Universitat Oberta de Catalunya

uoc.edu

Análisis y modelización estadística del control de la diabetes mellitus tipo II

Esteban Hernández Maldonado
Máster en Bioinformática y Bioestadística
Análisis de datos y técnicas de clustering

Dr. Daniel Fernández Martínez
Dr. Marc Maceira Duch

junio de 2021

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU Free Documentation License".

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis y modelización estadística del control de la diabetes mellitus tipo II*
Nombre del autor:	*Esteban Hernández Maldonado*
Nombre del consultor/a:	*Dr. Daniel Fernández Martínez*
Nombre del PRA:	*Dr. Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	*06/2021*
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Análisis de datos y técnicas de clustering*
Idioma del trabajo:	*Español*
Palabras clave	*Hemoglobina glicosilada, modelo de regresión, análisis multivariante*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

La diabetes mellitus tipo II es una enfermedad que puede ser controlada en ciertas ocasiones sin medicación si el paciente lleva unos hábitos saludables. Estos hábitos pueden ser medidos con el cuestionario DMSES.

En este trabajo se ha querido demostrar que llevando la dieta adecuada y realizando una actividad física adecuada los niveles de hemoglobina glicosilada pueden mantenerse en un nivel bajo.

Primero se ha realizado un análisis exploratorio de las variables utilizadas, para después proponer un modelo lineal. Por la naturaleza de los datos principales (preguntas de un cuestionario) no se consigue un modelo que explique la mayor parte de la variación de los datos. En cambio, estos cuestionarios sí que sirven para discriminar entre una diabetes bien o mal controlada por el paciente, como se verá aplicando los datos en regresión logística y análisis de conglomerados.

Con un análisis de componentes principales se podrá comprobar la relación inversa entre el nivel de hemoglobina glicosilada y las puntuaciones del cuestionario DMSES.

Abstract (in English, 250 words or less):

Type II diabetes mellitus is a disease that sometimes can be controlled without medication if the patient follows healthy habits. These habits can be measured with the DMSES questionnaire.

The goal of this Master's thesis was to demonstrate that by following the correct diet and doing adequate physical activity, the levels of glycosylated hemoglobin can be kept at a low level.

First, an exploratory analysis of the variables was done, and then a linear model was proposed. Due to the nature of the main data (questions in a questionnaire), a model that explains most of the variation in the data is not achieved. On the other hand, these questionnaires do serve to discriminate between diabetes that is good or bad controlled by the patient, as will be seen by applying the data in logistic regression and cluster analysis.

With a principal component analysis, the inverse relationship between the level of glycosylated hemoglobin and the scores of the DMSES questionnaire can be verified.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

La diabetes mellitus es, con diferencia, el más común de todos los trastornos endocrinos (trastornos hormonales) [1]. La enfermedad ha sido descrita como un estado de inanición en medio de la abundancia. Aunque hay mucho azúcar en la sangre, sin la acción adecuada de la insulina, el azúcar no llega a las células que la necesitan para obtener energía. La glucosa, la forma más simple de azúcar, es la principal fuente de energía para muchas funciones vitales. Privadas de glucosa, las células mueren de hambre y los tejidos comienzan a degenerarse. La glucosa no utilizada se acumula en el torrente sanguíneo, lo que conduce a una serie de complicaciones secundarias.

En los últimos años han aparecido aparatos que miden frecuentemente el nivel de azúcar en sangre de los pacientes con diabetes [2]. Existen además aparatos que aportan automáticamente la insulina necesaria para contrarrestar subidas de azúcar cuando éstas son detectadas [3]. La tecnología para almacenar y/o transmitir los datos obtenidos también existe por lo que técnicamente se disponen de todas las herramientas para controlar médicamente la enfermedad y de llevar un histórico de los niveles de azúcar de un paciente para, posteriormente, analizar clínicamente su caso.

Aunque la tecnología existe, ésta no se aplica a toda la población, o bien porque no todos los casos de diabetes necesitan de un seguimiento exhaustivo del azúcar en sangre, o bien porque los sistemas sanitarios no tienen los recursos necesarios, al menos para abarcar a todos los enfermos de diabetes.

Por otro lado, si el paciente lleva una vida saludable practicando ejercicio moderado y llevando una dieta adecuada, éste podrá llevar una vida normal y no desencadenará ninguna de las enfermedades crónicas asociadas (enfermedad cardiovascular, nefropatía, retinopatía y neuropatía) [4].

Para poder obtener información de cómo un enfermo de diabetes controla su enfermedad se puede recurrir al cuestionario DMSES (Diabetes Management Self-Efficacy Scale). Se trata de un cuestionario de 20 preguntas distribuidas en 4 grupos (hábitos alimenticios, actividad física, capacidad de medirse el nivel de azúcar y control médico) creado inicialmente para los pacientes de diabetes de Países Bajos y Estados Unidos [5] y que ya dispone de versiones locales para países como Italia, Grecia y Corea entre otros [6] [7] [8].

Las versiones locales pueden suponer desde pequeños cambios en las preguntas para adaptarlas a los diferentes idiomas hasta eliminar algunas de ellas o agruparlas de manera distinta. También la escala de las respuestas puede variar entre las diferentes versiones: de una escala de 1 a 5 a otra de 1 a 10. Siendo siempre el valor más bajo el que corresponde a “Totalmente en desacuerdo” y el más alto el que corresponde a “Totalmente de acuerdo”.

De manera adicional se dispone de las puntuaciones estandarizadas y centradas del total del cuestionario y de cada uno de los 4 grupos de preguntas.

La hemoglobina glicosilada en su versión más estable (HBA1c) es útil para medir clínicamente el nivel de azúcar de hasta los últimos dos meses [9]. Niveles por debajo de 5.7% son considerados normales, entre 5.7% y 6.4% son considerados como prediabetes y por encima de 6.5% son considerados como diabetes [10]. Cuanto más alto sea el porcentaje datos por la hemoglobina glicosilada peor estaría siendo gestionada la enfermedad por parte del paciente. Ésta será por tanto la variable que sirve para discriminar entre una diabetes bien o mal llevada por parte del paciente.

La meta de este trabajo será averiguar si la puntuación total del cuestionario DMSES (absoluta o estandarizada) o la de alguna de sus grupos puede explicar el nivel en sangre de la hemoglobina glicosilada y por tanto si se puede estimar que bien, o mal, está siendo gestionada la enfermedad por el paciente. Se considerará el uso de otras variables sociodemográficas, de laboratorio como el colesterol [11] y triglicéridos [12], así como la presencia de otras enfermedades.

Los datos del estudio se extraen de un fichero con información sobre 700 pacientes con diabetes mellitus tipo II de Tailandia. La lista de variables que contiene el fichero incluye, sin ser exhaustiva, las respuestas a las 20 preguntas de la versión tailandesa del cuestionario DMSES (T-DMSES), la puntuación estandarizada total y la de cada uno de los grupos de preguntas (hábitos alimenticios, actividad física, capacidad de medirse el nivel de azúcar y control médico), variables sociodemográficas (edad, sexo, peso, altura, religión, etc...), variables de laboratorio (hemoglobina glicosilada, colesterol) y presencia de otras enfermedades

1.2 Objetivos del Trabajo

Objetivos generales:

1. Generar un modelo donde la puntuación global del cuestionario DMSES o la puntuación de alguno de sus cuatro factores sirva para explicar el valor de la variable HBA1c, con ayuda o no de otros predictores.

2. Realizar un análisis multivariante de datos para detectar las variables posibles relaciones entre las variables.

Objetivos específicos:

1. comprobar si es posible crear un modelo de regresión que explique la hemoglobina glicosilada con alguna puntuación del cuestionario DMSES (absoluta/estandarizada, total/parcial) como predictor.

2. crear un modelo de regresión que ayude a discriminar entre diabetes controlada o no mediante la puntuación total o de algún factor del cuestionario DMSES y que mejore los ya existentes y que pueden ser hallado en la literatura accesible en la red.

3. Análisis multivariante de datos.

    - Hacer un análisis exploratorio multivariante de datos, buscando las correlaciones entre las variables o las distancias entre las observaciones.

    - Buscar que variables explican una mayor variabilidad de los datos utilizando el análisis de componentes principales.

    - Buscar posibles agrupaciones entre los pacientes.

1.3 Enfoque y método seguido

El método seguido se detalla en los siguientes puntos.

1.1 Exploración unidimensional de los datos.

Detectar la naturaleza de las variables (numérica discreta, numérica continua, carácter, factor...). Estudiar el rango de valores, media, percentiles y mediana para variables continuas y tabla de frecuencias para factores. Utilizar densidades o histogramas para la visualización de los datos.

1.2 Creación de variables derivadas.

El BMI da más información sobre la condición física de una persona que simplemente el peso. Esta variable no existe y puede formar parte en un futuro de la lista de variables predictoras.

Las puntuaciones totales y de cada uno de los dos factores se hallan disponibles en la base de datos y están estandarizadas. Puede que no sea de más utilidad tenerlas también calculadas de otras maneras.

1.3 Descarte de variables

Debido a la gran cantidad de variables que tiene el fichero de datos, un descarte de variables es necesario. Al trabajar con las puntuaciones totales o de alguno de los 4 factores del cuestionario puede que ya no sea necesario guardar o tener en cuenta las respuestas individuales.

Las variables factor con poca variabilidad serán estudiadas para valorar si se puede prescindir de ellas.

1.4 Creación del modelo y ajuste.

Crear un modelo lineal o transformable en lineal mediante el uso de logaritmos, raíces o potencias en los predictores o variable de respuesta. La variable respuesta será la hemoglobina glicosilada (HBA1c) y en la lista de predictores estaría alguna de las puntuaciones estandarizadas del cuestionario DMSES.

1.5 Exploración multidimensional de los datos

Mediante gráficos y correlaciones. Debido a la gran cantidad de variables se valorará particionar en análisis.

1.6 Análisis de componentes principales

Entre las variables HBA1c, puntuaciones estandarizadas del cuestionario DMSES y el resto de las variables continuas de los datos disponibles.

1.7 Análisis de conglomerados

Se utilizará K-medoids por ser más robusto y resistente a outliers, pero no se descarta el uso también de K-means. En caso de aplicar ambos se analizarán los resultados, ventajas y desventajas.

1.4 Planificación del Trabajo

Tareas

Las tareas principales que se desempeñarán a lo largo de este trabajo se pueden consultar en la siguiente lista.

Tarea		Tiempo (días)	Fecha límite
ID	Nombre		
1	**Búsqueda bibliográfica**	60	18.04.2021
2	**PEC 0: Definición de los contenidos de trabajo**	8	01.03.2021
3	**PEC1: Plan de trabajo**	14	16.03.2021
4	**Estado del arte**	3	20.03.2021
5	**Selección y creación de variables de interés**	2	23.03.2021
6	**Exploración unidimensional de datos**	1	26.03.2021
7	**Exploración multidimensional de datos**	5	28.03.2021
8	**Planteamiento del modelo de regresión**	5	02.04.2021
9	**Ajuste del modelo**	10	12.04.2021
10	**PEC 2: Desarrollo del trabajo - Fase 1**	14	19.04.2021
11	**Análisis de componentes principales**	5	26.04.2021
12	**Análisis de conglomerados**	10	05.05.2021
13	**PEC 3: Desarrollo del trabajo - Fase 2**	14	17.05.2021
14	**Conclusiones y resultados**	7	24.05.2021
15	**PEC4: Cierre de la memoria**	14	08.06.2021
16	**Elaboración de la presentación (PEC 5a)**	5	13.06.2021
17	**Preparación de la defensa pública (PEC5b)**	7	23.06.2021

Tabla 1: Tareas

Calendario

La lista de tareas se muestra a continuación de forma visual en el siguiente calendario

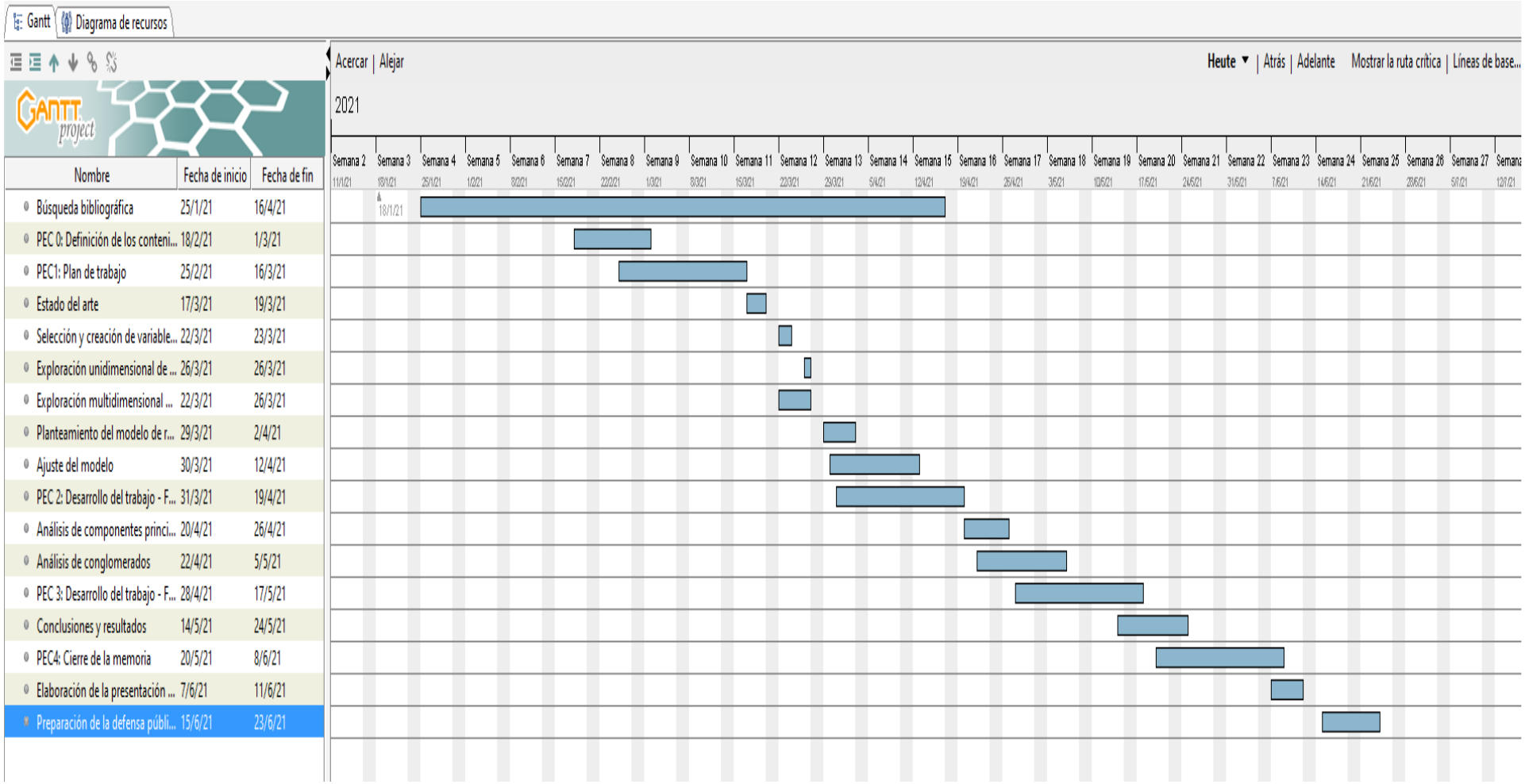


Figura 1: Calendario de tareas

1.5 Breve sumario de productos obtenidos

En esta memoria se espera entregar los siguientes productos:

- un modelo lineal que explique la concentración de la hemoglobina glicosilada en sangre a partir de una lista de predictores. Al menos una puntuación del cuestionario DMSES, la total o la de alguna de sus cuatro partes deberá hallarse en la lista de predictores.

- un modelo de regresión logística que discrimine entre pacientes según su hemoglobina glicosilada es mayor o igual que el 7% o no. Al menos una puntuación del cuestionario DMSES, la total o la de alguna de sus cuatro partes deberá hallarse en la lista de predictores.

- un estudio de análisis de componentes principales para buscar relaciones entre el cuestionario DMSES y la presencia de hemoglobina glicosilada en sangre.

- un análisis de conglomerados del conjunto de datos.

- un análisis discriminante lineal sobre el nivel de hemoglobina glicosilada superior o inferior al 7% utilizando las puntuaciones del cuestionario DMSES.

1.6 Breve descripción de los otros capítulos de la memoria

En el capítulo 2 se explicará qué datos serán utilizados, el número de registros y el listado de variables. El motivo por el que se descartan algunas variables y la razón por la que se crearán otras variables derivadas de las ya existentes. Se realizará un análisis descriptivo y se estudiará la presencia de datos faltantes y de outliers.

En el capítulo 3 se planteará un modelo lineal que explique el valor de la hemoglobina glicosilada a partir de, al menos, la puntuación del cuestionario DMSES.

En el capítulo 4 se planteará un modelo logístico que discrimine entre pacientes con un nivel de hemoglobina glicosilada superior al 7% de los que no.

En el capítulo 5 se hará un análisis multivariante de datos. Primero con un análisis de componentes principales, buscando relaciones entre la concentración de la hemoglobina glicosilada y el cuestionario DMSES. Segundo se hará un análisis de conglomerados y por último un análisis discriminante.

2. Material y métodos

2.1 Datos

Cómo se ha comentado previamente, los datos del estudio se extraen de un fichero con información sobre 700 pacientes con diabetes mellitus tipo II de Tailandia. El número de variables asciende a 145.

2.2 Selección de variables

Debido al alto número de variables sólo se seleccionará una parte del total de variables.

Las variables que contengan las respuestas a alguna pregunta concreta de alguno de los cuestionarios se descartarán debido a que ya se poseen las puntuaciones totales o parciales de cada cuestionario, siendo éstas de naturaleza continua. Las puntuaciones totales y parciales están centradas en el valor cero y han sido sometidas a un proceso de estandarización.

Se crearán por otra parte otras variables a partir de las existentes: BMI, sobrepeso y obesidad. Una vez obtenido el BMI, se descartarán para el análisis las variables peso y altura, ya que estas variables por sí mismas no aportan información sobre el estado de salud.

Las variables discretas con poca variabilidad no se toman en cuenta (accidente cerebrovascular, infarto cerebral, accidente cerebrovascular isquémico, accidente cerebrovascular, Hemorragia cerebral, ataque isquémico transitorio, angina de pecho, Insuficiencia cardíaca congestiva, revasculación coronaria, enfermedad arterial periférica, neuropatía).

#	Variable	Descripción [unidades]	Nota
1	id code	Número identificador del paciente	
2	hospital	Ciudad del hospital	1=phupaman, 2= srinakarin, 3=wegaronrat, 4= chula
3	gender	Sexo	1= hombre, 2= mujer
4	mstatus	Estado civil	1= soltero/a, 2=casado/a, 3=viudo/a, 4= divorciado/a, 5=separado/a
5	age	Edad [años]	>=20 años
6	edu	Nivel educativo	1=Sin estudios, 2=Elementales, 3=Secundaria, 4=Grado, 5=Máster, 6= Estudios superiores
7	religion	Religión	1=Budismo, 2= Islam, 3=Cristianismo, 4=Otra
8	income	Ingresos por hogar [moneda: Bath]	1= menos de 4,999, 2=5,000-9,999, 3=10,000-14,999, 4= 15,000-19,999, 5= 20,000-24,999, 6= más de 25,000
9	weight	Peso [kg]	
10	height	Altura [cm]	
11	dmdura	Duración de la diabetes [años]	>3 años
12	famhx	Antecedentes de T2DM en la familia	0= No, 1= Sí
13	famhxspec	Antecedentes de T2DM en la familia (detalle)	
14	comob	Comorbilidad	0= No, 1= Sí
15	comlip	Comorbilidad (lipidemia)	0= No, 1= Sí
16	comht	Comorbilidad (hipertensión)	0= No, 1= Sí
17	comchd	Comorbilidad (enfermedad coronaria)	0= No, 1= Sí
18	comkid	Comorbilidad (enfermedad renal)	0= No, 1= Sí
19	comoth	Otra comorbilidad	0= No, 1= Sí
20	comothx	Otra comorbilidad (detalle)	
21	dmrx	Tipo de tratamiento de T2DM	1= Nada,2= oral, 3= Insulina, 4= Oral e insulina
22	smk	Consumo de tabaco	1=Sí, 2= exfumador, 3= No
23	alcohol	Consumo de alcohol	1=Sí, 2= exbebedor, 3= No
24	hba1c	Hemoglobina glicosilada	
25	hba1cdate	Hemoglobina glicosilada (fecha de la muestra)	
26	ldl	Lipoproteína de baja densidad	
27	ldldate	Lipoproteína de baja densidad (fecha de la muestra)	
28	hdl	Lipoproteína de alta densidad	
29	hdldate	Lipoproteína de alta densidad (fecha de la muestra)	
30	trig	Triglicéridos	
31	trigdate	Triglicéridos (fecha de la muestra)	
32	sbp	Presión sanguínea sistólica	
33	dbp	Presión sanguínea diastólica	
34	bpdate	Presión sanguínea (fecha de la medida)	
35	compli	Complicación debida a la diabetes	0= No, 1= Sí
36	cva	Accidente cerebrovascular	0= No, 1= Sí
37	cereinfrac	Infarto cerebral	0= No, 1= Sí
38	ishemic	Accidente cerebrovascular isquémico	0= No, 1= Sí
39	stroke	Accidente cerebrovascular, sin especificar	0= No, 1= Sí
40	cerebhem	Hemorragia cerebral	0= No, 1= Sí
41	tia	Ataque isquémico transitorio	0= No, 1= Sí
42	angia	Angina de pecho	0= No, 1= Sí
43	chf	Insuficiencia cardíaca congestiva	0= No, 1= Sí
44	mi	Ataque de miocardio	0= No, 1= Sí
45	cororevas	Revasculación coronaria	0= No, 1= Sí
46	pad	Enfermedad arterial periférica	0= No, 1= Sí
47	neuropath	Neuropatía	0= No, 1= Sí
48	renal	Insuficiencia renal	0= No, 1= Sí
49	dn	Nefropatía diabética	0= No, 1= Sí
50	dr	Retinopatía diabética	0= No, 1= Sí
51	othcomp	Otra complicación	0= No, 1= Sí
52	DMSES.Diet	Puntuación estandarizada DMSES, factor dieta.	Variable continua
53	DMSES.Monitor	Puntuación estandarizada DMSES, factor monitor.	Variable continua
54	DMSES.Physical	Puntuación estandarizada DMSES, factor actividad física.	Variable continua
55	DMSES.Regimen	Puntuación estandarizada DMSES, factor régimen.	Variable continua
56	DMSES.Total	Puntuación total DMSES estandarizada.	Variable continua
57	DK.10	Puntuación estandarizada sobre conocimiento de la enfermedad	Variable continua entre 0 y 10
58	bmi	Índice de masa corporal	$Peso[kg]/(altura[m])^2$
59	sobrepeso	si BMI > 25	0= No, 1= Sí
60	obesidad	Si BMI > 30	0= No, 1= Sí
61	alerta	Diabetes mal controlada	0=No (hba1c <7%), 1=Sí (hba1c >=7%)

Tabla 2: Lista de variables

2.3 Análisis descriptivo de las variables

Variables categóricas

A continuación, se muestran todas las tablas de frecuencias de las variables categóricas para una mejor visualización.

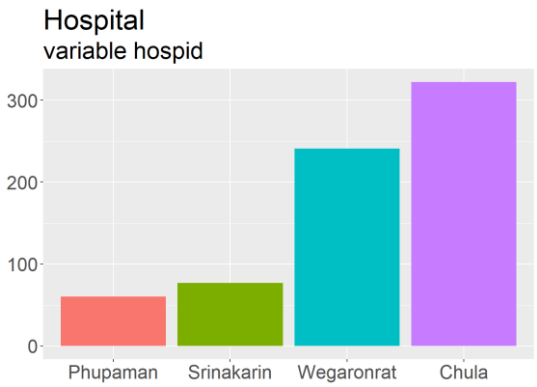


Figura 2: Frecuencia variable HOSPID

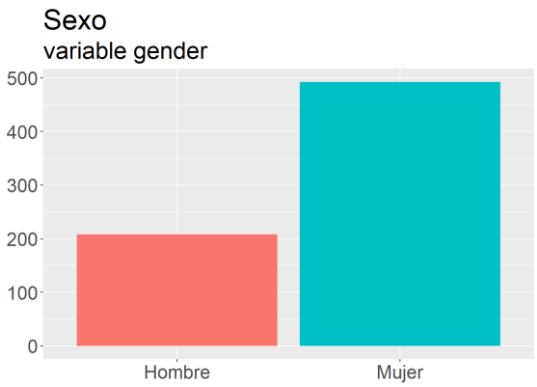


Figura 3: Frecuencia variable GENDER

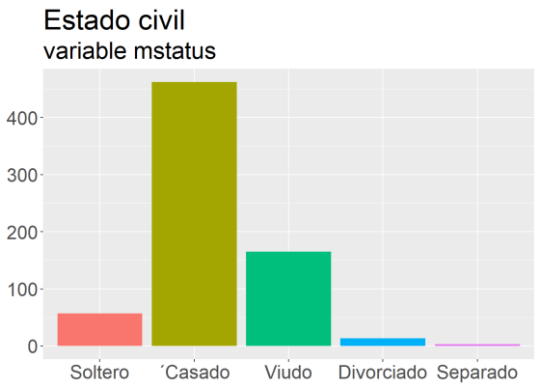


Figura 4: Frecuencia variable MSTATUS

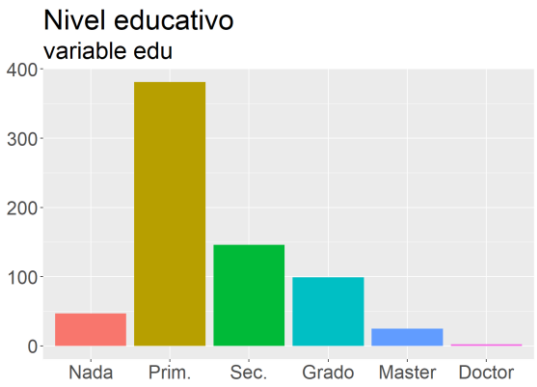


Figura 5: Frecuencia variable EDU

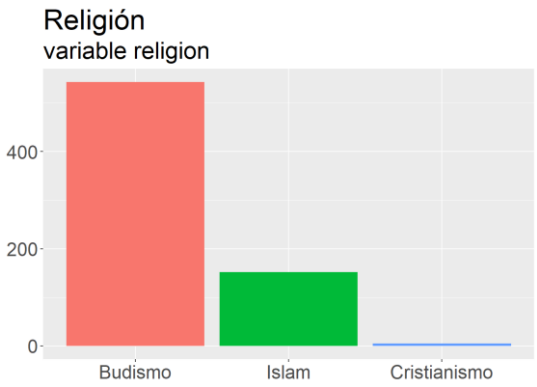


Figura 6: Frecuencia variable RELIGION

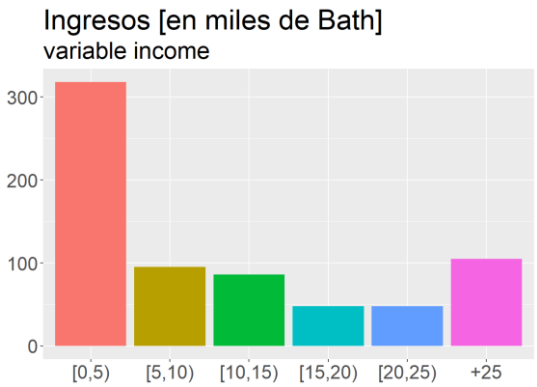


Figura 7: Frecuencia variable INCOME

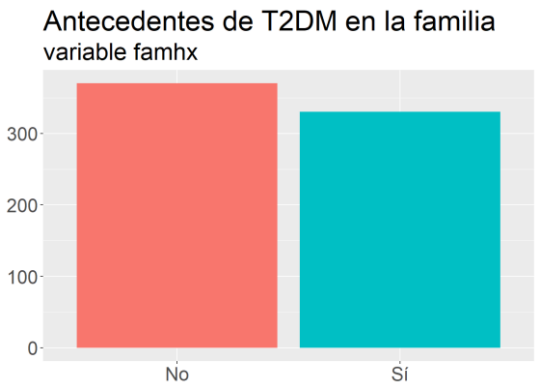


Figura 8: Frecuencia variable FAMHX

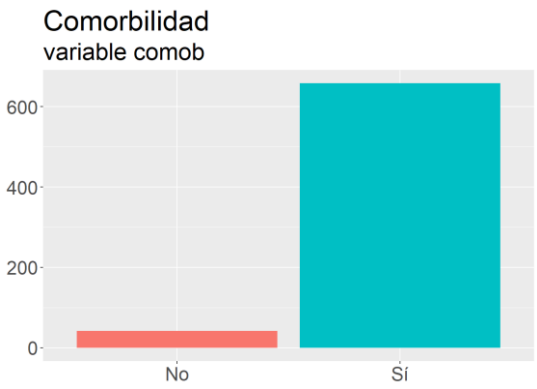


Figura 9: Frecuencia variable COMOB

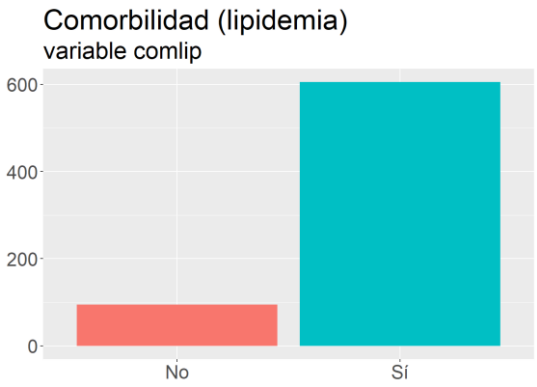


Figura 10: Frecuencia variable COMLIP

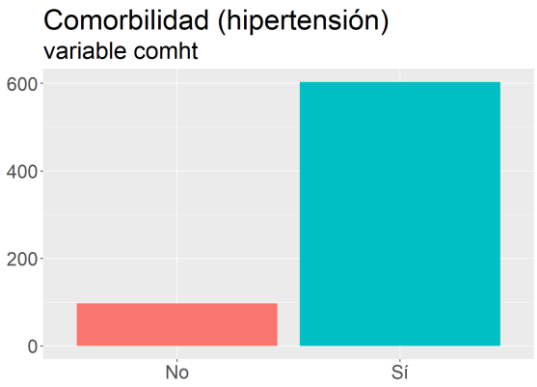


Figura 11: Frecuencia variable COMHT

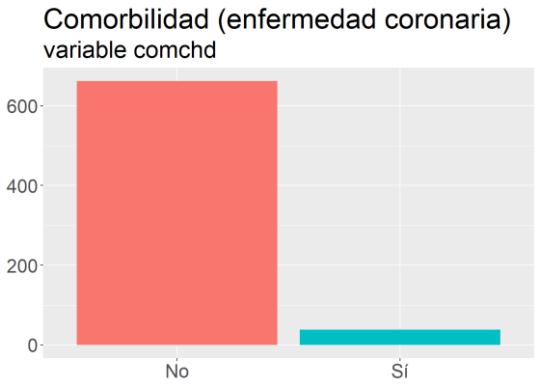


Figura 12: Frecuencia variable COMCHD

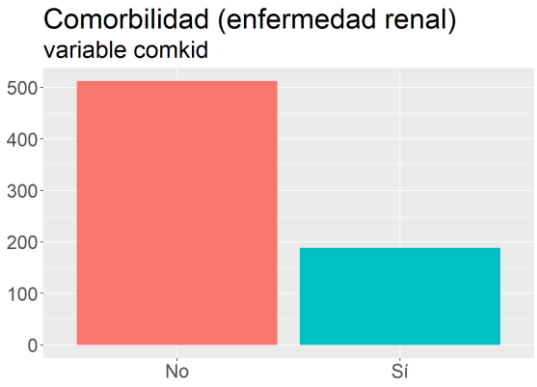


Figura 13: Frecuencia variable COMKID

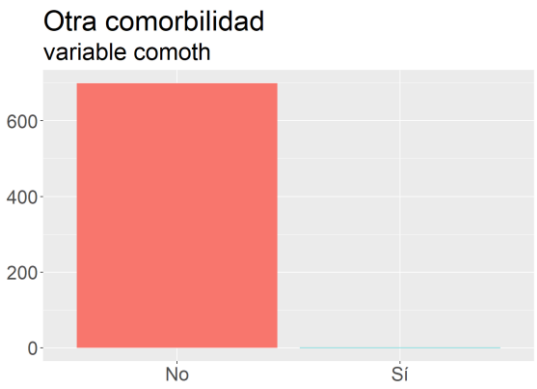


Figura 14: Frecuencia variable COMOTH

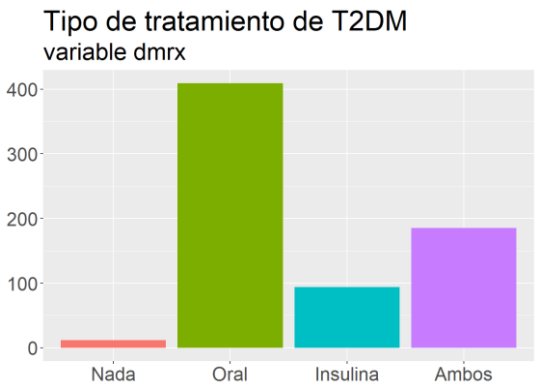


Figura 15: Frecuencia variable DMRX

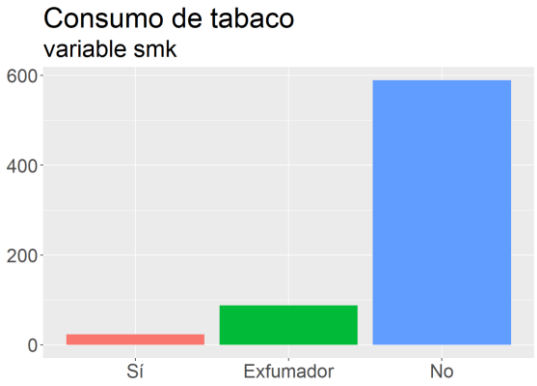


Figura 16: Frecuencia variable SMK

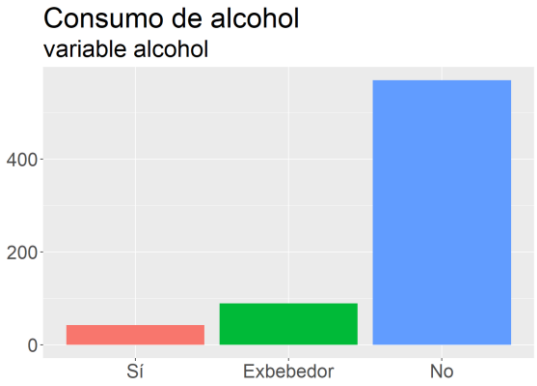


Figura 17: Frecuencia variable ALCOHOL

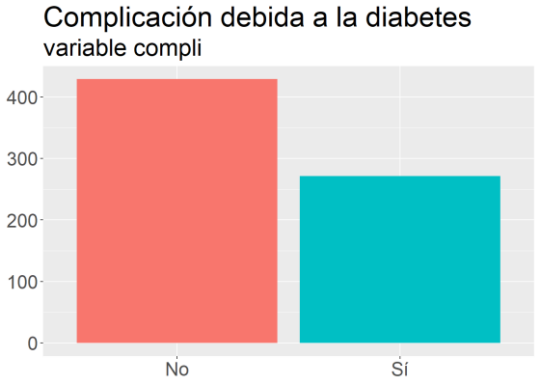


Figura 18: Frecuencia variable COMPLI

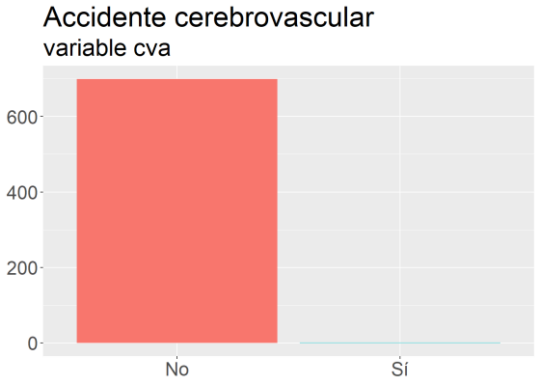


Figura 19: Frecuencia variable CVA

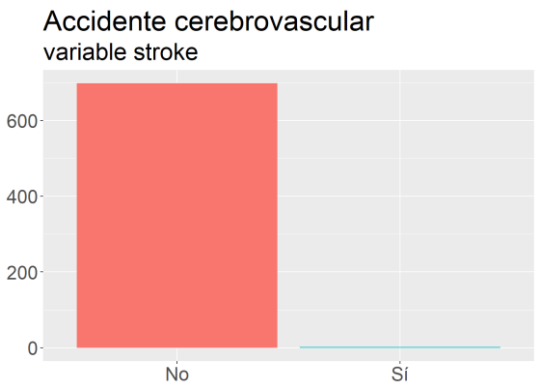


Figura 20: Frecuencia variable STROKE

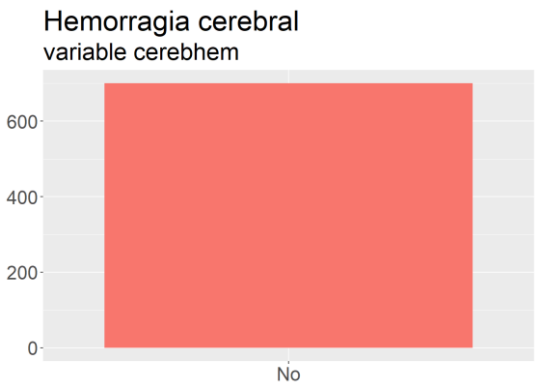


Figura 21: Frecuencia variable CEREBHEM

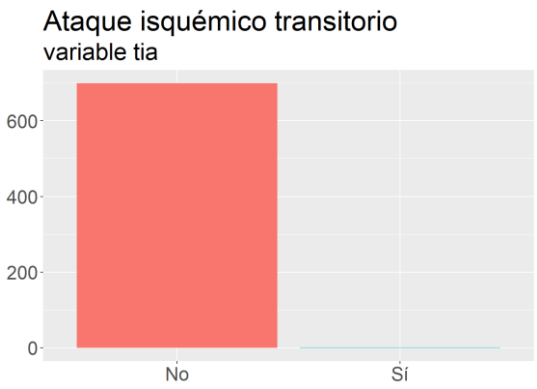


Figura 22: Frecuencia variable TIA

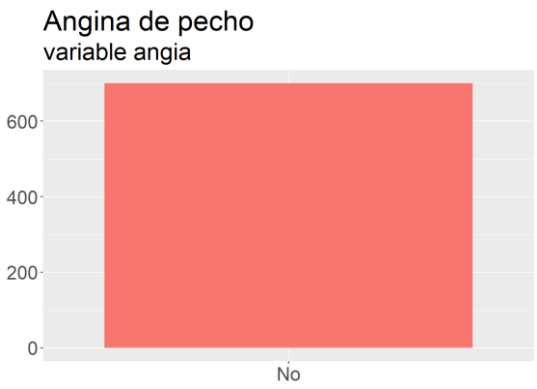


Figura 23: Frecuencia variable ANGIA

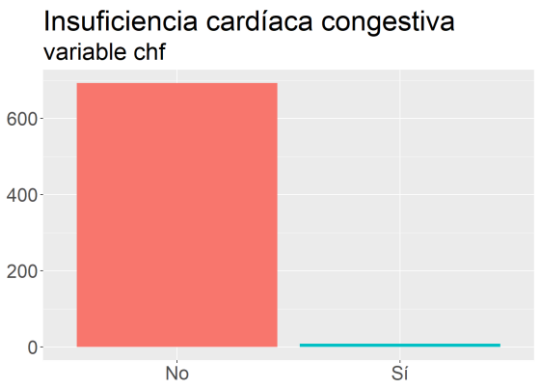


Figura 24: Frecuencia variable CHF

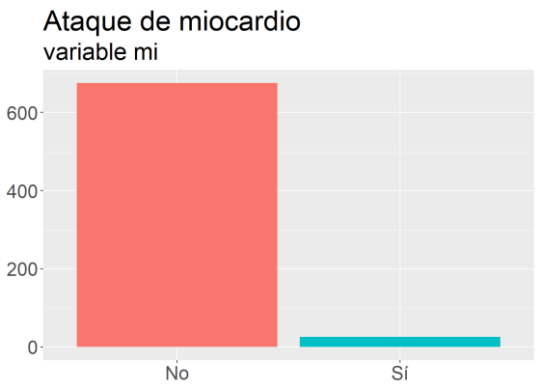


Figura 25: Frecuencia variable MI

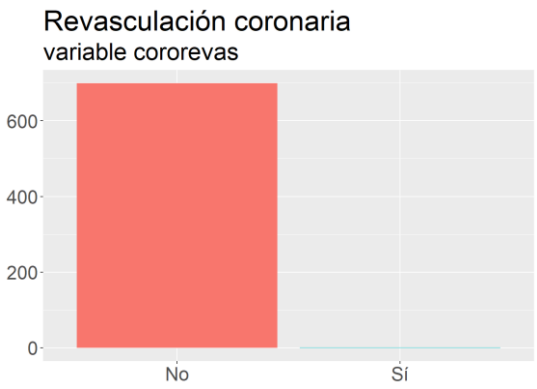


Figura 26: Frecuencia variable COROREVAS

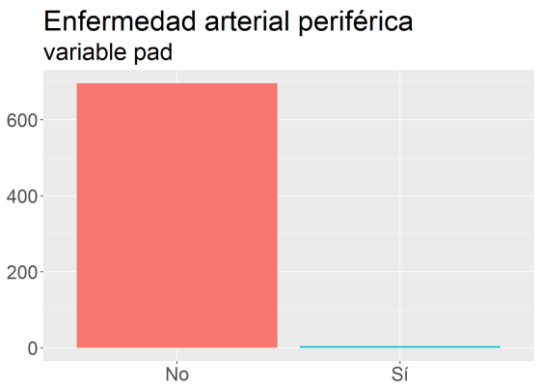


Figura 27: Frecuencia variable PAD

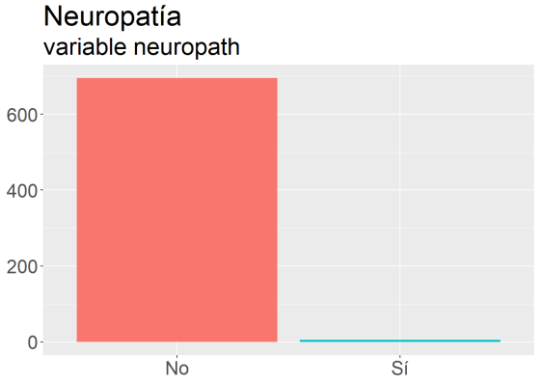


Figura 28: Frecuencia variable NEUROPATH

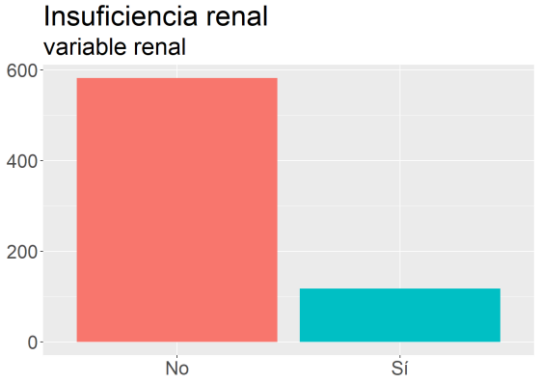


Figura 29: Frecuencia variable RENAL

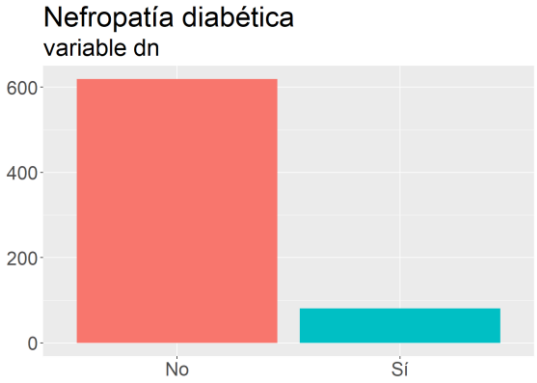


Figura 30: Frecuencia variable DN

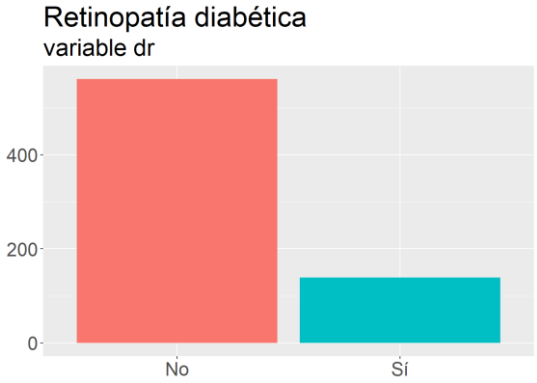


Figura 31: Frecuencia variable DR

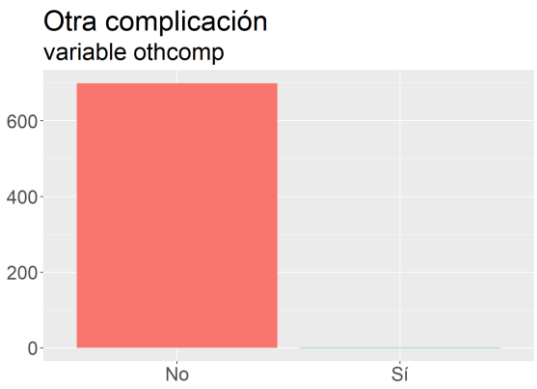


Figura 32: Frecuencia variable OTHCOMP

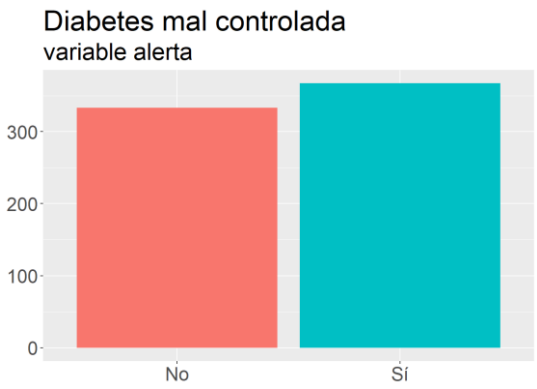


Figura 33: Frecuencia variable ALERTA

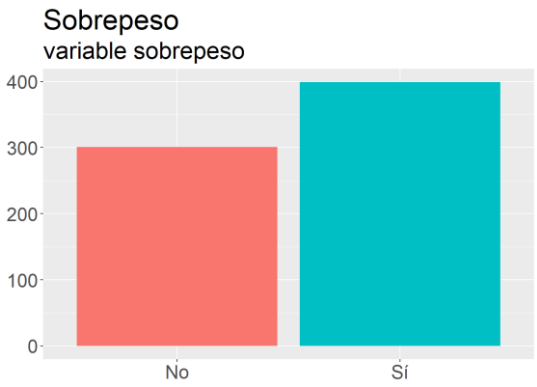


Figura 34: Frecuencia variable SOBREPESO

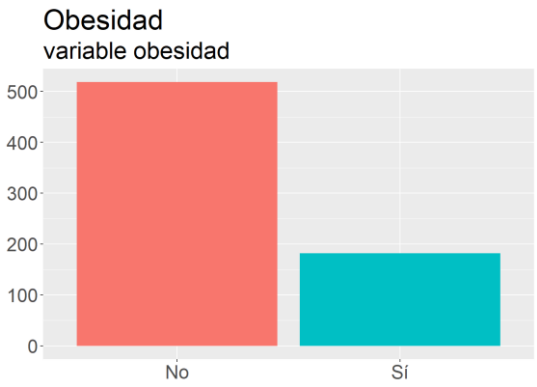


Figura 35: Frecuencia variable OBESIDAD

Variables continuas

	Mínimo	Q1	Mediana	Media	Q3	Máximo
Edad	26,00	59,00	65,00	65,16	73,00	95,00
Duración diabetes	4,00	7,00	10,00	13,53	20,00	45,00
Hemoglobina glicosilada	2,00	6,40	7,10	7,58	8,30	15,00
LDL	8,70	81,00	97,00	100,77	116,00	221,00
HDL	17,00	41,88	50,00	50,75	58,00	116,00
Triglicéridos	29,30	95,75	127,00	149,02	175,00	999,00
Presión sistólica	93,00	123,00	133,00	134,79	145,00	217,00
Presión diastólica	22,00	65,00	73,00	73,02	80,00	112,00
Puntuación DMSES (Dieta)	-1,46	-0,54	-0,07	0,00	0,52	1,03
Puntuación DMSES (Monitor)	-0,84	-0,23	-0,04	0,00	0,25	0,55
Puntuación DMSES (Act. Física)	-1,29	-0,29	0,02	0,00	0,29	0,79
Puntuación DMSES (Régimen)	-2,86	0,08	0,09	0,00	0,11	0,51
Puntuación DMSES (Total)	-5,63	-1,03	-0,03	0,00	1,04	2,51
Conocimiento de la enfermedad	0,00	4,22	5,49	5,53	6,92	10,00

Tabla 3: Variables continuas

Densidades de las variables continuas

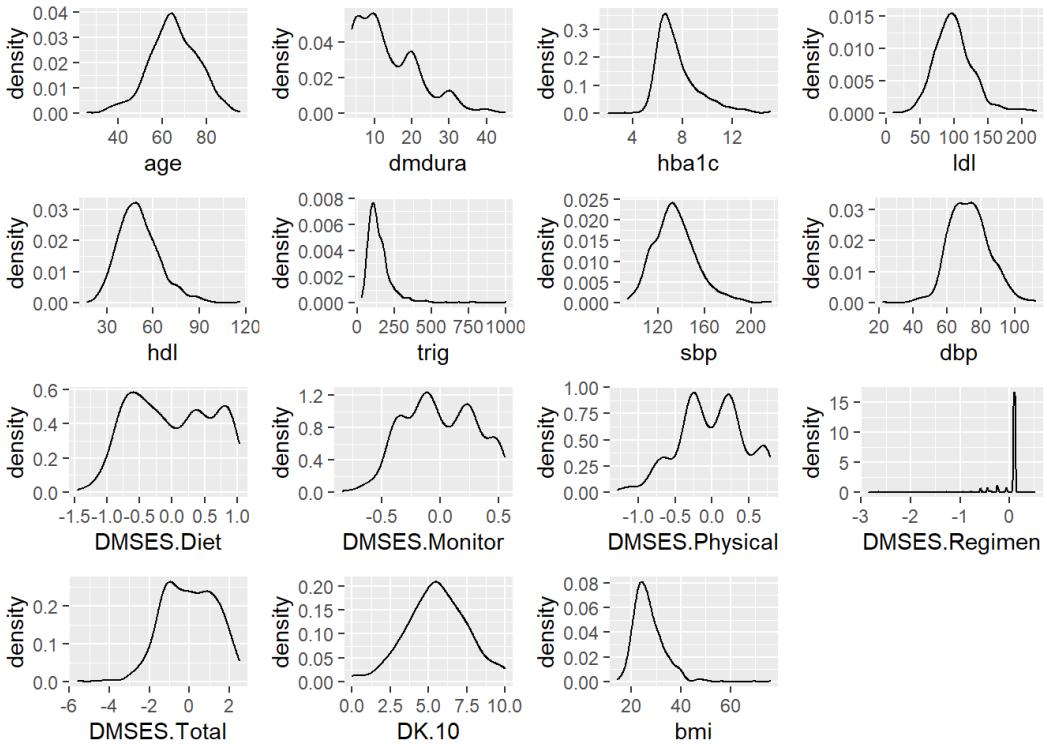


Figura 36:Densidades variables continuas

2.4 Datos faltantes

El conjunto de datos no presenta datos faltantes en ninguna de las variables de análisis. Sólo se tienen datos faltantes en dos variables fecha (fecha.ldl y fecha.bp) que finalmente fueron descartadas del análisis. En ambas variables los datos faltantes se dieron únicamente en un solo paciente

2.5 Outliers

Se estudia el caso de outliers mediante el uso de boxplot para las variables edad (age), duración de la diabetes en el paciente (dmdura), concentración de hemoglobina glicosilada en sangre (hba1c), lipoproteínas de baja densidad (ldl), lipoproteínas de alta densidad (hdl), presión sanguínea sistólica (sbp), presión sanguínea diastólica (dbp), las puntuaciones del cuestionario DMSES(DMSES.Diet, DMSES.Monitor, DMSES.Physical, DMSES.Regimen, DMSES.Total), conocimiento de la enfermedad (DK.10) y BMI.

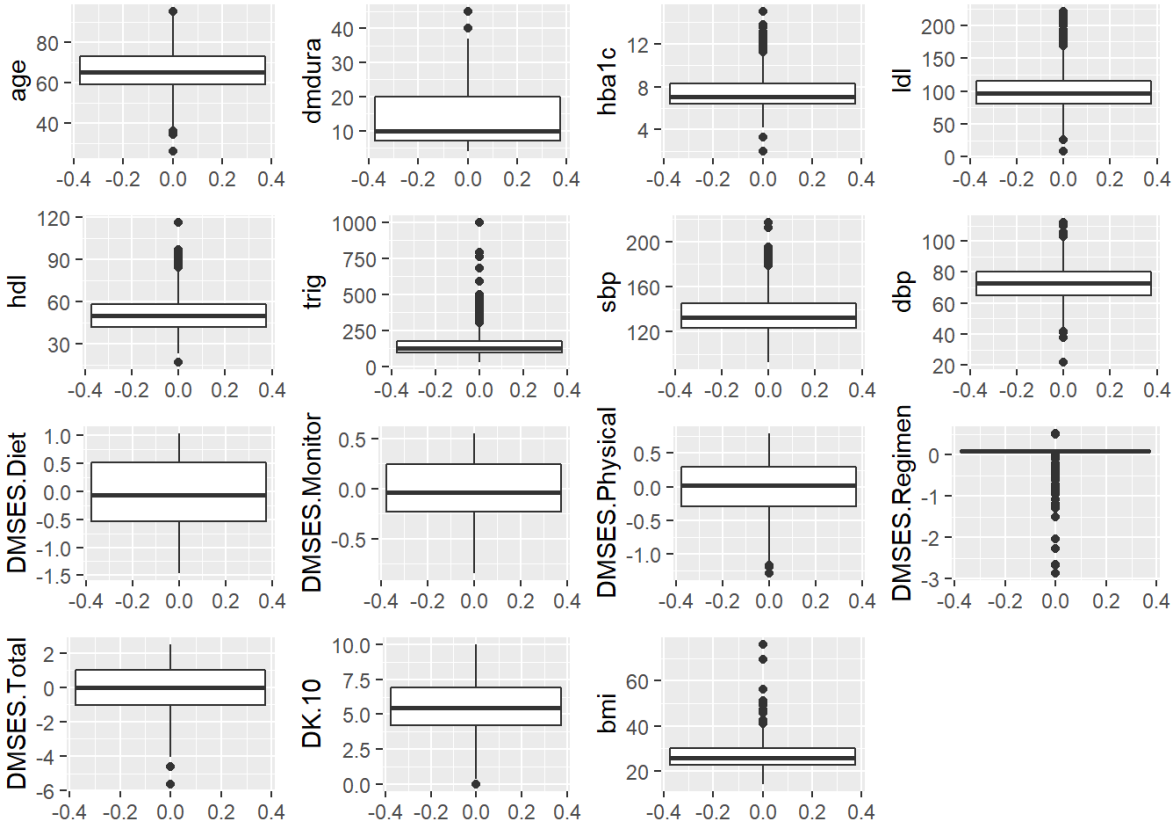


Figura 37: Outliers de variables continuas

Se observa una gran cantidad de outliers en la puntuación de apartado régimen del cuestionario DMSES.(DMSES.Regimen). Por ello, en un principio esta puntuación no será tomada como predictora de algún modelo si el resto de las puntuaciones tienen ya un buen comportamiento.

La variable predictora triglicéridos también tiene un elevado número de outliers. Se estudiará su caso si ésta es necesaria en el modelo final

3. Regresión lineal

3.1 Análisis de correlación de los predictores continuos.

Se analizan las correlaciones entre las variables continuas para detectar posible multicolinealidad entre los predictores.

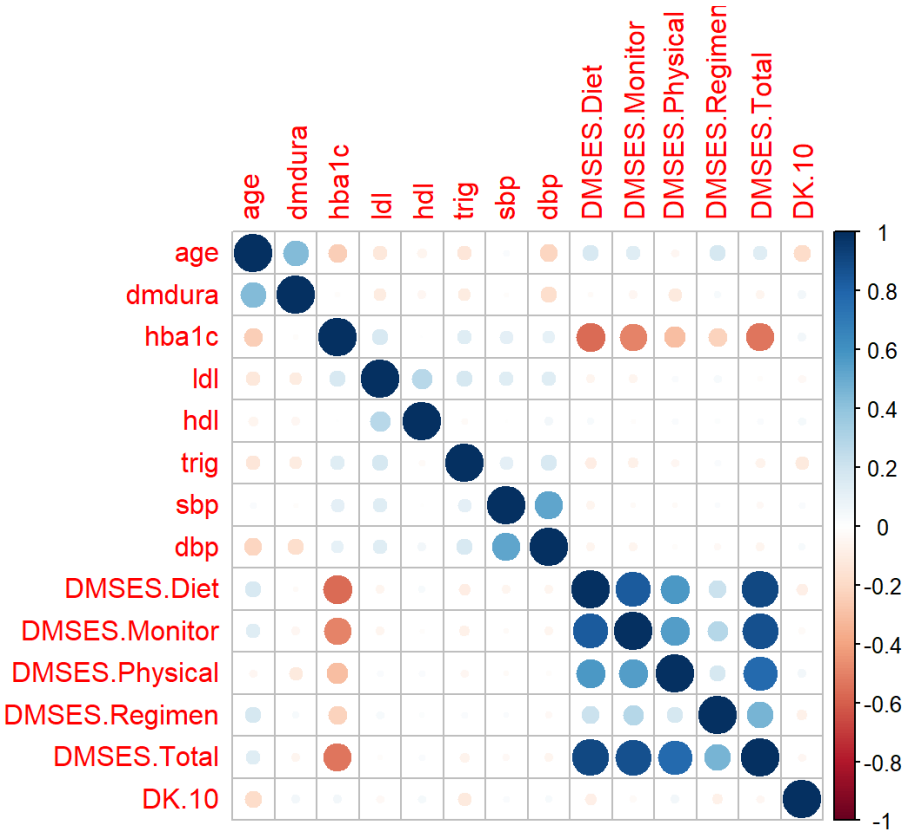


Figura 38: Correlaciones

Observamos una gran correlación entre las distintas puntuaciones del cuestionario DMSES. Y en menor medida de éstas con la variable de respuesta, hemoglobina glicosilada (hba1c).

Se tendrá que evitar usar más de una variable de puntuación del cuestionario DMSES en el mismo modelo.

Entre el resto de las variables predictoras continuas no se detectan correlaciones elevadas.

A continuación, se lista los 15 pares de variables con mayor correlación:

Variable 1	Variable 2	Correlación
DMSES.Diet	DMSES.Total	0. 908
DMSES.Monitor	DMSES.Total	0.876
DMSES.Diet	DMSES.Monitor	0.832
DMSES.Physical	DMSES.Total	0.779
DMSES.Diet	DMSES.Physical	0.578
hba1c	DMSES.Diet	-0.563
DMSES.Monitor	DMSES.Physical	0.558
hba1c	DMSES.Total	-0.534
sbp	dbp	0.530
hba1c	DMSES.Monitor	-0.499
DMSES.Regimen	DMSES.Total	0.461
age	dmdura	0.436
hba1c	DMSES.Physical	-0.303
DMSES.Monitor	DMSES.Regimen	0.281
ldl	hdl	0.271

Tabla 4: Variables con mayor correlación

3.2 Normalidad de la variable de respuesta

La hemoglobina glicosilada no sigue una distribución normal. Al aplicar logaritmos se obtiene una cierta mejora. Por ello se buscarán modelos donde la variable de respuesta sea el logaritmo de la hemoglobina glicosilada.

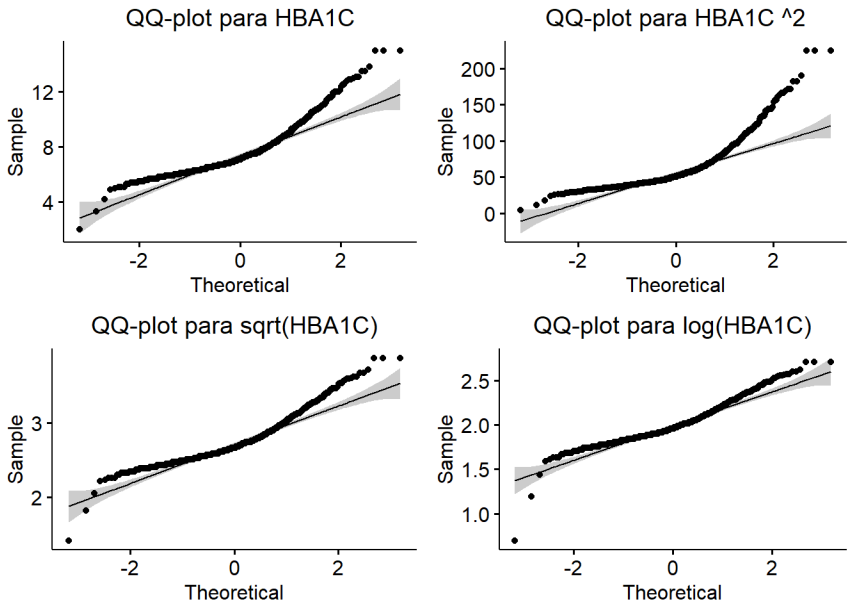


Figura 39: Normalidad de la variable respuesta y sus transformaciones

3.3 Planteamiento del modelo

El modelo que se plantea se muestra en la siguiente fórmula:

log(hba1c) ~ DMSES.Total + age + renal + ldl

```
##
## Call:
## lm(formula = log(hba1c) ~ DMSES.Total + age + renal + ldl, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32692 -0.09223 -0.01355  0.08516  0.59417
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1135308  0.0470599  44.911  < 2e-16 ***
## DMSES.Total -0.0786385  0.0050494 -15.574  < 2e-16 ***
## age         -0.0031851  0.0006003  -5.306 1.51e-07 ***
## renalSí      0.0684713  0.0175791   3.895 0.000108 ***
## ldl          0.0008448  0.0002106   4.012 6.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1695 on 695 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3466
## F-statistic:  93.7 on 4 and 695 DF,  p-value: < 2.2e-16
```

Obtenemos un modelo con las variables DMSES.Total, edad, insuficiencia renal (Sí/No) y LDL significativo, pero con una varianza explicada baja, R2 ajustado del 34.66%.

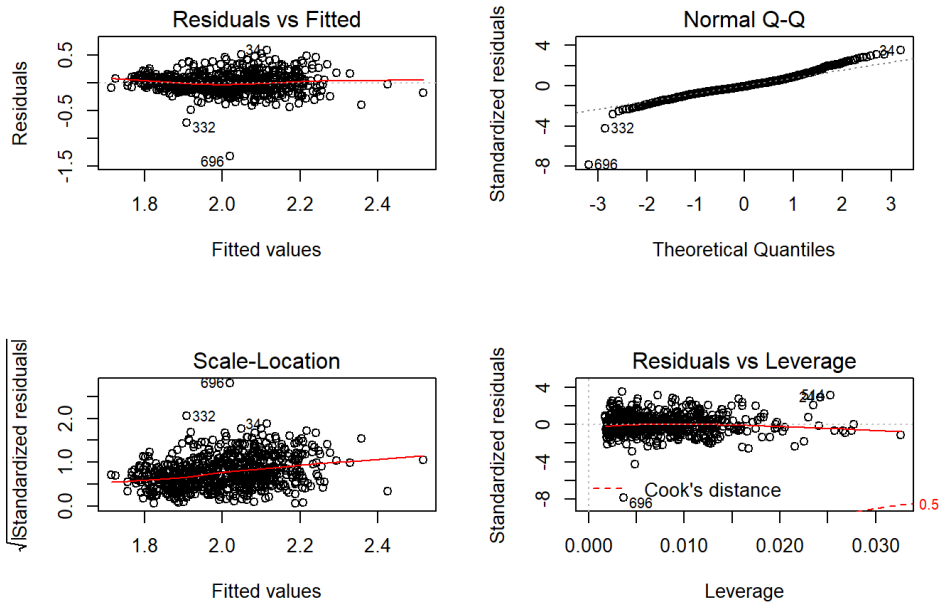


Figura 40: Plots modelo lineal

Normalidad de los residuos

Observando el gráfico QQ se puede deducir que los residuos siguen una distribución normal puesto que se ajustan muy bien a la recta de los cuantiles teóricos. En este caso, esto es muy importante porque, al aplicar el logaritmo a la hemoglobina glicosilada, se mejoraba la normalidad de la variable de respuesta pero aun existía una cierta divergencia respecto a la normal.

Homocedasticidad/ heterocedasticidad:

El gráfico scale-location confirma la homocedasticidad ya que sigue una línea recta casi horizontal y la nube de puntos tiene una amplitud homogénea.

• Multicolinealidad:

Los factores de inflación de la varianza están cercanos al 1. Por tanto, se puede descartar la multicolinealidad.

DMSES.Total	age	renalSí	LDL
1.066413	1.049026	1.054865	1.017554

4. Regresión logística

El objetivo de este apartado será comprobar si la puntuación del cuestionario DMSES (total o de alguno de sus apartados) puede ayudar a discriminar entre pacientes que controlan la diabetes y los que no.

Para ello crearemos modelos logísticos con cada una de las puntuaciones y la variable alerta que, recordemos, es una variable categórica con valor 1 si la hemoglobina glicosilada tiene un valor mayor que 7% (diabetes mal controlada) y 0 en caso contrario.

Modelo logístico 1: alerta ~ DMSES.Diet
Modelo logístico 2: alerta ~ DMSES.Monitor
Modelo logístico 3: alerta ~ DMSES.Regimen
Modelo logístico 4: alerta ~ DMSES.Physical

Para probar los diferentes modelos se particionará el fichero de datos tomando un 80% de los registros (560) para el conjunto de training y un 20% (140) para el conjunto de comprobación o test.

	DMSES Diet	DMSES Monitor	DMSES Regimen	DMSES Physical
Precisión :	0.8071	0.8286	0.6429	0.7
95% CI :	(0.7319, 0.8689)	(0.7197, 0.9082)	(0.5193, 0.7539)	(0.5787, 0.8038)
Tasa de No Información :	0.5357	0.5	0.5	0.5
-Valor P [Acc > NIR] :	1.787e-11	1.124e-08	0.01123	0.0005466
Kappa :	0.6111	0.6571	0.2857	0.4
Mcnemar's Test P-Value :	0.7003	1	1,59E-06	0.6625206
Sensibilidad :	0.7692	0.8286	0.2857	0.6571
Especifidad :	0.8400	0.8286	1.0000	0.7429
Pos Pred Value :	0.8065	0.8286	1.0000	0.7188
Neg Pred Value :	0.8077	0.8286	0.5833	0.6842
Predominio :	0.4643	0.5000	0.5000	0.5000
Tasa de detección :	0.3571	0.4143	0.1429	0.3286
Prevalencia de detección :	0.4429	0.5000	0.1429	0.4571
Precisión equilibrada :	0.8046	0.8286	0.6429	0.7000
Clase 'Positiva' :	No	No	No	No

Tabla 5: Precisión modelos logísticos

Cómo se puede observar, la mayoría de las puntuaciones del cuestionario DMSES son una excelente herramienta para discriminar entre pacientes que controlan la

enfermedad de los que no. Se obtienen precisiones por encima del 80% utilizando la puntuación de la parte de dieta y monitoreo.

Curvas ROC

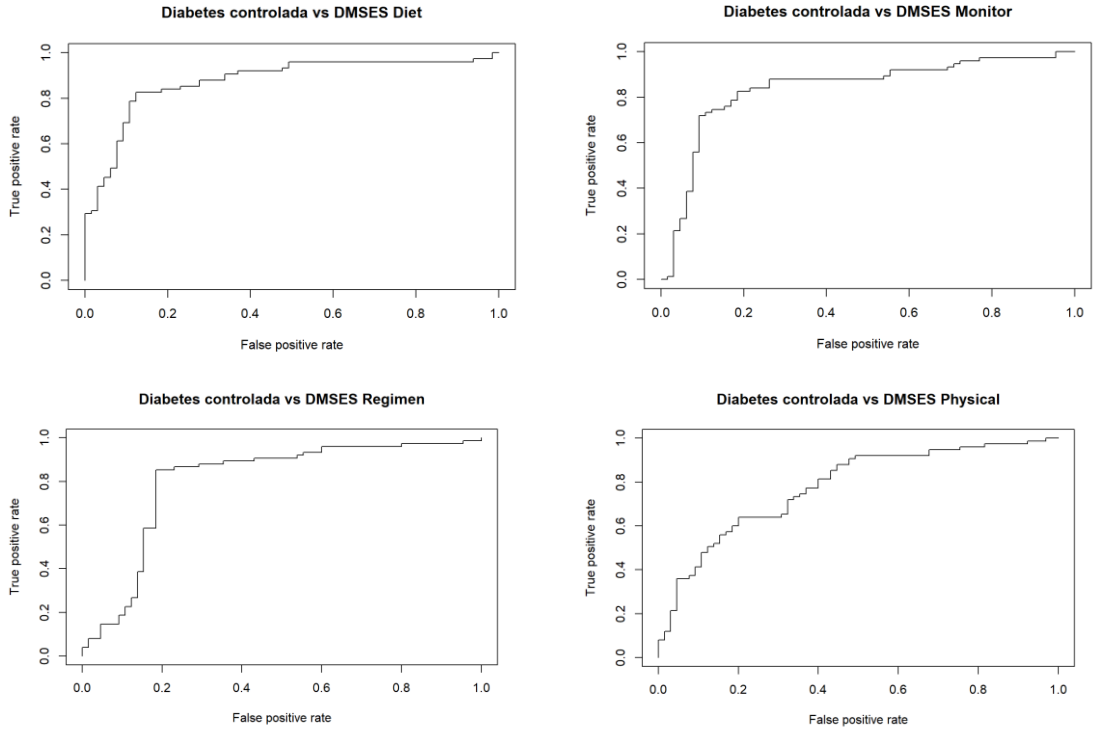


Figura 41: Curvas ROC

Una pregunta natural que se podría formular es si el añadir alguna variable más implica una mejora en la clasificación de los pacientes. Especialmente son de interés las variables que puedan ser obtenidas fácilmente en países subdesarrollados con escasa atención médica a la población.

Al comparar, por ejemplo, los modelos:

alerta~DMSES Diet y

alerta~DMSES Diet + bmi + dmdura + age

no se obtienen mejoras significativas en la precisión de los modelos por lo que el simple resultado del cuestionario es una buena herramienta para discriminar entre los pacientes con hemoglobina glicosilada mayor o menor del 7%.

5. Análisis multivariante

5.1 Análisis de componentes principales

En una primera aproximación al análisis de componentes principales se puede comprobar que ninguna componente principal acapara por sí misma un porcentaje notable de la variación de los datos. La primera componente explica únicamente un 27.5% de la variación y es necesario acudir a las tres primeras componentes principales para sobrepasar el 50% de la variación de los datos.

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.9614 1.3865 1.17696 1.0979 1.06861 0.94293 0.90553
## Proportion of Variance 0.2748 0.1373 0.09895 0.0861 0.08157 0.06351 0.05857
## Cumulative Proportion  0.2748 0.4121 0.51104 0.5971 0.67871 0.74221 0.80078
##                           PC8     PC9   PC10    PC11    PC12    PC13     PC14
## Standard deviation     0.8499 0.77263 0.7010 0.65595 0.62517 0.39650 1.46e-15
## Proportion of Variance 0.0516 0.04264 0.0351 0.03073 0.02792 0.01123 0.00e+00
## Cumulative Proportion  0.8524 0.89502 0.9301 0.96085 0.98877 1.00000 1.00e+00
```

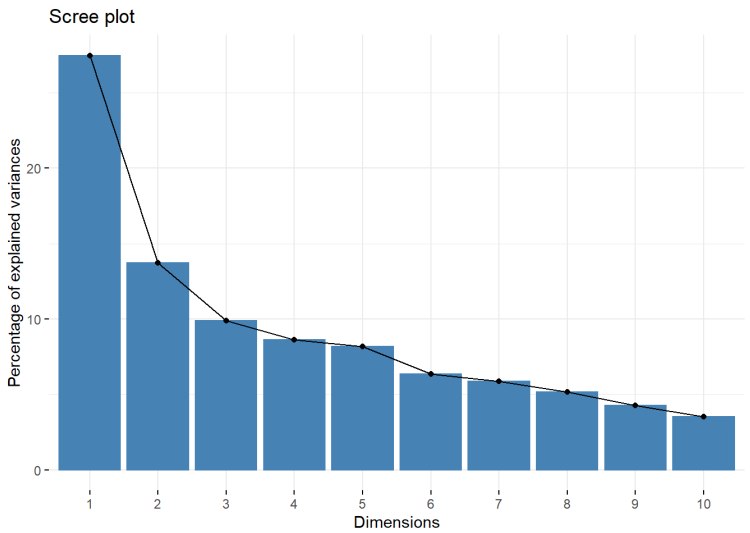


Figura 42: Varianza explicada por componente principal

No obstante, podemos ver que la hemoglobina glicosilada (-0.34) y los cuestionarios DMSES (0.46, 0.45, 0.37, 0.22 y 0.50) tienen una gran aportación a la primera componente principal, si miramos sus valores absolutos, y además se deduce una relación inversa, dado que tienen diferente signo.

Esto confirma el hecho empírico de que a mayor puntuación del cuestionario DMSES menor es el valor de la hemoglobina glicosilada en sangre.

En las siguientes dos componentes principales, la aportación en valor absoluto de estas variables es escaso o residual.

En la siguiente tabla se puede ver la aportación de cada variable continua a las primeras diez componentes principales.

```
##                   PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
## age              0.12  0.40 -0.51  0.11 -0.09  0.10 -0.08  0.01 -0.10  0.52
## dmdura           0.00  0.40 -0.39  0.02 -0.33 -0.25 -0.37 -0.02  0.18 -0.56
## hba1c           -0.34 -0.13  0.03  0.05 -0.02 -0.24 -0.11 -0.46  0.56  0.15
## ldl             -0.05 -0.35 -0.07  0.55 -0.19 -0.09 -0.18 -0.42 -0.56 -0.08
## hdl              0.00 -0.18  0.08  0.52 -0.51  0.30  0.07  0.43  0.38  0.08
## trig            -0.07 -0.28 -0.17  0.20  0.45 -0.43 -0.42  0.52  0.04  0.11
## sbp             -0.05 -0.36 -0.54 -0.30 -0.16  0.10  0.02 -0.10  0.04  0.34
## dbp             -0.07 -0.49 -0.32 -0.31 -0.07  0.19  0.09  0.12  0.01 -0.42
## DMSES.Diet       0.46 -0.05  0.01 -0.02  0.00  0.14 -0.19  0.02 -0.02 -0.07
## DMSES.Monitor    0.45 -0.07  0.01 -0.06  0.01  0.02 -0.11 -0.02 -0.02  0.00
## DMSES.Physical   0.37 -0.18  0.16 -0.06 -0.05 -0.14 -0.21 -0.25  0.33  0.13
## DMSES.Regimen    0.22 -0.01 -0.25  0.23  0.05 -0.49  0.73  0.00  0.08 -0.09
## DMSES.Total      0.50 -0.11  0.00  0.01  0.00 -0.09 -0.01 -0.08  0.11 -0.01
## DK.10           -0.04 -0.05  0.26 -0.36 -0.59 -0.50 -0.03  0.23 -0.26  0.21
```

De una manera más gráfica, si se representa las tres primeras componentes principales en gráficas bidimensionales, se puede observar que efectivamente es la primera componente principal (PC1) la que separa los pacientes que tienen un control de la diabetes de los que no.

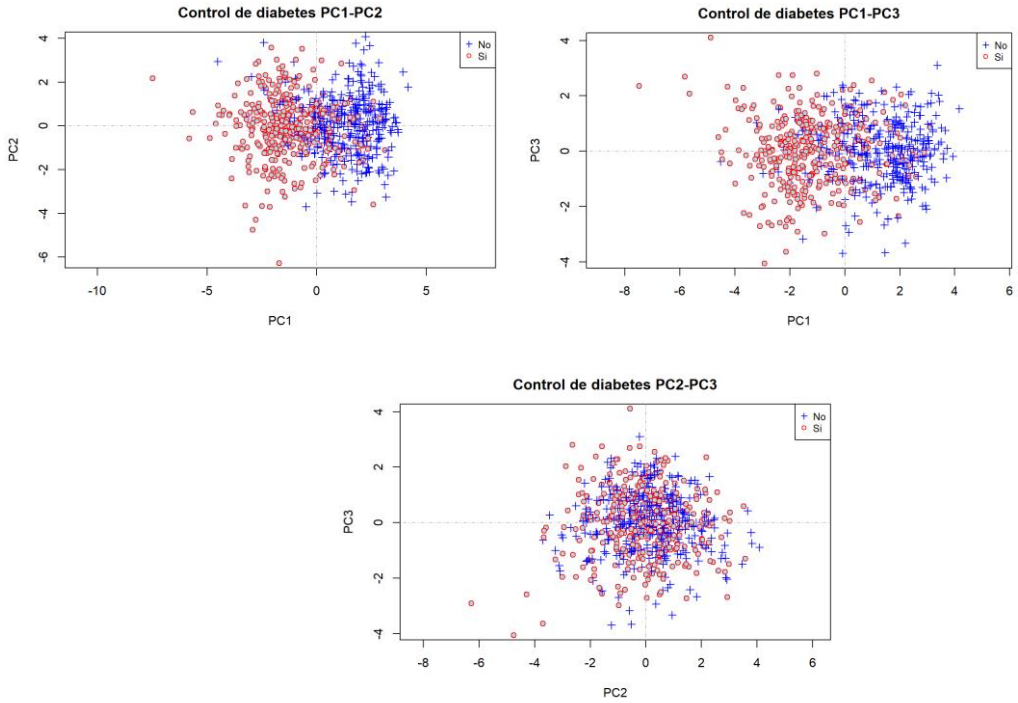


Figura 43: Representación de las tres primeras componentes principales

Y, por último, el análisis gráfico de la aportación de las variables en las dos primeras componentes principales pone de relieve como las puntuaciones del cuestionario DMSES se contraponen al valor de la hemoglobina glicosilada.

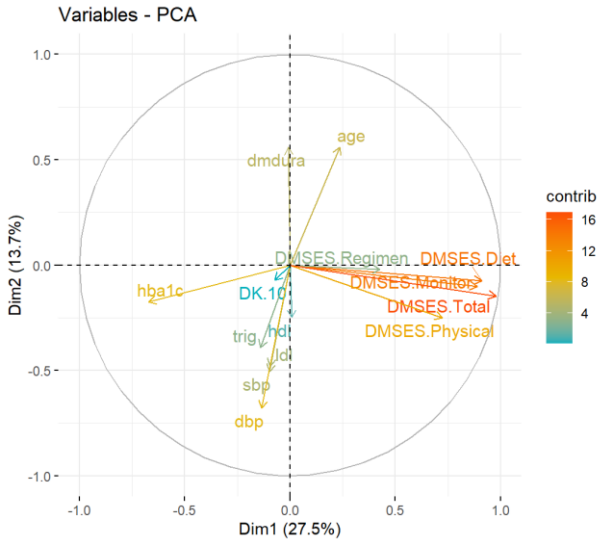


Figura 44: Aportación de cada variable a las dos primeras componentes principales

5.2 Análisis de conglomerados

En el fichero de datos contiene información sobre 700 pacientes, 317 de los cuales tienen la hemoglobina glicosilada por debajo del 7% y 383 por encima o igual a ese valor.

Si se ordena el fichero ascendentemente por el valor de la hemoglobina glicosilada, se toman las variables continuas y se escalan para evitar que un mayor rango de valores influya en las distancias, se obtienen el siguiente diagrama:

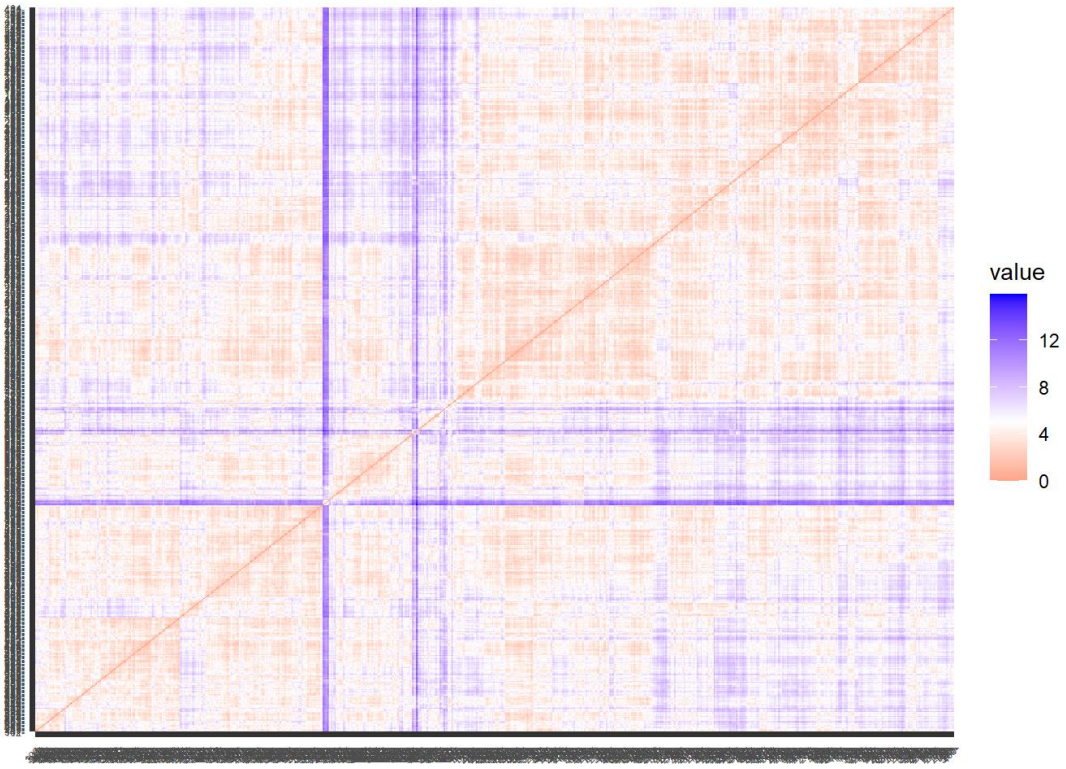


Figura 45: Mapa de distancias

Se observa en el cuadrante superior derecho un gran conglomerado que correspondería a la mitad aproximada de pacientes que tienen valores bajos de la hemoglobina glicosilada.

En cambio, las filas de la mitad de la imagen de las distancias, que correspondería a los pacientes con mayor nivel de hemoglobina glicosilada, no presentan un conglomerado unitario tan evidente.

La gran cantidad de pacientes dificulta el poder visualizar claramente el dendrograma obtenido, pero sí que se observan claramente dos brazos principales.

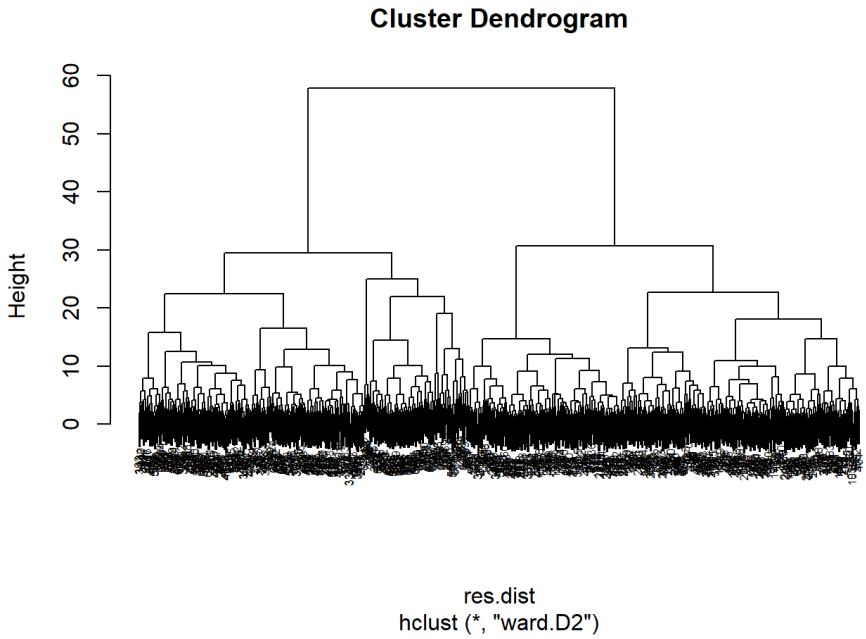


Figura 46: Dendrograma

Utilizando el método de conglomerados de K-Means y buscando la mejor manera de aglutinar los puntos optando de entre 2 a 12 conglomerados, se obtiene que la mejor manera es con 2 conglomerado.

La gran cantidad de puntos hace difícil distinguir como se reparten estos entre los dos conglomerados, pero una rápida observación de los números que sí son visibles nos hace ver que la mayoría de los puntos por debajo de 300 están en un conglomerado (los primeros 300 eran los de menor valor de hemoglobina glicosilada) y el resto en el otro.

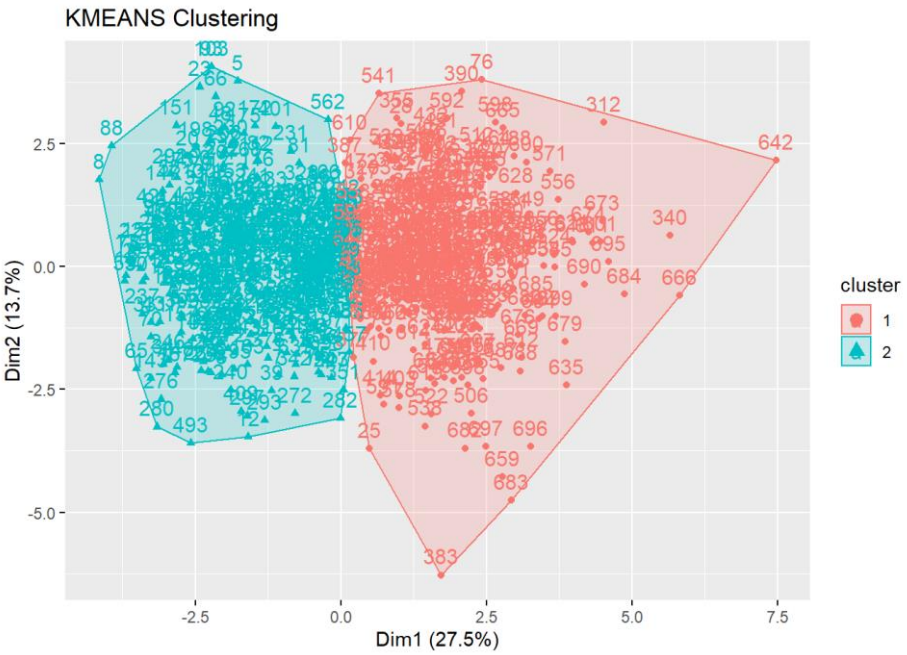


Figura 47: Conglomerados K-Means

5.3 Análisis discriminante

Para complementar el análisis hecho en la regresión logística, se procede a utilizar el análisis multivariante.

Se plantea el siguiente análisis lineal discriminante mediante la función *lda* de R

alerta ~ DMSES.Diet + age

Se obtiene una aparente buena clasificación, del 81.85%, pero hemos de hacer notar que este método necesita de la normalidad de las variables predictoras y este supuesto no se cumple para la componente DMSES.Diet.

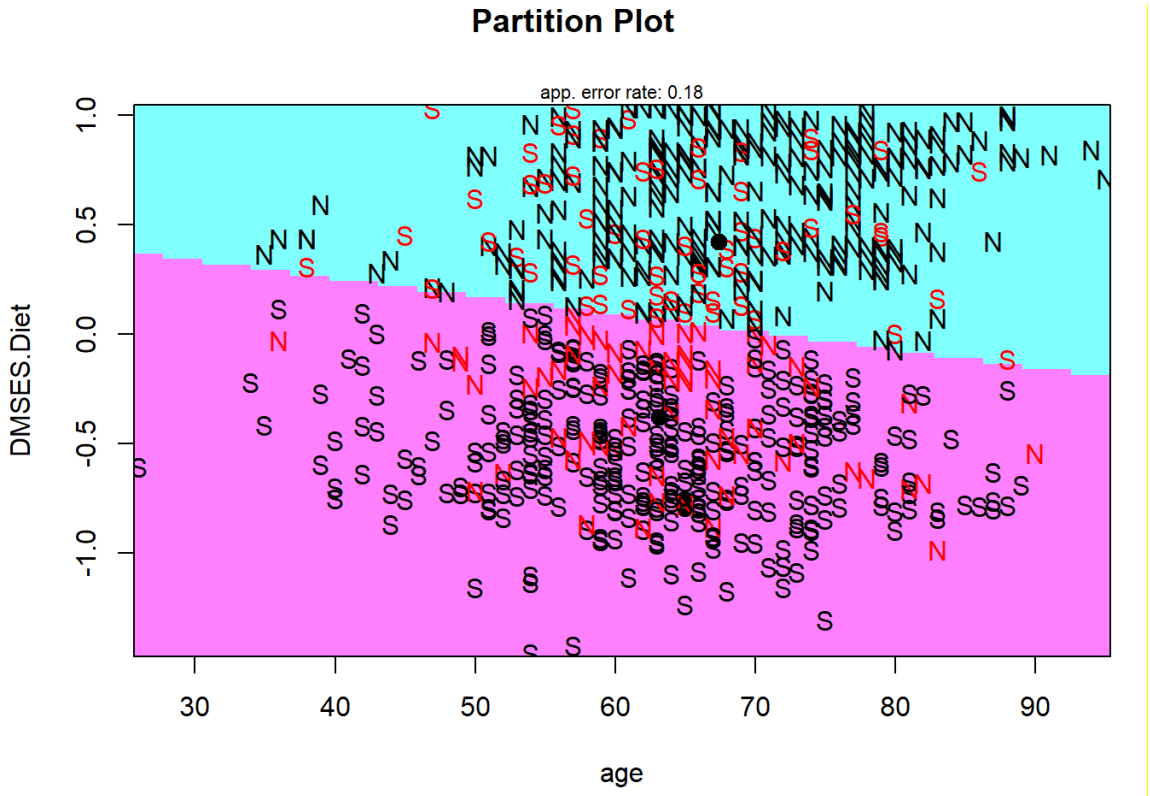


Figura 48: Partition Plot

6. Conclusiones

A lo largo de este trabajo final de máster se ha podido comprobar que una dieta adecuada y un ejercicio moderado, recogidos mediante el cuestionario DMSES, contribuyen a tener una diabetes controlada.

Se ha probado claramente que dichos cuestionarios ayudan a discriminar con un alto grado de precisión a los pacientes según el control sobre la enfermedad. Además, durante la comparación de varios modelos, se ha comprobado que añadir otras variables sociodemográficas no incrementa la precisión de la clasificación.

En cambio, los modelos lineales creados para poder predecir el valor numérico de la hemoglobina glicosilada han resultado tener poca capacidad para explicar la variabilidad de los datos, obteniendo valores máximos del R^2 ajustado por debajo del 40%. Por tanto, las puntuaciones de los cuestionarios DMSES no son válidas para predecir los valores numéricos de la variable respuesta. Una razón para este pobre resultado podría deberse a los propios datos del cuestionario. Las preguntas son respondidas por los propios pacientes y éstos puede no responder con la respuesta cierta de manera voluntaria o involuntaria. También se podría estudiar la diferencia temporal de las covariables. No queda claro cuando fueron tomados todos los datos, especialmente los del cuestionario y el desfase temporal que pudieran tener con otras covariables.

Ya en el análisis multivariante, el análisis de componentes principales nos confirma que la concentración de hemoglobina glicosilada tiene una relación inversa con las puntuaciones del cuestionario DMSES y el análisis de conglomerados nos indica, los pacientes con niveles bajos de hemoglobina glicosilada tienen una menor distancia entre ellos, considerando las variables continuas del conjunto de datos, formando un conglomerado.

Todo el código de R utilizado para realizar esta memoria se puede consultar en https://github.com/ehm2411/TFM.

7. Glosario

BMI:	Body Mass Index – índice de masa corporal
DMSES:	Diabetes Management Self-Efficacy Scale
HBA1c:	Hemoglobina glicosilada
HDL:	High density lipoprotein cholesterol
LDL:	Low density lipoprotein cholesterol
T2DM:	Diabetes Mellitus tipo II

8. Bibliografía

1. D.Whiting, L. Guariguata, C. Weil, and J. Shaw, “IDF diabetesatlas: global estimates of the prevalence of diabetes for 2011and 2030,” Diabetes Research and Clinical Practice, vol. 94,pp. 311–321, 2011
2. Annemaire Mol “What diagnostic devices do: the case of blood sugar measurement.” Theoretical Medicine and Bioethics volume 21, pages 9–22(2000)
3. https://www.fundaciondiabetes.org/infantil/185/bomba-de-insulina-ninos
4. International Diabetes Federation (IDF), “Update of mortalityattributable to diabetes for the IDF diabetes atlas: estimates forthe year 2013,” Diabetes Research and Clinical Practice, vol. 109, no. 3, pp. 461–465, 2015.
5. J.J. Van Der Bijl, A. Van Poelgeest-Eeltink, and L. Shortridge-Baggett, “The psychometric properties of the diabetes management self-efficacy scale for patients with type 2 diabetes mellitus,” Journal of Advanced Nursing, vol. 30, no. 2, pp. 352–359, 1999.
6. Messina R; Rucci P; Sturt J; Mancini T; Fantini MP “Assessing self-efficacy in type 2 diabetes management: validation of the Italian version of the Diabetes Management Self-Efficacy Scale (IT-DMSES).” Academic Journal | Publisher: BioMed Central
7. Fappa E; Efthymiou V; Landis G; Rentoumis A; Doupis J “Validation of the Greek Version of the Diabetes Management Self-Efficacy Scale (GR-DMSES).” Academic Journal | Publisher: Health Communications Inc
8. Lee, Eun-Hyun; van der Bijl, Jaap; Shortridge-Baggett, Lillie M.; Han, Seung J. “Psychometric Properties of the Diabetes Management Self-Efficacy Scale in Korean Patients with Type 2 Diabetes.” Academic Journal | International Journal of Endocrinology; 5/18/2015, Vol. 2015, p1-9, 9p
9. Hürmeydan, Özlem; Madenci, Özlem Çakır; Yildiz, Zeynep; Orçun, Asuman; Yücel, N. “Diagnostic performance of HbA1c for detecting prediabetes and diabetes in Turkish adults” Academic Journal | International Journal of Diabetes in Developing Countries: Incorporating Diabetes Bulletin. 40(4):585-590
10. https://medlineplus.gov/spanish/a1c.html
11. Luis, Daniel A. de; Izaola, Olatz; Primo, David; García Calvo, Susana; Gómez Ho... “Relación de la variante rs1800777 del gen CETP (proteína transportadora de ésteres de colesterol) con la masa grasa y HDL colesterol, en sujetos obesos con diabetes mellitus tipo 2 /...”Academic Journal | Nutrición Hospitalaria. December 2017 34(6):1328-1332
12. Elizabeth Woolley. “10 Causes of High Triglycerides in Diabetes” https://www.verywellhealth.com/what-causes-high-triglycerides-in-diabetes-1087722

9. Anexos

9.1 Cuestionario DMSES

Diet (9 items)	
(i) DMSES4	I can choose to eat good and healthy foods that are beneficial to my health
(ii) DMSES5	I can choose to eat various foods to maintain a healthy diet plan
(iii) DMSES9	I can maintain a healthy diet plan in the event that I get sick
(iv) DMSES10	I can follow a healthy diet plan regularly
(v) DMSES13	I can follow a healthy diet plan even when I am not at home
(vi) DMSES14	I can choose from various foods to maintain a healthy diet plan when I am not at home
(vii) DMSES15	I can follow a healthy diet plan during festivals, traditions, or rituals
(viii) DMSES16	I can choose to eat various foods to maintain a healthy diet plan when I eat foods at parties
(ix) DMSES17	I can maintain a healthy diet plan when I am feeling stressed or worried
Monitor (4 items)	
(i) DMSES1	I can check blood glucose levels by myself if necessary
(ii) DMSES2	I can reduce blood glucose levels when glucose levels in my blood are too high (e.g., changing the kinds of foods I eat)
(iii) DMSES3	I can increase blood glucose levels when glucose levels in my blood are too low (e.g., changing the kinds of foods I eat)
(iv) DMSES7	I can attend to my feet (e.g., cutting toenails and taking care of myself not causing wounds)
Physical (4 items)	
(i) DMSES6	I can control my body weight and maintain appropriate weight ranges
(ii) DMSES8	I can exercise and perform sufficient physical activity (e.g., walking, aerobic dancing, and muscle exercise)
(iii) DMSES11	I can increase the amount that I exercise if a doctor advises me to do so
(iv) DMSES12	In the case that I exercise more, I can modify my healthy diet plan
Regimen (3 items)	
(i) DMSES18	I can schedule an appointment to see a doctor four times a year to check my diabetes
(ii) DMSES19	I can take medicines as prescribed by a doctor
(iii) DMSES20	I can keep taking medicines continuously when I am sick

9.2 Detalle del análisis descriptivo de las variables.

```
##     id.code           hospid       gender          mstatus         age
##  1      :  1   Phupaman  : 60   Hombre:208   Soltero   : 57   Min.   :26.00
##  2      :  1   Srinakarin: 77   Mujer :492   ´Casado   :462   1st Qu.:59.00
##  3      :  1   Wegaronrat:241                Viudo     :165   Median :65.00
##  4      :  1   Chula     :322                Divorciado: 13   Mean   :65.16
##  5      :  1                                 Separado  :  3   3rd Qu.:73.00
##  6      :  1                                                  Max.   :95.00
##  (Other):694
##      edu               religion       income        dmdura      famhx
##  Nada  : 47   Budismo      :543   [0,5)  :318   Min.   : 4.00   No:370
##  Prim. :381   Islam        :152   [5,10) : 95   1st Qu.: 7.00   Sí:330
##  Sec.  :146   Cristianismo :  5   [10,15): 86   Median :10.00
##  Grado : 99   Otra religión:  0   [15,20): 48   Mean   :13.53
##  Master: 25                       [20,25): 48   3rd Qu.:20.00
##  Doctor:  2                      +25    :105   Max.   :45.00
##
##  comob    comlip   comht    comchd   comkid   comoth         dmrx
##  No: 42   No: 95   No: 97   No:662   No:512   No:699   Nada    : 12
##  Sí:658   Sí:605   Sí:603   Sí: 38   Sí:188   Sí:  1   Oral    :409
##                                                       Insulina: 94
##                                                       Ambos   :185
##
##
##
##         smk           alcohol        hba1c             ldl
##  Sí       : 23   Sí       : 42   Min.   : 2.000   Min.   :  8.7
##  Exfumador: 88   Exbebedor: 89   1st Qu.: 6.400   1st Qu.: 81.0
##  No       :589   No       :569   Median : 7.100   Median : 97.0
##                                  Mean   : 7.579   Mean   :100.8
##                                  3rd Qu.: 8.300   3rd Qu.:116.0
##                                  Max.   :15.000   Max.   :221.0
##
##       hdl              trig             sbp              dbp         compli
##  Min.   : 17.00   Min.   : 29.30   Min.   : 93.0   Min.   : 22.00   No:429
##  1st Qu.: 41.88   1st Qu.: 95.75   1st Qu.:123.0   1st Qu.: 65.00   Sí:271
##  Median : 50.00   Median :127.00   Median :133.0   Median : 73.00
##  Mean   : 50.75   Mean   :149.02   Mean   :134.8   Mean   : 73.02
##  3rd Qu.: 58.00   3rd Qu.:175.00   3rd Qu.:145.0   3rd Qu.: 80.00
##  Max.   :116.00   Max.   :999.00   Max.   :217.0   Max.   :112.00
##
##  cva      cereinfrac ishemic  stroke   cerebhem tia      angia    chf
##  No:699   No:700     No:688   No:698   No:700   No:699   No:700   No:692
##  Sí:  1   Sí:  0     Sí: 12   Sí:  2   Sí:  0   Sí:  1   Sí:  0   Sí:  8
##
##
##
##
##
##   mi      cororevas pad      neuropath renal     dn       dr      othcomp
##  No:675   No:699    No:696   No:695    No:582   No:619   No:561   No:699
##  Sí: 25   Sí:  1    Sí:  4   Sí:  5    Sí:118   Sí: 81   Sí:139   Sí:  1
##
##
##
##
##
##    DMSES.Diet       DMSES.Monitor      DMSES.Physical     DMSES.Regimen
##  Min.   :-1.45946   Min.   :-0.84225   Min.   :-1.28709   Min.   :-2.86027
##  1st Qu.:-0.53693   1st Qu.:-0.23065   1st Qu.:-0.28906   1st Qu.: 0.07856
##  Median :-0.06537   Median :-0.03718   Median : 0.01757   Median : 0.09450
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.52029   3rd Qu.: 0.24502   3rd Qu.: 0.29446   3rd Qu.: 0.10978
##  Max.   : 1.03434   Max.   : 0.55056   Max.   : 0.79336   Max.   : 0.51344
```

```
##
##   DMSES.Total          DK.10        alerta        bmi        sobrepeso obesidad
##  Min.   :-5.6289   Min.   : 0.000   No:333   Min.   :14.31   No:301    No:518
##  1st Qu.:-1.0268   1st Qu.: 4.224   Sí:367   1st Qu.:22.91   Sí:399    Sí:182
##  Median :-0.0258   Median : 5.486            Median :25.96
##  Mean   : 0.0000   Mean   : 5.535            Mean   :27.08
##  3rd Qu.: 1.0416   3rd Qu.: 6.919            3rd Qu.:30.14
##  Max.   : 2.5059   Max.   :10.000            Max.   :76.03
##
##   fecha.hba1c           fecha.ldl            fecha.hdl
##  Min.   :2006-02-01   Min.   :2012-11-08   Min.   :2006-04-23
##  1st Qu.:2016-03-01   1st Qu.:2016-02-04   1st Qu.:2016-01-13
##  Median :2016-05-04   Median :2016-04-21   Median :2016-03-29
##  Mean   :2016-03-07   Mean   :2016-02-19   Mean   :2016-01-26
##  3rd Qu.:2016-06-07   3rd Qu.:2016-05-31   3rd Qu.:2016-05-31
##  Max.   :2016-07-04   Max.   :2016-09-03   Max.   :2016-12-09
##                       NA's   :1
##    fecha.trig           fecha.bp          age_out dmdura_out hba1c_out ldl_out
##  Min.   :2012-11-08   Min.   :2006-04-27   0:692   0:692      0:666     0:674
##  1st Qu.:2016-02-02   1st Qu.:2016-04-22   1:  8   1:  8      1: 34     1: 26
##  Median :2016-04-21   Median :2016-05-19
##  Mean   :2016-02-27   Mean   :2016-05-14
##  3rd Qu.:2016-05-31   3rd Qu.:2016-06-16
##  Max.   :2016-12-21   Max.   :2035-03-27
##                       NA's   :1
##  hdl_out trig_out sbp_out dbp_out DMSES.Diet_out DMSES.Monitor_out
##  0:683   0:665    0:684   0:687   0:700          0:700
##  1: 17   1: 35    1: 16   1: 13
##
##
##
##
##
##  DMSES.Physical_out DMSES.Regimen_out DMSES.Total_out DK.10_out bmi_out
##  0:696              0:578             0:698           0:694     0:686
##  1:  4              1:122             1:  2           1:  6     1: 14
##
##
```

9.3 Otros modelos lineales estudiados.

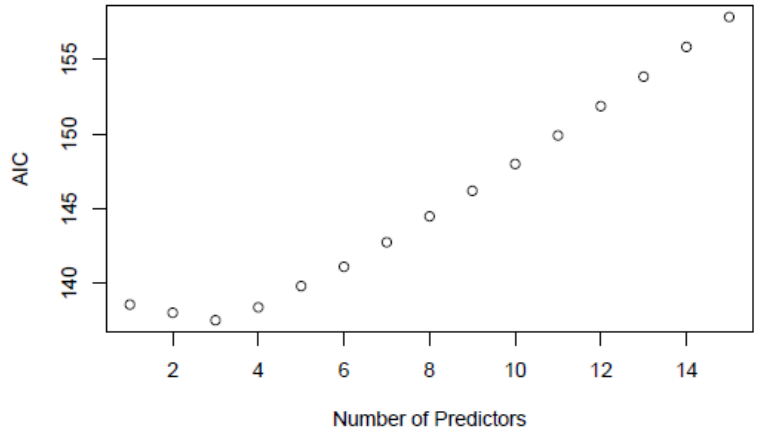
En el repositorio de Github hay otros modelos que se han probado, pero no se consigue mejores valores de R^2 ajustado.

En el siguiente cuadro se puede ver el intento de encontrar el modelo que obtenga un mayor R^2 ajustado.

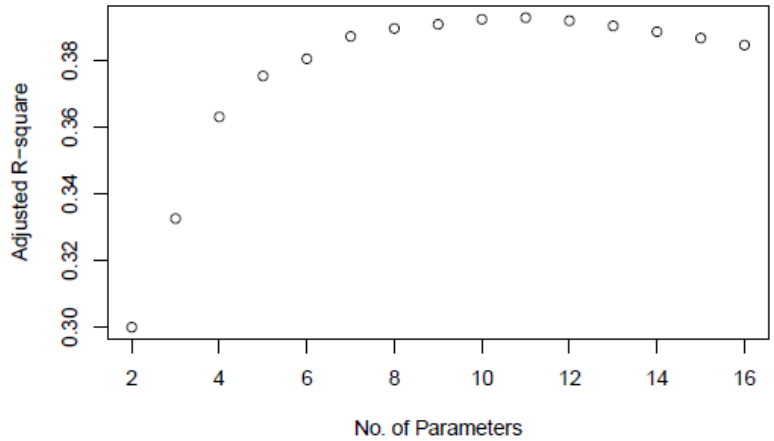
Primero probando que variables eran las más adecuadas para cada número de predictores.

```
#      (Intercept) obesidadSí sobrepesoSí genderMujer renalSí dmrx11 dmrx21 dmrx31  dnSí  drSí   age   hdl   sbp   dbp DMSES.Total DK.10
## 1         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE        TRUE FALSE
## 2         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE FALSE  TRUE FALSE FALSE FALSE FALSE        TRUE FALSE
## 3         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE  FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 4         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE   TRUE  FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE        TRUE FALSE
## 5         TRUE      FALSE       FALSE       FALSE   FALSE  FALSE  FALSE   TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 6         TRUE      FALSE        TRUE       FALSE   FALSE  FALSE   TRUE  FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 7         TRUE      FALSE        TRUE       FALSE   FALSE  FALSE  FALSE   TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 8         TRUE      FALSE        TRUE       FALSE   FALSE   TRUE   TRUE  FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 9         TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE  FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE        TRUE FALSE
## 10        TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE  FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 11        TRUE      FALSE        TRUE       FALSE    TRUE   TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 12        TRUE       TRUE        TRUE       FALSE    TRUE   TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 13        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE FALSE
## 14        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE        TRUE  TRUE
## 15        TRUE       TRUE        TRUE        TRUE    TRUE   TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE        TRUE  TRUE
```

Y luego comprobando el valor de AIC para cada combinación:



Y luego el R^2 ajustado.



Como se puede observar, el valor del R^2 ajustado apenas llega al 40% en el mejor de los casos.