UNIVERSIDAD
POLITECNICA
DE VALENCIA

Caracterización del perfil de los pacientes con enfermedades respiratorias que no se vacunan contra el virus de la influenza

José Nicolás Granero Moreno
Directora: Dr. Andrea Conchado Peiró

Septiembre de 2021
Departamento de Estadística e Investigación Operativa Aplicadas y Calidad
Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Índice

Índice de cuadros

Índice de figuras

Abreviaturas	Significado
Cort	Corticoesteroides
Vac neum	Vacuna antineumocócica
Hipolipe	Hipolipemiantes
Hosp_any	Hospitalizaciones el año anterior
$D < 1$	Dejó de fumar hace menos de un año
$D > 1$	Dejó de fumar hace más de un año
Md5c	Menos de 5 cigarros por día
Nnc	Nunca
Grp	Gerencia profesional
Ogyt	Ocupaciones gerenciales y técnicas
Snc	Sin clasificar
Anticoag	Anticoagulantes
Obes	Obesidad
Enf_cor	Enfermedad cardiovascular
Enf_renal	Enfermedad Renal
$Neuromusc\_des$	Desorden neuromuscular
$Hipo\_oral$	Hipoglucemiante oral
Enf_cer	Enfermedad Cerebrovascular
$Arterial\_per$	Arteriopatía periférica

Tabla 1: Tabla de abreviaturas

Resumen

La vacunación contra la gripe es muy importante, ya que causa muchas muertes todos los años. Teniendo esto en cuenta, se han utilizado unos datos cedidos por la empresa FISABIO para analizar las personas que no se vacunan contra la gripe aún perteneciendo a un colectivo de riesgo, como son las personas con enfermedades respiratorias. Para conseguir caracterizar este grupo, se dividirán los participantes en tres grupos de edad primero (menores de 18 años, adultos entre 18 y 65 años y mayores de 65 años, analizándose los dos últimos grupos) y se utilizarán métodos de Análisis Multivariante y Minería de Datos como la regresión logística, los árboles de clasificación y el bosque aleatorio (random forest) comparándose los métodos finalmente entre ellos. Finalmente, se ha podido observar como la gente que no se vacuna son las personas que no tienen ninguna enfermedad, exceptuando el problema respiratorio, o tratamiento.

Palabras claves: Vacunación, Gripe, Regresión logística, Minería de Datos, Árbol de Clasificación, Bosque Aleatorio.

Resum

La vacunació contra la grip és molt important, ja que causa moltes morts cada any. Tenint en compte això, s'han utilitzat unes dades cedides per l'empresa FISABIO per analitzar les persones que no es vacunen contra la grip tot i pertànyer a un colÂ·lectiu de risc, com són les persones amb malalties respiratòries. Per aconseguir caracteritzar aquest grup, es dividiran en tres grups de edat primer (menors de 18 anys, adults entre 18 i 65 anys i majors de 65 anys, analitzant els dos últims grups) i s'utilitzaran mètodes d'Anàlisi Multivariant i Mineria de Dades com la regressió logística, els arbres de classificació i el bosc aleatori (random forest) comparant els mètodes finalment entre ells. Finalment, s'ha pogut observar com la gent que no es vacuna són les persones que no tenen cap malaltia, exceptuant el problema respiratori, o tractament.

Paraules claus: Vacunació, Grip, Regresió logística, Minería de Dades, Arbre de Clasificació, Bosc Aleatori.

Abstract

Flu vaccination is very important, as it causes many deaths every year. Taking this into account, data provided by the FISABIO company has been used to analyze people who do not get vaccinated against the flu even though they belong to a risk group, such as people with respiratory diseases. In order to characterize this group, they will be divided into three age groups first (under 18 years old, adults between 18 and 65 years old and over 65 years old, analyzing the last two groups) and Multivariate Analysis and Data Mining methods will be used as the logistic regression, the classification trees and the random forest, finally comparing the methods between them. Finally, it has been observed how

the people who do not get vaccinated are the people who do not have any disease, except the respiratory problem, or treatment.

Keywords: Vaccination, Influenza, Logistic Regression, Data Mining, Classification Tree, Random Forest.

1. Introducción

La vacunación contra la gripe (influenza) es una práctica habitualmente recomendada por las instituciones sanitarias. Es especialmente importante para determinados grupos de riesgo por el riesgo de complicaciones respiratorias según las Organización Mundial de la Salud (WHO) [1]. Puesto que cada año se producen mutaciones en los virus de la influenza y la composición de las vacunas se modifica, la vacunación debe realizarse de forma anual. No obstante, la vacuna contra la influenza es una decisión a criterio del paciente y en ella intervienen aspectos como el estado físico, la convivencia con niños o enfermos, así como las características personales del paciente incluyendo su actitud hacia las vacunas. La participación estas personas pertenecientes a grupos de riesgo en las campañas de vacunación es vital para el sistema sanitario. Las epidemias de gripe (influenza) causan entre 3 y 5 millones de casos de enfermedad y entre 290000 y 650000 muertes anuales que podrían evitarse. [2].

Cada año, las autoridades sanitarias realizan una campaña informativa dirigida a la población general, destinada a fomentar la participación en campaãs de vacunación entre las personas pertenecientes a grupos de riesgo. A pesar de ello, anualmente se realiza un gran número de ingresos hospitalarios por complicaciones respiratorias derivadas de gripe, que podrían haberse evitado mediante la vacunación. La efectividad de esta vacuna en distintas poblaciones está ampliamente probada según los resultados de la Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO) [3, 4, 5].

El número total de enfermedades respiratorias crónicas se ha incrementado en un 39,5 % desde 1990 hasta 2017 que aunque la tasa de prevalencia estandarizada por edad y la tasa de incidencia estandarizada por edad actualmente muestran tendencias decrecientes [6]. En particular, España se caracteriza por una alta proporción de muertes por enfermedades respiratorias cróicas (9 %), y en especial, la Enfermedad Pulmonar Obstructiva Crónica (EPOC) que afecta al 10,2 % de la población adulta. Aproximadamente 18000 personas mueren anualmente debido a esta enfermedad, que requiere un 40 % del tiempo de atención al paciente en consultas de pulmonología. Cada año, su diagnosis, mantenimiento y tratamiento cuesta aproximadamente el 2 % del presupuesto sanitario[7].

Adicionalmente se sabe que el hábito tabáquico constituye un factor de riesgo en relación a influenza [8], además de ser una causa de riesgo a la hora de producir ciertas enfermedades crónicas respiratorias. Otros factores que pueden influir son la convivencia con niños que estudian en centros educativos o personas mayores, o circunstancias particulares del paciente.

1.1. Motivación

En base a estos resultados, el presente trabajo plantea que se estuidarán las características comunes de las personas que no se vacunan.

Se espera que los resultados del trabajo amplíen el conocimiento existente sobre las personas que rechazan participar en campañas de vacunación, incluso a pesar de pertenecer a grupos de riesgo. Por otro lado, los resultados esperados tienen importantes aplicaciones prácticas para el diseño de campañas de vacunación dirigidas específicamente a este colectivo.

Este trabajo pretende cubrir una necesidad de la empresa FISABIO, la cual cedió los datos, caracterizando las personas que no se vacunan contra la gripe.

Anteriormente se ha publicado un estudio similar entre personas mayores de 65 años [9] en el cual se analiza el perfil de las personas que sí participan en las campañas de vacunación mediante un modelo de regresión logística. Este trabajo aporta valor al estudio de la cuestión por dos razones. En primer lugar, pone el foco sobre la caracterizaci'on del perfil de pacientes en riesgo que deciden no vacunarse. Asimismo, se aplican técnicas estadísitcas más novedosas pertenecientes al ámbito del análisis multivariante y la minería de datos que se completa con una selección y validación del modelo más adecuado.

2. Objetivos

El objetivo general del trabajo es caracterizar el perfil de las personas que no se vacunan contra el virus de la gripe y fueron ingresados por una enfermedad respiratoria, asma o bronquitis.

Este objetivo general puede desglosarse en los siguientes objetivos específicos:

1. Identificar al colectivo de personas que no han asistido a la vacunación contra el virus de la gripe en tres campañas consecutivas y padecen de bronquitis o asma, segmentando por franjas de edad.

2. Examinar la relación entre la pertenencia a este colectivo y características personales y comportamiento del paciente.

3. Estimar y comparar distintos modelos predictivos de pertenencia a este colectivo, en función de factores seleccionados entre las características personas y comportamiento del paciente.

3. Metodología

3.1. Participantes

El estudio parte de una muestra inicial de 12806 personas, todas estas observaciones han sido recopiladas entre las temporadas de vacunación 2015/2016 y 2018/2019. Se han descartado 2018 datos ya que había un dato faltante en la variable de vacunación de la temporada correspondiente al año en curso y 1626 en los cuáles no se tenía la edad de la persona.

Como en el estudio se va a incluir únicamente a la población de adultos, entre los 18 y 65 años, y el de mayores de 65 años, el conjunto de datos final está formado por información de 8281 pacientes (1819 adultos y 6462 mayores de 65 años).

En la submuestra formada por adultos, el porcentaje de mujeres es $49,1\,\%$ y la edad media del grupo es de 55 años. Respecto a las hospitalizaciones del año anterior, se observa 0,62 hospitalizaciones por persona de media durante el año anterior, siendo la mediana 0, el valor del tercer cuartil 1 y el máximo 10, habiendo 2 datos faltantes no teniendose en cuenta esas dos observaciones en los análisis.

En el caso de la población de mayores de 65 años, la edad media es de 80,9 años, siendo el mínimo de 66 años, la mediana de 81 y el máximo de 104 años y el porcentaje de mujeres es de $46,3\,\%$. En cuestión de hospitalizaciones del año anterior, la muestra tuvo una media de 0,7 hospitalizaciones durante el año anterior, habiendo un $60,4\,\%$ que no tuvo ninguna y 7 casos con 10 hospitalizaciones. Al igual que en el caso anterior se descartan 17 observaciones con datos faltantes.

Además, no se ha tenido en cuenta a otras observaciones, ya que se han encontrado datos faltantes en otras variables, como por ejemplo en el hábito tabáquico.

3.2. Variables

Entre las variables disponibles en el conjunto de datos, se ha desestimado utilizar algunas variables como el número de identificación del paciente, la fecha de hospitalización de los pacientes, el hospital o las variables que tienen realación con el nacimiento, ya que no ofrecen ninguna información para el análisis. Otras variables como las variables de embarazo o las variables en relación con la dependencia (excepto la puntuación del índice de Barthel en sí), por el hecho de ser variables que se podrían utilizar para modelos específicos de mujeres embarazadas o de personas con dependencia. Y otras variables como si se han vacunado o no en la temporada 2019/2020 por tener muchos datos fantantes.

Las variables incluidas en los análisis son las siguientes:

Variable	Definición
Temporada	Temporada de vacunación en la que se realizó la encuesta ya que el paciente fue hospitalizado.
Hospital	Hospital en el que fue ingresado el paciente.
Sexo	Sexo del paciente hospitalizado.
Edad	Años del paciente en el momento de la hospitalización.
Peso	Peso del paciente en el momento de la hospitalización
Altura	Altura del paciente en el momento de la hospitalización
Cardiovascular	Enfermedad diagnosticada: Cardiovascular.
Cerebrovascular	Enfermedad diagnosticada: Cerebrovascular.
Arteriopatía periférica	Enfermedad diagnosticada: Arteriopatía periférica.
Asma	Enfermedad diagnosticada: Asma.
Bronquitis	Enfermedad diagnosticada: Bronquitis.
Diabetes	Enfermedad diagnosticada: Diabetes.
Otras Endocrinas	Enfermedad diagnosticada: Otra enfermedad endocrina, que no es la diabetes.
Anemia	Enfermedad diagnosticada: Anemia.
Hepática	Enfermedad diagnosticada: Alguna enfermedad hepática.
Renal	Enfermedad diagnosticada: Alguna enfermedad renal.
Desorden Neuromuscular	Enfermedad diagnosticada: Desorden neuromuscular.
Neoplasia	Enfermedad diagnosticada: Neoplasia.
Autoinmune	Enfermedad diagnosticada: Autoinmune.
Demencia	Enfermedad diagnosticada: Demencia.
Antihypertensivos	Tratamiento: Antihypertensivos.
Anticoagulantes	Tratamiento: Anticoagulantes.
Agregación Antiplaquetaria	Tratamiento: Agregación antiplaquetaria.
Hipolipemiantes	Tratamiento: Hipolipemiantes.
Insulina	Tratamiento: Insulina.
Hipoglucemiante Oral	Tratamiento: Hipoglucemiante de forma oral.
Inmunosupresores	Tratamiento: Inmunosupresores.
Corticoesteroides	Tratamiento: Corticoesteroides.
BMI	Índice de masa corporal del paciente en el momento de la hospotalización.
Obesida	Paciente es obeso y en caso afirmativo en qué grado.
Embarazo	Paciente embarazada o no.
Fumar	Persona fumadora y en caso afirmativo, cuánto. En caso negativo, si fumó, hace cuanto lo dejó.
Niños	Si el paciente tiene o no contacto estrecho con niños.
Clase Social	La clase social a la que pertenece el paciente hospitalizada basado en el trabajo que desempeña.
Índice barthel	Índice de barthel del paciente.
Consultas 3m	Las consultas en los últimos 3 meses del paciente.

Ingresos 12	Número de inresos en los últimos 12 meses del paciente.
Hospitalizaciones año anterior	Número de hospitalizaciones en el año anterior del paciente.
Huevo	Si el paciente tiene alergia al huevo.
r2018sfv	Si la parsona se vacunó o no la temporada 18/19.
r2017sfv	Si la parsona se vacunó o no la temporada 17/18.
r2016sfv	Si la parsona se vacunó o no la temporada 16/17.
r2015sfv	Si la parsona se vacunó o no la temporada 15/16.
r2014sfv	Si la parsona se vacunó o no la temporada 14/15.
r2013sfv	Si la parsona se vacunó o no la temporada 13/14.
Vacuna Neumocócica	Si el paciente tomó la vacuna antineumocócica o no.

Tabla 2: Variables introducidas en los análisis

Además de las citadas anteriormente, se han creado otras variables necesarias para los análisis tales como ver si se han vacunado en la temporada actual o no, si se vacunaron la temporada anterior o no, si se vacunaron hace dos temporadas o no, si se han vacunado en las últimas 3 temporadas y la última variable teniendo en cuenta si no se vacunaron ninguna temporada teniendo una enfermedad respiratoria, variable de interés (ser reticente a la vacunación a pesar de padecer una enfermedad respiratoria).

Para las tres primeras variables, se ha identificado cuál era la temporada en la que se había hospitalizado a la persona y las variables de esa temporada, la temporada anterior y la de dos temporadas anteriores. Se ha codificado el valor de la temporada actual, el valor de la temporada anterior y el valor de dos temporadas anteriores como los valores de las variables de las temporadas.

Con el fin de construir la variable relativa a la vacunación o negación a vacunarse durante las últimas tres temporadas, se han utilizado las tres variables creadas anteriormente. En caso de que en las tres variables la respuesta sea negativa, se codifica el valor como "No"(haciendo referencia a que no se han vacunado nunca). En caso de que alguna sea afirmativa, se codificará como "Si"(haciendo referencia a que se han vacunado en alguna ocasión).

Para finalizar con la variable de interés (ser reticente a la vacuna teniendo una enfermedad respiratoria), se ha creado teniendo en cuenta si una persona no se ha vacunado nunca y ha ingresado por una enfermedad respiratoria, ya sea asma o bronquitis, se ha codificado con una respuesta afirmativa, "Si", y en caso contrario como "No".

3.3. Análisis de datos

En este apratado, se describen las técnicas de análisis de datos empleadas, los modelos utilizados así como la validación y comparación entre modelos.

Para este propósito, se ha utilizado el software estadístico R[10] y para leer los archivos de Excel, el paquete readxl [11].

En primer lugar, se ha realizado un análisis univariante para saber el porcentaje de personas con una cierta enfermedad, el porcentaje de personas con un cierto tratamiento y otras características generales de la población. En segundo lugar, se ha realizado un análisis bivariante, para las poblaciones de adultos y mayores de 65 años.

En tercer lugar, se han ejecutado los modelos de clasificación para ambas poblaciones con las variables significativas del análisis bivariante y las variables del peso, altura y el número de hospitalizaciones del año anterior para observar qué variables se consideran significativas, junto con algunos estadísticos.

En el caso del ánalisis de correspondencias múltiples se observarán como se distribuyen las variables y cuales son más cercanas a nuestra variable de interés.

Para finalizar, se ha decidido hacer una comparación entre los modelos que se consideren posteriormente para observar cual de ellos es el mejor de todos para nuestro interés, el cual es predecir la clase de reticentes.

A continuación, se explicará en que consisten los modelos y cómo se ha realizado la comparación entre estos modelos.

Regresión Logística

Ésta es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados, GLM, que usa como función de enlace la función logit.

Se han ejecutado dos modelos de regresión, uno con las variables que se han considerado significativas del análisis bivariante más las citadas anteriormente y otro con un enfoque `stepwise`. El primero, se ha realizado con el comando `glm` y el segundo, con el comando `step`, ambas funciones dentro del paquete stats (dentro de R por defecto) y creándose variables `dummy`, con el paquete `fastDummies` [12].

Para medir la bondad de ajuste de estos modelos, se ha utilizado la tasa de acierto, el porcentaje de falso negativo, el porcentaje de verdadero positivo (estos se explicarán más adelante) y el pseudo-R^2. Este último se define como la $-2 \ln \frac{V_0}{V_r}$, siendo V_0 la verosimilitud del modelo sin variables significativas y V_r, la verosimilitud del modelo construido.

Árboles de Clasificación

El árbol de clasificación es una técnica de aprendizaje supervisado de clasificación que consiste en crear 'ramificaciones' y 'hojas' de manera que cada ramificación o nodo especifica el test de algún atributo concreto. Por lo que según el camino que siga el individuo cuya variable categórica se está tratando de predecir, se acaba en una 'hoja' u otra, lo que indicará la clase que se debe asociar a la predicción.

Respecto a los modelos basados en árboles, se han ejecutado tres árboles de clasificación. Para realizar los modelos de los árboles de clasificación, se han utilizado el paquete de `DMwR2` [13] y `rpart` [14] dibujándolos con el paquete `rpart.plot` [15].

Los dos primeros árboles de clasificación se han ejecutado con el paquete `rpart`, utilizando la función con el mismo nombre. En el primero, se ha dejado el parámetro de complejidad (cp) por defecto y en el segundo para realizar un árbol más completo y con más ramificaciones con el parámetro de complejidad igual a $0,001$. Siendo este parámetro la mínima mejora en el modelo necesaria en cada nodo.

Y el tercero es un árbol de clasificación podado con la función rpartXse del paquete `DMwR2`, con el argumento $se = 0,25$.La regla x-SE para la poda posterior de árboles se basa en las estimaciones de validación cruzada del error de los subárboles del árbol cultivado inicialmente, junto con los errores estándar de estas estimaciones. Estos valores se utilizan para seleccionar el modelo de árbol final. Es decir, el árbol seleccionado es el árbol más pequeño con un error estimado menor que $B + xSE$, donde B es la estimación más baja del error y SE es el error estándar de esta estimación B [13].

Bosque Aleatorio

El bosque aleatorio (random forest) consiste en la creación de árboles de clasificación en los que los nodos no se producen a partir del conjunto total de predictores, sino que se producen mediante un subconjunto elegido al azar. Después, se elige la clase por votación de los distintos árboles, es decir, se le asigna la clase más predicha por los distintos árboles.

Para utilizar los modelos del Random Forest, se ha utilizado el paquete `randomForest` [16] y la función con el mismo nombre. En esta función, como se aconseja en el paquete, para los modelos de clasificación, se ha ajustado el argumento del número de predictores elegidos aleatoriamente para cada nodo siendo igual a $\sqrt{número\ de\ variables}$.

A la hora de comparar los modelos, se realizará un hold out repetido. Este proceso consiste en fragmentar la muestra en un conjunto para entrenar el arlgoritmo, que se utilizará posteriormente, y en otro conjunto de validación, que servirá para obtener los estadísitcos de comparación. Este procedimiento se realiza, por el hecho de que un algoritmo entrenado con un conjunto de datos tiende a sobreajustar los datos con los

que se entrena. Con lo cual, se medirá de manera más realista. Este proceso, dividiéndose en una muestra del 70 % de entrenamiento y 30 % de validación, se repetirá en este caso 100 veces, ya que por puro azar un conjunto de datos lo puede ajustar mejor un modelo que en estadísticamente hablando tenga una menor media del estadístico a comparar.

Como la variable a predecir tiene dos clases y la clase de interés es minoritaria respecto a la otra, se remuestreará la clase minoritaria para que hayan los mismos casos de reticentes y de no reticentes. A continuación, se entrenarán los modelos con el conjunto de entrenamiento y se validarán con el conjunto de validación.

Como se ha dicho antes, la variable a predecir tiene dos clases se realizará la matriz de confusión para obtener algunos de los siguientes estadísticos. Así mismo, de todos se obtendrá la media y los Intervalos de Confianza al 95 %.

Figura 1: Matriz de Confusión genérica

tasa de acierto

Esta consiste en obtener cuántos datos que han sido clasificados como correctos, tanto si son de la clase de reticentes como en caso contrario, y después dividirlo respecto el total, es decir, Verdaderos Positivos (TP) más los Verdaderos Negativos (TN) entre la suma de éstas más los Falsos Positivos (FP) y Falsos Negativos (FN).

$$tasa\ de\ acierto = \frac{TP + TN}{TP + TN + FP + FN}$$

Coeficiente de Correlación de Matthews

El coeficiente de correlation de Matthews (también llamado mcc) consigue diferenciar los modelos cuando las clases de positivos y negativos están desbalanceadas. Este coeficiente se define con la siguiente expresión:

$$\frac{TPxTN - FNxFP}{\sqrt{(TP+FP)(FP+FN)(TN+FP)(TN+FN)}}$$

Área bajo la Curva ROC

Este estadístico se ha obtenido con la función de R performance, del paquete ROCR, para la medida `auc` (area under curve), esta medida es independiente del punto de corte. Cuanto más cerca esté el estadístico de 1, mejor será la discriminación y cuanto menor sea ésta, menor capacidad discriminatoria tendrá el modelo. Para calcular este estadístico, se ha utilizado el paquete `ROCR` [17].

Análisis de Correspondencias Múltiples

El análisis de correspondencia múltiple (MCA) es una extensión del análisis de correspondencia que permite analizar el patrón de relaciones de varias variables dependientes categóricas. Como tal, también puede verse como una generalización del análisis de componentes principales cuando las variables a analizar son categóricas en lugar de cuantitativas [18].

Para los ánalisis de correspondencias múltiples, se han utilizado los paquetes FactoMineR [19] y factoextra [20].

4. Resultados

4.1. Análisis de los factores influyentes en la reticencia a la vacuna con enfermedad respiratoria

En este subapartado, se ha realizado un análisis univariante de los datos para ver las características que tienen las personas de la muestra en general para después dividirlo por franjas de edad y realizar los análisis.

En primer lugar, se ha realizado un análisis univariante de éstos. Se observa que en los datos hay un 46,0 % de mujeres , un 9,6 % son personas menores de 18 años, un 19,8 % son personas entre 18 y 65 años y el resto, mayores de 65.

También, se aprecia un $2,2\,\%$ personas con obesidad morbida y el $19,8\,\%$ son personas con obesidad; hay 14 mujeres embarazadas; un $29,1\,\%$ tienen contacto con niños; hay $0,3\,\%$ de personas con alergia al huevo; un $34,4\,\%$ de personas han sido ingreasadas en los últimos 12 meses; $58,1\,\%$ se han vacunado alguna vez ($10,8\,\%$ en 1 temporada, $14,5\,\%$ en 2 temporadas y $32,8\,\%$ personas se vacunaron todas las temporadas) y $41,9\,\%$ en ninguna temporada, dentro de lo cual hay 916 con enfermedades respiratorias (bronquitis o asma), es decir, hay un $8,5\,\%$ que tiene enfermedades respiratorias y nos se han vacunado nunca. También, hay $56,1\,\%$ de personas que no se pusieron la vacuna antinemocócica registrada. Finalmente, la media de hospitalizaciones en el último año es de 0,5989, la edad media de las personas del estudio es de 67, 7, $14,7\,\%$ no han ido a ninguna consulta con el médico de cabezera en los últimos 3 meses, $12,6\,\%$ personas han ido en una ocasión, $15,1\,\%$ han ido en dos ocasiones, $14,7\,\%$ han ido en tres ocasiones, $48,8\,\%$ han ido en más de 3 ocasiones y el resto de personas no se acuerdan y el $8,5\,\%$, siendo un $17,6\,\%$ para la población de adultos y un $6,8\,\%$.

A continuación, se muestra una tabla de las proporciones de las personas con las distintas enfermedades y los distintos tratamientos:

Enfermedad	Ratio Enfermos
Problemas de Corazón	39,7 %
Enfermedad Cerebro.	5,0 %
Arteriopatía Periperiférica	2,7 %
Asma	9,0 %
Bronquitis	28,1 %
Diabetes	25,3 %
Otros Endocrinos	10,1 %
Anemia	8,6 %
Enfermedad Hepática	3,7 %
Enfermedad Renal Crónica	14,0 %
Autoinmune	1,3 %
Desorden Neuromuscular	4,4 %
Neoplasia	8,3 %
Enfermedad Autoinmune	3,2 %
Demencia	6,4 %

Tabla 3: Porcentaje de personas diagnosticadas con la enfermedad

Tratamiento	Ratio con trat.
Antihipertensivos	50,6 %
Anticoagulantes	18,7 %
Agregación antiplaquetaria	16,3 %
Hipolipemiantes	32,3 %
Insulina	9,8 %
Hipoglucemiante oral	19,2 %
Inmunosupresor	2,4 %
Corticoesteroides	23,9 %

Tabla 4: Porcentaje de personas con tratamiento

4.1.1. Análisis bivariante

En este subapartado, se han realizado estudios sobre las variables viendo cuales de éstas son significativas realizando un contraste de independencia son el estadístico χ^2. Se aceptará como significativo aquellos P-valores que sean menores a 0,1, aún siendo 0,05 el criterio más usado comunmente, ya que no se quiere ser tan restrictivo a la hora de seleccionar las variables para posteriores análisis

Población adulta

En primer lugar, se mostrará la tabla con los porcentajes del grupo de interés (no vacunados en ninguna campaña teniendo enfermedades respiratorias) para las enfermedades y otra con los tratamientos en personas adulta. Después, se verá otras tablas con variables de interés con varias categrías. Finalmente, una última tabla con varias variables de interés diferentes como la alergia al huevo o los ingresos en el último año.

En estas tablas con las variables de enfermedad y tratamientos, se aprecia que en general el porcentaje de vacunados es mayor en gente sin enfermedades y gente sin tratamientos, excepto en los corticoesteroides. Esto se debe a que es un tratamiento estrechamente ligado al asma. Con lo cual, la gente sin enfermedades y tratamientos tiende a vacunarse menos.

Como se puede ver, hay diferencias entre grupos y la gente con obesidad, sin llegar a tener una obesidad morbida, tiende a vacunarse menos en comparación con los otros dos grupos.

En el caso de fumar, se ve como las personas que no han fumado nunca en general se vacunan menos.

También, se concluye que las personas dedicadas a la gerencia profesional u ocupaciones gerenciales y técnicas son estadísticamente menos reacios a vacunarse teniendo enfermedades respiratorias.

Además, se aprecia como el número de consultas en los últimos tres meses, tener contacto cercano con niños, el sexo o estar embarazada, no influye en esta decisión. A diferencia de tener alergia al huevo, ya que algunas vacunas tienen proteínas de huevo, o haber estado ingresado en los últimos 12 meses que hace que sea significativamente estadístico las diferencias en estos grupos a favor de la no vacunación.

Enfermedad	% No Enfermos Retic.	% Enfermos Retic.	P-valor
Enfermedad Cardíovascular	$18,4\,\%$	$15,1\,\%$	-
Enfermedad Cerebrovascular	$18,0\,\%$	$9,5\,\%$	$0,085$
Arteriopatía periférica	$17,9\,\%$	$8,3\,\%$	$0,086$
Diabetes	$19,6\,\%$	$9,7\,\%$	$1,78x10^{-5}$
Otras Enfermedades Endocrinas	$18,1\,\%$	$14,0\,\%$	-
Anemia	$17,8\,\%$	$15,3\,\%$	-
Enfermedad Hepática Crónica	$17,9\,\%$	$15,2\,\%$	-
Enfermedad Renal Crónica	$18,5\,\%$	$9,2\,\%$	0,003
Autoinmune	$17,7\,\%$	$16,7\,\%$	-
Desorden Neuromuscular	$18,3\,\%$	$4,6\,\%$	0,001
Neoplasia	$18,2\,\%$	$12,5\,\%$	0,074
Enfermedad Autoinmune	$17,8$	$15,1\,\%$	-
Demencia	$17,8\,\%$	$0,0\,\%$	0.055

Tabla 5: Personas adultas con enfermedad, porcentaje de reticentes y no reticentes

Tratamiento	% No Tratados Retic.	% Tratados Retic.	P-valor
Antihipertensivos	$19,2\,\%$	$14,7\,\%$	0,018
Anticoagulantes	$18,1\,\%$	$13,0\,\%$	-
Agregación antiplaquetaria	$18,4\,\%$	$11,6\,\%$	0,02
Hipolipemiantes	$19,8\,\%$	$12,3\,\%$	$1,539x10^{-4}$
Insulina	$19,1\,\%$	$4,0\,\%$	$6,31x10^{-7}$
Hipoglucemiante oral	$18,6\,\%$	$12,1\,\%$	0,011
Inmunosupresor	$18,2\,\%$	$9,2\,\%$	0,024
Corticoesteroides	$11,2\,\%$	$34,8\,\%$	≈ 0
Vacuna Antineumocócica	$20,5\,\%$	$8,5\,\%$	$1,439x10^{-8}$

Tabla 6: Personas adultas con tratamientos, porcentaje de reticentes y no reticentes

Variable	% Reticentes	P-valor
Obesidad: No o medio	$15,9\,\%$	
Obesidad: Sí	$23,5\,\%$	$5,734x10^{-4}$
Obesidad: Morbida	$12,2\,\%$	

Tabla 7: Obesidad en población adulta y porcentaje reticentes

Variable	% Reticentes	P-valor
Fumar: Actualmente	$19,8\,\%$	
Fumar: Nunca	$12,3\,\%$	
Fumar: menos de 5 cigarros al día	$22,2\,\%$	$3,862x10^{-3}$
Fumar: Lo dejó hace menos de un año	$21,8\,\%$	
Fumar: Lo dejó hace más de un año	$19,0\,\%$	

Tabla 8: Hábito Tabáquico en población adulta y porcentaje reticentes

Variable	% Reticentes	P-valor
Clase Social: Cualificado (M)	$20,1\,\%$	
Clase Social: Cualificado (NM)	$17,1\,\%$	
Clase Social: Cualificado Parcialmente	$17,5\,\%$	
Clase Social: Gerencia Profesional	$9,1\,\%$	$0,057$
Clase Social: Inclasificable	$100\,\%$	
Clase Social: Ocupaciones Gerenciales y técnicas	$12,6\,\%$	
Clase Social: Sin cualificacion	$18,6\,\%$	

Tabla 9: Porcentaje reticentes en las clases sociales de la población adulta

Variable	% Reticentes	P-valor
Consultas: Ninguna	$16,0\,\%$	
Consultas: Una	$17,6\,\%$	
Consultas: Dos	$17,5\,\%$	-
Consultas: Tres	$20,3\,\%$	
Consultas: Más de Tres	$18,5\,\%$	
Consultas: No sabe, no se acuerda	$11,1\,\%$	

Tabla 10: Consultas en los últimos 3 meses en población adulta y porcentaje reticentes

Variable	% Reticentes clase 1	% Reticentes clase 2	P-valor
Niños	No= $17,4\,\%$	Sí= $18,3$	-
Sexo	$Hombre = 17,4\,\%$	$Mujer = 17,9\,\%$	-
Embarazo	No= $17,7\,\%$	Sí= $7,7\,\%$	-
Ingresos en los últimos 12 meses	No= $15,4\,\%$	Sí= $22,2\,\%$	$4,142x10^{-4}$
Alergia al huevo	No= $17,6\,\%$	Sí= $42,9\,\%$	0,08

Tabla 11: Porcentaje reticentes y no reticentes en otras variables de la población adulta

Población mayor de 65 años

Al igual que con la poblacion anterior, en primer lugar, se mostrará la tabla del análisis bivariante para las enfermedades y otra con los tratamientos en personas mayores de 65 años. Después, se muestran las variables binarias con el porcentaje de reticentes y no reticentes para la población de mayores de 65 años y variables con más de dos clases.

En cuanto a los tratamientos y a las enfermedades, se puede ver que en general la gente con ciertas enfermedades o con ciertos tratamientos se vacuna menos, pero sin ser algo general Esto se puede ver con la variable de otras enfermedades endocrinas, ya que el hecho de tener alguna enfermedad endocrina distinta a la diabetes marca diferencias significativas con el otro grupo vacunándose menos. En cambio, con los tratamientos sí que es similar al grupo anterior, ya que quitando el tratamiento de los corticoesteroides la gente sin tratamientos se vacuna menos, llegando a ser significativo.

Adicionalmente, se ve como las personas con obesidad morbida son más reticentes a la vacuna, al igual que la gente que fuma actualmente más o menos de 5 cigarros al día o que lo ha dejado hace menos de un año.

Así mismo, se ve como no hay diferencias entre las distintas clases sociales, el número de consultas realizadas en los últimos 3 meses, en tener contacto con niños o el sexo de los pacientes. A diferencia de los ingresos en los últimos 12 meses que se aprecian diferencias estadísticas entre los dos grupos vacunándose menos, aún teniendo enfermedades respiratorias, la gente que ha sido ingresada.

Finalmente a diferencia con el grupo de los adultos, no se aprecia diferencia estadística entre los pacientes con alergia al huevo o sin ella a la hora de vacunarse o no, aún siendo más alto el porcentaje de personas reticentes en el grupo de alérgicos.

Enfermedad	% No Enfermos Retic.	% Enfermos Retic.	P-valor
Enfermedad Cardíovascular	$7,6\,\%$	$6,1\,\%$	0.018
Enfermedad Cerebrovascular	$6,8\,\%$	$6,3\,\%$	-
Arteriopatía periférica	$6,8\,\%$	$6,6\,\%$	-
Diabetes	$7,1\,\%$	$6,2\,\%$	-
Otras Enfermedades Endocrinas	$6,6\,\%$	$8,2\,\%$	0.043
Anemia	$7,0\,\%$	$5,1\,\%$	0.04
Enfermedad Hepática Crónica	$6,7\,\%$	$7,2\,\%$	-
Enfermedad Renal Crónica	$7,1\,\%$	$5,4\,\%$	0.018
Autoinmune	$6,7\,\%$	$12,2\,\%$	-
Desorden Neuromuscular	$6,7\,\%$	$7,5\,\%$	-
Neoplasia	$6,8\,\%$	$6,2\,\%$	-
Enfermedad Autoinmune	$6,7\,\%$	$8,75\,\%$	-
Demencia	$7,1\,\%$	$4,0\,\%$	0.006

Tabla 12: Personas mayores de 65 con enfermedad, porcentaje reticentes y no reticentes

Tratamiento	% No Tratados Retic.	% Tratados Retic.	P-valor
Antihipertensivos	$8,3\,\%$	$6,25\,\%$	0.003
Anticoagulantes	$7,5$	$5,0\,\%$	$2,49x10^{-4}$
Agregación antiplaquetaria	$6,8\,\%$	$6,5\,\%$	-
Hipolipemiantes	$6,9\,\%$	$6,5\,\%$	-
Insulina	$7,0\,\%$	$5,5\,\%$	0.08
Hipoglucemiante oral	$7,0\,\%$	$6,2\,\%$	-
Inmunosupresor	$6,8\,\%$	$4,6\,\%$	-
Corticoesteroides	$4,5\,\%$	$12,1\,\%$	≈ 0
Vacuna antineumocócica	$7,6\,\%$	$5,1\,\%$	$8,271x10^{-5}$

Tabla 13: Personas mayores de 65 con tratamientos, porcentaje reticentes y no reticentes

Variable	% Reticentes	P-valor
Obesidad: No o medio	$6,1\,\%$	
Obesidad: Sí	$7,8\,\%$	$8,061x10^{-6}$
Obesidad: Morbida	$15,4\,\%$	

Tabla 14: Obesidad en población mayor y porcentaje reticentes

Variable	% Reticentes	P-valor
Fumar: Actualmente	$14,9\,\%$	
Fumar: Nunca	$5,3\,\%$	
Fumar: menos de 5 cigarros al día	$12,4\,\%$	$6,992x10^{-12}$
Fumar: Lo dejó hace menos de un año	$13,1\,\%$	
Fumar: Lo dejó hace más de un año	$7,1\,\%$	

Tabla 15: Hábito Tabáquico en población mayor de 65 y porcentaje reticentes

Variable	% Reticentes	P-valor
Clase Social: Cualificado (M)	$6,7\,\%$	
Clase Social: Cualificado (NM)	$8,8\,\%$	
Clase Social: Cualificado Parcialmente	$8,0\,\%$	
Clase Social: Gerencia Profesional	$5,7\,\%$	-
Clase Social: Inclasificable	$0\,\%$	
Clase Social: Ocupaciones Gerenciales y técnicas	$7,3\,\%$	
Clase Social: Sin cualificacion	$6,1\,\%$	

Tabla 16: Porcentaje reticentes en las clases sociales de la población mayor de 65 años

Variable	% Reticentes	P-valor
Consultas: Ninguna	$7,6\,\%$	
Consultas: Una	$8,3\,\%$	
Consultas: Dos	$6,1\,\%$	-
Consultas: Tres	$6,2\,\%$	
Consultas: Más de Tres	$6,4\,\%$	
Consultas: No sabe, no se acuerda	$4,3\,\%$	

Tabla 17: Consultas en los últimos 3 meses en población mayor de 65 años y porcentaje reticentes

Variable	% Reticentes clase 1	% Reticentes clase 2	P-valor
Niños	No= $6,7\,\%$	Sí= $7,2\,\%$	-
Sexo	$Hombre = 7,1\,\%$	$Mujer = 6,3\,\%$	-
Ingresos en los últimos 12 meses	No= $6,3\,\%$	Sí= $7,5\,\%$	$3,544x10^{-3}$
Alergia al huevo	No= $6,7\,\%$	Sí= $15,4\,\%$	-

Tabla 18: Porcentaje reticentes y no reticentes en otras variables de la población adulta

4.2. Modelización

Para caracterizar al grupo de interés (como se ha marcado en el segundo objetivo), se han realizado los siguientes modelos para ver qué factores son más importantes. En todos los modelos, se han introducido las variables que se han considerado más importantes según el análisis bivariante en base a la literatura.

4.2.1. Regresión logística

Para estos modelos, se han creado varibles `dummies` eliminando una, dejando la variable eliminada como la de referencia. Se han obtenido los logaritmos neperianos de las razones de ventajas, su significación y el pseudo-R^2. En primer lugar, se ha realizado una regresión logística y luego se ha realizado una regresión stepwise para realizar un modelo más parsimonioso e interpretable. Ésto se ha realizado para el conjunto de datos de la población adulta y para la población mayor de 65 años.

Población Adulta

Se ha realizado un primer modelo de regresión logística con los factores que se han considerado más importantes en el análisis de las variables en base a la revisión de la literatura, viendo que habían diferencias significativas entre grupos.

Regresión Logstica

Se ha realizado la regresión logística y se ha obtenido lo siguiente:

Variable	P-valor	$coef. = \ln OR$	$e^{coef.} = OR$
β_0	-	-4,41	$6,9x10^{-3}$
Enfermedad cerebrovascular	0,416	-0,40	0,6722
Arteriopatía periférica	0,502	-0,41	0,6668
Diabetes	0,241	-0,49	0,6144
Enfermedad Renal	0,361	-0,29	0,7476
Desorden Neuromuscular	0,01	-1,38	0,2519
Neoplasia	0,0378	-0,58	0,5627
Demencia	0,9646	-13,89	$9,2402x10^{-7}$
Antihypertensivos	0,153	-0,26	0.7735
Agregación antiplaquetaria	0,664	0,12	1,1328
Hipolipemiantes	0,02	-0,46	0,6341
Insulina	0,009	-1,23	0,2919
Hipoglucemiante oral	0,476	0,30	1,3451
Inmunosupresores	0,749	-0,13	0,8806
Cortiesteroides	$< 2x10^{-16}$	1,57	4,7825
Obesidad: No or mild (bmi$<$ 30)	0,314	0,55	1,7402
Obesidad: Sí ($>= 30 < 40$)	0,032	0,95	2,5848
Fumar: menos de 5 cigarros por día	0,767	0,08	1,0885
Fumar: nunca	0,035	-0,39	0,677
Fumar: Dejó hace menos de 1 year	0,695	-0,12	0,8904
Fumar: Dejó hace más de 1 year $>$ 1 year	0,588	-0,10	0,9070
Clase social: Cualificado parcialmente	0,837	-0,05	0,9481
Clase social: Gerencia profesional	0,11	-0,75	0,4734
Clase social: Ocupaciones gerenciales y técnicas	0,103	-0,57	0,5638
Clase social: Cualificado (NM)	0,772	-0,09	0,9164
Clase social: Inclasificable	0,991	16,86	$2,09x10^{7}$
Clase social: Sin cualificación	0,759	-0,07	0,9337
Hospitalizaciones año anterior	0,038	0,15	1,1661
Ingresos en los últimos 12 meses	0,322	0,2	1,2249
Alérgia al huevo	0,135	1,41	4,0983
Vacuna antineumocócica	$9,59x10^{-12}$	-1,42	0,2408
Peso	0,383	0,007	1,0066
Altura	0,286	0,011	1,0113

Tabla 19: Coeficientes de la Regresión Logística en adultos

Este modelo de regresión logística, tiene un $pseudo - R^2$ de un $17,86\,\%$, con una tasa de acierto de $83,54\,\%$, una tasa de falso negativo de $16,07\,\%$ y un porcentaje de verdadero positivo de $1,6\,\%$.

Como se puede apreciar, hay ciertos factores introducidos en el modelo que son significativos para un P-valor muy reducido que son los corticoides, tratamiento utilizado para el asma, y la vacuna antineumocócica, utilizada para prevenir la bronquitis.

Aparte, también hay otras, que en menor medida, son significativas como por ejemplo: tener las enfermedades de desorden neuromuscular y neoplasia, tomar hipolipemiantes o insulina, ser obeso sin tener obesidad morbida, no haber fumado nunca a diferencia de estar fumando actualmente y el número de hospitalizaciones el año anterior, aumentando la probabilidad de ser reticente éste último.

Regresión Logstica Stepwise

Continuando con el modelo de regresión logística, se ha realizado una regresión logística Stepwise, obteniéndose la siguiente tabla:

Variable	P-valor	$coef. = \ln OR$	$e^{coef.} = OR$
β_0	-	-5.04	0,0064
Desorden Neuromuscular	0,011	-1,36	0,2572
Neoplasia	0,033	-0,58	0,5584
Demencia	0,965	-13,93	≈ 0
Antihypertensivos	0,069	-0,31	0,7343
Hipolipemiantes	0,011	-0,47	0,6258
Insulina	$2x10^{-4}$	-1,52	0,2183
Cortiesteroides	$< 2x10-16$	1,58	4,8678
Obesidad: Sí ($>= 30 < 40$)	$1x10^{-4}$	0,59	1,8091
Fumar: Nunca	0,022	-0,38	0,6852
Clase social: Gerencia profesional	0,096	-0,71	2,0389
Clase social: Inclasificable	0,991	17,05	$2,53x10^{7}$
Clase social: Ocupaciones gerenciales y técnicas	0,098	-0,49	0,6121
Hospitalizaciones año anterior	$4x10^{-4}$	0,19	1,214
Vacuna antineumocócica	$5,78x10^{-12}$	-1,41	0,2431
Altura	0,021	0,02	1,0173

Tabla 20: Coeficientes de la Regresión Logística Stepwise para adultos

Este modelo tiene un $pseudo - R^2$ de $17,35\,\%$, con una tasa de acierto de $83,38\,\%$, una tasa de falso negativo de $16,13\,\%$ y una tasa de verdadero positivo de $1,54\,\%$.

Al igual que en el modelo previo, los corticoesteroides y la vacuna antineumocócica se

aprecian significativos con un P-valor muy pequeño, el primero aumentando la probabilidad de ser reticente y el segundo disminuyéndola.

Aparte, hay otros factores, al igual que con el modelo anterior, como tener enfermedades tales como desorden neuromuscular y neoplasia, descendiendo la probabilidad de ser reticente; tomar hipolipemiantes e insulina, decreciendo la probabilidad de reticencia a la vacuna; ser obeso, aumentando esta probabilidad; no haber fumado nunca, disminuyéndola respecto de estar fumando o haber fumado en alguna ocasión, y el número de hospitalizaciones el año anterior aumentando esta probabilidad por cada hospitalización.

Tanto en este modelo como el anterior, tienen unas categorías como la de tener demencia o estar en una clase social que no pertenece a ninguna de las otras, llamándo a ésta inclasificable, que en éste grupo tiene un número de casos muy reducidos y hace que tenga valores muy altos y muy bajos en estos modelos.

Población Mayor de 65 años

De manera análoga, se van a utilizar las variables que se han considerado significativas en el análisis bivariante para construir el modelo de regresión, en base a la revisión de la literatura, para caracterizar este grupo.

Regresión Logística

En primer lugar, se ha calculado el modelo de regresión para obtener la razón de ventajas, sus logaritmos neperianos, los niveles de significación de los coeficientes y posteriormente su pseudo-R^2. A continuación se muestra la tabla:

Variable	P-valor	$coef. = \ln OR$	$e^{coef.} = OR$
β_0	-	0,42	1,5244
Enfermedad Cardiovascular	0,971	$4x10^{-3}$	1,0044
Otras Enfermedades Endocrinas	0,016	0,34	1,3985
Anemia	0,099	-0,29	0,7515
Enfermedad Renal	0,256	-0,16	0,8510
Demencia	0,071	-0,38	0,6870
Antihypertensivos	0,038	-0,25	0,7793
Anticoagulantes	0,009	-0,35	0,7049
Insulina	0,033	-0,35	0,7018
Corticoesteroides	$< 2x10^{-16}$	1,01	2,7486
Obesidad: No o medio (bmi< 30)	0,016	-0,94	0,3889
Obesidad: Sí (>= 30 < 40)	0,033	-0,65	0,5242
Fumar: menos de 5 cigarros al día	0,926	-0,03	0,9695
Fumar: Nunca	$2,70x10^{-7}$	-0,96	0,384
Fumar:Dejó hace menos de 1 año	0,576	-0,17	0,8399
Fumar:Dejó hace más de 1 año	$8,55x10^{-5}$	-0,71	0,4899
Hospitalizaciones el año anterior	0,002	0,16	1,1751
Ingresos en los últimos 12 meses	0,42	-0,12	0,8866
Vacuna antineumocócica	$6x10^{-6}$	-0,53	0,5905
Peso	0,748	$2,1x10^{-3}$	1,0021
Altura	0,279	$-9,8x10^{-3}$	0,9902

Tabla 21: Coeficientes de la Regresión Logística para mayores de 65

Tiene un pseudo-R^2 de $7,38\,\%$, con una tasa de aciertos de un $93,23\,\%$, con una tasa de falso negativo de $6,75\,\%$ y una tasa de verdadero positivo del $0\,\%$.

Las variables que se han visto que son más discriminantes son tener algunas enfermedades endocrinas lo cual aumenta la probabilidad a ser reticente, tener los tratamientos de antihypertensivos, anticoagulantes, insulina y corticoesteroides teniendo todos estos,

excepto los corticoesteroides, una disminución de la razón de ventajas.

El hecho de no ser obeso o sí serlo en comparación a ser obeso morbido disminuye la probabilidad de ser reticente ante la vacuna teniendo una enfermedad respiratoria. No haber fumado nunca o habérselo dejado hace más de un año también disminuye esta probabilidad. Al igual que haberse puesto la vacuna antineumocócica. Al contrario que el número de hospitalizaciones el año anterior que hace aumentar esta probabilidad a mayor número de hospitalizaciones.

Regresión Logística Stepwise

Al igual que en el grupo de edad anterior, se ha realizado un proceso stepwise para la regresión logística para hacer el modelo más parsimonioso y más fácilmente interpretable.

Variable	P-valor	$coef. = \ln OR$	$e^{coef.} = OR$
β_0	-	-1,01	0,3644
Otras Enfermedades Endocrinas	0,014	0,34	1,4039
Anemia	0,063	-0,32	0,7274
Demencia	0,066	-0,38	0,6828
Antihypertensivos	0,026	-0,26	0,7748
Anticoagulantes	$3,7x10^{-3}$	-0,36	0,6954
Insulina	0,021	-0,38	0,6842
Corticoesteroides	$< 2x10^{-16}$	1,01	2,7519
Obesidad: No o medio (bmi$<$ 30)	$1,92x10^{-5}$	-1,07	0,3437
Obesidad: Sí ($>= 30 < 40$)	0,006	-0,7	0,4950
Fumar: Nunca	$6,86x10^{-9}$	-0,87	0,4202
Fumar: Dejó hace más de 1 año	$4,98x10^{-6}$	-0,68	0,5061
Hospitalizaciones del año anterior	0,001	0,13	1,1345
Vacuna antineumocócica	$3,29x10^{-6}$	-0,54	0,5836

Tabla 22: Coeficientes de la Regresión Logística Stepwise para mayores de 65

Este modelo tiene un pseudo-R^2 de $7,26\,\%$ y una tasa de acierto total de $93,25\,\%$, una tasa de falso negativo de $6,75\,\%$ y una tasa de positivo de $0\,\%$.

En este modelo, las variables más importantes son los corticoesteroides y no haber fumado nunca en comparación a haber dejado de fumar hace menos de un año o estar fumando en mayor o menor medida a día de hoy. La primera variable hace que aumente la probabilidad de ser reticente y la segunda hace que disminuya esta probabilidad.

Así mismo, otra variable significativa que hace que aumente esta probabilidad de reticencia son las hospitalizaciones el año anterior. A mayor número de hospitalizaciones el

año anterior, mayor probabilidad de no vacunarse aún teniendo enfermedades respiratorias.

Sin embargo, tener otras enfermedades endocrinas que no sean la diabetes, tomar antihypertensivos, anticoagulantes e insulina, no ser obeso o serlo sin ser obeso morbido, dejar de fumar hace más de un año y haber tomado la vacuna antineumocócica son variables significativas que hacen que descienda la probabilidad de ser reticente.

4.2.2. Árboles de Clasificación

Para los árboles de clasificación, se ha seguido una estrategia similar. Viendo las variables más importantes según el análisis bivariante en base a la literatura, se contruirán los árboles de clasificación para ambas poblaciones y se calculará la tasa de acierto que han tenido.

Se calculará un árbol por defecto que nos ofrece la función `rpart`, otro completo con esta misma función y el parámetro de complejidad igual a $0,001$ y el último será un árbol podado con la función `rpartXse`.

En el caso de los árboles de clasificación, las variables más importantes son las más cercanas al primer nodo ya que discriminan más en los grupos más grandes.

Población adulta

Los árboles de clasificación son los que se han citado anteriormente y tienen una tasa de clasificación de $84,98\,\%$, $86,9\,\%$ y $85,36\,\%$ respectivamente. Los árboles tienen una tasa de falsos negativos del $11,94\,\%$, $9,52\,\%$ y $11,67\,\%$. También, tienen unas tasas de positivos de $5,72\,\%$, $8,15\,\%$ y $6\,\%$.

Aunque el árbol completo tenga la mayor tasa de aciertos y la menor de falsos positivos, como se puede apreciar, éste tiene muchos nodos siendo muy extenso y poco parsimonioso, es decir, carece de interpretabilidad.

Como se puede ver en los otros dos árboles, las variables más importantes son tomar los corticoides y haber tomado la vacuna antineumocócica ya que son las variables que aparecen más arriba. Con estas variables, se aprecia que si no se toman corticoesteroides se le clasifica como no reticente a la vacuna o persona sin enfermedades respiratorias. En el caso de la vacuna neumocócica, es al contrario.

Así mismo, otras variables de importancia son tomar hipolipemiantes o el número de hospitalizaciones anuales. A continuación, se muestran en el orden anteriormente citado:

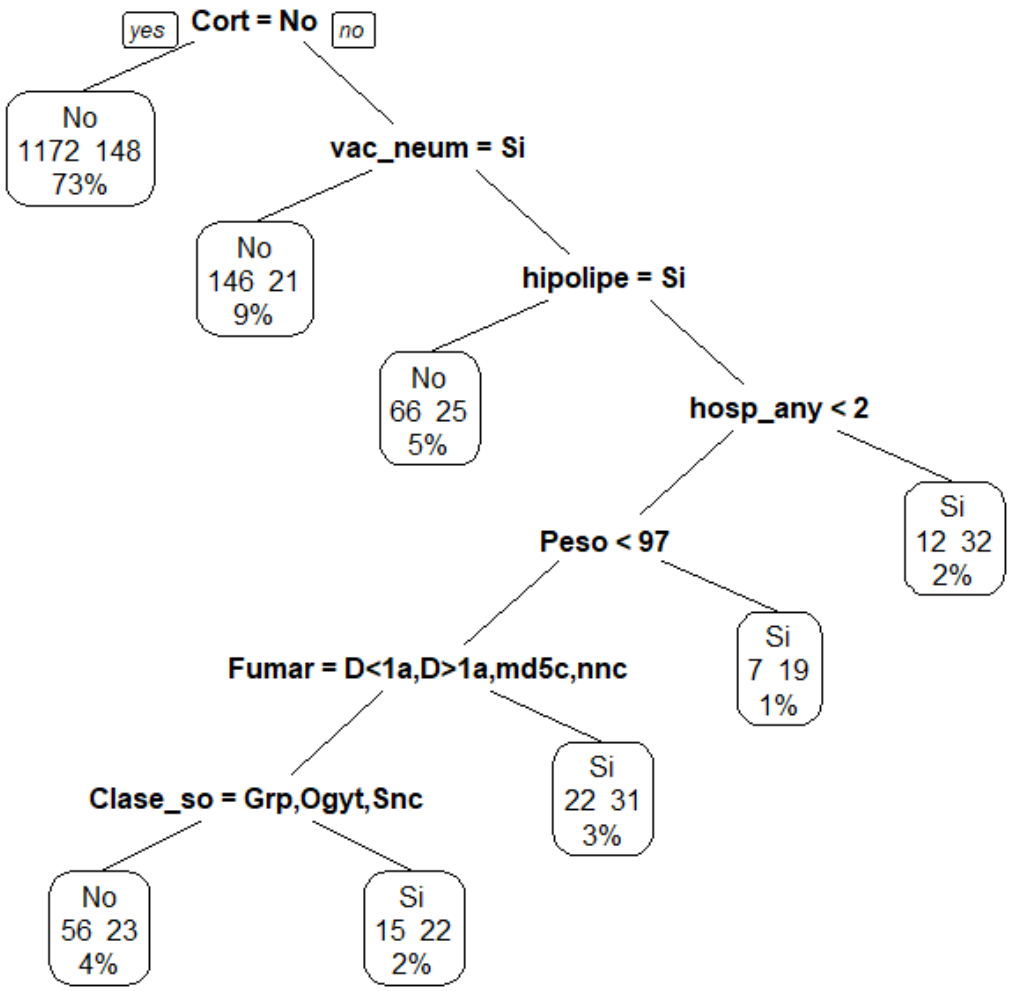

Figura 2: Arbol de clasificación por defecto para la población de adultos

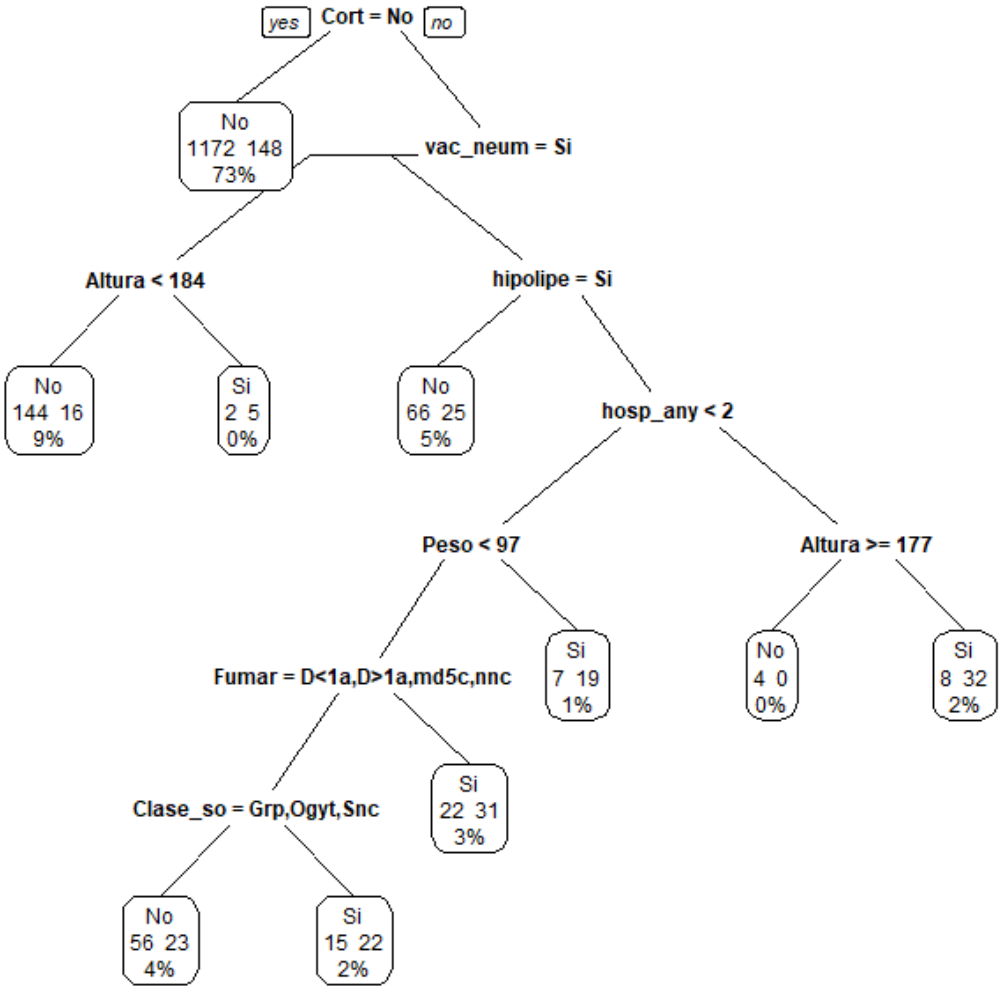

Figura 3: Arbol de clasificación podado para la población de adultos

Población mayor de 65 años

Para el modelo en este conjunto de datos, se ha utilizado únicamente el árbol de clasificación completo, ya que los otros dos predecían solo la clase mayoritaria siendo estos modelos de clasificación inservibles.

Este árbol de clasificación tiene una tasa de acierto de $93,31\,\%$, teniendo únicamente una tasa de verdaderos positivos del $0,09\,\%$ y de falsos negativos del $6,66\,\%$.

Para este árbol de clasificación al igual que con los anteriores, las variables más importantes son las que más arriba se encuentran ya que son las más discriminantes y esta son los corticoesteroides, haberse puesto o no la vacuna antineumocócica y el peso. Si la persona no toma corticoides, si se ha puesto la vacuna neumocócica tomando corticoides, no se ha puesto la vacuna neumocócica y pesa menos de 90 kilogramos, se le clasifica como persona que se vacuna o que no padece una enfermedad respiratoria.

El árbol de clasificación es el siguiente:

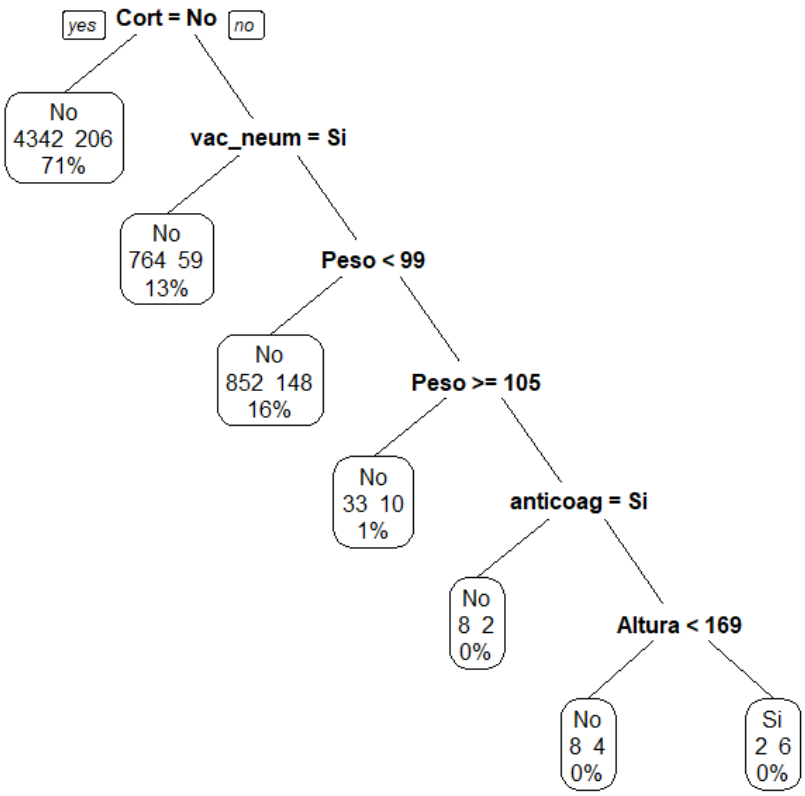

Figura 4: Arbol de clasificación completo para la población de mayores de 65 años

4.2.3. Random forest

Para caracterizar al grupo de interés con el modelo del random forest en ambas poblaciones, se ha realizado de manera similar a los otros modelos. En primer lugar, se han seleccionado las variables que se van a utilizar y en segundo lugar, se ha creado el modelo teniendo en cuenta que el número de variables a elegir en cada ramificación de manera aleatoria es el indicado en la metodología. El resto de parámetros, se han elegido por defecto. En ambos modelos al estar el número de variables comprendidas entre 16 y 25, se ha cogido el parámetro del número de variables por ramificación como 4. A continuación, se mostrará el gráfico de importancias para cada grupo, su tasa de acierto y su tasa de falsos postivos.

Población adulta

Para la población adulta, se ha obtenido una tasa de acierto del $93,23\,\%$, una tasa de falsos negativos del $6,71\,\%$ y una tasa de verdaderos positivos de $10,95\,\%$. A continuación se muestra el gráfico de importancias:

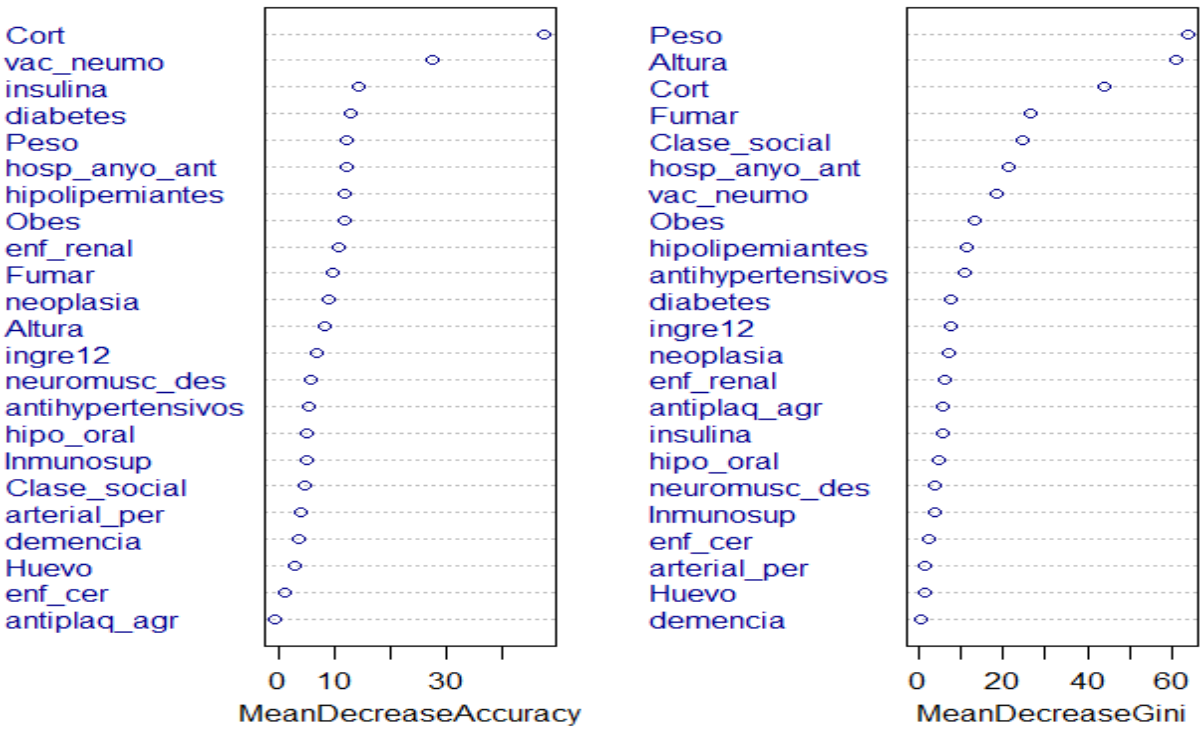

Figura 5: Gráfico de importancias del Random Forest para la población de adultos

En este modelo Random Forest, se ve como las variables más importantes en general para conseguir una mayor tasa de acierto son los corticoides, haberse vacunado del neumococo, tomar insulina, tener diabetes y el peso de la persona.

Población mayor de 65 años

Para los datos provenientes de esta población, se ha obtenido una tasa de acierto del $96,23\,\%$, una tasa de falsos negativos del $3,77\,\%$ y una tasa de positivos de $2,98\,\%$. A continuación, se muestra el gráfico de importancias:

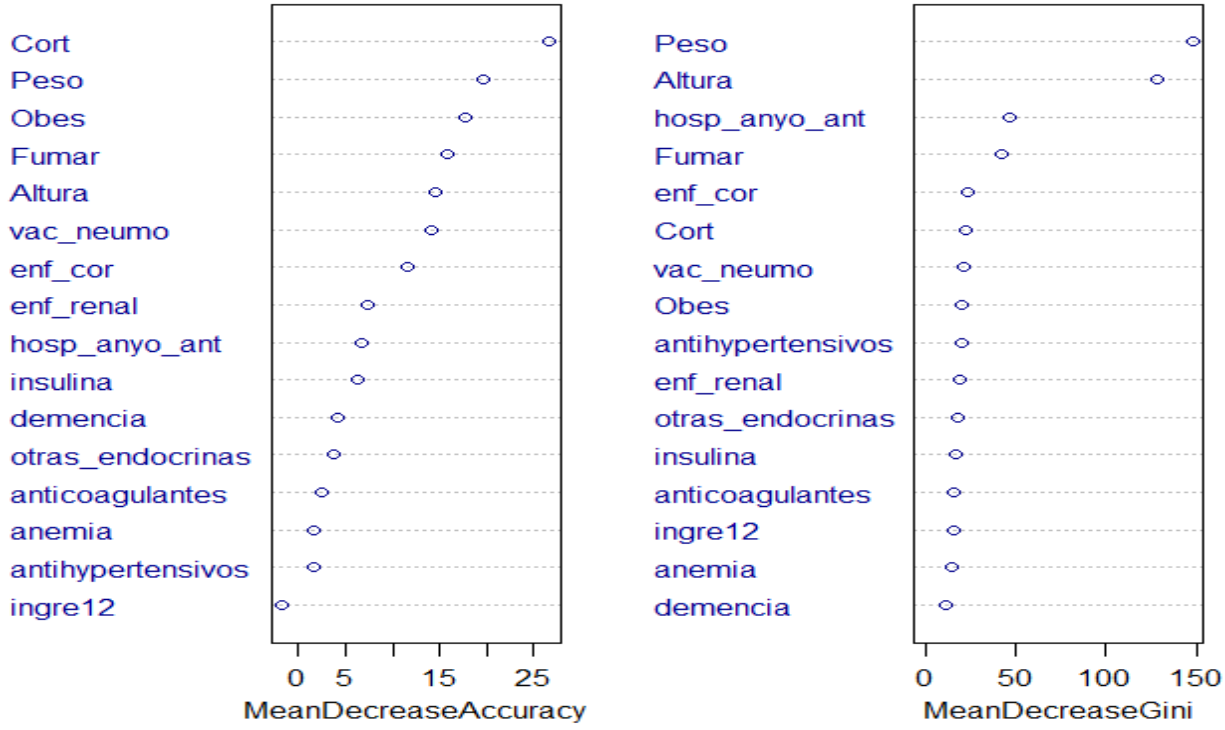

Figura 6: Gráfico de importancias del Random Forest para la población mayores de 65 años

Como se puede ver, las variables más importantes en este modelo son los corticoesteroides, el peso de la persona, el nivel de obesidad, si fuma (y en qué nivel) y la altura de la persona.

4.2.4. Análisis de Correspondencias Múltiples

En este apartado, se va a realizar un análisis de correspondencias múltiples, para observar qué variables explicativas son las más relacionadas con algunas variables de interés.

Las variables que se introducirán serán las mismas que en los modelos anteriores en ambas poblaciones. Además, la varible de hospitalizaciones el año anterior se ha tranformado en categórica con las siguientes clases: Ninguna hospitalización, entre 1 y 3 hospitalizaciones y más de tres hospitalizaciones.

Población adulta

En el siguiente análisis de correspondencias múltiples para la población adulta, se han seleccionado dos componentes observando el gráfico de varianzas explicadas.

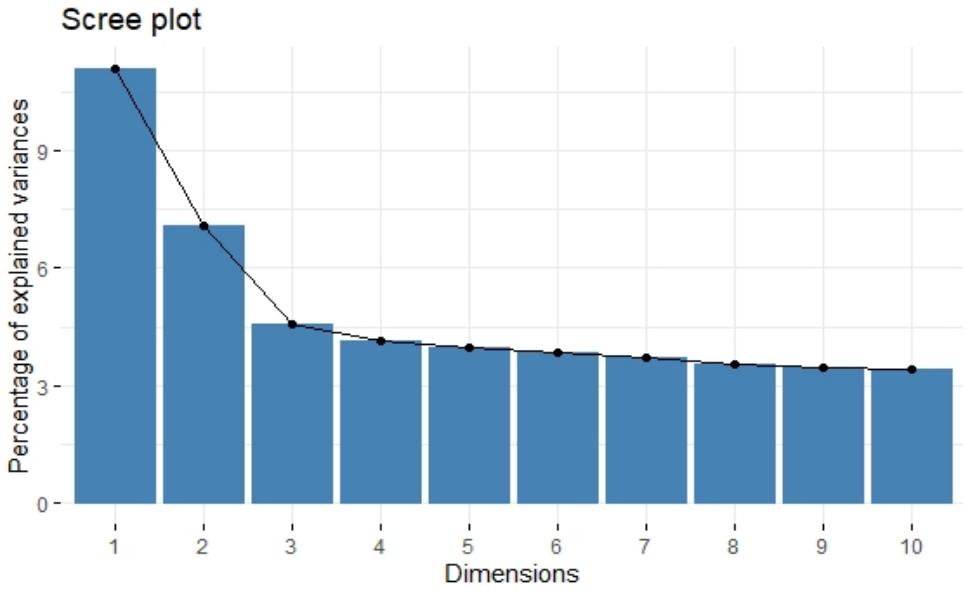

Figura 7: Varianza explicada por cada componente del MCA en adultos

Para finalizar, se han graficado las variables en las dos dimensiones que explican un 18,2 % de la variabilidad. En el siguiente gráfico, se puede observar como en el eje de horizontal, el de la primera componente, en la parte derecha se encuentran las categorías de las variables que representan tener una enfermedad o algún tratamiento, en cambio, en el mismo eje en la parte contraria, aunque muy centradas, se encuentran las categorías de las variables que representan no tener ninguna enfermedad, no tener ningún tratamiento y no fumar nunca o fumar actualmente. En el eje vertical, las variables más importantes se ve como son las hospitalizaciones y los ingresos, estando en la parte superior del mapa los que sí han sido ingresados en los últimos 12 meses y la gente que ha sufrido hospitlizaciones entre una y tres y más de tres, al contrario de la parte inferior en la cual se encuentran las categorías de no haber tenido ingresos ni hospitalizaciones. Para finalizar, se puede

ver como las variables de reticentes, a pesar de estar relativamente centrada, tiene cerca las clases de las variables que representan no padecer una enfermedad o no tener un tratamiento, excepto los corticoesteroides y el desorden neuromuscular, además de las mujeres, en caso contrario, la categoría de no reticentes está situada alrededor de la categoría de hombres y algunas categorías de clases sociales como pueden ser Ocupaciones Gerenciales y Técnicas o gente cualificada parcialmente, además de estar cerca también de algunas enfermedades o ser obeso o obeso morbido.

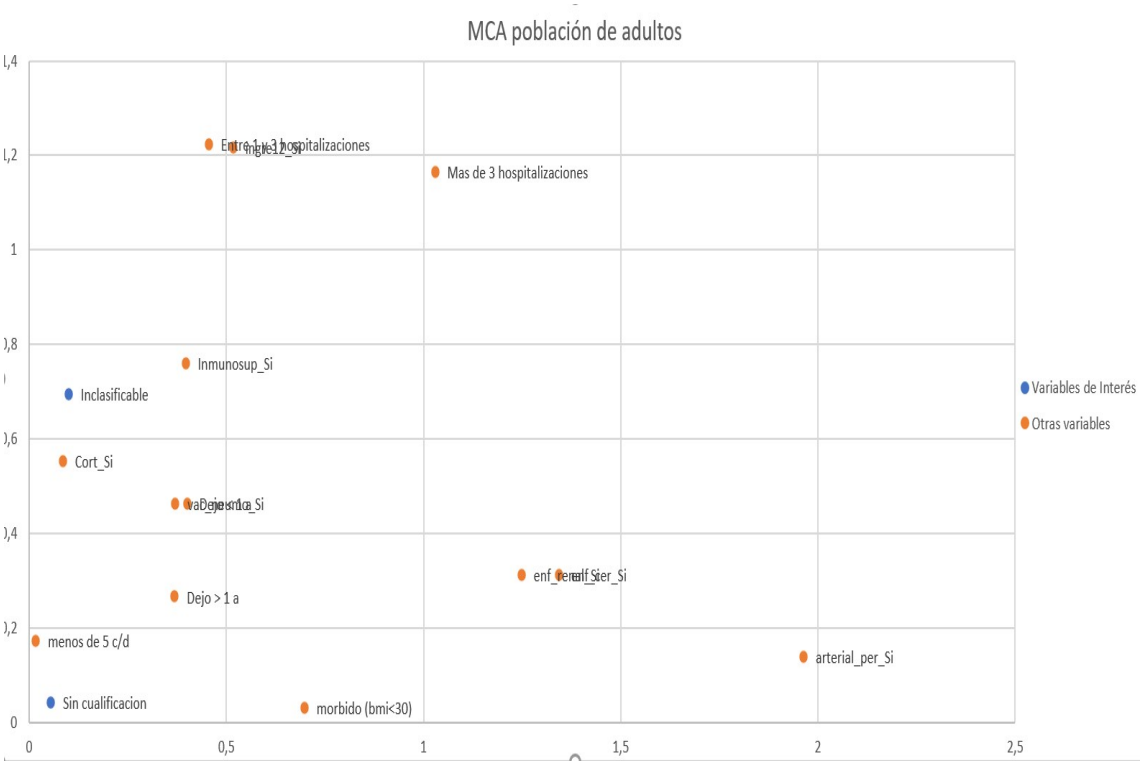

Figura 8: Mapa con las variables del MCA en adultos en el primer cuadrante

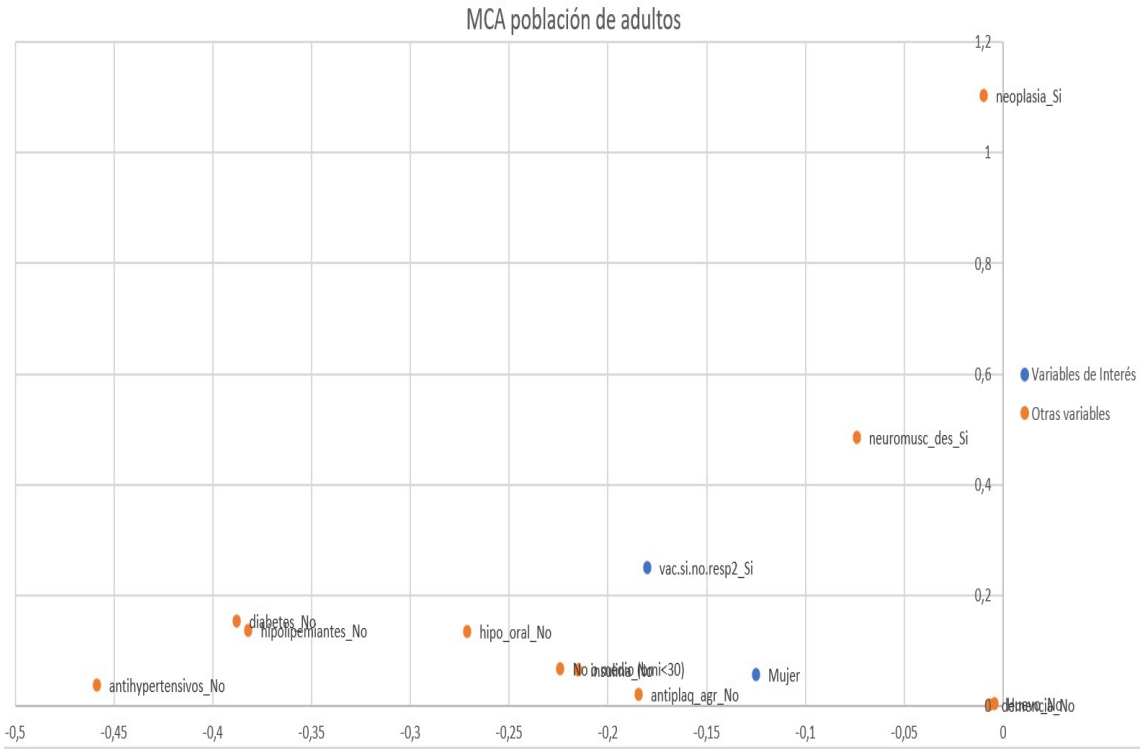

Figura 9: Mapa con las variables del MCA en adultos en el segundo cuadrante

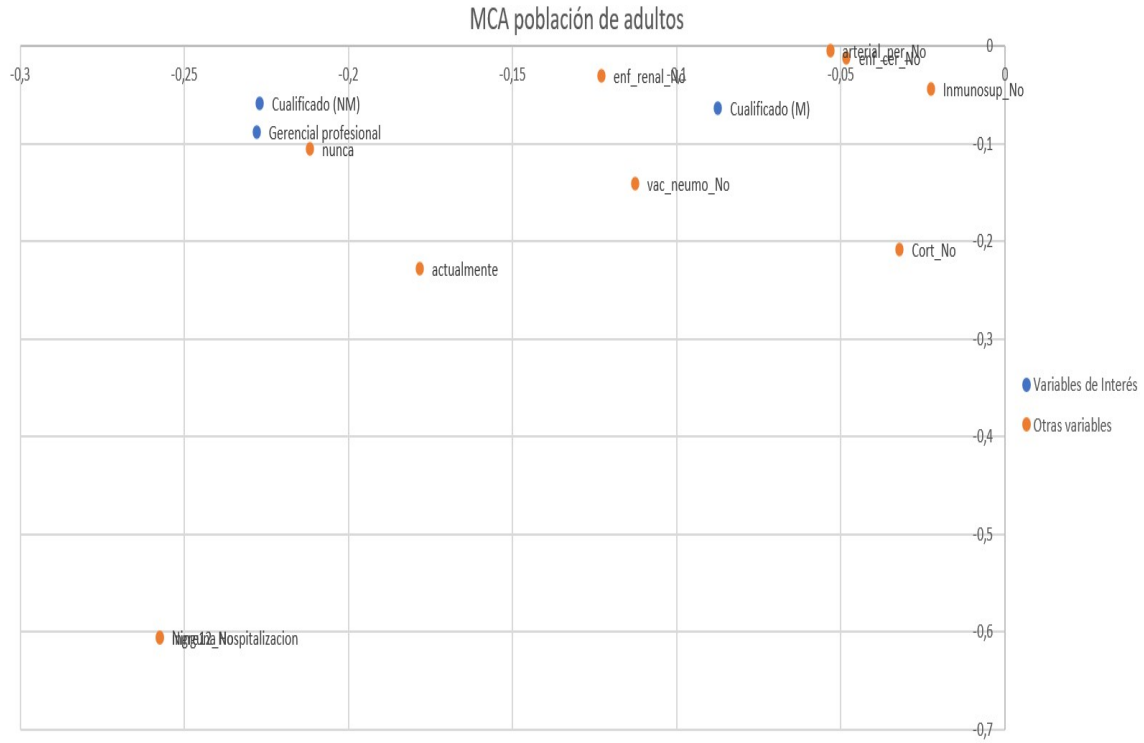

Figura 10: Mapa con las variables del MCA en adultos en el tercer cuadrante

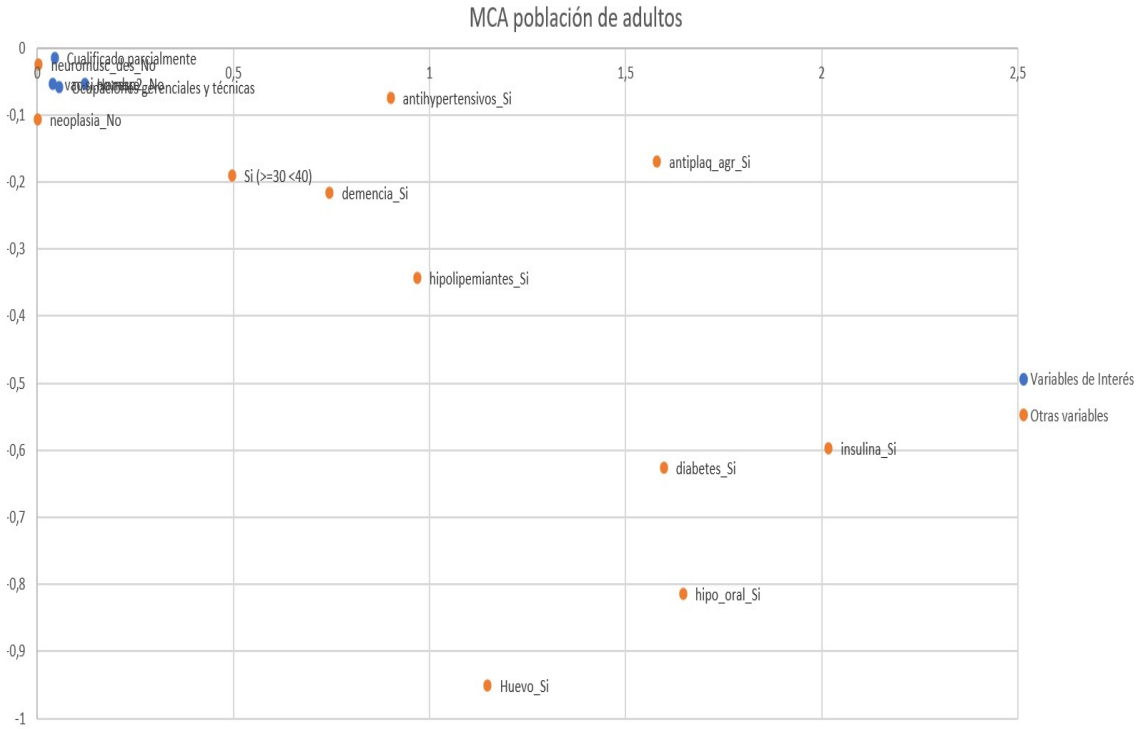

Figura 11: Mapa con las variables del MCA en adultos en el cuarto cuadrante

Población mayor de 65 años

En el siguiente análisis de correspondencias múltiples para la población de gente mayor de 65 años, se han seleccionado dos componentes observando el gráfico de varianzas explicadas.

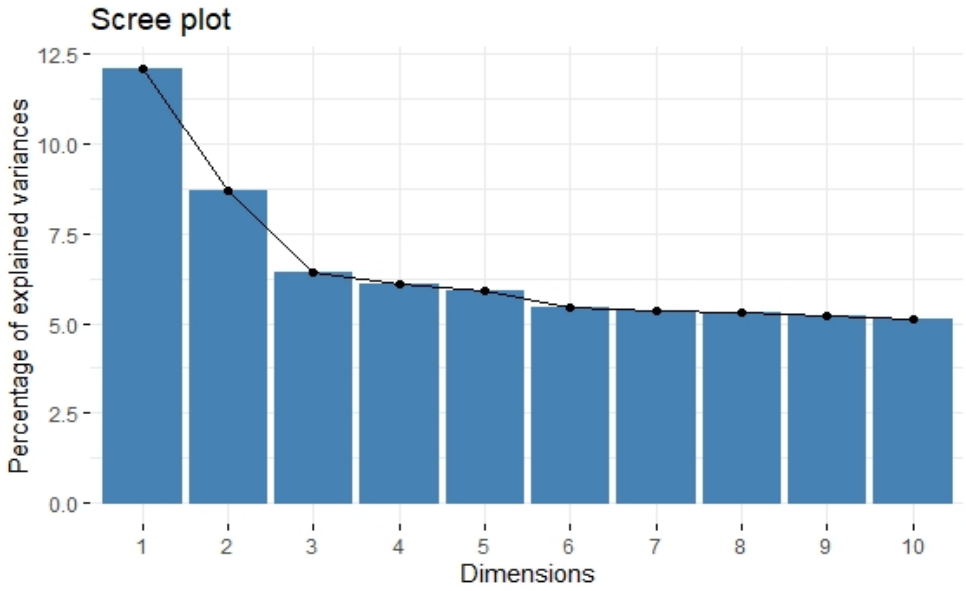

Figura 12: Varianza explicada por cada componente del MCA, población mayores de 65

Para finalizar, se han graficado las variables en las dos dimensiones que explican un 20,82 % de la variabilidad. En el siguiente gráfico, se puede observar como en el primer cuadrante, eje de la primera componente en la parte derecha y el eje de la segunda componente en la parte superior, se encuentran categorías como haber sido ingresado en los últimos 12 meses o haber sido hospitalizado entre 1 y 3 veces o más de 3 o tomar corticoesteroides. En el segundo cuadrante el cual es la parte izquierda del eje de la primera componente y la parte superior del eje de la segunda componente, la variables que más definen esta componente son haber dejado de fumar hace menos de 1 año o fumar, 5 cigarros o menos al día o fumar actualmente, no tomar antihypertensivos y no padecer enfermedades de corazón. En el tercer cuadrante el cual es la parte izquierda del eje horizontal y la parte inferior del eje vertical, se encuentran no haber sufrido hospitalizaciones en los últimos 3 meses, no haber tenido ingresos en los últimos 12 meses, ni haber fumado nunca. En el último cuadrante, se encuentran las categorías de tomar antihypertensivos, tener anemia, enfermedades cardíacas, enfermedades renales, alguna enfermedad endocrina que no sea la diabetes, tomar insulina, anticoagulantes, ser obeso u obeso morbido. Para finalizar, se puede ver como las variables de reticentes, a pesar de estar relativamente centrada, tiene cerca las clases de las variables de no tomar anticoagulantes, dejar de fumar hace más de un año, tomar corticoesteroides, ser hombre y no estar obeso. En caso contrario, la categoría de no reticentes está alrededor de las categorías de no padecer algunas enfermedades como la anemia o la demencia, no tomar corticoesteroides, ser gente sin cualificación o no haberse vacunado ante el neumococo.

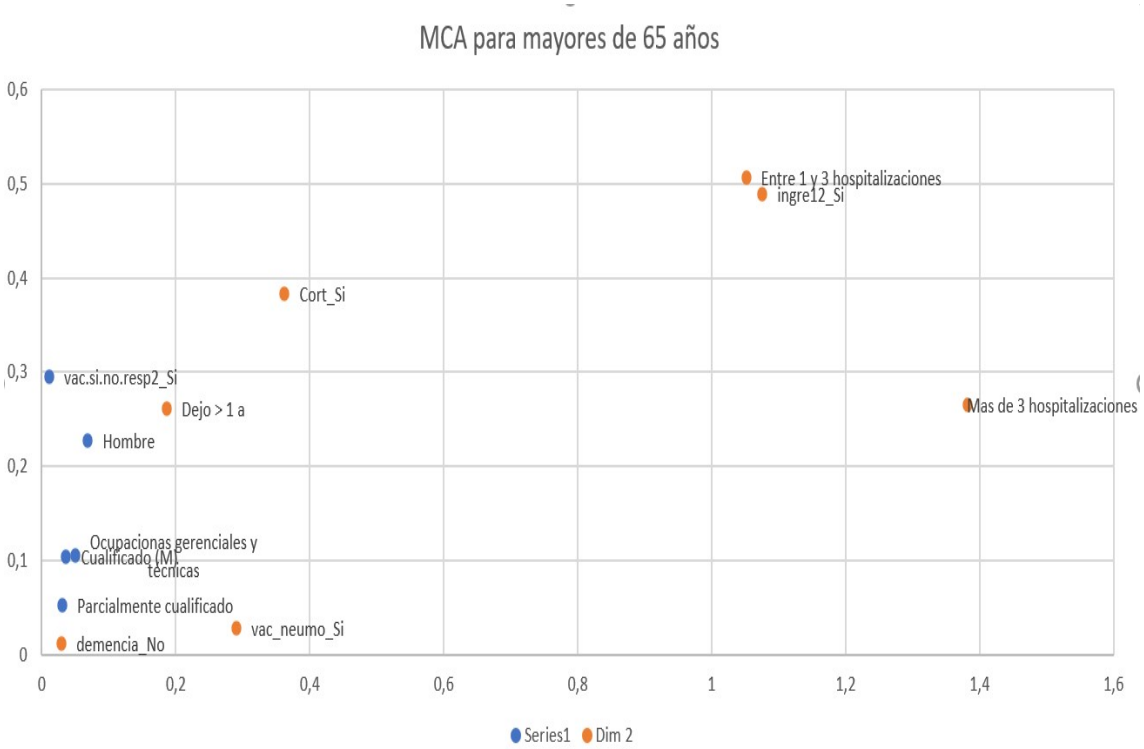

Figura 13: Mapa con las variables del MCA en mayores de 65 en el primer cuadrante

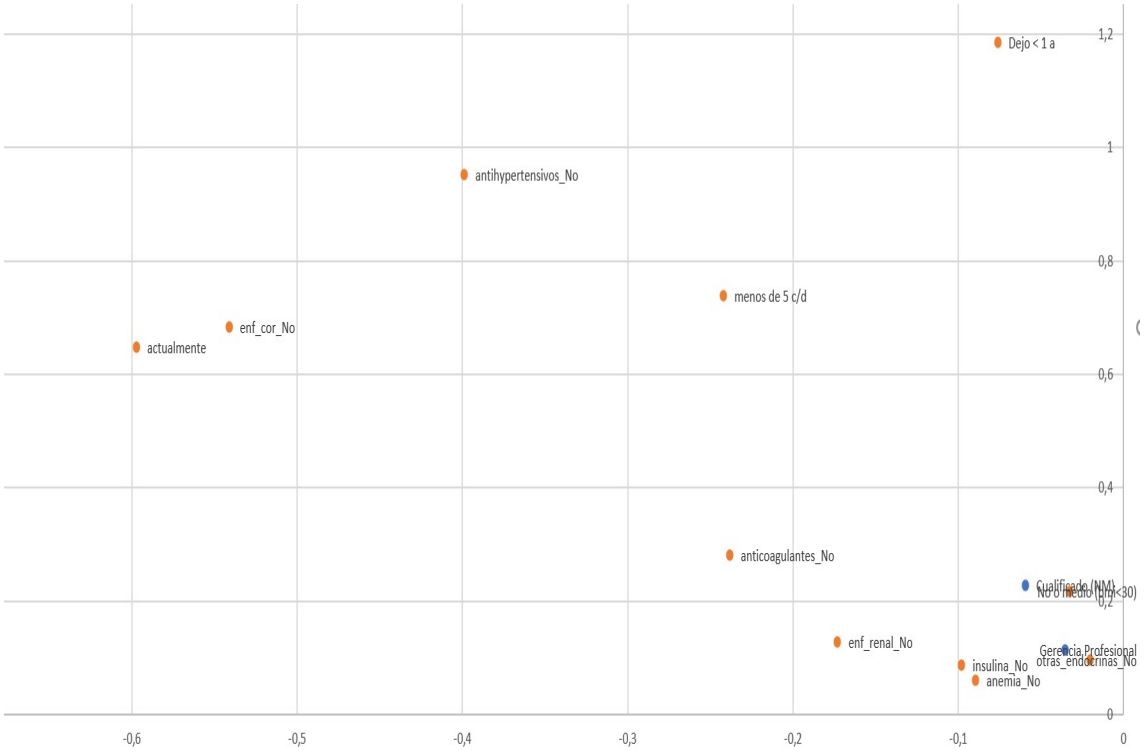

Figura 14: Mapa con las variables del MCA en mayores de 65 en el segundo cuadrante

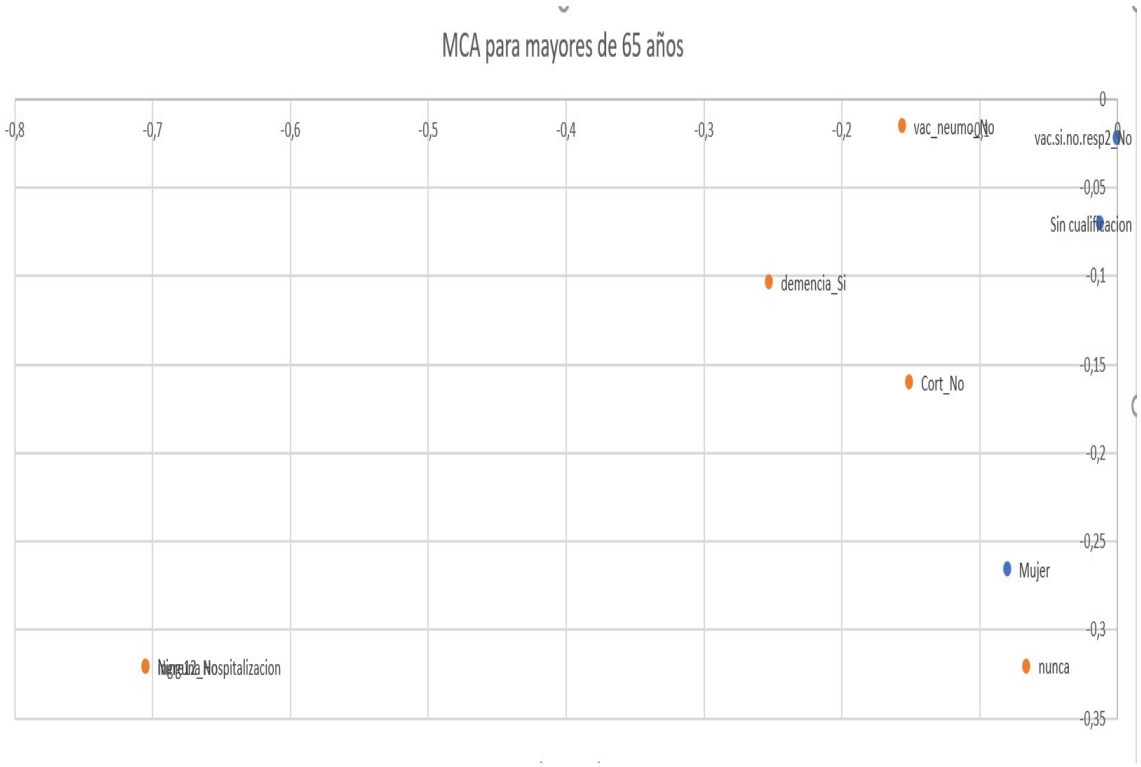

Figura 15: Mapa con las variables del MCA en mayores de 65 en el tercer cuadrante

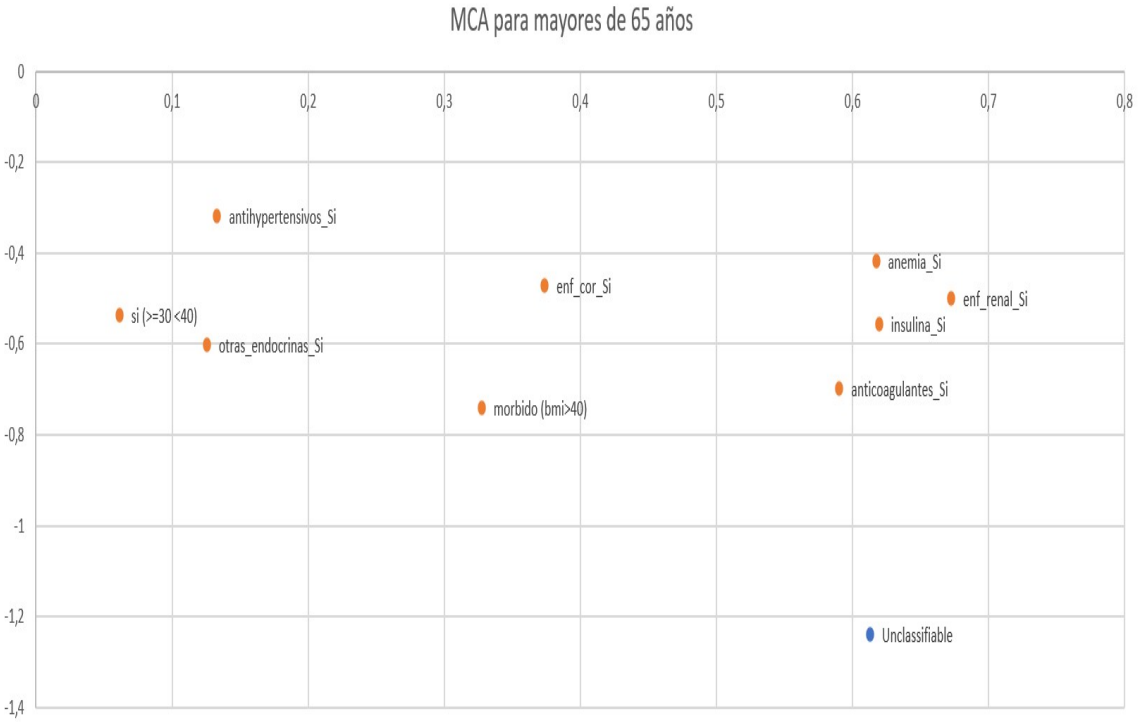

Figura 16: Mapa con las variables del MCA en mayores de 65 en el cuarto cuadrante

4.3. Comparación y selección del mejor modelo

En esta sección para abordar el tercer objetivo, se realizará una comparación de los distintos modelos explicados anteriormente. En este proceso, se realizará un hold out repetido 100 veces, para las dos poblaciones, dividiendo la muestra en un 70 % entrenamiento y un 30 % validación, elegidos aleatoriamente.

En el conjunto de entrenamiento, se ha sobremuestreado los casos de la variable de interés para igualar el número de casos en el conjunto de entrenamiento, dejando los conjuntos de validación sin sobremuestrear. Para finalizar, se obtendrán tres estadísticos de comparación como son la tasa de acierto del modelo, el área bajo la curva ROC del modelo y el coeficiente de correlation de Matthews para ambas poblaciones.

Población Adulta

Para los modelos de la población adulta, se ha visto como en ningún modelo la clase social inclasificable, es decir, una clase social a la cual no se le podía asignar ninguna otra, era significativa. Con lo cual, para los modelos siguientes no se tendrá en cuenta. Con la variable demencia, ya que hay muy pocos casos, tampoco se tendrá en cuenta. A continuación se muestran la media y los intervalos de confianza al 95 % por el método de los percentiles para los estadísticos anteriormente citados:

Modelo	Media Tasa acierto	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,7735	0,7309	0,8248
Regresión logística Stepwise	0,7743	0,7376	0,8239
Árbol de Clasificación por defecto	0,7151	0,6299	0,7908
Árbol de Clasificación Podado	0,7813	0,7513	0,8101
Random Forest	0,8137	0,778	0,8452

Tabla 23: Tasa de aciertos en media e intervalos de confianza para los modelos de adultos

Modelo	Media MCC	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,2897	0,1953	0,4122
Regresión logística Stepwise	0,2901	0,1887	0,4059
Árbol de Clasificación por defecto	0,2620	0,1726	0,3500
Árbol de Clasificación Podado	0,1439	0,0644	0,2360
Random Forest	0,2695	0,1653	0,3660

Tabla 24: Media de coeficiente de correlation de Matthews e intervalos de confianza para los modelos de adultos

Modelo	Media Área ROC	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,7234	0,6713	0,7872
Regresión logística Stepwise	0,7248	0,6727	0,7885
Árbol de Clasificación por defecto	0,6762	0,6124	0,7307
Árbol de Clasificación Podado	0,5613	0,5241	0,6031
Random Forest	0,7201	0,6646	0,7749

Tabla 25: Media del área bajo la curva ROC e intervalos de confianza para los modelos de adultos

Para acabar de comparar los modelos, se han realizado un test de la t de Student, con muestras emparejadas, comparando los modelos de la regresión logística, normal y stepwise, y el random forest entre sí con una probabilidad de falsa alarma del 5 %. No se tendrán en cuenta los dos árboles de clasificación porque ambos tienen un área bajo la curva ROC muy baja, teniendo uno la tasa de aciertos más baja y el otro un coeficiente de correlation de Matthews muy reducido. Al ser estadísticos de muestras aleatorias y tener una población de 100 casos por cada método y cada estadístico, se supone la normalidad de las muestras por el teorema central del límite.

En primer lugar, se aprecia que realizando los test no hay diferencias estadísticas entre los dos modelos de regresión, exceptuando el área bajo la curva ROC que se aprecia como la regresión logística stepwise es significativamente superior en media.

En segundo lugar, se aprecia como el random forest tiene una tasa de acierto significativamente más elevada y un coeficiente de correlation de Matthews significativamente más bajo, no habiendo diferencias significativas para el área bajo la curva ROC.

Con lo cual, al ver que la regresión logística stepwise tiene un coeficiente de correlation de Matthews más elevado que el random forest y este estadístico nos permite identificar la clase minoritaria, que en nuestro caso es la de interés, aún teniendo una tasa de acierto más baja, se puede concluir que la regresión logística stepwise es el mejor modelo.

Población Mayor de 65 años

Para esta población, se ha realizado de manera similar obteniéndose la media y los intervalos de confianza de los mismos estadísticos. Como el único árbol de clasificación que predecía algo que no fuera la clase mayoritaria era el árbol completo, se ha decidido no tener en cuenta los otros dos modelos trabajando con los dos modelos de regresión, el árbol de clasificación completo y el random Forest. A continuación se muestran los resultados:

Modelo	Media Tasa acierto	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,8045	0,7797	0,8258
Regresión logística Stepwise	0,8042	0,7784	0,8279
Árbol de Clasificación Completo	0,8426	0,8221	0,8622
Random Forest	0,9174	0,9061	0,9268

Tabla 26: Tasa de aciertos en media e intervalos de confianza para los modelos de mayores de 65

Modelo	Media MCC	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,1422	0,0915	0,1937
Regresión logística Stepwise	0,1416	0,0946	0,1930
Árbol de Clasificación Completo	0,052	0,0061	0,0992
Random Forest	0,0328	-0,0193	0,0867

Tabla 27: coeficiente de correlation de Matthews promedio e intervalos de confianza para los modelos de mayores de 65

Modelo	Media Área ROC	I.C. límite inf.	I.C. límite sup.
Regresión logística	0,6727	0,6329	0,7120
Regresión logística Stepwise	0,6723	0,6317	0,7112
Árbol de Clasificación Completo	0,5321	0,4981	0,5600
Random Forest	0,6135	0,5557	0,6568

Tabla 28: Media del área bajo la curva ROC e intervalos de confianza para los modelos de mayores de 65

Para acabar de comparar los modelos al igual que en el caso anterior, se han realizado un test de la t de Student, con muestras emparejadas, comparando todos los modelos anteriormente citados con $\alpha = 0,05$, se considera que todos son muestras normales por el

teorema central del límite.

En primer lugar, comparando la regresión logística con la regresión logística stepwise se aprecia que no hay significación que los estadísticos no son significativos exceptuando el área bajo la curva ROC que es significativamente mayor en la regresión logística, en comparación con el árbol de regresión y el random forest, los modelos de regresión tienen una menor tasa de acierto pero el área bajo la curva ROC y el coeficiente de correlation de Matthews es significativamente mayor.

En segundo lugar, se puede ver como el árbol de clasificación tiene un mayor coeficiente de correlation de Matthews, pero el Random Forest tiene una mayor tasa de acierto y una mayor área bajo la curva ROC.

En conclusión, la regresión logística predice peor en general que los modelos basados en árboles pero predice mejor la clase minoritaria que es la clase de interés, es decir, para nuestro interés la regresión logística es el mejor método.

5. Conclusiones

En conclusión, se ha conseguido identificar al colectivo de las personas que no han asistido a vacunarse en tres campañas de la gripe consecutivas padeciendo alguna enfermedad respiratoria segmentando este colectivo en 3 grupos de edad en el momento de la hospitalización. También, se ha conseguido examinar la relación que existen entre la pertenencia a este grupo y los segmentos de población adulta y de los mayores de 65 años caracterizando estos grupos, el grupo de los adultos que son reticentes a la vacuna padeciendo alguna enfermedad respiratoria son las personas sin enfermedades (exceptuando el asma o la bronquitis crónica), que toman corticoesteroides (tratamiento utilizado para el asma), obesas con un índice de masa corporal entre 30 y 40, las cuales han sufrido más hospitalizaciones el año anterior, sin haberse puesto la vacuna del neumococo y que fuman actualmente. Para el grupo de mayores de 65 años, se encuentra al grupo de personas sin enfermedades (más allá de la enfermedad respiratoria que le hacen pertenecer a este grupo), que toman corticoesteroides, con obesidad morbida, que fuman actualmente o han dejado de fumar hace menos de un año.

Como se ha podido ver, estas conclusiones son diferentes con el estudio citado anteriormente [9], ya que las variables que se obtenían de importancia era el sexo, la clase social y educación, las variables asociadas a la no vacunación era ser mujer, tener pocos estudios y ser de una clase social baja. En nuestro estudio para personas mayores de 65 años (aún siendo un grupo que padece enfermedades respiratorias), el sexo y la clase social han resultado no significativos.

Finalmente, se han comparado los modelos y se ha llegado a la conclusión que el mejor modelo para el grupo de adultos es la regresión logísitca stepwise ya tiene una área

significativamente mayor bajo la curva ROC que la regresión logística, sin tener el resto de estadísitco una diferencia significativa, y predice mejor la clase minoritaria, siendo esta la de reticentes que es la nuestra de interés, que el Random Forest aún teniendo una tasa de acierto significativamente menor. En el caso de la población mayor de 65 años, se ha llegado a la conclusión de que el mejor modelo es la regresión logística ya que a diferencia del caso anterior, la curva bajo la curva ROC es significativamente mayor en este modelo en comparación a la regresión logística stepwise (el resto de estadísticos no tienen diferencias estadísticas significativas) y al igual que antes, estos dos modelos tienen un coeficiente de correlation de Matthews y un área bajo la curva ROC significativamente mayor que los dos modelos basados en árboles (random forest y el árbol de clasificación, completo) prediciendo la clase minoritaria de mejor manera, ya que es la clase de interés, aún teniendo una tasa de acierto significativamente menor. Con lo cual, los modelos de regresión se ajustan de mejor manera a la hora de predecir los casos de nuestro interés, que son los reticentes, en ambos conjuntos de pacientes.

Este trabajo tiene importantes implicaciones para aportar nuevo conocimiento sobre la población que no decide vacunarse. Los resultados de este trabajo pueden utilizarse a nivel práctico en nuevas campañas de vacunación dirigiéndolas a personas que se sienten sanas, ya que en ambos grupos se ha visto que las personas que no se vacunan son aquellas que no están enfermas (exceptuando la enfermedad respiratoria) y que tienen un menor cuidado de su salud, ya que estamos hablando de personas obesas, que fuman y han sufrido varias hospitalizaciones el año anterior.

Caben mencionar distintas limitaciones asociadas al trabajo. En primer lugar, la muestra procede de un conjunto de datos observacional sobre ingresos hospitalarios dentro de la Comunidad Valenciana, lo que limita el alcance de las conclusiones a este ámbito. En segundo lugar, la información disponible es principalmente cualitativa. Por último, un mayor nivel de incertidumbre asociado al estudio del comportamiento humano y actitudes.

En este trabajo, se ha analizado la no vacunación asociada a los pacientes ingresados con enfermedades respiratorias crónicas como asma y bronquitis. En futuros trabajos, se estudiarán las causas por las cuales las personas no se vacunan sin asociarlo a ningún grupo en particular.

Por otro lado, la muestra era limitada, como ya se ha comentado, asociada únicamente a ingresos hospitalarios en la Comunidad Valenciana. Así mismo, para el futuro se requerirán una masa de datos más amplia pudiendo extender la muestra al resto del país. Además, se pueden considerar otras variables como el nivel de estudio que tiene la persona para homogeneizar más el estudio con otros realizados.

Agradecimientos

Agradezco muy sinceramente la ayuda que he recibido tanto de docentes como directivos de FISABIO. Y más concretamente, a la directora del trabajo, Andrea Conchado

Peiró, y a Javier Díez Domingo, jefe del Área de Investigación en Vacunas de FISABIO.

Referencias

[1] World Health Organization (WHO). Vaccines against influenza WHO position paper, November 2012. Weekly Epidemiological Record (2012), Volume 87, Issue 47, pages: 461-76.

[2] World Health Organization. 2021. Ïnfluenza (seasonal)". Accesed 9 September, 2021. https://www.who.int/es/news-room/fact-sheets/detail/influenza-(seasonal)

[3] Mira-Iglesias, A., López-Labrador, F.X., Guglieri-López, B., Tortajada-Girbés, M., Baselga-Moreno, V., Cano, L., et al. (2018). Influenza vaccine effectiveness in preventing hospitalisation of individuals 60 years of age and over with laboratory-confirmed influenza, Valencia Region, Spain, influenza season 2016/17. *Eurosurveillance* (2018), Volume 23, Issue 8, pii=17-00318. https://doi.org/10.2807/1560-7917.ES.2018.23.8.17-00318

[4] Mira-Iglesias A., López-Labrador F.X., Baselga-Moreno V., Tortajada-Girbés M., Mollar-Maseres J., Carballido-Fernández M., et al. (2019). Influenza vaccine effectiveness against laboratory-confirmed influenza in hospitalised adults aged 60 years or older, Valencia Region, Spain, 2017/18 influenza season. *Eurosurveillance* (2019), Volume 24, Issue 31, pii=1800461. https://doi.org/10.2807/1560-7917.ES.2019.24.31.1800461

[5] Mira-Iglesias, A., López-Labrador, F.X., García-Rubio, J., Mengual-Chuliá, B., Tortajada-Girbés, M., Mollar-Maseres, J., et al. (2021). Influenza Vaccine Effectiveness and Waning Effect in Hospitalized Older Adults. Valencia Region, Spain, 2018/2019 Season. *Enviromental Research and Public Health*, Volume 18, Issue 3, pii=1129. https://doi.org/10.3390/ijerph18031129

[6] Xie M., Liu X., Cao X. et al. . Trends in prevalence and incidence of chronic respiratory diseases from 1990 to 2017. *Respiratory Research* (2020), Volume 21, Issue 49. https://doi.org/10.1186/s12931-020-1291-8

[7] European Respiratory Journal. "2015/2016 Year of SEPAR". Accesed 9 September, 2021. http://www.ersnet.org/the-society/news/year-of-separ

[8] Wong C. M., et al. Cigarette smoking as a risk factor for influenza-associated mortality: evidence from an elderly cohort. Influenza and Other Respiratory Viruses. *Influenza Journal* (2012), Volume 7, Issue 4, 531-539. https://doi.org/10.1111/j.1750-2659.2012.00411.x

[9] Dios-Guerra, C., Carmona-Torres, J.M., López-Soto, P.J., Morales-Cané, I., Rodríguez-Borrego, M.A., Prevalence and factors associated with influenza vaccination of persons over 65 years old in Spain (2009-2014). *Vaccine* (2017), Volume 35, Issue 51, 7095-7100, ISSN 0264-410X. https://doi.org/10.1016/j.vaccine.2017.10.086

[10] R Core Team (2021). R: A language and environment for statistical computing. *RFoundation for Statistical Computing*, Vienna, Austria. URL: https://www.R-project.org/.

[11] Wickham, H. and Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

[12] Kaplan, J. (2020). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. R package version 1.6.3. https://CRAN.R-project.org/package=fastDummies

[13] Torgo, L. (2016). Data Mining with R, learning with case studies, 2nd edition, Chapman and Hall/CRC. URL: http://ltorgo.github.io/DMwR2

[14] Therneau, T. and Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

[15] Milborrow, S. (2020). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.9. https://CRAN.R-project.org/package=rpart.plot

[16] Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18-22.

[17] Sing T., Sander O., Beerenwinkel N. and Lengauer, T.. ROCR: visualizing classifier performance in R. *Bioinformatics* (2005), Volume 21, Issue 20, 3940-3941. https://doi.org/10.1093/bioinformatics/bti623

[18] Abdi, H. and Valentin, D. Multiple Correspondence Analysis. *Encyclopedia of measurement and statistics* (2007), 651-657.

[19] Le, S., Josse, J., Husson, F.. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* (2005), 25(1), 1-18. https://doi.org/10.18637/jss.v025.i01

[20] Kassambara, A. and Mundt, F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses* (2020). R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Anexos

Programa de Limpieza de Datos

```
library("readxl")
my_data <- read_excel("TFM_Nico_UPV_con_etiquetas.xls")
my_data<-my_data[,-c(2,4,7,8,9,10,11,12,13,14,43,44,46,47,49,50,52,53,54,55,
                     56,57,58,59,60,61,62,68,76,77,78,79,80,81,82)]
n<-nrow(my_data)
vac.temp.act<-NA
vac.temp.ant<-NA
vac.temp.ant2<-NA
for(i in 1:n){
  if (my_data$season[i]=="15/16"){
    vec.aux<-my_data$r2015sfv[i]
    vac.temp.act[i]<-vec.aux
    vec.aux<-my_data$r2014sfv[i]
    vac.temp.ant[i]<-vec.aux
    vec.aux<-my_data$r2013sfv[i]
    vac.temp.ant2[i]<-vec.aux
  }
  if (my_data$season[i]=="16/17"){
    vec.aux<-my_data$r2016sfv[i]
    vac.temp.act[i]<-vec.aux
    vec.aux<-my_data$r2015sfv[i]
    vac.temp.ant[i]<-vec.aux
    vec.aux<-my_data$r2014sfv[i]
    vac.temp.ant2[i]<-vec.aux
  }
  if (my_data$season[i]=="17/18"){
    vec.aux<-my_data$r2017sfv[i]
    vac.temp.act[i]<-vec.aux
    vec.aux<-my_data$r2016sfv[i]
    vac.temp.ant[i]<-vec.aux
    vec.aux<-my_data$r2015sfv[i]
    vac.temp.ant2[i]<-vec.aux
  }
  if (my_data$season[i]=="18/19"){
    vec.aux<-my_data$r2018sfv[i]
    vac.temp.act[i]<-vec.aux
    vec.aux<-my_data$r2017sfv[i]
    vac.temp.ant[i]<-vec.aux
    vec.aux<-my_data$r2016sfv[i]
    vac.temp.ant2[i]<-vec.aux
```

```
  }
}
filas<-NA
n1<-1
for(i in 1:n){
  if(is.na(vac.temp.act[i])){
    filas[n1]<-i
    n1<-n1+1
  }
}

vac.temp.act<-vac.temp.act[-filas]
vac.temp.ant<-vac.temp.ant[-filas]
vac.temp.ant2<-vac.temp.ant2[-filas]
my_data<-my_data[-filas,]
n<-nrow(my_data)
vac.si.no<-NA
for(i in 1:n){
  if(vac.temp.act[i]=="No"&&vac.temp.ant[i]=="No"&&vac.temp.ant2[i]=="No"){
    vac.si.no[i]<-"No"
  }
  else{
    vac.si.no[i]<-"Yes"
  }
}
vac.temp.act.aux<-NA
vac.temp.ant.aux<-NA
vac.temp.ant2.aux<-NA
for(i in 1:n){
  if(vac.temp.act[i]=="No"){
    vac.temp.act.aux[i]<-0
  }
  else{
    vac.temp.act.aux[i]<-1
  }
  if(vac.temp.ant[i]=="No"){
    vac.temp.ant.aux[i]<-0
  }
  else{
    vac.temp.ant.aux[i]<-1
  }
  if(vac.temp.ant2[i]=="No"){
    vac.temp.ant2.aux[i]<-0
  }
```

```
  else{
    vac.temp.ant2.aux[i]<-1
  }
}
vac.niveles<-NA
for(i in 1:n){
  if(vac.temp.act.aux[i]+vac.temp.ant.aux[i]+vac.temp.ant2.aux[i]==0){
    vac.niveles[i]<-"Nunca"
  }
  if(vac.temp.act.aux[i]+vac.temp.ant.aux[i]+vac.temp.ant2.aux[i]==1){
    vac.niveles[i]<-"1 Temporada"
  }
  if(vac.temp.act.aux[i]+vac.temp.ant.aux[i]+vac.temp.ant2.aux[i]==2){
    vac.niveles[i]<-"2 Temporadas"
  }
  if(vac.temp.act.aux[i]+vac.temp.ant.aux[i]+vac.temp.ant2.aux[i]==3){
    vac.niveles[i]<-"Todas"
  }
}
filas.no.vac<-NA
n1<-1
for(i in 1:n){
  if(vac.si.no[i]=="No"){
    filas.no.vac[n1]<-i
    n1<-n1+1
  }
}
vac.si.no.resp<-NA
for(i in 1:n){
  if(my_data$asthma[i]=="Yes"||my_data$abronchitis[i]=="Yes"){
    condition<-"Yes"
  }
  else{
   condition<-"No"
  }
  if(vac.si.no[i]=="Yes"&&condition=="Yes"){
    vac.si.no.resp[i]<-"Yes"
  }
  else{
    vac.si.no.resp[i]<-"No"
  }
}
my_data<-cbind(my_data[,c(1:40)],vac.temp.act,vac.temp.ant,vac.temp.ant2,
               vac.si.no,vac.niveles,vac.si.no.resp,my_data[,47])
```

```
write.csv2(my_data,"datos_TFM_con_etiquetas.csv")
n<-nrow(my_data)
m<-ncol(my_data)
for(i in 1:n){
  for(j in 1:m){
    if(is.na(my_data[i,j])){
      my_data[i,j]<-NA
    }
  }
}
my_data.vac.si<-my_data[-filas.no.vac,]
my_data.vac.no<-my_data[filas.no.vac,]
filas_menos18<-NA
k<-1
for(i in 1:n){
  if(is.na(my_data$yoa[i])){

  }
  else{
    if(my_data$yoa[i]<=18){
      filas_menos18[k]<-i
      k<-k+1
    }
  }
}
my_data_menos18<-my_data[filas_menos18,]

filas_adultos<-NA
k<-1
for(i in 1:n){
  if(is.na(my_data$yoa[i])){

  }
  else{
    if(my_data$yoa[i]>18&&my_data$yoa[i]<=65){
      filas_adultos[k]<-i
      k<-k+1
    }
  }
}
my_data_adultos<-my_data[filas_adultos,]

filas_mayores<-NA
k<-1
```

```
for(i in 1:n){
  if(is.na(my_data$yoa[i])){

  }
  else{
    if(my_data$yoa[i]>65){
      filas_mayores[k]<-i
      k<-k+1
    }
  }
}
my_data_mayores<-my_data[filas_mayores,]
write.csv2(my_data_menos18, "TFM_Nico_UPV_con_etiquetas_menos_18.csv",
           row.names=FALSE, na="")
write.csv2(my_data_adultos, "TFM_Nico_UPV_con_etiquetas_adultos.csv",
           row.names=FALSE, na="")
write.csv2(my_data_mayores, "TFM_Nico_UPV_con_etiquetas_mayores.csv",
           row.names=FALSE, na="")
```

Regresión Logística, Árboles de Regresión, Random Forest y comparación de métodos en población de adultos

```
library("readxl")
library("fastDummies")
library(MASS)
library(rpart)
library(DMwR2)
library(randomForest)
library(rpart.plot)
library(ROCR)
set.seed(1234)

my_data_adultos <- read_excel("TFM_Nico_UPV_con_etiquetas_adultos.xlsx")
n<-nrow(my_data_adultos)
for(i in 1:n){
  if(my_data_adultos$Cort[i]!="No"){
    my_data_adultos$Cort[i]<-"Si"
  }
}
n1<-1
h<-NA
for (i in 1:n) {
  if(is.na(my_data_adultos$hosp_anyo_ant[i])){
    h[n1]<-i
    n1<-n1+1
  }
}
my_data_adultos<-my_data_adultos[-h,]


########################## Analisis Adultos############################
adultos<-dummy_cols(
  my_data_adultos,
  select_columns = NULL,
  remove_first_dummy = TRUE
)
adultos<-adultos[,-c(1:27,30:38,40:46)]
adultos<-adultos[,-c(5:11,16,17,57:65)]

#Regresion logistica
logit.reg_adultos<-glm(factor(vac.si.no.resp2)~ .,
                       data =adultos[,c("vac.si.no.resp2","enf_cer_Si",
                        "arterial_per_Si","diabetes_Si",
```

```
                          "enf_renal_Si","neuromusc_des_Si",
                          "neoplasia_Si",demencia_Si",
                          "antihypertensivos_Si","antiplaq_agr_Si",
                          "hipolipemiantes_Si","insulina_Si",
                          "hipo_oral_Si","Inmunosup_Si",
                          "Cort_Si","Obes_No o medio (bmi<30)",
                          "Obes_Si (>=30 <40)","Fumar_Dejo > 1 a",
                          "Fumar_nunca",
                          "Fumar_Dejo < 1 a","Fumar_menos de 5 c/d",
                          "Clase_social_Cualificado parcialmente",
                          "Clase_social_Gerencial profesional",
                          "Clase_social_Ocupaciones gerenciales y tecnicas",
                          "Clase_social_Cualificado (NM)",
                          "Clase_social_Inclasificable",
                          "Clase_social_Sin cualificacion",
                          "hosp_anyo_ant","ingre12_Si","Huevo_Si",
                          "vac_neumo_Si","Peso","Altura")],
                        family = "binomial")
summary(logit.reg_adultos)
logit.reg_adultos_null<-glm(factor(vac.si.no.resp2)~1,
                            data =adultos[,c("vac.si.no.resp2","enf_cer_Si",
                      "arterial_per_Si","diabetes_Si",
                      "enf_renal_Si","neuromusc_des_Si","neoplasia_Si",
                      "demencia_Si","antihypertensivos_Si",
                      "antiplaq_agr_Si",
                      "hipolipemiantes_Si","insulina_Si",
                      "hipo_oral_Si",
                      "Inmunosup_Si","Cort_Si",
                      "Obes_No o medio (bmi<30)","Obes_Si (>=30 <40)",
                      "Fumar_Dejo > 1 a","Fumar_nunca",
                      "Fumar_Dejo < 1 a","Fumar_menos de 5 c/d",
                      "Clase_social_Cualificado parcialmente",
                      "Clase_social_Gerencial profesional",
                      "Clase_social_Ocupaciones gerenciales y tecnicas",
                      "Clase_social_Cualificado (NM)",
                      "Clase_social_Inclasificable",
                      "Clase_social_Sin cualificacion",
                      "hosp_anyo_ant","ingre12_Si","Huevo_Si",
                      "vac_neumo_Si","Peso","Altura")],
                            family = "binomial")
1-logLik(logit.reg_adultos)/logLik(logit.reg_adultos_null)

#Regresion logistica stepwise
logit.reg_step_adultos<-step(logit.reg_adultos)
```

```
vif(logit.reg_step_adultos)
logit.reg_step_adultos_null<-glm(factor(vac.si.no.resp2)~1,
                                 data =adultos[,c("vac.si.no.resp2",
                                 "neuromusc_des_Si",
                                 "neoplasia_Si","demencia_Si",
                                 "antihypertensivos_Si"
                                 ,"antiplaq_agr_Si",
                        "hipolipemiantes_Si","insulina_Si","Cort_Si",
                        "Obes_Si (>=30 <40)","Fumar_nunca",
                        "Clase_social_Gerencial profesional",
                        "Clase_social_Ocupaciones gerenciales y tecnicas",
                        "Clase_social_Inclasificable","hosp_anyo_ant",
                        "vac_neumo_Si","Altura")], family = "binomial")
1-logLik(logit.reg_step_adultos)/logLik(logit.reg_step_adultos_null)

#Class tree por defecto
default.ct_adultos <- rpart(factor(vac.si.no.resp2) ~ .,
                            data=my_data_adultos[,c("vac.si.no.resp2",
                            "enf_cer","arterial_per","diabetes",
                            "enf_renal","neuromusc_des",
                            "neoplasia","demencia",
                            "antihypertensivos","antiplaq_agr",
                            "hipolipemiantes",
                            "insulina","hipo_oral",
                            "Inmunosup","Cort",
                            "Obes",
                            "Fumar",
                            "Clase_social",
                            "hosp_anyo_ant","ingre12","Huevo",
                            "vac_neumo","Peso","Altura")],
                            method = "class")
X11()
prp(default.ct_adultos, extra=101)

#Class tree completo
treefull_adultos <- rpart(factor(vac.si.no.resp2)~ .,
                          data=my_data_adultos[,c("vac.si.no.resp2",
                          "enf_cer","arterial_per","diabetes",
                          "enf_renal","neuromusc_des","neoplasia",
                          "demencia",
                          "antihypertensivos","antiplaq_agr",
                          "hipolipemiantes",
                          "insulina","hipo_oral",
                          "Inmunosup","Cort",
```

```
                            "Obes",
                            "Fumar",
                            "Clase_social",
                            "hosp_anyo_ant","ingre12","Huevo",
                            "vac_neumo","Peso","Altura")],
                            method="class", cp=0.001)
X11()
prp(treefull_adultos, extra = 101)

#Class Tree podado
tree_adultos<-rpartXse(factor(vac.si.no.resp2)~ . , se = 0.25,
                         data=my_data_adultos[,c("vac.si.no.resp2",
                         "enf_cer","arterial_per","diabetes",
                         "enf_renal","neuromusc_des","neoplasia",
                         "demencia",
                         "antihypertensivos","antiplaq_agr",
                         "hipolipemiantes","insulina","hipo_oral",
                         "Inmunosup","Cort",
                         "Obes",
                         "Fumar",
                         "Clase_social",
                         "hosp_anyo_ant","ingre12","Huevo",
                         "vac_neumo","Peso","Altura")],
                         model=TRUE)
x11()
prp(tree_adultos, extra = 101)

#Random Forest
rf_adultos<-randomForest(factor(vac.si.no.resp2) ~ .,
                          data =my_data_adultos[,c("vac.si.no.resp2",
                          "enf_cer","arterial_per","diabetes",
                          "enf_renal","neuromusc_des","neoplasia",
                          "demencia","antihypertensivos",
                          "antiplaq_agr",
                          "hipolipemiantes","insulina","hipo_oral",
                          "Inmunosup","Cort",
                          "Obes",
                          "Fumar",
                          "Clase_social",
                          "hosp_anyo_ant","ingre12","Huevo",
                          "vac_neumo","Peso","Altura")],
                          mtry=4, method="class",importance=TRUE)
X11()
varImpPlot(rf_adultos, main="",col="dark blue")
```

```
n1<-1
h<-NA
for (i in 1:nrow(my_data_adultos)){
  if(my_data_adultos$Clase_social[i]=="Inclasificable"){
    h[n1]<-i
    n1<-n1+1
  }
}
my_data_adultos<-my_data_adultos[-h,]
adultos<-adultos[-h,]

#######Hold out oversampling (CURVAS ROC, TASA ACIERTO y MCC)########
perf.reg_log<-NA
tasa.aciertos.reg_log<-NA
mcc.reg_log<-NA

perf.reg_log_step<-NA
tasa.aciertos.reg_log_step<-NA
mcc.reg_log_step<-NA

perf.def_ct<-NA
tasa.aciertos.def_ct<-NA
mcc.def_ct<-NA

perf.tree_podado<-NA
tasa.aciertos.tree_podado<-NA
mcc.tree_podado<-NA

perf.rf<-NA
tasa.aciertos.rf<-NA
mcc.rf<-NA
for (i in 1:100) {
  train.index <- sample(c(1:dim(my_data_adultos)[1]), dim(my_data_adultos)[1]*0.7)
  train.df <- my_data_adultos[train.index, ]
  table<-table(train.df$vac.si.no.resp2)
  table

  valid.df <- my_data_adultos[-train.index, ]
  table(valid.df$vac.si.no.resp2)

  filas<-NA
  n1<-1
  for (j in 1:nrow(valid.df)) {
```

```
  if(train.df$vac.si.no.resp2[j]=="Si"){
    filas[n1]<-j
    n1<-n1+1
  }
}
s<-sample(filas,table[[1]], replace=TRUE)
train.bal.df<-rbind(train.df[train.df$vac.si.no.resp2=="No",], train.df[s,])
table(train.bal.df$vac.si.no.resp2)

datos.entrenamiento1<-train.bal.df
datos.validacion1<-valid.df

train.df <-adultos[train.index, ]

valid.df <- adultos[-train.index, ]

train.bal.df<-rbind(train.df[train.df$vac.si.no.resp2=="No",], train.df[s,])
table(train.bal.df$vac.si.no.resp2)

datos.entrenamiento2<-train.bal.df
datos.validacion2<-valid.df

#Regresion Logistica
logit.reg<-glm(factor(vac.si.no.resp2)~ .,
               data =datos.entrenamiento2[,c("vac.si.no.resp2",
               "enf_cer_Si","arterial_per_Si","diabetes_Si",
                 "enf_renal_Si","neuromusc_des_Si",
                 "neoplasia_Si","antihypertensivos_Si",
                 "antiplaq_agr_Si",
                 "hipolipemiantes_Si","insulina_Si",
                 "hipo_oral_Si",
                 "Inmunosup_Si","Cort_Si",
                 "Obes_No o medio (bmi<30)","Obes_Si (>=30 <40)",
                 "Fumar_Dejo > 1 a","Fumar_nunca",
                 "Fumar_Dejo < 1 a","Fumar_menos de 5 c/d",
                 "Clase_social_Cualificado parcialmente",
                 "Clase_social_Gerencial profesional",
                 "Clase_social_Ocupaciones gerenciales y tecnicas",
                 "Clase_social_Cualificado (NM)",
                 "Clase_social_Sin cualificacion",
                 "hosp_anyo_ant","ingre12_Si","Huevo_Si",
                 "vac_neumo_Si","Peso","Altura")],
               family = "binomial")
#Primer Modelo de Regresion logistica Stepwise
```

```
logit.reg_step<-step(logit.reg)
#Class tree por defecto
default.ct <- rpart(factor(vac.si.no.resp2) ~ .,
                data=datos.entrenamiento1[,c("vac.si.no.resp2",
                "enf_cer","arterial_per","diabetes",
                            "enf_renal","neuromusc_des",
                            "neoplasia",
                            "antihypertensivos","antiplaq_agr",
                            "hipolipemiantes","insulina",
                            "hipo_oral",
                            "Inmunosup","Cort",
                            "Obes",
                            "Fumar",
                            "Clase_social",
                            "hosp_anyo_ant","ingre12","Huevo",
                            "vac_neumo","Peso","Altura")],
                     method = "class")
#Class tree podado
tree_podado<-rpartXse(factor(vac.si.no.resp2)~ . , se = 0.25,
                      data=datos.entrenamiento1[,c("vac.si.no.resp2",
                      "enf_cer","arterial_per","diabetes",
                       "enf_renal","neuromusc_des","neoplasia",
                       "antihypertensivos","antiplaq_agr",
                       "hipolipemiantes","insulina","hipo_oral",
                       "Inmunosup","Cort",
                       "Obes",
                       "Fumar",
                       "Clase_social",
                       "hosp_anyo_ant","ingre12","Huevo",
                       "vac_neumo","Peso","Altura")],
                      model=TRUE)
#Random Forest
rf<-randomForest(factor(vac.si.no.resp2) ~ .,
                 data =datos.entrenamiento1[,c("vac.si.no.resp2",
                 "enf_cer","arterial_per","diabetes",
                   "enf_renal","neuromusc_des","neoplasia",
                   "antihypertensivos","antiplaq_agr",
                   "hipolipemiantes","insulina","hipo_oral",
                   "Inmunosup","Cort",
                   "Obes",
                   "Fumar",
                   "Clase_social",
                   "hosp_anyo_ant","ingre12","Huevo",
                   "vac_neumo","Peso","Altura")],
```

```
                  mtry=4, method="class",importance=TRUE)
#Reg. Log. Estadisticos
pred<-predict(logit.reg, datos.validacion2[,c("enf_cer_Si",
                  "arterial_per_Si","diabetes_Si",
                  "enf_renal_Si","neuromusc_des_Si",
                  "neoplasia_Si",
                  "antihypertensivos_Si","antiplaq_agr_Si",
                  "hipolipemiantes_Si","insulina_Si",
                  "hipo_oral_Si",
                  "Inmunosup_Si","Cort_Si",
                  "Obes_No o medio (bmi<30)","Obes_Si (>=30 <40)",
                  "Fumar_Dejo > 1 a","Fumar_nunca",
                  "Fumar_Dejo < 1 a","Fumar_menos de 5 c/d",
                  "Clase_social_Cualificado parcialmente",
                  "Clase_social_Gerencial profesional",
                  "Clase_social_Ocupaciones gerenciales y tecnicas",
                  "Clase_social_Cualificado (NM)",
                  "Clase_social_Sin cualificacion",
                  "hosp_anyo_ant","ingre12_Si","Huevo_Si",
                  "vac_neumo_Si","Peso","Altura")])
pred.reg_log <- prediction(pred,datos.validacion2$vac.si.no.resp2)
perf_dis.reg_log<-performance(pred.reg_log,"tpr","fpr")
perf.reg_log[i]<-performance(pred.reg_log,"auc")@y.values
n1<-nrow(datos.validacion2)
for(j in 1:n1){
  if(pred[j]<0.5){
    pred[j]<-0
  }
  else{pred[j]<-1}
}
tab <- table(pred, datos.validacion2$vac.si.no.resp2)
tasa.aciertos.reg_log[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.reg_log[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))*
sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))


#Performance-Reg. Log. Step Estadisticos
pred<-predict(logit.reg_step, datos.validacion2[,c("enf_cer_Si",
           "arterial_per_Si","diabetes_Si",
                           "enf_renal_Si","neuromusc_des_Si",
                           "neoplasia_Si",
                           "antihypertensivos_Si",
                           "antiplaq_agr_Si",
                           "hipolipemiantes_Si","insulina_Si",
```

```
                              "hipo_oral_Si",
                              "Inmunosup_Si","Cort_Si",
                              "Obes_No o medio (bmi<30)","Obes_Si (>=30 <40)",
                              "Fumar_Dejo > 1 a","Fumar_nunca",
                              "Fumar_Dejo < 1 a","Fumar_menos de 5 c/d",
                              "Clase_social_Cualificado parcialmente",
                              "Clase_social_Gerencial profesional",
                              "Clase_social_Ocupaciones gerenciales y tecnicas",
                              "Clase_social_Cualificado (NM)",
                              "Clase_social_Sin cualificacion",
                              "hosp_anyo_ant","ingre12_Si","Huevo_Si",
                              "vac_neumo_Si","Peso","Altura")])
pred.reg_log_step <- prediction(pred,datos.validacion2$vac.si.no.resp2)
perf_dis.reg_log_step<-performance(pred.reg_log_step,"tpr","fpr")
perf.reg_log_step[i]<-performance(pred.reg_log_step,"auc")@y.values
n1<-nrow(datos.validacion2)
for(j in 1:n1){
  if(pred[j]<0.5){
    pred[j]<-0
  }
  else{pred[j]<-1}
}
tab <- table(pred, datos.validacion2$vac.si.no.resp2)
tasa.aciertos.reg_log_step[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.reg_log_step[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt(tab[1,1]+tab[2,1])
*sqrt(tab[1,1]+tab[1,2])*sqrt(tab[2,2]+tab[2,1])*sqrt(tab[2,2]+tab[1,2]))


#Performance-Tree default Estadisticos
pred<-predict(default.ct, datos.validacion1[,c("enf_cer",
                                "arterial_per","diabetes",
                                "enf_renal","neuromusc_des",
                                "neoplasia",
                                "antihypertensivos",
                                "antiplaq_agr",
                                "hipolipemiantes","insulina",
                                "hipo_oral",
                                "Inmunosup","Cort",
                                "Obes",
                                "Fumar",
                                "Clase_social",
                                "hosp_anyo_ant","ingre12","Huevo",
                                "vac_neumo","Peso","Altura")])
pred.tree <- prediction(pred[,2],datos.validacion1$vac.si.no.resp2)
```

```
perf_dis.tree<-performance(pred.tree,"tpr","fpr")
perf.def_ct[i]<-performance(pred.tree,"auc")@y.values
n1<-nrow(datos.validacion1)
for(j in 1:n1){
  if(pred[j,1]<0.5){
    pred[j,1]<-"Si"
  }
  else{pred[j,1]<-"No"}
}
tab <- table(pred[,1], datos.validacion1$vac.si.no.resp2)
tasa.aciertos.def_ct[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.def_ct[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))
*sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))

#Performance-Tree podado Estadisticos
pred<-predict(tree_podado, datos.validacion1[,c("enf_cer",
                           "arterial_per","diabetes",
                           "enf_renal","neuromusc_des",
                           "neoplasia",
                           "antihypertensivos","antiplaq_agr",
                           "hipolipemiantes","insulina",
                           "hipo_oral",
                           "Inmunosup","Cort",
                           "Obes",
                           "Fumar",
                           "Clase_social",
                           "hosp_anyo_ant","ingre12","Huevo",
                           "vac_neumo","Peso","Altura")])
pred.tree_podado <- prediction(pred[,2],datos.validacion1$vac.si.no.resp2)
perf_dis.tree_podado<-performance(pred.tree_podado,"tpr","fpr")
perf.tree_podado[i]<-performance(pred.tree_podado,"auc")@y.values
n1<-nrow(datos.validacion1)
for(j in 1:n1){
  if(pred[j,1]<0.5){
    pred[j,1]<-"Yes"
  }
  else{pred[j,1]<-"No"}
}
tab <- table(pred[,1], datos.validacion1$vac.si.no.resp2)
tasa.aciertos.tree_podado[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.tree_podado[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))
*sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))

#Performance-rf Estadisticos
```

```
  pred<-predict(rf, datos.validacion1[,c("enf_cer",
                              "arterial_per","diabetes",
                              "enf_renal","neuromusc_des",
                              "neoplasia",
                              "antihypertensivos","antiplaq_agr",
                              "hipolipemiantes","insulina",
                              "hipo_oral",
                              "Inmunosup","Cort",
                              "Obes",
                              "Fumar",
                              "Clase_social",
                              "hosp_anyo_ant","ingre12","Huevo",
                              "vac_neumo","Peso","Altura")],type = "prob")
  pred.rf <- prediction(pred[,2],datos.validacion1$vac.si.no.resp2)
  perf_dis.rf<-performance(pred.rf,"tpr","fpr")
  perf.rf[i]<-performance(pred.rf,"auc")@y.values
  pred <- predict(rf, datos.validacion1[,c("enf_cer",
                                "arterial_per","diabetes",
                                "enf_renal","neuromusc_des",
                                "neoplasia",
                                "antihypertensivos","antiplaq_agr",
                                "hipolipemiantes","insulina",
                                "hipo_oral",
                                "Inmunosup","Cort",
                                "Obes",
                                "Fumar",
                                "Clase_social",
                                "hosp_anyo_ant","ingre12","Huevo",
                                "vac_neumo","Peso","Altura")])
  tab <- table(pred, datos.validacion1$vac.si.no.resp2)
  tasa.aciertos.rf[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
  mcc.rf[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))
  *sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))
}
#Tasas de acierto y graficos
tasa_acierto<-cbind(tasa.aciertos.reg_log,
                    tasa.aciertos.reg_log_step,
                    tasa.aciertos.def_ct,
                    tasa.aciertos.tree_podado,
                    tasa.aciertos.rf)
tasa_acierto
write.table(tasa_acierto, file = "tasas_acierto_Hold_out_adultos",
            append = FALSE, quote = TRUE, sep = " ", row.names = TRUE,
            col.names = TRUE)
```

```
mcc<-cbind(mcc.reg_log,
           mcc.reg_log_step,
           mcc.def_ct,
           mcc.tree_podado,
           mcc.rf)
mcc
write.table(mcc, file = "mcc_Hold_out_adultos", append = FALSE,
           quote = TRUE, sep = " ", row.names = TRUE, col.names = TRUE)


perf<-cbind(perf.reg_log,
            perf.reg_log_step,
            perf.def_ct,
            perf.tree_podado,
            perf.rf)
perf
write.table(perf, file = "perf_Hold_out_adultos", append = FALSE,
            quote = TRUE, sep = " ", row.names = TRUE, col.names = TRUE)
```

Regresión Logística, Árboles de Regresión, Random Forest y comparación de métodos en población de mayores de 65 años

```
library("readxl")
library("fastDummies")
library(MASS)
library(rpart)
library(DMwR2)
library(car)
library(randomForest)
library(rpart.plot)
set.seed(1234)

my_data_mayores <- read_excel("TFM_Nico_UPV_con_etiquetas_mayores_65.xlsx")
n<-nrow(my_data_mayores)
for(i in 1:n){
  if(my_data_mayores$Cort[i]!="No"){
    my_data_mayores$Cort[i]<-"Si"
  }
}
my_data_mayores<-my_data_mayores[-481,]
n1<-1
h<-0
for (i in 1:n) {
  if(is.na(my_data_mayores$hosp_anyo_ant[i])){
    h[n1]<-i
    n1<-n1+1
  }
}
my_data_mayores<-my_data_mayores[-h,]

##########################Analisis Mayores##########################
mayores<-dummy_cols(
  my_data_mayores,
  select_columns = NULL,
  remove_first_dummy = TRUE
)
mayores<-mayores[,-c(1:27,30:38,40:46)]
mayores<-mayores[,-c(5:11,16,17,57:65)]

#Regresion logistica
logit.reg_mayores<-glm(factor(vac.si.no.resp2)~ .,
                       data =mayores[,c("vac.si.no.resp2","enf_cor_Si",
```

```
                                        "otras_endocrinas_Si","anemia_Si",
                                        "enf_renal_Si",
                                        "demencia_Si",
                                        "antihypertensivos_Si",
                                        "anticoagulantes_Si",
                                        "insulina_Si","Cort_Si",
                                        "Obes_No o medio (bmi<30)",
                                        "Obes_si (>=30 <40)",
                                        "Fumar_menos de 5 c/d",
                                        "Fumar_nunca",
                                        "Fumar_Dejo < 1 a",
                                        "Fumar_Dejo > 1 a",
                                        "hosp_anyo_ant","ingre12_Si",
                                        "vac_neumo_Si","Peso",
                                        "Altura")],
                        family = "binomial")
summary(logit.reg_mayores)
vif(logit.reg_mayores)
logit.reg_mayores_null<-glm(factor(vac.si.no.resp2)~1,
                            data =mayores[,c("vac.si.no.resp2","enf_cor_Si",
                                     "otras_endocrinas_Si","anemia_Si",
                                      "enf_renal_Si",
                                      "demencia_Si","antihypertensivos_Si",
                                      "anticoagulantes_Si",
                                      "insulina_Si","Cort_Si",
                                       "Obes_No o medio (bmi<30)",
                                       "Obes_si (>=30 <40)",
                                       "Fumar_menos de 5 c/d",
                                       "Fumar_nunca",
                                       "Fumar_Dejo < 1 a",
                                       "Fumar_Dejo > 1 a",
                                       "hosp_anyo_ant","ingre12_Si",
                                       "vac_neumo_Si","Peso",
                                       "Altura")],
                            family = "binomial")
1-logLik(logit.reg_mayores)/logLik(logit.reg_mayores_null)

#Regresion logistica Stepwise
logit.reg_step_mayores<-step(logit.reg_mayores)
summary(logit.reg_step_mayores)
logit.reg_step_mayores_null<-glm(factor(vac.si.no.resp2)~1,
                                 data =mayores[,c("vac.si.no.resp2",
                                     "otras_endocrinas_Si","anemia_Si",
                                     "demencia_Si","antihypertensivos_Si",
```

```
                                        "anticoagulantes_Si",
                                        "insulina_Si","Cort_Si",
                                        "Obes_No o medio (bmi<30)",
                                        "Obes_si (>=30 <40)",
                                        "Fumar_nunca",
                                         "Fumar_Dejo < 1 a",
                                         "Fumar_Dejo > 1 a",
                                          "hosp_anyo_ant",
                                          "vac_neumo_Si")],
                                    family = "binomial")
1-logLik(logit.reg_step_mayores)/logLik(logit.reg_step_mayores_null)

#Class tree por defecto
default.ct_mayores<- rpart(factor(vac.si.no.resp2) ~ .,
                           data=my_data_mayores[,c("vac.si.no.resp2",
                           "enf_cor","otras_endocrinas","anemia",
                                                   "enf_renal",
                                                   "demencia","antihypertensivos",
                                                   "anticoagulantes",
                                                   "insulina","Cort",
                                                   "Obes",
                                                   "Fumar",
                                                   "hosp_anyo_ant","ingre12",
                                                   "vac_neumo","Peso",
                                                   "Altura")],
                            method = "class")
X11()
prp(default.ct_mayores, extra=101)

#Class tree completo
treefull_mayores <- rpart(factor(vac.si.no.resp2)~ .,
                          data=my_data_mayores[,c("vac.si.no.resp2","enf_cor",
                                                  "otras_endocrinas","anemia",
                                                  "enf_renal",
                                                  "demencia","antihypertensivos",
                                                  "anticoagulantes",
                                                  "insulina","Cort",
                                                  "Obes",
                                                  "Fumar",
                                                  "hosp_anyo_ant","ingre12",
                                                  "vac_neumo","Peso",
                                                  "Altura")],
                          method="class", cp=0.001)
X11()
```

```
prp(treefull_mayores, extra = 101)

#Class Tree podado
tree_mayores<-rpartXse(factor(vac.si.no.resp2)~ . , se = 0.25,
                       data=my_data_mayores[,c("vac.si.no.resp2","enf_cor",
                                               "otras_endocrinas","anemia",
                                               "enf_renal",
                                               "demencia","antihypertensivos",
                                               "anticoagulantes",
                                               "insulina","Cort",
                                               "Obes",
                                               "Fumar",
                                               "hosp_anyo_ant","ingre12",
                                               "vac_neumo","Peso",
                                               "Altura")],
                       model=TRUE)
X11()
prp(tree_mayores, extra = 101)

#Random Forest
rf_mayores<-randomForest(factor(vac.si.no.resp2) ~ .,
                         data =my_data_mayores[,c("vac.si.no.resp2","enf_cor",
                                                  "otras_endocrinas","anemia",
                                                  "enf_renal",
                                                  "demencia","antihypertensivos",
                                                  "anticoagulantes",
                                                  "insulina","Cort",
                                                  "Obes",
                                                  "Fumar",
                                                  "hosp_anyo_ant","ingre12",
                                                  "vac_neumo","Peso",
                                                  "Altura")],
                         mtry=4, method="class",importance=TRUE)
X11()
varImpPlot(rf_mayores, main="",col="dark blue")

#######Hold out Oversampling (CURVAS ROC, MCC, TASA ACIERTO)#######
perf.reg_log<-NA
tasa.aciertos.reg_log<-NA
mcc.reg_log<-NA

perf.reg_log_step<-NA
tasa.aciertos.reg_log_step<-NA
mcc.reg_log_step<-NA
```

```
perf.treefull<-NA
tasa.aciertos.treefull<-NA
mcc.treefull<-NA

perf.rf<-NA
tasa.aciertos.rf<-NA
mcc.rf<-NA
for (i in 1:100){
  train.index <- sample(c(1:dim(my_data_mayores)[1]), dim(my_data_mayores)[1]*0.7)
  train.df <- my_data_mayores[train.index, ]
  table<-table(train.df$vac.si.no.resp2)
  table

  valid.df <- my_data_mayores[-train.index, ]
  table(valid.df$vac.si.no.resp2)

  filas<-NA
  n1<-1
  for (j in 1:nrow(valid.df)) {
    if(train.df$vac.si.no.resp2[j]=="Si"){
      filas[n1]<-j
      n1<-n1+1
    }
  }
  s<-sample(filas,table[[1]], replace=TRUE)
  train.bal.df<-rbind(train.df[train.df$vac.si.no.resp2=="No",], train.df[s,])
  table(train.bal.df$vac.si.no.resp2)

  datos.entrenamiento1<-train.bal.df
  datos.validacion1<-valid.df

  train.df <-mayores[train.index, ]

  valid.df <- mayores[-train.index, ]

  train.bal.df<-rbind(train.df[train.df$vac.si.no.resp2=="No",], train.df[s,])
  table(train.bal.df$vac.si.no.resp2)

  datos.entrenamiento2<-train.bal.df
  datos.validacion2<-valid.df

  #Regresion Logistica
  logit.reg<-glm(factor(vac.si.no.resp2)~ .,
```

```
                data =datos.entrenamiento2[,c("vac.si.no.resp2",
                "enf_cor_Si","otras_endocrinas_Si","anemia_Si",
                "enf_renal_Si","demencia_Si","antihypertensivos_Si",
                "anticoagulantes_Si","insulina_Si","Cort_Si",
                "Obes_No o medio (bmi<30)","Obes_si (>=30 <40)",
                "Fumar_menos de 5 c/d","Fumar_nunca",
                "Fumar_Dejo < 1 a","Fumar_Dejo > 1 a",
                "hosp_anyo_ant","ingre12_Si","vac_neumo_Si","Peso",
                "Altura")], family = "binomial")
#Primer Modelo de Regresion logistica Stepwise
logit.reg_step<-step(logit.reg)
#Class tree completo 1
treefull<- rpart(factor(vac.si.no.resp2)~ .,
                 data=datos.entrenamiento1[,c("vac.si.no.resp2","enf_cor",
                                              "otras_endocrinas","anemia",
                                              "enf_renal",
                                              "demencia","antihypertensivos",
                                              "anticoagulantes",
                                              "insulina","Cort",
                                              "Obes",
                                              "Fumar",
                                              "hosp_anyo_ant","ingre12",
                                              "vac_neumo","Peso",
                                              "Altura")],
                 method="class", cp=0.001)
#Random Forest
rf<-randomForest(factor(vac.si.no.resp2) ~ .,
                        data =datos.entrenamiento1[,c("vac.si.no.resp2",
                        "enf_cor","otras_endocrinas","anemia",
                        "enf_renal","demencia","antihypertensivos",
                        "anticoagulantes","insulina","Cort",
                        "Obes","Fumar","hosp_anyo_ant","ingre12",
                        "vac_neumo","Peso","Altura")],
                        mtry=4, method="class",importance=TRUE)
#Performance-Reg. Log.
pred<-predict(logit.reg, datos.validacion2[,c("enf_cor_Si",
                                              "otras_endocrinas_Si","anemia_Si",
                                              "enf_renal_Si",
                                             "demencia_Si",
                                             "antihypertensivos_Si",
                                              "anticoagulantes_Si",
                                              "insulina_Si","Cort_Si",
                                              "Obes_No o medio (bmi<30)",
                                              "Obes_si (>=30 <40)",
```

```
                                                    "Fumar_menos de 5 c/d",
                                                    "Fumar_nunca",
                                                    "Fumar_Dejo < 1 a",
                                                    "Fumar_Dejo > 1 a",
                                                    "hosp_anyo_ant","ingre12_Si",
                                                    "vac_neumo_Si","Peso",
                                                    "Altura")])
pred.reg_log <- prediction(pred,datos.validacion2$vac.si.no.resp2)
perf_dis.reg_log<-performance(pred.reg_log,"tpr","fpr")
perf.reg_log[i]<-performance(pred.reg_log,"auc")@y.values
n1<-nrow(datos.validacion1)
for(j in 1:n1){
  if(pred[j]<0.5){
    pred[j]<-0
  }
  else{pred[j]<-1}
}
tab <- table(pred, datos.validacion2$vac.si.no.resp2)
tasa.aciertos.reg_log[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.reg_log[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))
*sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))

#Performance-Reg. Log. Step
pred<-predict(logit.reg_step, datos.validacion2[,c("enf_cor_Si",
                                                    "otras_endocrinas_Si",
                                                    "anemia_Si",
                                                    "enf_renal_Si",
                                                    "demencia_Si",
                                                   "antihypertensivos_Si",
                                                    "anticoagulantes_Si",
                                                    "insulina_Si",
                                                    "Cort_Si",
                                                    "Obes_No o medio (bmi<30)",
                                                    "Obes_si (>=30 <40)",
                                                    "Fumar_menos de 5 c/d",
                                                    "Fumar_nunca",
                                                    "Fumar_Dejo < 1 a",
                                                    "Fumar_Dejo > 1 a",
                                                    "hosp_anyo_ant",
                                                    "ingre12_Si",
                                                    "vac_neumo_Si","Peso",
                                                    "Altura")])
pred.reg_log_step <- prediction(pred,datos.validacion2$vac.si.no.resp2)
perf_dis.reg_log_step<-performance(pred.reg_log_step,"tpr","fpr")
```

```
perf.reg_log_step[i]<-performance(pred.reg_log_step,"auc")@y.values
n1<-nrow(datos.validacion1)
for(j in 1:n1){
  if(pred[j]<0.5){
    pred[j]<-0
  }
  else{pred[j]<-1}
}
tab <- table(round(pred), datos.validacion2$vac.si.no.resp2)
tasa.aciertos.reg_log_step[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.reg_log_step[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+
+tab[2,1]))*sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+
+tab[1,2])))

#performance-Tree completo
pred<-predict(treefull, datos.validacion1[,c("enf_cor",
                                             "otras_endocrinas","anemia",
                                             "enf_renal",
                                             "demencia","antihypertensivos",
                                             "anticoagulantes",
                                             "insulina","Cort",
                                             "Obes",
                                             "Fumar",
                                             "hosp_anyo_ant","ingre12",
                                             "vac_neumo","Peso",
                                             "Altura")])
pred.treefull <- prediction(pred[,2],datos.validacion1$vac.si.no.resp2)
perf_dis.treefull<-performance(pred.treefull,"tpr","fpr")
perf.treefull[i]<-performance(pred.treefull,"auc")@y.values

n1<-nrow(datos.validacion1)
for(j in 1:n1){
  if(pred[j,1]<0.5){
    pred[j,1]<-"Yes"
  }
  else{pred[j,1]<-"No"}
}
tab <- table(pred[,1], datos.validacion1$vac.si.no.resp2)
tasa.aciertos.treefull[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
mcc.treefull[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))*
sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))

#Performance-rf
pred<-predict(rf, datos.validacion1[,c("enf_cor",
```

```
                                          "otras_endocrinas","anemia",
                                          "enf_renal",
                                          "demencia","antihypertensivos",
                                          "anticoagulantes",
                                          "insulina","Cort",
                                          "Obes",
                                          "Fumar",
                                          "hosp_anyo_ant","ingre12",
                                          "vac_neumo","Peso",
                                          "Altura")],type = "prob")
  pred.rf <- prediction(pred[,2],datos.validacion1$vac.si.no.resp2)
  perf_dis.rf<-performance(pred.rf,"tpr","fpr")
  perf.rf[i]<-performance(pred.rf,"auc")@y.values

  pred <- predict(rf, datos.validacion1[,c("enf_cor",
                                           "otras_endocrinas","anemia",
                                           "enf_renal",
                                           "demencia","antihypertensivos",
                                           "anticoagulantes",
                                           "insulina","Cort",
                                           "Obes",
                                           "Fumar",
                                           "hosp_anyo_ant","ingre12",
                                           "vac_neumo","Peso",
                                           "Altura")])
  tab <- table(pred, datos.validacion1$vac.si.no.resp2)
  tasa.aciertos.rf[i]<-sum(tab[row(tab)==col(tab)])/sum(tab)
  mcc.rf[i]<-(tab[1,1]*tab[2,2]-tab[1,2]*tab[2,1])/(sqrt((tab[1,1]+tab[2,1]))*
  sqrt((tab[1,1]+tab[1,2]))*sqrt((tab[2,2]+tab[2,1]))*sqrt((tab[2,2]+tab[1,2])))
}
#Tasas de acierto y graficos
tasa_acierto<-cbind(tasa.aciertos.reg_log,
                    tasa.aciertos.reg_log_step,
                    tasa.aciertos.treefull,
                    tasa.aciertos.rf)
tasa_acierto
write.table(tasa_acierto, file = "tasas_acierto_Hold_out_mayores_65",
            append = FALSE, quote = TRUE, sep = " ",
            row.names = TRUE, col.names = TRUE)

mcc<-cbind(mcc.reg_log,
           mcc.reg_log_step,
           mcc.treefull,
           mcc.rf)
```

```
mcc
write.table(mcc, file = "mcc_Hold_out_mayores_65",
            append = FALSE, quote = TRUE, sep = " ", row.names = TRUE,
            col.names = TRUE)

perf<-cbind(perf.reg_log,
            perf.reg_log_step,
            perf.treefull,
            perf.rf)
perf
write.table(perf, file = "perf_Hold_out_mayores_65",
            append = FALSE, quote = TRUE, sep = " ", row.names = TRUE,
            col.names = TRUE)
```

Análisis de Correspondencias Múltiples en población de adultos

```
library("FactoMineR")
library("factoextra")
library(readxl)

my_data_adultos <- read_excel("TFM_Nico_UPV_con_etiquetas_adultos.xlsx")
n<-nrow(my_data_adultos)
for(i in 1:n){
  if(my_data_adultos$Cort[i]!="No"){
    my_data_adultos$Cort[i]<-"Si"
  }
}
attach(my_data_adultos)
for (i in 1:nrow(my_data_adultos)){
  if(is.na(my_data_adultos$hosp_anyo_ant[i])){
  }
  else{
    if(my_data_adultos$hosp_anyo_ant[i]<1){
      my_data_adultos$hosp_anyo_ant[i]<-"Ninguna Hospitalizacion"
    }else if(my_data_adultos$hosp_anyo_ant[i]<=3
             &&my_data_adultos$hosp_anyo_ant[i]>=1)
    {
      my_data_adultos$hosp_anyo_ant[i]<-"Entre 1 y 3 hospitalizaciones"
    }else{
      my_data_adultos$hosp_anyo_ant[i]<-"Mas de 3 hospitalizaciones"
    }
  }
}

Parametros <- na.omit(my_data_adultos[,c("enf_cer","arterial_per",
                                         "diabetes",
                                         "enf_renal","neuromusc_des",
                                         "neoplasia",
                                         "demencia","antihypertensivos",
                                         "antiplaq_agr","hipolipemiantes",
                                         "insulina",
                                         "hipo_oral",
                                         "Inmunosup",
                                         "Cort",
                                         "Obes",
                                         "Fumar","Clase_social",
```

```
                                          "hosp_anyo_ant",
                                          "ingre12","Huevo",
                                          "vac_neumo")])
Corresp <- MCA(Parametros, ncp = 2, graph = TRUE)
summary(Corresp)

# Extraer los autovalores en cada dimension
get_eigenvalue(Corresp)
# Visualizacion de autovalores
fviz_eig(Corresp)

# Proyeccion de la variable de Vacunacion en suplementario
Parametros_Sup <- na.omit(my_data_adultos[,c("vac.si.no.resp2","sexo",
                                             "Clase_social",
                                             "enf_cer",
                                             "arterial_per",
                                             "diabetes",
                                             "enf_renal",
                                             "neuromusc_des",
                                             "neoplasia",
                                             "demencia",
                                             "antihypertensivos",
                                             "antiplaq_agr",
                                             "hipolipemiantes",
                                             "insulina",
                                             "hipo_oral",
                                             "Inmunosup",
                                             "Cort",
                                             "Obes","Fumar",
                                             "hosp_anyo_ant",
                                             "ingre12","Huevo",
                                             "vac_neumo")])
Corresp_Sup <- MCA(Parametros_Sup,quali.sup = c(1,2,3), ncp=2, graph=TRUE)
summary(Corresp_Sup)
```

Análisis de Correspondencias Múltiples en población de mayores de 65 años

```
library("FactoMineR")
library("factoextra")
library(readxl)

my_data_mayores <- read_excel("TFM_Nico_UPV_con_etiquetas_mayores_65.xlsx")
n<-nrow(my_data_mayores)
for(i in 1:n){
  if(my_data_mayores$Cort[i]!="No"){
    my_data_mayores$Cort[i]<-"Si"
  }
}
attach(my_data_mayores)
for (i in 1:nrow(my_data_mayores)){
  if(is.na(my_data_mayores$hosp_anyo_ant[i])){
  }
  else{
    if(my_data_mayores$hosp_anyo_ant[i]<1){
      my_data_mayores$hosp_anyo_ant[i]<-"Ninguna Hospitalizacion"
    }else if(my_data_mayores$hosp_anyo_ant[i]<=3
             &&my_data_mayores$hosp_anyo_ant[i]>=1)
    {
      my_data_mayores$hosp_anyo_ant[i]<-"Entre 1 y 3 hospitalizaciones"
    }else{
      my_data_mayores$hosp_anyo_ant[i]<-"Mas de 3 hospitalizaciones"
    }
  }
}

Parametros <- na.omit(my_data_mayores[,c("enf_cor",
                                         "otras_endocrinas","anemia","enf_renal",
                                         "demencia","antihypertensivos",
                                         "anticoagulantes","insulina","Cort",
                                         "Obes","Fumar",
                                         "hosp_anyo_ant","ingre12","vac_neumo")])
Corresp <- MCA(Parametros, ncp = 2, graph = TRUE)
summary(Corresp)

# Extraer los autovalores en cada dimension
get_eigenvalue(Corresp)
# Visualizacion de autovalores
fviz_eig(Corresp)
```

```
# Proyeccion de la variable de Vacunacion en suplementario
Parametros_Sup <- na.omit(my_data_mayores[,c("vac.si.no.resp2","sexo",
                                             "Clase_social",
                                             "enf_cor","otras_endocrinas",
                                             "anemia","enf_renal",
                                             "demencia",
                                             "antihypertensivos",
                                             "anticoagulantes",
                                             "insulina","Cort",
                                             "Obes","Fumar",
                                             "hosp_anyo_ant",
                                             "ingre12","vac_neumo")])
Corresp_Sup <- MCA(Parametros_Sup, quali.sup = c(1,2,3), ncp=2, graph=TRUE)
summary(Corresp_Sup)
```