ANÁLISIS Y VISUALIZACIÓN DE LOS FACTORES DETERMINANTES DEL ESTADO DE SALUD EN LA UNIÓN EUROPEA

Eduardo Rivero Falcón

Máster Universitario de Ciencia de Datos

Área: Ciencia de datos aplicada a la salud

Tutora: Susana Pérez Álvarez

Profesora responsable: Ángels Rius Gavidia

Enero 2021

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

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis y visualización de los factores determinantes del estado de salud en la Unión Europea*
Nombre del autor:	*Eduardo Rivero Falcón*
Nombre del consultor/a:	*Susana Pérez Álvarez*
Nombre del PRA:	*Ángels Rius Gavidia*
Fecha de entrega (mm/aaaa):	01/2021
Titulación::	*Máster Universitario en Ciencia de Datos*
Área del Trabajo Final:	*Ciencia de datos aplicada a la salud*
Idioma del trabajo:	*Español*
Palabras clave	*Factores determinantes de salud, Machine Learning, Visualización, EHIS, SILC.*

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

La Organización Mundial de la Salud (OMS) define salud como un estado de perfecto (completo) bienestar físico, mental y social, y no sólo la ausencia de enfermedad. Alcanzar este estado es objetivo natural del ser humano y de las administraciones públicas y gobiernos. Establecer cuáles son los factores determinantes de la salud y su incidencia ayuda a conseguir este objetivo.

La Unión Europea ha llevado a cabo la "European Health Interview Survey" (EHIS), cuya segunda oleada se realizó entre 2013 y 2015, actualizada en Abril de 2019 y la "European Union Statistics on Income and Living Conditions" (EU-SILC), con datos hasta 2013. En estas encuestas se recogieron entre otros datos el estado de salud de la población, costumbres de alimentación y consumo (alcohol, tabaco) y apoyo social en el ámbito de la Unión Europea.

Los objetivos de este Trabajo Fin de Máster son a partir de estos datos analizar el grado de influencia de los factores determinantes del estado de salud mediante técnicas de Regresión Logística, además de detectar posibles patrones entre los países mediante Análisis de Componentes Principales y técnicas de Agrupamiento (clusterización).

Es también objetivo la visualización de los resultados obtenidos de forma que permita una comparación en el ámbito de la Unión Europea y un mejor entendimiento de las conclusiones obtenidas.

Como conclusión se analizó la influencia de los factores determinantes detectados, así como la segmentación en grupos de los países con características afines en cuanto a las variables de estudio.

Abstract (in English, 250 words or less):

The World Health Organization (WHO) defines health as a state of perfect (complete) physical, mental and social well-being, and not just the absence of disease. Achieving this state is a natural goal of any individual, public administrations and governments. Establishing the health determinants and its impact helps to achieve this goal.

The European Union has carried out the “European Health Interview Survey” (EHIS), whose second wave was collected between 2013 and 2015 and updated in April 2019, and the “European Union Statistics on Income and Living Conditions” (EU-SILC, 2013). Among other data these surveys gathered the health status of the population, consumption of food and products harmful to health (alcohol, tobacco) and social support in the whole European Union.

The aims of this Master's degree final project are, based on this data, to analyse the level of influence of the health determinants using Logistic Regression techniques, as well as to detect any patterns between countries using Principal Component Analysis and Clustering techniques.

It is also the aim of this Master's degrees final project to visualize the results obtained in such way it could allow us to compare them at European Union level and thus to have a better understanding of the conclusions of this project.

To conclude, the influence of the determining factors detected was analysed, as well as the segmentation into groups of countries with similar characteristics in terms of the study variables.

ÍNDICE

1. Introducción 4

1.1 Contexto y justificación del Trabajo 4

1.1.1 Encuestas sobre salud. Eurostat. European Health Interview Survey (EHIS). European Union Statistics of Income and Living Conditions (EU-SILC)..... 4

1.1.2 Motivación personal..... 6

1.2 Objetivos del Trabajo..... 6

1.3 Enfoque y método seguido..... 7

1.4 Planificación del Trabajo..... 8

2. ESTADO DEL ARTE. 11

2.1 SALUD PERCIBIDA (SELF-PERCEIVED HEALTH, SPH) 11

2.2 RESUMEN DE LA BIBLIOGRAFÍA ANALIZADA: DETERMINANTES DEL SPH. 12

2.3 RESUMEN DE LA BIBLIOGRAFÍA ANALIZADA: ESTUDIOS COMPARATIVOS ENTRE PAÍSES. 14

2.4 CONCLUSIONES SOBRE EL ESTADO DEL ARTE. 17

3. CAPTURA Y PREPROCESADO. 19

3.1 Captura de los datos. 19

3.2 Preprocesado de los datos. 22

4. ANÁLISIS EXPLORATORIO ("EDA"). 27

4.1 Análisis exploratorio unidimensional. 27

4.2 Análisis exploratorio bidimensional: correlación entre las variables numéricas. 36

4.3 Análisis exploratorio bidimensional: distribución de las variables independientes según la variable dependiente. 39

5. ANÁLISIS DE LA INFLUENCIA DE LAS DISTINTAS VARIABLES: REGRESIÓN LOGÍSTICA. VALIDACIÓN DEL MODELO. 44

5.1 Regresión logística. 44

5.2 Validación del modelo: Devianza. 45

5.3 Validación del modelo: Multicolinealidad. 46

5.4 Validación del modelo: Métricas de clasificación. 47

5.5 Cálculo de los odds-ratio 48

6. ANÁLISIS DE LOS DATOS POR PAÍSES. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA). AGRUPAMIENTO. 49

6.1 Factor "F1". 49

6.2 Factor "F3". 50

6.3 Agrupación a partir de los componentes principales seleccionados (F1, F3). 51

7. CONCLUSIONES. 55

7.1 Análisis global de la influencia de las distintas variables en el contexto europeo. Regresión logística, Variables determinantes, valoración de los odds-ratio. 55

7.2 Análisis por países. Agrupamiento (clustering) a partir de los factores 'F1' y 'F3' del Análisis de componentes principales (PCA). 58

7.3 Visualización interactiva de los datos del proyecto (descriptivos y resultado del análisis) en Tableau. 63

7.4 Líneas de trabajo futuras. 64

8. ANEXO I. Correlación y pruebas de significación estadística de las variables independientes numéricas iniciales. 65

9. ANEXO II. Trazas de los procesos de selección del mejor modelo en la regresión logística... 68

9.1 Proceso de selección en la dirección 'backward'. 68

9.2 Proceso de selección en la dirección 'both'. 70

9.3 Proceso de selección en la dirección 'forward. 73

10. ANEXO III. Cálculo de los componentes principales (PCA) de los datos por país y gráficas de los pesos de cada componente. 74

11. ANEXO IV: GRÁFICAS DE AGRUPACIÓN (CLUSTERIZACIÓN) POR PAÍSES, CON DISTINTOS "LINKAGES" PROBADOS. 79

11.1 "LINKAGE COMPLETE" 79

11.2 "LINKAGE AVERAGE" 80

11.3 "LINKAGE WARD" 81

12. ANEXO V: CÓDIGO. 82

13. BIBLIOGRAFÍA. 83

LISTA DE FIGURAS

Figura 1: Página web de la oficina Europea de estadística (Eurostat). 5

Figura 2: Tablas de datos sobre la salud (Eurostat) 6

Figura 3: Esquema metodología CRISP-DM 7

Figura 4: DIAGRAMA de Gantt del proyecto. Fuente: propia 10

Figura 5: Gráfica situación de los países en función de los primeros factores 16

Figura 6: Árbol de directorios con las tablas de datos. 19

Figura 7: DIAGRAMA DE BLOQUES CAPTURA Y PREPROCESADO 25

Figura 8: Ejemplo de la evaluación de lectura de datos de R 27

Figura 9: ejemplo resumen estadístico R 27

Figura 10: distribucion variables categóricas "sex" y "age" 28

Figura 11: distribución variables del índice de masa corporal 29

Figura 12: distribución variables tipos de fumador 30

Figura 13: distribución de las variables 'actividad fisica' y 'bebedor habitual' 30

Figura 14: distribución variables consumo diario de frutas y verduras 32

Figura 15: distribucion variables apoyo social 32

Figura 16: distribución variables nivel de dolor crónico 33

Figura 17: distribución variables población con enfermedad crónica y gasto sanitario 34

Figura 18: distribución variables uso de medicinas 34

Figura 19: distribución variable objetivo 'sph buena o muy buena' 35

Figura 20: variables sobrepeso y actividad física por autovaloración de salud 39

Figura 21: variables consumo diario de frutas y verduras por autovaloracion de salud 40

Figura 22: variables consumo de tabaco por autovaloración de salud 40

Figura 23: variables apoyo social por autovaloración de salud 41

Figura 24: variables población con enfermedad crónica y uso de medicinas no prescribidas por autovaloración de salud 42

Figura 25: variables gasto sanitario y bebedores habituales por autovaloración positiva de salud 42

Figura 26: variables categóricas age y sex por autovaloración positiva de salud. 43

Figura 27: interacción variables age y sex 43

Figura 28: VALIDACIÓN, CURVA ROC 48

Figura 29: PESOS DEL FACTOR 'F1' DEL PCA 49

Figura 30: PESOS DEL FACTOR 'F3' DEL PCA 50

Figura 31: GRÁFICA DE SITUACIÓN DE LOS PAÍSES SEGÚN SUS FACTORES 'F1' Y 'F3' 51

Figura 32: DENDOGRAMA LINKAGE 'COMPLETE' 52

Figura 33: DENDOGRAMA LINKAGE 'AVERAGE' 53

Figura 34: DENDOGRAMA LINKAGE 'WARD' 53

Figura 35: AGRUPACIÓN (CLUSTERING) PAÍSES FACTORES 'F1' Y 'F3' 55

Figura 36: análisis clustering por países 58

Figura 37: PCA PESOS FACTOR 1 75

Figura 38: PCA PESOS FACTOR 3 75

Figura 39: PCA PESOS FACTOR 2 75

Figura 40: PCA PESOS FACTOR 5 76

Figura 41: PCA PESOS FACTOR 4 76

Figura 42: PCA PESOS FACTOR 6 76

Figura 43: PCA PESOS FACTOR 7 77

Figura 44: PCA PESOS FACTOR 8 77

Figura 45: PCA PESOS FACTOR 9 77

Figura 46: PCA PESOS FACTOR 10 78

Figura 47: PCA PESOS FACTOR 11 78

Figura 48: PCA PESOS FACTOR 12 78

Figura 49: CLUSTERING LINKAGE "COMPLETE” 79

Figura 50: CLUSTERING LINKAGE "AVERAGE" 80

Figura 51: CLUSTERING LINKAGE "WARD" 81

LISTA DE TABLAS

Tabla 1: Planificación del proyecto. Fuente: propia 9

Tabla 2: Disposición de los datos en las tablas originales (eurostat) 20

Tabla 3: variables finales del dataset una vez terminado el preprocesado 26

Tabla 4: variables numéricas del dataset final, tras eliminar las variables correladas 38

Tabla 5: resumen mejor modelo regresión logística 45

Tabla 6: VALIDACIÓN, FACTOR DE INFLACIÓN DE LA VARIANZA 46

Tabla 7: VALIDACIÓN, MATRIZ DE CONFUSIÓN 47

Tabla 8: VALIDACIÓN, MÉTRICAS DE CLASIFICACIÓN (REGRESIÓN LOGÍSTICA) 47

Tabla 9: odds-ratio variables estadísticamente significativas 48

Tabla 10: dendogramas, asignación de países 53

Tabla 11: factores determinantes, análisis odds-ratio 55

1. Introducción

1.1 Contexto y justificación del Trabajo

La salud es un bien universal al que tienen derecho todas las personas. Como tal es recogido por nuestra Constitución en el artículo 43 del título 1 "De los derechos y deberes fundamentales". En el mismo artículo también dicta que *"Compete a los poderes públicos organizar y tutelar la salud pública a través de medidas preventivas y de las prestaciones y servicios necesarios".* (Congreso de los Diputados, 2003)

De forma similar se establece la salud como derecho en la Declaración Universal de los Derechos Humanos en su artículo 25:*"Toda persona tiene derecho a un nivel de vida adecuado que le asegure, así como a su familia, la salud y el bienestar (…)"* (Declaración Universal de los Derechos Humanos, 2020). En cuanto a la Unión Europea, su Carta de los derechos fundamentales sostiene en su artículo 35 que *"Toda persona tiene derecho a acceder a la prevención sanitaria y a beneficiarse de la atención sanitaria en las condiciones establecidas por las legislaciones y prácticas nacionales. Al definirse y ejecutarse todas las políticas y acciones de la Unión se garantizará un nivel elevado de protección de la salud humana"* (Agencia de la Unión Europea para los Derechos Fundamentales, 2007). Queda pues constatado que es papel de las distintas administraciones públicas y gobiernos garantizar el derecho a la salud de forma activa con medidas preventivas y políticas de promoción.

Para ello es cuestión de vital importancia establecer los factores determinantes de la salud y su grado de impacto, de forma que las políticas y campañas que se adopten desde la administración incidan sobre estos factores. Su conocimiento ayuda además a la planificación de la asignación de los recursos y del gasto público necesario.

1.1.1 Encuestas sobre salud. Eurostat. European Health Interview Survey (EHIS). European Union Statistics of Income and Living Conditions (EU-SILC).

Como indicador de la importancia de las encuestas públicas de salud, la Organización Mundial de la Salud (WHO, World Health Organization) actuando como referencia creó el proyecto EUROHIS (2003) para mejorar la cooperación internacional en el campo de las encuestas de salud, así como desarrollar instrumentos comunes que permitieran construir indicadores provenientes de estas encuestas.

La oficina de estadística de la Unión Europea (Eurostat) ha promovido en este sentido dos iniciativas:

- *European Union Statistics on Income and Living Conditions (EU-SILC),* que recoge datos de población adulta a partir de 16 años, entre otros el propio estado de salud auto percibido y dolencias o enfermedades crónicas. Los últimos datos fueron recogidos por los países en distintos años, del 2003 al 2013.

- *European Health Interview Survey (EHIS).*Los datos de su segunda oleada se recogieron entre 2013 y 2015, y la base de datos disponible fue actualizada en Abril de 2019. Recoge datos sobre actividad física, índice de obesidad, consumo de frutas y verduras, consumo de tabaco y alcohol y apoyo social, entre otros.

La oficina estadística de la Unión Europea *(Eurostat)* facilita el acceso público a estos datos (Eurostat, 2020).

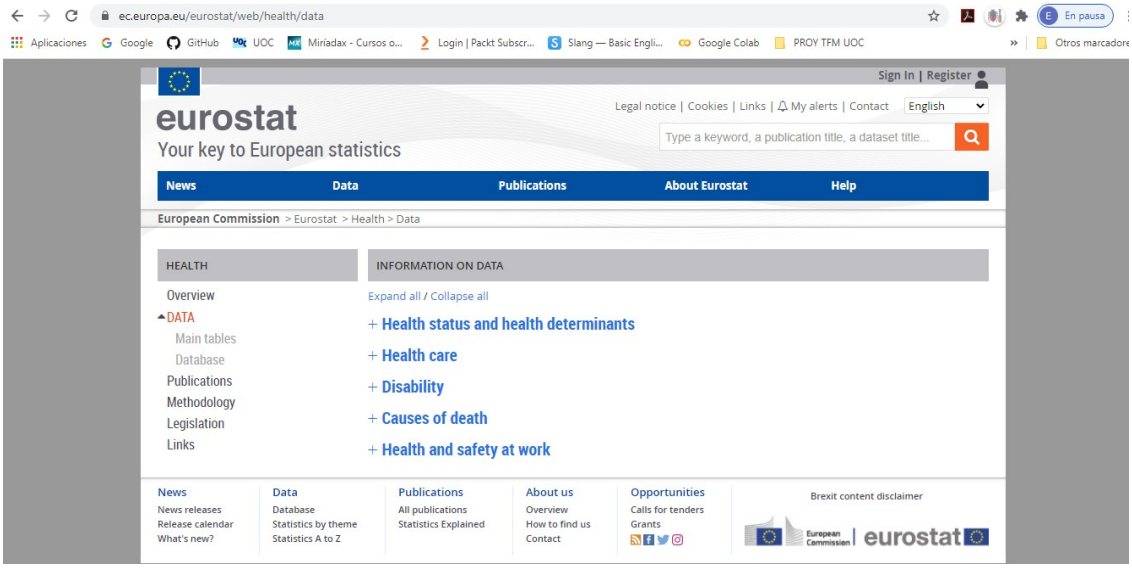

FIGURA 1: PÁGINA WEB DE LA OFICINA EUROPEA DE ESTADÍSTICA (EUROSTAT).

ESPACIO DE ACCESO A DATOS DEL ÁMBITO DE LA SALUD

La Eurostat permite acceder, a partir de diversos enlaces a las tablas que contienen la información. La misma página facilita también un explorador (inspección previa) de esos datos y la descarga de los mismos. Los datos están repartidos entre varias tablas (archivos CSV).

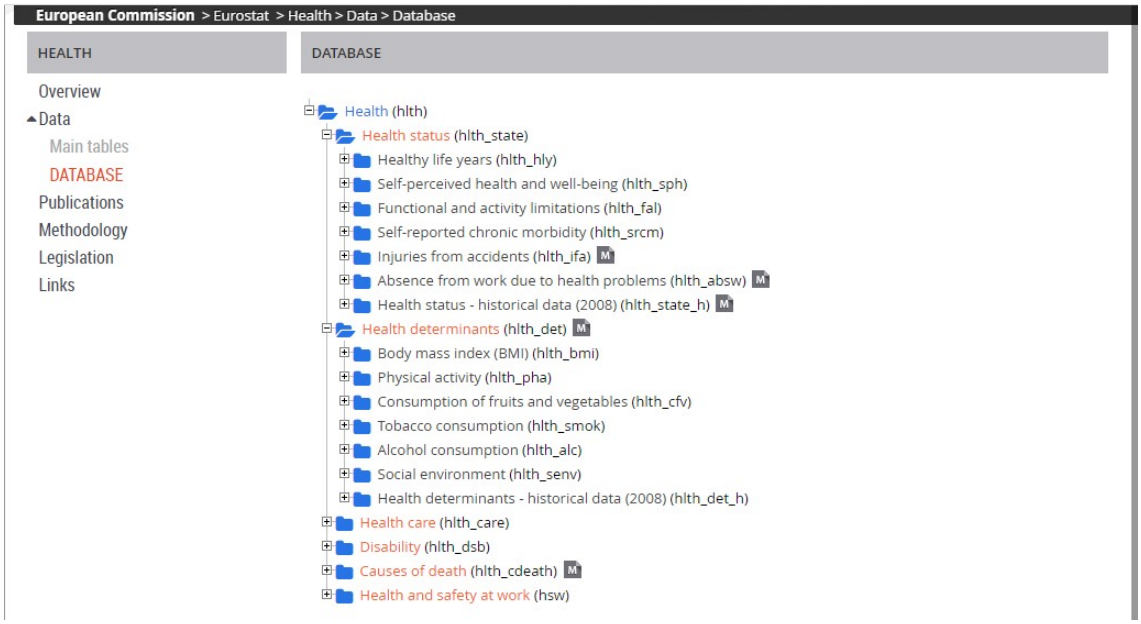

FIGURA 2: TABLAS DE DATOS SOBRE LA SALUD (EUROSTAT)

1.1.2 Motivación personal.

Entre las numerosas aplicaciones de la Ciencia de Datos siempre me han llamado la atención las relativas a la salud por sus repercusiones y beneficios. Conclusiones como las que pudieran salir de análisis como el de este Trabajo Fin de Máster son directamente aplicables y suponen un beneficio general para la sociedad.

Otra motivación a la hora de elegir esta línea de proyecto es enfrentarme a un escenario real en cuanto a la recopilación e integración de la información necesaria, así como aplicarlos métodos estadísticos y de aprendizaje automático aprendidos en el transcurso del Máster, pudiendo ampliar estos conocimientos para la consecución de los objetivos de este proyecto.

1.2 Objetivos del Trabajo

Los objetivos principales son:

- analizar el grado de influencia de los factores determinantes del estado de salud a nivel europeo, a partir de los datos facilitados por Eurostat.
- detectar posibles patrones entre estos factores y países, ayudándonos de técnicas de Machine Learning.

- presentar los resultados de manera gráfica, permitiendo una mejor comprensión a la vez que facilitar la comparación entre distintos países.

1.3 Enfoque y método seguido

En este proyecto utilizaremos la metodología de trabajo dictada por la Cross Industry Standard Process for Data Mining (CRISP-DM). Se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que se utilizan en este campo. Es prácticamente el estándar de facto para los proyectos de descubrimiento de conocimiento y el desarrollo de la minería de datos. Las seis fases de las que se compone son:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Despliegue.

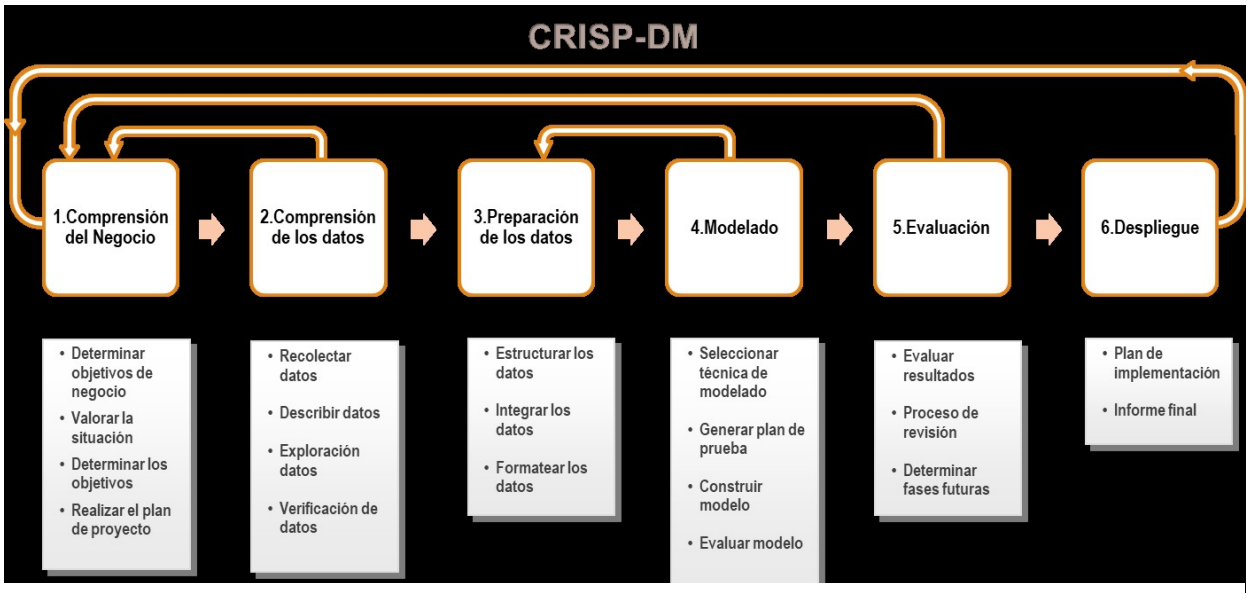

FIGURA 3: ESQUEMA METODOLOGÍA CRISP-DM

Como se puede ver la sucesión de fases no es necesariamente rígida, pudiendo volver en un determinado momento a fases anteriores en función de los resultados parciales, errores encontrados o previsión de no cumplimiento de los objetivos.

Partiendo de los datos proporcionados por Eurostat (archivos mencionados en el apartado 1.1.1), para la fase de preparación de los datos (más conocido por su acrónimo en inglés "*ETL*", de "*Extract, Transform, and Load"*) utilizaremos el lenguaje *Python* con las librerías "pandas" y "numpy". En la fase de análisis y modelado utilizaremos el lenguaje *R* por su mayor capacidad y número de librerías especializadas en métodos estadísticos. Para la visualización de los resultados utilizaremos el programa *Tableau*.

Dada la naturaleza de Trabajo Fin de Máster, en la última fase el informe final consiste en la redacción de la memoria, a la que se acompañará de una presentación de diapositivas.

1.4 Planificación del Trabajo

En este apartado se exponen cada una de las fases y subfases del trabajo delimitadas en tiempo por las fechas de inicio y final, con el objetivo de obtener como resultado final el adecuado. Esta planificación respeta las fechas indicadas para cada una de las PECs en el aula del Trabajo Fin de Máster.

En la siguiente tabla se reflejan los aspectos comentados.

TABLA 1: PLANIFICACIÓN DEL PROYECTO. FUENTE: PROPIA

Actividad	Descripción	Fecha de Inicio	Tiempo (días)	Fecha Final
1	Definición y planificación del TFM (PEC1)	16/09/2020	12	27/09/2020
1.1	Definición (título, objetivos, metodología) y planificación del TFM	16/09/2020	9	24/09/2020
1.2	Redacción de la PEC1	23/09/2020	5	27/09/2020
2	Estado del arte o análisis del mercado del proyecto (PEC2)	28/09/2020	21	18/10/2020
2.1	Revisión del estado del arte en análisis de los factores determinantes de la salud	28/09/2020	12	09/10/2020
2.2	Búsqueda de bibliografía relacionada	10/10/2020	5	14/10/2020
2.3	Redacción de la PEC2	12/10/2020	7	18/10/2020
3	Diseño e implementación del trabajo (PEC3)	19/10/2020	56	13/12/2020
3.1	Comprensión de los datos	19/10/2020	5	23/10/2020
3.1.1	Obtención de los datos: tablas EHIS	19/10/2020	2	20/10/2020
3.1.2	Exploración de los datos	21/10/2020	3	23/10/2020
3.2	Preparación de los datos	24/10/2020	14	06/11/2020
3.2.1	Limpieza, validación y selección	24/10/2020	3	26/10/2020
3.2.2	Integración de los datos desde las distintas tablas; Formateado	27/10/2020	6	01/11/2020
3.2.3	Análisis descriptivo de los datos	02/11/2020	5	06/11/2020
3.3	Modelado	07/11/2020	24	30/11/2020
3.3.1	Selección de técnicas de modelado	07/11/2020	4	10/11/2020
3.3.2	Generar plan de prueba	11/11/2020	3	13/11/2020
3.3.3	Construcción de los modelos	14/11/2020	17	30/11/2020
3.4	Evaluación e interpretación de los resultados	30/11/2020	14	13/12/2020
3.5	Redacción de la PEC3	02/11/2020	42	13/12/2020
4	Redacción de la Memoria (PEC4)	14/12/2020	20	02/01/2021
4.1	Integración de las PECS 1,2 y 3 en la Memoria	14/12/2020	5	18/12/2020
4.2	Escritura de los apartados restantes de la Memoria	19/12/2020	10	28/12/2020
4.3	Revisión de la Memoria (PEC4)	29/12/2020	5	02/01/2021
5	Presentación y defensa del Proyecto (PEC5)	03/01/2021	8	10/01/2021
5.1	Realización de la presentación	03/01/2021	8	10/01/2021

El diagrama de Gantt correspondiente es el siguiente:

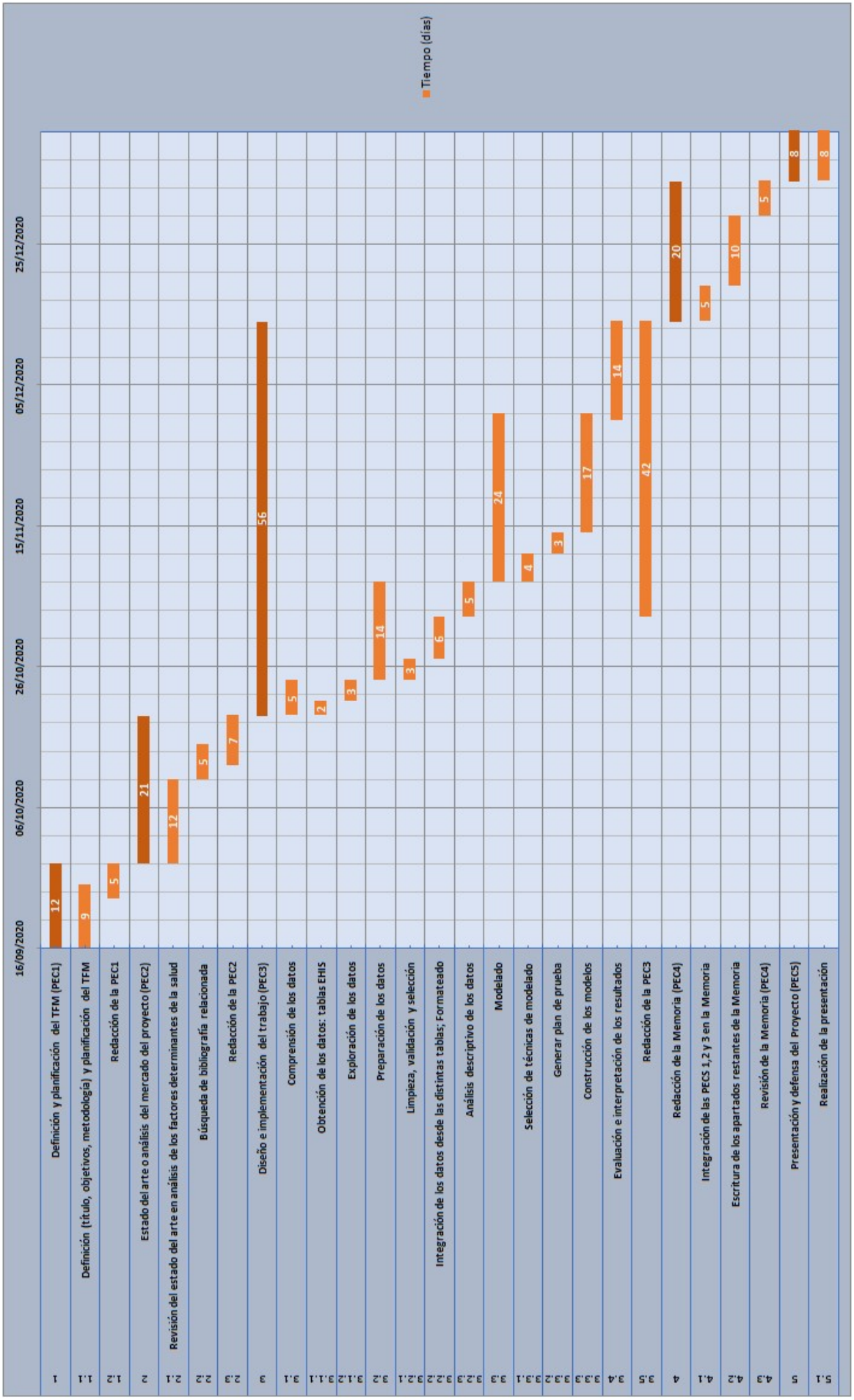

FIGURA 4: DIAGRAMA DE GANTT DEL PROYECTO. FUENTE: PROPIA

2. ESTADO DEL ARTE.

Para la exploración del actual Estado del Arte se ha realizado una búsqueda amplia por internet, utilizando entre otros los términos "health determinants" y "health status" por la temática de nuestro estudio, "self-rated health" por la variable objetivo de nuestro análisis, y "eurostat ehis" y "eurostat silc" para buscar posibles precedentes. También se utilizó algunos recomendados por Google Scholar como "cross-country comparison".

Cabe señalar un mayoritario uso de la salud percibida, nombrada como Self-Perceived Health (SPH) o Self-Rated Health (SRH), como indicador del estado de salud, frente a las más tradicionales de mortalidad y morbilidad.

A continuación, se resume brevemente la situación actual así como las observaciones obtenidas para la evaluación de percepción de salud así como las comparativas entre países.

2.1 SALUD PERCIBIDA (SELF-PERCEIVED HEALTH, SPH)

Las dos denominaciones de la salud auto percibida encontradas en los resultados de nuestra búsqueda (SPH/SRH) se refieren al mismo concepto, respuesta del individuo a la pregunta "¿cómo diría usted que es su estado actual de salud?", con 5 (en algún de las referencias encontradas 6) posibles respuestas, graduadas en niveles desde "Bad" hasta "Very Good". De aquí en adelante adoptaremos el término SPH por ser el utilizado por Eurostat.

El uso de la SPH como indicador de salud es ampliamente justificado en las referencias encontradas. Citando algunas entre ellas (Daviña, 2010) señala que "La relevancia del indicador de salud percibida, sobre todo, en países desarrollados, se debe a dos razones fundamentales: en primer lugar, además de los indicadores de mortalidad y morbilidad, son necesarios otros indicadores que proporcionen información adicional sobre la salud de la población en un contexto de aumento de las enfermedades crónicas y descenso de la mortalidad; en segundo lugar, la salud percibida es un indicador que adopta un enfoque multidimensional (físico, emocional, social, etc.) y que está asociado con características sociodemográficas y con otros indicadores como la mortalidad y la morbilidad". El mismo autor también nos indica que "Hay numerosos trabajos que han mostrado una asociación entre la salud percibida y otros indicadores de salud como la enfermedad crónica, discapacidad, supervivencia en personas mayores, mortalidad, etc."

En la misma línea (Cott, 1999) remarca que "La Salud Auto Percibida ha sido identificada como un importante indicador multidimensional del estado de salud. Las auto valoraciones del estado de salud se han encontrado como predictoras significativos de la mortalidad y morbilidad." Así mismo (Murata, 2006) afirma que "La utilidad de la Salud Auto Percibida en la predicción de mortalidad y decaimiento del estado de salud ha sido bien documentada en estudios realizados en países asiáticos y occidentales"

Destacar por último en (Fernández Martínez, 2012) se menciona que "el estado de Salud Auto Percibida es una medida ampliamente usada en muchos estudios y está estrechamente asociado con el estado de salud objetivo (…) Es un indicador útil de las necesidades de atención de salud en el diseño de programas y es muy sensible a factores sociales que causan desigualdades en el estado de salud."

En definitiva, la SPH es un indicador de salud multidimensional que tiene en cuenta además la situación socio-económica, emocional y el entorno del individuo, adecuado a la situación de los países desarrollados (baja mortalidad y alta morbilidad) y con una alta relación con éstos dos indicadores clásicos.

2.2 RESUMEN DE LA BIBLIOGRAFÍA ANALIZADA: DETERMINANTES DEL SPH.

Tomando pues como indicador la SPH encontramos en primer lugar varios estudios acerca de sus factores determinantes y el grado de influencia de los mismos, enfocados en países como España, Canadá y Japón.

En (Daviña, 2010) el autor parte de datos de la Encuesta Nacional de Salud (España) de 2006, con población a partir de los 16 años. De estos datos estudia la relación de factores con la SPH en cuatro dominios diferentes:

- Variables socio-demográficas: sexo, edad (categorizado en 11 grupos), clase social (según define la Sociedad Española de Epidemiología), nivel educativo (tres categorías) y situación laboral.

- Variables relacionadas con enfermedad e incapacidad: enfermedades crónicas, estado de salud mental (12-item General Health Questionnaire—GHQ-12), restricciones en las actividades diarias cotidianas, problemas de visión y audición.

- Estilo de vida: frecuencia en los hábitos de tabaco y alcohol (últimos 12 meses), actividad física en el trabajo, actividad física en tiempo libre, Índice de Masa Corporal (IMC)

- Uso de los servicios de salud: tiempo desde la última consulta médica, si han estado hospitalizados los últimos 12 meses, tiempo desde la última visita al dentista, número de medicinas consumidas en las últimas dos semanas.

La variable dependiente SPH (con valores posibles "Muy bueno", "Bueno", "Regular", "Malo" y "Muy malo") fue recategorizada en dos niveles, Positivo ("Muy bueno", "Bueno") y Negativo (resto de valores).

Con estas variables fueron calculados los "Odds Ratios" (OR) para SPH positiva en comparación con negativa, usando regresión logística multivariable. Se hizo previamente análisis univariante en relación con la SPH para la selección de las variables explicativas. El estudio se realizó para el total de la población y posteriormente diferenciando dos grupos de edad, 16–49 y 50 años en adelante.)

Como conclusiones generales se encuentra un empeoramiento de la SPH con la edad hasta los 49 años, con comportamiento inverso a partir de los 50 (mejor a mayor edad). Así mismo se observó en el primer grupo de edad que la obesidad, tabaco y alcohol están asociados con peor SPH, mientras que en el segundo grupo son más determinantes el ejercicio físico en tiempo libre y suficientes horas de sueño en la mejora de SPH.

En (Cott, 1999) el autor toma los datos de la "National Population Health Survey interview" (Canadá, 1994/95), con individuos a partir de los 20 años. Hay que destacar que a pesar de centrarse en población con padecimiento crónico Cott afirma en el artículo que esto no supone posible sesgo.

Los factores incluidos en el análisis fueron enfermedad crónica relatada, discapacidad a largo y corto plazo, variables demográficas (edad, estado civil, género, educación, estado de empleo e ingresos), estilo de vida (fumar, actividad física, bebida) estado psicológico (estrés crónico, angustia, autoestima) y apoyo social. Se utilizó regresión logística nominal multivariante para el análisis de los determinantes de SPH excelente en comparación con mala y buena.

En sus conclusiones las variables relativas a enfermedad crónica se asociaron con mala SPH, con contribuciones más pequeñas pero significativas de factores demográficos y estilo de vida. El estado sicológico, en especial la autoestima, se asocian con una mejor SPH en las personas con enfermedades crónicas o discapacidad.

Por otro lado (Murata, 2006) estudia la población anciana (a partir de 65 años) de Matsukawa, una ciudad en el sur de la prefectura de Nagano, en el centro de Japón. Tiene como objetivos el análisis de la SPH y mortalidad. Para la primera, entre las variables determinantes se encuentran enfermedad crónica, decaimiento del estado físico, apoyo social (amigos, familia), ingresos económicos y sensación de aislamiento. Para evaluar el impacto de estas variables realiza un análisis de regresión logística, con la SPH dicotomizada en BUENA para valores "buena/media" del estudio, y POBRE para valores "escasa/pobre".

Los resultados finales del mismo indican que factores psicosociales como el aislamiento se asociaron con una SPH pobre entre las mujeres. Entre los hombres la situación económica es más determinante. Para ambos sexos como era de esperar se encuentra relación con las variables relativas a la salud física.

En otro artículo (Fernández Martínez, 2012) se parte de una encuesta de calidad de vida en una muestra representativa de la población española a partir de 60 años en 2008.

Las variables independientes incluyen sociodemográficas (edad, sexo, estado civil, nivel de ingresos, región de residencia) y de estado de salud (enfermedad crónica según escala "Cumulative Illness Rating Scale (CIRS) for geriatrics", independencia en la realización de las tareas diarias según el índice de Barthel, depresión reportada según la escala "Hospital Anxiety and Depression Scale" (HADS-D). En este caso, se encontró una alta proporción de datos faltantes ("missings") de nivel de ingresos, situación solventada tomando un análisis de componentes principales (PCA) de situación económica auto percibida, satisfacción con la

situación económica actual y seguridad con la futura en las sucesivas evaluaciones económicas.

Fueron utilizados modelos de regresión logística binaria para explicar el efecto sobre la SPH de las variables anteriormente mencionadas. En la selección de variables se usó un enfoque "forward stepwise" partiendo de un modelo vacío y añadiendo variables en orden de mayor correlación con la variable dependiente.

Como conclusiones del estudio los encuestados más jóvenes eran más propensos a tener una SPH positiva. Con independencia de la edad la falta de condiciones crónicas, independencia en la realización de actividades de vida diaria y menor nivel de depresión también se asociaron con SPH positiva. Los residentes en el sur de España mostraron SPH más negativo que los de otras regiones. El sexo y nivel educativo en combinación de los otros factores demostraron ser no significativos. La dependencia funcional y la depresión afectan de manera negativa entre los adultos mayores independientemente de los factores socios demográficos y económicos.

Por último y como aportación de estudios enfocados sólo en la salud mental, en (Arias de la Torre, 2018) se utilizan datos del EHIS (Eurostat). Con variables demográficas (edad, sexo, etc.) y socioeconómicas (nivel educativo, lugar de residencia, estado laboral, etc.) investiga el efecto de estas variables sobre el estado de salud mental según escala "DSM-IV Major Depressive Disorder (MDD, a 8-item version of the Patients Health Questionnaire."

Similar a los otros estudios mencionados analiza los valores OddsRatio (OR) calculados de modelos de regresión logística con las variables mencionadas. Los resultados arrojaron una mayor prevalencia de MDD (depresión) en mujeres, fuertemente relacionada con la situación personal y desventajas socioeconómicas.

2.3 RESUMEN DE LA BIBLIOGRAFÍA ANALIZADA: ESTUDIOS COMPARATIVOS ENTRE PAÍSES.

Los datos facilitados por Eurostat están dispuestos por países, dando por tanto oportunidad para realizar estudios comparativos. En la búsqueda de bibliografía encontramos las siguientes referencias de interés.

En (Schmidt, 2002) se estudia el impacto de factores determinantes en la salud física y mental en 10 países (United Kingdom, France, Germany, Lithuania, Latvia, the Czech Republic, Croatia, Romania, Slovakia and Israel) con un total de 4849 encuestados, datos provenientes del proyecto EUROHIS.

Las variables explicativas escogidas fueron condiciones crónicas (asma, diabetes, hipertensión), apoyo social medido por la escala Oslo, consumo de alcohol (volumen diario y mensual), actividad física (número de días y tiempo empleado), uso de medicinas prescritas y uso de los servicios médicos (preventivos y curativos).

Como variables dependientes se tomaron el índice de salud mental MHI5 y un indicador de salud general (GENERAL HEALTH, similar al SPH, con rango 1-5). Para identificar los factores

determinantes se emplearon modelos de regresión lineal múltiple, analizando sus coeficientes. Primero fueron analizados los datos de todos los países conjuntamente como aproximación universal, y posteriormente por grupos de países para explorar posibles efectos culturales.

En sus conclusiones el estudio encontró como factores con contribuciones significativas a una salud general más pobre fueron mayor edad, presencia de enfermedades crónicas y el uso de servicios médicos curativos. Las variables que pronosticaban una mejor salud general fueron la educación superior, el empleo, el uso de servicios médicos preventivos, actividad vigorosa regular y un consumo menor de alcohol.

Señalar que hubo un efecto significativo en la división de países en grupos. En el grupo de los países de Europa Occidental hay asociaciones más directas entre las condiciones crónicas, características sociodemográficas y actividad vigorosa con la salud general, mientras que en el este los aspectos emocionales e interpersonales parecen ejercer un efecto sobre la salud general.

A continuación, se describen dos estudios con una metodología similar. En el primero de ellos (Pacakova, 2018) se parte de la base de datos online de la "OECD Health Statistics 2017" (Organisation for Economic Co-operation and Development), con datos disponibles hasta 2015 de 25 países europeos.

Entre las variables independientes escogidas se encuentran esperanza de vida, esperanza de vida saludable, mortalidad por problemas circulatorios, mortalidad por cáncer, gasto per cápita en salud, número de empleados en salud y sistemas sociales, número de consultas anuales por doctor, camas de hospital por mil habitantes y estancia media anual de los pacientes en hospitales.

Como variable dependiente se escogió la denominada "Perceived health status" (PHS), de concepto muy similar a la SPH. Se realizó un estudio para una PHS Buena/muy buena, y otro para mala/muy mala.

Partiendo de la matriz formada por los 25 países seleccionados y las variables anteriormente mencionadas se le realiza un análisis de componentes principales (conocido por su acrónimo en inglés, "PCA"). En este estudio calculan las cargas de las variables iniciales sobre los ejes resultantes de la descomposición ("loadings", pesos de las variables en los nuevos ejes). A partir de ellos y según en cada caso las variables iniciales con mayor peso de cada factor, se dota a los primeros factores resultantes de la descomposición de un determinado sentido mayoritariamente positivo y negativo para determinadas métricas como el buen cuidado de la salud o mortalidad. De esta forma se les otorga cierto significado a estos primeros factores resultantes de la PCA. El estudio realiza una comparativa a nivel gráfico de los componentes (factores F1 a F4) por país. Al mismo tiempo estudia la correlación entre estos factores y la variable dependiente. El estudio confirma en las conclusiones relaciones causales significativas entre el estado de salud, los gastos sanitarios y los servicios de cuidado de salud. También indican desigualdades significativas en los resultados de salud de los países europeos vigilados.

Como ejemplo mostramos a continuación **una de las gráficas del estudio** que ilustra la posición de los países según los **factoresF1 (interpretado por los autores, en la forma**

comentada en el párrafo anterior, como buenas condiciones de cuidado de salud) y F2 (interpretado como mal estado de salud):

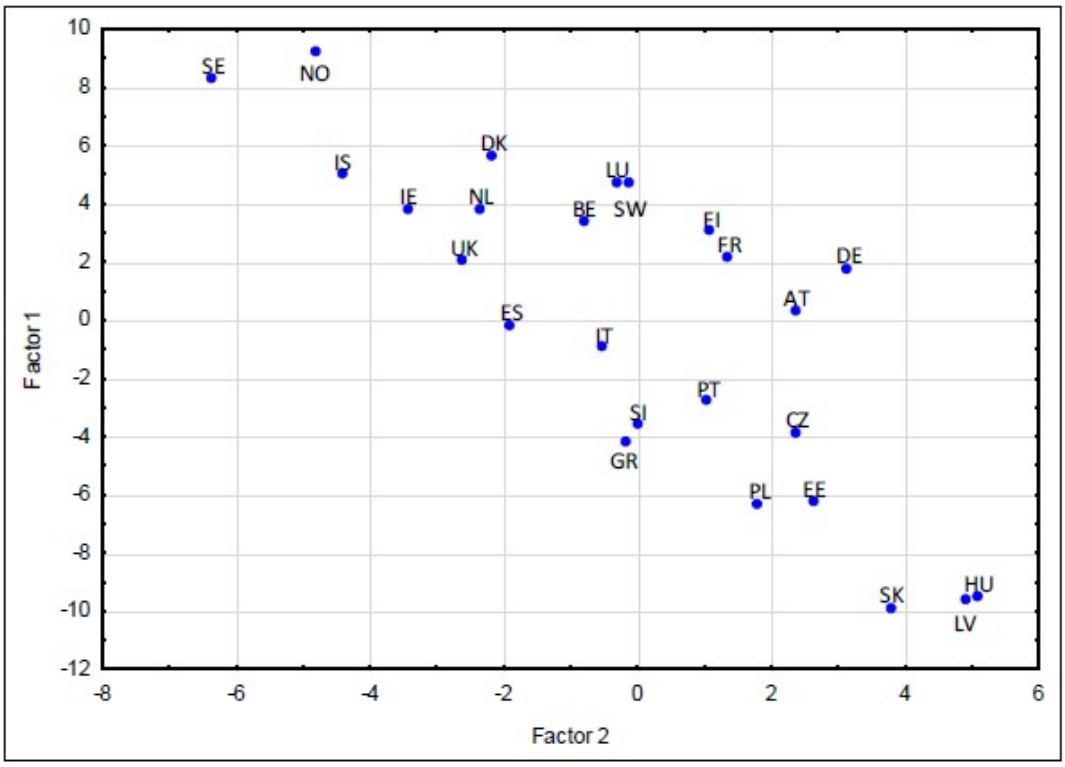

FIGURA 5: GRÁFICA SITUACIÓN DE LOS PAÍSES EN FUNCIÓN DE LOS PRIMEROS FACTORES DE LA DESCOMPOSICIÓN PCA. FUENTE: ARTÍCULO CITADO

En el segundo estudio (Jindrova, 2016) se trabaja con datos de la WHO, en concreto 53 países de la "WHO European Region". Para el estudio se han seleccionado todos los países de la Unión Europea completado con Noruega y Suiza, 30 países en total.

Entre los factores a estudiar seleccionados se encuentra información económica (gasto per cápita en salud, producto interior bruto (GDP)), estado de salud (esperanza de vida en el nacimiento y a los 65 años), padecimiento de enfermedades severas (sistema circulatorio, corazón, cerebro-vasculares), factores de riesgo (situación de sobrepeso, tanto por ciento de fumadores regulares, consumo de alcohol) y recursos del cuidado de la salud (camas de hospital por cien mil habitantes, media de estancia de los pacientes en hospital). La variable a explicar es el nivel de cuidado sanitario de los países.

En la primera fase se realiza al igual que en el anterior estudio un análisis de componentes principales, con estudio y representación gráfica de los factores resultantes de la misma. La diferencia en este caso es que con las variables iniciales mencionadas por países implementa posteriormente la técnica de clustering aglomerativo, utilizando como "linkage" (definición de distancia entre dos clústeres) el método "Ward".

En el dendograma resultante se observa un clúster formado por países con antigüedad en la UE (Francia, Alemania, Dinamarca etc.) junto con Noruega y Suiza unido a países occidentales con menor antigüedad (España, Grecia, Chipre). En el otro clúster se encuentran países del este (antigua influencia soviética, Polonia, República Checa, Hungría) y países bálticos (Estonia, Letonia y Lituania)

Como conclusiones la situación sanitaria en el grupo de los miembros antiguos y de antigüedad media mencionados anteriormente es notablemente diferente a la de los miembros nuevos, en perjuicio de éstos últimos.

Por último tenemos en (Rathmann, 2015) los datos fueron recogidos del "Health Behaviour in School-aged Children (HBSC) study" en 2006, que incluye adolescentes de 27 países de Europa y América del Norte, encuesta realizada en colaboración con la WHO.

Los factores determinantes estudiados fueron el producto interior bruto (GNP), el gasto sanitario en tanto por ciento del GNP, desigualdades en los ingresos económicos según el índice Gini y una variable categórica "tipo de régimen de bienestar" con valores "Social Democratic, Conservative, Liberal, Southern, and Eastern European".

La variable dependiente es una variable calculada a partir de un cuestionario a los adolescentes sobre diversos síntomas (quejas) percibidos en los últimos 6 meses. Los valores de esta variable que miden la frecuencia de estos síntomas son "diariamente", "varias veces por semana", "al menos una vez por semana", "alrededor de una vez al mes", "raramente o nunca". Una suma final indicando el número de síntomas con frecuencia mínima semanal forma el valor de la variable, con rango 0-8.

Con estas variables se realiza un análisis multivariante (regresión lineal), analizando el valor de los coeficientes resultado del mismo. Entre las conclusiones se observa que los adolescentes de países con mayor desigualdad de ingresos y con tradición liberal de bienestar se asociaron con más quejas de salud en comparación con los adolescentes de países con menor desigualdad de ingresos o el régimen socialdemócrata. Los países con mayor renta, gasto en salud pública y desigualdad de ingresos mostraron asociaciones más fuertes entre el estatus socioeconómico y las quejas de salud.

2.4 CONCLUSIONES SOBRE EL ESTADO DEL ARTE.

Las referencias encontradas en este apartado de estado del arte nos han mostrado los distintos enfoques, ámbitos geográficos y metodologías utilizadas hasta el momento en estudios relativos a la temática de nuestro TFM, los factores determinantes del estado de salud.

Hemos comprobado la idoneidad de la SPH como indicador global multidisciplinario del estado de salud y los distintos posibles factores que pueden influir en ella. Todo esto nos sirve para poder establecer un punto de partida para formar la estructura y metodología de nuestro proyecto.

Tomando como referencia todo lo anterior, los datos que vamos a utilizar en nuestro proyecto son los facilitados online de la base de datos de Eurostat sobre salud ya mencionados, con datos del EHIS y del EU-SILC. Entre ellos disponemos de datos por países sobre hábitos de consumo saludables (frutas y verduras), antropométricos (BMI), consumo de sustancias nocivas (alcohol y tabaco), actividad física realizada, entorno social (ayuda cercana disponible), uso de medicinas y gasto realizado en salud.

También se encuentran disponibles datos sobre la SPH por países. Disponemos pues de un amplio abanico de datos de partida para poder seleccionar las variables de nuestro estudio.

El plan y metodología a seguir para nuestro trabajo es el siguiente:

1.- Captura de los datos mencionados de la página web del Eurostat y preprocesado (limpieza, formato, etc.) de los mismos.

2.- Análisis previo exploratorio (EDA) de las variables de nuestro estudio.

3.- Realizar un análisis a nivel global (datos de todos los países) del SPH y sus determinantes, empleando modelos de regresión logística.

4.- Posterior aplicación PCA dado el número de variables, estudiar la "representatividad" de los factores resultantes de la reducción y comparar los valores para los distintos países.

5.- Aplicar técnicas de clustering para poder segmentar los países. Analizar los posibles grupos resultantes.

3. CAPTURA Y PREPROCESADO.

3.1 Captura de los datos.

El Eurostat pone a disposición pública datos de la Unión Europea en su página web. Los relativos a la salud para nuestro estudio se encuentran en la dirección ya comentada anteriormente (Eurostat, 2020). La información se reparte en numerosas tablas siguiendo un diseño multidimensional. La página posee una interfaz que permite inspeccionar los datos y conocerlos (metadata), así como seleccionar campos de interés y niveles de detalle de las dimensiones antes de su descarga. Pasamos a describir la disposición de los datos para situar en contexto las siguientes fases en el preprocesado que describiremos, necesarias para resolver ciertas situaciones con las que nos enfrentamos. Las tablas se presentan en un árbol de directorios tal como se ilustra en la imagen XX a continuación:

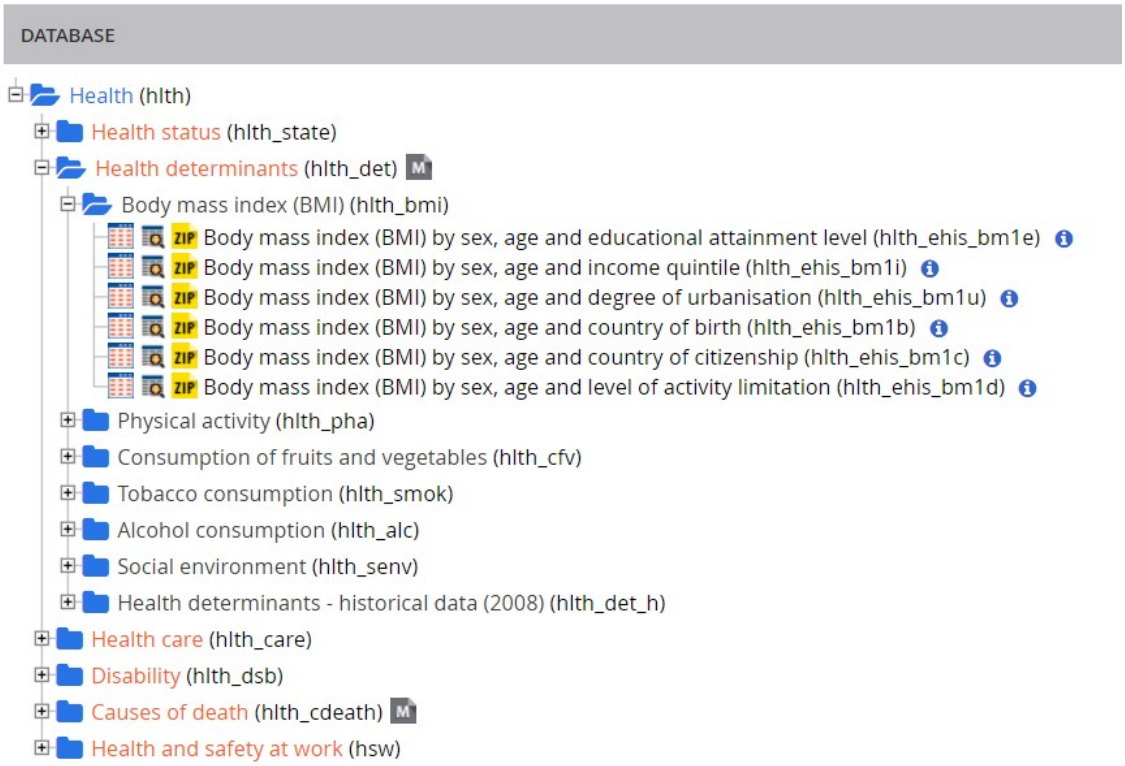

FIGURA 6: ÁRBOL DE DIRECTORIOS CON LAS TABLAS DE DATOS.

Podemos localizar los datos (provenientes de las encuestas comentadas EHIS y SILC) de las variables independientes y la dependiente cada una de ellas en su carpeta, conteniendo varias tablas. En cada una de estas tablas se encuentran los datos de la misma variable en cuestión (medida en porcentaje de población de esa franja bajo esa condición) y dividida por 3 dimensiones comunes: la geográfica (por país), la de género (masculino/femenino) y por grupos de edad. Al mismo tiempo se dividen entre los distintos niveles en los que se presenta la variable en cuestión. La diferencia entre las tablas se encuentra en una cuarta dimensión adicional, relativa a ingresos o nivel educativo, por poner algunos casos.

Para entender mejor todo lo anterior pongamos como ejemplo el de la figura anterior para la variable BMI (BodyMassIndex) que tiene como niveles (Underweight, Normal, Overweight). Fijando la cuarta dimensión en ingresos (Low, Middle, High), algunas de las observaciones (filas de las tablas) podrían ser como se muestra en la tabla a continuación:

TABLA 2: DISPOSICIÓN DE LOS DATOS EN LAS TABLAS ORIGINALES (EUROSTAT)

BMI	GEO	SEX	AGE	INCOME	VALUE	
….	……	…..	…..	…..	…..	
Overweight	Italy	Females	From 18 to 24 years	High	7.8	(*1)
Overweight	Italy	Females	From 25 to 34 years	High	12.5	
….	……	…..	…..	…..	…..	
Normal	Spain	Males	From 45 to 64 years	Low	21.9	(*2)
……	……	……	……	…..	…..	

(*1) En Italia, el grupo de población femenina entre 18 y 24 años de edad, con ingresos altos y en situación de sobrepeso representa el 7.8% de la población en esa franja.

(*2) En España, el grupo de población masculina entre 45 y 64 años de edad, con ingresos bajos y en situación de peso normal representa el 21.9% de la población en esa franja.

En primer lugar, debido a esta disposición de los datos sólo podíamos aprovechar como variable una entre las distintas opciones de la cuarta dimensión. Por ejemplo si quisiéramos haber tenido "Nivel de ingresos" y "Nivel educativo" no disponemos de los datos tal como nos lo presentan para las distintas combinaciones (entre los de nivel medio de ingresos como se reparten los porcentajes correspondientes a los distintos niveles educativos).

Al seleccionar por lo tanto una de las cuartas dimensiones disponibles, nos encontramos con la sorpresa de que las tablas de las distintas variables independientes no las tienen todas ni muchísimo menos, es más, solo hay una (nivel educativo) que está presente en todas y lo más importante, en las tablas de la variable dependiente SPH. Al elegir esta opción y analizar el conjunto de datos ("dataset") nos encontramos con una considerable cantidad de nulos en varias variables incluida la objetivo, nulos que abarcaban en casos a países enteros (no disponibilidad de datos en varios países.

Visto lo anterior, cualquier intento de imputación o eliminación de observaciones, desvirtuaría nuestro estudio por el gran volumen de observaciones faltantes. Por este motivo, **se tomó como decisión de diseño no seleccionar esa cuarta dimensión**, ya que además el número de variables disponibles en la página era más que suficiente.

Finalmente, los valores seleccionados para las tres dimensiones comunes según el nombre con el que se identifican en el conjunto de datos son los siguientes:

- **GEO** (31 países): Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Turkey, United Kingdom.

  En este caso, cabe señalar que fueron los 31 países presentes en la mayoría de las tablas, nos encontramos con variables en cuyas tablas no estaban ciertos países.

- **SEX** (2 grupos): Males; Females.

- **AGE** (5 grupos de edad):From 15 to 24 years; From 25 to 34 years; From 35 to 44 years; From 45 to 64 years; 65 years or over.

  Estos grupos de edad se establecieron para poder tener un número suficiente de observaciones en nuestro estudio, con el objetivo de tener después del preprocesado 31 países x 2 géneros x 5 grupos de edad (310 observaciones finales)

Con estos criterios, se descargaron las tablas de las variables independientes y dependientes:

- **Independientes**: Body Mass Index; Physical Activity; Daily Consumption of fruits and vegetables; Smoking of tobacco; Frequency of alcohol consumption; Overall perceived Social Support; Self reported use of Prescribed Medicines; Self reported use of Non-Prescribed Medicines; Health Care Expenditure; Self Reported Chronic Morbidity (People having a long-standing illness or health problem); Severity of bodily pain.

- **Dependiente**: Self-Perceived Health.

3.2 Preprocesado de los datos.

Esta fase se llevó a cabo en el lenguaje Python mediante "Jupyter notebooks" (a partir de ahora 'notebook'). La primera acción por acometer en el preprocesado es cambiar el formato de presentación de los datos. Las distintas categorías de la variable independiente (recuérdense por ejemplo, "Underweight", "Normal", "Overweight" para la variable "BMI") se encontraban en una columna categórica, con el valor correspondiente del porcentaje de población en esa franja afectada en otra columna titulada como "Value".

Había pues que realizar una acción de "pivotaje" para crear en su lugar una columna por categoría con sus valores correspondientes (3 columnas en el ejemplo comentado). Debido a que esta columna inicial de categorías añadía un valor por categoría a cada una de las combinaciones de país, género y grupo de edad, hubo que reconfigurar el formato de la tabla para tener sólo las combinaciones de estas últimas tres (las dimensiones comunes) y los valores de la variable resultado del pivotaje en su adecuado lugar.

Este proceso se implementa en el notebook de nombre **"automatiz_prx_indep_EHIS.ipynb"**. Para cada tabla original se obtenía como resultado una tabla con columnas las tres dimensiones comunes (GEO<país>, SEX y AGE) y tantas columnas como categorías de la variable en cuestión con sus correspondientes valores. En el notebook se implementa una función al que se pasa como parámetro un diccionario de tuplas (nombre_archivo_descarga_variable, nombre_columna_categorias, categorías_por_orden), que permite procesar variables por lotes.

Este proceso se pudo aplicar a la mayoría de las variables, pero hubo excepciones que necesitaban de añadidos o cambios en el preprocesado. Las señalamos a continuación:

- Variables con cambios en el preprocesado seguido: Son variables con una sola categoría, que no necesitaban de ciertas operaciones intermedias: "Self reported use of Prescribed Medicines", "Self reported use of Non-Prescribed Medicines" y "Self Reported Chronic Morbidity". Estas variables fueron procesadas por el notebook **"procesado_otros.ipynb"**

-Variables donde se detectaron después de los anteriores procesados datos faltantes: Fueron las correspondientes a "Physical Activity" y "Frequency of alcohol consumption", ambas tratadas en el notebook **"procesado_variables_datos_faltantes.ipynb"**. En ambos casos los datos faltantes se corresponden con pocos países y/o grupos de edad. Pasamos a describir el proceso para cada una. Al ser relativamente pocos casos se tomó como decisión buscar datos existentes de fuentes de confianza en internet. Se decidió esta opción frente a técnicas tradicionales de imputación de nulos (imputación por valores centrales, kNN) dado su bajo número y porque nos pareció más fiel para nuestro estudio. Se descartó también la posibilidad de eliminar por completo un país por la pérdida de información que si está disponible para el resto de variables.

Encontramos datos proporcionados por la Oficina Europea de la OMS, aunque no en la misma disposición de las tablas del estudio. Pasamos a describir cada caso:

A) "Physical Activity".

Datos faltantes para Bélgica, grupo "65 years or over" (ambos sexos) y datos no disponibles de Holanda. Los datos fueron tomados en este caso de las siguientes direcciones web:

- *https://www.euro.who.int/__data/assets/pdf_file/0011/382574/netherlands-eng.pdf*
- *https://www.euro.who.int/__data/assets/pdf_file/0010/382339/belgium-eng.pdf*

Señalar que no se pudieron encontrar datos para el 2014 (año de nuestro estudio). En la fuente estaban más cercanos los de 2018.

En el caso de Bélgica se disponían de los datos para el grupo de edad necesitados, pero no diferenciados por sexo. Para el reparto pues entre sexos se tomó la decisión de tomar un país del entorno, con similares extensiones y situación socio-económica, Dinamarca. En este caso en esa franja de edad los valores por sexo eran similares (diferencia de poco más de un punto porcentual. Por lo cual para Bélgica lo dividimos por igual entre ambos sexos.

Para Holanda los datos de la OMS si diferenciaban por género, pero en cuanto a los grupos de edad solo coincidía en el último (mayor de 65 años) mientras que los restantes de nuestro estudio los englobaba la fuente en una sola categoría. Se tomó también la decisión de tomar Dinamarca como país de referencia. En este caso se calcularon para los datos de Dinamarca las proporciones de reparto entre grupos de edad que nos faltaban para Holanda, de cada uno de los dos géneros. Conocidas estas proporciones se aplicaron sobre el total de la única categoría proporcionada por la fuente para repartirlas por los grupos de edad faltantes.

B) "Frequency of alcohol consumption".

Para esta variable faltan los datos completos de Holanda y Francia. Las fuentes de la OMS a las que se recurrieron son las siguientes:

- https://www.who.int/substance_abuse/publications/en/netherlands.pdf?ua=1
- https://www.who.int/substance_abuse/publications/global_alcohol_report/profiles/fra.pdf?ua=1

No se pudieron encontrar datos para el 2014 (año de nuestro estudio). En la fuente estaban más cercanos para 2016, para Holanda 2002.

Señalar en primer lugar que, en este caso que dada su diversidad, para que hubiera concordancia entre los datos de las fuentes (Consumed at least 60 grams/six glases or more of pure alcohol on at least one occasion in the past 30 days) y los de nuestra tabla

tuvimos que unir (sumar) los dos niveles de la variable, convirtiéndola en una única con el sentido de “bebedor habitual”

Los datos en este caso también diferenciaban por género, pero no por grupos de edad. Se tomaron esta vez como países de referencia para Holanda la misma Dinamarca, y para Francia Alemania. Al igual que en el caso anterior se calcularon las proporciones de reparto existentes entre los grupos de edad de los países de referencia, y se aplicaron posteriormente al total del dato de la fuente para repartirlo por los distintos grupos de edad faltantes.

Una vez procesadas todas las tablas y resueltas las incidencias acaecidas como hemos descrito, solo faltaba combinar las variables en un único dataset con las 3 dimensiones comunes. Tal acción se llevó a cabo en el notebook **“crear_dataset_final.ipynb”.** Para ello se implementó una función que tomaba como argumento un diccionario con los nombres de los archivos ya procesados en el formato adecuado y sin faltantes que se habían creado.

En este punto es necesario comentar que en la primera ejecución se produjo un error en la incorporación de las variables “Self Reported Chronic Morbidity” y en la dependiente “Self Perceived Health”. Para nuestra sorpresa en estas dos variables el primer grupo de edad era de “**16** a 24” mientras que en las restantes es de “**15** a 24”. Dada la escasa diferencia de un año, considerando el pequeño impacto en la población a evaluar tanto dentro de su franja como respecto a la totalidad, en aras de maximizar la información posible y una vez consultada la tutora para su aprobación, se decidió adecuar el nombre del primer grupo de edad en estas dos variables al de las 9 restantes.

Con esta corrección ya se pudo ejecutar de nuevo la función y se creó el dataset final, con dos últimas pequeñas modificaciones:

- Uniformidad en el número de decimales. Se encontraron diferencias en algunas variables. Siendo la unidad el porcentaje se decidió unificar el formato a 2 decimales.

- Nombre del país Alemania: este país en las tablas de Eurostat aparece como 'Germany (until 1990 former territory of the FRG)' para constatar suponemos históricamente la reunificación de este país en el contexto de la “caída del muro” y el desplome del bloque soviético.

  Por simplicidad y para poder ser un valor detectable como geográfico en la visualización de Tableau que forma parte de este Trabajo Fin de Máster, se deja sólo como ‘Germany’.

Presentamos a continuación el diagrama de bloques que resume las etapas de captura y preprocesado y la tabla con las variables del dataset final (archivo “**health_determ_EU_v0.csv”)**:

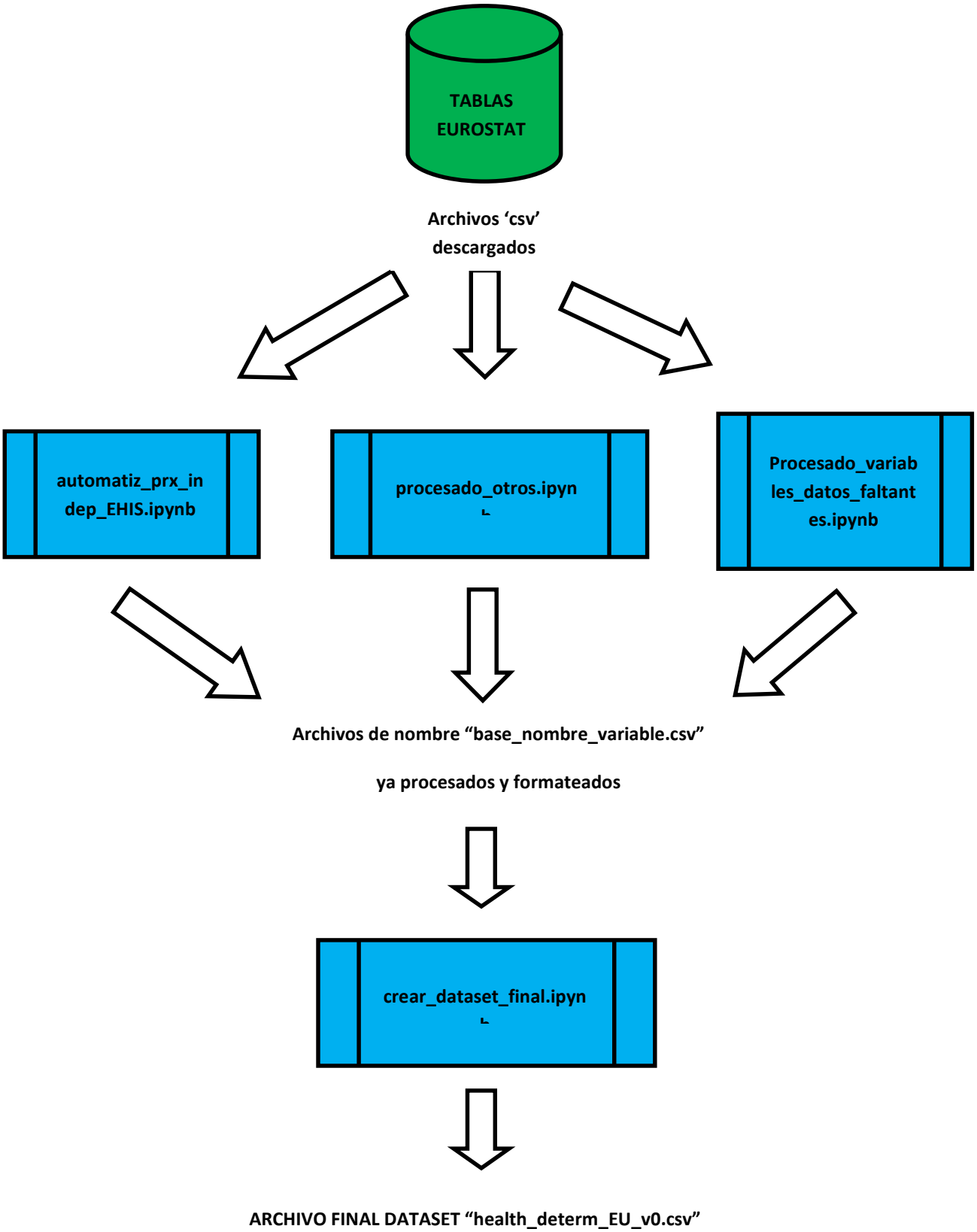

FIGURA 7: DIAGRAMA DE BLOQUES CAPTURA Y PREPROCESADO

TABLA 3: VARIABLES FINALES DEL DATASET UNA VEZ TERMINADO EL PREPROCESADO

NOMBRESVARIABLES	SIGNIFICADO	UNIDADES
“BMI_Underweight”; “BMI_Normal”; “BMI_Overweight”.	Índice de Masa Corporal(distintos grados)	Porcentaje de la franja de población (país/sexo/edad) bajo cada condición(*1)
“Physical_activity”	Actividad Física: Ejercicio físico al menos una vez en semana	(*1)
“DailyFyV_0.portions”; “DailyFyV_From.1.to.4.portions”; “DailyFyV_5.portions.or.more”	Consumos diarios Frutas y Verduras	(*1)
“smoking_Non.smoker”; “smoking_Daily.smoker”; “smoking_Occasional.smoker”	Tipos de fumador	(*1)
“Frequent_drink_pct”	Bebedor ‘habitual’	(*1)
“social_support_Poor”; “social_support_Intermediate”; “social_support_Strong”	Apoyo social auto declarado(distintos grados)	(*1)
“Chron_ill_pct”	Población con enfermedad o dolencia crónica	(*1)
“Sev_body_pain_None”; “Sev_body_pain_Mild.or.very.mild”; “Sev_body_pain_Moderate”; “Sev_body_pain_Severe.or.very.severe”	Niveles de dolor crónico auto declarado	(*1)
“Health_Expend_._GDP”	Gasto sanitario %PIB	Porcentaje sobre el PIB del país
“presc_med_use_pct”	Uso de medicinas prescribidas	(*1)
“non_presc_med_use_pct”	Uso de medicinas no prescribidas	(*1)
“SPH_G_or_VG_pct”	SPH valorada como Buena o muy buena	(*1)

4. ANÁLISIS EXPLORATORIO (“EDA”).

A partir de este momento las fases posteriores de análisis, que comienzan con el análisis exploratorio (“EDA”) y continúan con las distintas técnicas de aprendizaje automático que vamos a utilizar, las implementaremos en R por la mayor capacidad y expresividad de sus librerías de métodos estadísticos, incluyendo las técnicas de regresión.

4.1 Análisis exploratorio unidimensional.

En primer lugar, cargamos el archivo resultado de la fase de preprocesado. Se comprueba que se han cargado todas las variables y con el tipo adecuado. Posteriormente se comprueba también la inexistencia de valores nulos. Tenemos pues las 310 observaciones con 26 variables, las 23 de la tabla anterior más las 3 variables correspondientes a las dimensiones geográfica (GEO), de género (SEX) y de grupos de edad (AGE).

```
## 'data.frame':    310 obs. of  26 variables:
##  $ GEO                                : Factor w/ 31 levels
"Austria”, “Belgium",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ SEX                                : Factor w/ 2 levels
"Females”, “Males": 2 2 2 2 2 1 1 1 1 1 ...
##  $ AGE                                : Factor w/ 5 levels "65
years or over",..: 2 3 4 5 1 2 3 4 5 1 ...
##  $ BMI_Underweight                    : num  11.9 2.4 0.3 0.4 0.6
13.2 7.7 2.5 2.4 2.7 ...
##  $ BMI_Normal                         : num  69.4 56.5 45.6 33.8
33.7 71.2 55.4 62.1 51 42.8 ...
```

FIGURA 8: EJEMPLO DE LA EVALUACIÓN DE LECTURA DE DATOS DE R

Comprobamos a continuación en el resumen estadístico del dataset que los valores de las variables categóricas (dimensiones) son correctas y el rango de valores numéricos es compatible con las unidades de medida (porcentajes).

```
##        GEO           SEX                       AGE     BMI_Underweight
##  Austria : 10   Females:155   65 years or over   :62   Min.   : 0.000
##  Belgium : 10   Males  :155   From 15 to 24 years:62   1st Qu.: 0.600
##  Bulgaria: 10                 From 25 to 34 years:62   Median : 1.500
##  Croatia : 10                 From 35 to 44 years:62   Mean   : 3.217
##  Cyprus  : 10                 From 45 to 64 years:62   3rd Qu.: 4.475
##  Czechia : 10                                          Max.   :18.500
##  (Other) :250
```

FIGURA 9: EJEMPLO RESUMEN ESTADÍSTICO R

*** VARIABLES CATEGÓRICAS.**

Pasamos a examinar las variables, empezando por las categóricas. Recordar que para cada país están presentes los dos géneros y los mismos grupos de edad comentados (dimensiones), por lo cual sus valores están repartidos por igual.

- Variables género y edad.

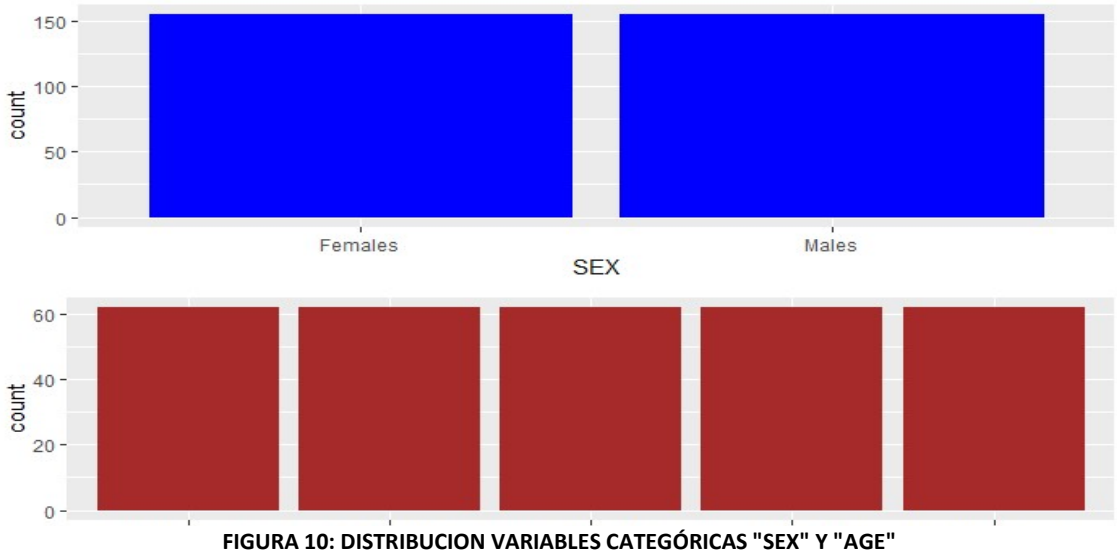

FIGURA 10: DISTRIBUCION VARIABLES CATEGÓRICAS "SEX" Y "AGE"

- **Variables del Índice de Masa Corporal ('BMI')**

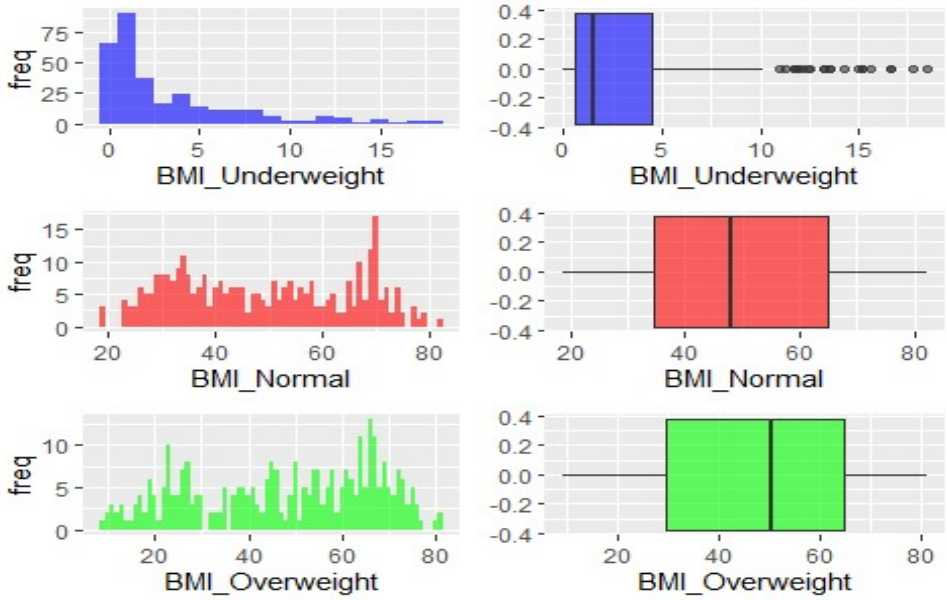

FIGURA 11: DISTRIBUCIÓN VARIABLES DEL ÍNDICE DE MASA CORPORAL

Observamos valores bajos en comparación para la variable 'BMI_Underweight', con presencia de atípicos por encima del 10% que pueden entrar dentro de la lógica en países del entorno europeo. También observamos valores de distribución del sobrepeso (con sesgo a la derecha) algo superiores a los del peso normal, alcanzando valores del 80%.

Como hemos comentado, los valores atípicos (outliers) que nos encontramos para la variable 'BMI_Underweight' no parecen fruto de error sino de valores posibles de la variable para el entorno europeo.

Para el posterior análisis (regresión logística) donde estos outliers pueden constituir ruido que altere los resultados, decidimos tomar como valor límite el máximo/mínimo valor existente no atípico más/menos un 20%. Los valores que sobrepasen por encima/debajo estos límites serán truncados a dichos valores. De esta forma no eliminamos observaciones y queda reflejada su tendencia original (valores altos/bajos). Se prefiere esta alternativa a eliminar las observaciones correspondientes dado el relativo bajo número de observaciones, o realizar algún tipo de imputación que altere dicha tendencia.

Por lo general y a partir de este momento realizaremos siempre esta corrección de outliers, a menos que observemos en los mismos algún posible origen de error y no puedan ser compatibles con una distribución de la población en el ámbito de nuestro estudio.

- **Variables relacionadas con los tipos de fumador ‘smoking’.**

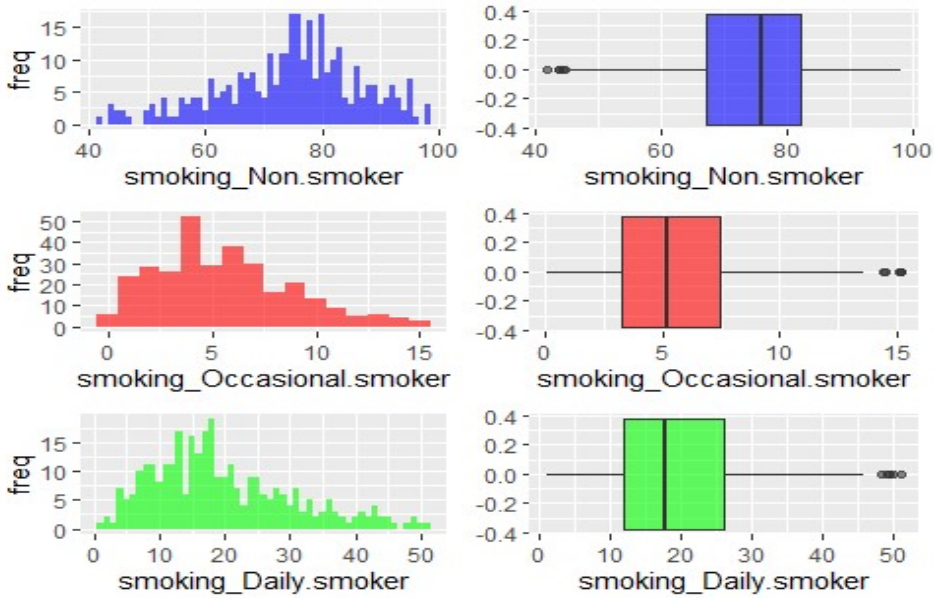

FIGURA 12: DISTRIBUCIÓN VARIABLES TIPOS DE FUMADOR

Se observan valores altos de porcentajes de no fumadores (primer cuartil alcanza casi el 70%) frente a valores más pequeños de ocasionales y diarios, si bien en estos últimos se alcanzan valores que alcanzan el 50%. Los números parecen decir que el consumo de tabaco en Europa empieza a ser minoritario en proporción. Se observan valores atípicos en bajos números presentes en todas las categorías, no muy distanciados. De nuevo no parecen fruto de errores.

- **Variables relacionadas con la Actividad Física y Bebedor habitual (‘Physical_activity’ y ‘Frequent_drink_pct’).**

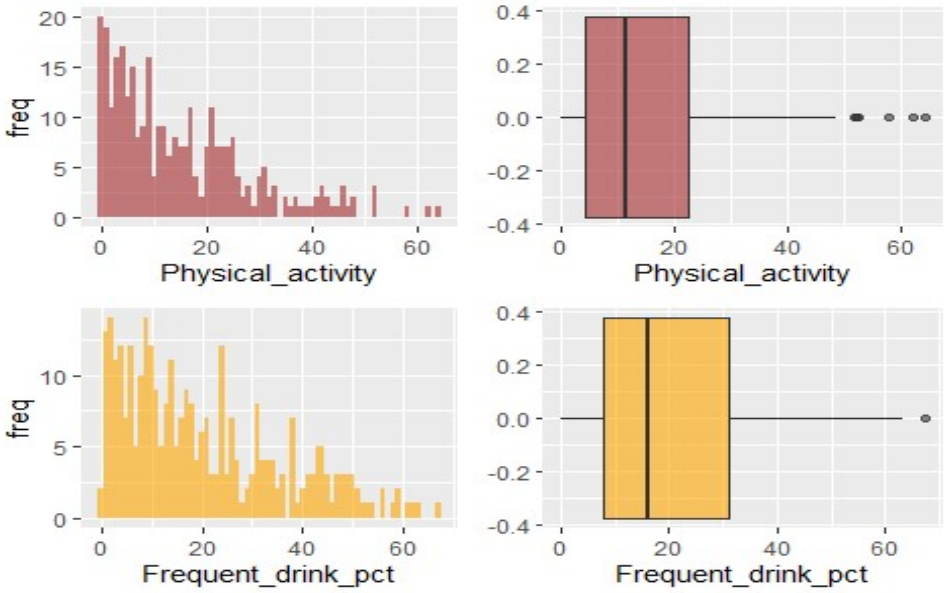

FIGURA 13: DISTRIBUCIÓN DE LAS VARIABLES 'ACTIVIDAD FISICA' Y 'BEBEDOR HABITUAL'

En este caso contemplamos valores de la Actividad Física que alcanzan hasta poco más del 60% (tercer cuartil algo más de 20). Para los Bebedores Frecuentes se alcanza casi el 65% (tercer cuartil 30 aprox.). Se observa sesgo a la izquierda en ambas distribuciones. Presencia de valores atípicos a la derecha en ambos, con mayor número y separación para el primero y un solo valor no muy alejado en el segundo.

- Variables relacionadas con el Consumo diario de Frutas y Verduras 'DailyFyV'.

Se observan porcentajes considerables de la población que no consumen diariamente frutas y verduras (1er Q por encima de 25, 2do Q 45 aprox.), llegando a casi 70% en algunos casos. Valores considerables entre los del porcentaje de población que la consume en menor cantidad de la recomendada (1er Q 40, 2do Q 60), hasta poco más de 80 %. Sin atípicos en ambos.

A resaltar valores bajos por lo general para el consumo recomendado de al menos 5 piezas diarias (1er Q 7 aproximadamente, 2do Q algo más de 15), con valores atípicos que llegan a sobrepasar ligeramente 40%.

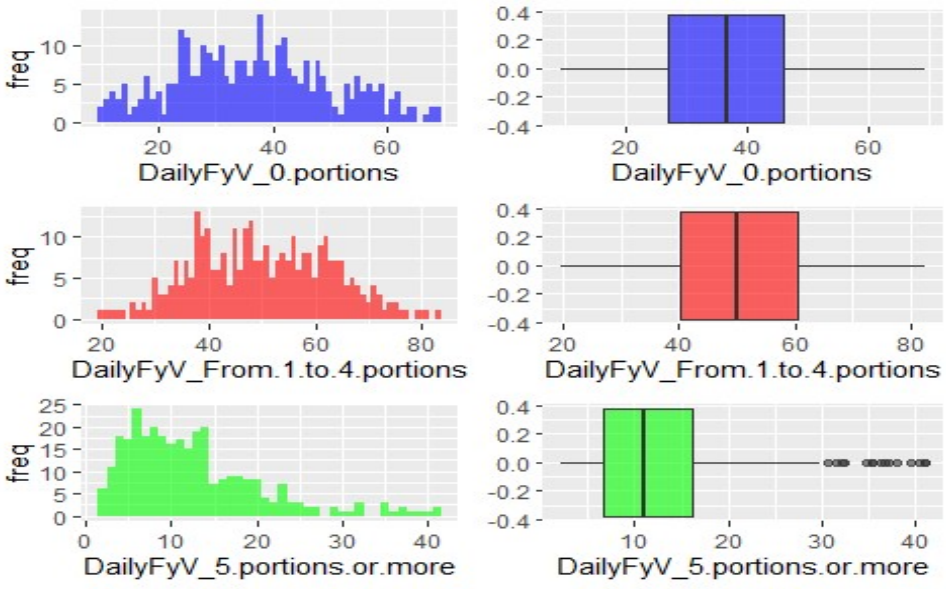

FIGURA 14: DISTRIBUCIÓN VARIABLES CONSUMO DIARIO DE FRUTAS Y VERDURAS

- Variables relativas al Apoyo Social ('social _support').

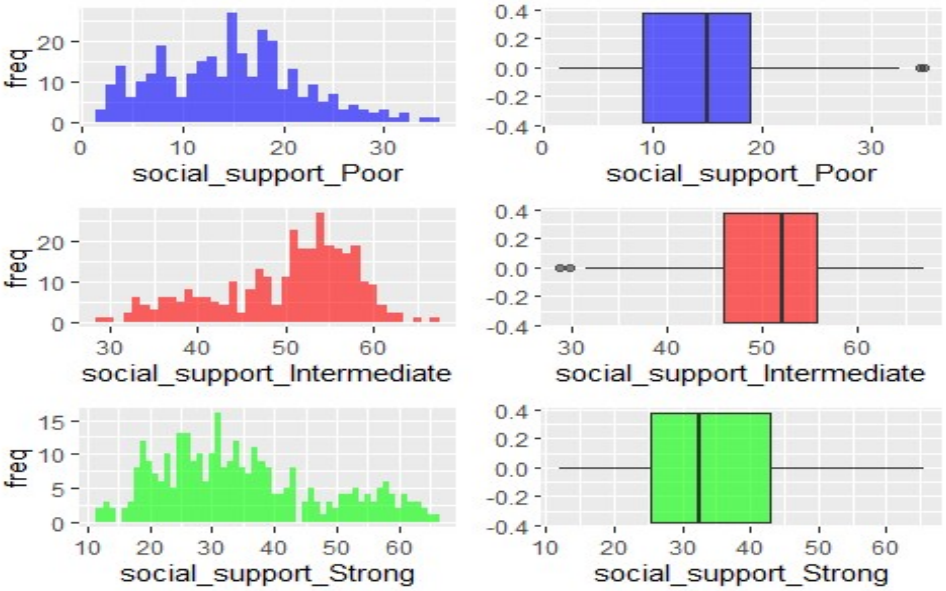

FIGURA 15: DISTRIBUCION VARIABLES APOYO SOCIAL

Entre las variables relativas al "Apoyo Social" destacar que los valores para la categoría 'Poor' son menores a los otros dos, siendo el de 'Intermediate' algo superior al de 'Strong'. Presencia de valores extremos en los dos primeros niveles comentados, escasos y no muy alejados en cualquier caso.

Parece por estos valores que en el entorno europeo el mayor problema es el apoyo social calificado de presente pero no suficiente ('Intermediate') que la de apoyo social prácticamente inexistente ('Poor').

- Variables relativas a los Niveles de dolor crónico 'Sev_body_pain'.

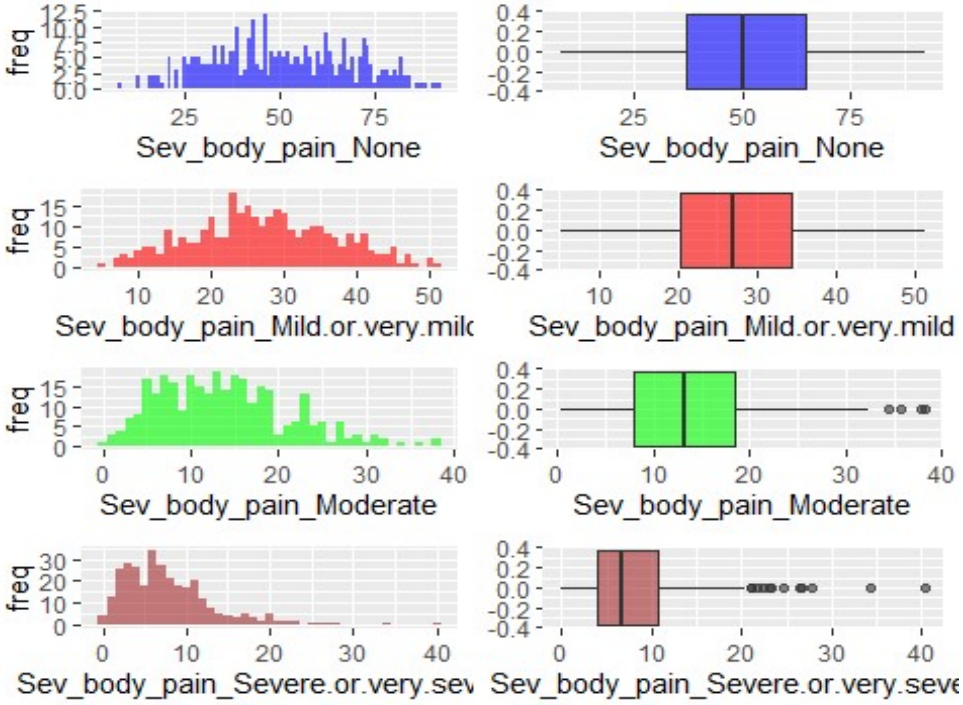

FIGURA 16: DISTRIBUCIÓN VARIABLES NIVEL DE DOLOR CRÓNICO

Entre los niveles de dolor, los mayores valores de porcentaje se encuentran en dolor leve o muy leve y ausencia de dolor físico, con medianas respectivas de 27 y 50 aprox., llegando a alcanzar valores alrededor de 50 y 90. Sin valores extremos ambas categorías.

Los porcentajes de nivel moderado son más bajos con mediana cercana a 15, alcanzando valores que sobrepasan 30 y con valores atípicos hasta cerca de los 40. Aun menor los porcentajes para el dolor severo y muy severo. El 3er cuartil apenas sobrepasa los 10 con valores típicos hasta 20 y atípicos concentrados la gran mayoría en menos de 30.

La distribución de los niveles de dolor parece la usual entre una población genérica.

- Variables relativas a la Población con enfermedad crónica y Gasto sanitario (% del PIB).

Nos encontramos con una tasa de enfermedad crónica con una distribución extendida, con sesgo a la derecha. Con una mediana de 25%, 1er cuartil de 12 y 3er cuartil de 40 aproximadamente, los máximos alcanzan hasta 80% aproximadamente. Presencia de un atípico testimonial.

En cuanto al gasto sanitario el histograma sin 'picos' locales ni generales se explica debido a que dicho valor no depende de los grupos de edad ni género, son el mismo para todos los grupos de un mismo país. Los valores se reparten desde poco más de 4 hasta el 11.5% del PIB aprox., con 1er Q de casi el 7% y 3er Q de 10. Se percibe ligero sesgo a la izquierda. Sin valores atípicos. Vemos que el gasto en sanidad medido como porcentaje del PIB de los países del estudio se mueve en una franja pequeña de valores.

FIGURA 17: DISTRIBUCIÓN VARIABLES POBLACIÓN CON ENFERMEDAD CRÓNICA Y GASTO SANITARIO

- Variables relativas al Uso de Medicinas ("presc_med_use_pct" y "non_presc_med_use_pct").

En el uso de medicinas prescritas y no prescritas se observan primeros cuartiles y medianas semejantes (alrededor de 25 y de 37.5 respectivamente), si bien el rango intercuartílico es mayor para la primera. Ambas alcanzan en sus valores máximos cerca de 90, tratándose de valores atípicos en el caso de las no prescritas. Distribuciones muy similares las de ambas variables.

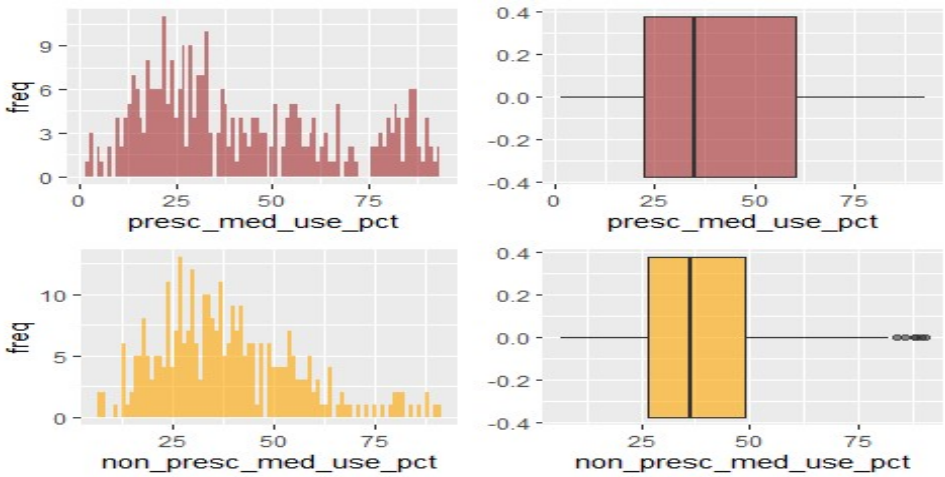

FIGURA 18: DISTRIBUCIÓN VARIABLES USO DE MEDICINAS

- **Variable objetivo SPH valorada como Buena o muy buena 'SPH_G_or_VG_pct'.**

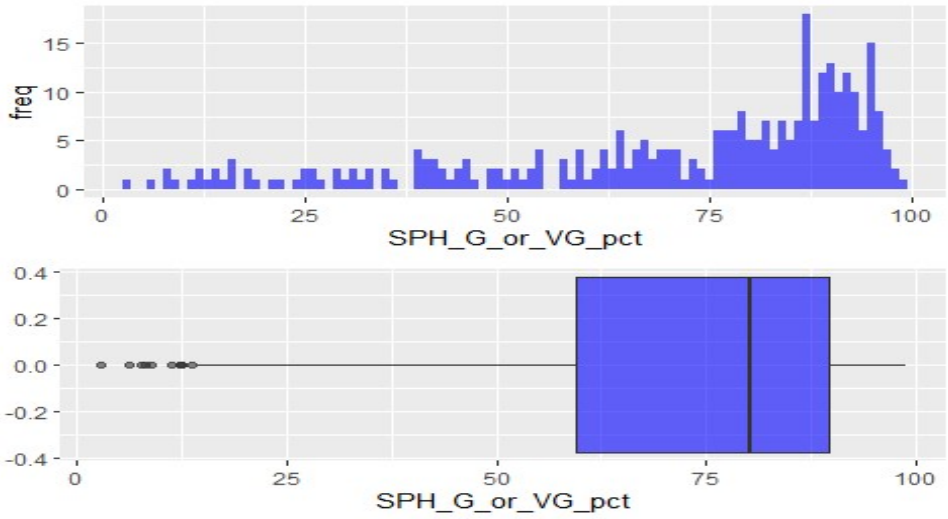

FIGURA 19: DISTRIBUCIÓN VARIABLE OBJETIVO 'SPH BUENA O MUY BUENA'

En primer lugar, los diagramas muestran una distribución de valores por encima del 50% la gran mayoría, con primer cuartil de 59.5 y mediana de 80.1 (media 71.4, del resumen estadístico inicial). Valores máximos llegan casi a la centena. Se observa por lo tanto una clara tendencia a valorar positivamente el estado de salud en el entorno europeo.

En el caso de la variable objetivo lo que nos ocupa es decidir el valor umbral ('threshold') a la hora de convertir esta variable (porcentajes) en la dicotómica necesaria para el posterior análisis aplicando regresión logística. De esta forma además podremos proseguir con el EDA bidimensional, la correspondencia de las variables con los valores de la variable objetivo dicotomizada final de nuestro estudio.

Hay que tener en cuenta que para calificar de forma clara e inequívoca en un bloque de población (país/género/grupo_edad, delimitado por los valores de esas columnas/dimensiones) la percepción positiva o muy positiva como característica o predominante de dicho grupo, no valdrían valores del 50% o ligeramente superiores (¿que pasa con la otra mitad?). Incluso valores cercanos al 60% podrían ser discutibles. Se toma finalmente, apoyado además por la distribución vista con sesgo claro a la izquierda, como **valor umbral el 65**, a falta de ver los resultados del análisis.

Una vez dicotomizada la variable con este umbral, la distribución en porcentaje de la nueva variable objetivo fue: SPHpos = 0 (30%); SPHpos = 1 (70%).

Finalmente ordenamos los niveles de la variable 'AGE' para una mejor presentación y disposición en el EDA bidimensional. Tipificamos como categórica la nueva variable objetivo (dicotimizada) y eliminamos la antigua (en porcentaje) para acometer el análisis exploratorio bidimensional, comportamiento de las variables independientes frente a nuestra variable objetivo.

4.2 Análisis exploratorio bidimensional: correlación entre las variables numéricas.

Vamos a comprobar las correlaciones existentes entre las variables numéricas. La presencia de correlación en estas variables influye negativamente en los resultados de la regresión y su interpretación, dando lugar a valores erróneos de los coeficientes y odds-ratio.

Realizamos entonces el estudio de la correlación entre estas variables con la significación estadística de las mismas mediante el p-valor. La función "rcorr" de la librería "Hmisc" muestra los resultados del estudio de la correlación con prueba de significancia estadística.

Dichos resultados por extensos (gran número de variables numéricas) se muestran en el **ANEXO I**. Iremos nombrando a continuación las variables encontradas correladas (alto valor del coeficiente) que hayan superado la prueba de significación estadística (de hecho todas las mencionadas con p-valor prácticamente cero), y las medidas que hemos decidido tomar al respecto.

- Variables relativas al Índice de masa corporal.

Las 3 variables relativas a BMI (Underweight, Normal y Overweight) están correladas, con valores de 0.78 a -0.99. Por interés a la hora de estudiar su influencia y determinación en el estado de salud conservamos la de sobrepeso (BMI_Overweight) y eliminamos las 2 restantes.

- Variables relativas al Consumo diario de Frutas y Verduras.

Hay correlación entre la de '0_portions' y la de '1 to 4 portions' (-0.82). Para eliminar esta situación decidimos eliminar la primera (cero porciones) y renombrar la segunda a 'Less_5_portions'. De esta forma aún conservamos información sobre el consumo recomendado (5 porciones o más diarias) y parte del menor al recomendado. Señalar que se probó con la alternativa de juntar (sumar) los dos primeros niveles, pero se mantenía la correlación.

- Variables relativas al Tipo de Fumador.

Hay correlación entre 'smoking_Non.smoker' y 'smoking_Daily.smoker' (-0.96). En este caso por mayor interés en estudiar su importancia eliminaremos la primera, manteniendo la segunda junto con la otra existente de 'smoking_Occasional.smoker'.

- Variables relativas al Apoyo Social.

En este caso la decisión se nos presenta más difícil. Existe correlación entre 'social_support_Strong' y 'social_support_Intermediate' (-0.90) y entre 'social_support_Strong' y 'social_support_Poor' (-0.88). Decidimos por lo tanto eliminar la variable 'social_support_Strong'. Quedan pues 'social_support_Poor' y 'social_support_Intermediate', este último con su significado original de "no del todo satisfactorio", variables de interés y creemos suficientes para el estudio de la influencia del apoyo social.

-Interacciones entre las variables relativas a la Población con Enfermedad Crónica, las relativas a Nivel de Dolor Crónico y las de Uso de medicinas prescribidas.

Estamos ante una situación donde se ven afectadas bastantes variables.

Por un lado la variable relativa la Población con Enfermedad crónica ("Chron_ill_pct", de gran interés para el estudio) está correlacionada con todas las relativas al Nivel de Dolor (excepto la de dolor leve, "Sev_body_pain_Mild.or.very.mild") con valores de 0.73 a 0.80. Esta relación tampoco resulta una gran sorpresa.

Por otro lado, también encontramos correlación entre la variable relativa al Uso de Medicinas Prescribidas ("presc_med_use_pct") y tres de las relativas al Nivel de Dolor (-0.73 a 0.77, de nuevo exceptuando "Sev_body_pain_Mild.or.very.mild").

Dada entonces esta situación junto con la condición de que es usual que las enfermedades crónicas manifiesten algún tipo de dolor, se decide eliminar todas las variables relativas a Dolor Corporal (dado su poco interés eliminamos también la de dolor leve "Sev_body_pain_Mild.or.very.mild").

Finalmente se comprueba que existe también alta correlación entre la Población con Enfermedad Crónica ("Chron_ill_pct"), y la relativa al Uso de medicinas prescribidas (+0.89). Siendo una decisión difícil por querer mantenerla en el estudio, pero dada según nuestra opinión la mayor importancia de estudiar los efectos de la Enfermedad Crónica, y dado también que todavía quedaría la variable relativa a Medicinas No Prescribidas, con todo lo anterior decidimos eliminar la variable Uso de Medicinas Prescribidas.

Procedemos entonces a realizar las operaciones descritas y formar el nuevo dataset, que quedará con las siguientes 13 variables que se añaden a las categóricas ya descritas ('GEO, 'SEX', 'AGE') correspondientes a las dimensiones, que no se han modificado:

TABLA 4: VARIABLES NUMÉRICAS DEL DATASET FINAL, TRAS ELIMINAR LAS VARIABLES CORRELADAS

NOMBRES VARIABLES	SIGNIFICADO	UNIDADES
"BMI_Overweight".	Porcentaje de población con sobrepeso	Porcentaje de la franja de población (país/sexo/edad) bajo cada condición (*1)
"Physical_activity"	Actividad Física: Ejercicio físico al menos una vez en semana	(*1)
"DailyFyV_Less.5.portions"; "DailyFyV_5.portions.or.more"	Consumos diarios Frutas y Verduras	(*1)
"smoking_Daily.smoker"; "smoking_Occasional.smoker"	Tipos de fumador	(*1)
"Frequent_drink_pct"	Bebedor 'habitual'	(*1)
"social_support_Poor"; "social_support_Intermediate";	Apoyo social auto declarado (distintos grados)	(*1)
"Chron_ill_pct"	Población con enfermedad o dolencia crónica	(*1)
"Health_Expend_._GDP"	Gasto sanitario %PIB	Porcentaje sobre el PIB del país
"non_presc_med_use_pct"	Uso de medicinas no prescribidas	(*1)
"SPHpos"	SPH positiva (dicotomizada)	Binaria, valores '0' y '1'

Continuamos con el análisis bidimensional.

4.3 Análisis exploratorio bidimensional: distribución de las variables independientes según la variable dependiente.

- Variables relativas al sobrepeso (BMI_Overweight) y la Actividad Física (Physical Activity).

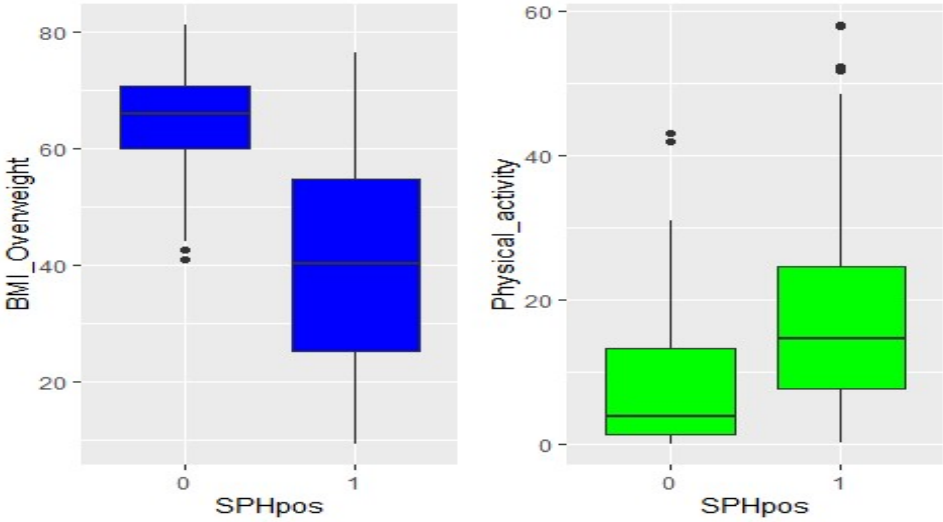

FIGURA 20: VARIABLES SOBREPESO Y ACTIVIDAD FÍSICA POR AUTOVALORACIÓN DE SALUD

En el primer caso se observa una clara diferencia entre las distribuciones que se corresponden con SPH bajo, que presentan mayor porcentaje de población con sobrepeso, y los de valoración del SPH positiva con menores valores de porcentaje en su distribución. Aún con cierto solapamiento se puede percibir el porcentaje de sobrepeso como posible elemento de influencia.

En cuanto a la actividad física el grupo de valoración positiva presenta porcentajes de población mayores en su distribución, si bien en este caso el solapamiento es más acusado. Habrá que verificar su relevancia en el análisis posterior.

Estas relaciones entran dentro de lo esperable, hacer ejercicio y mantener un peso adecuado parecen mejorar la auto percepción de salud.

- Variables relativas al Consumo Diario de Frutas y Verduras (DailyFyV).

En cuanto al consumo diario de frutas y verduras, parece no haber prácticamente diferencias en las distribuciones de valores (porcentajes) en función de la SPH en el caso donde el consumo es el recomendado (5 raciones o más, variable 'DailyFyV_5.portions.or.more'). Para los de consumo menor se observa una pequeña diferencia, con ligeramente mayores valores porcentuales los de SPH negativa.

Estudiados las distribuciones por separado y debido al solapamiento en las distribuciones, habrá que esperar al análisis junto con el resto de las variables.

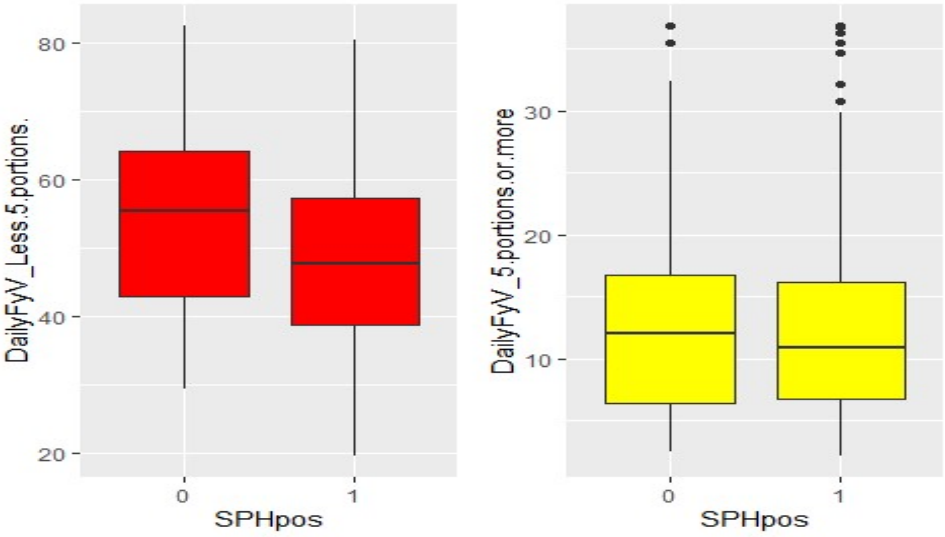

FIGURA 21: VARIABLES CONSUMO DIARIO DE FRUTAS Y VERDURAS POR AUTOVALORACION DE SALUD

- Variables relativas al Tipo de Fumador (smoking).

En el Tipo de Fumador observamos con sorpresa para ambas variables una correspondencia entre valores mayores de porcentaje de población consumidora y valoración positiva del SPH, existiendo mayor solapamiento de las distribuciones en el caso de fumador diario.

El carácter de "auto percibido" de la variable objetivo puede relativizar esta aparente contradicción. A pesar de la certeza del perjuicio de consumo y de las campañas de sensibilización existentes, la propia población fumadora no parece valorarla como factor negativo para su estado de salud o simplemente la ignora. Parecería tener mayor influencia la variable de Fumadores Ocasionales. Deberemos esperar a la fase de análisis.

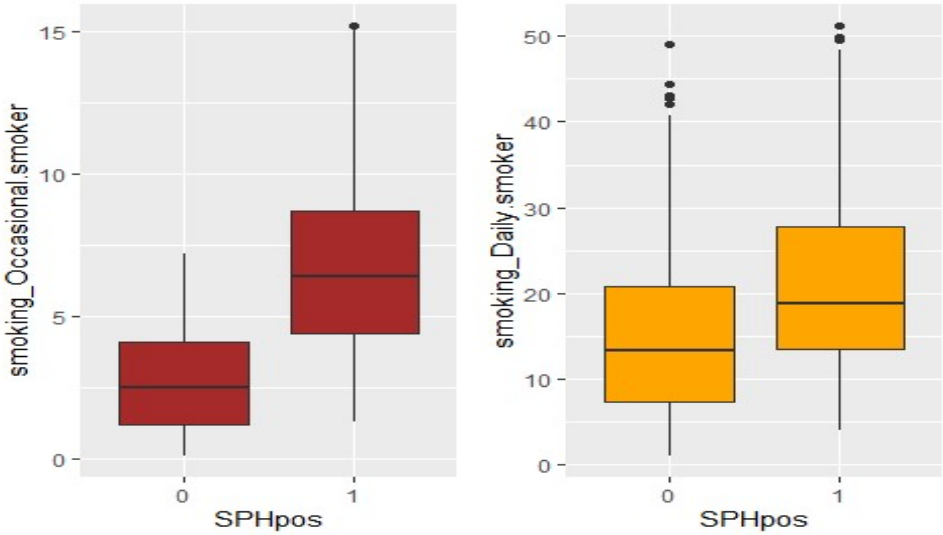

FIGURA 22: VARIABLES CONSUMO DE TABACO POR AUTOVALORACIÓN DE SALUD

- Variables relativas al Apoyo Social del entorno (social_support).

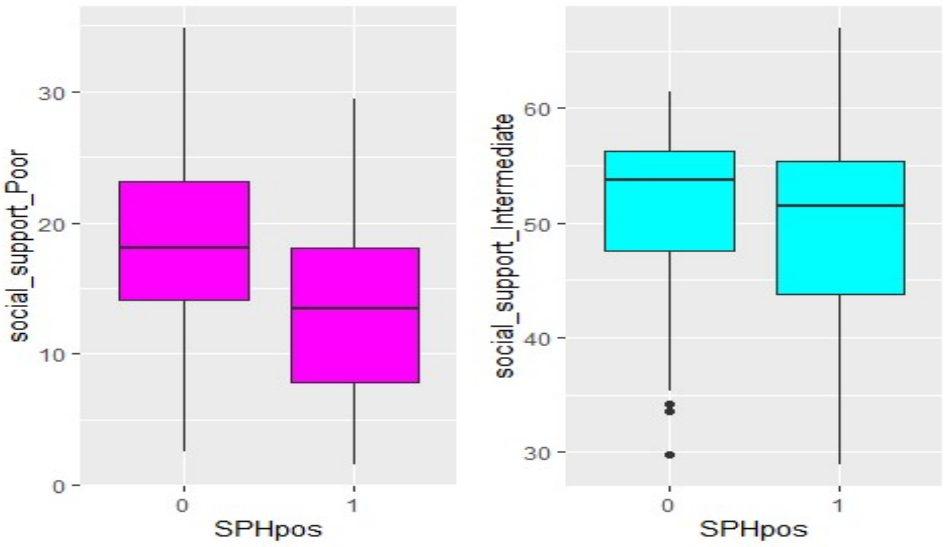

FIGURA 23: VARIABLES APOYO SOCIAL POR AUTOVALORACIÓN DE SALUD

Vemos en estas gráficas primero que los valores mayores porcentuales de Apoyo Social Pobre se asocian más con una percepción negativa del estado de salud, con una perceptible diferencia (aún con cierto solapamiento) entre las distribuciones de porcentajes. Para el caso de Apoyo Intermedio nos encontramos con distribuciones respecto a la SPH con diferencias pero menores (más "concentrada" la distribución para la valoración negativa).

Como factor parece entonces que el apoyo social pobre podría ser determinante, habrá que corroborar la influencia o no en el análisis.

- Variables Población con Enfermedad Crónica (Chron_ill_pct) y Uso de Medicinas No Prescribidas (non_presc_med_use_pct)

Se observa (por otro lado del todo esperable) una clara diferencia en las distribuciones de porcentaje para la Población con Enfermedad Crónica, con valores mayores de porcentaje de población con tal condición relacionado con SPH negativa y relativamente poco solapamiento entre ambas. Es entonces una candidata clara a factor determinante, a la espera de los resultados del análisis.

En cuanto al uso de medicinas no prescribidas, no existen grandes diferencias en las distribuciones, acaso estando más concentrada la de SPH positiva. Duda sobre su influencia, a la espera del análisis.

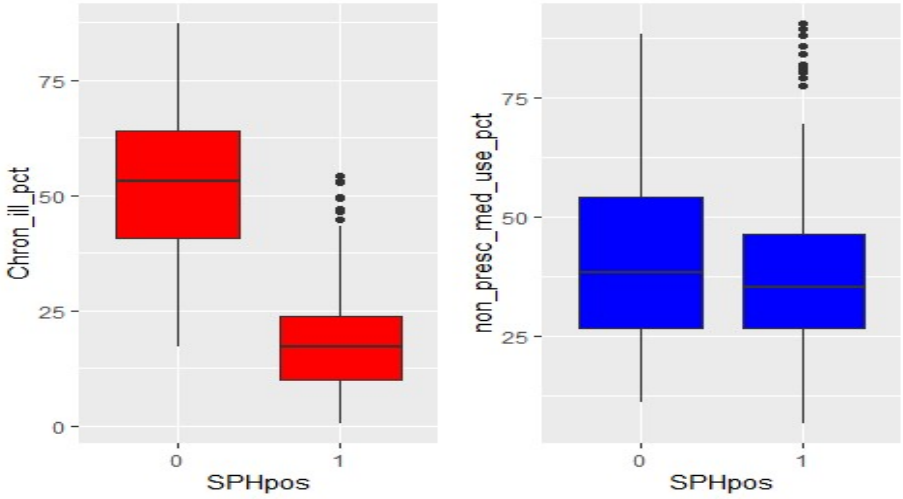

FIGURA 24: VARIABLES POBLACIÓN CON ENFERMEDAD CRÓNICA Y USO DE MEDICINAS NO PRESCRIBIDAS POR AUTOVALORACIÓN DE SALUD

- Variables Gasto Sanitario (Health_Expend_._GDP) y Bebedores habituales (Frequent_drink_pct).

Dentro del rango pequeño de valores en el que se mueve el gasto sanitario (en % del PIB), se distingue una pequeña diferencia en las distribuciones para cada valor de SPH, siendo mayores los porcentajes de gasto entre los de valoración positiva. Dada el contexto de rango corto de valores, podría ser posible que un mayor gasto en sanidad influya aún en forma ligeramente favorable, a falta de confirmarse en el análisis.

En cuanto a Bebedores Habituales, con cierto solapamiento en las distribuciones de valores, se observa al contrario de lo que en un principio se pudiera suponer que: mayores porcentajes de población bebedora habitual se asocian con SPH positiva. De nuevo aquí al igual que indicamos para el consumo de tabaco recordar el concepto de "auto percibida" de la variable objetivo, y la no valoración como negativa del consumo por parte de los bebedores frecuentes.

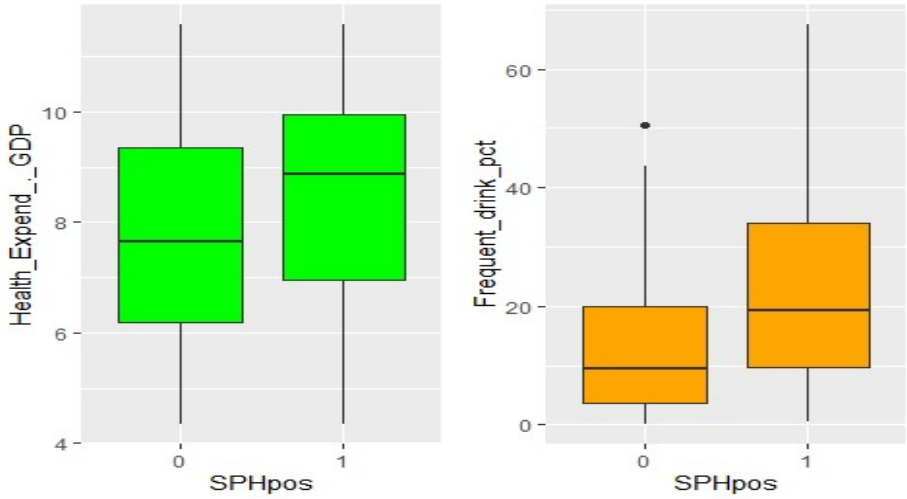

FIGURA 25: VARIABLES GASTO SANITARIO Y BEBEDORES HABITUALES POR AUTOVALORACIÓN POSITIVA DE SALUD

- Relación de la SPH con las variables categóricas AGE y SEX.

En primer lugar vemos que las distribuciones por género entre valoración positiva y negativa de salud presentan proporciones similares en hombres y mujeres (cerca del 75%, casi tres cuartas partes valoración positiva SPH)

En cuanto a los grupos de edad observamos hasta los 34 años la práctica ausencia de población que manifieste SPH negativa. Ésta aparece a partir de esos 35 años, aumentando en cada uno de los 3 grupos restantes con el aumento de edad. La franja entre 45 y 64 alcanza el 50% aproximadamente y en la tercera edad (65 o más) la valoración negativa es mayoritaria, sobre el 90%. Vemos pues una tendencia clara a valorar peor el estado de salud en los últimos intervalos de edad, previa a la tercera edad y en ella misma.

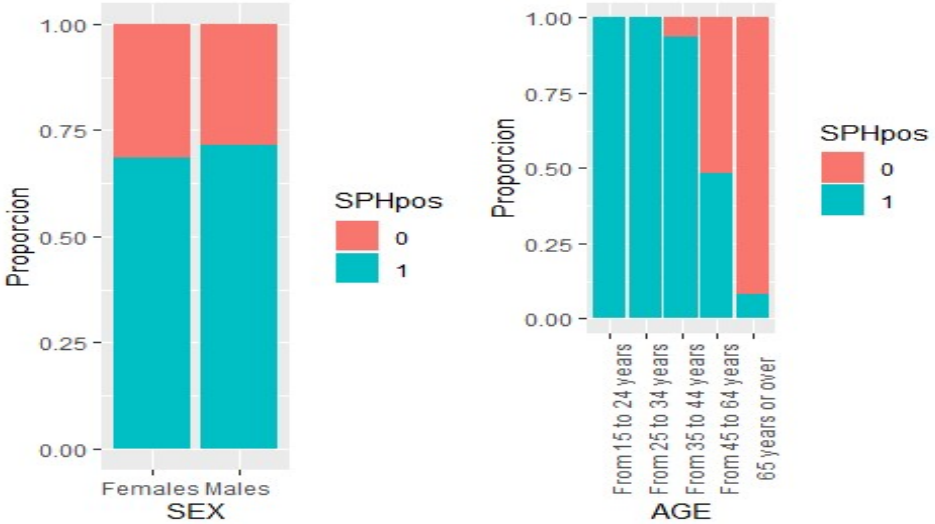

FIGURA 26: VARIABLES CATEGÓRICAS AGE Y SEX POR AUTOVALORACIÓN POSITIVA DE SALUD.

Veamos ahora la interacción entre las dos dimensiones.

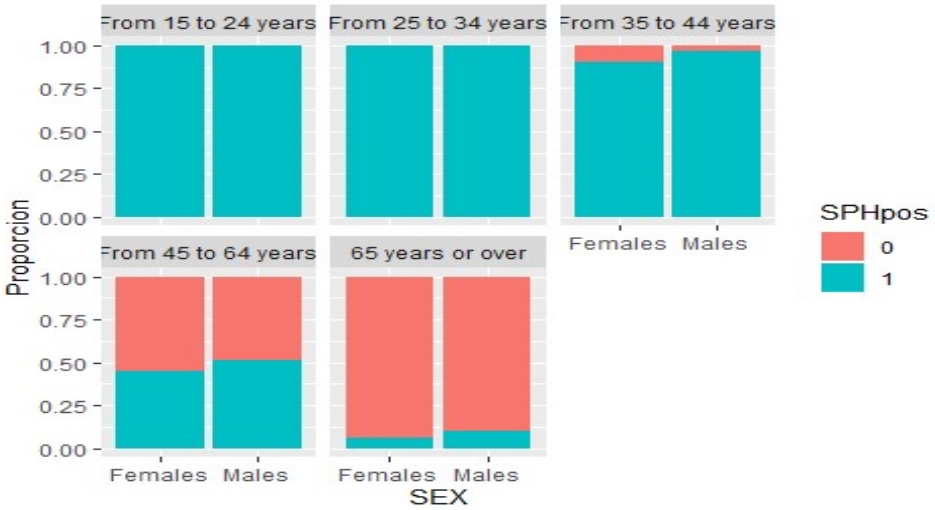

FIGURA 27: INTERACCIÓN VARIABLES AGE Y SEX

No se observa diferencia en las proporciones por sexo en todos los grupos de edad, con valores prácticamente idénticos.

En un principio y a falta de los resultados del análisis con el resto de variables, parece que el género tiene menor influencia mientras que la edad si pudiera ser determinante.

5. ANÁLISIS DE LA INFLUENCIA DE LAS DISTINTAS VARIABLES: REGRESIÓN LOGÍSTICA. VALIDACIÓN DEL MODELO.

5.1 Regresión logística.

Para descubrir cuáles son, en el ámbito global europeo, las variables determinantes de la auto valoración del estado de salud (SPH), prescindiremos primero de la variable GEO (país). En posterior apartado haremos un estudio por país de la materia, además de la imposibilidad de agregar 30 variables (31 países menos el de referencia) a la regresión logística, ya con 15 variables y la restricción de sólo 310 observaciones.

La influencia de cada variable vendrá dada por su significación estadística (p-valor asociado a su coeficiente) en la regresión logística (variable objetivo 'SPHpos'). Para aquellas variables cuyos coeficientes sean estadísticamente significativos (sean por lo tanto explicativas) calcularemos y valoraremos sus odds-ratio.

En las tareas de preparación particionaremos nuestro dataset en conjuntos TRAIN y TEST de forma estratificada (en ambos las proporciones de la variable objetivo sean similares a las del conjunto original) y estableceremos los niveles de referencia para las variables explicativas categóricas ('Males' para 'SEX' y 'From 15 to 24 years' para 'AGE').

En la búsqueda del mejor modelo (según criterio de información AIC) y las variables con influencia estadísticamente significativa, realizaremos procesos de selección de variables del modelo ('stepwise') en las direcciones 'backward', 'both' y 'forward'. El proceso 'stepwise' consiste en, partiendo del modelo sin/con todas las variables, añadir/eliminar una variable en cada paso de entre las disponibles/existentes, evaluando con cuál de las variables añadida/eliminada se obtiene un mejor modelo. A la que parte del modelo sin variables se le denomina 'forward' y al que parte del modelo con todas las variables 'backward'. En el caso de 'both' se parte del modelo completo y en cada iteración se prueba añadiendo y después quitando de este conjunto una variable. **Las trazas** (secuencia de modelos intermedios durante el proceso 'stepwise' comentado) **de cada modalidad** por su longitud pueden ser consultadas en el **ANEXO II**.

Se llegó al <u>mismo mejor modelo</u> (con las mismas variables) <u>con las direcciones 'backward' y 'both'</u>. Cabe señalar que se hicieron pruebas siguiendo la traza con modelos intermedios de valores de AIC algo mayores pero cercanos, para ver si podíamos incluir más variables estadísticamente significativas aún con peor valor AIC para el modelo, pero no hubo éxito.

El **resumen del mejor modelo encontrado, con los coeficientes, error estándar y pruebas de significación estadística (p-valor asociado) de las variables** es el siguiente:

TABLA 5: RESUMEN MEJOR MODELO REGRESIÓN LOGÍSTICA

VARIABLES	COEFICIENTE	ERROR ESTÁNDAR	P-VALOR ASOCIADO
BMI_Overweight	0.06026	0.03665	0.1001
DailyFyV_Less.5.portions	0.04738	0.03197	0.1384
DailyFyV_5.portions.or.more	0.08674	0.05103	0.0892
smoking_Occasional.smoker	0.50580	0.21277	0.0174
social_support_Poor	-0.11658	0.04746	0.0140
Chron_ill_pct	-0.13709	0.03309	3.44e-05
Health_Expend_._GDP	0.37567	0.18155	0.0385
SEX_RFemales	-1.53159	0.91629	0.0946
Null deviance	304.717 on 248 degrees of freedom		
Residual deviance	85.039 on 240 degrees of freedom		
AIC	103.04		

5.2 Validación del modelo: Devianza.

En primer lugar para comprobar la bondad del ajuste de nuestro modelo logístico utilizamos la Devianza, que se define como menos el doble del logaritmo de la razón de verosimilitud (log-likelihood), es decir, Devianza=−2× log-likelihood y se representa como -2LL.

La Devianza tiene una distribución $\chi 2$ y compara los valores de la predicción observados en dos momentos:

- El modelo sin variables independientes, sólo con la constante (modelo referencia).

- El modelo con las variables predictoras introducidas.

Simplemente de los valores devueltos en el resumen de nuestro modelo tomamos la Devianza del modelo referencia (valor 'Nulldeviance') y le restamos la Devianza del nuevo modelo (valor 'Residual Deviance'). Estos valores son accesibles desde la propia variable donde almacenamos nuestro modelo.

Esta diferencia se conoce como “likelihood-ratio” y tiene una distribución χ2 con k-1 grados de libertad, el número de parámetros del nuevo modelo, menos el número de parámetros del modelo referencia que es siempre 1.

Para calcular la probabilidad asociada al estadístico chi-cuadrado utilizamos la función <pchisq(Chi, gl)>, cuyos argumentos son el estadístico chi-cuadrado y sus grados de libertad. La probabilidad que queremos es 1 menos el valor de la función <pchisq ()>. Este valor nos indica si nuestro modelo explica mejor la variable dependiente de lo que lo haría el azar, si nuestro modelo en conjunto es significativo.

Hechos los cálculos el p-valor asociado es prácticamente cero, menor que la significancia escogida del 5 % (Alpha = 0.05), por lo tanto podemos afirmar que **el modelo en conjunto es significativo**.

5.3 Validación del modelo: Multicolinealidad.

Aparte del estudio de correlación ya practicado entre las variables numéricas, comprobaremos la posible existencia de multicolinealidad, que afectaría gravemente a la interpretación de los resultados de la regresión. Lo haremos mediante el cálculo del factor de inflación de la varianza ‘VIF’ de todas las variables independientes:

TABLA 6: VALIDACIÓN, FACTOR DE INFLACIÓN DE LA VARIANZA

VARIABLES	VIF
BMI_Overweight	2.468416
DailyFyV_Less.5.portions	2.052280
DailyFyV_5.portions.or.more	2.099753
smoking_Occasional.smoker	1.888218
social_support_Poor	1.345774
Chron_ill_pct	2.109755
Health_Expend_._GDP	1.531077
SEX_RFemales	2.710311

El valor mínimo posible de ‘VIF’ es 1, se suele tomar como umbral máximo indicador de multicolinealidad el valor 10, aunque en algunos estudios ya toman por encima de 5 como indicador posible de multicolinealidad. En nuestro caso todos los valores se sitúan por debajo de ambos umbrales, luego podemos en este aspecto **dar nuestro modelo como válido**.

5.4 Validación del modelo: Métricas de clasificación.

Por último, en la validación de nuestro modelo analizaremos las métricas usuales de clasificación (SPHpos '0' o '1').Para ello debemos antes que nada realizar sobre el conjunto TEST las mismas operaciones de preparación (definir los mismos niveles de referencia) que realizamos en cuanto a las variables categóricas para el conjunto TRAIN.

A continuación, calcularemos la matriz de confusión y principales métricas de bondad del ajuste (clasificación) que se derivan de ella. Para ello realizamos las predicciones de nuestro modelo sobre los datos del conjunto TEST y los contrastaremos con los valores reales (TEST) del SPHpos. Los resultados fueron los siguientes:

TABLA 7: VALIDACIÓN, MATRIZ DE CONFUSIÓN

PREDICC OBSERV	'0'	'1'
'0'	15	2
'1'	3	41

TABLA 8: VALIDACIÓN, MÉTRICAS DE CLASIFICACIÓN (REGRESIÓN LOGÍSTICA)

ACCURACY	0.918
PRECISION	0.9318
RECALL	0.9535
F1	0.9425

En la matriz de confusión podemos observar solo 5 clasificaciones erróneas de entre las 61 observaciones totales del conjunto TEST.

Como vemos nuestro modelo alcanza buenos valores tanto del primer objetivo de precisión ('accuracy', 91.8%) como de entre otros los valores 'precision', 'recall' y 'f1', usuales métricas de clasificación, todas por encima del 93%.

Para terminar con la validación del modelo dibujamos la curva ROC correspondiente a nuestro modelo y calculamos el área bajo esa curva:

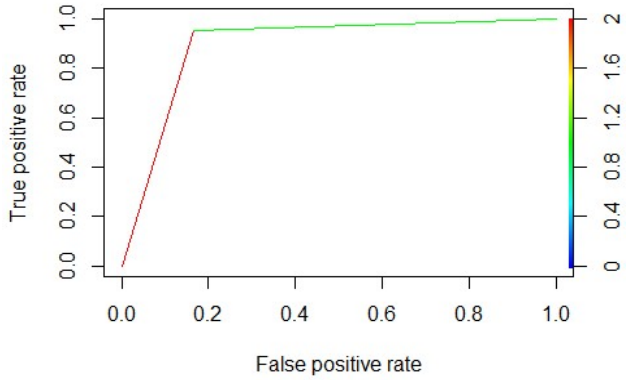

FIGURA 28: VALIDACIÓN, CURVA ROC

Como se puede comprobar con la curva ROC cercana a la esquina superior izquierda y un valor calculado del área bajo la curva (AUC) de 0.89, **se corrobora la bondad de nuestro modelo**.

5.5 Cálculo de los odds-ratio

Finalmente, superadas las pruebas de validación de nuestro modelo con éxito, pasamos al **cálculo de los odds-ratio con sus intervalos de confianza de las variables estadísticamente significativas** (considerando un nivel de significación de 0.1, 10%, TABLA 5). Es necesario en este punto señalar que incluimos la variable de sobrepeso (BMI_Overweight) con un p-valor que sobrepasa ligeramente el nivel (0.1001) dada la pequeña diferencia y por la importancia de este factor en el contexto de nuestro análisis, que va más allá de su significación. Los resultados son:

TABLA 9: ODDS-RATIO VARIABLES ESTADÍSTICAMENTE SIGNIFICATIVAS

VARIABLES	ODDS-RATIO	IC: 5%	IC: 95%
BMI_Overweight	*0.9415207*	*0.88230542*	*0.9957348*
DailyFyV_5.portions.or.more	*1.0906083*	*1.00584828*	*1.1912773*
smoking_Occasional.smoker	*1.6583096*	*1.19118242*	*2.4084185*
social_support_Poor	*0.8899614*	*0.82009911*	*0.9599426*
Chron_ill_pct	*0.8718911*	*0.82149695*	*0.9166357*
Health_Expend_._GDP	*1.4559706*	*1.08753314*	*1.9848585*
SEX_RFemales	*0.2161924*	*0.04470736*	*0.4306222*

Podemos ver que los factores que afectan **positivamente** al SPH (OR>1) son Consumo de al menos 5 porciones de frutas y verduras, Fumador Ocasional y Gasto Sanitario. Los que afectan **negativamente** al SPH (OR<1) son el Sobrepeso, el Apoyo Social Pobre, la Población con Enfermedad Crónica y el Género Femenino. En el apartado **7.1**de las **Conclusiones** los analizaremos con detalle.

6. ANÁLISIS DE LOS DATOS POR PAÍSES. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA). AGRUPAMIENTO.

Para estudiar la situación por país de la variable objetivo (auto valoración del estado de salud, 'SPHpos'), dado el número elevado de variables de las que disponemos (15) y que estudiando por país se nos quedan 31 observaciones, decidimos realizar un análisis PCA y estudiar la carga (pesos de las variables originales) en los factores resultado de la misma.

Agrupando primero los datos por país (promediando los valores de las variables numéricas de las distintas franjas de género y edad pertenecientes a cada uno) y con el previo estandarizado de estos valores, se procede al cálculo de los componentes principales (PCA). El objetivo es encontrar entre estos componentes (factores) aquellos que por sus cargas (pesos de las variables originales en ellos) y teniendo en cuenta si estas variables afectan de forma positiva o negativa al SPH (como vimos en el apartado anterior con los odds-ratio), puedan considerando el aporte en conjunto de los pesos proporcionar una orientación clara, bien positiva o negativa, del factor en cuestión respecto a la variable objetivo 'SPHpos'.

El **cálculo de los componentes, sus pesos y gráficas, por su elevado número se acompañan en el ANEXO III.** Aquí mostraremos las gráficas y analizaremos los dos factores encontrados con significado u orientación global clara respecto a la valoración positiva de salud. De entre los factores encontramos con las condiciones descritas los **factores 'F1'y 'F3'**:

6.1 Factor "F1".

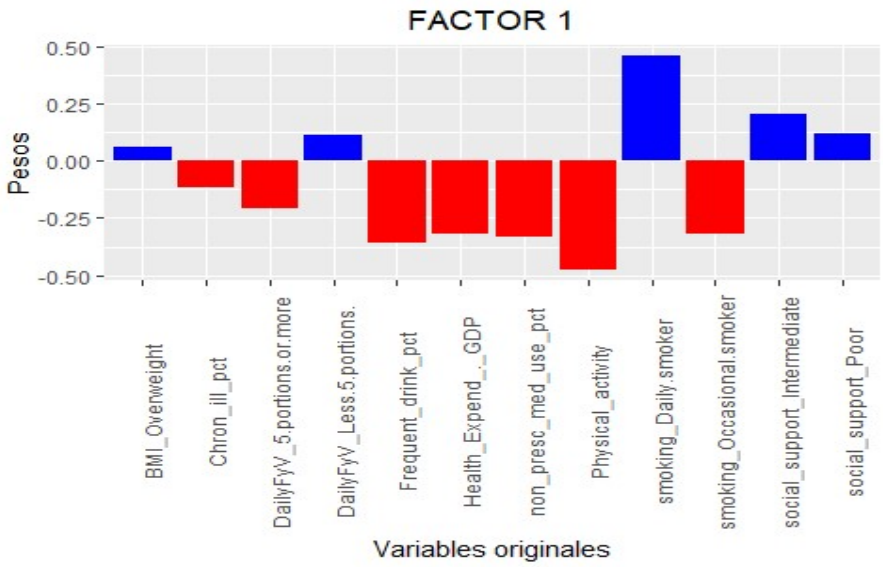

FIGURA 29: PESOS DEL FACTOR 'F1' DEL PCA

Las variables con mayores aportaciones con el signo positivo son "Smoking_Daily.smoker" y "social_support_intermediate". Con el signo negativo por orden de

mayor a menor peso “Physical_activity”, “Frequent_drink_pct”, prácticamente iguales “Health_Expend_._GDP”, “Smoking_Occasional.smoker” y “non_prescrib_med_use_pct”, y por último “DailyFyV_5.portions.or.more”. Para ambos signos se han despreciado las variables con valores inferiores o iguales a 0.125, por ser prácticamente la cuarta parte de los valores máximos y por ser las variables con esos pesos pocas en número respecto al resto, con lo cual su aportación no es comparable.

Para el signo positivo señalar que nos encontramos con solo 2 variables a considerar. Aunque ninguna de ellas resultó significativa en la regresión logística la variable “social_support_intermediate” (existencia de apoyo social pero insuficiente) da un claro tono negativo respecto del SPHpos.

De entre las del signo negativo nos encontramos con las 3 variables significativas (regresión logística) que incrementaban la valoración positiva del estado de salud: “Health_Expend_._GDP”, “DailyFyV_5.portions.or.more” y “Smoking_Occasional.smoker”. Del resto en el EDA mostraban mayores valores en la distribución correspondiente a la valoración positiva, o prácticamente similar en el caso de uso de las medicinas.

Valorando estas aportaciones en conjunto, teniendo en cuenta el significado de las variables, la orientación sobre la SPHpos de cada signo y el mayor número de variables y significancia de las mismas para el signo negativo, podemos concluir que **el factor F1 resumido tendría el significado de “Apoyo social insuficiente frente al Gasto sanitario y hábitos de vida considerados positivos”, con el carácter de cuanto más negativo mejor la autovaloración de salud (SPHpos).**

6.2 Factor “F3”.

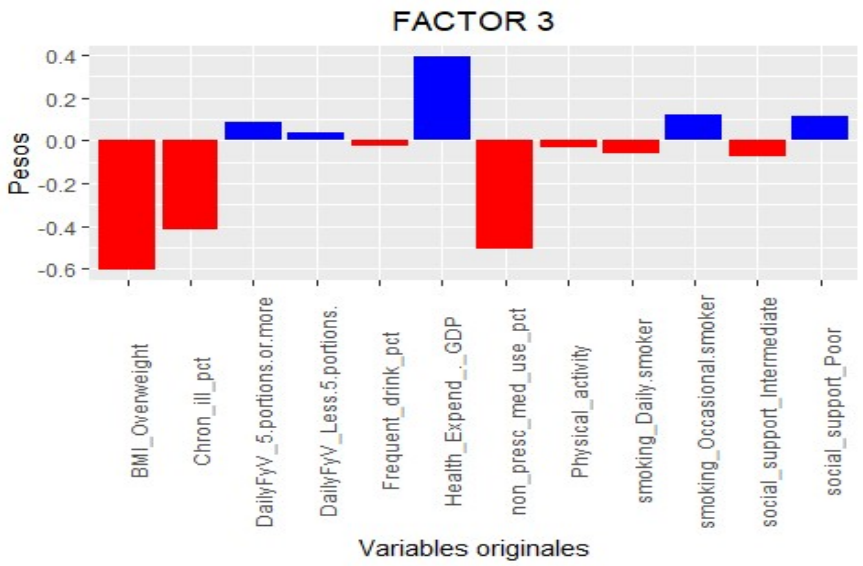

FIGURA 30: PESOS DEL FACTOR 'F3' DEL PCA

Para este factor las variables a considerar son menos ya que nos encontramos con bastantes de ellas con pesos inferiores o como mucho iguales a la cuarta parte de las 4 variables a considerar. Con signo positivo y aportación única según lo comentado anteriormente nos encontramos con "Health_Expend_._GDP", variable significativa y de aporte positivo a la SPHpos. Con el signo negativo tenemos a "BMI_Overweight", "Chron_ill_pct" y "non_prescrib_med_use_pct", las dos primeras significativas y con aporte negativo considerable a la SPHpos, la tercera no significativa y neutra (sin poder observarse orientación en las distribuciones por SPHpos en el EDA), aunque podría esta tercera incluso verse como una consecuencia de la condición de enfermedad crónica ya contemplada.

Valorando globalmente estas aportaciones, teniendo en cuenta el significado de las variables, la orientación sobre la SPHpos de cada signo, podemos concluir que **el factor F3 resumido tendría el significado de "Gasto sanitario frente a mala condición física (sobrepeso, enfermedad crónica)", con el carácter de cuanto más positivo mejor la autovaloración de salud (SPHpos).** Como resumen:

- **F1**: "Apoyo social insuficiente frente al Gasto sanitario y hábitos de vida considerados positivos". La valoración positiva del estado de salud (SPHpos) se corresponde con el signo negativo de este factor.
- **F3**: "Gasto sanitario frente a mala condición física (sobrepeso, enfermedad crónica)". La valoración positiva del estado de salud (SPHpos) se corresponde con el signo positivo de este factor.

6.3 Agrupación a partir de los componentes principales seleccionados (F1, F3).

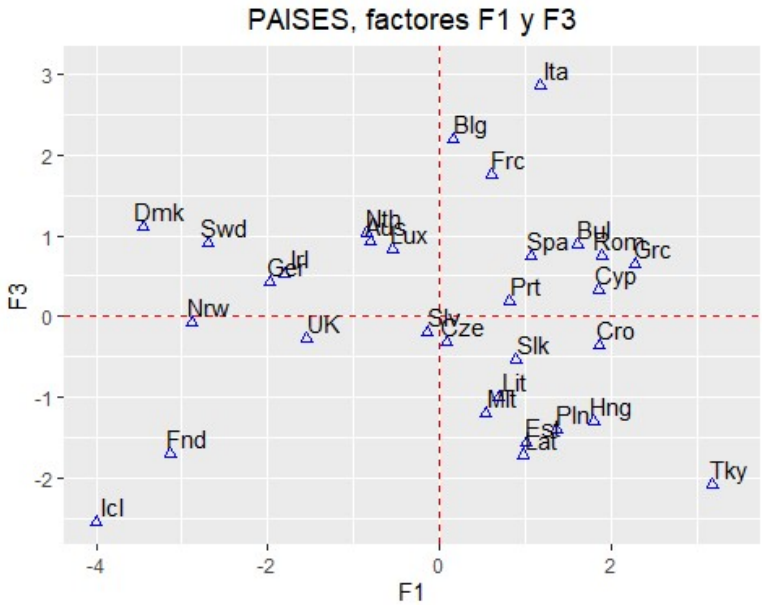

FIGURA 31: GRÁFICA DE SITUACIÓN DE LOS PAÍSES SEGÚN SUS FACTORES 'F1' Y 'F3'

Nos encontramos en primer lugar con 2 agrupaciones más separadas del resto, la formada solo por Turquía en la esquina inferior derecha y la formada por Islandia y Finlandia en la inferiorizquierda. Con menos separación se observa también el grupo de 3 formado por Italia, Bélgica y Francia en la parte superior derecha.

Existe además una serie de puntos más separados en el cuadrante superior izquierda (con Reino Unido y Noruega cruzando el límite ligeramente, valor muy poco negativo de F3), destacando como muy cercanos entre sí a Holanda, Austria y Luxemburgo. El resto de este cuadrante son Dinamarca, Suecia Alemania e Irlanda.

Por último en la parte derecha junto con dos puntos en el centro se encuentran una serie de países: España, Portugal, Grecia, Chipre, Rumanía, Bulgaria, Eslovaquia, Estonia, Letonia, Lituania, Malta, Polonia y Hungría. Lo de la parte central son Eslovenia y la República Checa.

Nuestra objetivo es aplicar técnicas de clusterización a los puntos (países) y poder comparar y sacar conclusiones si es posible sobre dichos clústeres. Al no tener un número claro de clústeres predeterminado y dado el relativo número pequeño de puntos, pensamos que los dendogramas nos pueden dar una idea muy clara de las posibles agrupaciones y distancias entre ellas. Por esta razón nos decantamos por implementar **clustering jerárquico acumulativo**.

Dentro de estas opción de clustering probaremos con los "linkages" Ward (basado en la similitud entre puntos de un mismo cluster, idea similar al método k-means), el complete (suele formar grupos compactos) y el average (utilizado como un método a medio camino). Observaremos los dendogramas para poder tener una idea inicial de un rango de número de clústeres a probar. En la elección final influirá el que sean distinguibles los grupos más aislados comentados en la descripción general, y a ser posible que exista uniformidad de signos en un mismo cluster, para una interpretación más clara según los factores F1 y F3.Veamos pues los dendogramas:

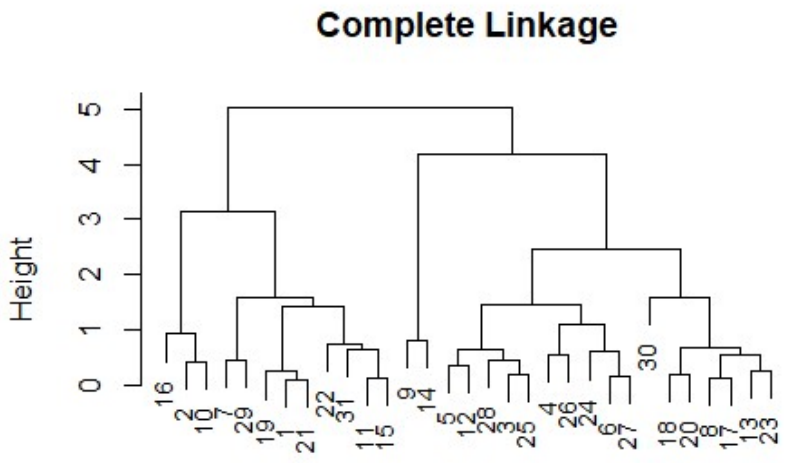

FIGURA 32: DENDOGRAMA LINKAGE 'COMPLETE'

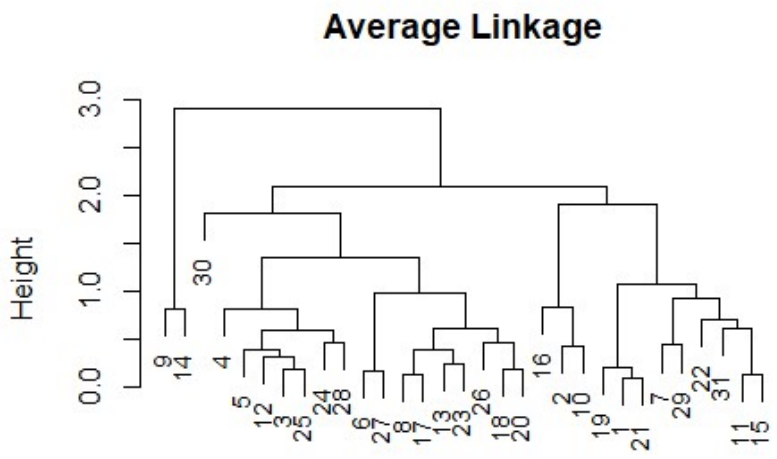

FIGURA 33: DENDOGRAMA LINKAGE 'AVERAGE'

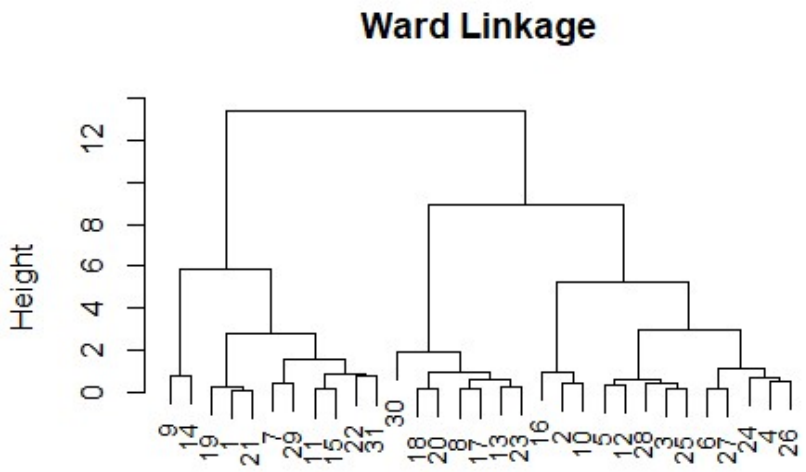

FIGURA 34: DENDOGRAMA LINKAGE 'WARD'

TABLA 10: DENDOGRAMAS, ASIGNACIÓN DE PAÍSES

Austria	**Belgium**	**Bulgaria**	**Croatia**	**Cyprus**	**Czechia**	**Denmark**	**Estonia**
1	2	3	4	5	6	7	8
Finland	**France**	**Germany**	**Greece**	**Hungary**	**Iceland**	**Ireland**	**Italy**
9	10	11	12	13	14	15	16
Latvia	**Lithuania**	**Luxembourg**	**Malta**	**Netherlands**	**Norway**	**Poland**	**Portugal**
17	18	19	20	21	22	23	24
Romania	**Slovakia**	**Slovenia**	**Spain**	**Sweden**	**Turkey**	**U.K**	
25	26	27	28	29	30	31	

Observando el dendograma correspondiente al linkage 'complete' parece que un buen punto de partida sería 5 clústeres. Da la sensación de grupos más concentrados. Aun así se distinguirían con ese punto de partida ya los más aislados, los correspondientes a los números {2, 10,16}, al {9,14} y el 30 (Bélgica, Francia, Italia, zona superior derecha, los dos siguientes Finlandia e Islandia, esquina inferior izquierda, y el último Turquía aislado en la esquina inferior derecha).

En el caso del "average" parece un buen punto de partida 6 clústeres. Se siguen distinguiendo los grupos comentados más separados. Por último para el "Ward" se necesitaría un valor más alto de partida, entre 6 y 7, de hecho más alto todavía para distinguir los 3 grupos aislados.

Decidimos para los 3 tomar en un primer momento el número máximo a probar de 8 clústeres, excepto para el "Ward" comentado que llegamos en principio hasta 9. Para comparar se dibujan las gráficas para todos los valores de clústeres (corte indicado a la función 'cutree'). Las **distintas gráficas** correspondientes a distintos números de clústeres **para cada tipo de 'linkage'** (por su elevado número) se pueden consultar en el **ANEXO IV**.

Vistas todas las gráficas en primer lugar descartamos el linkage "Ward" ya que para llegar a la distinción de los 3 grupos particulares necesita un número tan alto (8) que implica en nuestra opinión una excesiva división en la parte derecha (sin grandes diferencias en las coordenadas).

En el caso del "complete" con 6 ya distingue las particularidades. Tiene como negativo que la disposición mayoritaria a la derecha la divide en 2 grupos, el superior con disparidad destacable del signo de F3 dentro del grupo.

Por el contrario, para **el caso "average"** observamos incluso más posibilidades. Con 6 ya se diferencian los grupos particulares más separados, y al contrario del anterior lo hace con mayor uniformidad de signo entre los dos grupos de la derecha. Si llegamos a 7 a la anterior se añade en el cuadrante superior izquierda el distinguir Holanda, Austria y Luxemburgo como grupo con coordenadas prácticamente iguales. Y si llegamos a 8 a todo lo anterior distingue dentro de la distribución de la derecha aparte de dos grupos con mayor uniformidad de signo, el grupo formado por República Checa y Eslovaquia, prácticamente ambos en el centro (origen de coordenadas) de la gráfica. Podríamos por lo tanto tomar los 2 extremos de división, con 6 y con 8.

Por lo anteriormente comentado nos quedamos finalmente con la **agrupación correspondiente a la gráfica del caso linkage "average" con 8 clústeres** ya que, mediante uniones de ciertos grupos, nos permitirá también hacernos una idea en el caso de disminuir hasta 6 clústeres. La agrupación será comentada en el **apartado 7.2** de las **Conclusiones**, vemos a continuación la gráfica correspondiente:

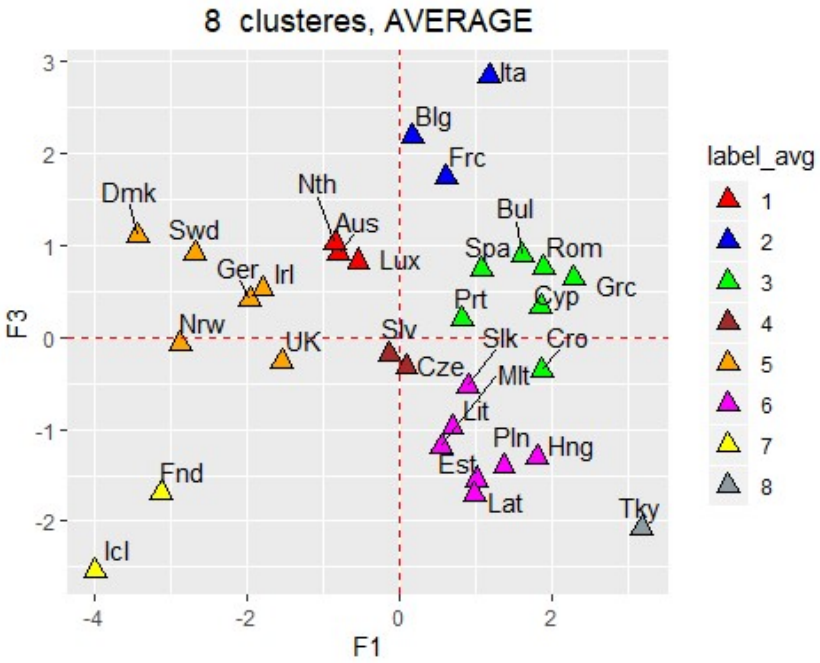

FIGURA 35: AGRUPACIÓN (CLUSTERING) PAÍSES FACTORES 'F1' Y 'F3'

7. CONCLUSIONES.

7.1 Análisis global de la influencia de las distintas variables en el contexto europeo. Regresión logística, Variables determinantes, valoración de los odds-ratio.

Comprobada ya anteriormente la bondad de nuestro modelo de regresión logística que nos permitirá detectar las variables determinantes, pasamos a la evaluación de sus odds-ratio calculados. Es necesario recordar que las unidades de estas variables son porcentajes, sobre el PIB para la variable 'Gasto sanitario' y para el resto porcentaje de población que cumple o se encuentra en la situación que indica el nombre de la variable.

TABLA 11: FACTORES DETERMINANTES, ANÁLISIS ODDS-RATIO

BMI_Overweight	*DailyFyV_5.portions.or.more*	*smoking_Occasional.smoker*
0.9415207	*1.0906083*	*1.6583096*
social_support_Poor	*Chron_ill_pct*	*Health_Expend_._GDP*
0.8899614	*0.8718911*	*1.4559706*
	SEX_RFemales	
	0.2161924	

Como vemos **las variables que afectan negativamente a la auto valoración positiva del estado de salud ('SPHpos')** con odds-ratio inferior a la unidad **son 'BMI_Overweight', 'social_support_Poor', 'Chron_ill_pct' y 'SEX_RFemales'** (en referencia al género masculino).

Las que afectan en forma positiva con odds-ratio mayor que la unidad **son 'DailyFyV_5.portions.or.more', 'Health_Expend_._GDP' y sorpresivamente 'smoking_Occasional.smoker'**

Pasamos a comentar cada una de ellas:

- Índice de masa corporal – Sobrepeso (BMI_Overweight)

Como vemos por su odds-ratio (0.9415) el sobrepeso afecta negativamente a la autovaloración del estado de salud (SPHpos). El valor nos indica que por cada aumento en la variable independiente de un punto porcentual la probabilidad de tener valoración positiva se multiplica por 0.94, es decir baja casi un 6% respecto a la anterior del aumento del punto porcentual.

Entre los factores negativos es el que influye de forma más leve.

- DailyFyV_5.portions.or.more

Tiene un odds-ratio de 1.0906. Por lo tanto, tomar al menos la dosis diaria recomendada de frutas y verduras es positivo para la autovaloración del estado de salud, si bien la influencia es de grado leve.

Por cada aumento de un punto porcentual en el consumo la probabilidad de tener valoración positiva aumenta casi en un 9.06% respecto a antes del aumento.

- smoking_Occasional.smoker

Con un odds-ratio de 1.6583, nos indica para nuestra sorpresa que tal condición (porcentaje de población con ésta) es un factor positivo para la autovaloración del estado de salud. Situando este dato en su contexto debemos recordar que ya en el análisis exploratorio, y tanto para los fumadores ocasionales como los diarios, los valores más altos de porcentaje fumadora se encontraban en la distribución de los que referían valoración positiva de salud.

Dado el carácter de 'auto percibida' del estado de salud de la SPH, junto con lo anterior comentado y dado el comprobado perjuicio sobre la salud del tabaco, todo esto podría indicar una mala o deficiente percepción por parte de la población fumadora (no achacable a la falta de información, presente de sobra) del perjuicio de su consumo, o simplemente de su decisión de ignorar este factor.

Cada aumento de un punto porcentual de población fumadora ocasional multiplica la probabilidad de tener valoración positiva por 1.66 respecto a antes del aumento. Aumento que se pudiera parecer excesivo y que situamos dentro del contexto explicado anteriormente.

- social_support_Poor

En este caso tenemos un odds-ratio de prácticamente 0.89. Como era esperable y se dejaba entrever la posibilidad ya en el análisis exploratorio, la percepción de falta de apoyo social (del entorno) afecta negativamente a la autovaloración del estado de salud.

Este valor nos dice que por cada aumento de un punto porcentual de la población con esta circunstancia la probabilidad de tener valoración positiva baja en un 11% respecto a antes del aumento.

- Chron_ill_pct

Su odds-ratio es de 0.8719. Del todo esperable, la condición de vivir bajo una enfermedad o dolencia crónica afecta negativamente a la autovaloración del estado de salud.

Por cada aumento de un punto porcentual de la población en esta condición la probabilidad de tener valoración positiva baja en un 12.81% respecto al anterior.

La enfermedad crónica junto con la anterior de apoyo social escaso son los factores que más inciden en una valoración negativa de la salud.

- Health_Expend_._GDP

Tiene un odds-ratio de 1.4559. Este valor parece indicar que el esfuerzo en aumento del gasto en sanidad obtiene resultados, reflejándose en una autovaloración positiva del estado de salud de la población.

El aumento de un punto porcentual sobre el PIB del Gasto Sanitario supone aumentar la probabilidad de tener valoración positiva en un nada despreciable 45.6% sobre el anterior. Este grado de influencia pudiera justificarse con los mejores y en mayor número recursos sanitarios ofrecidos a su población si se aumenta el gasto y la percepción de la población ante esta mejora.

- SEX_RFemales

Esta variable tiene un odds-ratio de 0.2161924, tomado como referencia al masculino. La disminución en la auto percepción del estado de salud de las mujeres respecto a la de los hombres es considerable. Llama la atención la diferencia con lo observado en el análisis exploratorio, en el cual no había diferencia por sexo.

Quizás esta circunstancia se deba a que en el exploratorio considerábamos a la población en conjunto, mientras que ahora estamos en el nuevo contexto de las variables encontradas significativas. Una diferencia similar entre los resultados finales y el análisis exploratorio en el caso de variables sociodemográficas fue encontrada en (Camarero, Almazán, & Mañas).

La probabilidad de tener valoración positiva al ser mujer disminuye en un 78.4% respecto a la del hombre.

Como resumen final dentro de los factores determinantes de la auto valoración positiva del estado de salud, los de mayor influencia son a favor el Gasto Sanitario y en contra la situación de padecer Enfermedad Crónica y la de tener un Apoyo Social del entorno pobre.

A favor tenemos en menor medida el Consumo Diario de Frutas y Verduras recomendado (5 porciones o más), y de mayor influencia la sorpresa de la condición de Fumador Ocasional, debida al carácter ya comentado de autovaloración del indicador.

En contra en menor medida la situación de Sobrepeso y señalar la considerable diferencia del género femenino respecto al masculino.

7.2 Análisis por países. Agrupamiento (clustering) a partir de los factores 'F1' y 'F3' del Análisis de componentes principales (PCA)

Partiendo de la mejor opción encontrada como vimos anteriormente, agrupamiento jerárquico aglomerativo con linkage "average" y número de clústeres 8, pasamos a analizar los distintos grupos encontrados:

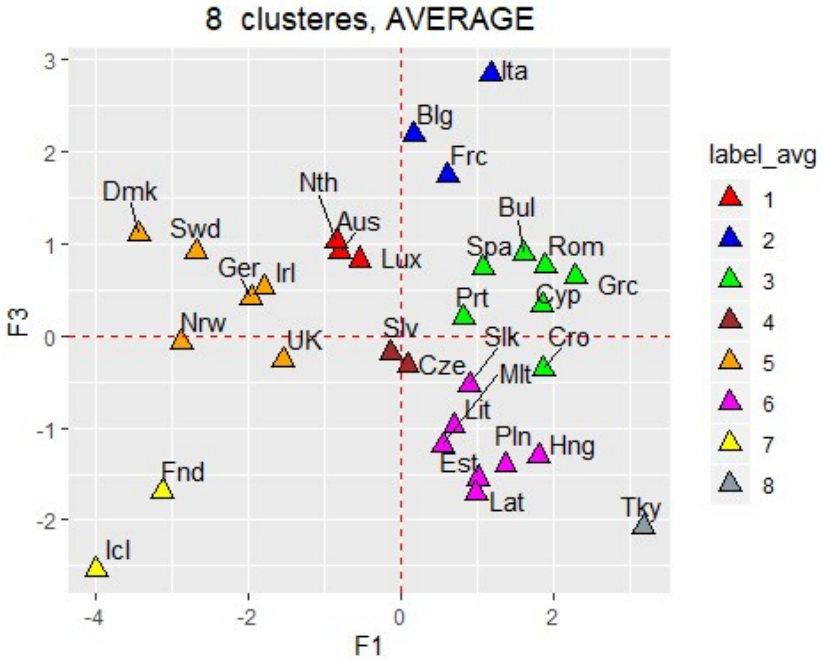

FIGURA 36: ANÁLISIS CLUSTERING POR PAÍSES

Recordemos antes el **carácter de los factores 'F1' y 'F3'**:

- **F1:** "Apoyo social insuficiente frente al Gasto sanitario y hábitos de vida considerados positivos". La valoración positiva del estado de salud (SPHpos) se corresponde con el signo negativo.

- **F3:** “Gasto sanitario frente a mala condición física (sobrepeso, enfermedad crónica)”. La valoración positiva del estado de salud (SPHpos) se corresponde con el signo positivo.

Dadas estas relaciones de los signos de ‘F1’ y ‘F3’ con la valoración positiva de salud, en la gráfica la mejor situación la tendrán los países que se encuentren en el cuadrante superior izquierda, y se encontrarán con la peor valoración los del cuadrante inferior derecha. Pasamos a continuación a analizar los clústeres:

1.- En la situación privilegiada del cuadrante superior izquierda nos encontramos con 2 grupos:

A) {Alemania, Dinamarca, Noruega, Suecia, Reino Unido e Irlanda}: Mejor grupo con los valores óptimos, valores negativos medios-altos de F1 y positivos medio-bajos de F3.

Todos se encuentran entre los valores más altos de Gasto Sanitario.

Entre los hábitos considerados positivos también todos con valores altos-medios de Actividad Física, Bebedores Habituales y Fumadores Ocasionales (en este último caso a excepción de Reino Unido que es muy bajo). En cuanto al Consumo de Frutas y Verduras recomendado encontramos dos comportamientos dentro del grupo, de los valores más altos (los 3 de cabeza) para Reino Unido, Irlanda y Dinamarca, y en el grupo de los medios-bajos para Alemania, Noruega y Suecia. Este doble comportamiento vendría a explicar el que este grupo sea menos concentrado, esté algo más estirado. En global buenos valores para los hábitos considerados positivos.

Apoyo Social insuficiente en valores medio en el caso de Alemania y bajos en el resto. Hablando del Sobrepeso tenemos valores bajos excepto para Irlanda y Reino Unido. En cuanto a la tasa de Enfermedad crónica valores medios-altos excepto para Irlanda y Dinamarca. Estas condiciones de sobrepeso y enfermedad crónica explicarían los valores medio-bajos de F2. Al igual las excepciones comentadas justificarían que este clúster esté menos concentrado, un poco más estirado.

El grupo mejor situado con todos valores en general óptimos para la valoración positiva.

B) {Holanda, Austria y Luxemburgo}: valores bajos, negativos para F1 y positivos para F3.

Presenta la particularidad para el Gasto Sanitario de valores entre los más altos para Holanda y Austria y de los menores para Luxemburgo. Esta diferencia se deba seguramente al muy pequeño (aún frente a sus compañeros) tamaño del país y la unidad de medida como porcentaje del gasto sobre el PIB, que será también inferior.

Valores medios-altos para la Actividad Física y Fumadores Ocasionales. Valores altos para Holanda y Luxemburgo y bajo para Austria en el Consumo de Frutas y Verduras, y medio bajo para Austria y Holanda en Bebedores Frecuentes, con la excepción de alta en Luxemburgo, no sabemos si influenciado por su particularidad comentada.

Valores medios-bajos de Apoyo Social insuficiente para los 3. Valores entre los más bajos de sobrepeso y medio-altos (excepto Luxemburgo bajo) de tasa de enfermedad crónica.

Los valores relativos a los hábitos de vida considerados positivos más bajos respecto al anterior grupo, y mayor tasa de enfermedad crónica justificarían la buena posición pero con valores bajos de F1 y F3 óptimos.

2.- Excepción del cuadrante inferior derecha:

C) {Islandia y Finlandia}: grupo con comportamiento particular, valores óptimos de F1 y malos (medio-altos negativos) para F3.

Valores alto y medio-alto respectivamente de Gasto Sanitario.

Ambos tienen dentro de los hábitos auto considerados positivos de los más altos los correspondientes a Fumador Ocasional, Bebedor Frecuente y Actividad Física, con valores medios de Consumo de Frutas y Verduras al menos el recomendado (igual o mayor a 5 porciones).

Valores de apoyo social insuficiente medio para Finlandia y el más bajo de todos los países para Islandia. En cuanto al sobrepeso valor muy alto (segundo del ranking) para Islandia y medio-alto para Finlandia. Tasa de Enfermedad Crónica la más alta (primera del ranking) para Finlandia y media para Islandia.

Son las condiciones físicas (sobrepeso y enfermedad crónica), con ambos países en primeros lugares y medio-altos los que explicarían el signo tan negativo de F3 y el aislamiento en la esquina inferior izquierda de este grupo, con el resto de factores (gasto sanitario y hábitos) óptimos.

3.- Cuadrante superior derecha:

D) {Francia, Bélgica e Italia}: grupo con valores ligeramente malos de F1 (valores cortos positivos) y de los mejores (positivos altos) para F3.

Valores altos de Gasto Sanitario para los dos primeros (Francia primero del ranking de hecho) y medio para Italia.

En cuanto a los hábitos considerados positivos valores medios y medios bajos en el Consumo de Frutas y Verduras, Actividad Física y Fumadores Ocasionales. Presenta particularidad para los Bebedores Habituales, con valor medio alto para Bélgica y bajos para Francia e Italia.

Presentan valores medio-altos de Apoyo Social insuficiente. En cambio en Sobrepeso ocupan los últimos valores del ranking y en cuanto a la Enfermedad crónica valores bastante bajos excepto para Francia con medio-alto.

El apoyo Social insuficiente medio-alto y valores en general medios-bajos de los hábitos considerados positivos explican los valores de F1 ligeramente malos. El predominante alto Gasto sanitario y la buena condición física (última en sobrepeso y baja excepto Francia en enfermedad Crónica) explican que tengan los mejores valores de F3.

E) {España, Portugal, Grecia, Chipre, Croacia, Bulgaria y Rumanía}: grupo con valores cortos y medios positivos de F1 (leve-medio mal comportamiento) y cortos óptimos (positivos cortos) para F3.

Valores en el rango entre medios y bajos, con los valores superiores (por encima de la mitad en el ranking) para España y Portugal y el peor valor (penúltimo del ranking) para Rumanía. Disparidad por lo tanto en este aspecto.

Valores de nuevo entre medios y bajos de Actividad Física (liderando en medio del ranking España). Para el Consumo de Frutas y Verduras valores medio-altos para España, Portugal y Chipre y de los últimos del ranking para el resto. Valores en general medios-bajos de bebedores frecuentes con la excepción de Rumania (altos) y por último para los Fumadores Ocasionales valores bajos excepto para Bulgaria, Rumania y Grecia (medio-altos).

En el Apoyo Social insuficiente se observan dos subgrupos. España, Croacia y Chipre con excelente comportamiento (de los últimos en el ranking) y el resto con valores medio-altos. Para el sobrepeso valores en general medios-bajos (excepto Grecia y Croacia altos). Por último la tasa de Enfermedad Crónica presenta valores en general bajos excepto en Portugal y Chipre.

Observamos pues en este grupo tendencias de la mitad hacia abajo en Gasto Sanitario, en general medios-bajos (con excepciones) en los Hábitos considerados positivos y medio bajos (de nuevo excepciones) en Sobrepeso y Enfermedad Crónica. Comportamiento dispar en el clúster en cuanto al Apoyo Social insuficiente. Son los valores en general medios-bajos de Gasto Sanitario y Hábitos y la aportación que pueda hace el Apoyo Social los que hacen tener valor leve-moderadamente malo de F1. El buen comportamiento en general de la condición física Sobrepeso, Enfermedad crónica) se ven seguramente frenados por al Gasto sanitario medio-bajo. Importante señalar que las numerosas excepciones explicarían cierto grado de disgregación en el clúster (y la variación de comportamiento que conlleva), con España en una posición privilegiada dentro del mismo.

4.-Peor situación correspondiente al cuadrante inferior derecha:

F) {Estonia, Letonia, Lituania, Polonia, Hungría, Eslovaquia y Malta}: grupo con valores cortos y medios positivos de F1 (leve-medio mal comportamiento) y de cortos a medio-altos negativos (mal comportamiento) para F3.

Valores del gasto sanitario entre los últimos de ranking a excepción de Malta, que se explicaría al igual que comentamos con Luxemburgo con el pequeño tamaño del país y la unidad del gasto en porcentaje del PIB.

Presentan valores entre medios y bajos de Actividad Física y Bebedores Habituales, igual que para Fumadores Ocasionales (excepción en este último caso de Eslovenia medio-alto). En el Consumo de Frutas y Verduras comportamiento en general medio excepto Eslovenia, Malta y Lituania (medio-altos)

Valores entre los más altos del ranking de Apoyo Social insuficiente (excepción de Hungría y Malta medios). En cuanto al Sobrepeso valores medio-altos y el mayor valor del ranking para Malta. La tasa de Enfermedad Crónica presenta por lo general valores medios con 2 excepciones a destacar, Estonia (segunda en el ranking) y Letonia (quinta).

Justificación clara de la mala situación de este clúster, con Gastos Sanitarios bajos, valores en general medio-bajos de Hábitos considerados positivos y altos de Apoyo Social insuficiente, con valores medio-altos y algunas altas posiciones en el ranking para la mala condición física (Sobrepeso y Enfermedad Crónica). De nueva las excepciones justificarían cierto disgregamiento en el clúster.

G) {Turquía}: aislada en la esquina inferior derecha con valor alto positivo de F1 y alto negativo para F3 (ambos el peor comportamiento)

Para empezar presenta el último valor en el ranking de Gasto Sanitario.

Último en el ranking para el Consumo de Frutas y Verduras, penúltimo en el de Actividad Física y Bebedores Habituales, sólo presenta valor medio-bajo en Fumadores Ocasionales en cuanto a los Hábitos considerados positivos.

Quinto en el ranking de Apoyo Social insuficiente y tercero tanto en el de Sobrepeso como en el de tasa de Enfermedad Crónica terminan de situar a Turquía en la peor de las posiciones de autovaloración positiva de salud.

5.- Excepción origen de coordenadas: valores de F1 y F3 prácticamente cero.

H) {Eslovenia y República Checa}: comportamiento singular, grupo con valores prácticamente cero para ambos F1 y F3.Estos países presentan coordenadas (-0.14,-0.18) Eslovenia y (0.08,-0.31) para la República Checa.

Presentan ambos valores medios de Gasto Sanitario.

Entre los hábitos considerados positivos valores bajos de Consumo de Frutas y Verduras y medios-bajos de Bebedores Habituales, con Eslovenia algo mejor (medio-alto) respecto al medio de la República Checa en Actividad Física y mayor diferencia a favor de la República Checa (segundo en el ranking respecto valor medio) en Fumadores Ocasionales. En global preponderancia de valores medios-bajos en los hábitos.

Valores de Apoyo Social insuficiente medio-bajos, al igual que de tasa de Enfermedad Crónica. Finalmente, valores medio-altos de Sobrepeso.

La tendencia a presentar valores medios o cercanos a los medios para todas las variables, con algunas excepciones que parecen compensarse entre ellas, vendría a justificar la posición intermedia respecto a los factores. Hablando hipotéticamente dada su posición en la gráfica podría tratarse de países que han evolucionado positivamente en transición del grupo (F) al (E).

6.- Resumen final.

Los países pertenecientes a **los grupos "A" y "B" (Alemania, Dinamarca, Noruega, Suecia, Reino Unido, Irlanda, Holanda, Austria y Luxemburgo), los mejor posicionados en cuanto a la totalidad de las variables** (con alguna excepción por otro lado lógica) **están en el Centro y Norte de Europa**. **Faltarían** por la región para completar el **grupo "C" con Islandia y Finlandia**,

ambos lastrados y separados de los anteriores por la condición física (valores de Sobrepeso y Enfermedad Crónica), compartiendo la bondad de las restantes.

A continuación nos encontramos con **el grupo "D" (Francia, Bélgica e Italia),del centro de Europa excepto el último**, con algo peores valores de hábitos de vida considerados positivos y Apoyo Social insuficiente medio-alto, compartiendo con los anteriores el Gasto Sanitario alto y buena condición física (bajo Sobrepeso y tasa de Enfermedad Crónica) en general.

Seguimos con el grupo "E" (España, Portugal, Grecia, Chipre, Croacia, Bulgaria y Rumanía), **países del sur de Europa a excepción de los dos últimos (este).** El Gasto Sanitario entre medio y bajo, en general valores medios-bajos de Hábitos considerados positivos y valores medios de Apoyo Social insuficiente son los que, a pesar de una no mala condición física (valores medios-bajos de Sobrepeso y bajos en general de tasa de Enfermedad Crónica) sitúan a este grupo un peldaño por debajo del anterior.

El grupo "F" (Estonia, Letonia, Lituania, Polonia, Hungría, Eslovaquia y Malta) presenta como características Gastos Sanitarios bajos, valores en general medio-bajos de Hábitos considerados positivos y altos de Apoyo Social insuficiente. Acompañado de valores medio-altos y algunas altas posiciones en el ranking de mala condición física (Sobrepeso y Enfermedad Crónica), sitúan a este grupo entre los últimos en autovaloración positiva. Señalar que **con la excepción de Malta todas fueron pertenecientes al antiguo bloque soviético.**

Llegando al final nos encontramos por un lado **aislada** a **Turquía (grupo "G"),** con los peores resultados en todas las variables: Gasto y Hábitos considerados positivos en las últimas y penúltimas posiciones del ranking, mala Condición Física con terceros puestos en el ranking de Sobrepeso y Enfermedad Crónica, y quinta posición en el de Apoyo Social insuficiente.

Por otro finalmente tenemos como excepción al **grupo "H" (Eslovenia y República Checa)**con en general valores medios o cercanos a los medios para todas las variables, con algunas excepciones que parecen compensarse para dar como final una posición intermedia en general para todos los factores. **Su posición en el este de Europa y condición de ex componentes del bloque soviético** hacen compararse con el grupo "F" estando un peldaño por encima y relativamente próximos al grupo "E".

En general parecen continuar las diferencias entre Norte y Centro Europa por un lado, Sur de Europa por otro, y finalmente los países del Este, siendo menores entre los dos primeros grupos y más palpable en el caso del Este.

7.3 Visualización interactiva de los datos del proyecto (descriptivos y resultado del análisis) en Tableau.

Como parte de este Trabajo Fin de Máster se ha preparado una visualización interactiva de los datos, tanto los descriptivos a partir de los datos originales como los resultantes del análisis realizado. Esta visualización se compone de 4 vistas o "dashboards".

En la primera y como punto de partida se presenta una panorámica global con los valores promedio por país de SPH, así como su distribución por género y grupos de edad. De esta

forma y con la interactividad podremos comparar la SPH de distintos países y descubrir para cada uno su distribución por las dos dimensiones mencionadas de género y edad.

En la segunda, para la selección de país o países hecha en el lateral derecho, se distribuyen las 12 variables independientes del estudio (previas al análisis de los odds-ratio), agrupadas por Sobrepeso y Hábitos saludables (Actividad Física, Consumo Diario de Frutas y Verduras); Hábitos nocivos (alcohol, tabaco); Enfermedad Crónica, Apoyo Social y Uso de medicinas no prescribidas y por último Gasto sanitario. Si la selección es múltiple podremos comparar dichas variables entre distintos países.

En una tercera vista lo que se pretende es poder establecer para cada una de las 12 variables un ranking de países del estudio. Seleccionando la variable en cuestión de interés en la esquina superior derecha podremos ver de mayor a menor los valores de la variable para cada país y ver su distribución por el ámbito europeo. Se permitirá también la selección simple o múltiple de países en el mapa y de esta forma distinguir y comparar sus posiciones en dicho ranking.

En la cuarta vista se muestran los resultados del análisis realizado en "R", valor de las variables detectadas como (factores) determinantes y clusterización de los países según factores 'F1' y 'F3' de la descomposición PCA. Podremos contextualizar cada país dentro de la asignación de clústeres, y al hacer selección en el mapa podremos ver los valores de los factores determinantes para dicho país en la parte inferior. Si la selección es múltiple los podremos comparar entre países. La vista también nos permitirá hacer selección de clúster o clústeres (esquina superior izquierda) para ver su situación en el mapa y en la gráfica de los factores (F1, F3).

El enlace para acceder a dicha visualización es el siguiente:

https://public.tableau.com/profile/eduardo.rivero#!/vizhome/TFM_Eduardo_Rivero/Visualizacin_TFM?publish=yes

7.4 Líneas de trabajo futuras.

Algunas líneas de trabajo futuras podrían ser en primer lugar y dependiendo de la completitud los datos recopilados por Eurostat ampliar el número de países (Suiza, Serbia, Bosnia y Herzegovina, Ucrania).

También se podría estudiar en este caso la mortalidad por número de habitantes y verificar la relación que hay con la auto valoración del estado de salud (SPH) como mencionan bastantes autores de la bibliografía consultada.

Por último, otra opción sería realizar un estudio comparativo de los distintos sistemas de salud existentes en Europa y explorar posibles efectos en la SPH y la mortalidad.

8. ANEXO I. Correlación y pruebas de significación estadística de las variables independientes numéricas iniciales.

```
rcorr(as.matrix(var.numericas.HD))

##                                    BMI_UnderweightBMI_Normal          BMI_Overweight
## BMI_Underweight                              1.00       0.78                   -0.85
## BMI_Normal                                   0.78       1.00                   -0.99
## BMI_Overweight                              -0.85      -0.99                   1.0
## Physical_activity                            0.18       0.36                   -0.34
## DailyFyV_0.portions                          0.04       0.13                   -0.12
## DailyFyV_From.1.to.4.portions               -0.06      -0.17                   0.16
## DailyFyV_5.portions.or.more                  0.03       0.05                   -0.04
## smoking_Non.smoker                           0.12      -0.06                   0.03
## smoking_Daily.smoker                        -0.21      -0.09                   0.11
## smoking_Occasional.smoker                    0.25       0.49                   -0.46
## Frequent_drink_pct                          -0.10       0.02                   0.01
## social_support_Intermediate                 -0.05      -0.06                   0.06
## social_support_Poor                         -0.26      -0.31                   0.31
## social_support_Strong                        0.16       0.20                   -0.20
## Chron_ill_pct                               -0.46      -0.67                   0.65
## Sev_body_pain_Mild.or.very.mild             -0.18      -0.20                   0.21
## Sev_body_pain_Severe.or.very.severe         -0.31      -0.47                   0.46
## Sev_body_pain_Moderate                      -0.33      -0.49                   0.48
## Sev_body_pain_None                           0.33       0.46                   -0.45
## Health_Expend_._GDP                          0.03       0.08                   -0.07
## presc_med_use_pct                           -0.41      -0.63                   0.61
## non_presc_med_use_pct                       -0.07      -0.03                   0.04

##                    Physical_activity        DailyFyV_0.portions
## BMI_Underweight                            0.18                      0.04
## BMI_Normal                                 0.36                      0.13
## BMI_Overweight                             -0.34                     -0.12
## Physical_activity                          1.00                      0.17
## DailyFyV_0.portions                        0.17                      1.00
## DailyFyV_From.1.to.4.portions              -0.19                     -0.82
## DailyFyV_5.portions.or.more                0.01                      -0.41
## smoking_Non.smoker                          -0.01                    -0.43
## smoking_Daily.smoker                        -0.15                    0.31
## smoking_Occasional.smoker                  0.53                      0.52
## Frequent_drink_pct                          0.49                     0.40
## social_support_Intermediate                 -0.28                    0.36
## social_support_Poor                         -0.26                    0.21
## social_support_Strong                      0.30                      -0.33
## Chron_ill_pct                               -0.30                    -0.38
## Sev_body_pain_Mild.or.very.mild             0.22                     0.08
## Sev_body_pain_Severe.or.very.severe         -0.27                    -0.51
## Sev_body_pain_Moderate                      -0.19                    -0.34
## Sev_body_pain_None                         0.05                      0.25
## Health_Expend_._GDP                        0.39                      -0.20
## presc_med_use_pct                           -0.26                    -0.52
## non_presc_med_use_pct                       0.30                     -0.02
##                                    DailyFyV_From.1.to.4.portions      DailyFyV_5.portions.or.more
## BMI_Underweight                                            -0.06                         0.03
## BMI_Normal                                                 -0.17                         0.05
## BMI_Overweight                                              0.16                         -0.04
## Physical_activity                                          -0.19                         0.01
## DailyFyV_0.portions                                        -0.82                         -0.41
## DailyFyV_From.1.to.4.portions                               1.00                         -0.19
## DailyFyV_5.portions.or.more                                -0.19                         1.00
## smoking_Non.smoker                                          0.28                         0.30
## smoking_Daily.smoker                                       -0.15                         -0.29
## smoking_Occasional.smoker                                  -0.48                         -0.12
## Frequent_drink_pct                                         -0.36                         -0.12
## social_support_Intermediate                                -0.28                         -0.18
## social_support_Poor                                        -0.20                         -0.03
## social_support_Strong                                       0.27                         0.12
## Chron_ill_pct                                               0.32                         0.14
## Sev_body_pain_Mild.or.very.mild                            -0.27                         0.28
## Sev_body_pain_Severe.or.very.severe                         0.44                         0.18
## Sev_body_pain_Moderate                                      0.30                         0.10
## Sev_body_pain_None                                         -0.12                         -0.24
## Health_Expend_._GDP                                         0.02                         0.32
## presc_med_use_pct                                           0.43                         0.20
## non_presc_med_use_pct                                      -0.10                         0.19


##                                    smoking_Non.smoker        smoking_Daily.smoker
## BMI_Underweight                                 0.12                  -0.21
## BMI_Normal                                     -0.06                  -0.09
```

```
## BMI_Overweight                                    0.03                0.11
## Physical_activity                                -0.01                -0.15
## DailyFyV_0.portions                              -0.43                0.31
## DailyFyV_From.1.to.4.portions                     0.28                -0.15
## DailyFyV_5.portions.or.more                       0.30                -0.29
## smoking_Non.smoker                                1.00                -0.96
## smoking_Daily.smoker                             -0.96                1.00
## smoking_Occasional.smoker                        -0.42                0.15
## Frequent_drink_pct                               -0.28                0.14
## social_support_Intermediate                      -0.18                0.18
## social_support_Poor                               0.00                0.03
## social_support_Strong                             0.11                -0.12
## Chron_ill_pct                                     0.55                -0.41
## Sev_body_pain_Mild.or.very.mild                   0.26                -0.32
## Sev_body_pain_Severe.or.very.severe               0.55                -0.44
## Sev_body_pain_Moderate                            0.54                -0.45
## Sev_body_pain_None                               -0.54                0.49
## Health_Expend_._GDP                               0.16                -0.22
## presc_med_use_pct                                 0.61                -0.47
## non_presc_med_use_pct                             0.25                -0.29

##                                      smoking_Occasional.smoker          Frequent_drink_pct
## BMI_Underweight          0.25                        -0.10
## BMI_Normal               0.49                        0.02
## BMI_Overweight          -0.46                        0.01
## Physical_activity        0.53                        0.49
## DailyFyV_0.portions     0.52                         0.40
## DailyFyV_From.1.to.4.portions                    -0.48                           -0.36
## DailyFyV_5.portions.or.more                      -0.12                           -0.12
## smoking_Non.smoker-0.42                    -0.28
## smoking_Daily.smoker     0.15                        0.14
## smoking_Occasional.smoker                         1.00                           0.52
## Frequent_drink_pct       0.52                        1.00
## social_support_Intermediate                       0.08                           0.03
## social_support_Poor     -0.07                        0.00
## social_support_Strong   -0.01                        -0.02
## Chron_ill_pct           -0.61                        -0.32
## Sev_body_pain_Mild.or.very.mild                    0.10                          0.14
## Sev_body_pain_Severe.or.very.severe               -0.55                          -0.41
## Sev_body_pain_Moderate  -0.47                        -0.23
## Sev_body_pain_None       0.31                        0.15
## Health_Expend_._GDP      0.14                        0.10
## presc_med_use_pct       -0.65                        -0.35
## non_presc_med_use_pct    0.07                        0.04

##                                      social_support_Intermediate        social_support_Poor
## BMI_Underweight                                          -0.05                 -0.26
## BMI_Normal                                               -0.06                 -0.31
## BMI_Overweight                                            0.06                 0.31
## Physical_activity                                        -0.28                 -0.26
## DailyFyV_0.portions                                       0.36                 0.21
## DailyFyV_From.1.to.4.portions                            -0.28                 -0.20
## DailyFyV_5.portions.or.more                              -0.18                 -0.03
## smoking_Non.smoker                                       -0.18                 0.00
## smoking_Daily.smoker                                      0.18                 0.03
## smoking_Occasional.smoker                                 0.08                 -0.07
## Frequent_drink_pct                                        0.03                 0.00
## social_support_Intermediate                               1.00                 0.59
## social_support_Poor                                       0.59                 1.00
## social_support_Strong                                    -0.90                 -0.88
## Chron_ill_pct                                            -0.04                 0.26
## Sev_body_pain_Mild.or.very.mild                           0.04                 0.28
## Sev_body_pain_Severe.or.very.severe                      -0.06                 0.20
## Sev_body_pain_Moderate                                   -0.01                 0.27
## Sev_body_pain_None                                       -0.01                 -0.32
## Health_Expend_._GDP                                      -0.32                 -0.19
## presc_med_use_pct                                        -0.15                 0.20
## non_presc_med_use_pct                                    -0.05                 -0.06
          _                  support_Strong     Chron_ill_pct      Sev_body_pain_Mild.or.very.mild
## BMI_Underweight      0.16         -0.46             -0.18
## BMI_Normal           0.20         -0.67             -0.20
## BMI_Overweight      -0.20          0.65             0.21
## Physical_activity    0.30         -0.30             0.22
## DailyFyV_0.portions -0.33         -0.38             0.08
## DailyFyV_From.1.to.4.portions        0.27          0.32                -0.27
## DailyFyV_5.portions.or.more          0.12          0.14                0.28
## smoking_Non.smoker                   0.11          0.55                0.26
## smoking_Daily.smoker                -0.12         -0.41                -0.32
## smoking_Occasional.smoker          -0.01         -0.61                0.10
## Frequent_drink_pct                  -0.02         -0.32                0.14
## social_support_Intermediate         -0.90         -0.04                0.04
## social_support_Poor                 -0.88          0.26                0.28
## social_support_Strong                1.00         -0.11                -0.17
## Chron_ill_pct                       -0.11          1.00                0.31
```

```
## Sev_body_pain_Mild.or.very.mild     -0.17          0.31                1.00
## Sev_body_pain_Severe.or.very.severe -0.07          0.80                0.20
## Sev_body_pain_Moderate              -0.14          0.76                0.44
## Sev_body_pain_None   0.18         -0.73             -0.77
## Health_Expend_._GDP  0.29          0.03              0.28
## presc_med_use_pct   -0.02          0.89              0.33
## non_presc_med_use_pct0.06          0.20              0.39

##                                   Sev_body_pain_Severe.or.very.severe           Sev_body_pain_Moderate
## BMI_Underweight                  -0.31                                  -0.33
## BMI_Normal                        -0.47                                 -0.49
## BMI_Overweight                                    0.46                                     0.48
## Physical_activity                                -0.27                                     -0.19
## DailyFyV_0.portions                              -0.51                                     -0.34
## DailyFyV_From.1.to.4.portions                    0.44                                      0.30
## DailyFyV_5.portions.or.more                       0.18                                     0.10
## smoking_Non.smoker                               0.55                                      0.54
## smoking_Daily.smoker                             -0.44                                     -0.45
## smoking_Occasional.smoker                        -0.55                                     -0.47
## Frequent_drink_pct    -0.41                                     -0.23
## social_support_Intermediate                      -0.06                                     -0.01
## social_support_Poor                               0.20                                     0.27
## social_support_Strong                            -0.07                                     -0.14
## Chron_ill_pct                                     0.80                                     0.76
## Sev_body_pain_Mild.or.very.mild                   0.20                                     0.44
## Sev_body_pain_Severe.or.very.severe               1.00                                     0.77
## Sev_body_pain_Moderate                            0.77                                     1.00
## Sev_body_pain_None-0.74                             -0.88
## Health_Expend_._GDP                              -0.01                                     0.06
## presc_med_use_pct                                 0.77                                     0.77
## non_presc_med_use_pct                             0.22                                     0.33

                                   Sev_body_pain_NoneHealth_Expend_._GDP         presc_med_use_pct
## BMI_Underweight                          0.33                0.03                      -0.41
## BMI_Normal                               0.46                0.08                      -0.63
## BMI_Overweight                           -0.45               -0.07                     0.61
## Physical_activity                         0.05                0.39                     -0.26
## DailyFyV_0.portions                      0.25               -0.20                      -0.52
## DailyFyV_From.1.to.4.portions            -0.12                0.02                      0.43
## DailyFyV_5.portions.or.more              -0.24                0.32                      0.20
## smoking_Non.smoker                       -0.54                0.16                      0.61
## smoking_Daily.smoker                     0.49               -0.22                      -0.47
## smoking_Occasional.smoker                0.31                0.14                      -0.65
## Frequent_drink_pct                       0.15                0.10                      -0.35
## social_support_Intermediate              -0.01               -0.32                     -0.15
## social_support_Poor                      -0.32               -0.19                     0.20
## social_support_Strong                    0.18                0.29                      -0.02
## Chron_ill_pct                     -0.73                 0.03                      0.89
## Sev_body_pain_Mild.or.very.mild          -0.77                0.28                      0.33
## Sev_body_pain_Severe.or.very.severe      -0.74               -0.01                      0.77
## Sev_body_pain_Moderate                   -0.88                0.06                      0.77
## Sev_body_pain_None                       1.00               -0.16                      -0.73
## Health_Expend_._GDP                      -0.16                1.00                      0.14
## presc_med_use_pct                        -0.73                0.14                      1.00
## non_presc_med_use_pct                    -0.40                0.01                      0.20

##                                     non_presc_med_use_pct
## BMI_Underweight                                    -0.07
## BMI_Normal                                         -0.03
## BMI_Overweight                                      0.04
## Physical_activity                                   0.30
## DailyFyV_0.portions                                -0.02
## DailyFyV_From.1.to.4.portions                      -0.10
## DailyFyV_5.portions.or.more                         0.19
## smoking_Non.smoker                                  0.25
## smoking_Daily.smoker                               -0.29
## smoking_Occasional.smoker                           0.07
## Frequent_drink_pct                                  0.04
## social_support_Intermediate                        -0.05
## social_support_Poor                                -0.06
## social_support_Strong                               0.06
## Chron_ill_pct                                       0.20
## Sev_body_pain_Mild.or.very.mild                     0.39
## Sev_body_pain_Severe.or.very.severe                 0.22
## Sev_body_pain_Moderate                              0.33
## Sev_body_pain_None                                 -0.40
## Health_Expend_._GDP                                 0.01
## presc_med_use_pct                                   0.20

## non_presc_med_use_pct                               1.00
```

9. ANEXO II. Trazas de los procesos de selección del mejor modelo en la regresión logística.

9.1 Proceso de selección en la dirección 'backward'.

```
mod.logist.completo<-glm(SPHpos~., data =dfHD_train, family = binomial)

mod.stepwise.backward<-stepAIC(mod.logist.completo, direction ="backward", trace =TRUE)

## Start:  AIC=115.08
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     non_presc_med_use_pct + AGE_R + SEX_R
##
##                                Df Deviance    AIC
## - AGE_R                         4   82.925 110.92
## - BMI_Overweight                1   79.130 113.13
## - social_support_Poor           1   79.137 113.14
## - non_presc_med_use_pct         1   79.159 113.16
## - SEX_R                         1   79.337 113.34
## - Frequent_drink_pct            1   79.575 113.58
## - DailyFyV_5.portions.or.more   1   79.594 113.59
## - smoking_Daily.smoker          1   80.133 114.13
## - Physical_activity             1   80.277 114.28
## - DailyFyV_Less.5.portions.     1   80.529 114.53
## <none>                              79.079 115.08
## - smoking_Occasional.smoker     1   81.696 115.70
## - social_support_Intermediate   1   81.821 115.82
## - Health_Expend_._GDP           1   84.827 118.83
## - Chron_ill_pct                 1   87.985 121.98
##
## Step:  AIC=110.92
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     non_presc_med_use_pct + SEX_R
##
##                                Df Deviance    AIC
## - non_presc_med_use_pct         1   82.925 108.92
## - Frequent_drink_pct            1   82.961 108.96
## - Physical_activity             1   83.268 109.27
## - smoking_Daily.smoker          1   83.581 109.58
## - social_support_Poor           1   84.030 110.03
## - social_support_Intermediate   1   84.106 110.11
## - DailyFyV_Less.5.portions.     1   84.344 110.34
## - DailyFyV_5.portions.or.more   1   84.455 110.45
## <none>                              82.925 110.92
## - SEX_R                         1   85.179 111.18
## - BMI_Overweight                1   85.668 111.67
## - Health_Expend_._GDP           1   86.124 112.12
## - smoking_Occasional.smoker     1   89.163 115.16
## - Chron_ill_pct                 1  104.532 130.53
##
## Step:  AIC=108.93
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R
##
##                                Df Deviance    AIC
## - Frequent_drink_pct            1   82.961 106.96
## - Physical_activity             1   83.397 107.40
## - smoking_Daily.smoker          1   83.581 107.58
## - social_support_Intermediate   1   84.176 108.18
## - social_support_Poor           1   84.219 108.22
```

```
## - DailyFyV_5.portions.or.more  1   84.474 108.47
## - DailyFyV_Less.5.portions.    1   84.477 108.48
## <none>                             82.925 108.92
## - SEX_R                        1   85.421 109.42
## - BMI_Overweight               1   85.671 109.67
## - Health_Expend_._GDP          1   86.613 110.61
## - smoking_Occasional.smoker    1   89.179 113.18
## - Chron_ill_pct                1  106.451 130.45
##
## Step:  AIC=106.96
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + social_support_Poor + social_support_Intermediate +
##     Chron_ill_pct + Health_Expend_._GDP + SEX_R
##
##                                Df Deviance    AIC
## - Physical_activity             1   83.402 105.40
## - smoking_Daily.smoker          1   83.632 105.63
## - social_support_Intermediate  1   84.177 106.18
## - social_support_Poor           1   84.272 106.27
## - DailyFyV_Less.5.portions.    1   84.478 106.48
## - DailyFyV_5.portions.or.more  1   84.617 106.62
## <none>                             82.961 106.96
## - BMI_Overweight               1   85.684 107.68
## - SEX_R                        1   86.544 108.54
## - Health_Expend_._GDP          1   86.679 108.68
## - smoking_Occasional.smoker    1   89.949 111.95
## - Chron_ill_pct                1  107.178 129.18
##
## Step:  AIC=105.4
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + smoking_Daily.smoker + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R
##
##                                Df Deviance    AIC
## - smoking_Daily.smoker          1   83.925 103.92
## - social_support_Intermediate  1   84.264 104.26
## - social_support_Poor           1   84.965 104.97
## - DailyFyV_Less.5.portions.    1   85.128 105.13
## <none>                             83.402 105.40
## - BMI_Overweight               1   85.758 105.76
## - DailyFyV_5.portions.or.more  1   85.843 105.84
## - Health_Expend_._GDP          1   86.679 106.68
## - SEX_R                        1   86.717 106.72
## - smoking_Occasional.smoker    1   90.121 110.12
## - Chron_ill_pct                1  107.353 127.35
##
## Step:  AIC=103.92
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + social_support_Poor + social_support_Intermediate +
##     Chron_ill_pct + Health_Expend_._GDP + SEX_R
##
##                                Df Deviance    AIC
## - social_support_Intermediate  1   85.039 103.04
## - DailyFyV_Less.5.portions.    1   85.469 103.47
## - social_support_Poor           1   85.488 103.49
## <none>                             83.925 103.92
## - DailyFyV_5.portions.or.more  1   86.162 104.16
## - SEX_R                        1   86.904 104.90
## - Health_Expend_._GDP          1   87.659 105.66
## - BMI_Overweight               1   87.696 105.70
## - smoking_Occasional.smoker    1   90.447 108.45
## - Chron_ill_pct                1  113.113 131.11
##
## Step:  AIC=103.04
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + social_support_Poor + Chron_ill_pct +
##     Health_Expend_._GDP + SEX_R
##
##                                Df Deviance    AIC
## <none>                             85.039 103.04
## - DailyFyV_Less.5.portions.    1   87.293 103.29
## - SEX_R                        1   88.025 104.03
```

```
## - DailyFyV_5.portions.or.more  1   88.162 104.16
## - BMI_Overweight               1   88.209 104.21
## - Health_Expend_._GDP          1   89.563 105.56
## - social_support_Poor          1   91.580 107.58
## - smoking_Occasional.smoker    1   91.739 107.74
## - Chron_ill_pct                1  113.434 129.43
```

```
summary(mod.stepwise.backward)
```

```
##
## Call:
## glm(formula = SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     social_support_Poor + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R, family = binomial, data = dfHD_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.26688  -0.04643   0.04430   0.15504   2.29398
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   2.31409    3.67165   0.630   0.5285
## BMI_Overweight               -0.06026    0.03665  -1.644   0.1001
## DailyFyV_Less.5.portions.     0.04738    0.03197   1.482   0.1384
## DailyFyV_5.portions.or.more   0.08674    0.05103   1.700   0.0892 .
## smoking_Occasional.smoker     0.50580    0.21277   2.377   0.0174 *
## social_support_Poor          -0.11658    0.04746  -2.457   0.0140 *
## Chron_ill_pct                -0.13709    0.03309  -4.142 3.44e-05 ***
## Health_Expend_._GDP           0.37567    0.18155   2.069   0.0385 *
## SEX_RFemales                 -1.53159    0.91629  -1.672   0.0946 .
## ---
## Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 304.717  on 248  degrees of freedom
## Residual deviance:  85.039  on 240  degrees of freedom
## AIC: 103.04
##
## Number of Fisher Scoring iterations: 8
```

9.2 Proceso de selección en la dirección 'both'.

```
mod.stepwise.both<-stepAIC(mod.logist.completo, direction ="both", trace =TRUE)
```

```
## Start:  AIC=115.08
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     non_presc_med_use_pct + AGE_R + SEX_R
##
##                                Df Deviance    AIC
## - AGE_R                         4   82.925 110.92
## - BMI_Overweight                1   79.130 113.13
## - social_support_Poor           1   79.137 113.14
## - non_presc_med_use_pct         1   79.159 113.16
## - SEX_R                         1   79.337 113.34
## - Frequent_drink_pct            1   79.575 113.58
## - DailyFyV_5.portions.or.more   1   79.594 113.59
## - smoking_Daily.smoker          1   80.133 114.13
## - Physical_activity             1   80.277 114.28
## - DailyFyV_Less.5.portions.     1   80.529 114.53
## <none>                              79.079 115.08
## - smoking_Occasional.smoker     1   81.696 115.70
## - social_support_Intermediate   1   81.821 115.82
## - Health_Expend_._GDP           1   84.827 118.83
## - Chron_ill_pct                 1   87.985 121.98
##
```

```
## Step:  AIC=110.92
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     non_presc_med_use_pct + SEX_R
##
##                               Df Deviance    AIC
## - non_presc_med_use_pct        1   82.925 108.92
## - Frequent_drink_pct           1   82.961 108.96
## - Physical_activity            1   83.268 109.27
## - smoking_Daily.smoker         1   83.581 109.58
## - social_support_Poor          1   84.030 110.03
## - social_support_Intermediate  1   84.106 110.11
## - DailyFyV_Less.5.portions.    1   84.344 110.34
## - DailyFyV_5.portions.or.more  1   84.455 110.45
## <none>                             82.925 110.92
## - SEX_R                        1   85.179 111.18
## - BMI_Overweight               1   85.668 111.67
## - Health_Expend_._GDP          1   86.124 112.12
## + AGE_R                        4   79.079 115.08
## - smoking_Occasional.smoker    1   89.163 115.16
## - Chron_ill_pct                1  104.532 130.53
##
## Step:  AIC=108.93
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + Frequent_drink_pct + social_support_Poor +
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R
##
##                               Df Deviance    AIC
## - Frequent_drink_pct           1   82.961 106.96
## - Physical_activity            1   83.397 107.40
## - smoking_Daily.smoker         1   83.581 107.58
## - social_support_Intermediate  1   84.176 108.18
## - social_support_Poor          1   84.219 108.22
## - DailyFyV_5.portions.or.more  1   84.474 108.47
## - DailyFyV_Less.5.portions.    1   84.477 108.48
## <none>                             82.925 108.92
## - SEX_R                        1   85.421 109.42
## - BMI_Overweight               1   85.671 109.67
## - Health_Expend_._GDP          1   86.613 110.61
## + non_presc_med_use_pct        1   82.925 110.92
## + AGE_R                        4   79.159 113.16
## - smoking_Occasional.smoker    1   89.179 113.18
## - Chron_ill_pct                1  106.451 130.45
##
## Step:  AIC=106.96
## SPHpos ~ BMI_Overweight + Physical_activity + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     smoking_Daily.smoker + social_support_Poor + social_support_Intermediate +
##     Chron_ill_pct + Health_Expend_._GDP + SEX_R
##
##                               Df Deviance    AIC
## - Physical_activity            1   83.402 105.40
## - smoking_Daily.smoker         1   83.632 105.63
## - social_support_Intermediate  1   84.177 106.18
## - social_support_Poor          1   84.272 106.27
## - DailyFyV_Less.5.portions.    1   84.478 106.48
## - DailyFyV_5.portions.or.more  1   84.617 106.62
## <none>                             82.961 106.96
## - BMI_Overweight               1   85.684 107.68
## - SEX_R                        1   86.544 108.54
## - Health_Expend_._GDP          1   86.679 108.68
## + Frequent_drink_pct           1   82.925 108.92
## + non_presc_med_use_pct        1   82.961 108.96
## + AGE_R                        4   79.635 111.64
## - smoking_Occasional.smoker    1   89.949 111.95
## - Chron_ill_pct                1  107.178 129.18
##
## Step:  AIC=105.4
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + smoking_Daily.smoker + social_support_Poor +
```

```
##     social_support_Intermediate + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R
##
##                                Df Deviance    AIC
## - smoking_Daily.smoker          1   83.925 103.92
## - social_support_Intermediate  1   84.264 104.26
## - social_support_Poor          1   84.965 104.97
## - DailyFyV_Less.5.portions.    1   85.128 105.13
## <none>                             83.402 105.40
## - BMI_Overweight               1   85.758 105.76
## - DailyFyV_5.portions.or.more  1   85.843 105.84
## - Health_Expend_._GDP          1   86.679 106.68
## - SEX_R                        1   86.717 106.72
## + Physical_activity            1   82.961 106.96
## + non_presc_med_use_pct        1   83.268 107.27
## + Frequent_drink_pct           1   83.397 107.40
## - smoking_Occasional.smoker    1   90.121 110.12
## + AGE_R                        4   80.480 110.48
## - Chron_ill_pct                1  107.353 127.35
##
## Step:  AIC=103.92
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + social_support_Poor + social_support_Intermediate +
##     Chron_ill_pct + Health_Expend_._GDP + SEX_R
##
##                                Df Deviance    AIC
## - social_support_Intermediate  1   85.039 103.04
## - DailyFyV_Less.5.portions.    1   85.469 103.47
## - social_support_Poor          1   85.488 103.49
## <none>                             83.925 103.92
## - DailyFyV_5.portions.or.more  1   86.162 104.16
## - SEX_R                        1   86.904 104.90
## + smoking_Daily.smoker         1   83.402 105.40
## + Physical_activity            1   83.632 105.63
## - Health_Expend_._GDP          1   87.659 105.66
## - BMI_Overweight               1   87.696 105.70
## + non_presc_med_use_pct        1   83.837 105.84
## + Frequent_drink_pct           1   83.924 105.92
## - smoking_Occasional.smoker    1   90.447 108.45
## + AGE_R                        4   81.087 109.09
## - Chron_ill_pct                1  113.113 131.11
##
## Step:  AIC=103.04
## SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. + DailyFyV_5.portions.or.more +
##     smoking_Occasional.smoker + social_support_Poor + Chron_ill_pct +
##     Health_Expend_._GDP + SEX_R
##
##                                Df Deviance    AIC
## <none>                             85.039 103.04
## - DailyFyV_Less.5.portions.    1   87.293 103.29
## + social_support_Intermediate  1   83.925 103.92
## - SEX_R                        1   88.025 104.03
## - DailyFyV_5.portions.or.more  1   88.162 104.16
## - BMI_Overweight               1   88.209 104.21
## + smoking_Daily.smoker         1   84.264 104.26
## + non_presc_med_use_pct        1   84.955 104.95
## + Physical_activity            1   85.029 105.03
## + Frequent_drink_pct           1   85.039 105.04
## - Health_Expend_._GDP          1   89.563 105.56
## - social_support_Poor          1   91.580 107.58
## - smoking_Occasional.smoker    1   91.739 107.74
## + AGE_R                        4   82.747 108.75
## - Chron_ill_pct                1  113.434 129.43
```

```
summary(mod.stepwise.both)
```

```
##
## Call:
## glm(formula = SPHpos ~ BMI_Overweight + DailyFyV_Less.5.portions. +
##     DailyFyV_5.portions.or.more + smoking_Occasional.smoker +
##     social_support_Poor + Chron_ill_pct + Health_Expend_._GDP +
##     SEX_R, family = binomial, data = dfHD_train)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.26688  -0.04643   0.04430   0.15504   2.29398
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   2.31409    3.67165   0.630   0.5285
## BMI_Overweight               -0.06026    0.03665  -1.644   0.1001
## DailyFyV_Less.5.portions.     0.04738    0.03197   1.482   0.1384
## DailyFyV_5.portions.or.more  0.08674    0.05103   1.700   0.0892 .
## smoking_Occasional.smoker    0.50580    0.21277   2.377   0.0174 *
## social_support_Poor          -0.11658    0.04746  -2.457   0.0140 *
## Chron_ill_pct                -0.13709    0.03309  -4.142 3.44e-05 ***
## Health_Expend_._GDP           0.37567    0.18155   2.069   0.0385 *
## SEX_RFemales                 -1.53159    0.91629  -1.672   0.0946 .
## ---
## Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 304.717  on 248  degrees of freedom
## Residual deviance:  85.039  on 240  degrees of freedom
## AIC: 103.04
##
## Number of Fisher Scoring iterations: 8
```

9.3 Proceso de selección en la dirección 'forward.

```
mod.logist.cero<-glm(SPHpos~1, data =dfHD_train, family = binomial)

mod.stepwise.forward<-stepAIC(mod.logist.cero, direction ="forward", trace =TRUE)

## Start:  AIC=306.72
## SPHpos ~ 1

summary(mod.stepwise.forward)

##
## Call:
## glm(formula = SPHpos ~ 1, family = binomial, data = dfHD_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5492  -1.5492   0.8466   0.8466   0.8466
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8416     0.1381   6.092 1.11e-09 ***
## ---
## Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 304.72  on 248  degrees of freedom
## Residual deviance: 304.72  on 248  degrees of freedom
## AIC: 306.72
##
## Number of Fisher Scoring iterations: 4
```

10. ANEXO III. Cálculo de los componentes principales (PCA) de los datos por país y gráficas de los pesos de cada componente.

```
# calculo
pca.pais<-prcomp(dfHD.pais.num%>%dplyr::select(-GEO), scale =TRUE)

names(pca.pais)
## [1] "sdev"     "rotation" "center"   "scale"    "x"
# Resumen
summary(pca.pais)
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.8707 1.5381 1.2790 1.1405 1.02689 0.81302 0.68514
## Proportion of Variance 0.2916 0.1972 0.1363 0.1084 0.08788 0.05508 0.03912
## Cumulative Proportion  0.2916 0.4888 0.6251 0.7335 0.82138 0.87647 0.91558
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.55825 0.52017 0.47449 0.39532 0.22214
## Proportion of Variance 0.02597 0.02255 0.01876 0.01302 0.00411
## Cumulative Proportion  0.94155 0.96410 0.98286 0.99589 1.00000

# recuperamos los pesos

df.loadings<-data.frame(pca.pais$rotation[,1:12])

x.names<-row.names(df.loadings)


# grafica de los pesos de cada factor principal (PCA)

for(columnain1:12){

plot(ggplot(mapping =aes(x =x.names, y =df.loadings[,columna])) +
geom_col(fill=ifelse(df.loadings[,columna]>=0,"blue","red")) +theme(axis.text.x
=element_text(angle =90)) +
labs(x ="Variables originales", y ="Pesos")  +
ggtitle(paste("FACTOR",columna)) +
theme(plot.title =element_text(hjust =0.5))
    )
```

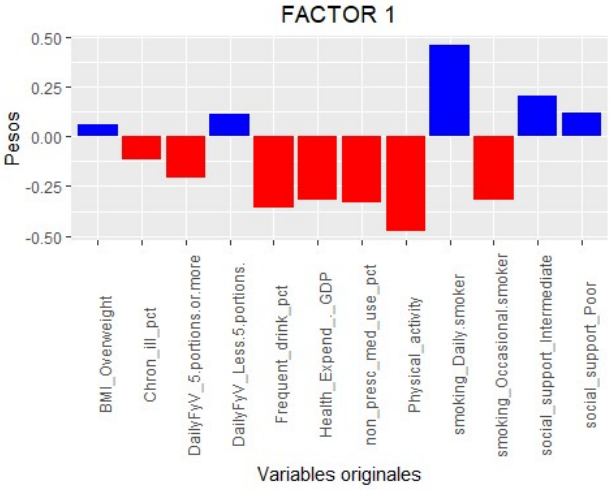

FIGURA 37: PCA PESOS FACTOR 1

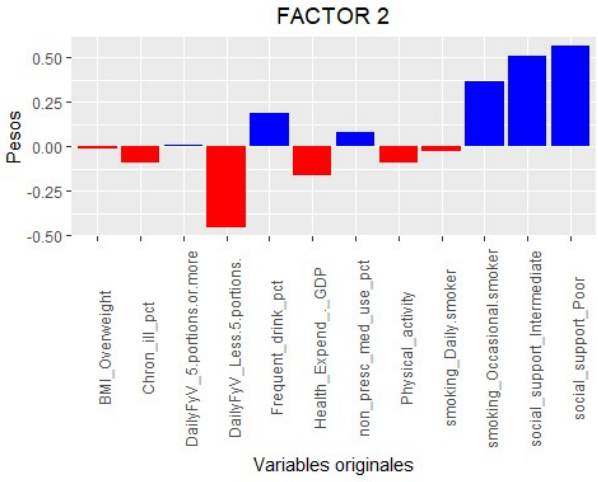

FIGURA 39: PCA PESOS FACTOR 2

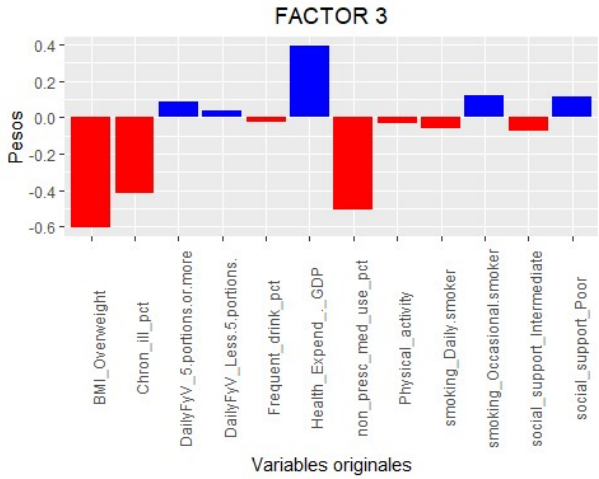

FIGURA 38: PCA PESOS FACTOR 3

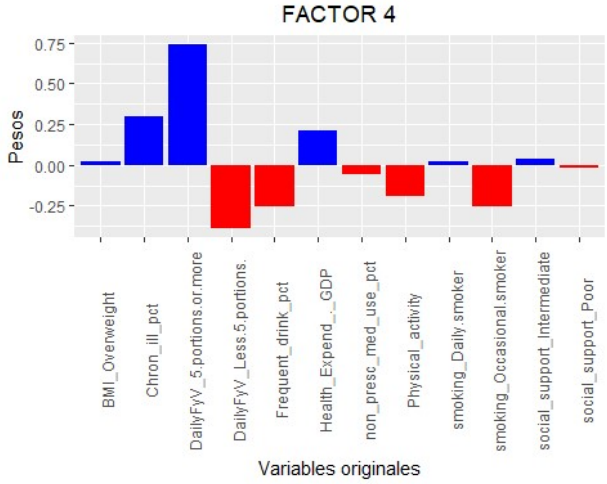

FIGURA 41: PCA PESOS FACTOR 4

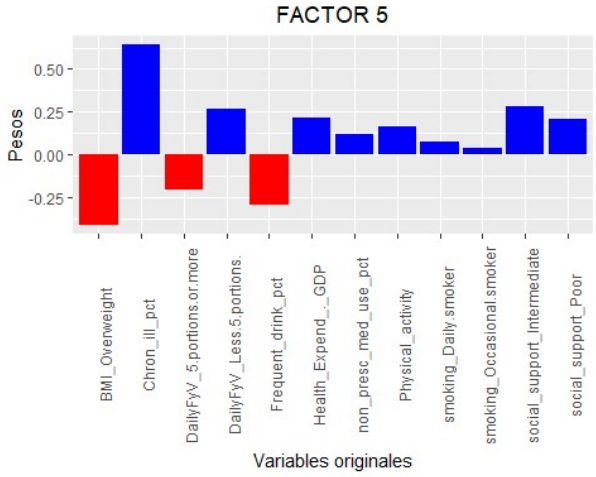

FIGURA 40: PCA PESOS FACTOR 5

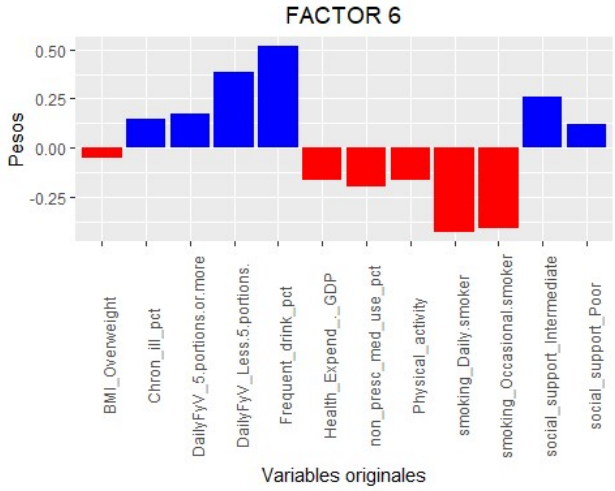

FIGURA 42: PCA PESOS FACTOR 6

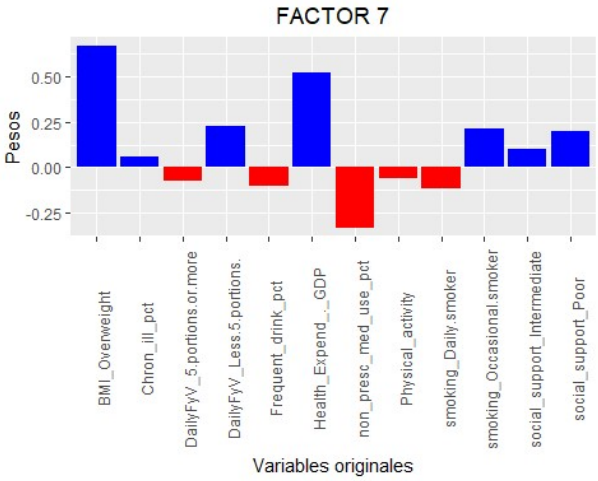

FIGURA 43: PCA PESOS FACTOR 7

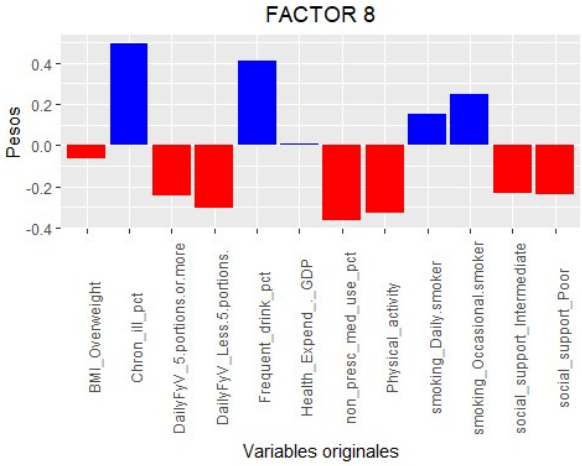

FIGURA 44: PCA PESOS FACTOR 8

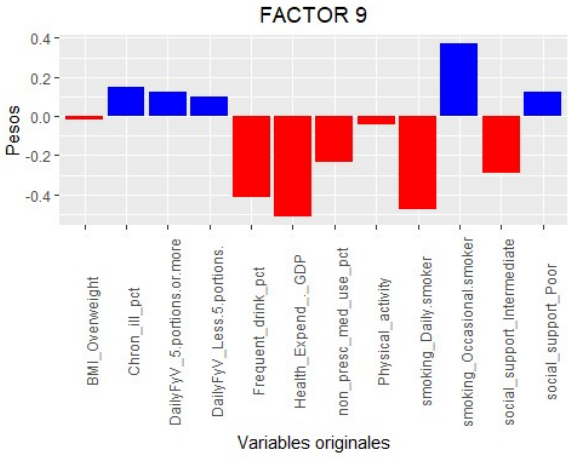

FIGURA 45: PCA PESOS FACTOR 9

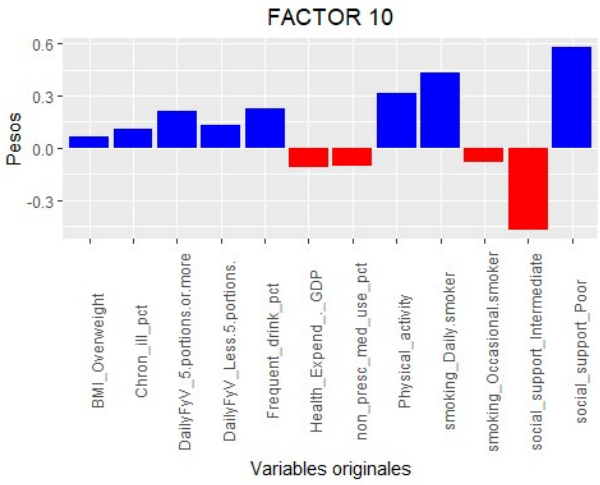

FIGURA 46: PCA PESOS FACTOR 10

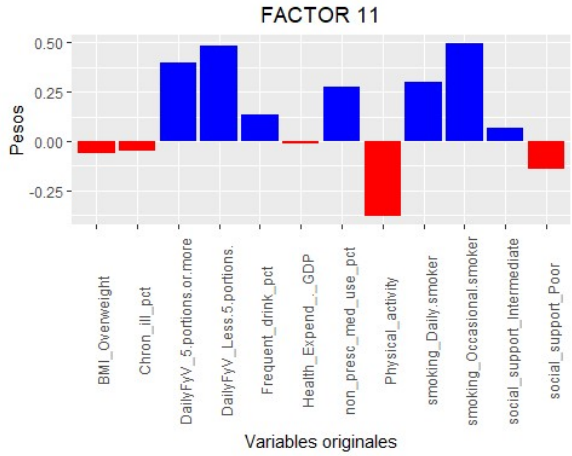

FIGURA 47: PCA PESOS FACTOR 11

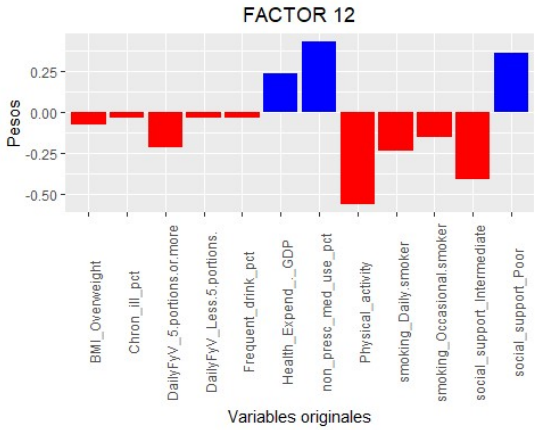

FIGURA 48: PCA PESOS FACTOR 12

11. ANEXO IV: GRÁFICAS DE AGRUPACIÓN (CLUSTERIZACIÓN) POR PAÍSES, CON DISTINTOS "LINKAGES" PROBADOS.

11.1 "LINKAGE COMPLETE"

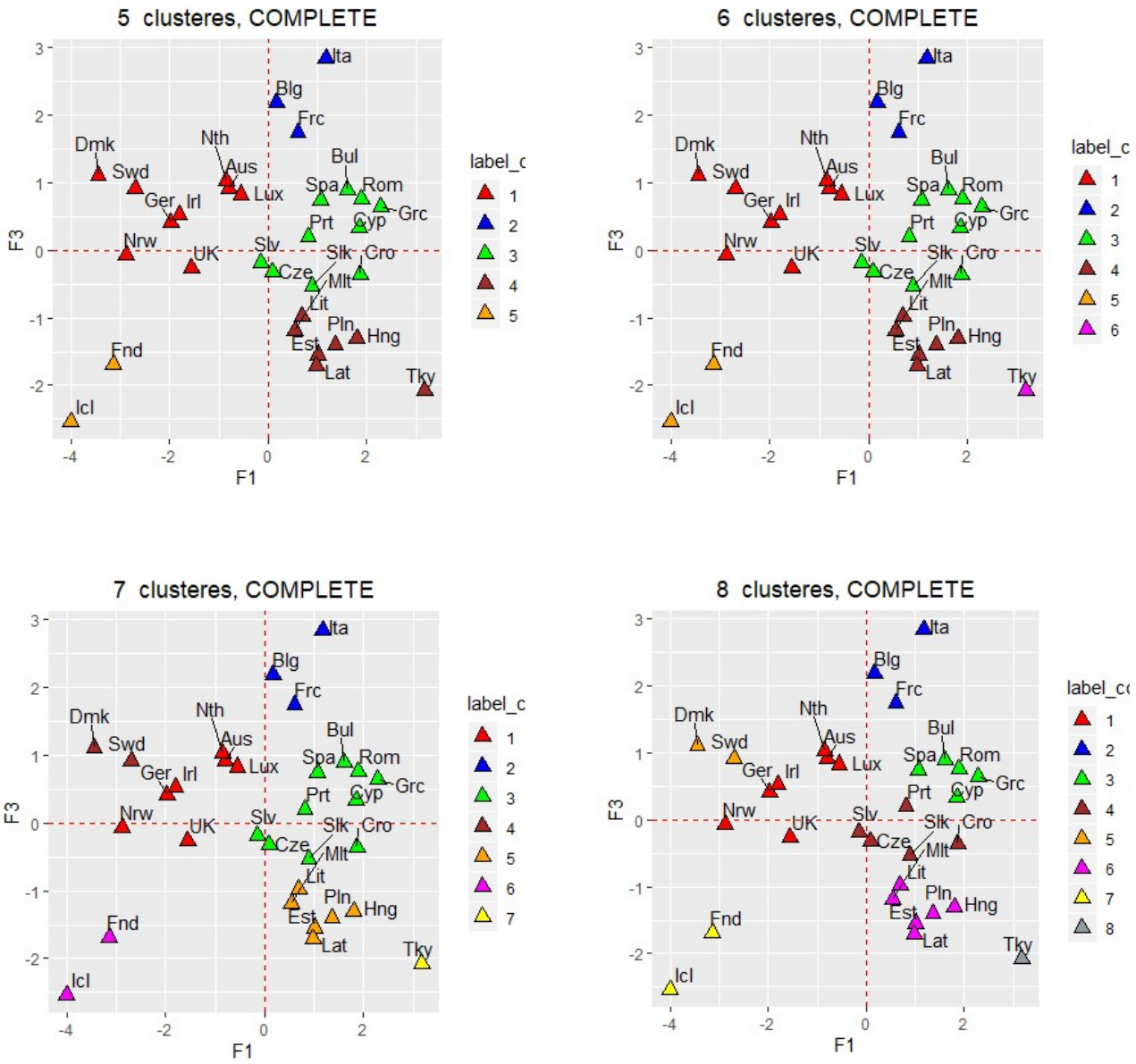

FIGURA 49: CLUSTERING LINKAGE "COMPLETE"

11.2 “LINKAGE AVERAGE”

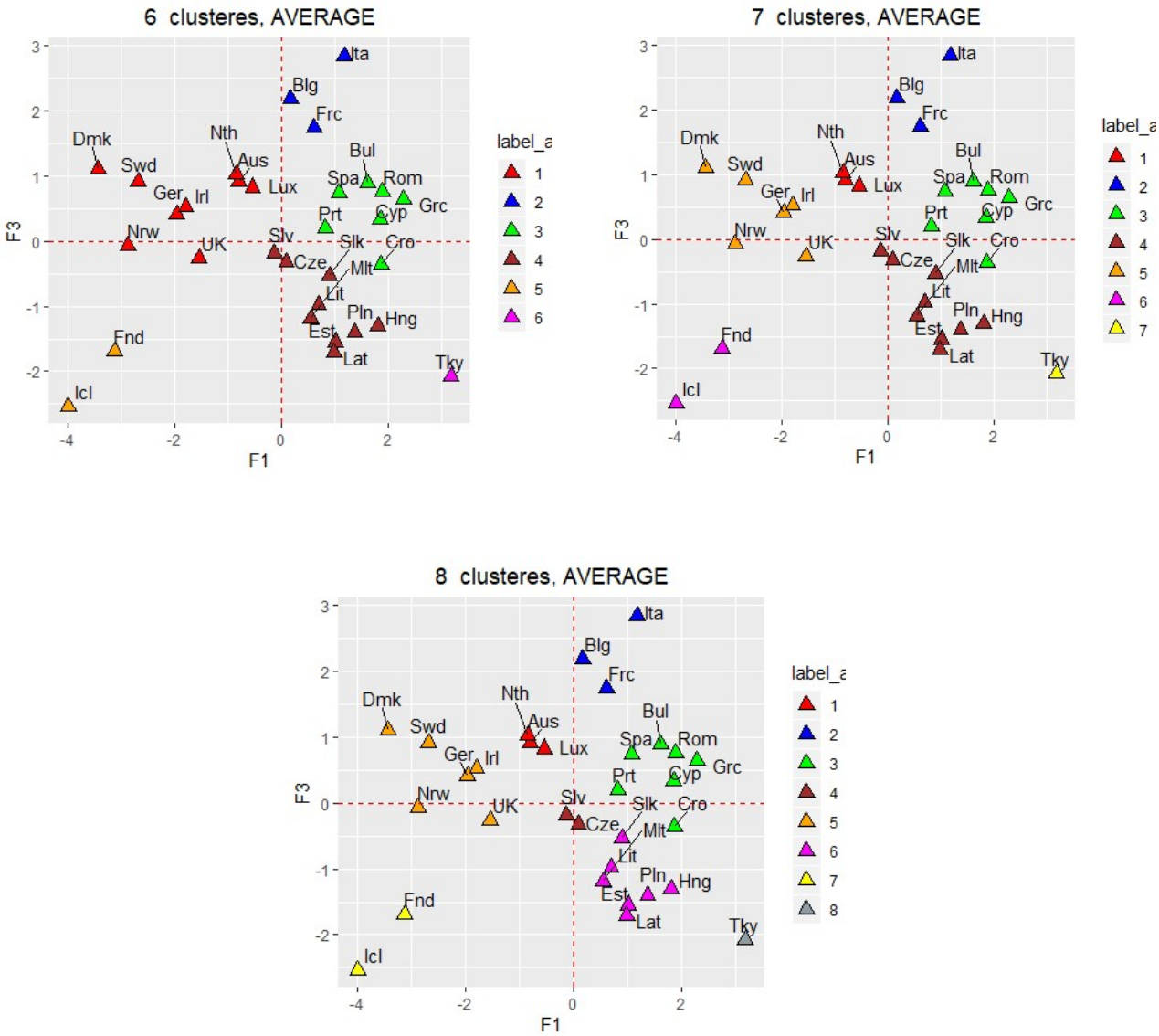

FIGURA 50: CLUSTERING LINKAGE "AVERAGE"

11.3 “LINKAGE WARD”

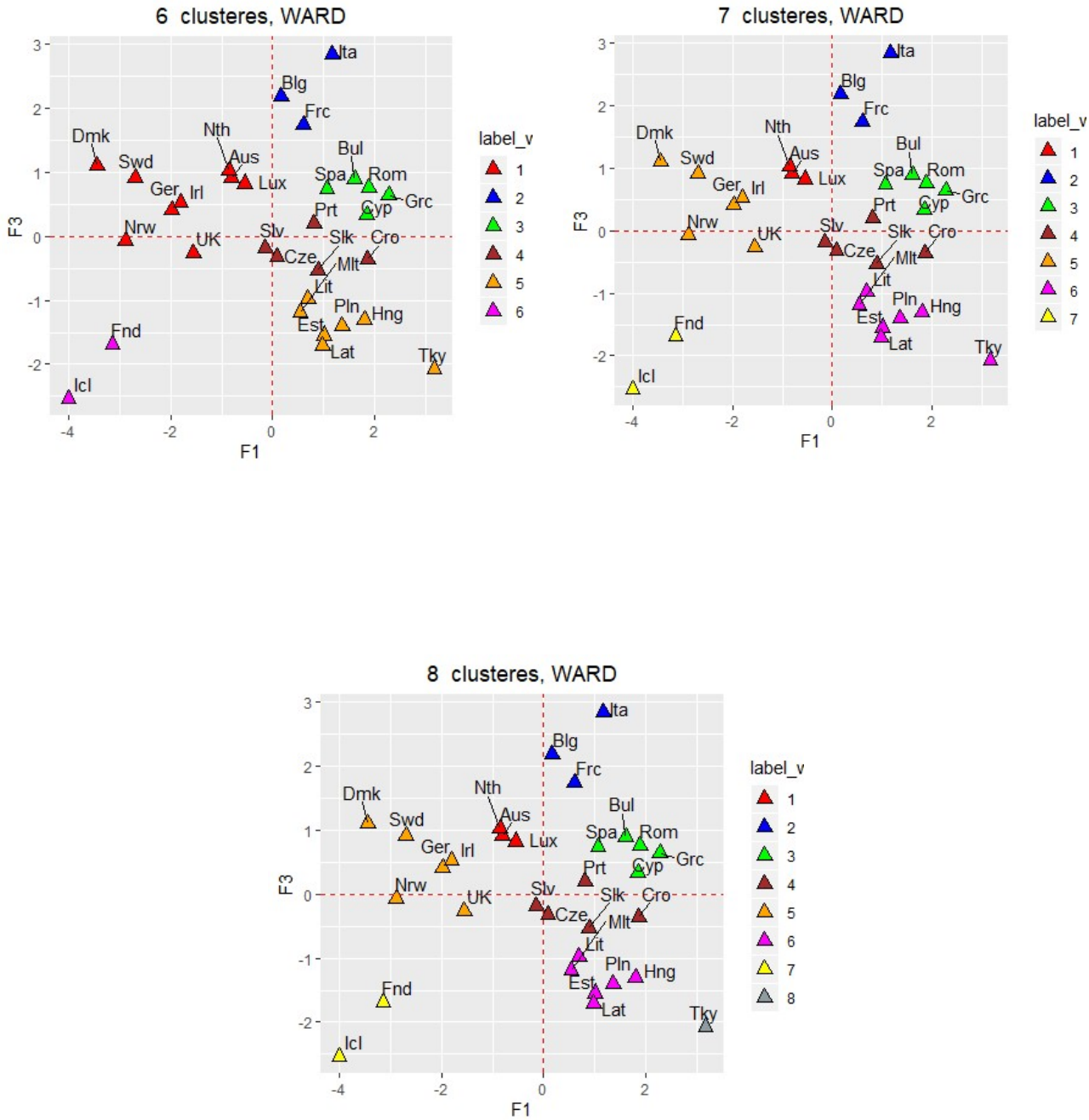

FIGURA 51: CLUSTERING LINKAGE "WARD"

12. ANEXO V: CÓDIGO.

El código de este TFM junto con los archivos necesarios, incluido el de la visualización, están disponibles en el siguiente enlace:

https://github.com/edurfpy/TFM_Ciencia_Datos_Eduardo_Rivero_Falcon.git

13. BIBLIOGRAFÍA.

- Agencia de la Unión Europea para los Derechos Fundamentales. (2007). *Página web de la Agencia de la Unión Europea para los Derechos Fundamentales. Artículo 35: Protección de la salud.*Visitada en Octubre 2020: https://fra.europa.eu/es/eu-charter/article/35-proteccion-de-la-salud

- Arias de la Torre, V. M. (2018). Prevalence of major depressive disorder and association with personal and socio-economic factors. Results for Spain of the European Health Interview Survey 2014-2015. *Journal of Affective Disorders, Vol.239* , 203-207.

- Camarero, L., Almazán, A., & Mañas, B. (n.d.). Regresión logística: fundamentos y aplicación a la investigación sociológica. *Departamento de Sociología I, UNED* , 47-48.

- Congreso de los Diputados. (2003). *Página web oficial del Congreso de los Diputados. Constitución Española*. Visitada en Octubre 2020: https://app.congreso.es/consti/constitucion/indice/titulos/articulos.jsp?ini=43&tipo=2

- Cott, G. &. (1999). Determinants of self rated health for Canadians with chronic disease an disability. *Journal of Epidemiology & Community Health* , 731-736.

- Daviña, P. G. (2010). *Tésis Doctoral: Los determinantes de la salud percibida en España.* Escuela Universitaria de Enfermería, Fisioterapia y Podología. Universidad Complutense de Madrid.

- Declaración Universal de los Derechos Humanos. (2020). *Página web en español de la Declaración Universal de los Derechos Humanos*. Visitada en Octubre 2020: https://dudh.es/25/?doing_wp_cron=1608990060.4794330596923828125000

- Eurostat. (2020). *Página de datos abiertos sobre salud del Eurostat*. Visitadaen Octubre de 2020, de https://ec.europa.eu/eurostat/web/health/data

- Fernández Martínez, P. F. (2012). Self-perceived health status in older adults: regional and sociodemographic inequalities in Spain. *Saúde Pública [online], vol. 46* , 310-319.

- Jindrova, P. Z. (2016). *Comparison of Health Care Results in Public Health Systems of European Countries.* Czech Republic: European Financial systems. Proceedings of the 13th International Scientific Conference.

- Murata, K. &. (2006). Determinants of self-rated health:Could health status explain the association between self-rated health and mortality? *Archives of Gerontology and Geriatrics, Volume 43* , 369-380.

- Pacakova, K. (2018). Comparing inequalities in health outcomes in european countries. *Journal of International Studies* , 215-227.

- Rathmann, O. H. (2015). Macro-level determinants of young people's subjective health and health inequelities: A multilevel analysis in 27 welfare states. *Maturitas, vol. 80* , 414-420.

- Schmidt, P. B. (2002). *Cross-cultural analysis of relationships of health indicators across Europe:first results based on the EUROHIS project.* Report on a WHO study, WHO Regional Office for Europe.