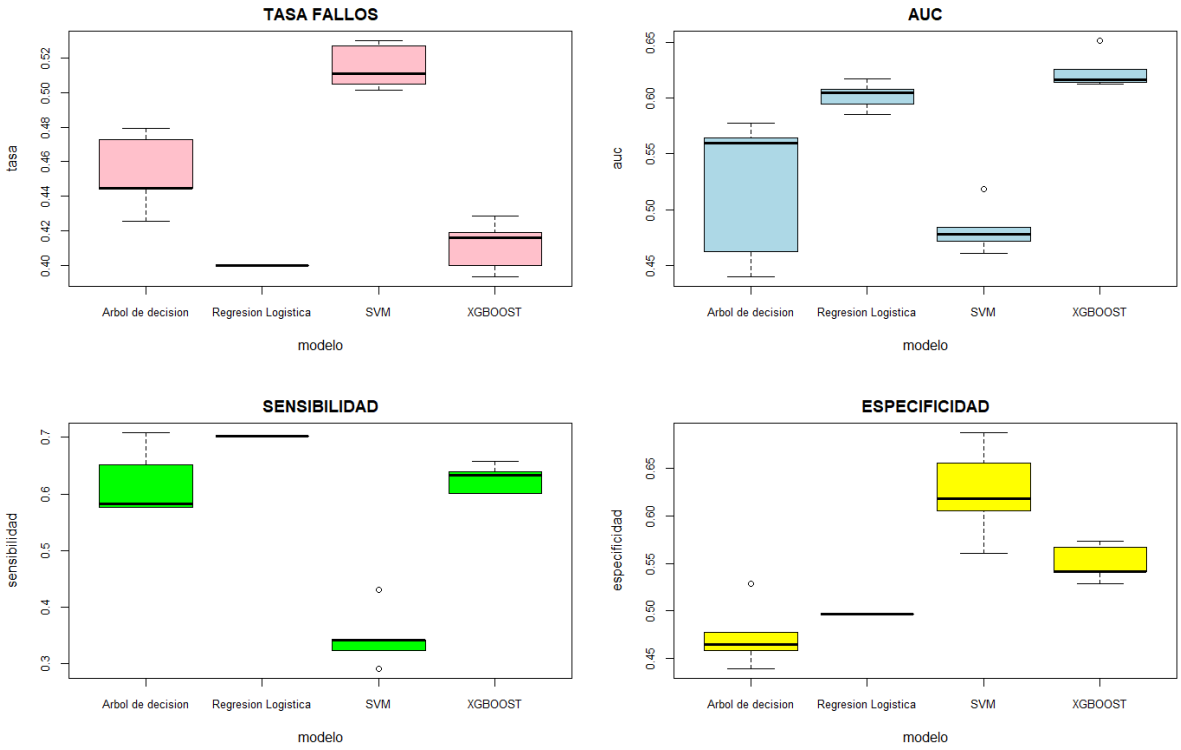
6.1. Evaluación de la Xerostomía usando el conjunto de validación

Para el problema de la xerostomía, se construyen cuatro clases de modelos, uno por uno, en cada uno de los cinco niveles de fármacos, con los parámetros de los modelos utilizando los valores óptimos que han sido ajustado, de manera que se puede comparar el rendimiento respectivo de cada modelo bajo diferentes criterios de evaluación en cada nivel.

Nivel 1

Aquí se puede ver que el modelo de regresión logística es el mejor modelo, debido a su alta precisión, valor AUC y sensibilidad. El siguiente mejor modelo es el modelo XGBoost, aunque tiene el AUC más alto, se puede notar que hay un valor atípico en el gráfico de caja sobre él y, por lo tanto, este modelo no se considera estable.

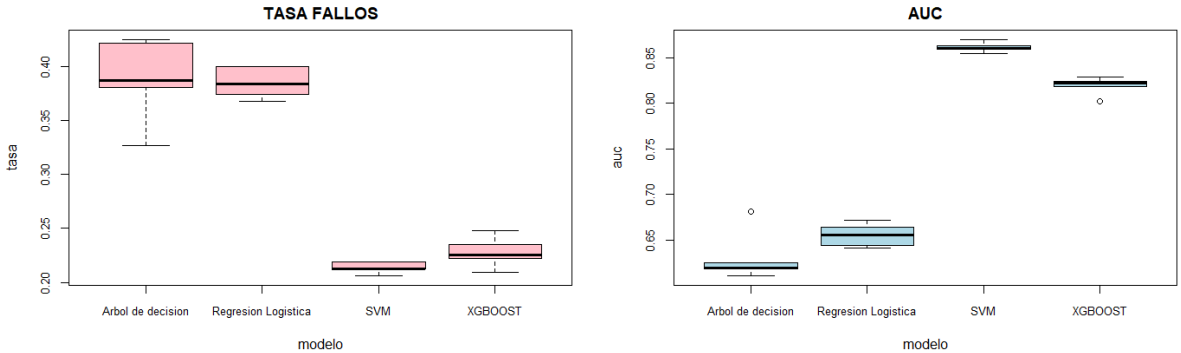
Figura 4. Comparación de modelos en nivel 1

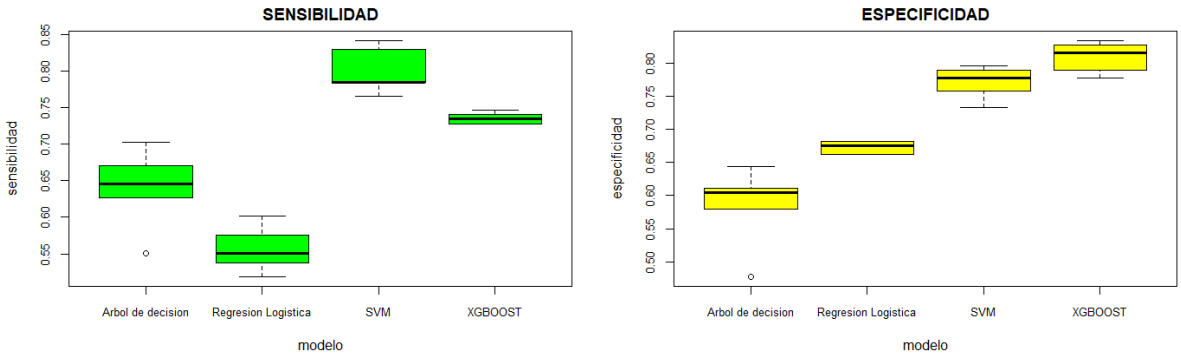


Nivel 2

En el nivel dos el SVM es el modelo más exitoso, con mejores valores para cada métrica. También se observa que la especificidad de XGBoost es ligeramente superior a los resultados obtenidos por SVM, lo que indica que el modelo es mejor para diagnosticar a los no pacientes que el modelo SVM.

Figura 5. Comparación de modelos en nivel 2

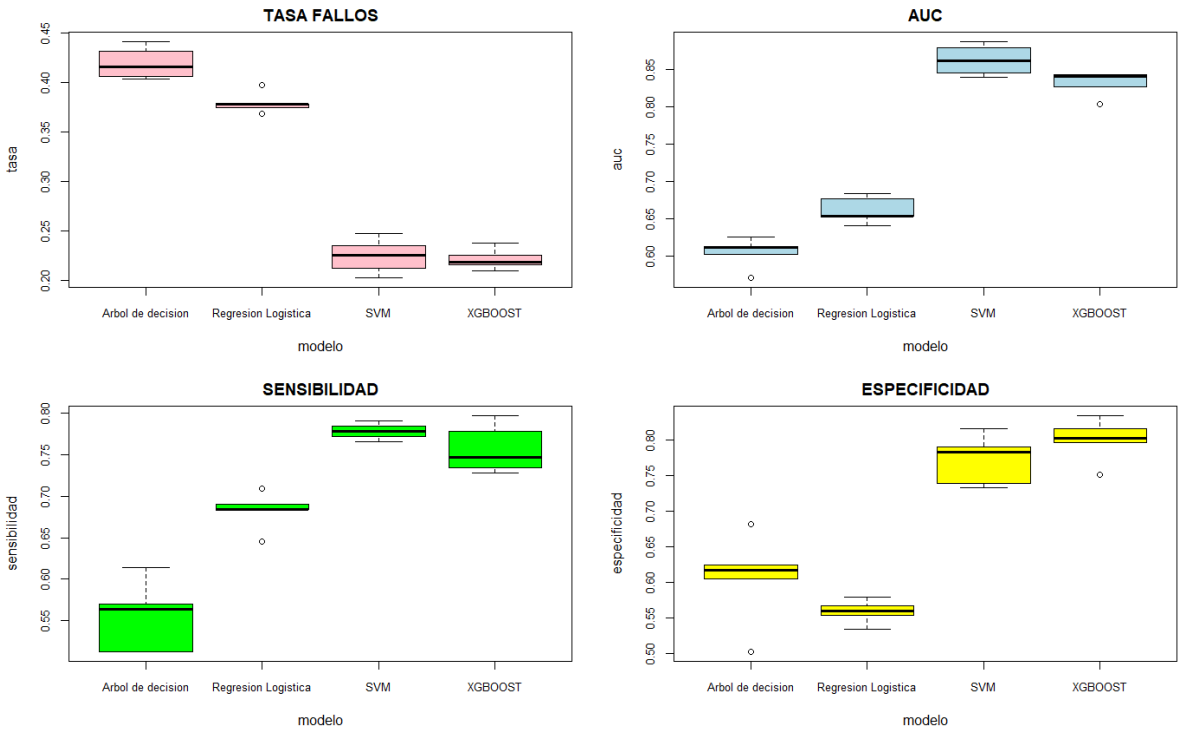




Nivel 3

En este nivel es difícil obtener un modelo ganador ya que la precisión de predicción de SVM y XGBoost son muy similares. SVM tiene un AUC más alto que XGBoost. Sin embargo, la variabilidad de la precisión de XGBoost es ligeramente mejor que la de SVM. Por otro lado, en cuanto a la capacidad de diagnosticar pacientes frente a no pacientes, SVM sobresale en la predicción de TP, proporcionando una mejor sensibilidad. XGBoost, por el contrario, presenta buenos resultados para la predicción de TN. Así pues, considerando que la sensibilidad y la especificidad son métricas que no se basan en la precisión, entonces, es más apropiado seleccionar el modelo ganador por los criterios de baja tasa de fallo y alta AUC. Por lo tanto, el modelo ganador es el SVM.

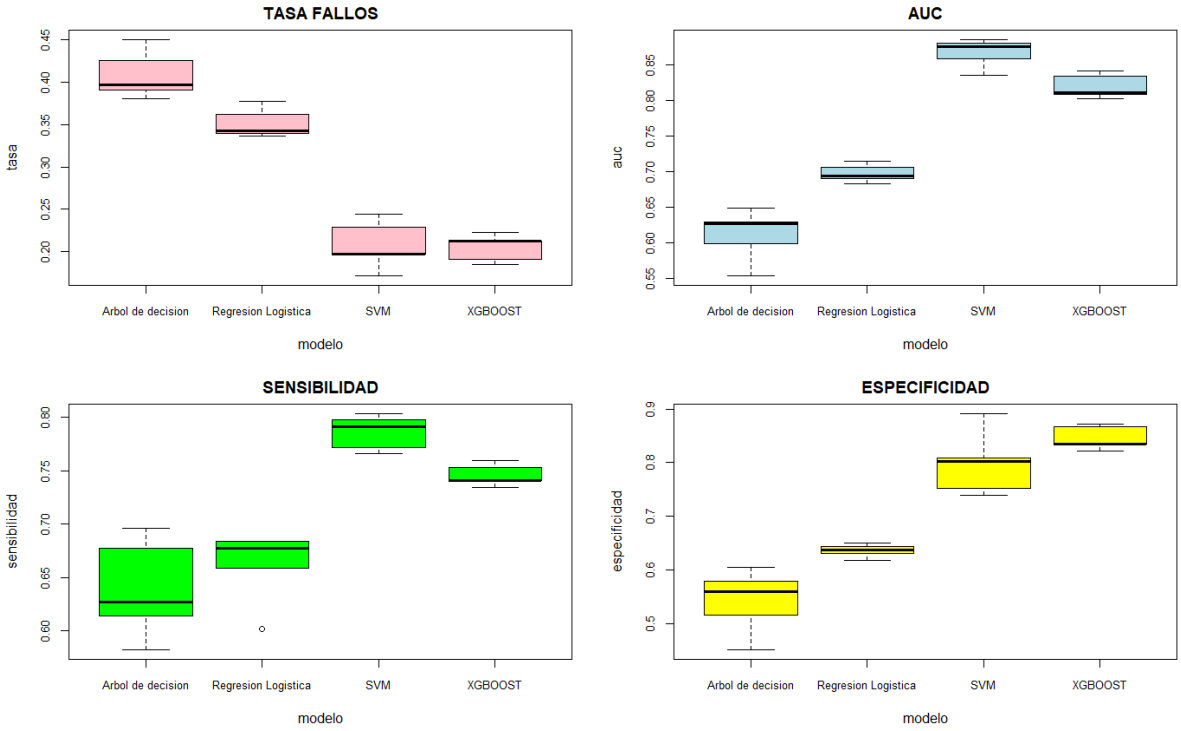
Figura 6. Comparación de modelos en nivel 3



Nivel 4

Al igual que en el nivel tres, se elige el SVM como modelo ganador, teniendo en cuenta el análisis anterior, que SVM tiene un mejor AUC y comporta una buena tasa de fallo, además, su sensibilidad también es excelente en comparación con otros modelos.

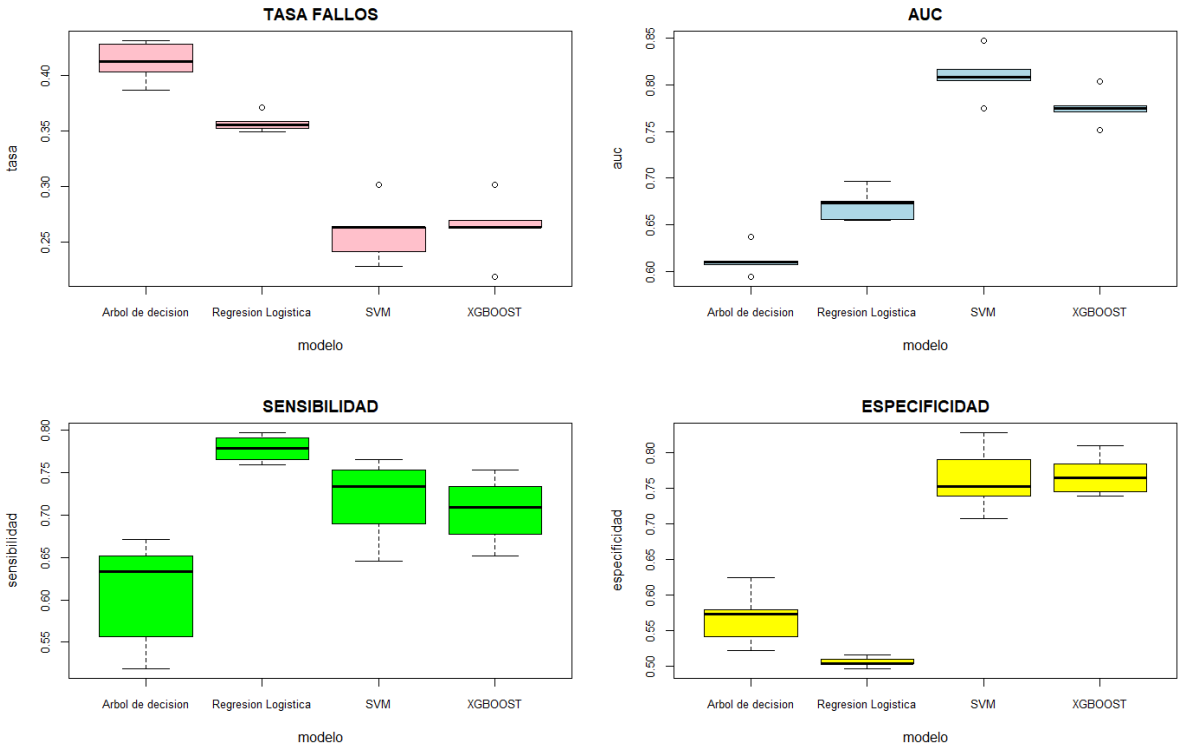
Figura 7. Comparación de modelos en nivel 4



Nivel 5

Usando razonamientos similares a los realizados en los niveles tres y cuatro, el modelo ganador en nivel cinco es el SVM.

Figura 8. Comparación de modelos en nivel 5



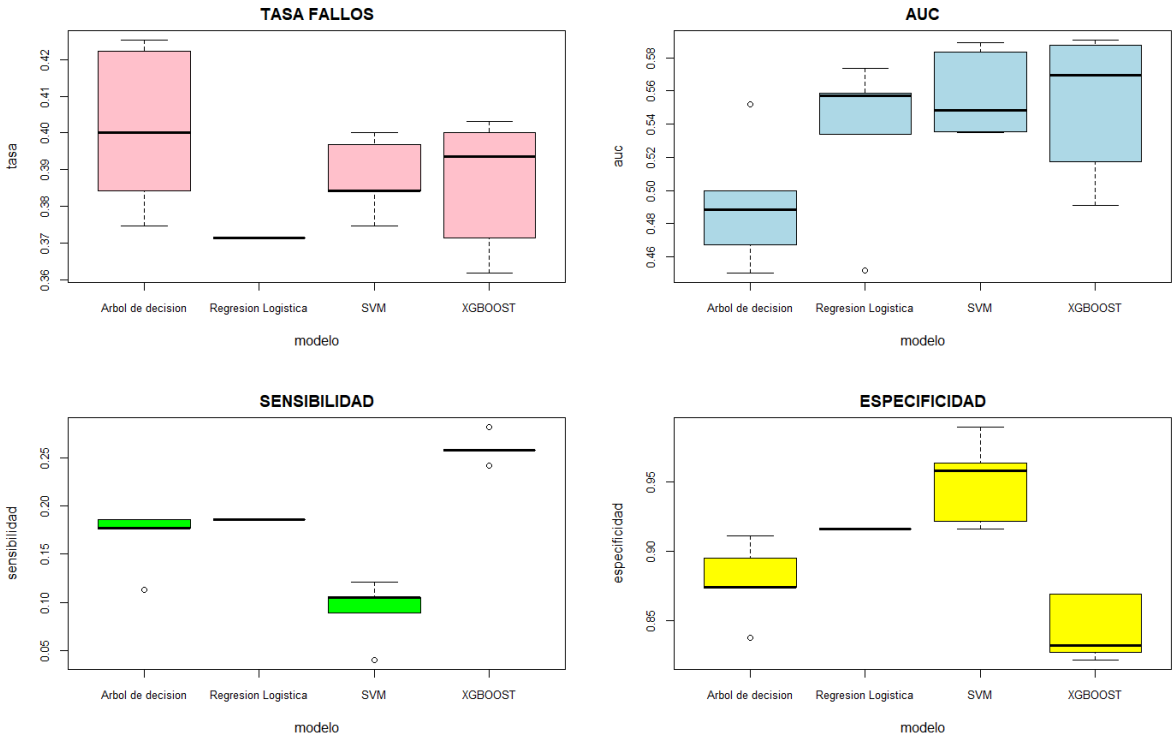
6.2. Evaluación de la Hiposalivación usando el conjunto de validación

En la comparación del modelo aquí, los modelos utilizan los parámetros óptimos solamente relaciona al problema de hiposalivación en cuestión, a continuación, se traza de diagramas de caja basados en los cinco resultados de cada modelo.

Nivel 1

En el caso de la hiposalivación, la regresión logística es el modelo ganador en el nivel uno, dado que presenta la median del AUC de regresión logística sitúa en la segunda mejor posición y tiene significativamente menos variabilidad del AUC que los demás modelos. Además, se observa una tasa de fallos más baja y no existencia de la variabilidad. A pesar de ser el mejor modelo local, la regresión logística tiene un par de defecto, ya que su poder (sensibilidad y especificidad) de diagnóstico para conocer los pacientes y no pacientes es débil. Y su capacidad de clasificación correcta no es fuerte dado que la tasa de acierto es solo de 63%.

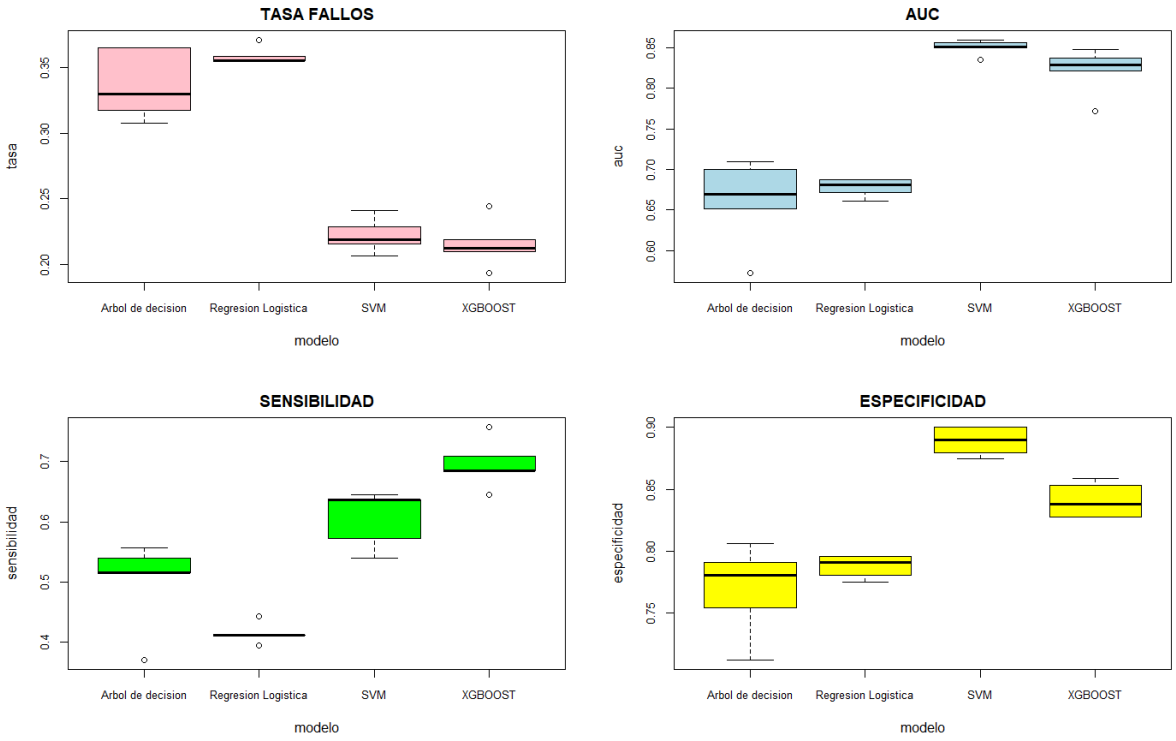
Figura 9. Comparación de modelos en nivel 1



Nivel 2

Se piensa que el SVM debería ser el modelo ganador, aunque el máximo AUC de SVM es el mismo que el de XGBoost, SVM tiene mejor estabilidad dado que su variabilidad de AUC es menor que la de XGBoost. Lo mismo ocurre en el caso de tasa de fallo. Pero existe un cambio en el diagnóstico para las muestras de pacientes y no pacientes, en el caso de xerostomía, SVM siempre tiene un mejor resultado de sensibilidad y en presente caso, SVM es el mejor modelo en la identificación de los no pacientes y XGBoost se centra en la predicción de las personas a punto de sufrir hiposalivación.

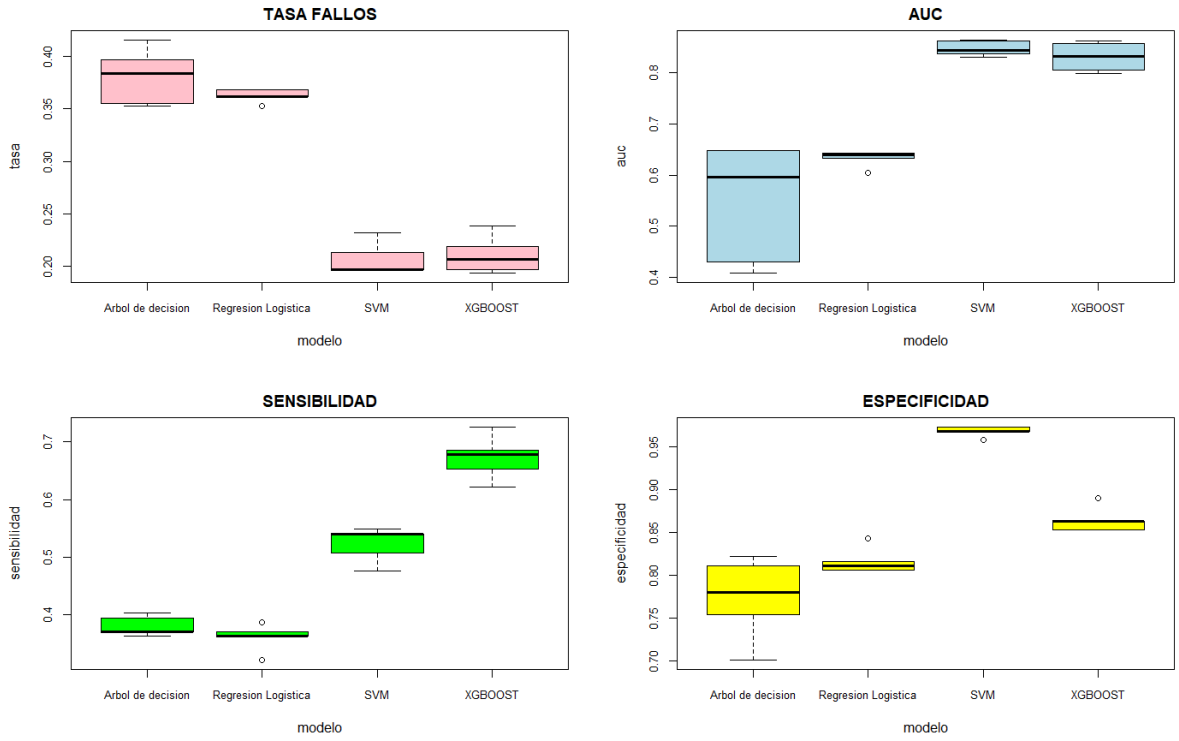
Figura 10. Comparación de modelos en nivel 2



Nivel 3

Tras observar los gráficos, el SVM supera al modelo XGBoost, dado que tiene mejor accuracy, valor de AUC y menor variabilidad. Además, el SVM tiene una especificidad muy alta en este nivel, superior a 0,95, lo que lo convierte en un modelo muy fiable para predecir que un individuo no sufrirá la hiposalivación.

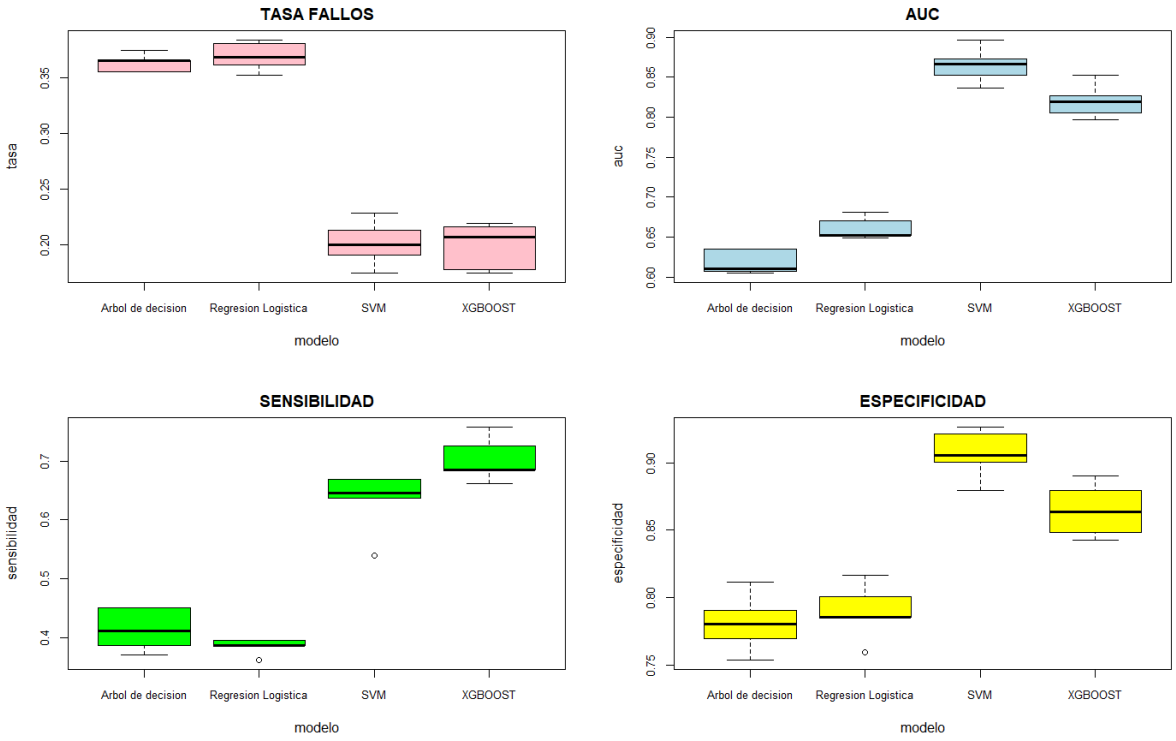
Figura 11. Comparación de modelos en nivel 3



Nivel 4

El comportamiento del SVM en este nivel es muy bueno dado que presenta un AUC de casi 0.9, siendo muy preciso (más del 90%) en la identificación de aquellos casos que no tienen riesgo de hiposalivación.

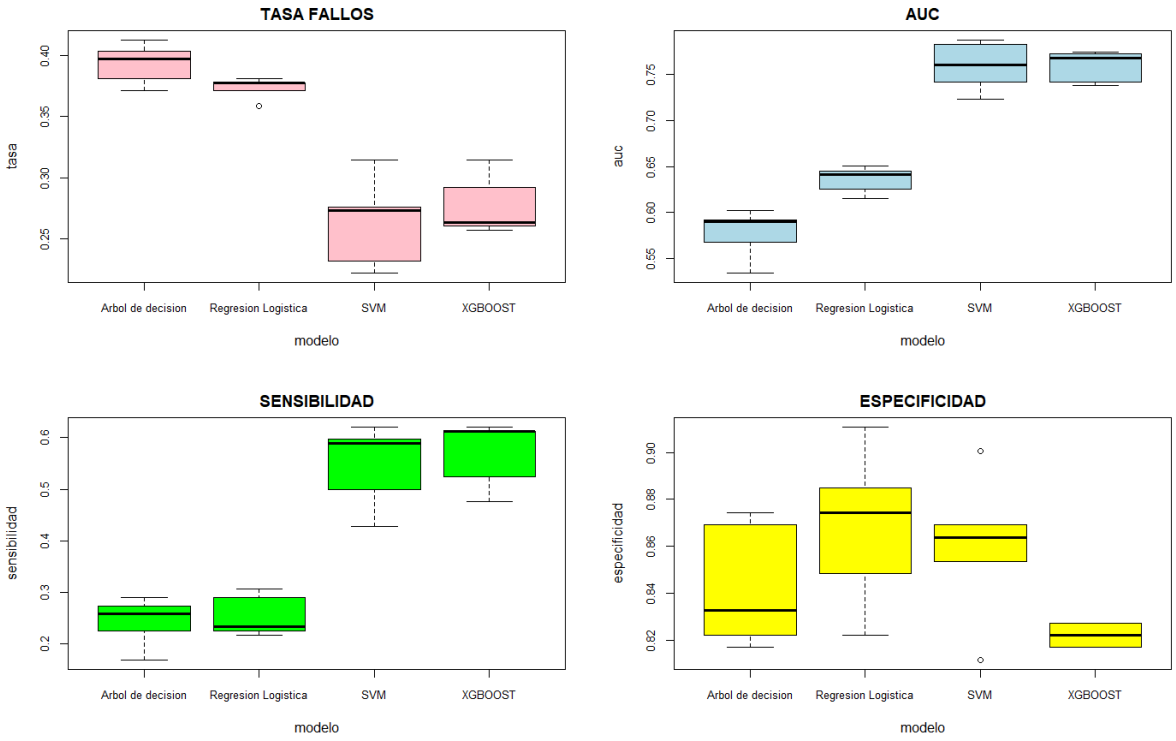
Figura 12. Comparación de modelos en nivel 4



Nivel 5

XGBoost presenta una clara fortaleza en el nivel cinco, tanto en términos de tasa de fallo como en los valores AUC como en su excelente variabilidad, que es completamente superior a los otros modelos.

Figura 13. Comparación de modelos en nivel 5



Las figuras y los análisis anteriores se obtuvieron al ejecutar los modelos usando el conjunto de validación sobre la base de que los modelos utilizaban los parámetros óptimos. En este sentido, observar:

- El análisis muestra que el modelo de árbol de decisión y el modelo de regresión logística sólo son adecuados para su uso en el primer nivel (el conjunto de datos del nivel uno tiene menos características y contiene menos información). Además, proporcionan un tiempo de ejecución más rápido y garantizan una buena precisión.
- A partir del segundo nivel, SVM y XGBoost se convierten en los mejores modelos de predicción, dado que proporcionan una mayor precisión y los resultados de la predicción son estables.

No basta con comparar y analizar los modelos sobre el conjunto de validación, y también es necesario estudiar los resultados de los modelos sobre un conjunto de pruebas. A continuación, se muestran los resultados del testeo de los modelos sobre el conjunto de pruebas para los dos casos de xerostomía e hiposalivación.

Tabla 8. Resultado del testeo sobre el conjunto de prueba para el caso de la Xerostomía

Nivel 1 ATC				
Algoritmos	**Accuracy**	**AUC**	**Sensibilidad**	**Especificidad**
Stepwize Regresión Logística	0.571	0.564	0.603	0.538

Árbol de Decisión	0.591	0.593	0.603	0.576
Support Vector Machine	0.638	0.640	0.509	0.769
XGBoost	0.657	0.716	0.698	0.614
Nivel 2 ATC				
Stepwize Regresión Logística	0.609	0.617	0.584	0.634
Árbol de Decisión	0.647	0.670	0.754	0.538
Support Vector Machine	0.761	0.844	0.830	0.692
XGBoost	0.771	0.863	0.773	0.769
Nivel 3 ATC				
Stepwize Regresión Logística	0.714	0.773	0.717	0.711
Árbol de Decisión	0.647	0.651	0.622	0.673
Support Vector Machine	0.752	0.839	0.738	0.769
XGBoost	0.752	0.859	0.811	0.692
Nivel 4 ATC				
Stepwize Regresión Logística	0.657	0.708	0.735	0.576
Árbol de Decisión	0.666	0.707	0.717	0.615
Support Vector Machine	0.771	0.821	0.735	0.807
XGBoost	0.761	0.859	0.735	0.788
Nivel 5 ATC				
Stepwize Regresión Logística	0.628	0.644	0.8113	0.442
Árbol de Decisión	0.609	0.643	0.773	0.442
Support Vector Machine	0.8	0.892	0.735	0.865
XGBoost	0.771	0.801	0.811	0.730

Table 9. Resultado del testeo del conjunto de prueba sobre el caso de la Hiposalivación

Nivel 1 ATC				
Algoritmos	**Accuracy**	**AUC**	**Sensibilidad**	**Especificidad**
Stepwize Regresión Logística	0.581	0.547	0.095	0.904
Árbol de Decisión	0.561	0.552	0.119	0.857
Support Vector Machine	0.591	0.508	0.071	0.936
XGBoost	0.542	0.503	0.214	0.761
Nivel 2 ATC				
Stepwize Regresión Logística	0.581	0.589	0.357	0.730
Árbol de Decisión	0.619	0.617	0.428	0.746
Support Vector Machine	0.771	0.796	0.667	0.841
XGBoost	0.752	0.807	0.714	0.777
Nivel 3 ATC				
Stepwize Regresión Logística	0.619	0.595	0.333	0.809
Árbol de Decisión	0.628	0.624	0.333	0.825

Support Vector Machine	0.866	0.852	0.666	1
XGBoost	0.809	0.871	0.738	0.857
Nivel 4 ATC				
Stepwize Regresión Logística	0.609	0.581	0.404	0.746
Árbol de Decisión	0.666	0.630	0.428	0.825
Support Vector Machine	0.885	0.896	0.833	0.921
XGBoost	0.857	0.911	0.833	0.087
Nivel 5 ATC				
Stepwize Regresión Logística	0.571	0.552	0.190	0.825
Árbol de Decisión	0.628	0.613	0.428	0.761
Support Vector Machine	0.8	0.815	0.761	0.825
XGBoost	0.781	0.837	0.667	0.857

Las conclusiones que se obtienen sobre el conjunto de pruebas son casi idénticas a las obtenidas sobre el conjunto de validación. Sin embargo, la selección del modelo ganador es diferente. En el caso de la Xerostomía como se presenta en tabla 8, el XGBoost sustituye al modelo SVM en el nivel dos, tres y cuatro como el modelo con mayor poder predictivo. A pesar de su tasa de fallo es muy similar, el AUC de XGBoost es ligeramente mejor. En cuanto al quinto nivel, tabla 9 muestra que el modelo SVM es óptimo cómo se comporta en el conjunto de validación, dado que se destaca el buen valor de AUC. Y en el caso de hiposalivación, el modelo ganador también es el XGBoost a partir del segundo nivel. Aunque el modelo SVM presenta en general una mayor precisión que XGBoost, sin embargo, XGBoost presenta un AUC más alto que los otros modelos. Teniendo en cuenta que el objetivo tiene una mayor necesidad de conocer el grado de riesgo de que una persona desarrolle xerostomía o hiposalivación, una mayor sensibilidad es más importante dada la proximidad de las dos AUC de los dos distintos modelos, por lo tanto, SVM debería ser el modelo más apropiado en el quinto nivel.

Por último, observar que comparando el cambio en los valores de AUC del modelo entre los conjuntos de validación y de prueba, se puede verificar que no hay un aumento o disminución significativa en el AUC, lo que demuestra la buena capacidad de generalización del modelo para hacer predicciones muy precisas incluso cuando se expone a nuevas muestras desconocidas.

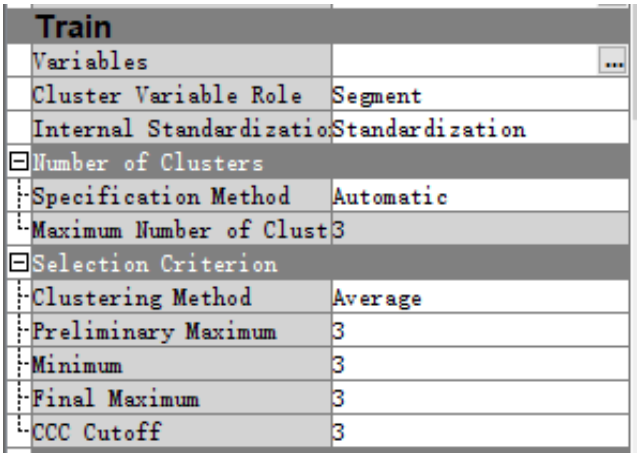
6.3. Análisis de clustering con herramienta SAS Enterprise Miner

En esta sección se va a crear una serie de perfiles con la ayuda de técnicas de clustering, para tratar de distinguir entre grupos que pertenecen a pacientes y grupos que no pertenecen a

pacientes, a continuación, describir y estudiar posibles factores patógenos en función de las características de ese grupo.

El programa informático del SAS ofrece tres métodos de agrupación (SAS Institute Inc, 2017), se elige el uso de Average Linkage para todos los datos de nivel uno a cinco, y sólo se generan tres perfiles al final. Una de las razones es que no había suficientes datos, y la segunda es que un gran número de perfiles no sería útil para analizar características patológicas comunes y la ingesta de fármacos. Por lo tanto, se utilizarán los siguientes ajustes de parámetros en SAS como se muestra en figura 14, fijando el número preliminar máximo y mínimo para el número de clusters, para que sólo tenga tres perfiles el resultado.

Figura 14. Configuración de parámetros para Clustering jerárquico en SAS Enterprise Miner



Una vez calculados los clusters, los resultados se resumen en la Tabla A1 – Tabla A10 del anexo, donde la columna del código de los fármacos es el porcentaje de la población que utiliza el fármaco correspondiente en esa categoría, considerándose prevalente en más del 50% en ese perfil. Además, para facilitar la comprensión, cada perfil está etiquetado según el porcentaje de personas que presentan la enfermedad. Así, en el caso de la xerostomía, se considera que un grupo no tiene riesgo de desarrollarla si el porcentaje de personas que la padecen es inferior al 40%. Si es superior al 60%, existe una probabilidad alta de sufrir la enfermedad. Y si el porcentaje se encuentra entre el 40% y el 60%, se denominan no definidos, son ambiguos y es no es posible analizarlos. Los mismos razonamientos se han aplicado al problema de la hiposalivación, salvo algunos ajustes que se han realizado en los umbrales para tener en cuenta la especificidad de la proporción de muestras positivas y negativas de 4:6. Así, del 60% se reduce al 55% el límite derecho, es decir, se considera que un grupo tiene hiposalivación cuando la prevalencia es superior al 55%.

A. Xerostomía

Para realizar las agrupaciones en caso de xerostomía en diferentes niveles de fármacos, se plantea construir el modelo de cluster jerárquico mediante un flujo de trabajo en SAS Enterprise Miner (véase en la figura A25 de anexo).

Nive1

Hay dos etiquetas sin padecer xerostomía, pero no hay importancia analítica porque el número de personas en esta categoría es demasiado reducido, solo hay 12 personas y 29 personas en los dos grupos, respectivamente.

Nivel 2

Hay una variabilidad significativa en fármaco A11, con un gran número de casos de enfermedad en el grupo que no toma y ninguno en el grupo que sí lo toma. Es decir, el fármaco A11 probablemente ofrece protección. Sin embargo, este fármaco no se mencionó en la tabla 4, al ser excluida la influencia por el modelo de regresión logística y el modelo de árbol de decisión.

Nivel 3

La administración de C09C y N02A es muy destacada aquí. El grupo que toma estos fármacos no sufre xerostomía, entonces, ambos los dos fármacos comportan buena protección, pero se observa que C09C es consistente con los resultados de la tabla 4, y se considera que N02A contribuye a la xerostomía.

Nivel 4

El C10AA se utiliza comúnmente en los grupos dos y tres, y ambos son etiquetados como enfermos, por lo que la posibilidad de un efecto xerostomizante puede existir para el C10AA (aunque no se encuentra en la tabla 4).

Nivel 5

La distribución del consumo de medicamentos en esta clase es demasiado fragmentada para ofrecer un análisis valioso.

B. Hiposalivación

El flujo de trabajo para generar los perfiles de hiposalivación se puede encontrar en la figura A26 de anexo, se ha modelado en cada uno de los niveles independientes.

Nivel 1

Hay dos grupos con etiqueta de no definido y otro grupo con muy pocos elementos que no dan un análisis, en dicho grupo solo hay 10 observaciones.

Nivel 2

Para la hiposalivación A11 no estaba teóricamente asociada o proporcionaba protección, ya que el grupo que tomaba y el que no tomaba el fármaco estaba etiquetado como no enfermo. Más del 50% de los miembros del grupo enfermo toman B01, mientras que su consumo en los otros grupos sólo es un pequeño número de personas y sin coger este problema, se supone que B01 tiene efecto de reducción de salivación.

Nivel 3

Debido a los dos grupos no definidos y la falta de apoyo de datos en el grupo restante, no se puede derivar un análisis sólido. Dado que la tabla A8 de anexo muestra que, en el único grupo sin enfermedad, ninguno de los consumos de drogas es significativo para indicar su asociación con el problema de hiposialia.

Nivel 4

El C09AA se toma comúnmente en el cluster dos, más del 60%, y no se toma en el grupo tres. Ambos están etiquetados como sin riesgo de hiposalivación, lo que permite deducir que presenta un efecto protector del fármaco o que éste no tiene un efecto directo, en consonancia con los resultados de la tabla 4.

Nivel 5

En el cluster 2 el uso de C09DB01 y C09DB02 es del 100% y 90,7% respectivamente, lo que infiriere protección. No obstante, no está presente en la tabla 4, lo que puede estar relacionado con el bajo número de usuarios.

Los resultados de la clustering jerárquico fueron exitosos en algunos niveles, ya que adicionalmente dan ciertos factores farmacológicos que pueden estar relacionados con la xerostomía y la hiposalivación, aunque no están registrados en la tabla 4 y 5.

CAPÍTULO 7. Discusión

7.1. Discusión

El objetivo principal de este trabajo fue explorar qué fármacos se asociaban a dos patologías orales relativamente frecuentes como son la xerostomía y la hiposalivación, así como predecir el riesgo de que una persona sufra este problema al consumir diferentes medicamentos. En este trabajo se proponen cuatro modelos para lograr los objetivos definidos en apartados anteriores. Al interpretar los resultados de la Regresión Logística y el Árboles de Decisión podemos ver como ciertos fármacos influyen en dichas patologías como se puede observar en la Tabla 4 y la Tabla 5 que resumen dichos resultados. Debido a las deficiencias del conjunto de datos utilizados para el trabajo (que utilizaba en su totalidad una clasificación binaria, y que además recogía un número limitado de observaciones) las predicciones de estos dos modelos de Regresión Logística y Árbol de Decisión no cumplieron las expectativas. Por ello, se utilizaron también los modelos Support Vector Machine y Extreme Gradiente Boosting para intentar obtener aún mejores predicciones.

En base a las métricas de evaluación dadas por los modelos en el conjunto de pruebas, no se pueden hacer predicciones para el primer nivel de fármacos en ninguna de las dos patologías, ya que los resultados fueron malos. En los niveles del segundo a cuarto, XGBoost fue el modelo ganador, pero no ofreció ventajas absolutas sobre SVM dado que su tasa de fallo fue similar y no fue tampoco mejor en términos de sensibilidad y especificidad. Se recomienda utilizar XGBoost para predecir la xerostomía y la hiposalivación dada la prioridad del AUC como criterio de evaluación, siendo el modelo líder en AUC en cada categoría.

En el nivel cinco, SVM fue el modelo más adecuado para ambas patologías, aunque en hiposalivación XGBoost obtuvo un mejor AUC, pero su sensibilidad y especificidad no están tan bien equilibradas como las de SVM.

Los factores dados por los análisis de la regresión logística y del árbol de decisión para la enfermedad de la xerostomía y la hiposalivación son razonables. Hay suficientes estudios para justificar los fármacos identificados y su dirección de la acción en este trabajo (Pérez Espinosa et al., 2016; Ramírez Martínez-Acitores et al., 2020; Cappetta et al., 2018; Wolf et al., 2017). El presente estudio destaca los siguientes resultados, en primer lugar, existe una proporción de fármacos que reducen la producción de saliva, aumentando la sensación de boca seca. En nuestro caso no se observó como los bloqueantes de canales de calcio se asociaban a xersotimía, lo que contradice los resultados del estudio de Nonzee y cols. (2012). También, se observó

como algunos fármacos se asocian a la xerostomía, pero que no reducían el flujo salival. Por último, también los resultados muestran como existen fármacos con efecto protector.

Con respecto a los fármacos que tienen efecto xerostomizante, deben evitarse al prescribir medicamentos siempre que sea posible, cuando el paciente se queja ya de sequedad bucal. A la inversa, los fármacos que tienen la capacidad de suprimir los problemas salivales o tienen cierto efecto protector, pueden estar indicados en aquellos pacientes que ya sufren xerostomía para sustituir, siempre que estén indicados, a aquellos que reducen el flujo salival o aumentan el riesgo de xerostomía.

Por otro lado, se utilizó el clustering jerárquico para obtener una segmentación de estas personas observadas, agrupándolas en tres perfiles en función de los comportamientos similares. La aplicación del clustering compensa algunas de las deficiencias del modelo predictivo, por ejemplo, el C09CA no se ha agrupado como medicamento protector en la Tabla 3, pero según los resultados de la Tabla A4, la mayoría de las personas que tomaron el medicamento fueron clasificados en el grupo de no xerostomía, lo que indica de forma indirecta un efecto protector. Según el estudio de Masajtis et al. (2009) el uso de losartán (C09CA01) podría reducir la aparición de xerostomía en pacientes hipertensos. Estos resultados están en línea con los del presente trabajo. No obstante, es importante subrayar que el análisis de clustering no puede inferir casualidad y que su comprensión es subjetiva. Este análisis podría proporcionar una orientación terapéutica para los profesionales sanitarios en un futuro.

7.2. Limitaciones

En este estudio existen ciertas limitaciones, en primer lugar, la capacidad de clasificación del modelo explicativo es débil, de acuerdo con las evidencias de los comportamientos de modelo en conjunto de prueba, presenta que los modelos de regresión logística tienen un mejor resultado de tasa de acierto de 0,714 en diferentes cinco clases de medicamentos y en las dos patologías, y su mejor AUC es ligeramente superior, de 0,773 (se encuentra en el caso de xerostomía, Nivel 3 de tabla 8). Sobre árbol de decisión, tiene una mejor tasa de acierto de 0,666 y mejor AUC de 0,707 (se encuentra en el caso de xerostomía, Nivel 4 de tabla 8). Es decir, no basta con uso de los factores detectados por la regresión logística y los árboles de decisión para cuantificar las probabilidades de que un individuo esté a punto de desarrollar xerostomía o hiposalivación. En segundo lugar, se encuentra en el análisis factorial de ciertos fármacos derivado del modelo en este trabajo es inconsistente con los hallazgos de los estudios médicos, por ejemplo, enalapril (C09AA02) en caso de xerostomía, los antidepresivos (N06A) y los Bloqueantes de receptores de angiotensina II (C09CA) en el caso de hiposalivación. Sobre

todo, los cálculos del modelo se basan en los datos recogidos y en funciones matemáticos, por lo que las muestras utilizadas en este trabajo pueden diferir de las utilizadas en otros estudios, y la diversidad del consumo de drogas da lugar a resultados de clasificación inversos.

La regresión logística y los árboles de decisión dan interpretaciones controvertidas de ciertos fármacos, dando resultados protectores en un modelo e induciendo xerostomía o reduciendo el flujo salival en otro modelo. Por ejemplo, el fármaco de psicolépticos (N05) se asocia a hiposalivación en regresión logística, pero obtiene el efecto contrario en el modelo árbol de decisión. Otro escenario es que un modelo da una explicación para la variable y otro modelo no la considera como significativa, por lo tanto, esta variable no fue seleccionado ni apareció en resultados. Algunos ejemplos de estos fármacos son los psicoanalépticos (N06), antidepresivos (N06A) y derivados de la benzodiazepina (N05BA), etc.

En conjunto, la novedad de este trabajo es la presentación sistemática de la dirección de los efectos de los fármacos sobre la xerostomía y la hiposalivación. Asimismo, dos modelos predictivos adicionales de SVM y XGBoost tienen como objetivo acelerar la obtención de una estimación del riesgo de una persona de padecer estas dos enfermedades con la información disponible sobre el consumo de drogas.

CAPÍTULO 8. Conclusiones y líneas de trabajo futuro

8.1. Conclusiones

En este trabajo se ha realizado un estudio sobre la influencia de algunos fármacos sobre la xerostomía y la hiposalivación, llegándose a las siguientes conclusiones:

A. Xerostomía:

- Los factores que influyen en la xerostomía en el primer nivel de fármacos no son importantes.
- En segundo nivel, los medicamentos de la clase de diuréticos (C03), agentes betabloqueantes (C07) y psicoanalépticos (N06) son asociados a la sensación de boca seca. los agentes antitrombóticos (B01), los bloqueantes de canales de calcio (C08) y los agentes que actúan sobre el sistema renina-angiotensina (C09) son fármacos que suelen tener un efecto protector y pueden reducir el riesgo de desarrollar la enfermedad.
- En tercer nivel, los grupos de fármacos de bloqueantes selectivos de canales de calcio con efectos principalmente vasculares (C08A), inhibidores de la ECA de monoterapia (C09A) y bloqueantes de receptores de angiotensina II (BRA) de monofármacos tienen protección significativa. Los opioides (N02A) y antidepresivos (N06A) que van a aumentar la probabilidad de la enfermedad.
- En cuarto nivel, los inhibidores de la agregación plaquetaria sin heparina (B01AC), derivados de la dihidropiridina (C08CA) e inhibidores de la ECA de monodrogas (C09AA) presentan buena capacidad protectora. A la inversa, los agentes betabloqueantes selectivos (C07AB) e inhibidores selectivos de la recaptación de serotonina (N06AB) provocan la xerostomía.
- En quinto nivel, los fármacos específicos de ácido acetilsalicílico (B01AC06), amlodipino (C08CA01) y calcio carbonato o colecalciferol (A12AXP1) son protecciones para xerostomía. El acenocumarol (B01AA07) se asocia a sequedad bucal.

B. Hiposalivación:

- Los factores que influyen en la hiposalivación en el primer nivel de fármacos no son importantes.

- En segundo nivel, existen tres grupos de fármacos que pueden disminuir la producción de saliva, los agentes para el tratamiento de alteraciones causadas por ácidos (A02), fármacos usados en diabetes (A10) y agentes betabloqueantes (C07), respectivamente. Los bloqueantes de canales de calcio (C08) y los agentes modificadores de los lípidos (C10) tuvieron un efecto protector.
- En tercer nivel, los inhibidores de la ECA de monoterapia (C09A) y los antagonistas de angiotensina II de combinaciones (C09D) se asociaron a la afección protectora. Los agentes betabloqueantes (C07A) y los ansiolíticos (N05B) son factores que causan la reducción de saliva.
- En cuarto nivel, tres categorías de fármacos que tienen efecto protector son: derivados de la dihidropiridina (C08CA), inhibidores de la HMG-COA reductasa (C10AA) y derivados de la benzodiazepina (N05BA). Y los únicos inhibidores de la bomba de protones aumentan el riesgo de hiposalivación.
- En quinto nivel, también son tres fármacos los que tienen un efecto protector: paracetamol (N02BE01), amlodipino (C08CA01) y simvastatina (C10AA01). El uso de omeprazol (A02BC01) se asocia a una reducción del flujo salival.

C. Modelos

- Se considera que, del segundo al cuarto nivel de fármacos, se recomienda el uso de XGBoost como modelo predictivo para estimar la probabilidad de que un individuo desarrolle xerostomía o hiposalivación, de forma que no sólo se garantice una predicción correcta, sino también tenga una mayor sensibilidad. Cuando se trata del quinto nivel, se debe utilizar la SVM para consolidar un alto grado de precisión y sensibilidad.
- La técnica de jerárquico clustering es una buena herramienta para distinguir los diferentes aspectos de los grupos enfermos y no enfermos en este estudio, pero no proporciona la agrupación ideal.

8.2. Líneas de trabajo futuro.

El análisis realizado se puede mejorar en varios sentidos, de manera que se pueden definir varias líneas de trabajo futuro:

- Expandir el tamaño muestral: Se interesa en recopilar más muestra médica de forma que el modelo explicativo pueda tener más información y datos sobre el consumo de

fármacos, lo que permitirá que el modelo ofrezca una explicación más convincente de las variables. Se plantea recoger más datos relativos a los fármacos junto con los resultados de cuestionarios de xerostomía y recogida de flujo salival. Para ello, se buscará la cooperación de múltiples centros.

- Mejorar la precisión predictiva: Con respecto al aumento de la precisión de la predicción de la patología, se pueden utilizar métodos más avanzados, y la viabilidad del modelo stacking ensemble modelo (Charoenkwan, et al., 2021). El stacking funciona utilizando múltiples clasificadores primarios para hacer predicciones y asignando diferentes pesos para obtener la respuesta predicha final. También existe alternativa, utilizando la votación para determinar la clasificación de la salida final. Por lo tanto, se puede diseñar un modelo para corregir la precisión de la siguiente manera:

$$y_{pred} = 0.75 * Pred.XGBoost + 0.25 * Pred.SVM$$

- Crear mejores perfiles: Hay que intentar utilizar la técnica de K-mode clustering (Papachristou et al., 2018) para identificar de subgrupos de xerostomía y hiposalivación con varios perfiles, ya que las variables del conjunto de datos son todas valores discretos.
- Obtener explicación más fiable: Los modelos predictivos avanzados tienen fortaleza de alta precisión, sin embargo, es muy difícil interpretar las variables como influir la respuesta, se puede implementar alguna técnica especial como DALEX para estudiar los cambios internos de modelos.

Bibliografía

Abdullah, M. J. (2015). Prevalence of xerostomia in patients attending Shorish dental speciality in Sulaimani city. *Journal of Clinical and Experimental Dentistry*, 7, e45-e53. https://doi.org/10.4317/ jced.51867

Aliko, A., Wolff, A., Dawes, C., Aframian, D., Proctor, G., Ekström, J., Narayana, N., Villa, A., Sia, Y. W., Joshi, R. K., McGowan, R., Beier Jensen, S., Kerr, A. R., Lynge Pedersen, A. M., & Vissink, A. (2015). World Workshop on Oral Medicine VI, clinical implications of medication-induced salivary gland dysfunction. *Oral Surgery Oral Medicine Oral Pathology Oral Radiology*, 120, 185–206. https://doi. org/10.1016/j.oooo.2014.10.027

ATC de la OMS, (2021). *WHO Collaborating Centre for Drug Statistics Methodology, Norwegian Institute of Public Health.* https://www.whocc.no/atc_ddd_index/

Calviño, A. (2021). Técnicas y Metodología de la Minería de Datos (SEMMA). *Material didáctico de master en minería de datos e inteligencia de negocios, UCM*

Cappetta, K., Beyer, C., Johnson, J. A., & Bloch, M. H. (2018). Meta-analysis: Risk of dry mouth with second generation antidepressants. *Progress in neuro-psychopharmacology & biological psychiatry*, *84*(Pt A), 282–293. https://doi.org/10.1016/j.pnpbp.2017.12.012

Carramolino-Cuéllar, E., Lauritano, D., Silvestre, F. J., Carinci, F., Lucchese, A., & Silvestre-Rangil, J. (2018). Salivary flow and xerostomia in patients with type 2 diabetes. *Journal of Oral Pathology & Medicine*, 47, 526–530. https://doi.org/10.1111/jop.12712

Carvalho, H. N., Dos Santos, Y. L., Bernardino, Í. M., de Lima, K. C., Granville-Garcia, A. F., & Melo de Brito Costa, E. M. (2020). Accuracy of a questionnaire on xerostomia as a screening tool for hyposalivation. *International dental journal*, 70(6), pp. 427–434. https://doi.org/10.1111/idj.12586

Chao, M., El Naqa, I., Bakst, R. L., Lo, Y. C., & Peñagarícano, J. A. (2022). Cluster model incorporating heterogeneous dose distribution of partial parotid irradiation for radiotherapy induced xerostomia prediction with machine learning methods. *Acta oncologica (Stockholm, Sweden)*, *61*(7), 842–848. https://doi.org/10.1080/0284186X.2022.2073187

Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M. M., Manavalan, B., & Shoombuatong, W. (2021). StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings in bioinformatics*, *22*(6), bbab172. https://doi.org/10.1093/bib/bbab172

Código ATC. (2021). Wikipedia, *La enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=C%C3%B3digo_ATC&oldid=138571760.

Dalodom, S., Lam-ubol, A., Jeanmaneechotechai, S., Takamfoo, L., Intachai, W., Duangchada, K., Hongsachum, B., Kanjanatiwat, P., Vacharotayangul, P., & Trachootham, D. (2016). Influence of oral moisturizing jelly as a saliva substitute for the relief of xerostomia in elderly patients with hypertension and diabetes mellitus. *Geriatric Nursing*, 37, 101e–109e. https://doi.org/10.1016/j.gerin urse.2015.10.014

Daniel, G. G. (2021). Métodos Emsemble. *Material didáctico de master en minería de datos y inteligencia de negocios, UCM.*

Enrique, G. (2004). Medicamentos esenciales. *Farmacia Profesional*, 18(8), pp. 6-11.

Gareth, J., Daniela, W., Trevor, H. & Robert, T. (2013). An introduction to statistical learning: with applications in R. *Spinger*. https://link.springer.com/book/10.1007/978-1-4614-7138-7

Han, P., Suarez-Durall, P., & Mulligan, R. (2015). Dry mouth: a critical topic for older adult patients. J*ournal of prosthodontic research*, 59(1), 6–19. https://doi.org/10.1016/j.jpor.2014.11.001

Huang, J. & Ling, C.X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), pp. 299-310. Doi: 10.1109/TKDE.2005.50.

Ivanovski, K., Pesevka, S., Ristoska, S., Dirjanska, K., Mindova, S., Pandilova, M., & Eftimoska, M. (2015). The impact of antihypertensive medications on quantitative and qualitative characteristics of saliva. *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, 6, 1356–1364.

Liu, B., Dion, M. R., Jurasic, M. M., Gibson, G., & Jones, J. A. (2012). Xerostomia and salivary hypofunction in vulnerable elders: prevalence and etiology. *Oral surgery, oral medicine, oral pathology and oral radiology*, 114(1), 52–60. https://doi.org/10.1016/j.oooo.2011.11.014

Looström, H., Akerman, S., Ericson, D., Tobin, G. y Götrick, B. (2011). Tramadol-induced oral dryness and pilocarpine treatment: effects on total protein and IgA. *Archives of oral biology*, 56(4), pp. 395–400. https://doi.org/10.1016/j.archoralbio.2010.10.019

López-Pintor, R. M., Martínez-Acitores, L. R., Valle, J. S., González-Serrano, J., Casañas, Arriba, de L., Hernández, G. (2022). Xerostomia and Hyposalivation. *Oral Health and Aging*. https://doi.org/10.1007/978-3-030-85993-0_5

Masajtis-Zagajewska, A., & Nowicki, M. (2009). Influence of dual blockade of the renin-

angiotensin system on thirst in hemodialysis patients. *Nephron. Clinical practice*, 112(4), pp. 242–247. https://doi.org/10.1159/000224790

Matsumoto, N., Ushikoshi-Nakayama, R., Yamazaki, T. et al. (2020). What Are the Major Causes of Dry Mouth in Elderly Adults? *Curr Oral Health Reports*, pp. 165–167. https://doi.org/10.1007/s40496-020-00262-6

Men, K., Geng, H., Zhong, H., Fan, Y., Lin, A., & Xiao, Y. (2019). A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial. *International journal of radiation oncology, biology, physics*, *105*(2), 440–447. https://doi.org/10.1016/j.ijrobp.2019.06.009

Nederfors, T., Nauntofte, B., & Twetman, S. (2004). Effects of furosemide and bendroflumethiazide on saliva flow rate and composition. *Archives of Oral Biology*, 49, 507–513. https://doi.org/10.1016/j.archoralbio.2004.01.007

Ng, A. (2017). *Lecture 12.5 — Support Vector Machines | (Kernels-II)* [Video]. Youtube. https://www.youtube.com/watch?v=XfyR_49hfi8&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&index=74

Nonzee, V., Manopatanakul, S., & Khovidhunkit, S. O. (2012). Xerostomia, hyposalivation and oral microbiota in patients using antihypertensive medications. *Journal of the Medical Association of Thailand*, 95, 96–104.

Ostermann, G., Brisgand, B., Schmitt, J. & Fillastre, J. P. (1988). Efficacy and acceptability of rilmenidine for mild to moderate systemic hypertension. *The American journal of cardiology*, 61(7), pp. 76–80. https://doi.org/10.1016/0002-9149(88)90470-5

Papachristou, N., Barnaghi, P., Cooper, B. A., Hu, X., Maguire, R., Apostolidis, K., Armes, J., Conley, Y. P., Hammer, M., Katsaragakis, S., Kober, K. M., Levine, J. D., McCann, L., Patiraki, E., Paul, S. M., Ream, E., Wright, F., & Miaskowski, C. (2018). Congruence Between Latent Class and K-Modes Analyses in the Identification of Oncology Patients With Distinct Symptom Experiences. *Journal of pain and symptom management*, *55*(2), 318–333.e4. https://doi.org/10.1016/j.jpainsymman.2017.08.020

Pereira, L. J., Foureaux, C., Periera, C. V., Alves, M. C., Campos, C. H., Rodrigues Garcia, M., & Andrade, E. F. (2016). Oral physiology, nutrition and quality of life in diabetic patients associated or not with hypertension and beta-blockers therapy. *Journal of Oral Rehabilitation*, 43(7), 511–518. https://doi.org/10.1111/joor.12398

Pérez Espinosa, Y., Ureña Espinosa, M., Rodríguez González, Y., Bosch Utra, K., & Portelles Morales, T. (2016). Xerostomía causada por el consumo de diuréticos en pacientes hipertensos. *Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta*, *41*(10).

http://revzoilomarinello.sld.cu/index.php/zmv/article/view/944

Portela, J. (2021). Cruzada 60 árbol binaria, XGBoosting binaria, SVM binaria, logística binaria. *Material didáctico de master en minería de datos ye inteligencia de negocios, UCM.*

Prasanthi, B., Kannan, N., & Patil, R. (2014). Effect of diuretics on salivary flow, composition and oral health status, A clinico-biochemical study. *The Annals of Medical and Health Sciences Research*, 4, 549-553. https://doi.org/10.4103/2141-9248.139311

Ramírez Martínez Acitores, L., Hernández Ruiz de Azcarate, F., Casañas, E., Serrano, J., Hernández, G., & López Pintor, R. M. (2020). Xerostomia and salivary flow in patients taking antihypertensive drugs. *International Journal of Environmental Research and Public Health*, 17, 2478. https://doi.org/10.3390/ijerph1707247

Ramírez, L., Sánchez, I., Muñoz, M., Martínez-Acitores, M. L., Garrido, E., Hernández, G., & López-Pintor, R. M. (2021). Risk factors associated with xerostomia and reduced salivary flow in hypertensive patients. *Oral diseases*, 10.1111/odi.14090. Advance online publication. https://doi.org/10.1111/odi.14090

Rhodus, N. L., & Brown, J. (1990). The association of xerostomia and inadequate intake in older adults. *Journal of the American Dietetic Association*, *90*(12), 1688–1692.

Saleh, J., Figueiredo, M. A. Z., Cherubini, K., & Salum, F. G. (2015). Salivary hypofunction, An update on aetiology, diagnosis and therapeutics. *Archives of Oral Biology*, 60, 242–255. https://doi.org/10.1016/j. archoralbio.2014.10.004

SAS Institute Inc, (2017). Introduction to SEMMA. *SAS® Enterprise Miner™ 14.3: Reference Help.*
https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

SAS Institute Inc, (2017). Cluster Node. *SAS® Enterprise Miner™ 14.3: Reference Help.*
https://documentation.sas.com/doc/en/emref/14.3/n1vjatb74dundbn12d2ecb09juak.htm

Sreebny, L. M., & Schwartz, S. S. (1997). A reference guide to drugs and dry mouth--2nd edition. *Gerodontology*, 14(1), pp. 33–47. https://doi.org/10.1111/j.1741-2358.1997.00033.x

Tahrir, N. N., & Aldelaimi, B. D. S. (2006). The effect of atenolol (B-blocker) on salivary composition in patients with essential hypertension. *Journal of Baghdad College of Dentistry*, 18, 57–59.

Thomson, W. M., Chalmers, J. M., Spencer, A. J., Slade, G. D., & Carter, K. D. (2006). A longitudinal study of medication exposure and xerostomia among older people. *Gerodontology*, 23(4), 205–213. https:// doi.org/10.1111/j.1741-2358.2006.00135.x

Villa, A., Connell, C. L., & Abati, S. (2014). Diagnosis and management of xerostomia and hyposalivation. *Therapeutics and Clinical Risk Management*, 11, 45–51. https://doi.org/10.2147/TCRM. S76282

What is minimum child weight in XGBoost? (2019). *MullOverThing.* https://mulloverthing.com/what-is-minimum-child-weight-in-xgboost/#What_is_minimum_child_weight_in_XGBoost

Wolff, A., Joshi, R. K., Ekström, J., Aframian, D., Pedersen, A. M., Proctor, G., Narayana, N., Villa, A., Sia, Y. W., Aliko, A., McGowan, R., Kerr, A. R., Jensen, S. B., Vissink, A., & Dawes, C. (2017). A Guide to Medications Inducing Salivary Gland Dysfunction, Xerostomia, and Subjective Sialorrhea: A Systematic Review Sponsored by the World Workshop on Oral Medicine VI. *Drugs in R&D*, *17*(1), 1–28. https://doi.org/10.1007/s40268-016-0153-9

Yusniyanti,A. L., Virgantari, F., & Faridhan, Y. E. (2021). Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators *Journal of Physics: Conference Series.* https://iopscience.iop.org/article/10.1088/1742-6596/1863/1/012071/meta

Zheng, J. H. (1995). A Dictionary of Statistical Theory and Practice. *China Statistics Press*.

Anexo I: Tablas y gráficas de resultados

Tabla A1. Resumen de Clustering Jerárquico asociado a la Xerostomía en Nivel 1

Cluster	Recuento	Xerostomía	Xerostomía prop.	Sistema A	Sistema B	Sistema D
1	12	No	16.7%	100.0%	8.3%	16.7%
2	379	Indefinido	52.2%	58.6%	24.1%	100.0%
3	29	No	37.9%	65.7%	35.9%	0.0%

	Sistema G	Sistema L	Sistema M	Sistema N	Sistema R	Sistema S
1	0.0%	100.0%	41.7%	33.3%	0.0%	16.7%
2	24.1%	0.0%	17.2%	41.4%	24.1%	24.1%
3	16.4%	0.0%	15.0%	58.0%	15.3%	6.3%

Tabla A2. Resumen de Clustering Jerárquico asociado a la Xerostomía en Nivel 2

Cluster	Recuento	Xerostomía	Xerostomía prop.	A02	A10	A11
1	251	Indefinido	43.8%	39.4%	28.3%	0.0%
2	65	No	40.0%	55.4%	26.2%	55.4%
3	104	Yes	72.1%	57.7%	25.0%	0.0%

	A12	B01	C03	C07	C08	C09
1	16.7%	22.3%	12.4%	31.5%	13.5%	79.3%
2	20.0%	30.8%	9.2%	9.2%	10.8%	75.4%
3	3.8%	47.1%	40.4%	17.3%	19.2%	85.6%

	C10	G04	N02	N05	N06	R03
1	58.6%	9.6%	24.7%	22.7%	0.8%	0.0%
2	58.5%	12.3%	41.5%	23.1%	18.5%	0.0%
3	57.7%	28.8%	43.3%	34.6%	42.3%	36.5%

	S01
1	0.0%
2	46.2%
3	2.9%

Tabla A3. Resumen del Clustering Jerárquico asociado a la Xerostomía en Nivel 3

Cluster	Recuento	Xerostomía	Xerostomía prop.	A02B	A10B	A12A
1	85	No	37.6%	40.0%	30.6%	27.1%
2	136	Indefinido	43.4%	55.9%	44.9%	5.1%
3	199	Yes	60.3%	39.7%	14.1%	11.1%

	B01A	C07A	C08C	C09A	C09B	C09C
1	30.6%	5.9%	7.1%	18.8%	7.1%	60.0%
2	50.7%	38.2%	30.1%	8.1%	44.9%	4.4%
3	15.1%	9.0%	2.0%	37.7%	1.0%	0.0%

	C09D	C10A	G04C	N02A	N02B	N05B
1	8.2%	62.4%	4.7%	51.8%	55.3%	42.4%
2	24.3%	58.1%	31.6%	1.5%	9.6%	10.3%
3	30.2%	49.7%	5.0%	0.0%	31.2%	27.1%

	N06A
1	10.6%
2	8.1%
3	18.6%

Tabla A4. Resumen de Clustering Jerárquico asociado a la Xerostomía en Nivel 4

Cluster	Recuento	Xerostomía	Xerostomía prop.	A02BC	A10BA	A10BD
1	202	No	24.3%	39.6%	1.0%	2.5%
2	128	Yes	78.9%	47.7%	34.4%	0.0%
3	90	Yes	67.8%	43.3%	0.0%	43.3%

	B01AC	C07AB	C08CA	C09AA	C09BA	C09CA
1	14.4%	2.0%	10.4%	29.2%	10.9%	24.3%
2	15.6%	5.5%	10.9%	25.0%	23.4%	6.3%
3	26.7%	34.4%	17.8%	12.2%	7.8%	0.0%

	C09DA	C09DB	C10AA	G04CA	N02BB	N02BE
1	11.4%	3.0%	39.6%	1.5%	13.4%	34.7%
2	18.0%	2.3%	71.1%	34.4%	3.1%	7.8%
3	16.7%	53.3%	61.1%	6.7%	5.6%	25.6%

	N05BA	N06AB

1	20.3%	3.0%
2	24.2%	21.1%
3	35.6%	5.6%

Tabla A5. Resumen de Clustering Jerárquico asociado a la Xerostomía en Nivel 5

Cluster	Recuento	Xerostomía	Xerostomía prop.	A02BC01	A10BA02	A12AXP1
1	129	No	23.3%	53.5%	1.6%	30.2%
2	235	Yes	61.3%	19.1%	14.9%	1.3%
3	56	Yes	66.1%	33.9%	14.3%	0.0%

	B01AA07	B01AC06	C08CA01	C09AA02	C09AA03	C09BA03
1	0.0%	19.4%	18.6%	3.1%	16.3%	19.4%
2	1.3%	14.5%	1.3%	12.3%	10.6%	5.1%
3	64.3%	16.1%	10.7%	8.9%	0.0%	3.6%

	C09DB01	C09DB02	C10AA01	C10AA05	D09DB06	G04CA02
1	0.0%	0.0%	22.5%	23.3%	0.0%	0.0%
2	24.3%	20.9%	25.1%	23.4%	20.4%	1.3%
3	0.0%	0.0%	8.9%	30.4%	0.0%	46.4%

	N02BB02	N02BE01	N05BA06	N05BA08
1	14.7%	51.2%	15.5%	13.2%
2	4.3%	11.9%	6.8%	6.8%
3	12.5%	16.1%	5.4%	16.1%

Tabla A6. Resumen de Clustering Jerárquico asociado a la Hiposalivación en Nivel 1

Cluster	Recuento	Hiposalivación	Hiposalivación prop.	Sistema A	Sistema B	Sistema D
1	12	Indefinido	41.7%	100.0%	8.3%	16.7%
2	10	No	0.0%	30.0%	60.0%	90.0%
3	398	Indefinido	40.5%	66.1%	34.4%	5.0%

	Sistema G	Sistema L	Sistema M	Sistema N	Sistema R	Sistema S
1	0.0%	100.0%	41.7%	33.3%	0.0%	16.7%
2	80.0%	0.0%	50.0%	0.0%	70.0%	80.0%
3	15.3%	0.0%	14.3%	58.3%	14.6%	5.8%

Tabla A7. Resumen de Clustering Jerárquico asociado a la Hiposalivación en Nivel 2

Cluster	Recuento	Hiposalivación	Hiposalivación prop.	A02	A10	A11
1	218	No	23.4%	34.9%	4.1%	0.0%
2	166	Yes	59.6%	60.8%	59.0%	0.0%
3	36	No	44.4%	50.0%	19.4%	100.0%

	A12	B01	C03	C07	C08	C09
1	20.2%	13.8%	11.5%	7.8%	16.5%	88.5%
2	4.8%	53.6%	29.5%	49.4%	14.5%	69.9%
3	19.4%	16.7%	13.9%	11.1%	2.8%	77.8%

	C10	G04	N02	N05	N06	R03
1	51.4%	8.3%	39.9%	17.4%	2.3%	13.8%
2	71.1%	25.3%	18.1%	38.6%	27.7%	4.8%
3	41.7%	5.6%	47.2%	16.7%	19.4%	0.0%

	S01
1	7.8%
2	9.0%
3	2.8%

Tabla A8. Resumen de Clustering Jerárquico asociado a la Hiposalivación en Nivel 3

Cluster	Recuento	Hiposalivación	Hiposalivación prop.	A02B	A10B	A12A
1	109	Indefinido	51.4%	46.8%	37.6%	17.4%
2	191	No	26.2%	37.2%	6.3%	14.1%
3	120	Indefinido	50.0%	55.8%	51.7%	5.0%

	B01A	C07A	C08C	C09A	C09B	C09C
1	25.7%	10.1%	7.3%	14.7%	4.6%	46.8%
2	13.6%	0.5%	5.8%	38.2%	15.7%	0.0%
3	59.2%	52.5%	26.7%	10.8%	28.3%	5.0%

	C09D	C10A	G04C	N02A	N02B	N05B
1	13.8%	66.1%	3.7%	27.5%	42.2%	64.2%
2	23.0%	45.0%	3.7%	7.3%	37.2%	8.4%
3	34.2%	60.8%	38.3%	1.7%	4.2%	15.0%

	N06A
1	27.5%
2	5.8%
3	13.3%

Tabla A9. Resumen de Clustering Jerárquico asociado a la Hiposalivación en Nivel 4

Cluster	Recuento	Hiposalivación	Hiposalivación prop.	A02BC	A10BA	A10BD
1	175	Indefinido	48.0%	52.0%	19.4%	12.0%
2	146	No	39.0%	38.4%	2.1%	6.2%
3	99	No	25.3%	33.3%	9.1%	14.1%

	B01AC	C07AB	C08CA	C09AA	C09BA	C09CA
1	30.9%	9.1%	19.4%	1.1%	32.6%	26.9%
2	6.2%	10.3%	5.5%	68.5%	0.0%	0.0%
3	10.1%	11.1%	9.1%	0.0%	2.0%	10.1%

	C09DA	C09DB	C10AA	G04CA	N02BB	N02BE
1	10.3%	1.7%	61.7%	25.1%	1.1%	16.0%
2	0.0%	0.0%	48.6%	4.1%	4.1%	23.3%
3	43.4%	54.5%	47.5%	3.0%	28.3%	41.4%

	N05BA	N06AB
1	22.3%	13.7%
2	18.5%	7.5%
3	38.4%	3.0%

Tabla A10. Resumen de Clustering Jerárquico asociado a la Hiposalivación en Nivel 5

Cluster	Recuento	Hiposalivación	Hiposalivación prop.	A02BC01	A10BA02
1	56	Indefinido	44.6%	33.9%	14.3%
2	54	No	35.2%	33.3%	9.3%
3	310	No	39.4%	31.0%	10.3%

	A12AXP1	B01AA07	B01AC06	C08CA01	C09AA02
1	0.0%	64.3%	16.1%	10.7%	8.9%
2	5.6%	5.6%	18.5%	3.7%	0.0%

3	12.6%	0.0%	15.8%	8.1%	10.6%

	C09AA03	**C09BA03**	**C09DB01**	**C09DB02**	**C10AA01**
1	0.0%	3.6%	0.0%	0.0%	8.9%
2	0.0%	0.0%	100.0%	90.7%	22.2%
3	14.8%	11.9%	1.0%	0.0%	24.5%

	C10AA05	**D09DB06**	**G04CA02**	**N02BB02**	**N02BE01**
1	30.4%	0.0%	46.4%	12.5%	16.1%
2	22.2%	88.9%	5.6%	13.0%	35.2%
3	23.5%	0.0%	0.0%	7.1%	24.2%

	N05BA06	**N05BA08**
1	5.4%	16.1%
2	7.4%	14.8%
3	10.3%	8.1%

Figura A1. Coeficiente de SLR asociada a Xerostomía en Nivel 1

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1989     0.1940   1.025   0.3051
SistemaD      1.1910     0.5478   2.174   0.0297 *
SistemaR     -0.7254     0.3447  -2.104   0.0354 *
SistemaL      1.3051     0.8030   1.625   0.1041
SistemaN     -0.3572     0.2371  -1.507   0.1319
```

Figura A2. Coeficiente de SLR asociada a Xerostomía en Nivel 2

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3506     0.1801   1.947  0.05154 .
A10          -0.4898     0.2827  -1.732  0.08325 .
B01           0.7348     0.2990   2.457  0.01400 *
C03          -1.0414     0.3349  -3.110  0.00187 **
C07          -0.7373     0.3238  -2.277  0.02279 *
C08           0.7389     0.3441   2.147  0.03178 *
N06          -1.0113     0.3744  -2.701  0.00691 **
R03          -0.9980     0.4626  -2.157  0.03098 *
```

Figura A3. Coeficiente de SLR asociada a Xerostomía en Nivel 3

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3461     0.1882  -1.839  0.06586 .
A12A          0.8481     0.3992   2.125  0.03362 *
B01A          0.4846     0.3048   1.590  0.11190
C07A         -0.5525     0.3670  -1.505  0.13226
C08C          1.1592     0.3762   3.082  0.00206 **
C09A          0.6777     0.2956   2.292  0.02188 *
C09C          1.4648     0.4521   3.240  0.00119 **
N02A         -1.1206     0.4345  -2.579  0.00991 **
N06A         -1.2126     0.3885  -3.121  0.00180 **
```

Figura A4. Coeficiente de SLR asociada a Xerostomía en Nivel 4

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3843     0.2104  -1.827 0.067770 .
A02BC        -0.3839     0.2585  -1.485 0.137536
B01AC         1.8010     0.4080   4.414 1.01e-05 ***
C07AB        -1.6203     0.4890  -3.313 0.000922 ***
C08CA         1.3592     0.3971   3.423 0.000619 ***
C09AA         0.7072     0.2933   2.411 0.015898 *
C09CA         1.0645     0.4421   2.408 0.016060 *
N02BB         0.7555     0.4038   1.871 0.061361 .
N06AB        -1.4096     0.4934  -2.857 0.004278 **
```

Figura A5. Coeficiente de SLR asociada a Xerostomía en Nivel 5

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2318     0.1656  -1.399 0.161671
B01AC06       1.3198     0.3563   3.704 0.000212 ***
B01AA07      -1.0871     0.4501  -2.415 0.015715 *
C09AA02      -1.0425     0.4633  -2.250 0.024426 *
C08CA01       0.9674     0.4522   2.140 0.032395 *
A12AXP1       0.8479     0.4532   1.871 0.061327 .
N02BB02       0.9530     0.4194   2.272 0.023082 *
C09DB02     -16.8672   739.3669  -0.023 0.981799
D09DB06      16.5469   739.3669   0.022 0.982145
```

Figura A6. Coeficiente de SLR asociada a Hiposalivación en Nivel 1

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4327     0.1329   3.256  0.00113 **
SistemaG     -0.6386     0.3113  -2.051  0.04025 *
SistemaR      0.8525     0.3705   2.301  0.02139 *
```

Figura A7. Coeficiente de SLR asociada a Hiposalivación en Nivel 2

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8261     0.2754   3.000 0.002703 **
A02          -0.4366     0.2619  -1.667 0.095498 .
A10          -0.6384     0.2939  -2.172 0.029819 *
A11          -0.6954     0.4450  -1.563 0.118145
C03          -0.6412     0.3188  -2.011 0.044311 *
C07          -1.0807     0.3037  -3.558 0.000373 ***
C08           0.8237     0.3955   2.083 0.037253 *
C10           0.6095     0.2659   2.293 0.021859 *
N02           0.4408     0.2880   1.531 0.125800
N05          -0.6365     0.2830  -2.249 0.024512 *
```

Figura A8. Coeficiente de SLR asociada a Hiposalivación en Nivel 3

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3715     0.3332  -1.115  0.26489
A10B         -0.4213     0.2812  -1.498  0.13408
C07A         -0.8126     0.3397  -2.392  0.01674 *
C08C          0.6513     0.3867   1.684  0.09213 .
C09A          0.8754     0.3631   2.411  0.01591 *
C09B          1.0317     0.3995   2.582  0.00981 **
C09C          0.6524     0.4491   1.453  0.14630
C09D          0.8272     0.3549   2.331  0.01977 *
C10A          0.6509     0.2600   2.503  0.01231 *
N02B          0.6187     0.2923   2.117  0.03427 *
N05B         -0.7741     0.2857  -2.710  0.00673 **
```

Figura A9. Coeficiente de SLR asociada a Hiposalivación en Nivel 4

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2391     0.2801  -0.854  0.39336
A02BC        -0.4423     0.2654  -1.666  0.09565 .
B01AC         0.6528     0.3615   1.806  0.07097 .
C07AB        -1.3610     0.4586  -2.968  0.00300 **
C08CA         0.7139     0.4033   1.770  0.07670 .
C09AA         0.8468     0.3270   2.589  0.00961 **
C09BA         0.8208     0.3936   2.085  0.03705 *
C09DB         0.6794     0.3838   1.770  0.07671 .
C10AA         0.7403     0.2598   2.849  0.00439 **
N02BE         0.8935     0.3215   2.779  0.00545 **
N05BA        -0.4672     0.3017  -1.549  0.12145
N06AB        -0.8275     0.4548  -1.820  0.06882 .
```

Figura A10. Coeficiente de SLR asociada a Hiposalivación en Nivel 5

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1052     0.1947   0.541  0.58878
N02BE01       0.9891     0.3067   3.225  0.00126 **
B01AA07      -0.9647     0.4007  -2.408  0.01605 *
C08CA01       1.0288     0.4779   2.153  0.03134 *
A02BC01      -0.4461     0.2648  -1.684  0.09209 .
C10AA01       0.5877     0.3091   1.901  0.05728 .
C10AA05       0.5537     0.3003   1.844  0.06518 .
```

Figura A11. Visualización de árbol de decisión asociada a Xerostomía en Nivel 1

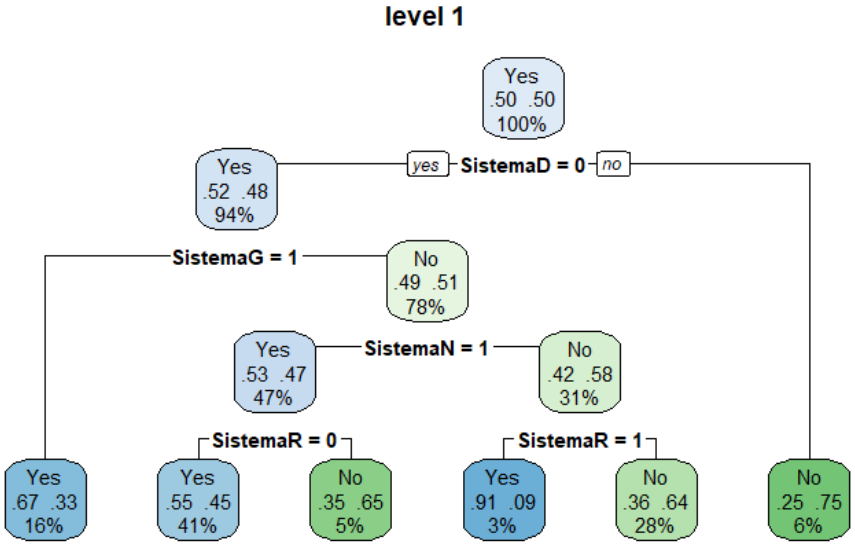


Figura A12. Visualización de árbol de decisión asociada a Xerostomía en Nivel 2

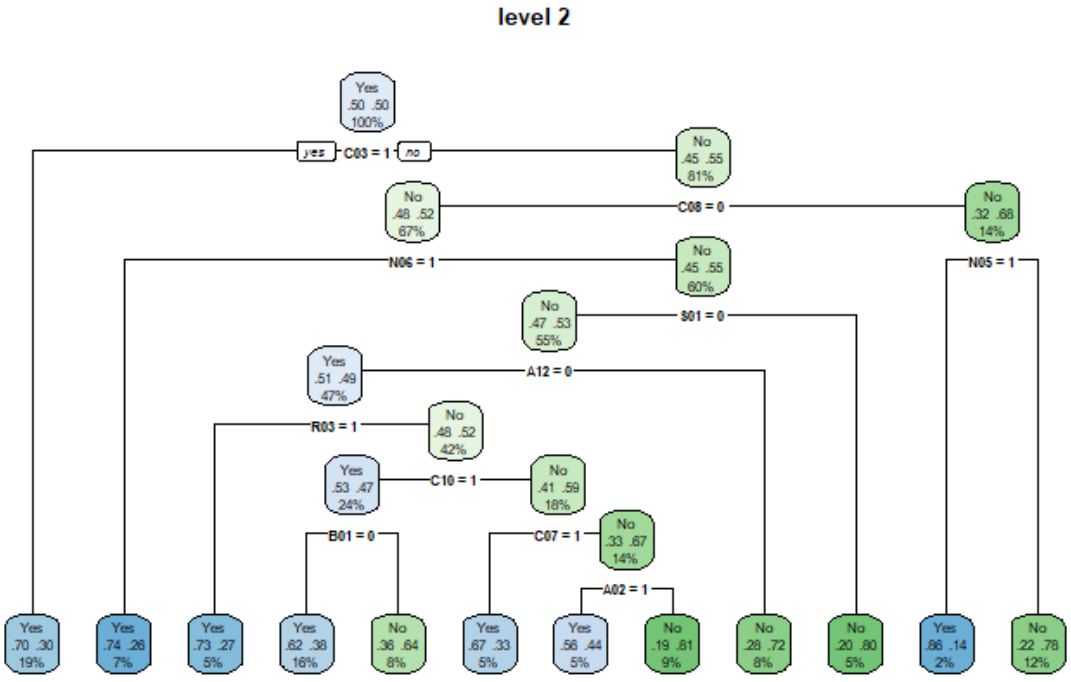


Figura A13. Visualización de árbol de decisión asociada a Xerostomía en Nivel 3

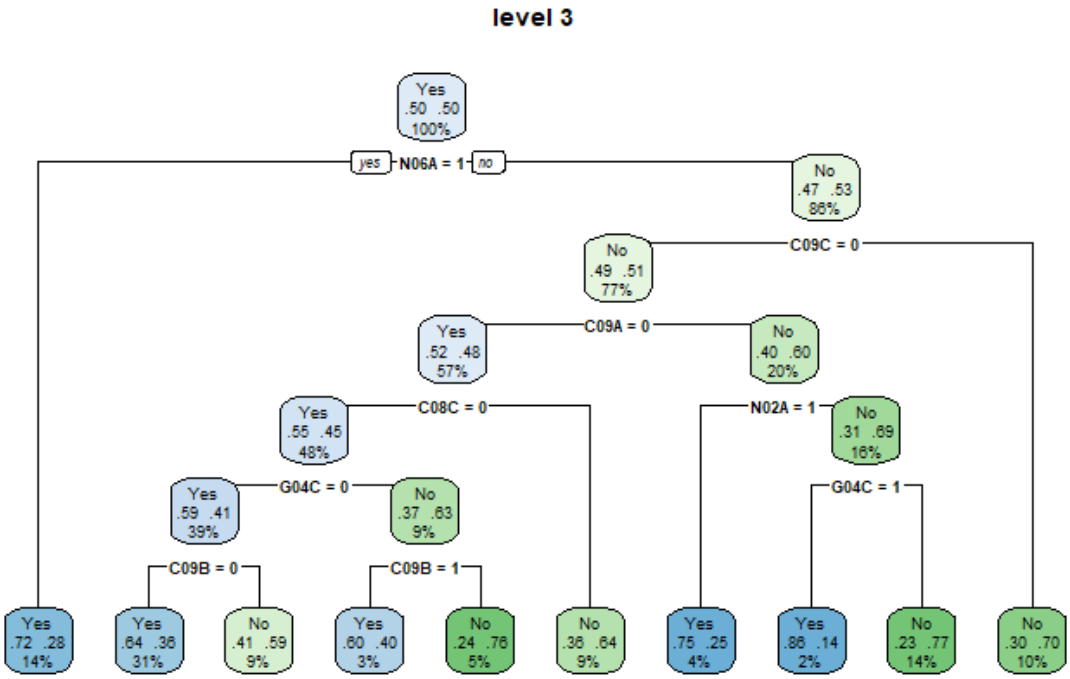


Figura A14. Visualización de árbol de decisión asociada a Xerostomía en Nivel 4

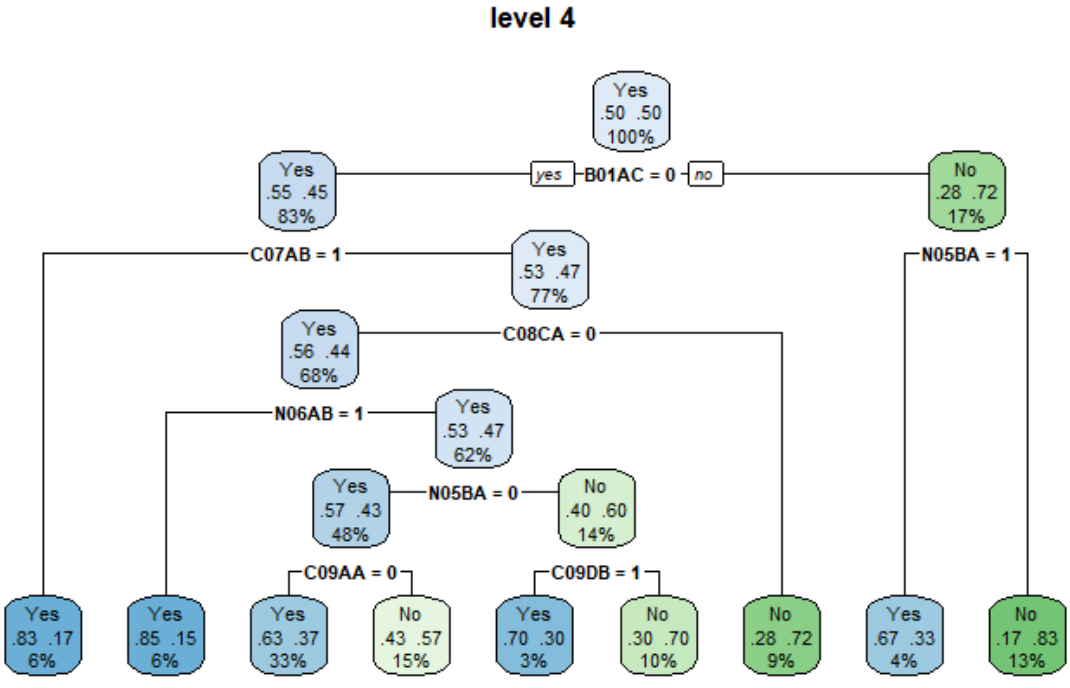


Figura A15. Visualización de árbol de decisión asociada a Xerostomía en Nivel 5

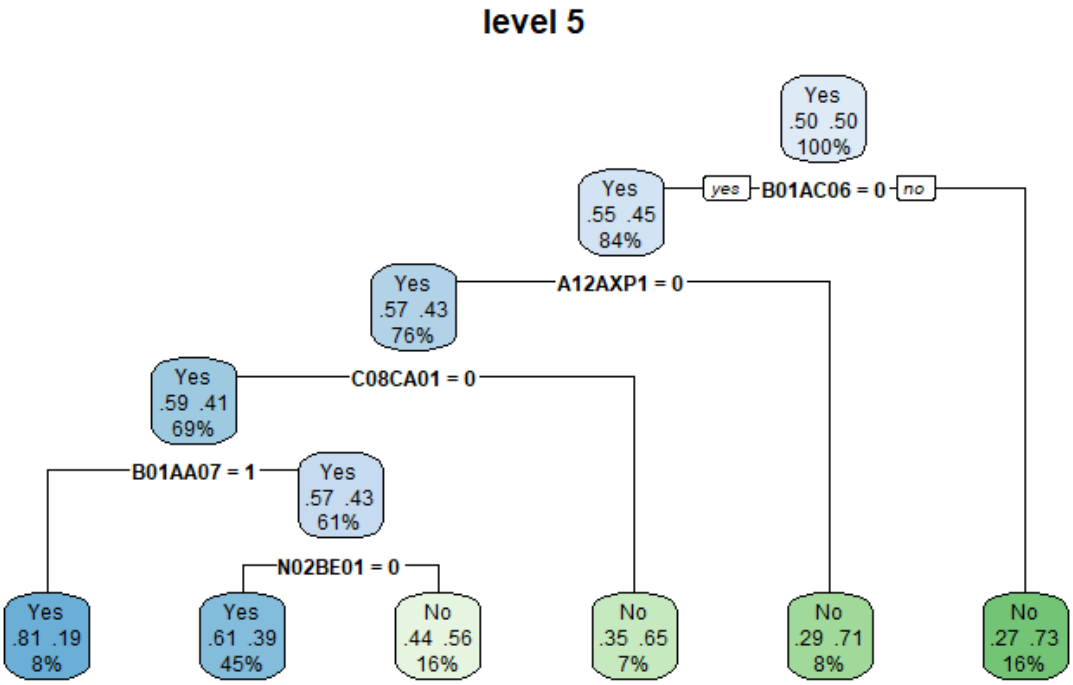


Figura A16. Visualización de árbol de decisión asociada a Hiposalivación en Nivel 1

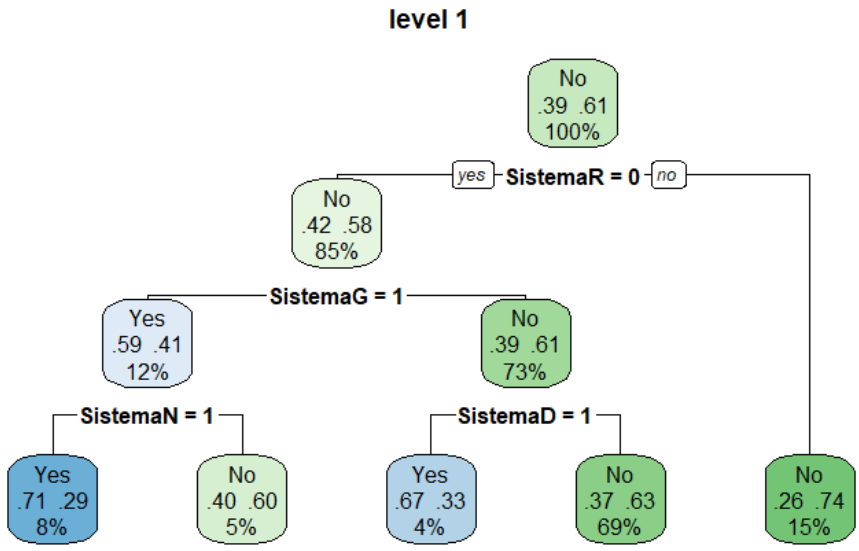


Figura A17. Visualización de árbol de decisión asociada a Hiposalivación en Nivel 2

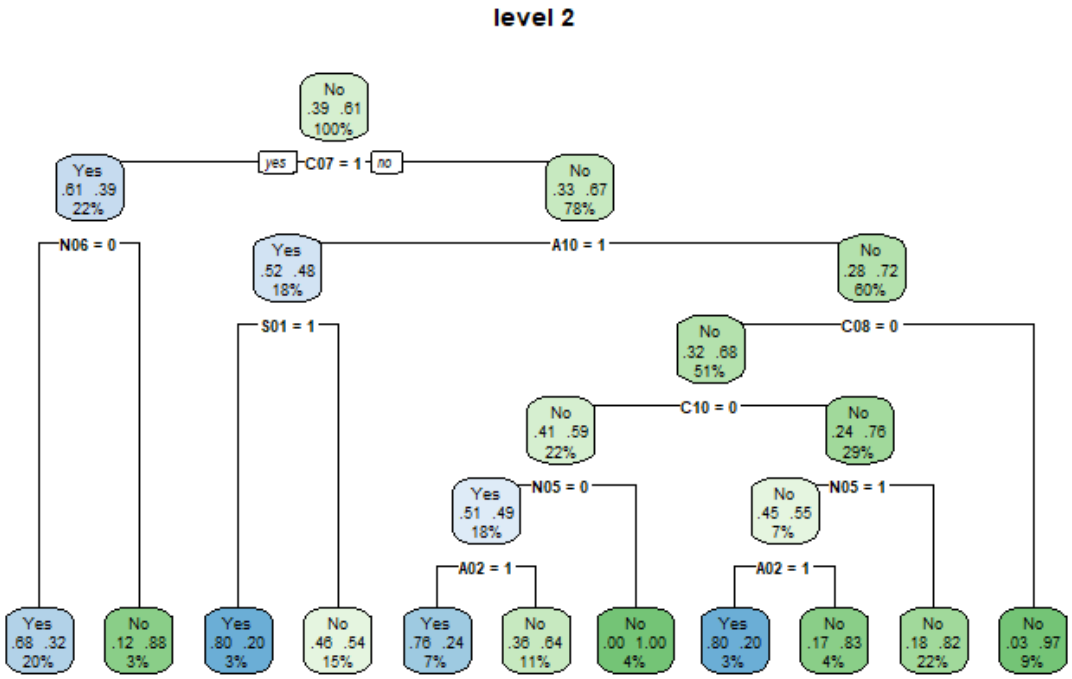


Figura A18. Visualización de árbol de decisión asociada a Hiposalivación en Nivel 3

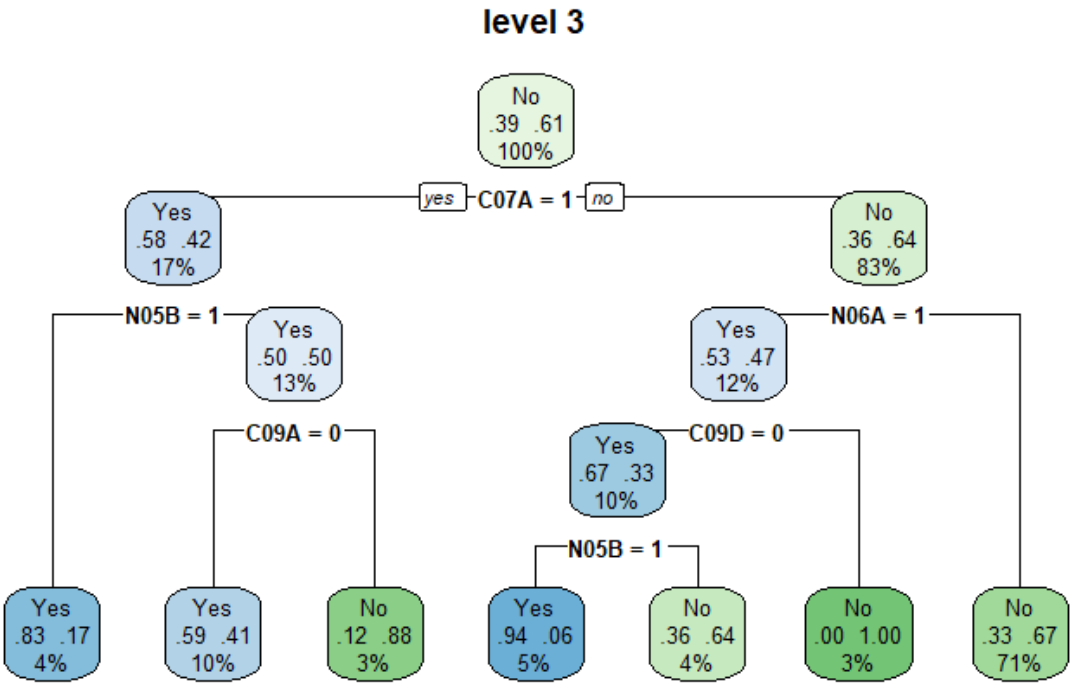


Figura A19. Visualización de árbol de decisión asociada a Hiposalivación en Nivel 4

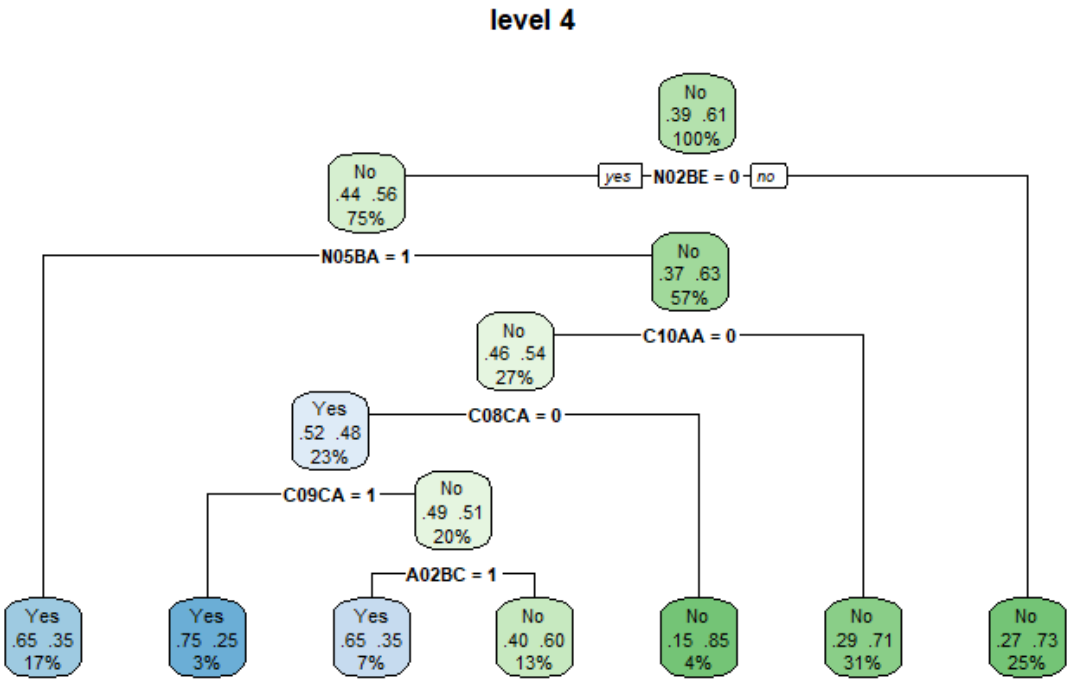


Figura A20. Visualización de árbol de decisión asociada a Hiposalivación en Nivel 5

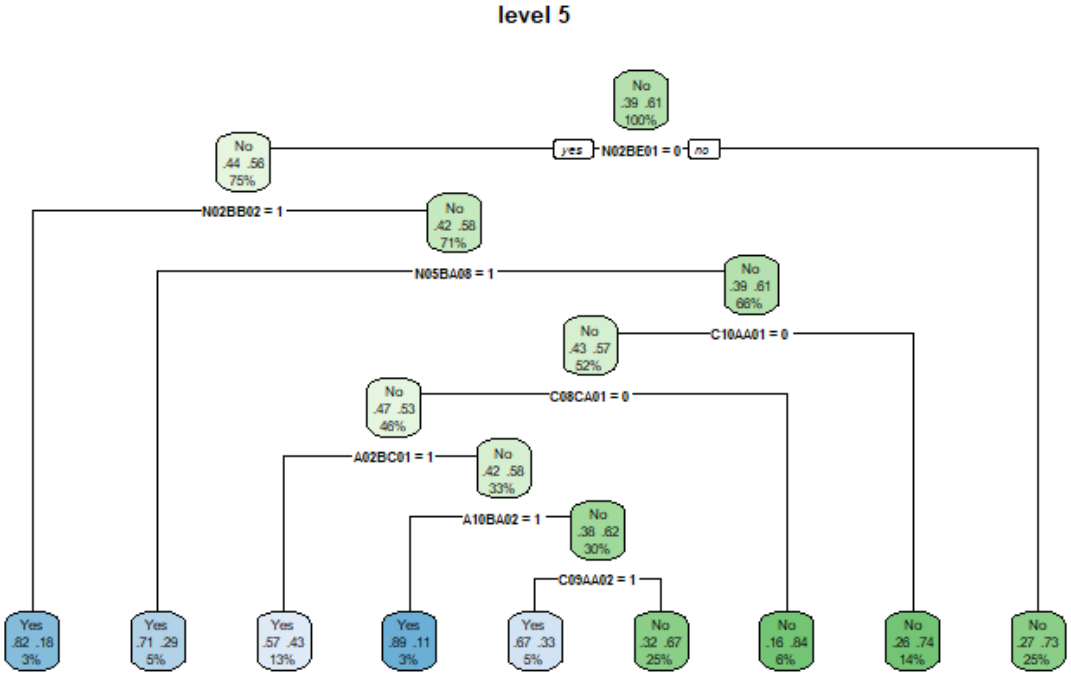


Figura A21. Combinación de parámetros utilizados en SVM para nivel 1 - 5 de Xerostomía

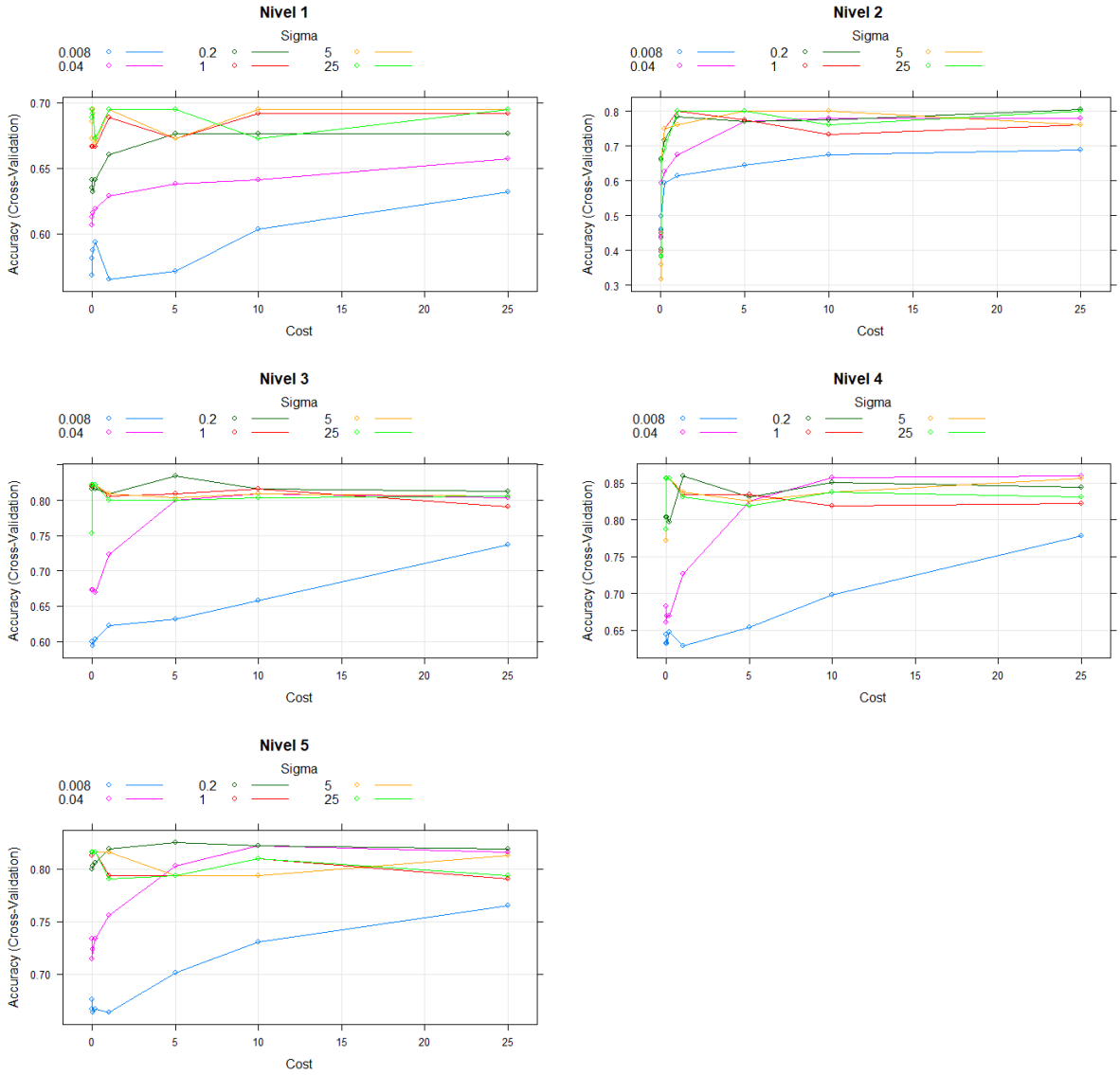
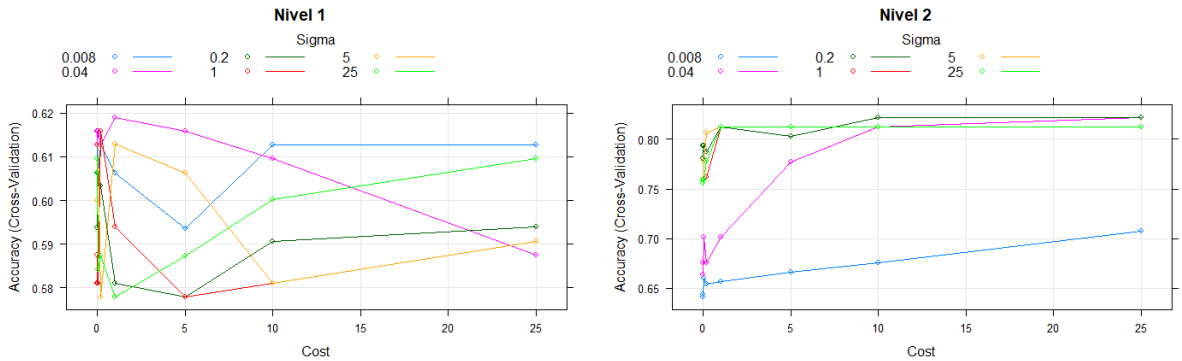


Figura A22. Combinación de parámetros utilizados en SVM para nivel 1 - 5 de Hiposalivación



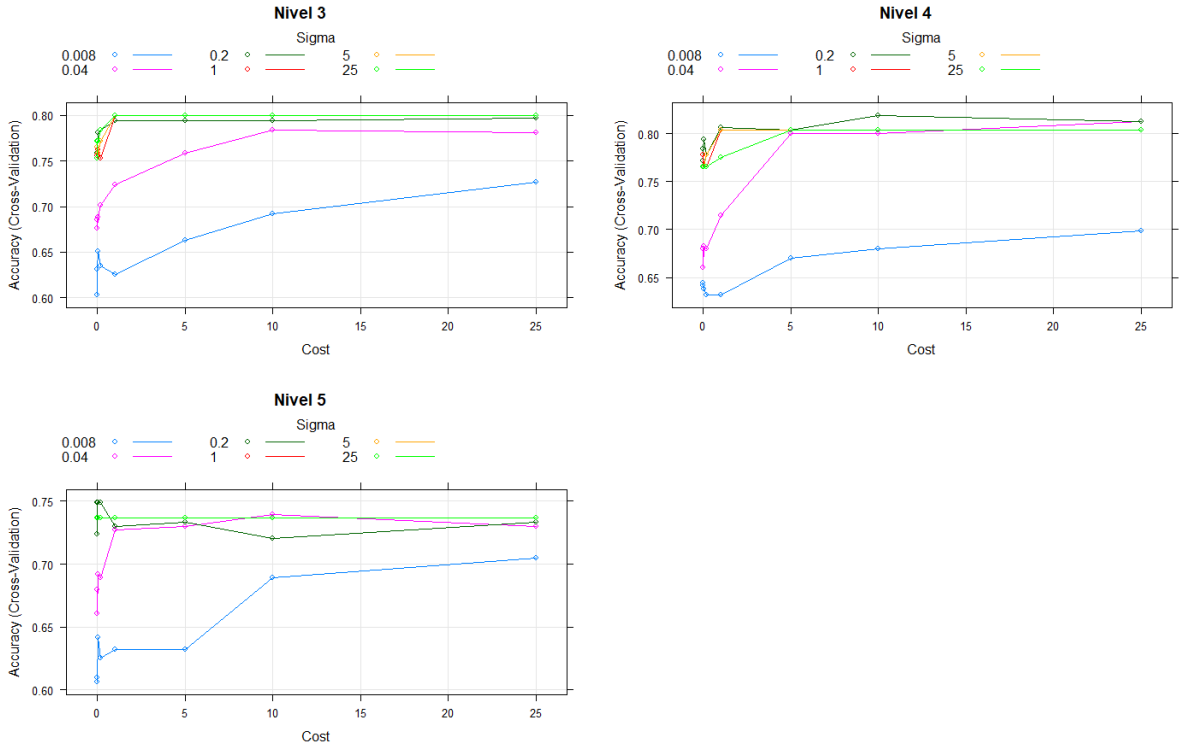
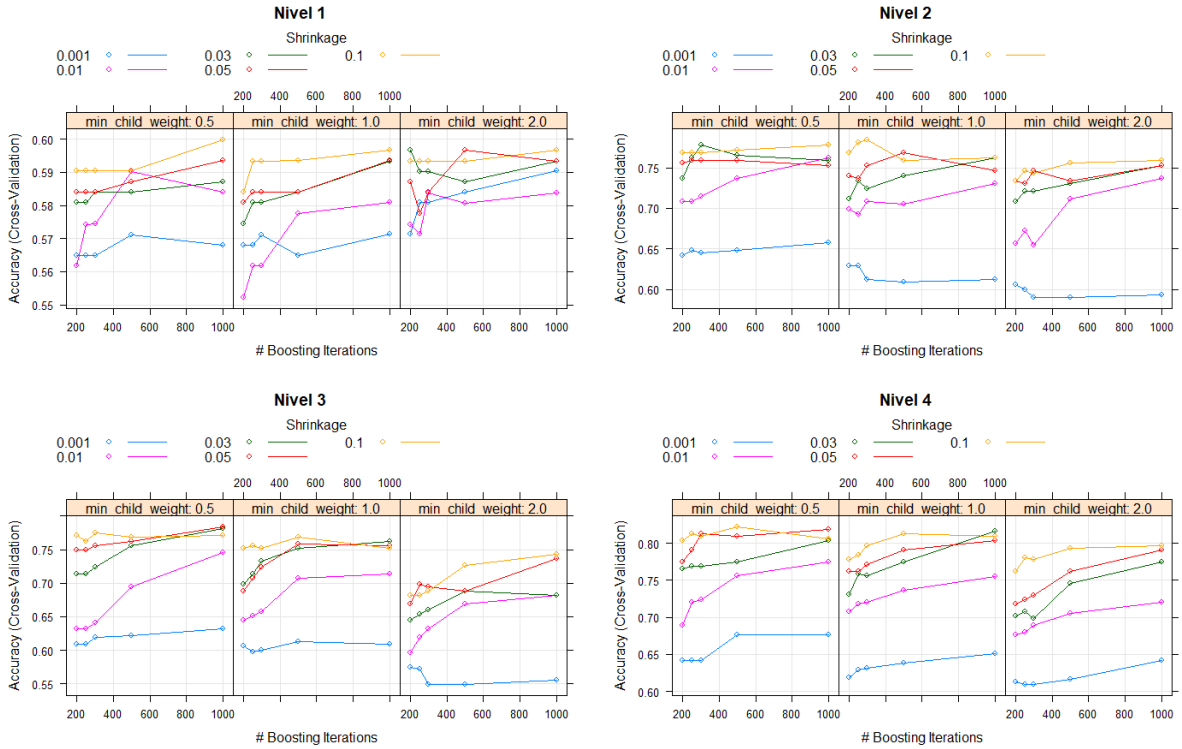


Figura A23. Combinación de parámetros utilizados en XGBoost para nivel 1 - 5 de Xerostomía



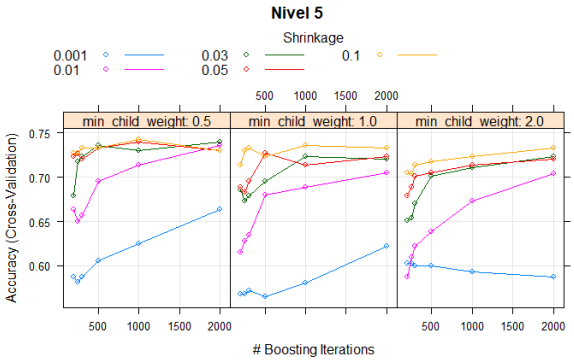


Figura A24. Combinación de parámetros utilizados en XGBoost para nivel 1 - 5 de Hiposalivación

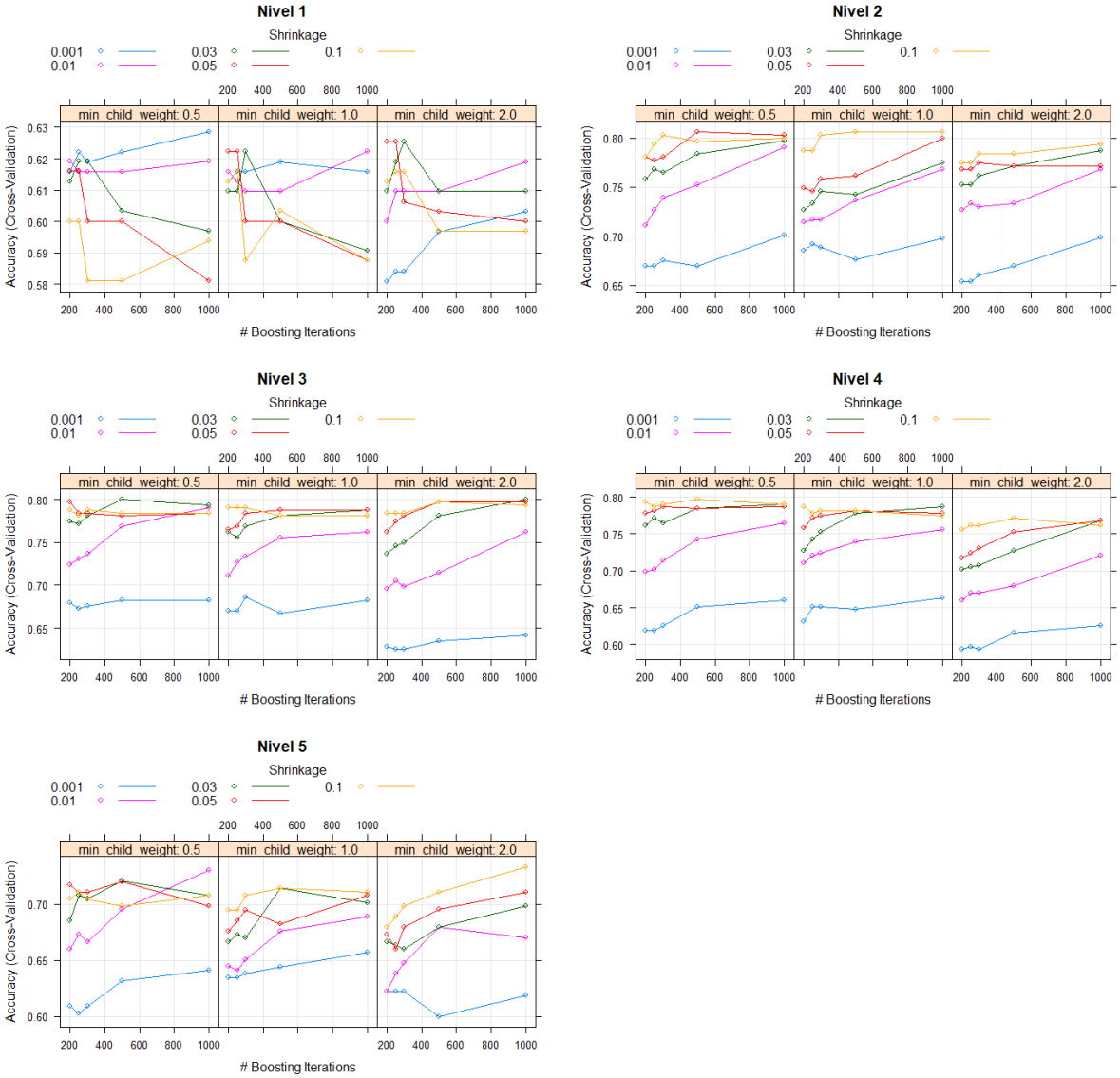


Figura A25. Work flujo para crear cluster en SAS Enterprise Miner en caso de xerostomía

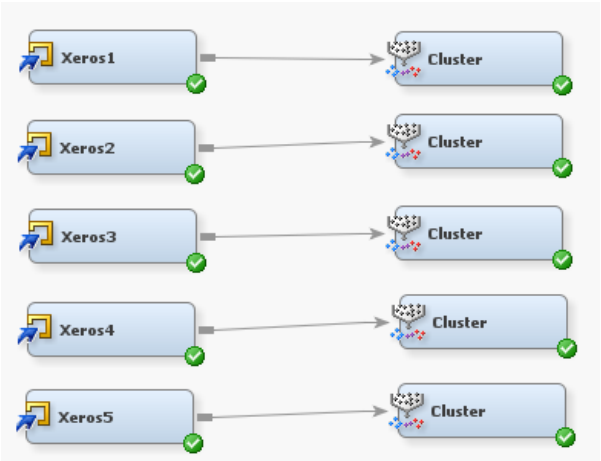


Figura A26. Work flujo para crear cluster en SAS Enterprise Miner en caso de hiposalivación

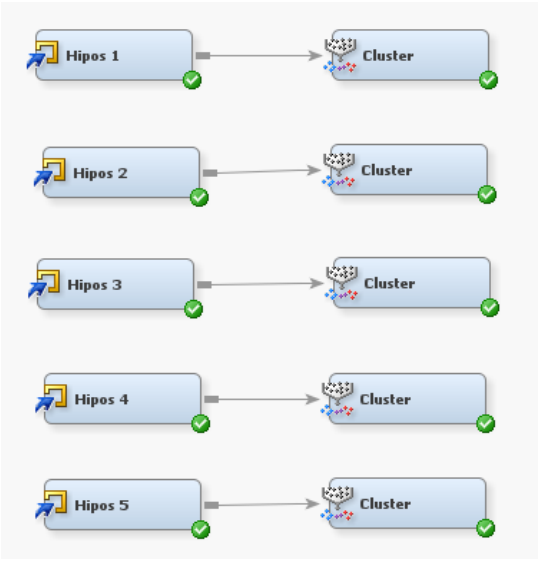
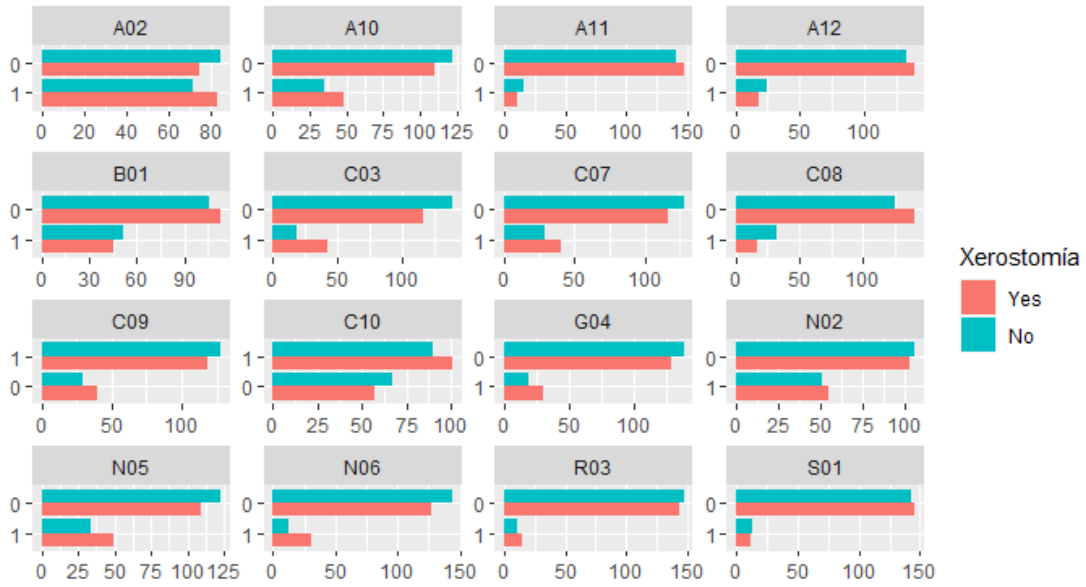


Figura A27. Exploración descriptiva de los pacientes en el caso de xerostomía

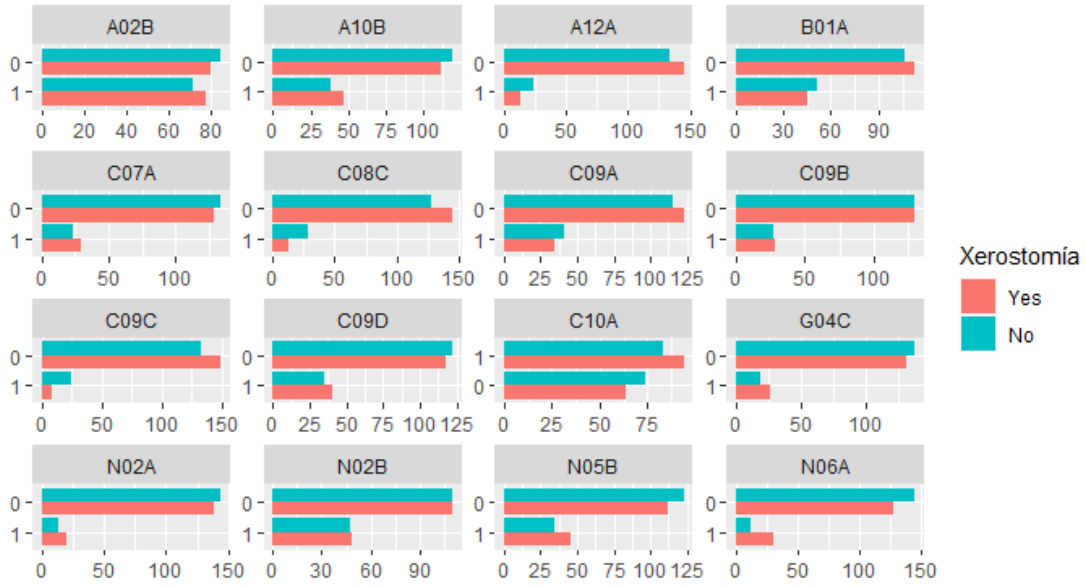
NIVEL 1


SistemaA
SistemaB
SistemaD
SistemaG
SistemaL
SistemaM
SistemaN
SistemaR
SistemaS
Xerostomía
Yes
No

NIVEL 2



NIVEL 3



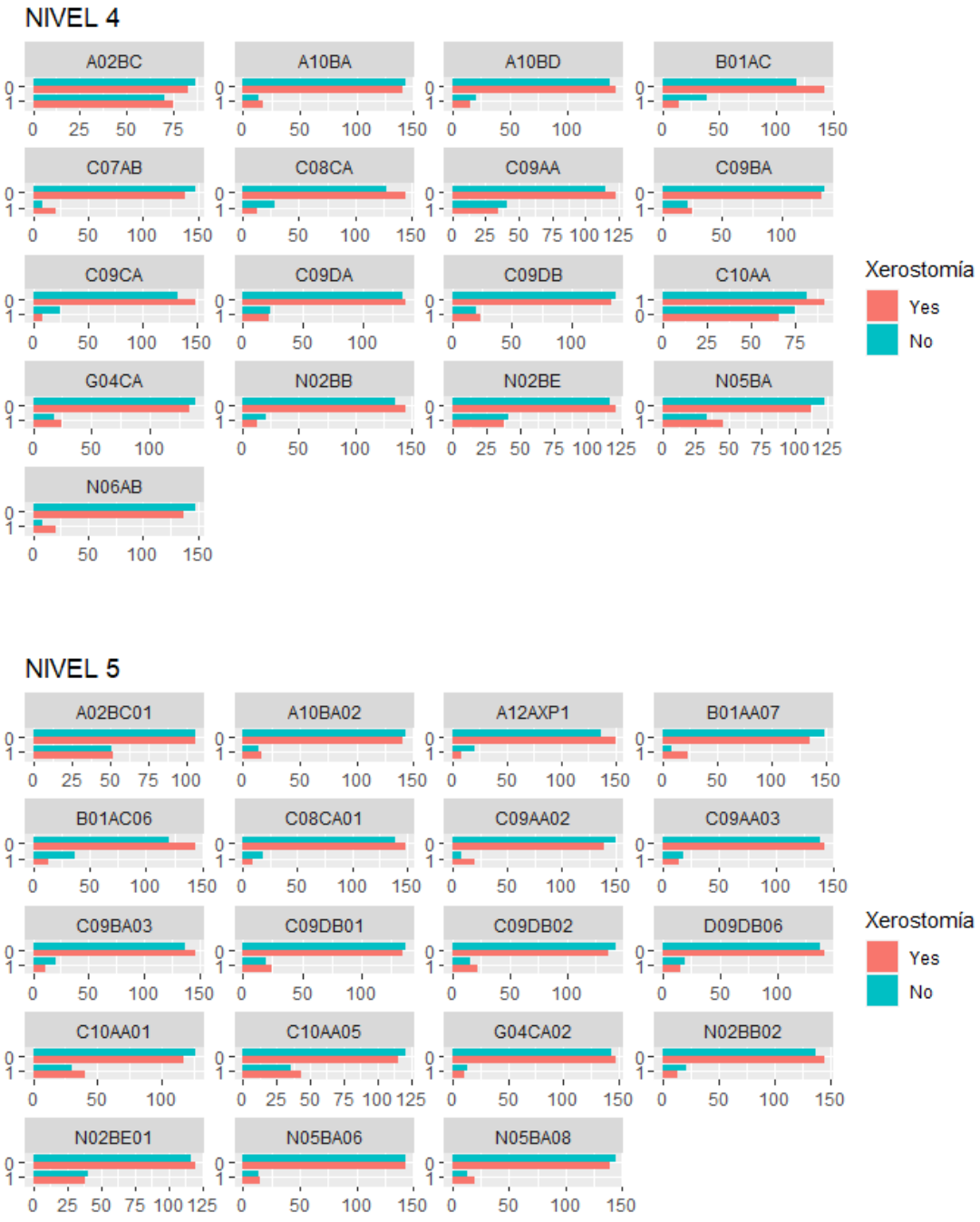
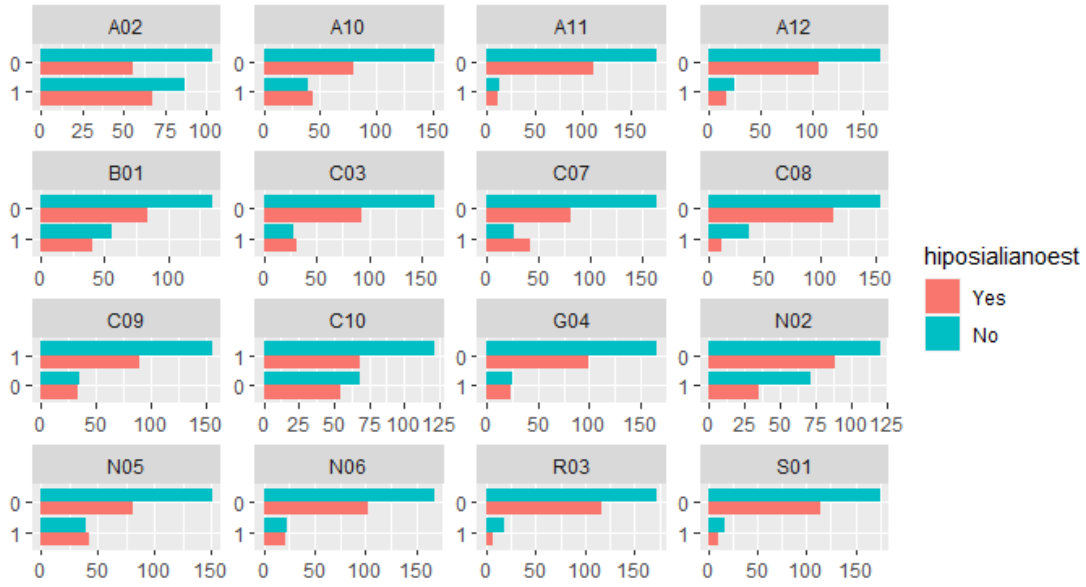


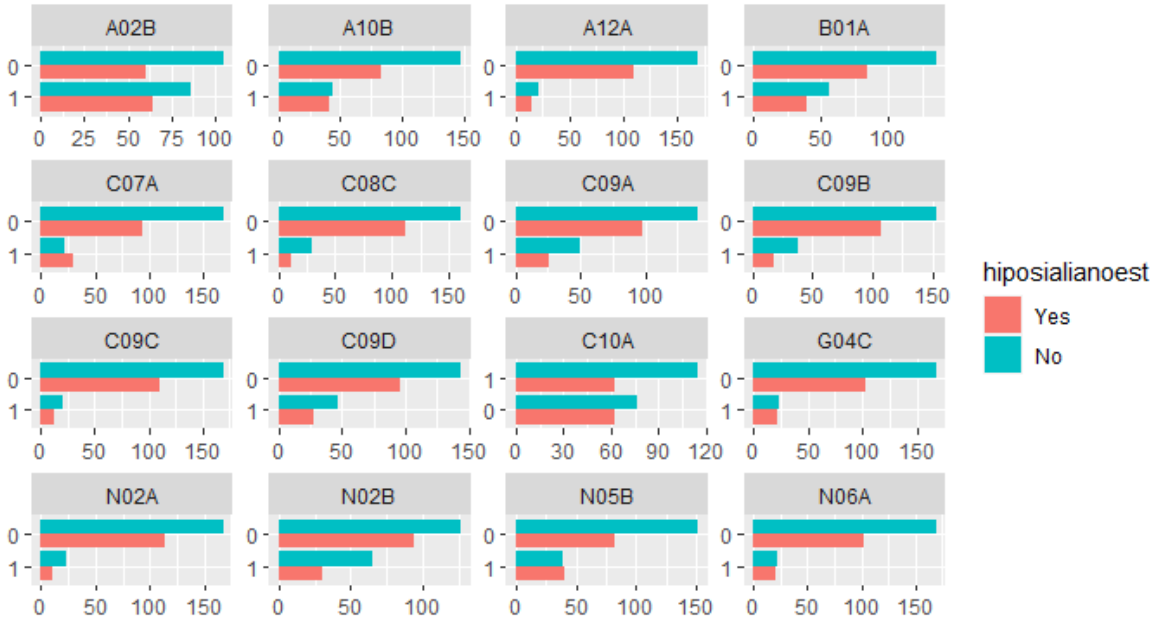
Figura 28. Exploración descriptiva de los pacientes en el caso de hiposalivación

NIVEL 1

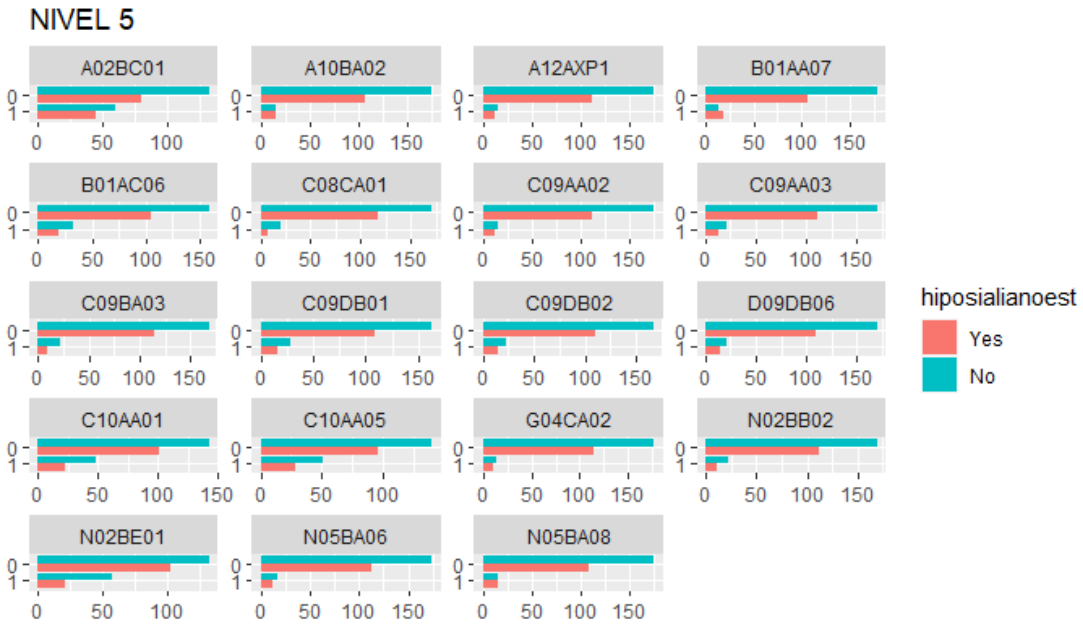
NIVEL 2



NIVEL 3



NIVEL 4



Anexo II: Códigos en R

Códigos en Rstudio:

---
title: "TFM-Xerostomia"
author: "Yuang Huang"
date: "2/1/2022"
output: html_document

```{r}
library(readxl);library(tidymodels);library(tidyverse);library(ggplot2);library(caret);library(DataExplorer); #DALEX
```

Resample for increasing the dataset
```{r}
datas.original<-read_excel("E:/UCM/TFM/TFM Xerostomia Farmacos.xlsx")
set.seed(100384772)
new_sample_ind <- sample( x= c(1:nrow(datas.original)),size = 200, replace = TRUE)
datas <-rbind.data.frame(datas.original,datas.original[new_sample_ind,])
dim(datas)
aaa=names(datas)
fco<-aaa[grep(pattern = "º",names(datas))]
datas<-datas%>%select(-(1:4),-6,-8,-9,-fco)
d.names<- names(datas)
len.names<-nchar(d.names)
set.seed(100384772)
train_ind <- sample(x= c(1:420),size = 420*0.75, replace = FALSE)
datas <- datas%>%mutate(Xerostomía = ifelse(Xerostomía == 1, "Yes" , "No" ),
                        Xerostomía = factor(Xerostomía,levels = c("Yes","No")),
                        hiposialianoest = ifelse(hiposialianoest == 1, "Yes" , "No" ),
                        hiposialianoest = factor(hiposialianoest,levels = c("Yes","No")))
head(datas)
#quantile(datas$Edad,probs = seq(0,1,1/3))
# datas<- datas%>%mutate(Edad = case_when(Edad <= 71 ~"48-71",
```

```
#                       Edad > 71 & Edad<=79 ~"72-79",
#                       Edad > 79~"80+") )
```

```{r}
xeros.percepcion1<-datas%>%select(1,which(len.names>=12))%>%select(-
"hiposialianoest")%>%slice(train_ind)
names(xeros.percepcion1) <-
c("Xerostomía","SistemaA","SistemaB","SistemaC","SistemaD","SistemaG","SistemaH","Sis
temaJ","SistemaL","SistemaM","SistemaN",
                    "SistemaR","SistemaS")
test.percepcion1 <-datas%>%select(1,which(len.names>=12))%>%select(-
"hiposialianoest")%>%slice(-train_ind)
names(test.percepcion1) <-
c("Xerostomía","SistemaA","SistemaB","SistemaC","SistemaD","SistemaG","SistemaH","Sis
temaJ","SistemaL","SistemaM","SistemaN",
                    "SistemaR","SistemaS")

xeros.percepcion2<-datas%>%select(1,which(len.names==3))%>%slice(train_ind)
test.percepcion2<-datas%>%select(1,which(len.names==3))%>%slice(-train_ind)

xeros.percepcion3<-datas%>%select(1,which(len.names==4))%>%slice(train_ind)
test.percepcion3<-datas%>%select(1,which(len.names==4))%>%slice(-train_ind)

xeros.percepcion4<-datas%>%select(1,which(len.names==5))%>%slice(train_ind)
test.percepcion4<-datas%>%select(1,which(len.names==5))%>%slice(-train_ind)

xeros.percepcion5<-datas%>%select(1,which(len.names==7))%>%slice(train_ind)
test.percepcion5<-datas%>%select(1,which(len.names==7))%>%slice(-train_ind)
```

```{r}
#name.delete<-grep("fco",names(datas),value = TRUE)
```

#datas<-datas%>%select(-`Nº de fcos` , -`Nº fcos A`, - `A02 nº fcoss`, -`A10 Nº fcos`,-`a12 nº fcos`, -`Nº fcos B` ,- `B01 nº fcos`,-`B03 Nº fcos`, -`Nº fcos C` , - `C01 nº fcos` ,-`C03 nº fcos`, - `C04 nº fcos` , -`C08 nº fcos` ,- `C09 nºfcos` ,- `C10 nº fcos` ,-`Nº fcos D` , - `G04 nº fcos` , -`Nº fcos H` , - `Nº fcos M` ,-`Nº fcos N` ,-`N02 nº fcos`,-`N03 nº fcos` ,-`n04 nº fcos` ,- `N05 nº fcos`, -`N06 nº fcos`)
#dim(datas)

#datas.simple<-datas%>%select(`Nº paciente`,Sexo,`Edad (años)`,`Nº de fcos`,`nº Cigarros/dia`,`Dosis alcohol`,`Nº fcos A`,`Nº fcos B`,`Nº fcos C`,`Nº fcos D`,`Nº Fcos G`,`Nº fcos H`,`Nº fcos M`,`Nº fcos N`,`Nº Fcos R`,XEROSTOMÍA,`Saliva Estimulada (ml/min)`,`Saliva no estimulada (ml/min)`) #datas$`Hiposialia saliva no estimulada`
#names(datas.simple)[3]<-"Edad"
#names(datas.simple)[16]<-"Xerostomia"

#sk<-skimr::skim(datas.simple)
```

# near 0 Varianza para sensacion
```{r}
cut0var1<-caret::nearZeroVar(xeros.percepcion1,freqCut =nrow(xeros.percepcion1)/10,uniqueCut = 1) #freqCut: the cutoff for the ratio of the most common value to the second most common value
#uniqueCut: the cutoff for the percentage of distinct values out of the number of total samples
length(xeros.percepcion1)
length(cut0var1)
datas_0var1<-xeros.percepcion1[,-cut0var1]#%>%select(-"SistemaH",-"SistemaJ",-"SistemaL" )
#view(colSums(datas[,cut0var],na.rm = TRUE))
dim(datas_0var1)
#hemos quitado sistema H, J y L, porque ellos tiene menos de 6 muestras.

cut0var2<-caret::nearZeroVar(xeros.percepcion2,freqCut =nrow(xeros.percepcion2)/25,uniqueCut = 1)

```
length(xeros.percepcion2)
length(cut0var2)
datas_0var2<-xeros.percepcion2[,-cut0var2]

cut0var3<-caret::nearZeroVar(xeros.percepcion3,freqCut
=nrow(xeros.percepcion3)/25,uniqueCut = 1)
length(xeros.percepcion3)
length(cut0var3)
datas_0var3<-xeros.percepcion3[,-cut0var3]

cut0var4<-caret::nearZeroVar(xeros.percepcion4,freqCut
=nrow(xeros.percepcion4)/25,uniqueCut = 1)
length(xeros.percepcion4)
length(cut0var4)
datas_0var4<-xeros.percepcion4[,-cut0var4]

cut0var5<-caret::nearZeroVar(xeros.percepcion5,freqCut
=nrow(xeros.percepcion5)/25,uniqueCut = 1)
length(xeros.percepcion5)
length(cut0var5)
datas_0var5<-xeros.percepcion5[,-cut0var5]
```

EDA para data sensacion

```{r}
datas_0var1 %>% plot_bar(by='Xerostomía', by_position='dodge',nrow = 4,title = 'NIVEL 1')
datas_0var2 %>% plot_bar(by='Xerostomía', by_position='dodge',nrow = 5,ncol = 4,title =
'NIVEL 2')
datas_0var3 %>% plot_bar(by='Xerostomía', by_position='dodge',nrow = 5,ncol = 4,title =
'NIVEL 3')
datas_0var4 %>% plot_bar(by='Xerostomía', by_position='dodge',nrow = 5,ncol = 4,title =
'NIVEL 4')
datas_0var5 %>% plot_bar(by='Xerostomía', by_position='dodge',nrow = 5,ncol = 4,title =
'NIVEL 5')
```

```

# Logit Regression para data fake

```{r}
#control<-trainControl(method = "CV",number=4,summaryFunction =
twoClassSummary,classProbs=TRUE,savePredictions = "all")
control<-trainControl(method = "CV",number=4,classProbs=TRUE,savePredictions = "all")
set.seed(100384772)
logit.1<- train(Xerostomía~.,data =
datas_0var1,trControl=control,method="glmStepAIC",direction = "forward")
logit.2<- train(Xerostomía~.,data =
datas_0var2,trControl=control,method="glmStepAIC",direction = "both")
logit.3<- train(Xerostomía~.,data =
datas_0var3,trControl=control,method="glmStepAIC",direction = "back")
logit.4<- train(Xerostomía~.,data =
datas_0var4,trControl=control,method="glmStepAIC",direction = "back")
logit.5<- train(Xerostomía~.,data =
datas_0var5,trControl=control,method="glmStepAIC",direction = "forward")

plot(pROC::roc(response=ifelse(logit.1$pred$obs
=="Yes",1,0),predictor=logit.1$pred$Yes),main = "LOGIT 1")
plot(pROC::roc(response=ifelse(test.percepcion1$Xerostomía =="Yes",1,0),predictor=1-
predict(logit.1,test.percepcion1,type = "prob")$Yes),col = "purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.2$pred$obs
=="Yes",1,0),predictor=logit.2$pred$Yes),main = "LOGIT 2")
plot(pROC::roc(response=ifelse(test.percepcion2$Xerostomía
=="Yes",1,0),predictor=predict(logit.2,test.percepcion2,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```
```

```
plot(pROC::roc(response=ifelse(logit.3$pred$obs
=="Yes",1,0),predictor=logit.3$pred$Yes),main = "LOGIT 3")
plot(pROC::roc(response=ifelse(test.percepcion3$Xerostomía
=="Yes",1,0),predictor=predict(logit.3,test.percepcion3,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.4$pred$obs
=="Yes",1,0),predictor=logit.4$pred$Yes),main = "LOGIT 4")
plot(pROC::roc(response=ifelse(test.percepcion4$Xerostomía
=="Yes",1,0),predictor=predict(logit.4,test.percepcion4,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.5$pred$obs
=="Yes",1,0),predictor=logit.5$pred$Yes),main = "LOGIT 5")
plot(pROC::roc(response=ifelse(test.percepcion5$Xerostomía
=="Yes",1,0),predictor=predict(logit.5,test.percepcion5,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(logit.1,test.percepcion1,type = "raw"),test.percepcion1$Xerostomía)
confusionMatrix(predict(logit.2,test.percepcion2,type = "raw"),test.percepcion2$Xerostomía)
confusionMatrix(predict(logit.3,test.percepcion3,type = "raw"),test.percepcion3$Xerostomía)
confusionMatrix(predict(logit.4,test.percepcion4,type = "raw"),test.percepcion4$Xerostomía)
confusionMatrix(predict(logit.5,test.percepcion5,type = "raw"),test.percepcion5$Xerostomía)

#模型解释是基于其余变量不变的情况下，具有局限性。说明这模型并不科学。

```

```
# SVM para data fake
```

```{r}
library(e1071)
SVMgrid<-expand.grid(C= c(5^c(-4:2),10), sigma = 5^c(-3:2))

#SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2),degree=c(2,3),scale=c(0.1,0.5,1,2,5))

#SVMgrid<-expand.grid(C=5^c(-3:2))
set.seed(100384772)
svm.1<- train(Xerostomía~.,data =
datas_0var1,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.2<- train(Xerostomía~.,data =
datas_0var2,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.3<- train(Xerostomía~.,data =
datas_0var3,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.4<- train(Xerostomía~.,data =
datas_0var4,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.5<- train(Xerostomía~.,data =
datas_0var5,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)

plot(pROC::roc(response=ifelse(svm.1$pred$obs
=="Yes",1,0),predictor=svm.1$pred$Yes),main = "SVM1")
plot(pROC::roc(response=ifelse(test.percepcion1$Xerostomía
=="Yes",1,0),predictor=predict(svm.1,test.percepcion1,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.2$pred$obs
=="Yes",1,0),predictor=svm.2$pred$Yes),main = "SVM2")
plot(pROC::roc(response=ifelse(test.percepcion2$Xerostomía
=="Yes",1,0),predictor=predict(svm.2,test.percepcion2,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
plot(pROC::roc(response=ifelse(svm.3$pred$obs
=="Yes",1,0),predictor=svm.3$pred$Yes),main = "SVM3")
plot(pROC::roc(response=ifelse(test.percepcion3$Xerostomía
=="Yes",1,0),predictor=predict(svm.3,test.percepcion3,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.4$pred$obs
=="Yes",1,0),predictor=svm.4$pred$Yes),main = "SVM4")
plot(pROC::roc(response=ifelse(test.percepcion4$Xerostomía
=="Yes",1,0),predictor=predict(svm.4,test.percepcion4,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.5$pred$obs
=="Yes",1,0),predictor=svm.5$pred$Yes),main = "SVM5")
plot(pROC::roc(response=ifelse(test.percepcion5$Xerostomía
=="Yes",1,0),predictor=predict(svm.5,test.percepcion5,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(svm.1,test.percepcion1,type = "raw"),test.percepcion1$Xerostomía)
confusionMatrix(predict(svm.2,test.percepcion2,type = "raw"),test.percepcion2$Xerostomía)
confusionMatrix(predict(svm.3,test.percepcion3,type = "raw"),test.percepcion3$Xerostomía)
confusionMatrix(predict(svm.4,test.percepcion4,type = "raw"),test.percepcion4$Xerostomía)
confusionMatrix(predict(svm.5,test.percepcion5,type = "raw"),test.percepcion5$Xerostomía)

plot(svm.1, main = "Nivel 1");plot(svm.2,main = "Nivel 2");plot(svm.3,main = "Nivel
3");plot(svm.4,main = "Nivel 4");plot(svm.5,main = "Nivel 5");
```

#Las SVM se basan en support vector para construir el modelo, por lo que no tenemos forma de detectar la importancia de las variables
```

# XGBOOST para data fake
```{r}
#Early stopping
xgbmgrid_es<-expand.grid(eta=c(0.1),
 min_child_weight=c(0.5),
 nrounds=c(200,250,300,500,1000,2000,3000,5000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)
x.es1<- train(Xerostomía~.,data = datas_0var1,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es2<- train(Xerostomía~.,data = datas_0var2,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es3<- train(Xerostomía~.,data = datas_0var3,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es4<- train(Xerostomía~.,data = datas_0var4,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es5<- train(Xerostomía~.,data = datas_0var5,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
plot(x.es1);plot(x.es2);plot(x.es3);plot(x.es4);plot(x.es5)

xgbmgrid1<-expand.grid(eta=c(0.1,0.05,0.03,0.01,0.001),
 min_child_weight=c(0.5,1,2),
 nrounds=c(200,250,300,500,1000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

xgbmgrid2<-expand.grid(eta=c(0.1,0.05,0.03,0.01,0.001),
 min_child_weight=c(0.5,1,2),
 nrounds=c(200,250,300,500,1000,2000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

set.seed(100384772)
```
```

```
xgbm1<- train(Xerostomía~.,data = datas_0var1,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid1,verbose=FALSE)
xgbm2<- train(Xerostomía~.,data = datas_0var2,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid1,verbose=FALSE)
xgbm3<- train(Xerostomía~.,data = datas_0var3,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid1,verbose=FALSE)
xgbm4<- train(Xerostomía~.,data = datas_0var4,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid1,verbose=FALSE)
xgbm5<- train(Xerostomía~.,data = datas_0var5,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid2,verbose=FALSE)

plot(pROC::roc(response=ifelse(xgbm1$pred$obs
=="Yes",1,0),predictor=xgbm1$pred$Yes),main = "XGBOOST1")
plot(pROC::roc(response=ifelse(test.percepcion1$Xerostomía
=="Yes",1,0),predictor=predict(svm.1,test.percepcion1,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm2$pred$obs
=="Yes",1,0),predictor=xgbm2$pred$Yes),main = "XGBOOST2")
plot(pROC::roc(response=ifelse(test.percepcion2$Xerostomía
=="Yes",1,0),predictor=predict(svm.2,test.percepcion2,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm3$pred$obs
=="Yes",1,0),predictor=xgbm3$pred$Yes),main = "XGBOOST3")
plot(pROC::roc(response=ifelse(test.percepcion3$Xerostomía
=="Yes",1,0),predictor=predict(svm.3,test.percepcion3,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
plot(pROC::roc(response=ifelse(xgbm4$pred$obs
=="Yes",1,0),predictor=xgbm4$pred$Yes),main = "XGBOOST4")
plot(pROC::roc(response=ifelse(test.percepcion4$Xerostomía
=="Yes",1,0),predictor=predict(svm.4,test.percepcion4,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm5$pred$obs
=="Yes",1,0),predictor=xgbm5$pred$Yes),main = "XGBOOST5")
plot(pROC::roc(response=ifelse(test.percepcion5$Xerostomía
=="Yes",1,0),predictor=predict(svm.5,test.percepcion5,type = "prob")$Yes),col =
"purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(xgbm1,test.percepcion1,type = "raw"),test.percepcion1$Xerostomía)
confusionMatrix(predict(xgbm2,test.percepcion2,type = "raw"),test.percepcion2$Xerostomía)
confusionMatrix(predict(xgbm3,test.percepcion3,type = "raw"),test.percepcion3$Xerostomía)
confusionMatrix(predict(xgbm4,test.percepcion4,type = "raw"),test.percepcion4$Xerostomía)
confusionMatrix(predict(xgbm5,test.percepcion5,type = "raw"),test.percepcion5$Xerostomía)

plot(xgbm1, main = "Nivel 1");plot(xgbm2, main = "Nivel 2");plot(xgbm3, main = "Nivel
3");plot(xgbm4, main = "Nivel 4");plot(xgbm5, main = "Nivel 5");

explainer_gbm <- DALEX::explain(xgbm2, data = test.percepcion2 %>% dplyr::select(-
Xerostomía),
                        y = test.percepcion2$Xerostomía, label = "GBM")
explainer_gbm %>%
    model_profile(variables = c(names(datas_0var2))[2:17])%>%plot()
tester <- test.percepcion2[4, ]
tester
```

```
# Break down Features
bd_gbm <- explainer_gbm %>%
DALEX::predict_parts(new_observation = tester,
type = "break_down")
plot(bd_gbm)
```

DT CON BUCLE para data fake

```{r}
library(rpart.plot)
arbolgrid <-  expand.grid(cp=c(0.01))
store_dt_prm <- c()
set.seed(100384772)
for (minbu in seq(from=10, to=80, by=10)){
  for (minsp in seq(from = 4, to= 24, by = 4)){
  print(c(minbu,minsp))
  cat("\n")
  arbol1 <-  train(Xerostomía~.,data = datas_0var1,method="rpart",minsplit =
minsp,minbucket=minbu,trControl=control,tuneGrid=arbolgrid)
  arbol2 <-  train(Xerostomía~.,data = datas_0var2,method="rpart",minsplit =
minsp,minbucket=minbu,trControl=control,tuneGrid=arbolgrid)
  arbol3 <-  train(Xerostomía~.,data = datas_0var3,method="rpart",minsplit =
minsp,minbucket=minbu,trControl=control,tuneGrid=arbolgrid)
  arbol4 <-  train(Xerostomía~.,data = datas_0var4,method="rpart",minsplit =
minsp,minbucket=minbu,trControl=control,tuneGrid=arbolgrid)
  arbol5 <-  train(Xerostomía~.,data = datas_0var5,method="rpart",minsplit =
minsp,minbucket=minbu,trControl=control,tuneGrid=arbolgrid)
  store_dt_prm<-
data.frame(rbind(store_dt_prm,c(minbu,minsp,arbol1$results[[2]],arbol2$results[[2]],arbol3$r
esults[[2]],arbol4$results[[2]],arbol5$results[[2]])))
  }
}
names(store_dt_prm) <- c('minbucket','minsplit','Accuracy in dt1','Accuracy in dt2','Accuracy
in dt3','Accuracy in dt4','Accuracy in dt5')
```

```
print(store_dt_prm[which.max(store_dt_prm$`Accuracy in dt1`),])
print(store_dt_prm[which.max(store_dt_prm$`Accuracy in dt2`),])
print(store_dt_prm[which.max(store_dt_prm$`Accuracy in dt3`),])
print(store_dt_prm[which.max(store_dt_prm$`Accuracy in dt4`),])
print(store_dt_prm[which.max(store_dt_prm$`Accuracy in dt5`),])

set.seed(100384772)
dt1 <- train(Xerostomía~.,data = datas_0var1,method="rpart",minsplit =
12,minbucket=10,trControl=control,tuneGrid=arbolgrid)
dt2 <- train(Xerostomía~.,data = datas_0var2,method="rpart",minsplit =
4,minbucket=60,trControl=control,tuneGrid=arbolgrid)
dt3 <- train(Xerostomía~.,data = datas_0var3,method="rpart",minsplit =
16,minbucket=80,trControl=control,tuneGrid=arbolgrid)
dt4 <- train(Xerostomía~.,data = datas_0var4,method="rpart",minsplit =
4,minbucket=80,trControl=control,tuneGrid=expand.grid(cp=0.025))
dt5 <- train(Xerostomía~.,data = datas_0var5,method="rpart",minsplit =
24,minbucket=50,maxdepth = 2,trControl=control,tuneGrid=expand.grid(cp=0.0135))

plot(pROC::roc(response=ifelse(dt1$pred$obs =="Yes",1,0),predictor=dt1$pred$Yes),main =
"Decision Tree1")
plot(pROC::roc(response=ifelse(test.percepcion1$Xerostomía
=="Yes",1,0),predictor=predict(dt1,test.percepcion1,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt2$pred$obs =="Yes",1,0),predictor=dt2$pred$Yes),main =
"Decision Tree2")
plot(pROC::roc(response=ifelse(test.percepcion2$Xerostomía
=="Yes",1,0),predictor=predict(dt2,test.percepcion2,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
plot(pROC::roc(response=ifelse(dt3$pred$obs =="Yes",1,0),predictor=dt3$pred$Yes),main =
"Decision Tree3")
plot(pROC::roc(response=ifelse(test.percepcion3$Xerostomía
=="Yes",1,0),predictor=predict(dt3,test.percepcion3,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt4$pred$obs =="Yes",1,0),predictor=dt4$pred$Yes),main =
"Decision Tree4")
plot(pROC::roc(response=ifelse(test.percepcion4$Xerostomía
=="Yes",1,0),predictor=predict(dt4,test.percepcion4,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt5$pred$obs =="Yes",1,0),predictor=dt5$pred$Yes),main =
"Decision Tree5")
plot(pROC::roc(response=ifelse(test.percepcion5$Xerostomía
=="Yes",1,0),predictor=predict(dt5,test.percepcion5,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(dt1,test.percepcion1,type = "raw"),test.percepcion1$Xerostomía)
confusionMatrix(predict(dt2,test.percepcion2,type = "raw"),test.percepcion2$Xerostomía)
confusionMatrix(predict(dt3,test.percepcion3,type = "raw"),test.percepcion3$Xerostomía)
confusionMatrix(predict(dt4,test.percepcion4,type = "raw"),test.percepcion4$Xerostomía)
confusionMatrix(predict(dt5,test.percepcion5,type = "raw"),test.percepcion5$Xerostomía)

rpart.plot(dt1$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 1")
rpart.plot(dt2$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 2")
rpart.plot(dt3$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 3")
rpart.plot(dt4$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 4")
```

```
rpart.plot(dt5$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 5")

plot(varImp(dt1),main = "NIVEL 1");plot(varImp(dt2),main = "NIVEL
2");plot(varImp(dt3),main = "NIVEL 3")
plot(varImp(dt4),main = "NIVEL 4");plot(varImp(dt5),main = "NIVEL 5")
```

REAL DATA

```{r}
library(ROSE)

set.seed(100384772)
xeros.real1<- datas_0var1%>%select(-1)%>%mutate(hiposialianoest =
datas[train_ind,]$hiposialianoest)
#xeros.real1.bal<- ovun.sample(hiposialianoest~., data = xeros.real1,method = "over")$data
test.real1<- test.percepcion1%>%select(-1)%>%mutate(hiposialianoest = datas[-
train_ind,]$hiposialianoest)

xeros.real2<- datas_0var2%>%select(-1)%>%mutate(hiposialianoest =
datas[train_ind,]$hiposialianoest)
#xeros.real2.bal<- ovun.sample(hiposialianoest~., data = xeros.real2,method = "over")$data
test.real2<- test.percepcion2%>%select(-1)%>%mutate(hiposialianoest = datas[-
train_ind,]$hiposialianoest)

xeros.real3<- datas_0var3%>%select(-1)%>%mutate(hiposialianoest =
datas[train_ind,]$hiposialianoest)
#xeros.real3.bal<- ovun.sample(hiposialianoest~., data = xeros.real3,method = "over")$data
test.real3<- test.percepcion3%>%select(-1)%>%mutate(hiposialianoest = datas[-
train_ind,]$hiposialianoest)

xeros.real4<- datas_0var4%>%select(-1)%>%mutate(hiposialianoest =
datas[train_ind,]$hiposialianoest)
#xeros.real4.bal<- ovun.sample(hiposialianoest~., data = xeros.real4,method = "over")$data
```

```
test.real4<- test.percepcion4%>%select(-1)%>%mutate(hiposialianoest = datas[-train_ind,]$hiposialianoest)

xeros.real5<- datas_0var5%>%select(-1)%>%mutate(hiposialianoest = datas[train_ind,]$hiposialianoest)
#xeros.real5.bal<- ovun.sample(hiposialianoest~., data = xeros.real5,method = "over")$data
test.real5<- test.percepcion5%>%select(-1)%>%mutate(hiposialianoest = datas[-train_ind,]$hiposialianoest)
```

EDA para data real

```{r}
xeros.real1 %>% plot_bar(by='hiposialianoest', by_position='dodge',nrow = 4,title = 'NIVEL 1')
xeros.real2 %>% plot_bar(by='hiposialianoest', by_position='dodge',nrow = 5,ncol = 4,title = 'NIVEL 2')
xeros.real3 %>% plot_bar(by='hiposialianoest', by_position='dodge',nrow = 5,ncol = 4,title = 'NIVEL 3')
xeros.real4 %>% plot_bar(by='hiposialianoest', by_position='dodge',nrow = 5,ncol = 4,title = 'NIVEL 4')
xeros.real5 %>% plot_bar(by='hiposialianoest', by_position='dodge',nrow = 5,ncol = 4,title = 'NIVEL 5')
```

LR para real data

```{r}
control<-trainControl(method = "CV",number=4,classProbs=TRUE,savePredictions = "all")
set.seed(100384772)
logit.6<- train(hiposialianoest~.,data = xeros.real1,trControl=control,method="glmStepAIC",direction = "back")
logit.7<- train(hiposialianoest~.,data = xeros.real2,trControl=control,method="glmStepAIC",direction = "both")
logit.8<- train(hiposialianoest~.,data = xeros.real3,trControl=control,method="glmStepAIC",direction = "back")
```

```
logit.9<- train(hiposialianoest~.,data =
xeros.real4,trControl=control,method="glmStepAIC",direction = "back")
logit.10<- train(hiposialianoest~.,data =
xeros.real5,trControl=control,method="glmStepAIC",direction = "forward")

plot(pROC::roc(response=ifelse(logit.6$pred$obs
=="Yes",1,0),predictor=logit.6$pred$Yes),main = "logit.6")
plot(pROC::roc(response=ifelse(test.real1$hiposialianoest =="Yes",1,0),predictor=1-
predict(logit.6,test.real1,type = "prob")$Yes),col = "purple",add = TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.7$pred$obs
=="Yes",1,0),predictor=logit.7$pred$Yes),main = "logit.7")
plot(pROC::roc(response=ifelse(test.real2$hiposialianoest
=="Yes",1,0),predictor=predict(logit.7,test.real2,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.8$pred$obs
=="Yes",1,0),predictor=logit.8$pred$Yes),main = "logit.8")
plot(pROC::roc(response=ifelse(test.real3$hiposialianoest
=="Yes",1,0),predictor=predict(logit.8,test.real3,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.9$pred$obs
=="Yes",1,0),predictor=logit.9$pred$Yes),main = "logit.9")
plot(pROC::roc(response=ifelse(test.real4$hiposialianoest
=="Yes",1,0),predictor=predict(logit.9,test.real4,type = "prob")$Yes),col = "purple",add =
TRUE)
```

```
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(logit.10$pred$obs
=="Yes",1,0),predictor=logit.10$pred$Yes),main = "logit.10")
plot(pROC::roc(response=ifelse(test.real5$hiposialianoest
=="Yes",1,0),predictor=predict(logit.10,test.real5,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(logit.6,test.real1,type = "raw"),test.real1$hiposialianoest)
confusionMatrix(predict(logit.7,test.real2,type = "raw"),test.real2$hiposialianoest)
confusionMatrix(predict(logit.8,test.real3,type = "raw"),test.real3$hiposialianoest)
confusionMatrix(predict(logit.9,test.real4,type = "raw"),test.real4$hiposialianoest)
confusionMatrix(predict(logit.10,test.real5,type = "raw"),test.real5$hiposialianoest)
```

DT 2 CON BUCLE para data real

```{r}
library(rpart.plot)
arbolgrid <-  expand.grid(cp=c(0.01))
store_dt_prm2 <- c()
set.seed(100384772)
for (minbu2 in seq(from=10, to=80, by=10)){
  for (minsp2 in seq(from = 4, to= 24, by = 4)){
  print(c(minbu2,minsp2))
  cat("\n")
  arbol6 <-  train(hiposialianoest~.,data = xeros.real1,method="rpart",minsplit =
minsp2,minbucket=minbu2,trControl=control,tuneGrid=arbolgrid)
  arbol7 <-  train(hiposialianoest~.,data = xeros.real2,method="rpart",minsplit =
minsp2,minbucket=minbu2,trControl=control,tuneGrid=arbolgrid)
  arbol8 <-  train(hiposialianoest~.,data = xeros.real3,method="rpart",minsplit =
minsp2,minbucket=minbu2,trControl=control,tuneGrid=arbolgrid)
```

```
  arbol9 <-  train(hiposialianoest~.,data = xeros.real4,method="rpart",minsplit =
minsp2,minbucket=minbu2,trControl=control,tuneGrid=arbolgrid)
  arbol10 <-  train(hiposialianoest~.,data = xeros.real5,method="rpart",minsplit =
minsp2,minbucket=minbu2,trControl=control,tuneGrid=arbolgrid)
  store_dt_prm2<-
data.frame(rbind(store_dt_prm2,c(minbu2,minsp2,arbol6$results[[2]],arbol7$results[[2]],arbo
l8$results[[2]],arbol9$results[[2]],arbol10$results[[2]])))
  }
}
names(store_dt_prm2) <- c('minbucket','minsplit','Accuracy in dt6','Accuracy in
dt7','Accuracy in dt8','Accuracy in dt9','Accuracy in dt10')
print(store_dt_prm2[which.max(store_dt_prm2$`Accuracy in dt6`),])
print(store_dt_prm2[which.max(store_dt_prm2$`Accuracy in dt7`),])
print(store_dt_prm2[which.max(store_dt_prm2$`Accuracy in dt8`),])
print(store_dt_prm2[which.max(store_dt_prm2$`Accuracy in dt9`),])
print(store_dt_prm2[which.max(store_dt_prm2$`Accuracy in dt10`),])

set.seed(100384772)
dt6 <- train(hiposialianoest~.,data = xeros.real1,method="rpart",minsplit =
20,minbucket=60,trControl=control,tuneGrid=arbolgrid)
dt7 <- train(hiposialianoest~.,data = xeros.real2,method="rpart",minsplit =
20,minbucket=80,trControl=control,tuneGrid=expand.grid(cp=0.02111))
dt8 <- train(hiposialianoest~.,data = xeros.real3,method="rpart",minsplit =
24,minbucket=20,trControl=control,tuneGrid=expand.grid(cp=0.02222))
dt9 <- train(hiposialianoest~.,data = xeros.real4,method="rpart",minsplit =
20,minbucket=60,trControl=control,tuneGrid=expand.grid(cp=0.02111))
dt10 <- train(hiposialianoest~.,data = xeros.real5,method="rpart",minsplit =
16,minbucket=70,trControl=control,tuneGrid=expand.grid(cp=0.025))

plot(pROC::roc(response=ifelse(dt6$pred$obs =="Yes",1,0),predictor=dt6$pred$Yes),main =
"Decision Tree 6")
plot(pROC::roc(response=ifelse(test.real1$hiposialianoest =="Yes",1,0),predictor=1-
predict(dt6,test.real1,type = "prob")$Yes),col = "purple",add = TRUE)
```

```
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt7$pred$obs =="Yes",1,0),predictor=dt7$pred$Yes),main =
"Decision Tree 7")
plot(pROC::roc(response=ifelse(test.real2$hiposialianoest
=="Yes",1,0),predictor=predict(dt7,test.real2,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt8$pred$obs =="Yes",1,0),predictor=dt8$pred$Yes),main =
"Decision Tree 8")
plot(pROC::roc(response=ifelse(test.real3$hiposialianoest
=="Yes",1,0),predictor=predict(dt8,test.real3,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt9$pred$obs =="Yes",1,0),predictor=dt9$pred$Yes),main =
"Decision Tree 9")
plot(pROC::roc(response=ifelse(test.real4$hiposialianoest
=="Yes",1,0),predictor=predict(dt9,test.real4,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(dt10$pred$obs
=="Yes",1,0),predictor=dt10$pred$Yes),main = "Decision Tree 10")
plot(pROC::roc(response=ifelse(test.real5$hiposialianoest
=="Yes",1,0),predictor=predict(dt10,test.real5,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
confusionMatrix(predict(dt6,test.real1,type = "raw"),test.real1$hiposialianoest)
confusionMatrix(predict(dt7,test.real2,type = "raw"),test.real2$hiposialianoest)
confusionMatrix(predict(dt8,test.real3,type = "raw"),test.real3$hiposialianoest)
confusionMatrix(predict(dt9,test.real4,type = "raw"),test.real4$hiposialianoest)
confusionMatrix(predict(dt10,test.real5,type = "raw"),test.real5$hiposialianoest)

rpart.plot(dt6$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 1")
rpart.plot(dt7$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 2")
rpart.plot(dt8$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 3")
rpart.plot(dt9$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 4")
rpart.plot(dt10$finalModel,type = 2,extra = 104,fallen.leaves = TRUE, main = "level 5")

plot(varImp(dt6),main = "NIVEL 1");plot(varImp(dt7),main = "NIVEL
2");plot(varImp(dt8),main = "NIVEL 3")
plot(varImp(dt9),main = "NIVEL 4");plot(varImp(dt10),main = "NIVEL 5")
```

SVM

```{r}
library(e1071)
SVMgrid<-expand.grid(C= c(5^c(-4:2),10), sigma = 5^c(-3:2))
#SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2),degree=c(2,3),scale=c(0.1,0.5,1,2,5))
#SVMgrid<-expand.grid(C=5^c(-3:2))
set.seed(100384772)
svm.6<- train(hiposialianoest~.,data =
xeros.real1,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.7<- train(hiposialianoest~.,data =
xeros.real2,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.8<- train(hiposialianoest~.,data =
xeros.real3,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
svm.9<- train(hiposialianoest~.,data =
xeros.real4,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)
```

```
svm.10<- train(hiposialianoest~.,data =
xeros.real5,method="svmRadial",trControl=control,tuneGrid=SVMgrid,verbose=FALSE)

plot(pROC::roc(response=ifelse(svm.6$pred$obs
=="Yes",1,0),predictor=svm.6$pred$Yes),main = "SVM 6")
plot(pROC::roc(response=ifelse(test.real1$hiposialianoest
=="Yes",1,0),predictor=predict(svm.6,test.real1,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.7$pred$obs
=="Yes",1,0),predictor=svm.7$pred$Yes),main = "SVM 7")
plot(pROC::roc(response=ifelse(test.real2$hiposialianoest
=="Yes",1,0),predictor=predict(svm.7,test.real2,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.8$pred$obs
=="Yes",1,0),predictor=svm.8$pred$Yes),main = "SVM 8")
plot(pROC::roc(response=ifelse(test.real3$hiposialianoest
=="Yes",1,0),predictor=predict(svm.8,test.real3,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(svm.9$pred$obs
=="Yes",1,0),predictor=svm.9$pred$Yes),main = "SVM 9")
plot(pROC::roc(response=ifelse(test.real4$hiposialianoest
=="Yes",1,0),predictor=predict(svm.9,test.real4,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
plot(pROC::roc(response=ifelse(svm.10$pred$obs
=="Yes",1,0),predictor=svm.10$pred$Yes),main = "SVM 10")
plot(pROC::roc(response=ifelse(test.real5$hiposialianoest
=="Yes",1,0),predictor=predict(svm.10,test.real5,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(svm.6,test.real1,type = "raw"),test.real1$hiposialianoest)
confusionMatrix(predict(svm.7,test.real2,type = "raw"),test.real2$hiposialianoest)
confusionMatrix(predict(svm.8,test.real3,type = "raw"),test.real3$hiposialianoest)
confusionMatrix(predict(svm.9,test.real4,type = "raw"),test.real4$hiposialianoest)
confusionMatrix(predict(svm.10,test.real5,type = "raw"),test.real5$hiposialianoest)

plot(svm.6, main = "Nivel 1");plot(svm.7,main = "Nivel 2");plot(svm.8,main = "Nivel
3");plot(svm.9,main = "Nivel 4");plot(svm.10,main = "Nivel 5");
explainer_svm7 <- DALEX::explain(svm.7, data = test.real2 %>% dplyr::select(-
hiposialianoest),
                         y = test.real2$hiposialianoest, label = "svm")
explainer_svm7 %>%
    DALEX::model_profile(variables = c(names(xeros.real2))[1:16],variable_type =
"categorical")%>%plot()
```

XGBOOST para data real

```{r}
#Early stopping
xgbmgrid_es<-expand.grid(eta=c(0.1),
 min_child_weight=c(0.5),
 nrounds=c(200,250,300,500,1000,2000,3000,5000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)
x.es6<- train(hiposialianoest~.,data = xeros.real1,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
```

```
x.es7<- train(hiposialianoest~.,data = xeros.real2,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es8<- train(hiposialianoest~.,data = xeros.real3,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es9<- train(hiposialianoest~.,data = xeros.real4,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
x.es10<- train(hiposialianoest~.,data = xeros.real5,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid_es,verbose=FALSE)
plot(x.es6);plot(x.es7);plot(x.es8);plot(x.es9);plot(x.es10)

xgbmgrid<-expand.grid(eta=c(0.1,0.05,0.03,0.01,0.001),
 min_child_weight=c(0.5,1,2),
 nrounds=c(200,250,300,500,1000),
  max_depth=6,gamma=0,colsample_bytree=1,subsample=1)

set.seed(100384772)
xgbm6<- train(hiposialianoest~.,data = xeros.real1,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid,verbose=FALSE)
xgbm7<- train(hiposialianoest~.,data = xeros.real2,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid,verbose=FALSE)
xgbm8<- train(hiposialianoest~.,data = xeros.real3,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid,verbose=FALSE)
xgbm9<- train(hiposialianoest~.,data = xeros.real4,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid,verbose=FALSE)
xgbm10<- train(hiposialianoest~.,data = xeros.real5,
method="xgbTree",trControl=control,tuneGrid=xgbmgrid,verbose=FALSE)

plot(pROC::roc(response=ifelse(xgbm6$pred$obs
=="Yes",1,0),predictor=xgbm6$pred$Yes),main = "XGBOOST 6")
plot(pROC::roc(response=ifelse(test.real1$hiposialianoest
=="Yes",1,0),predictor=predict(xgbm6,test.real1,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))
```

```
plot(pROC::roc(response=ifelse(xgbm7$pred$obs
=="Yes",1,0),predictor=xgbm7$pred$Yes),main = "XGBOOST 7")
plot(pROC::roc(response=ifelse(test.real2$hiposialianoest
=="Yes",1,0),predictor=predict(xgbm7,test.real2,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm8$pred$obs
=="Yes",1,0),predictor=xgbm8$pred$Yes),main = "XGBOOST 8")
plot(pROC::roc(response=ifelse(test.real3$hiposialianoest
=="Yes",1,0),predictor=predict(xgbm8,test.real3,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm9$pred$obs
=="Yes",1,0),predictor=xgbm9$pred$Yes),main = "XGBOOST 9")
plot(pROC::roc(response=ifelse(test.real4$hiposialianoest
=="Yes",1,0),predictor=predict(xgbm9,test.real4,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

plot(pROC::roc(response=ifelse(xgbm10$pred$obs
=="Yes",1,0),predictor=xgbm10$pred$Yes),main = "XGBOOST 10")
plot(pROC::roc(response=ifelse(test.real5$hiposialianoest
=="Yes",1,0),predictor=predict(xgbm10,test.real5,type = "prob")$Yes),col = "purple",add =
TRUE)
legend(x= "bottomright", legend = c("ROC de validacion","ROC de test"), col =
c("black","purple"), lwd = 2, lty = 0,pch = c(19,19))

confusionMatrix(predict(xgbm6,test.real1,type = "raw"),test.real1$hiposialianoest)
```

```
confusionMatrix(predict(xgbm7,test.real2,type = "raw"),test.real2$hiposialianoest)
confusionMatrix(predict(xgbm8,test.real3,type = "raw"),test.real3$hiposialianoest)
confusionMatrix(predict(xgbm9,test.real4,type = "raw"),test.real4$hiposialianoest)
confusionMatrix(predict(xgbm10,test.real5,type = "raw"),test.real5$hiposialianoest)

plot(xgbm6, main = "Nivel 1");plot(xgbm7, main = "Nivel 2");plot(xgbm8, main = "Nivel 3");plot(xgbm9, main = "Nivel 4");plot(xgbm10, main = "Nivel 5");
```

#Comparacion de modelos Nivel 1 xerostomia

```{r}
listclass1<-c("SistemaA",
"SistemaB" ,"SistemaD" ,"SistemaG" ,"SistemaL" ,"SistemaM" ,"SistemaN" ,"SistemaR",
"SistemaS")
source("E:/UCM/TECNICAS DE MACHINE LEARNING/TodosLosProgramasYdatasetsRArboles/cruzada xgboost binaria_ultimate.R")
source("E:/UCM/TECNICAS DE MACHINE LEARNING/SVM/svm scripts/cruzada SVM binaria RBF_ultimate.R")
source("E:/UCM/TECNICAS DE MACHINE LEARNING/TodosLosProgramasYdatasetsRArboles/cruzadas avnnet y glmAIC binaria_ultimate.R")
source("E:/UCM/TECNICAS DE MACHINE LEARNING/TodosLosProgramasYdatasetsRArboles/cruzada arbolbin_ultimate.R")
medias1.1<-cruzadalogistica(data=datas_0var1,vardep="Xerostomía",
  listconti=c(),listclass=c( "SistemaD", "SistemaR", "SistemaL" ,"SistemaN"),
  grupos=4,sinicio=1234,repe=5)
medias1.1$modelo="Regresion Logistica"

medias1.2<-cruzadaarbolbin(data=datas_0var1,vardep="Xerostomía",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,cp=c(0.01), minbucket = 10, minsplit = 12)
medias1.2$modelo="Arbol de decision"
```

```
medias1.3<-cruzadaSVMbinRBF(data=datas_0var1,vardep="Xerostomía",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,
  C=0.0016,sigma=25)
medias1.3$modelo="SVM"

medias1.4<-cruzadaxgbmbin(data=datas_0var1,vardep="Xerostomía",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.1,nrounds=1000,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias1.4$modelo="XGBOOST"
union1 <- c()
union1<-rbind(medias1.1[2:6],medias1.2[2:6],medias1.3,medias1.4[2:6])
par(cex.axis=0.8)
boxplot(data=union1,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union1,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union1,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union1,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 2 xerostomia

```{r}
listclass2=c("A02", "A10" ,"A11", "A12", "B01", "C03", "C07",
"C08" ,"C09" ,"C10" ,"G04", "N02" ,"N05" ,"N06", "R03" ,"S01")
medias2.1<-cruzadalogistica(data=datas_0var2,vardep="Xerostomía",
  listconti=c(),listclass=c( "A10", "B01", "C03" ,"C07" ,"C08" ,"N06" ,"R03"),
  grupos=4,sinicio=1234,repe=5)
medias2.1$modelo="Regresion Logistica"

medias2.2<-cruzadaarbolbin(data=datas_0var2,vardep="Xerostomía",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,cp=c(0.01), minbucket = 60, minsplit = 4)
medias2.2$modelo="Arbol de decision"
```

```
medias2.3<-cruzadaSVMbinRBF(data=datas_0var2,vardep="Xerostomía",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,
  C=25,sigma=0.2)
medias2.3$modelo="SVM"

medias2.4<-cruzadaxgbmbin(data=datas_0var2,vardep="Xerostomía",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=1,eta=0.1,nrounds=300,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias2.4$modelo="XGBOOST"
union2 <- c()
union2<-rbind(medias2.1[2:6],medias2.2[2:6],medias2.3,medias2.4[2:6])
par(cex.axis=0.8)
boxplot(data=union2,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union2,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union2,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union2,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 3 xerostomia

```{r}
listclass3=c("A02B", "A10B", "A12A" ,"B01A", "C07A" ,"C08C" ,"C09A", "C09B",
"C09C", "C09D", "C10A", "G04C" ,"N02A" ,"N02B", "N05B", "N06A")
medias3.1<-cruzadalogistica(data=datas_0var3,vardep="Xerostomía",
  listconti=c(),listclass=c( "A12A", "B01A", "C07A", "C08C", "C09A", "C09C", "N02A",
"N06A"),
  grupos=4,sinicio=1234,repe=5)
medias3.1$modelo="Regresion Logistica"

medias3.2<-cruzadaarbolbin(data=datas_0var3,vardep="Xerostomía",
  listconti=c(),listclass=listclass3,
```

```
  grupos=4,sinicio=1234,repe=5,cp=c(0.01), minbucket = 80, minsplit = 16)
medias3.2$modelo="Arbol de decision"

medias3.3<-cruzadaSVMbinRBF(data=datas_0var3,vardep="Xerostomía",
  listconti=c(),listclass=listclass3,
  grupos=4,sinicio=1234,repe=5,
  C=5,sigma=0.2)
medias3.3$modelo="SVM"

medias3.4<-cruzadaxgbmbin(data=datas_0var3,vardep="Xerostomía",
  listconti=c(),listclass=listclass3,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.05,nrounds=1000,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias3.4$modelo="XGBOOST"
union3 <- c()
union3<-rbind(medias3.1[2:6],medias3.2[2:6],medias3.3,medias3.4[2:6])
par(cex.axis=0.8)
boxplot(data=union3,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union3,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union3,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union3,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 4 xerostomia

```{r}
listclass4<-c("A02BC", "A10BA", "A10BD", "B01AC" ,"C07AB", "C08CA" ,"C09AA",
"C09BA" ,"C09CA", "C09DA" ,"C09DB", "C10AA", "G04CA", "N02BB", "N02BE",
"N05BA", "N06AB")
medias4.1<-cruzadalogistica(data=datas_0var4,vardep="Xerostomía",
  listconti=c(),listclass=c( "A02BC", "B01AC", "C07AB" ,"C08CA", "C09AA", "C09CA",
"N02BB", "N06AB"),
  grupos=4,sinicio=1234,repe=5)
medias4.1$modelo="Regresion Logistica"
```

```
medias4.2<-cruzadaarbolbin(data=datas_0var4,vardep="Xerostomía",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,cp=c(0.025), minbucket = 80, minsplit = 4)
medias4.2$modelo="Arbol de decision"

medias4.3<-cruzadaSVMbinRBF(data=datas_0var4,vardep="Xerostomía",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,
  C=1,sigma=0.2)
medias4.3$modelo="SVM"

medias4.4<-cruzadaxgbmbin(data=datas_0var4,vardep="Xerostomía",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.1,nrounds=500,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias4.4$modelo="XGBOOST"
union4 <- c()
union4<-rbind(medias4.1[2:6],medias4.2[2:6],medias4.3,medias4.4[2:6])
par(cex.axis=0.8)
boxplot(data=union4,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union4,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union4,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union4,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 5 xerostomia

```{r}
listclass5<-c("A02BC01", "A10BA02" ,"A12AXP1", "B01AA07", "B01AC06","C08CA01",
"C09AA02" ,"C09AA03", "C09BA03" ,"C09DB01", "C09DB02", "D09DB06", "C10AA01",
"C10AA05", "G04CA02","N02BB02" ,"N02BE01", "N05BA06" ,"N05BA08")
medias5.1<-cruzadalogistica(data=datas_0var5,vardep="Xerostomía",


```

```
  listconti=c(),listclass=c("B01AC06","B01AA07" ,"C09AA02", "C08CA01",
"A12AXP1" ,"N02BB02", "C09DB02", "D09DB06" ),
  grupos=4,sinicio=1234,repe=5)
medias5.1$modelo="Regresion Logistica"

medias5.2<-cruzadaarbolbin(data=datas_0var5,vardep="Xerostomía",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,cp=c(0.0135), minbucket = 50, minsplit = 24)
medias5.2$modelo="Arbol de decision"

medias5.3<-cruzadaSVMbinRBF(data=datas_0var5,vardep="Xerostomía",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,
  C=5,sigma=0.2)
medias5.3$modelo="SVM"

medias5.4<-cruzadaxgbmbin(data=datas_0var5,vardep="Xerostomía",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.1,nrounds=1000,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias5.4$modelo="XGBOOST"
union5 <- c()
union5<-rbind(medias5.1[2:6],medias5.2[2:6],medias5.3,medias5.4[2:6])
par(cex.axis=0.8)
boxplot(data=union5,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union5,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union5,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union5,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 1 hiposialia

```{r}
medias6.1<-cruzadalogistica(data=xeros.real1,vardep="hiposialianoest",
```

```
  listconti=c(),listclass=c("SistemaG","SistemaR" ),
  grupos=4,sinicio=1234,repe=5)
medias6.1$modelo="Regresion Logistica"

medias6.2<-cruzadaarbolbin(data=xeros.real1,vardep="hiposialianoest",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,cp=c(0.01), minbucket = 60, minsplit = 20)
medias6.2$modelo="Arbol de decision"

medias6.3<-cruzadaSVMbinRBF(data=xeros.real1,vardep="hiposialianoest",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,
  C=1,sigma=0.04)
medias6.3$modelo="SVM"

medias6.4<-cruzadaxgbmbin(data=xeros.real1,vardep="hiposialianoest",
  listconti=c(),listclass=listclass1,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.001,nrounds=1000,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias6.4$modelo="XGBOOST"
union6 <- c()
union6<-rbind(medias6.1[2:6],medias6.2[2:6],medias6.3,medias6.4[2:6])
par(cex.axis=0.8)
boxplot(data=union6,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union6,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union6,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union6,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 2 hiposialia

```{r}
medias7.1<-cruzadalogistica(data=xeros.real2,vardep="hiposialianoest",
  listconti=c(),listclass=c("A02", "A10" ,"A11", "C03", "C07" ,"C08" ,"C10", "N02" ,"N05" ),


```

```
  grupos=4,sinicio=1234,repe=5)
medias7.1$modelo="Regresion Logistica"

medias7.2<-cruzadaarbolbin(data=xeros.real2,vardep="hiposialianoest",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,cp=c(0.02111), minbucket = 80, minsplit = 20)
medias7.2$modelo="Arbol de decision"

medias7.3<-cruzadaSVMbinRBF(data=xeros.real2,vardep="hiposialianoest",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,
  C=10,sigma=0.2)
medias7.3$modelo="SVM"

medias7.4<-cruzadaxgbmbin(data=xeros.real2,vardep="hiposialianoest",
  listconti=c(),listclass=listclass2,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=1,eta=0.1,nrounds=500,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias7.4$modelo="XGBOOST"
union7<- c()
union7<-rbind(medias7.1[2:6],medias7.2[2:6],medias7.3,medias7.4[2:6])
par(cex.axis=0.8)
boxplot(data=union7,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union7,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union7,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union7,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 3 hiposialia

```{r}
medias8.1<-cruzadalogistica(data=xeros.real3,vardep="hiposialianoest",
  listconti=c(),listclass=c("A10B", "C07A" ,"C08C" ,"C09A" ,"C09B" ,"C09C",
"C09D" ,"C10A", "N02B", "N05B" ),


```

```
  grupos=4,sinicio=1234,repe=5)
medias8.1$modelo="Regresion Logistica"

medias8.2<-cruzadaarbolbin(data=xeros.real3,vardep="hiposialianoest",
  listconti=c(),listclass=listclass3,
  grupos=4,sinicio=1234,repe=5,cp=c(0.02222), minbucket = 20, minsplit = 24)
medias8.2$modelo="Arbol de decision"

medias8.3<-cruzadaSVMbinRBF(data=xeros.real3,vardep="hiposialianoest",
  listconti=c(),listclass=listclass3,
  grupos=4,sinicio=1234,repe=5,
  C=1,sigma=25)
medias8.3$modelo="SVM"

medias8.4<-cruzadaxgbmbin(data=xeros.real3,vardep="hiposialianoest",
  listconti=c(),listclass=listclass3,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.03,nrounds=500,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias8.4$modelo="XGBOOST"
union8<- c()
union8<-rbind(medias8.1[2:6],medias8.2[2:6],medias8.3,medias8.4[2:6])
par(cex.axis=0.8)
boxplot(data=union8,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union8,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union8,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union8,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 4 hiposialia

```{r}
medias9.1<-cruzadalogistica(data=xeros.real4,vardep="hiposialianoest",
  listconti=c(),listclass=c("A02BC", "B01AC" ,"C07AB",
"C08CA" ,"C09AA","C09BA" ,"C09DB" ,"C10AA" ,"N02BE", "N05BA", "N06AB" ),
```

```
  grupos=4,sinicio=1234,repe=5)
medias9.1$modelo="Regresion Logistica"

medias9.2<-cruzadaarbolbin(data=xeros.real4,vardep="hiposialianoest",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,cp=c(0.02111), minbucket = 60, minsplit = 20)
medias9.2$modelo="Arbol de decision"

medias9.3<-cruzadaSVMbinRBF(data=xeros.real4,vardep="hiposialianoest",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,
  C=10,sigma=0.2)
medias9.3$modelo="SVM"

medias9.4<-cruzadaxgbmbin(data=xeros.real4,vardep="hiposialianoest",
  listconti=c(),listclass=listclass4,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=0.5,eta=0.1,nrounds=500,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias9.4$modelo="XGBOOST"
union9<- c()
union9<-rbind(medias9.1[2:6],medias9.2[2:6],medias9.3,medias9.4[2:6])
par(cex.axis=0.8)
boxplot(data=union9,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union9,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union9,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union9,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

#Nivel 5 hiposialia

```{r}
medias10.1<-cruzadalogistica(data=xeros.real5,vardep="hiposialianoest",
  listconti=c(),listclass=c("N02BE01", "B01AA07", "C08CA01", "A02BC01" ,"C10AA01",
"C10AA05"),


```

```
  grupos=4,sinicio=1234,repe=5)
medias10.1$modelo="Regresion Logistica"

medias10.2<-cruzadaarbolbin(data=xeros.real5,vardep="hiposialianoest",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,cp=c(0.025), minbucket = 70, minsplit = 16)
medias10.2$modelo="Arbol de decision"

medias10.3<-cruzadaSVMbinRBF(data=xeros.real5,vardep="hiposialianoest",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,
  C=0.008,sigma=0.2)
medias10.3$modelo="SVM"

medias10.4<-cruzadaxgbmbin(data=xeros.real5,vardep="hiposialianoest",
  listconti=c(),listclass=listclass5,
  grupos=4,sinicio=1234,repe=5,
  min_child_weight=2,eta=0.1,nrounds=1000,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1)
medias10.4$modelo="XGBOOST"
union10<- c()
union10<-rbind(medias10.1[2:6],medias10.2[2:6],medias10.3,medias10.4[2:6])
par(cex.axis=0.8)
boxplot(data=union10,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union10,auc~modelo,main="AUC",col = "light blue")
boxplot(data=union10,sensibilidad~modelo,main="SENSIBILIDAD",col = "green")
boxplot(data=union10,especificidad~modelo,main="ESPECIFICIDAD",col = "yellow")
```

Exportar 1-5 datasets

```{r}
write.csv(rbind.data.frame(xeros.percepcion1,test.percepcion1)%>%select(Xerostomía,listcla
ss1),file = "E:/UCM/TFM/Split Dataset/Xeros1.csv")
```

```
write.csv(rbind.data.frame(xeros.percepcion2,test.percepcion2)%>%select(Xerostomía,listcla
ss2),file = "E:/UCM/TFM/Split Dataset/Xeros2.csv")
write.csv(rbind.data.frame(xeros.percepcion3,test.percepcion3)%>%select(Xerostomía,listcla
ss3),file = "E:/UCM/TFM/Split Dataset/Xeros3.csv")
write.csv(rbind.data.frame(xeros.percepcion4,test.percepcion4)%>%select(Xerostomía,listcla
ss4),file = "E:/UCM/TFM/Split Dataset/Xeros4.csv")
write.csv(rbind.data.frame(xeros.percepcion5,test.percepcion5)%>%select(Xerostomía,listcla
ss5),file = "E:/UCM/TFM/Split Dataset/Xeros5.csv")

write.csv(rbind.data.frame(xeros.real1,test.real1%>%select(hiposialianoest,listclass1)),file =
"E:/UCM/TFM/Split Dataset/Hipos1.csv")
write.csv(rbind.data.frame(xeros.real2,test.real2%>%select(hiposialianoest,listclass2)),file =
"E:/UCM/TFM/Split Dataset/Hipos2.csv")
write.csv(rbind.data.frame(xeros.real3,test.real3%>%select(hiposialianoest,listclass3)),file =
"E:/UCM/TFM/Split Dataset/Hipos3.csv")
write.csv(rbind.data.frame(xeros.real4,test.real4%>%select(hiposialianoest,listclass4)),file =
"E:/UCM/TFM/Split Dataset/Hipos4.csv")
write.csv(rbind.data.frame(xeros.real5,test.real5%>%select(hiposialianoest,listclass5)),file =
"E:/UCM/TFM/Split Dataset/Hipos5.csv")
```

```{r}
pROC::roc(response=ifelse(test.real1$hiposialianoest
=="Yes",1,0),predictor=predict(svm.6,test.real1,type = "prob")$Yes)$auc
pROC::roc(response=ifelse(test.real2$hiposialianoest
=="Yes",1,0),predictor=predict(svm.7,test.real2,type = "prob")$Yes)$auc
pROC::roc(response=ifelse(test.real3$hiposialianoest
=="Yes",1,0),predictor=predict(svm.8,test.real3,type = "prob")$Yes)$auc
pROC::roc(response=ifelse(test.real4$hiposialianoest
=="Yes",1,0),predictor=predict(svm.9,test.real4,type = "prob")$Yes)$auc
pROC::roc(response=ifelse(test.real5$hiposialianoest
=="Yes",1,0),predictor=predict(svm.10,test.real5,type = "prob")$Yes)$auc
```