FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2020/2021

Trabajo de Fin de Máster

TÍTULO: Predicción de la felicidad en el mundo

***Alumno*: Belén Arellano Clemente**

***Tutor*: Juana María Alonso Revenga**

Junio de 2021

UNIVERSIDAD COMPLUTENSE
MADRID

AGRADECIMIENTOS

A mis compañeros y docentes por formar parte de esta etapa de mi vida.

A mi tutora Juana por guiarme y ayudarme siempre que lo he necesitado.

A mis padres por su apoyo incondicional.

ÍNDICE

1. RESUMEN

Según la RAE (Real Academia Española), la felicidad es un estado de grata satisfacción espiritual. Por tanto, es un factor cualitativo que debido a su alto grado de subjetividad supone un gran obstáculo para su medición. Existen diversos estudios y son varias las encuestas que se llevan a cabo con el fin de poder cuantificar esta variable.

En este trabajo se abarcaron tres objetivos: buscar la relación entre la felicidad y los distintos factores sociales, económicos y políticos; sacar un modelo de predicción y determinar el número de países parecidos entre sí para ayudar en la predicción.

Se manipuló la base de datos del World Happiness Report 2021 añadiendo más variables para intentar mejorar el modelo predictivo. Se tuvieron 3 unidades de análisis derivadas de la felicidad permitiendo, de esta manera, la utilización de los métodos estadísticos Análisis Factorial de Datos Mixtos (AFDM), regresión lineal, regresión logística binaria y K vecinos más próximos (KNN).

Como resultados fueron encontrados relaciones entre la felicidad y las variables input siendo el indicador regional una variable clave para la parte de aprendizaje supervisado. El continente americano y europeo encabezaba el ranking de felicidad más alta mientras que en África están situados los países menos felices.

2. ABSTRACT

According to the RAE (Real Academia Española), happiness is a state of pleasant spiritual satisfaction. Therefore, it is a qualitative factor that, due to its high degree of subjectivity, represents a great obstacle to its measurement. There are several studies and surveys that are carried out in order to quantify this variable.

Three objectives were covered in this paper: to seek the relationship between happiness and the different social, economic and political factors; pull out a prediction model and determine the number of countries similar to each other to aid in the prediction

The World Happiness Report 2021 database was manipulated by adding more variables to try to improve the predictive model. There were 3 units of analysis derived from happiness allowing, in this way, the use of the statistical methods Factorial Analysis of Mixed Data (FAMD), linear regression, binary logistic regression and K-nearest neighbours (KNN)

As results, relationships were found between happiness and the input variables, with the regional indicator being a key variable for the supervised learning part. American and European continents led the ranking of highest happiness while in Africa are located the least happy countries.

3. INTRODUCCIÓN

La felicidad es un concepto emocional complejo y difícil de definir debido a que posee un componente muy subjetivo. Por ejemplo, una situación feliz para una persona puede resultar triste para otra por lo que hay una variante personal que hace más difícil su medición. Por consiguiente, han sido muchos los estudios relacionados con esta temática y, sobre todo, en intentar entender cuáles son los factores que influyen en él.

A nivel individual, en 2004, se dio por finalizado uno de los estudios de más larga duración de todos los tiempos elaborado por la Universidad de Harvard y dirigido por el doctor psiquiatra George Viallant y el profesor Robert Waldinger. Se pretendió encontrar las claves de la felicidad analizando varias variables a los mismos individuos durante 75 años. Una de las conclusiones del estudio es que las relaciones cercanas, más que el dinero o la fama, son las que mantienen felices a las personas durante toda su vida. Además, el aislamiento o el alcohol provoca problemas emocionales que influyen negativamente.

A nivel gubernamental o político, el PIB ha sido siempre el indicador por excelencia que se ha utilizado para medir el nivel de actividad, el desarrollo global de la sociedad, el progreso y el bienestar. Nunca la felicidad. No obstante, existen nuevas demandas que han ido apareciendo en la actualidad y el PIB ya no puede medirlas y tampoco orientar las políticas para su logro. En consecuencia, se han comenzado a plantearse preguntas del tipo “¿Los gobiernos deben orientar sus políticas hacia el crecimiento o hacia la felicidad?”.(Montuschi, s. f.). Desde 2012, y con una periodicidad anual, un comité de expertos de las Naciones Unidas realiza el World Happiness Report (WHR) para medir la felicidad en los diferentes países del mundo. Para ello, se basa en factores como el PIB, el apoyo social, la libertad individual de los ciudadanos, la generosidad, la corrupción y la esperanza de vida saludable.

En este Trabajo Fin de Máster, además de contar con las variables del WHR, han sido incluidas otras como la tasa de depresión (prevalencia) en el país o las horas de sol al año, entre otras.

De esta manera, lo que se pretende es analizar qué variables son las que más influyen en la escala de la felicidad en el mundo para, más adelante, poder realizar una predicción. Además, poder mejorar este pronóstico se ha podido determinar

La importancia de este trabajo estriba en que, además de poder medir la felicidad, algunos de los factores añadidos resultaron significativos para la construcción del modelo predictivo.

4. OBJETIVOS

1. Analizar la relación existente entre la escala de la felicidad y los distintos factores sociales, económicos y políticos, en los diferentes países del mundo.
2. Predecir la escala de la felicidad en el mundo en función de los distintos factores sociales, económicos y políticos.
3. Determinar el número de países parecidos entre sí para predecir la felicidad.

5. METODOLOGÍA

La **base de datos** que se utilizó proviene de la del WHR de 2021 pero fue obtenida de la página web Kaggle. Éste contiene 149 países y 20 variables de las que sólo se utilizaron 9 puesto que las demás eran producto del análisis del WHR como errores, valores máximos o mínimos, entre otros.

En el WHR la felicidad se obtuvo como la respuesta media nacional de los encuestados a la pregunta:

Imagine una escalera, con escalones numerados del 0 (parte inferior) al 10 (parte superior). La parte superior de la escalera representa la mejor vida posible para usted y la parte inferior de la escalera representa la peor vida posible para usted. ¿En qué escalón de la escalera diría que personalmente siente que se encuentra en este momento?

En este trabajo, la felicidad se ha dividido en las siguientes **unidades de análisis**:

Nombre	Descripción	Rango	Tipo
Happiness	Escala o puntuación de la felicidad. Denominada *Ladder* en el WHR.	0-10	Cuantitativa
Happiness_cat	Valoración de la felicidad: *muy infeliz* (1-3), *infeliz* (4), *neutro* (5), *feliz* (6-7) y *muy feliz* (8-10)	5 categorías	Nominal
Happines_bin	Valoración de la felicidad: *país feliz* (1) o *infeliz* (0)	0 ó 1	Binaria

Tabla 5-1. Unidades de análisis

Las **variables** que hicieron posible el cumplimiento de los objetivos fueron:

	Nombre	Descripción	Rango/ categorías	Tipo
1	Country name	Nombre del país. Único	único	ID
2	Corruption	Percepciones de corrupción.	0-1	Cuantitativa
3	Freedom	Derecho individual a la libertad que tiene un ciudadano para elegir por si mismo sus propias acciones o patrón de vida, siempre y cuando no perjudique a la libertad de elección de los demás.	0-1	Cuantitativa
4	Generosity	Generosidad del país.	(-1) - 1	Cuantitativa
5	Ln(PIB_pc)	logaritmo natural del PIB per cápita del país de 2019 a 2020	0-100	Cuantitativa
6	Regional indicator	Región del continente al que pertenece (10): *Central and Eastern Europe, Commonwealth of Independent States,*	10	Cualitativa

		East Asia, Latin America and Caribbean, Middle East and North Africa, North America and ANZ, South Asia, Southeast Asia, Sub-Saharan Africa y *Western Europe*.		
7	Social_support	Apoyo social.	0-1	Cuantitativa
9	Healthy_life_ expectancy	Esperanza de vida saludable, es decir, la media esperada de años que una persona pueda vivir con salud plena.	40-80	Cuantitativa
9	Clima	Tipo de clima del país: *cálido* (ecuatorial, tropical, desértico), *templado* (mediterráneo, subtropical, oceánico, continental), *frío* y *variado*	15	Cualitativa
10	Desempleo	Tasa de desempleo en el país (en porcentaje).	0-100	Cuantitativa
11	Esperanza_vida	Esperanza de vida 2021.	50-90	Cuantitativa
12	Gobierno	Tipo de gobierno actual en el país: estado socialista unipartidista, monarquía (absoluta, constitucional, constitucional electiva o parlamentaria), república (constitucional electiva, federal presidencialista, islámica parlamentaria, islámica presidencialista, parlamentaria, presidencialista o semipresidencialista).	12	Cualitativa
13	Horas_sol	Horas de sol al año en media (número entero)	600-4100	Cuantitativa
14	IDH	Índice de desarrollo humano (2019) mide los niveles de desarrollo de un país (compuesto por la esperanza de vida, la educación e indicadores de ingreso per cápita).	0-1	Cuantitativa
15	Peligroso	Las 20 ciudades más peligrosas según el Global Peace Index 2019.	0,1	Cualitativa
16	Prevalencia	Prevalencia de depresión en cada país.	0-100	Cuantitativa

Tabla 5-2. Descripción de las variables

Las variables de la 1 a la 7 provienen de la base de datos del WHR. De entre éstas, la 2, 3, 4 y 7 y son la media nacional de las preguntas:

- *Corruption*: *"¿La corrupción está generalizada en todo el gobierno o no" y "¿La corrupción está generalizada en las empresas o no?"*
- *Freedom*: *"¿Está satisfecho o insatisfecho con su libertad de elegir lo que hace con su vida?*
- *Generosity*: en este caso se trata el residuo de la regresión del promedio nacional sobre el PIB per cápita. *"¿Ha donado dinero a una organización benéfica en el último mes?"*
- *Soccial_support:* pregunta de respuesta binaria *Sí* o *No,* pero también se hace la media como en los demás casos. *¿Si tuviera problemas familiares, tiene familia o amigos en los que pueda contar cuando lo necesite?*

Las variables de la 8 a la 16 fueron incluidas en la base de datos y provienen de distintas fuentes o estudios relacionados con la felicidad. Se ha pensado que pueden resultar interesantes para el modelo predictivo.

Fueron utilizados distintos **métodos estadísticos** según los diferentes apartados:

- Análisis descriptivo: tablas de frecuencias, diagrama de barras y diagrama de tartas para las variables cualitativas. Media, mediana, desviación típica, coeficiente de variación, curtosis, asimetría, rango y rango intercuartílico para las variables cuantitativas.
- Depuración de los datos: corrección de errores, tratamiento de datos atípicos y datos faltantes. Además, para la medición de la relación entre variables, gráfico de correlación de Pearson, gráfico de valor, histogramas, gráficos de dispersión y mapa de calor.
- Aprendizaje no supervisado: Análisis Factorial de Datos Mixtos (AFDM; siglas en inglés FAMD).
- Aprendizaje supervisado: métodos de selección de variables stepwise, forward y backward, regresión lineal, validación cruzada repetida (VCR), regresión logística binaria y K-Nearest Neighbors (KNN).

Han sido varios los **programas estadísticos** con los que se ha trabajado a lo largo de este TFM: SAS MINER en la parte de depuración de los datos y algoritmo KNN, R para el AFMD y SAS (base) en los modelos de regresión lineal y logística.

6. ANÁLISIS DESCRIPTIVO

Independientemente de los diferentes tipos de variables que se tenga en tu conjunto de datos, el paso siguiente es el del análisis descriptivo. Se trata de explorar los datos con el fin de identificar sus principales características tanto de forma numérica como gráfica.

6.1. VARIABLES CUALITATIVAS

Como variables de tipo cualitativo se tienen: *regional_indicator*, *clima, gobierno, peligroso* y *happiness_cat.*

Happiness_cat

Happiness_cat				
Happiness_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	4	2.68	4	2.68
4	20	13.42	24	16.11
5	50	33.56	74	49.66
6	50	33.56	124	83.22
7	21	14.09	145	97.32
8	4	2.68	149	100.00

Tabla de frecuencias 1. Happiness_cat

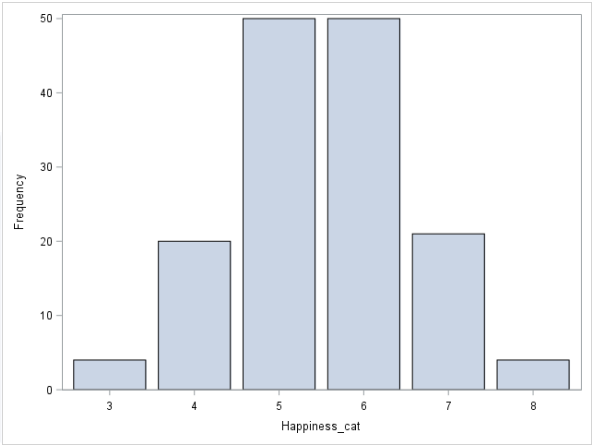


Gráfico histograma 1 Happiness_cat

NOTA: Al tratarse de una variable ordinal, se dibuja el gráfico de barras en lugar del de tartas.

- Si se consideraran estas categorías como números, se observa que es una variable con una distribución normal puesto que en los extremos esta frecuencia es menor frente al centro que es donde se acumulan gran cantidad de los países.
- La mayoría de éstos poseen una felicidad Neutra (5) o Feliz (6).
- 24 de los países son muy infelices o infelices, es decir, más del 16% del conjunto de datos. Por el contrario, son 4 países los que se consideran muy felices (2.68%).

Regional_indicator

Regional_indicator				
Regional_indicator	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Central and Eastern Europe	17	11.41	17	11.41
Commonwealth of Independent States	12	8.05	29	19.46
East Asia	6	4.03	35	23.49
Latin America and Caribbean	20	13.42	55	36.91
Middle East and North Africa	17	11.41	72	48.32
North America and ANZ	4	2.68	76	51.01
South Asia	7	4.70	83	55.70
Southeast Asia	9	6.04	92	61.74
Sub-Saharan Africa	36	24.16	128	85.91
Western Europe	21	14.09	149	100.00

Tabla de frecuencias 2. Regional_indicator

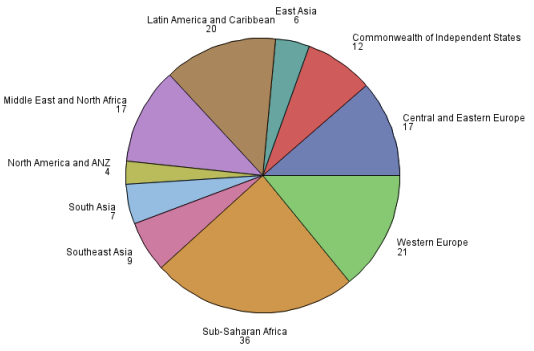


Gráfico de tartas 1. Regional_indicator

- Las **regiones más frecuentes** son *Sub-saharian Africa* (24.16%) y *Western Europe* (14.09%). Si lo viésemos por continentes se tiene que la mayoría de los

países son de África o Europa pues conjuntamente suponen más del 60% de a base de datos.

- Las **regiones menos frecuentes** son *North America and ANZ* (Australia and New Zeland) y *East Asia* que son conjuntamente 10 países.

Clima

Clima				
Clima	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Calido desertico	16	10.74	16	10.74
Calido ecuatorial	10	6.71	26	17.45
Calido ecuatorial y Frio	2	1.34	28	18.79
Calido ecuatorial y tropical	2	1.34	30	20.13
Calido tropical	41	27.52	71	47.65
Calido tropical y desertico	4	2.68	75	50.34
Templado continental	14	9.40	89	59.73
Templado continental y Frio	8	5.37	97	65.10
Templado continental y mediterraneo	1	0.67	98	65.77
Templado continental y subtropical	2	1.34	100	67.11
Templado mediterraneo	18	12.08	118	79.19
Templado oceanico	14	9.40	132	88.59
Templado oceanico y Frio	2	1.34	134	89.93
Templado subtropical	3	2.01	137	91.95
Variado	12	8.05	149	100.00

Tabla de frecuencias 3. Clima

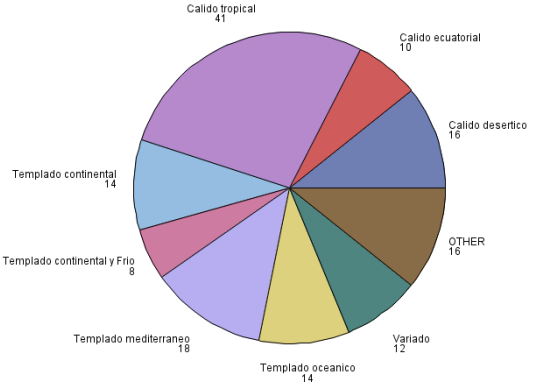


Gráfico de tartas 2. Clima

- Los **climas mayoritarios** en la base de datos son cálido tropical, templado mediterráneo y cálido desértico. Éstos comprenden la mitad de los países.
- Los **climas minoritarios** son los que están en la categoría de *Otros* (con una frecuencia de 4 o menos): templado continental y mediterráneo, cálido ecuatorial y Frío, cálido ecuatorial y tropical, templado continental y subtropical, templado oceánico y Frio, templado subtropical y cálido tropical y desértico.
- **Un 27.52% de los países son de clima tropical** y corresponden con 41 observaciones siendo, sin duda, el clima más habitual.
- **12 de 149 países poseen un clima variado** entre tropical y cálido, es decir, un 8.05% del conjunto de datos.

Gobierno

Gobierno				
Gobierno	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Estado socialista unipartidista	3	2.01	3	2.01
Monarquia absoluta	2	1.34	5	3.36
Monarquia constitucional	11	7.38	16	10.74
Monarquia constitucional electiva	2	1.34	18	12.08
Monarquia parlamentaria	9	6.04	27	18.12
Republica constitucional electiva	1	0.67	28	18.79
Republica federal presidencialista	1	0.67	29	19.46
Republica islamica parlamentaria	1	0.67	30	20.13
Republica islamica presidencialista	2	1.34	32	21.48
Republica parlamentaria	42	28.19	74	49.66
Republica presidencialista	48	32.21	122	81.88
Republica semipresidencialista	27	18.12	149	100.00

Tabla de frecuencias 4. Gobierno

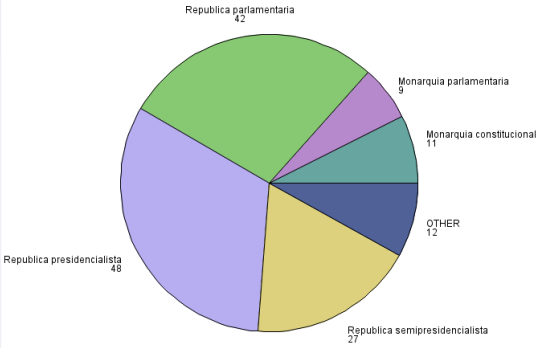


Gráfico de tartas 3. Gobierno

- En este caso, **el gobierno en la mayoría de los países es la república** y más en concreto la presidencialista, la semipresidencialista y la parlamentaria. Conjuntamente, abarcan casi el 80% de los datos.
- Como **gobiernos minoritarios** se encuentran: las repúblicas parlamentaria o presidencialista de tipo islámica y la república federal presidencialista. Con una frecuencia unitaria cada una.
- Un 6.04% de los países poseen una monarquía parlamentaria frente al 7.38% que es monarquía constitucional. Mientras, el 1.34% son monarquías absolutas.

Peligroso

Peligroso				
Peligroso	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	129	86.58	129	86.58
1	20	13.42	149	100.00

Tabla de frecuencias 5. Peligroso

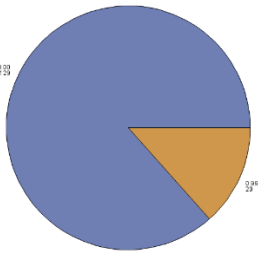


Gráfico de tartas 4. Peligroso

Se introdujeron 20 países peligrosos y el resto fue considerado como no peligroso, es decir, un 13.42% de nuestras observaciones están en la lista de los más peligrosos.

6.2. VARIABLES CUANTITATIVAS

A continuación, se presenta una tabla (6-1.) con los estadísticos descriptivos más importantes para cada variable de esta tipología:

La única que posee **valores perdidos** en la base de datos es *Prevalencia* por lo que va a ser necesaria su imputación en pasos futuros.

Si la **media** y la **mediana** no están muy distanciadas entre sí, entonces significa que la variable es casi simétrica. En general, esto se cumple en casi todas las variables a excepción de *Desempleo*, *Generosity* y *Corruption* donde se puede sospechar asimetría. Además, para ver la normalidad de las variables nos podemos fijar en la **asimetría** y la **curtosis**. Una variable se considera de distribución Normal cuando ambas se encuentran en el intervalo [-2,2]. Incumplen esta condición: *Desempleo* y *Corruption* cuya curtosis > 2, es decir, son leptocúrticas (la distribución es más empinada que la curva normal).

Si la **desviación típica** es pequeña indica buena señal. Si la relativizamos se obtiene el **coeficiente de variación** o desviación típica relativizada que es el error que se comete al sustituir la variable con la media. En términos medios la variación de los datos en relación con la media es del 19.22%. El menor error se sitúa en *Generosity* (10.067%) mientras que la variable con mayor dispersión es *Desempleo* (76.576%).

Por último, cuando la diferencia entre el **rango** y el **rango intercuartílico** es grande, es motivo suficiente para sospechar que puede haber atípicos. En general, se diría que sí que los hay debido a que existen discrepancias entre estas métricas.

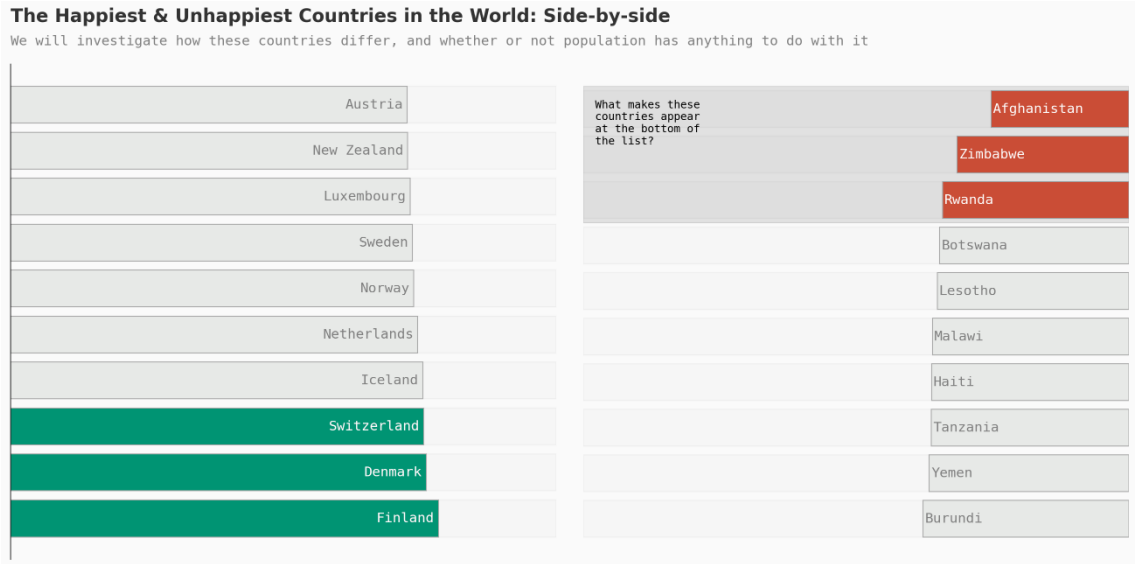
Variable	nmiss	Media	Mediana	Desv. típica	Coef. de variación	Kurtosis	Asimetría	Mínimo	Máximo	Rango	Rango Intercuartílico
Horas_sol	0	2447.034	2445	573.588	23.44	-0.949	-0.045	1268	3605	2337	856
Prevalencia	4	4.513	4.5	0.633	14.034	-0.292	0.187	3.2	6.3	3.1	0.9
IDH	0	0.737	0.756	0.151	20.539	-0.857	-0.419	0.394	0.957	0.563	0.253
Desempleo	0	8.347	6.4	6.392	76.576	2.976	1.649	0.5	33.3	32.8	6.8
Esperanza_vida	0	73.59	75.234	7.442	10.113	-0.435	-0.597	54.622	85.026	30.404	10.636
Happiness	0	5.533	5.534	1.074	19.41	-0.369	-0.104	2.523	7.842	5.319	1.403
Ln_PIB_per_capita_	0	9.432	9.569	1.159	12.283	-0.815	-0.352	6.635	11.647	5.012	1.88
Social_support	0	0.815	0.832	0.115	14.101	0.395	-0.938	0.463	0.983	0.52	0.155
Healthy_life_expectancy	0	64.993	66.603	6.762	10.404	-0.564	-0.522	48.478	76.953	28.475	9.798
Freedom	0	0.792	0.804	0.113	14.317	0.408	-0.755	0.382	0.97	0.588	0.159
Generosity	0	-0.015	-0.036	0.151	10.067	1.636	1.01	-0.288	0.542	0.83	0.205
Corruption	0	0.727	0.781	0.179	24.638	2.25	-1.577	0.082	0.939	0.857	0.178

Tabla 6-1. Estadísticos descriptivos variables cuantitativas

$$\boldsymbol{Media} = \bar{x} = \frac{\sum_{i=1}^{159} x_i}{159} \qquad \boldsymbol{Mediana} = Me = L_{i-1} + \frac{159/2 - F_{i-1}}{f_i} \cdot a \qquad \boldsymbol{Desviación\ típica} = S = \sqrt{\sum_{i=1}^{159} \frac{(x_i - \bar{x})^2}{159}} \qquad \boldsymbol{Coeficiente\ de\ Variación} = r = \frac{S_i}{\bar{x}_i} \cdot 100$$

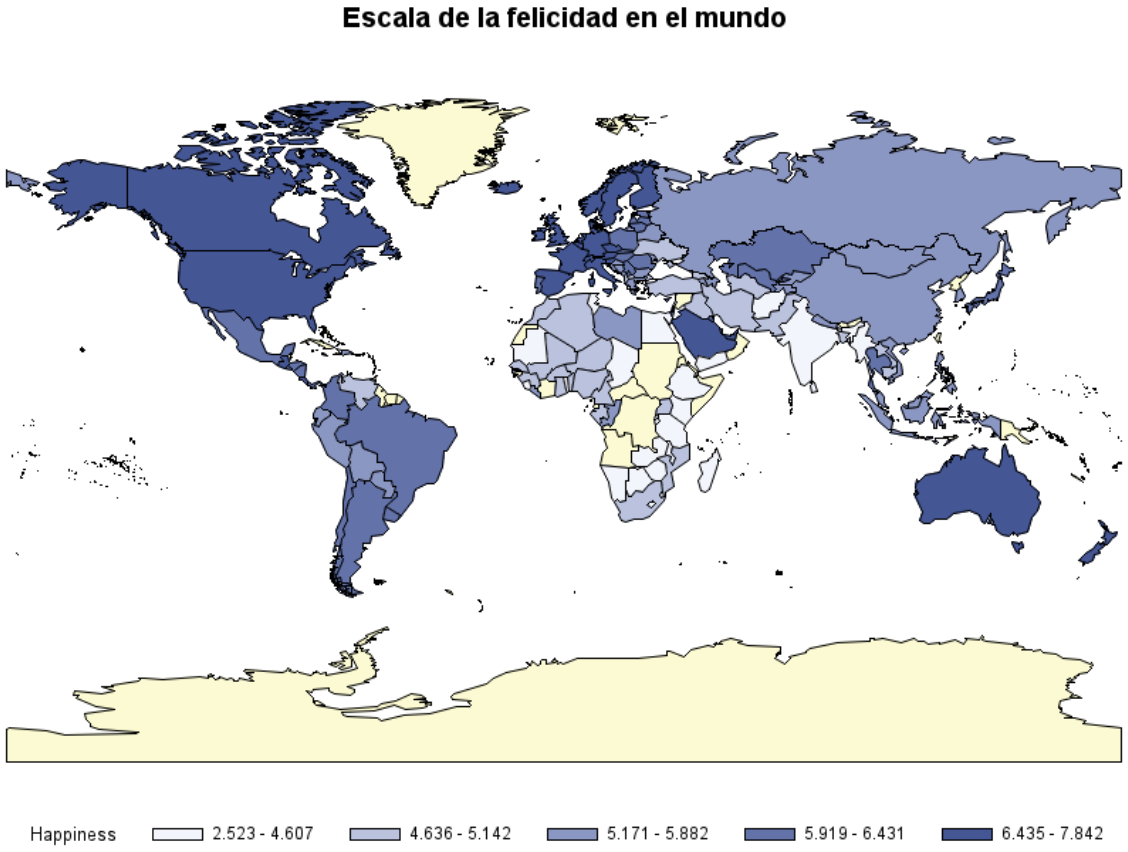
$$\boldsymbol{Kurtosis} = \frac{\sum_{i=1}^{160} \frac{(x_i - \bar{x})^4}{160}}{S^4} - 3 \qquad \boldsymbol{Asimetría} = \frac{\sum_{i=1}^{160} \frac{(x_i - \bar{x})^3}{160}}{S^3} \qquad \boldsymbol{Rango} = Máximo - Mínimo \qquad \boldsymbol{Rango\ intercuartílico} = RI = Q_3 - Q_1$$

Con respecto a nuestra **variable objetivo *Happiness***, la media se encuentra en 5.53. El país más feliz se sitúa en 7.84 en la escala de la felicidad mientras que la media del país más infeliz es 2.23. Por análisis anteriores hechos en Kaggle con esta misma base de datos, se tiene:



Fuente: Kaggle
Ilustración 1. Ranking países más felices y menos felices en el mundo

- Los 3 países con mayor felicidad son Finlandia, Dinamarca y Suiza, Islandia y Países Bajos siendo este primero el más feliz.
- Sin embargo, Ruanda, Zimbabue y Afganistán son los países más infelices, encabezando Afganistán como el que menor puntuación posee.



Mapa 1. Escala de la felicidad de en mundo

OBS: Los países en amarillo claro son las que no poseen data.

A la vista del mapa, es el **continente americano** (y ANZ) **o europeo** donde la mayoría de los países son de tono más oscuro, es decir, se consideran bastante felices. Es en este último donde se sitúan los valores más cercanos a 8: Islandia, Dinamarca, Finlandia y Suiza.

En **Asia** encontramos variedad. Desde países muy o poco felices (Afganistán, India, Birmania), pasando por índices medios (China o Indonesia) hasta felices (Japón o Arabia saudita, entre otros).

En **África**, son varios los países de los que no se tienen datos. Sin embargo, de los que poseemos, su escala de la felicidad es como mucho media. Aunque, en general éstos no se sienten ni felices ni infelices (cercanos a 5), es el continente que toma el tono más claro y no se observa ninguna que sea extremadamente feliz.

7. DEPURACIÓN DE DATOS

Este apartado será llevado a cabo con el software estadístico SAS MINER.

7.1. Corrección de errores

Por un lado, comenzaremos con las **variables cuantitativas**. Ya vimos por el análisis descriptivo, que las métricas de nuestras variables estaban correctamente a excepción de *Prevalencia* que era la que presentaba valores faltantes. Además, no hay problemas con los máximos y mínimos porque *Generosity* es la única variable que puede tomar valores negativos.

Por otro lado, en las **variables cualitativas** tampoco existen valores ausentes. Sin embargo, hay que ver si hay categorías que supongan un porcentaje igual o menor al 5% debido a que, en ese caso, habría que recategorizarlas.

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
Clima	Clima	C	15	0
Gobierno	Gobierno	C	12	0
Peligroso	Peligroso	N	2	0
Regional_indicator	Regional ...	C	10	0

Tabla 7-1. Niveles y ausentes en las variables cualitativas

Clima

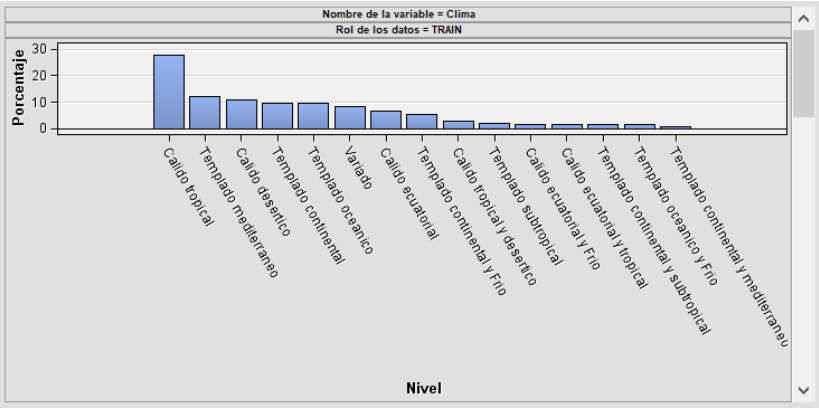


Gráfico de frecuencias (en %) 1. Clima

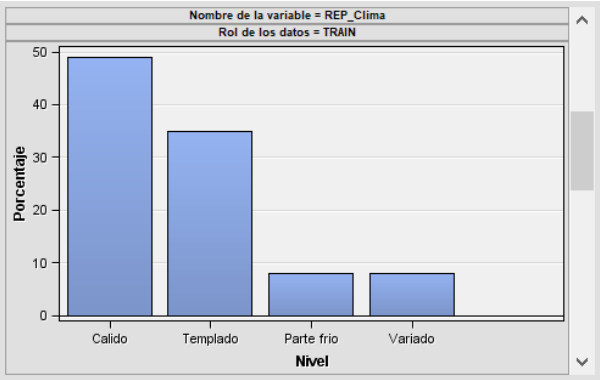


Gráfico de frecuencias (en %) 2. REP_Clima

Como existen categorías con poca frecuencia, entonces éstas serán recategorizadas:

- *Calido* (47.65) sin diferenciar en subclimas dentro del mismo. Estos seran: *calido tropical* (27.52), *calido desertico* (10.74), *calido ecuatorial* (6.71), *calido tropical y desértico* (2.68) y, *calido ecuatorial y tropical* (1.34).
- *Templado* (34.09) también sin diferenciar en subclimas. Estos serán: *templado mediterráneo* (12.08), *templado continental* (9.4), *templado oceánico* (9.4), *templado subtropical* (2.01), *templado continental y subtropical* (1.34) y, *templado continental y mediterráneo* (0.67).
- *Variado* (8.05), el país contiene clima cálido y templado. Sin cambios.
- *Parte frío* (8.05), es decir, aquellos países cuyo clima templado o cálido tiene una parte con clima frío (de montaña o polar). Estos serán *templado continental y Frio* (5.37), *cálido ecuatorial y Frío* (1.34) y, *templado oceánico y Frio* (1.34).

Gobierno

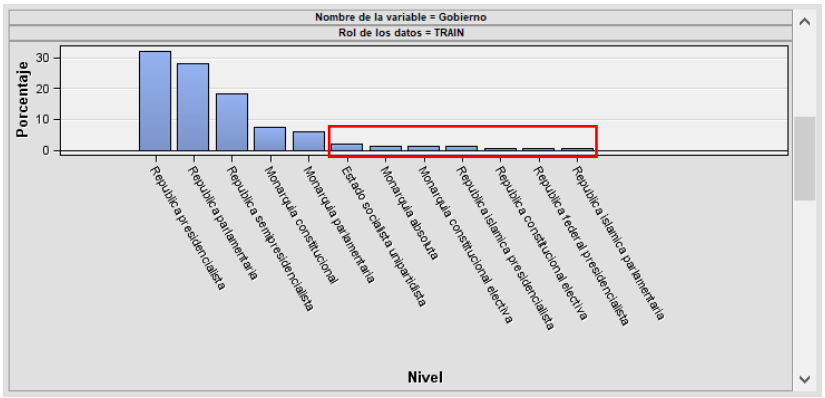


Gráfico de frecuencias (en %) 3. Gobierno

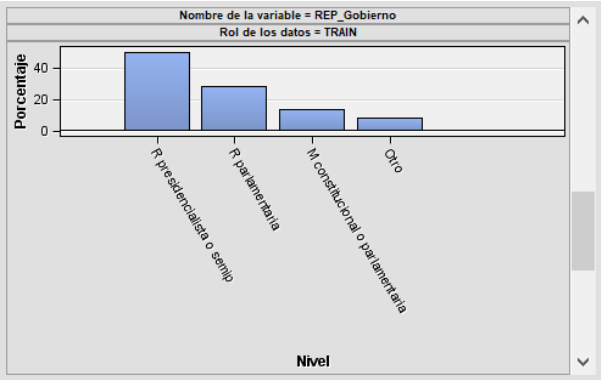


Gráfico de frecuencias (en %) 4. REP_Gobierno

- *R presidencialista o semipresidencialista* (50.33): países cuyo tipo de gobierno es *república presidencialista* (32.21) o *semipresidencialista* (18.12).
- *R parlamentaria (28.19):* república parlamentaria. Sin cambios.
- *M constitucional o parlamentaria* (13.42): *monarquía constitucional* (7.38) y *monarquia parlamentaria* (6.04).
- *Otros* (7.37): formado por todas las categorías recuadradas en rojo.

Regional_indicator

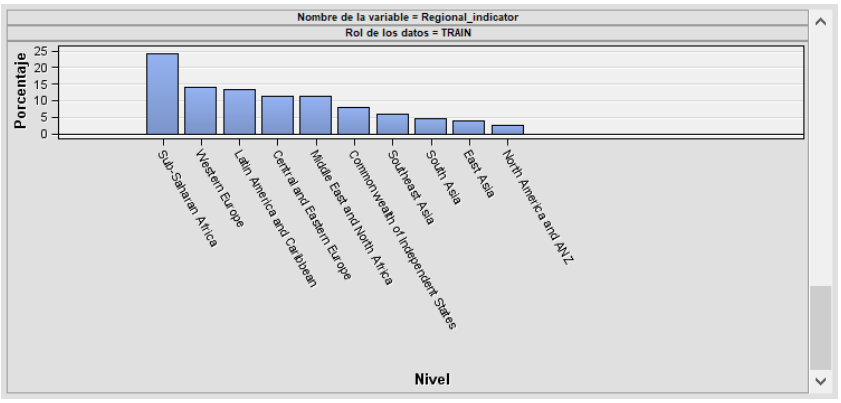


Gráfico de frecuencias (en %) 5. Regional_indicator

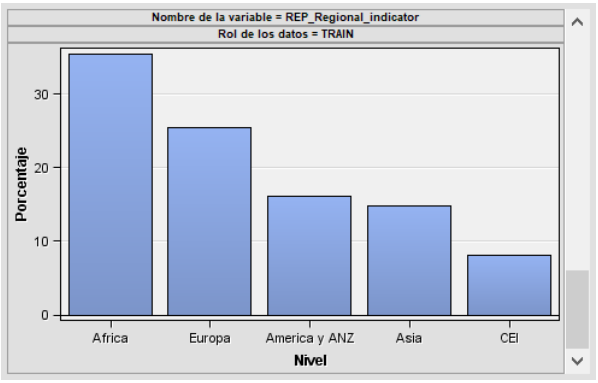


Gráfico de frecuencias (en %) 6. REP_Regional_indicator

- *Africa* (35.57): *Sub-Saharan Africa* (24.16) y *Middle East and North Africa* (11.41).
- *Asia* (14.77): *Southest Asia* (6.04), *South Asia* (4.7) y *East Asia* (4.03).

- *America y ANZ* (16.1): *Latin America and Caribbean* (13.42) y N*orth America y ANZ* (2.68).
- *Europa* (25.5): *Western Europe* (14.09) y *Central and Eastern Europe* (11.41).
- *CEI* (8.05): Comunidades de Estados Independientes. Sin cambios.

Peligroso

No necesita cambios.

7.2. Atípicos

En primer lugar, procedemos a la **partición de los datos** en train (70%) y test (30%) quedando 100 y 49 observaciones respectivamente. Esto se debe a que los valores atípicos se deben mirar en el fichero de entrenamiento.

Un dato es **atípico** (u outlier) cuando es numéricamente distante del resto de los datos. Por lo tanto, solo se estudiarán en las variables cuantitativas o nominales. Para poder determinarlo, hay que fijarse en la simetría de las variables. Para el caso de la depuración, consideraremos que una variable es simétrica cuando su coeficiente de asimetría pertenece al intervalo [-1,1]. En caso contrario, se dice que es asimétrica lo que significa que hay atípicos en la variable.

Inputs ordenados	Rol de los datos	Variable	Asimetría ▼	Mediana
2	TRAIN	Desempleo	1.422694	6.4
1	TRAIN	Generosity	1.226838	-0.034
8	TRAIN	Prevalencia	0.003747	4.5
3	TRAIN	Horas sol	-0.06318	2421
9	TRAIN	Ln PIB per capita	-0.3033	9.557
5	TRAIN	IDH	-0.38173	0.756
10	TRAIN	Healthy life expectancy	-0.42985	66.402
11	TRAIN	Esperanza vida	-0.57385	75.21
6	TRAIN	Freedom	-0.72808	0.788
7	TRAIN	Social support	-1.01066	0.83
4	TRAIN	Corruption	-1.59895	0.776

Tabla 7-2. Asimetría y mediana variables cuantitativas

A la vista de la tabla *Tabla 7-3.* se tiene que todas las variables seleccionadas con tono más claro cumplen la simetría. Las demás son asimétricas. A continuación, el tratamiento de datos atípicos fue:

- Para variables simétricas se utilizará el método de desviación estándar.
- Para variables asimétricas se utilizará el método MAD (Median Absolute Deviation) debido a que en todos los casos la mediana es distinta de 0: *Desempleo, Generosity, Social_Support* y *Corruption*.

Cuentas de reemplazo total

Variable	Rol	Etiqueta	Entrenamiento	COL2
Corruption	INPUT	Corrupti...	0	0
Desempleo	INPUT	Desem...	0	0
Esperanza_vida	INPUT	Espera...	0	0
Freedom	INPUT	Freedom	1	0
Generosity	INPUT	Genero...	0	0
Healthy_life_e	INPUT	Healthy...	0	0
Horas_sol	INPUT	Horas ...	0	0
IDH	INPUT	IDH	0	0
Ln_PIB_per_cap	INPUT	Ln(PIB ...	0	0
Prevalencia	INPUT	Prevale...	0	1
Social_support	INPUT	Social ...	0	0

Tabla 7-4. Cuentas de reemplazo total atípicos

Únicamente ha habido dos reemplazos de atípicos: uno en train en *Freedom* con el método de desviación estándar y el otro en test *Prevalencia* con el método MAD.

A continuación, se estudiarán las **observaciones** que son **potencialmente atípicas** ya que deben representar una proporción pequeña del conjunto de datos. En esta ocasión, como el número de observaciones detectadas atípicas en el fichero de entrenamiento en *Freedom* es inferior al 5%, entonces se debe de transformar en ausente.

7.3. Datos faltantes

Al igual que con los datos atípicos, la presencia de datos faltantes (o missings) debe de ser analizada ya que muchos de los procedimientos estadísticos no aplican si éstos existen en la base de datos. Además, puede ocurrir que los missing no estén de forma aleatoria en los datos.

Las estrategias que se siguieron fueron:

- Eliminación de variables y/o observaciones cuando exista más de 50% de missing.
- Imputación: se sustituirá el missing por un valor válido.
- Recategorización: incluir el missing como una categoría en el caso de variables de clase. En este caso, tampoco va a ser necesario porque ninguna poseía datos ausentes.

La variable *REP_prevalencia* tiene 2 ausentes mientras que *REP_Freedom* 1. En cuanto a las variables categóricas, no hay. Por tanto, no se rechaza ninguna variable de intervalo porque no ninguna supera el 50% de datos ausentes.

Variables de intervalo

Variable	Etiqueta	Ausente ▼	N
REP_Prevalencia	Replace...	2	98
REP_Freedom	Replace...	1	99
Happiness	Happiness	0	100
REP_Corruption	Replace...	0	100
REP_Desempleo	Replace...	0	100
REP_Esperanza_vida	Replace...	0	100
REP_Generosity	Replace...	0	100
REP_Healthy_life_expectancy	Replace...	0	100
REP_Horas_sol	Replace...	0	100
REP_IDH	Replace...	0	100
REP_Ln_PIB_per_capita	Replace...	0	100
REP_Social_support	Replace...	0	100

Tabla 7-5. Ausentes variables cuantitativas

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
Happiness_cat	Happines...	N	6	0
Peligroso	Peligroso	N	2	0
REP_Clima	Replace...	C	4	0
REP_Gobierno	Replace...	C	4	0
REP_Regional_indicator	Replace...	C	5	0

Tabla 7-6. Ausentes variables cualitativas

En cuanto a volumen de ausentes por observaciones, al haber sólo 3 en toda la base de datos, no es necesaria la creación de la variable *NumMissing* porque como mucho van a haber missing en una misma observación (<50%).

Luego, no se eliminarán variables ni observaciones así que pasamos a la *Imputación* de dichos valores mediante el método de *Distribución*.

Se verifica que ya no haya ausentes y que todos los estadísticos descriptivos estén correctos:

Estadísticos descriptivos de la variable de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
Happiness	Happiness	0	100	2.52	7.57	5.49	1.070	-0.17790	-0.23240
IMP_REP_Freedom	Imputed: Replacement: Freedom	0	100	0.53	0.97	0.79	0.108	-0.47244	-0.50063
IMP_REP_Prevalencia	Imputed: Replacement: Prevalencia	0	100	3.20	5.90	4.48	0.598	-0.03313	-0.53896
REP_Corruption	Replacement: Corruption	0	100	0.08	0.94	0.73	0.176	-1.59895	2.44473
REP_Desempleo	Replacement: Desempleo	0	100	0.50	33.30	8.28	5.978	1.42269	2.53452
REP_Esperanza_vida	Replacement: Esperanza_vida	0	100	54.79	85.03	73.31	7.599	-0.57385	-0.49519
REP_Generosity	Replacement: Generosity	0	100	-0.26	0.54	-0.01	0.156	1.22684	2.13074
REP_Healthy_life_expectancy	Replacement: Healthy_life_expectancy	0	100	48.70	76.95	64.71	6.938	-0.42985	-0.68689
REP_Horas_sol	Replacement: Horas_sol	0	100	1268.00	3605.00	2430.83	594.783	-0.06318	-0.96465
REP_IDH	Replacement: IDH	0	100	0.39	0.96	0.73	0.155	-0.38173	-0.92437
REP_Ln_PIB_per_capita_	Replacement: Ln(PIB_per capita)	0	100	6.64	11.49	9.41	1.174	-0.30330	-0.86982
REP_Social_support	Replacement: Social_support	0	100	0.46	0.98	0.81	0.117	-1.01066	0.68036

Estadísticos de sumarización de la variable de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
Happiness_cat	Happiness_cat	N	6	0
Peligroso	Peligroso	N	2	0
REP_Clima	Replacement: Clima	C	4	0
REP_Gobierno	Replacement: Gobierno	C	4	0
REP_Regional_indicator	Replacement: Regional_indicator	C	5	0

Tabla 7-7. Verificación ausentes y errores

Las variables se encuentran en perfectas condiciones para proseguir con el trabajo.

7.4. Medición de relación entre variables

Una vez que los datos están limpios, se debe estudiar la relación existente entre las variables input y la objetivo (escala o categoría de la felicidad).

Para ello, es necesaria la creación de dos variables aleatorias que nos servirá como punto de corte o como comparación para saber la utilidad de una variable a la hora de predecir.

Código de entrenamiento

```
DATA &EM_EXPORT_TRAIN;
SET &EM_IMPORT_DATA;
aleat1=rand("uniform");
aleat2=rand("uniform");
RUN;
```

```
DATA &EM_EXPORT_TEST;
SET &EM_IMPORT_TEST;
aleat1=rand("uniform");
aleat2=rand("uniform");
RUN;
```

Ilustración 2. Código de variables aleatorias en ficheros de entrenamiento y test

- Gráfico de correlación de Pearson: si está cerca del 0 entonces es que no hay correlación ni directa (próximo a 1) ni indirecta (próximo a -1), es decir, no sirve para predecir.

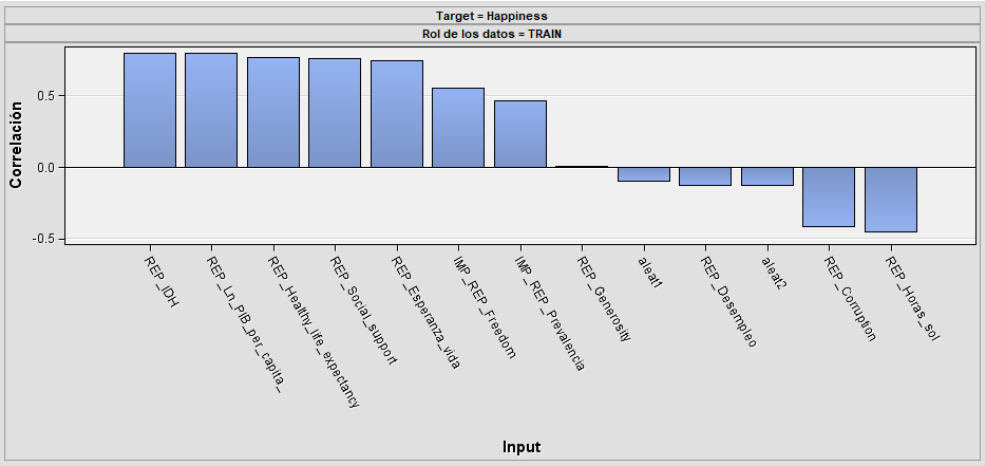


Gráfico de correlación de Pearson 1. Happiness

A la vista del grafico se observa que las variables *REP_Generosity* y *REP_Desempleo* no tienen mucha correlación con la felicidad de un país ya que aportan lo mismo (o incluso menos) que dos variables aleatorias cualquiera.

- Gráfico de valor: para ver la importancia de las variables.

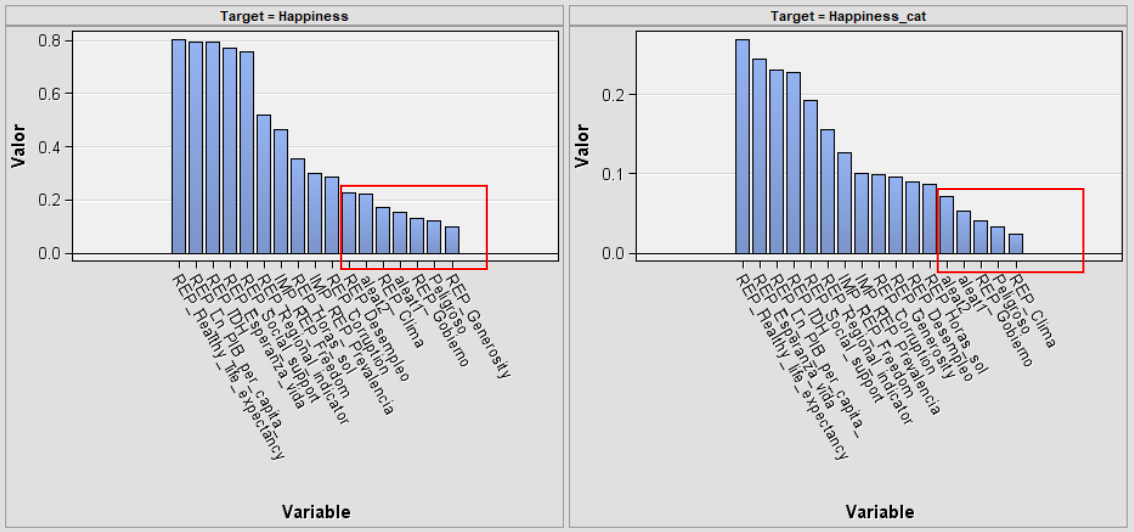


Gráfico de valor 1. Happiness y Happiness_cat

Independientemente si la variable objetivo es nominal o categórica, las 5 variables más importantes son la esperanza de vida saludable, el PIB per cápita, el índice de desarrollo del país, la esperanza de vida y el apoyo social. En las menos importantes ya se encuentra cierta discrepancia aunque coinciden en *REP_clima*, *Peligroso* y *REP_gobierno*. Es en *Happiness* donde también está *REP_Generosity* (la de menor valor), que en *Hapiness_cat* está considerado de mediana importancia.

- Diagramas de dispersión/barras: para ver correlación.

A continuación, se presentarán los **diagramas de dispersión o barras** para ver si existe o no relación lineal con nuestra variable objetivo.

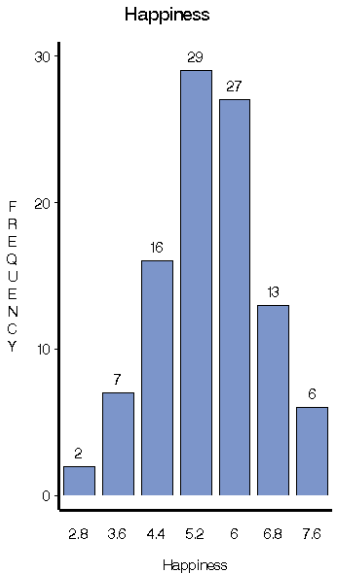


Gráfico histograma 2. Happiness

Los países con una escala de 5.2 a 6 son los más frecuentes, es decir, están en una felicidad neutra o felices.

Dos países son los que se encuentran en los valores más bajos de la variable (2.8) mientras que, por el contrario, 6 son los más felices (7.6).

7.4.1. Variables cualitativas

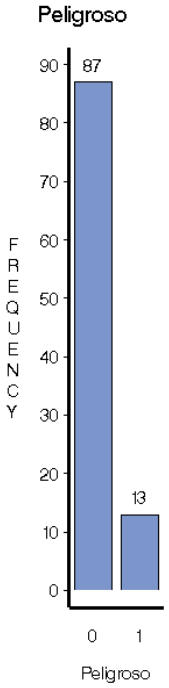


Gráfico de frecuencias (absolutos) 1. Peligroso

Gráfico 1. Peligroso vs Happiness (media)

En el fichero de entrenamiento se tienen 13 de los 20 países más peligrosos. Al parecer, la media en felicidad de que un país peligroso o seguro solo se diferencia en una unidad por lo que parece que influye pero no en gran medida.

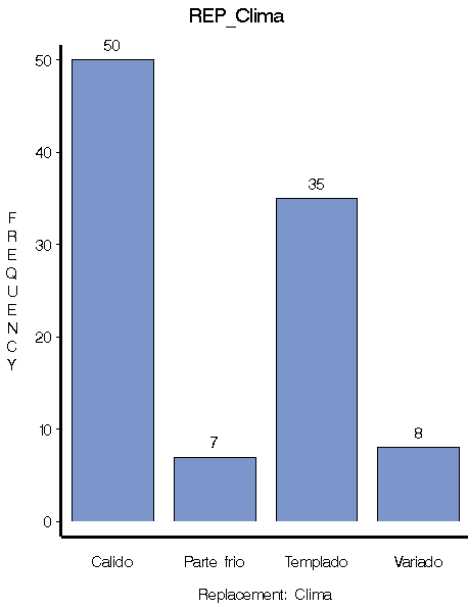


Gráfico de frecuencias (absolutos) 2. REP_Clima

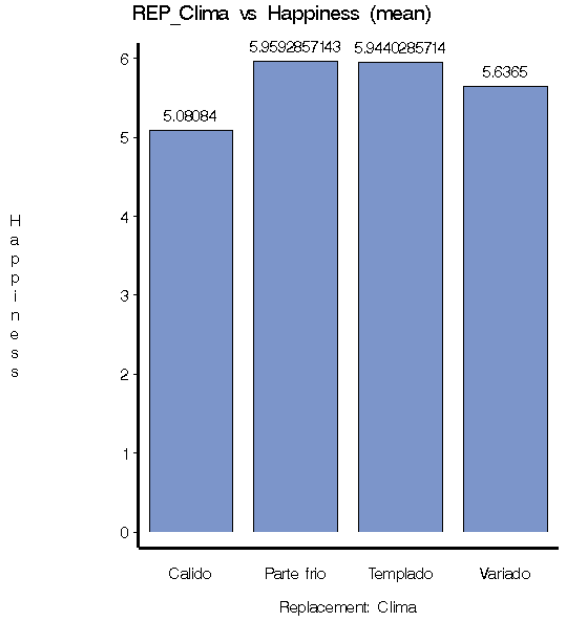


Gráfico 2. REP_Clima vs Happiness (media)

A la vista del gráfico, no hay mucha dispariedad en media entre un país y otro. Si que se observa que los países con un clima cálido (el más frecuente) son los menos felices en comparación con el resto pero tampoco hay diferencias significativas.

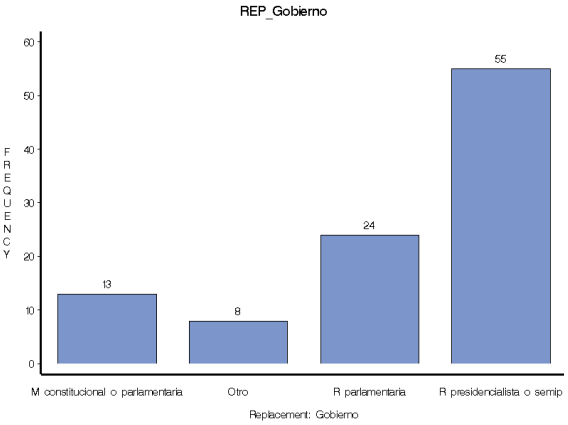


Gráfico de frecuencias (absolutos) 3. REP_ Gobierno

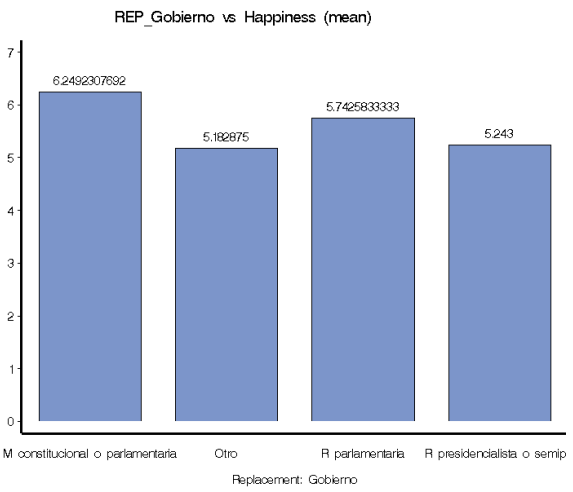


Gráfico 3. REP_Gobierno vs Happiness (media)

Los países cuyo tipo de gobierno es monarquía constitucional o parlamentaria o una república parlamentaria son los más felices (medias 6.25 y 5.74, respectivamente) mientras que los que poseen una república presidencialista o semipresidencialista, o tienen otro tipo de gobierno, su escala de felicidad promedio se encuentra en 5.2.

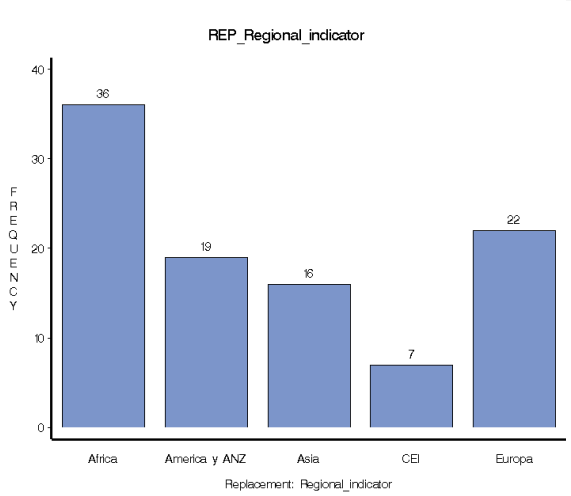


Gráfico de frecuencias (absolutos) 4. REP_Regional_indicator

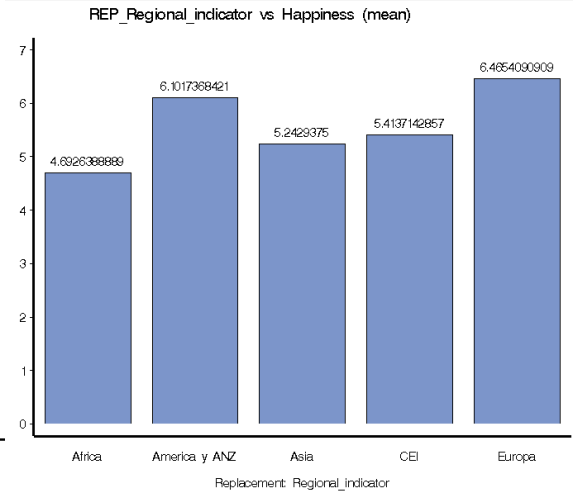


Gráfico 4. REP_Regional_indicator vs Happiness (media)

En este caso las regiones si que parecen tener relación con la variable objetivo *Happiness*.

Los países pertenecientes a la región africana son los más infelices mientras que los de Europa los más felices seguidos del continente america y ANZ. Los que se encuentran rondando el punto medio son Asia y las CEI.

7.4.2. Variables cuantitativas

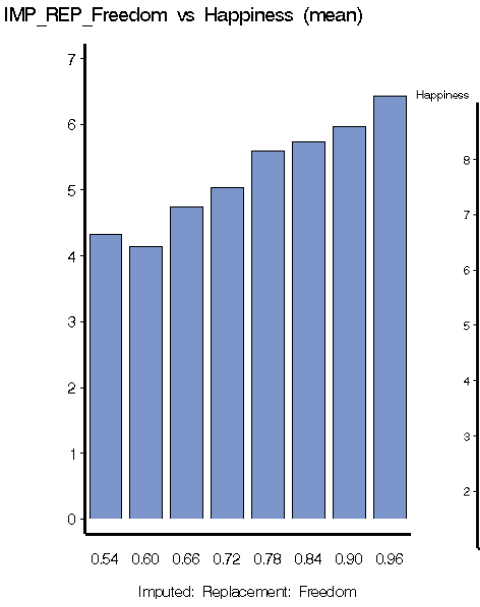


Gráfico de frecuencias (absolutos) 5. IMP_REP_Freedom vs Happiness (media)

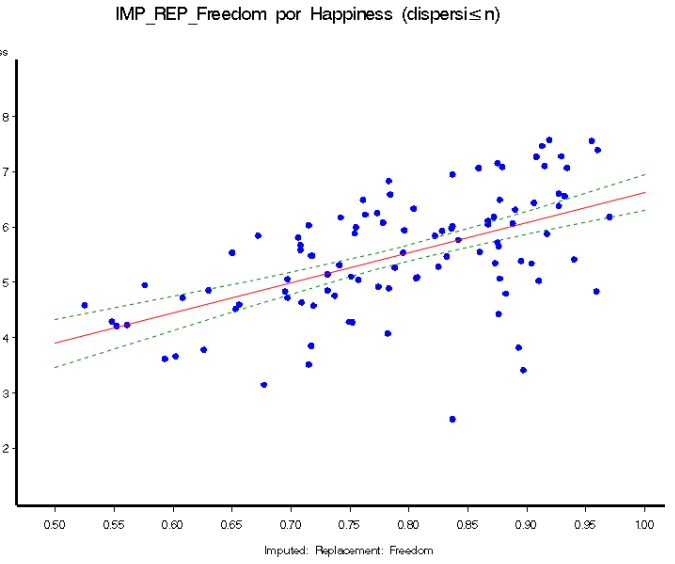


Gráfico de dispersión 1. IMP_REP_Freedom vs Happiness

La mayoría de países se sitúan en una media de libertad entre 0.72 y 0.9, siendo esta última la más frecuente, es decir, 26 de los 100 países en el fichero de entrenamiento considera que si que está satisfecho con respecto a la libertad de toma de decisiones en su vida.

A la derecha, se observa que *IMP_REP_Freedom* si que tiene correlación directa con *Happiness*, es decir, cuanta más libertad mayor es la escala de la felicidad.

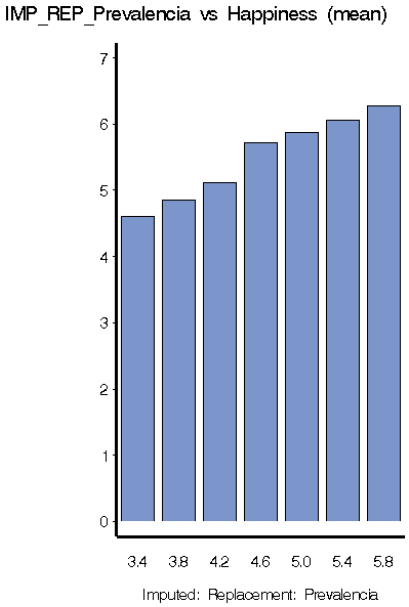


Gráfico de frecuencias (absolutos) 6. IMP_REP_Prevalencia vs Happiness (media)

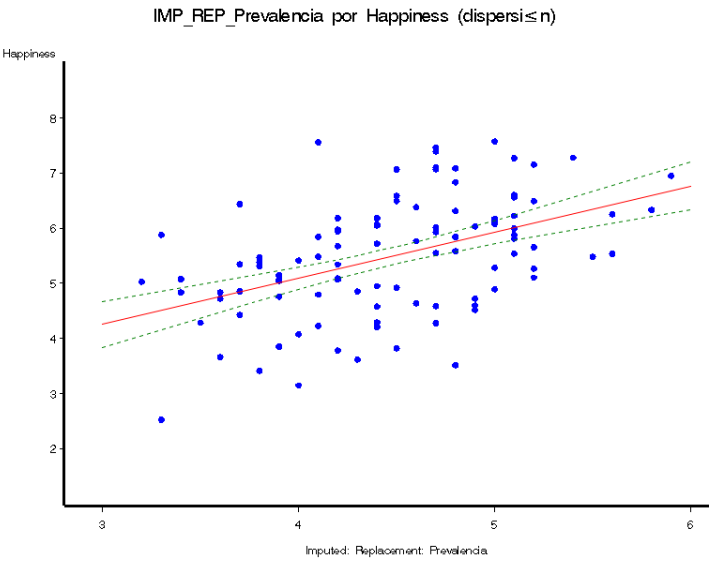


Gráfico de dispersión 2. IMP_REP_Prevalencia vs Happiness

Si que existe una clara correlación entre la prevalencia de depresión de cada país con la felicidad. A pesar que, cuanta más prevalencia mayor felicidad. Según el último estudio de la OMS, y con un margen de error entre 2% y 6%, los 10 países con una tasa más alta y baja del mundo son, repectivamente:

1. Ukraine (6.3%)	1. Solomon Islands (2.9%)
2. United States (5.9%)	2. Papua New Guinea (3.0%)
3. Estonia (5.9%)	3. Timor – Leste (3.0%)
4. Australia (5.9%)	4. Vanuatu (3.1%)
5. Brazil (5.8%)	5. Kiribati (3.1%)
6. Greece (5.7%)	6. Tonga (3.2%)
7. Portugal (5.7%)	7. Samoa (3.2%)
8. Belarus (5.6%)	8. Laos (3.2%)
9. Finland (5.6%)	9. Nepal (3.2%)
10. Lithuania (5.6%)	10. Philippines (3.3%)

Fuente: World Popular review 2021
Ilustración 3. Ranking países con alta y baja tasa de prevalencia

Es decir, se ve que los países más felices son, a su vez, los que poseen una prevalencia más alta. Esto también se puede deber a que existen países en los que es difícil contabilizar esta métrica.

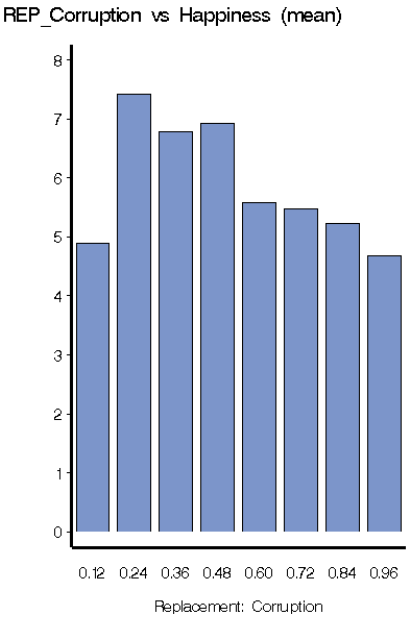


Gráfico de frecuencias (absolutos) 7. REP_Corruption vs Happiness (media)

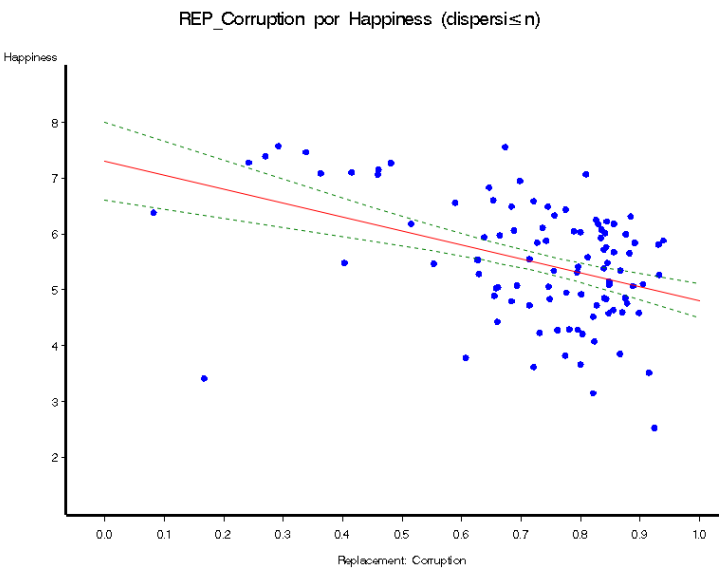


Gráfico de dispersión 3. REP_Corruption vs Happiness

Existe correlacion negativa entre la escala de la felicidad con la percepción que se tiene de la corrupción, es decir, cuando dicha sensación es alta entonces Happiness toma valores bajos.

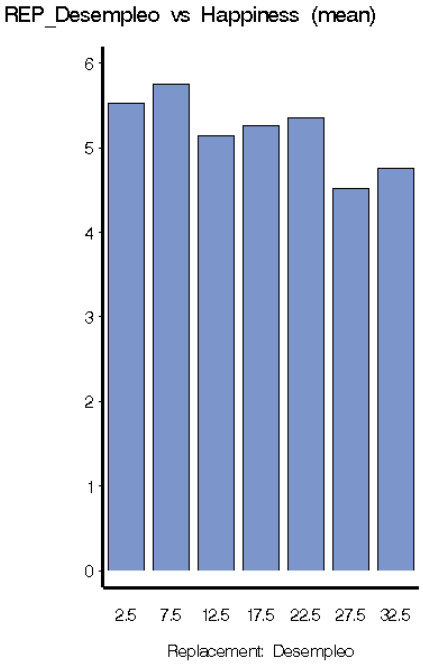


Gráfico de frecuencias (absolutos) 8. REP_Desempleo vs Happiness (media)

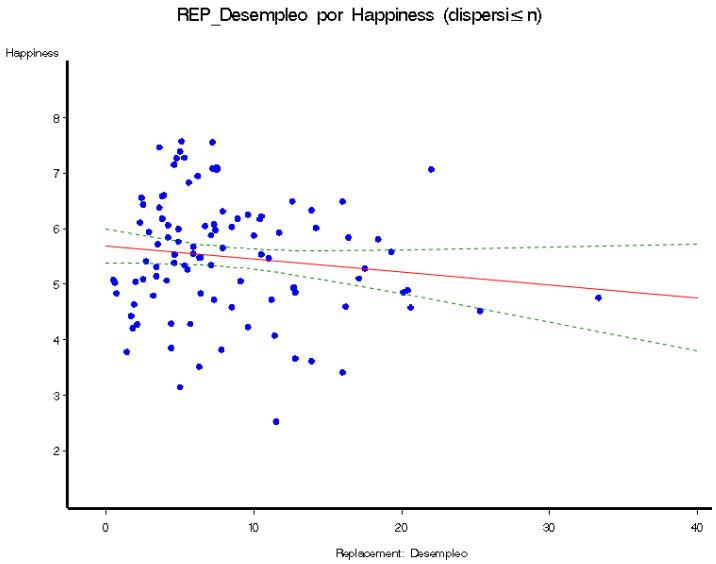


Gráfico de dispersión 4. REP_Desempleovs Happiness

Ligera relación entre el desempleo y la felicidad. Se ve que, a medida que para valores pequeños de la tasa de desempleo, la felicidad es mayor.

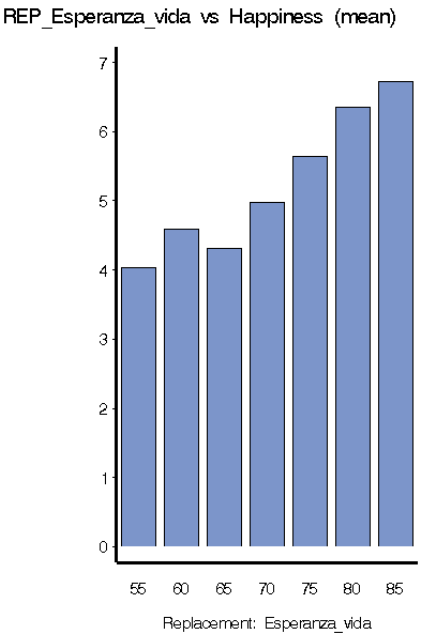


Gráfico de frecuencias (absolutos) 9. REP_Esperanza_vida vs Happiness (media)

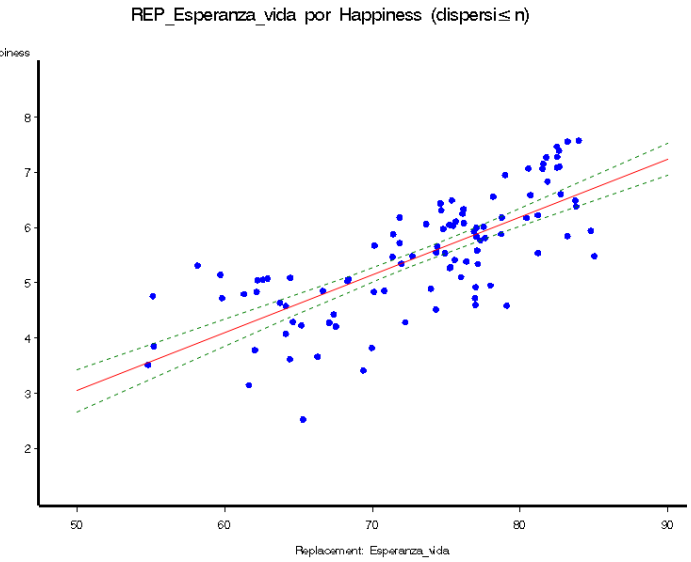


Gráfico de dispersión 5. REP_Esperanza_vida vs Happiness

Existe clara correlación positiva entre la esperanza de vida del país con la felicidad.

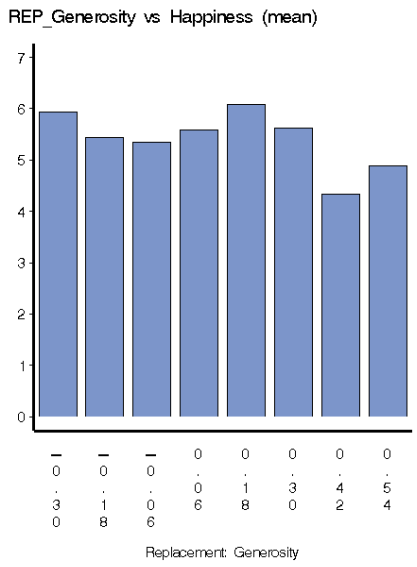


Gráfico de frecuencias (absolutos) 10. REP_Generosity vs Happiness (media)

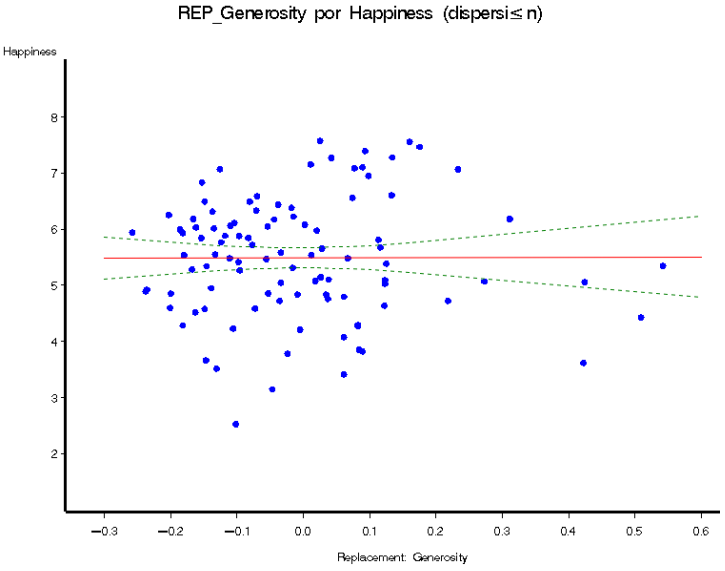


Gráfico de dispersión 6. REP_Generosity vs Happiness

No se ve una clara correlación. Ni directa ni indirecta. Todas las demás parece ser que si que están correlacionadas con *Happiness*

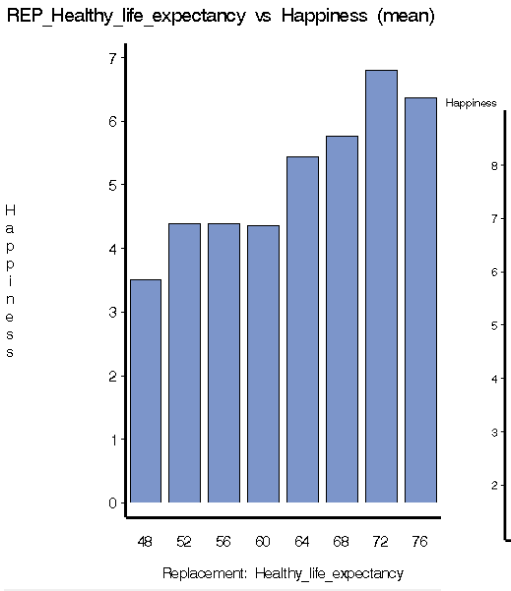


Gráfico de frecuencias (absolutos) 11. REP_Healthy_life_expectancy vs Happiness (media)

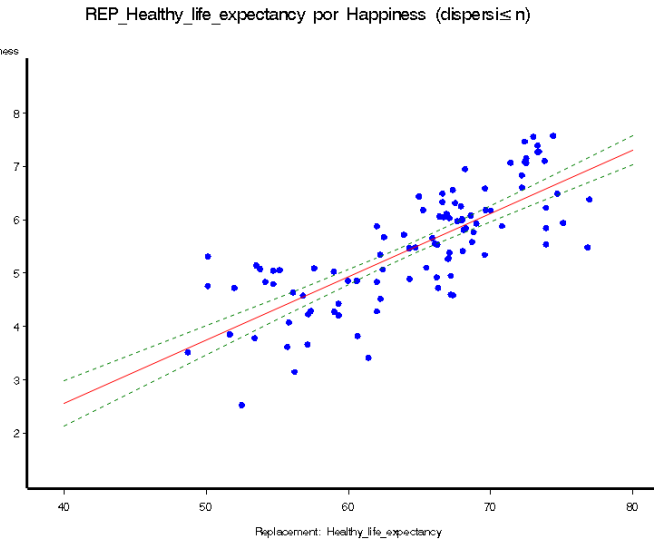


Gráfico de dispersión 7. REP_Healthy_life_expectancy vs Happiness

A la vista de ambos gráficos, existe un destacable relación positiva ya que cuando la edad media de años que una persona pueda vivir con salud plena es cada vez mayor, entonces la escala media de la felicidad en el país aumenta.

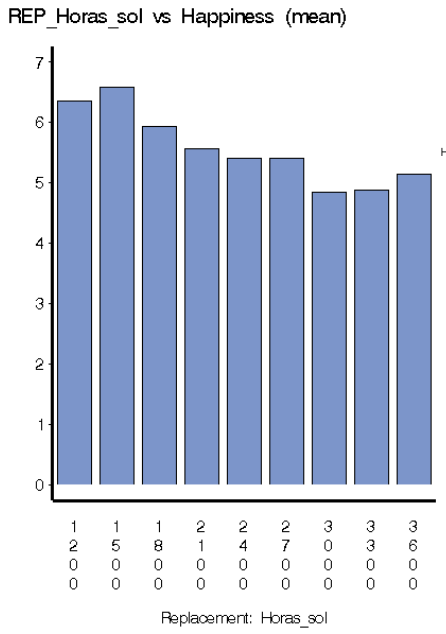


Gráfico de frecuencias (absolutos) 12. REP_Horas_sol vs Happiness (media)

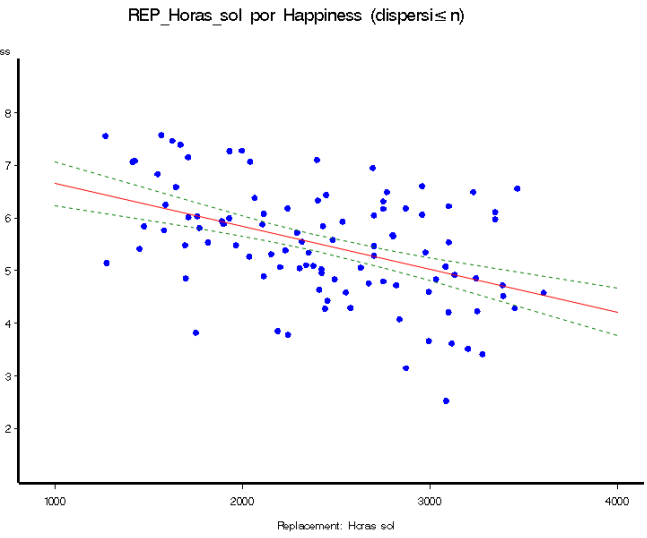


Gráfico de dispersión 8. REP_Horas_sol vs Happiness

Correlación negativa. Al parecer, a medida que las horas de sol al año son menores, entonces la media de la felicidad en esos países es mayor.

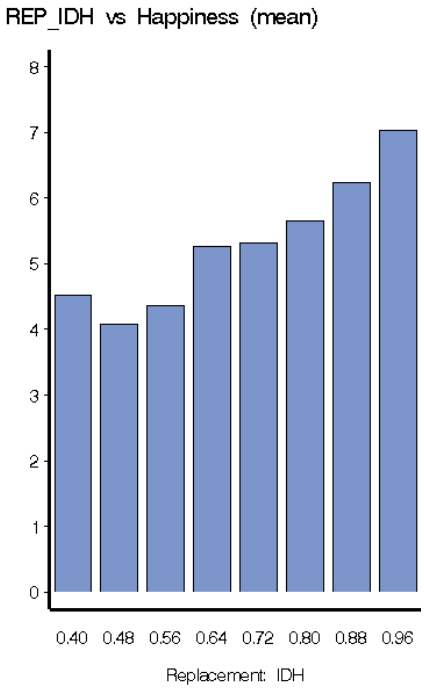


Gráfico de frecuencias (absolutos) 13. REP_IDH vs Happiness (media)

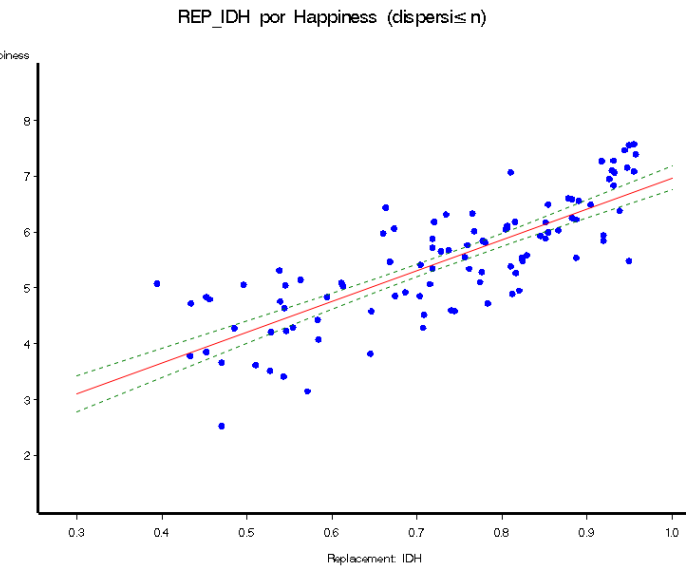


Gráfico de dispersión 9. REP_IDH vs Happiness

Correlación positiva, cuanto mayor es el índice de desarrollo del país mayor es *Happiness*.

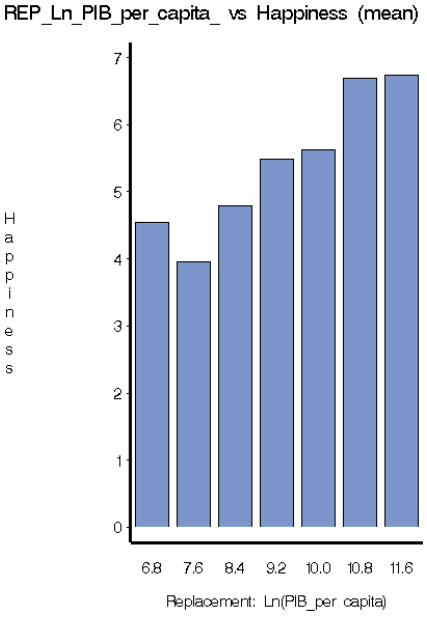


Gráfico de frecuencias (absolutos) 14. REP_Ln_PIB_per_capita vs Happiness (media)

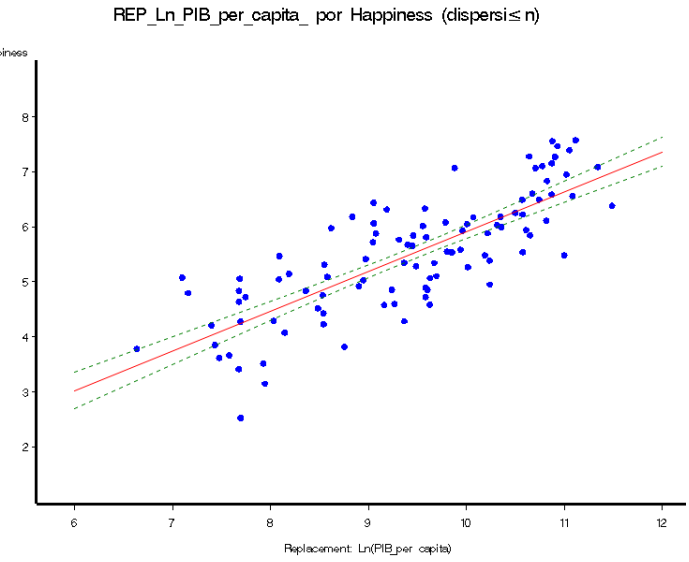


Gráfico de dispersión 10. REP_Ln_PIB_per_capita vs Happiness

Clara asociación lineal directa entre la felicidad y el PIB, es decir, son los páises con una tasa de PIB más alto los más felices mientras que los más infelices poseen un tasa mucho menor.

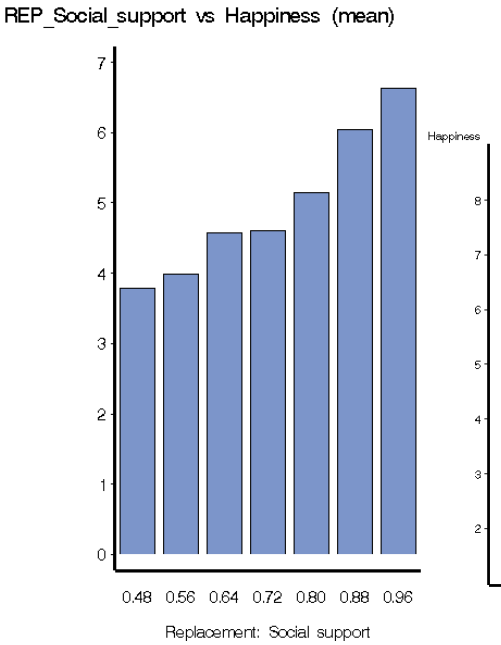


Gráfico de frecuencias (absolutos) 15. REP_Social_support vs Happiness (media)

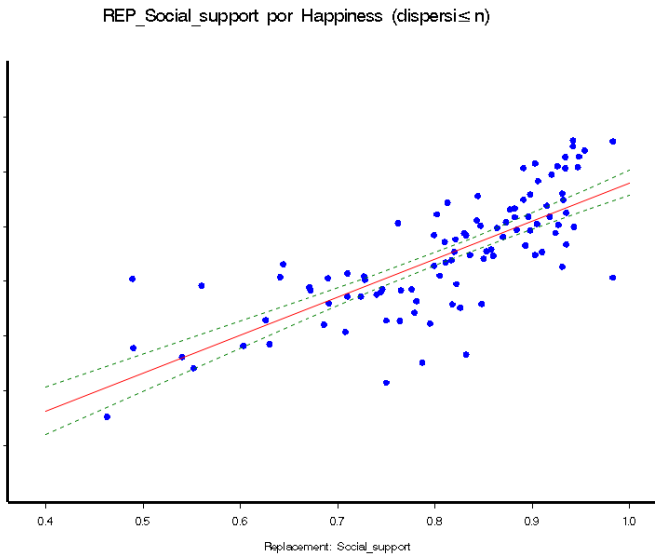


Gráfico de dispersión 11. REP_Social_support vs Happiness

Los países que cuyo valor en *REP_Social_support* es próximo a 1 coincide en que toma valores en la felicidad medio-altos. Por el contrario, son más infelices los que no sienten que tengan ese apoyo social.

8. APRENDIZAJE NO SUPERVISADO - FAMD

En la actualidad es muy común trabajar con datos mixtos, es decir, con conjuntos de datos de tipo cualitativo y cuantitativo. Sobre todo, es muy frecuente encontrarlo en datos que provienen de encuestas donde existen preguntas tanto de tipo cerrada como abierta (Bécue-Bertaut & Pagès, 2008). Este problema ya fue estudiado en 1971 a través del coeficiente de similaridad de Gower (Gower, 1971) comúnmente utilizado en el Análisis Clúster cuando se tiene una base de datos multivariada de naturaleza mixta.

En este proyecto esta situación fue tratada con el AFMD. Se trata una técnica de aprendizaje no supervisado consistente en el Análisis de Componentes Principales (ACP) pero en el cual se pueden utilizar variables mixtas. Con este tipo de análisis es posible analizar la similitud entre individuos teniendo en cuenta que se tienen datos tanto continuos como nominales para poder reducir la dimensión. Además, también se puede ver la asociación entre ambos tipos. Por tanto, esta técnica se podría considerar como una mezcla entre el ACP y el ACM (Análisis de Correspondencias Múltiple).

En este caso, el AFMD será llevado a cabo con R con las librerías `"FactoMineR"` y `"factoextra"`.

Antes de comenzar el análisis, en el caso de las variables cuantitativas, se puede sacar la matriz de correlación (R) para ver qué grado de asociación tienen. Si el valor es próximo a ±1 entonces es que existe una fuerte correlación entre esas dos variables mientras que si es cercano a 0 hay incorrelación.

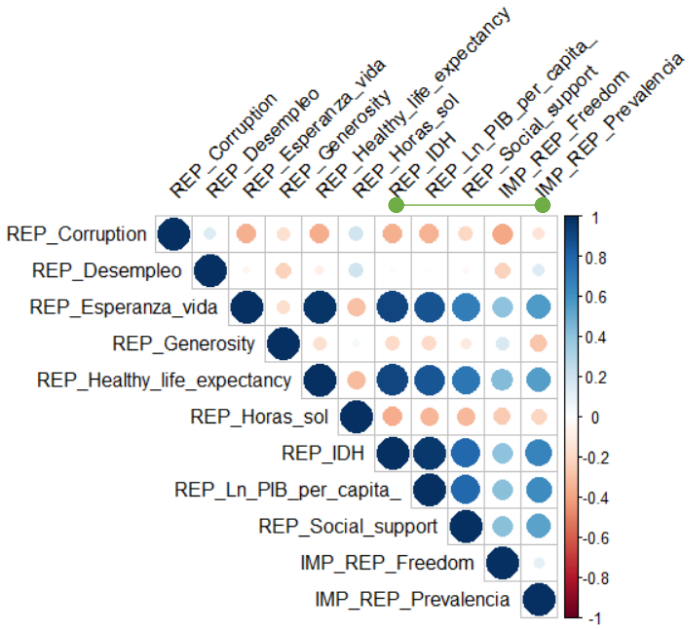


Ilustración 4. Matriz de correlaciones

Se observa que las asociaciones más significativas son en **positivo**. La variable *REP_Esperanza_vida* posee una correlación bastante alta y positiva con todas (claramente la más fuerte con *REP_Healthy_life_expectancy*) las variables excepto con *REP_Generosity* y *REP_Horas_sol*. *REP_Healthy_life_expectancy, REP_Ln_PIB_per_capita* y *REP_social_support* correlacionadas con las variables subrayadas en verde.

En la parte **negativa**, las variables con todo o casi todo de tono más rojizo se encuentran en *REP_Corruption* y *REP_Horas_sol*.

Por último, las que están **incorreladas** *REP_Desempleo* con *REP_Esperanza_vida*, *REP_Healthy_life_expectancy, REP_Ln_PIB_per_capita* y *REP_social_support; y REP_Generosity* con *REP_horas_sol.*

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.76349639	30.7431654	30.74317
Dim.2	2.20785341	10.0356973	40.77886
Dim.3	1.68040772	7.63821689	48.41708
Dim.4	1.5066314	6.84832456	55.2654
Dim.5	1.36976611	6.2262096	61.49161
Dim.6	1.14661211	5.21187324	66.70349
Dim.7	0.9318556	4.23570729	70.93919
Dim.8	0.82432	3.74690907	74.6861
Dim.9	0.81281236	3.69460165	78.38071
Dim.10	0.70803874	3.21835793	81.59906
...	...	...	...
Dim.18	0.23392552	1.0632978	98.54458
Dim.19	0.17534093	0.79700424	99.34158
Dim.20	0.11165751	0.50753414	99.84911
Dim.21	0.01916167	0.0870985	99.93621
Dim.22	0.01403323	0.06378741	100

Tabla 8-1. Autovalores, % varianza

Con el criterio de los autovalores se escogerían 6 factores explicando un 66.70% de la varianza mientras que con el criterio de la varianza explicada (>80%) nos quedaríamos con 10 factores.

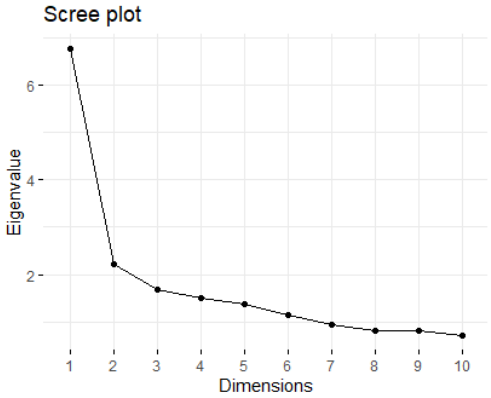


Gráfico 5. Dimensiones vs Autovalores

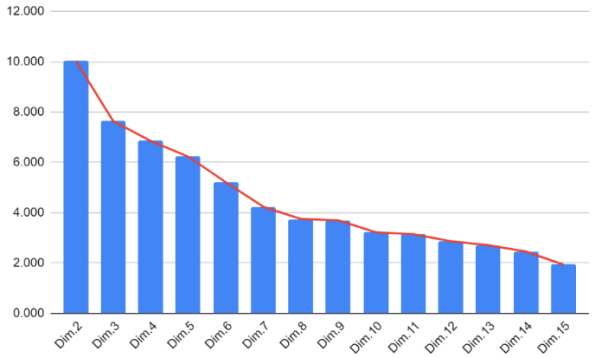


Gráfico 6. Dimensiones vs Autovalores

A la vista del gráfico de sedimentación, el punto de codo se aprecia en 2 factores que, en este caso, no es factible puesto que no se llegaría ni a explicar la mitad de la variabilidad de los datos. Para tener una mayor visibilidad, creamos otro gráfico eliminando la primera dimensión (gráfico 6). En él, podrían considerarse como puntos de codo: dim.3, dim.8 y dim.10. El de 3 dimensiones no es posible ya que explicaríamos muy poca variabilidad (48.42%). **Nos quedaremos con 8 factores** explicando de esta manera un 74.69%.

Se repite la función FAMD indicándole los 8 factores. Después, se hace un summary para obtener la siguiente información descritas posteriormente: las coordenadas (*dim.i*), las contribuciones (*ctr*) y las cargas (*cos2*). Todas estas métricas a nivel individual como del tipo de variable.

Correlaciones	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
REP_Corruption	-0.400	0.405	0.226	-0.039	0.242	-0.393	-0.147	-0.023
REP_Desempleo	-0.051	0.588	0.240	0.101	0.031	0.353	0.294	0.070
REP_Esperanza_vida	0.917	-0.037	0.172	-0.067	-0.048	0.040	-0.111	0.076
REP_Generosity	-0.173	-0.535	-0.247	-0.120	-0.014	-0.002	0.402	0.460
REP_Healthy_life_expectancy	0.923	-0.063	0.149	-0.021	-0.053	-0.001	-0.124	0.098
REP_Horas_sol	-0.467	0.105	0.271	-0.002	-0.313	0.406	-0.042	0.039
REP_IDH	0.956	0.044	0.109	-0.002	-0.007	0.085	-0.083	0.053
REP_Ln_PIB_per_capita_	0.919	0.032	0.134	-0.064	-0.031	0.113	-0.092	0.044
REP_Social_support	0.816	0.022	0.052	0.143	0.028	-0.036	-0.060	0.177
IMP_REP_Freedom	0.493	-0.612	0.034	0.127	0.044	0.079	0.150	-0.126
IMP_REP_Prevalencia	0.692	0.408	0.041	0.161	-0.123	0.017	0.097	-0.150
0	0.259	-0.182	-0.189	0.064	-0.138	-0.122	0.021	-0.179
1	-1.670	1.176	1.217	-0.411	0.893	0.787	-0.133	1.156
Calido	-1.825	-0.349	0.005	0.044	-0.312	0.063	-0.137	0.061
Parte Frio	1.961	-0.723	-1.479	1.928	1.564	1.282	0.445	-0.520
Templado	2.018	0.838	-0.334	-0.579	-0.119	-0.267	-0.234	0.112
Variado	0.396	-0.786	2.898	0.313	0.848	-0.512	1.401	-0.336

M constitucional o parlamentaria	2.108	-1.037	-0.164	-0.215	-1.564	0.874	-0.187	0.137
Otro	-1.260	-1.216	1.833	-1.541	1.121	1.182	-0.506	-1.277
R parlamentaria	1.399	0.775	-0.487	-0.671	0.643	-0.087	0.660	0.039
R presidencialista o semip	-1.144	0.037	0.023	0.680	-0.122	-0.373	-0.239	0.146
Africa	-2.443	0.516	-0.211	-0.222	-0.504	0.403	0.052	-0.096
America y ANZ	0.644	-0.494	1.733	1.448	-0.731	-0.698	0.305	0.096
Asia	-0.242	-1.987	0.579	-1.245	1.059	-0.200	-0.563	0.292
CEI	0.489	0.261	-1.126	2.321	1.928	0.787	-0.908	0.321
Europa	2.986	0.661	-0.781	-0.617	-0.057	-0.254	0.347	-0.197

Tabla 8-2. Correlaciones. 8 componentes

En la *tabla 8-2.* se tiene las correlaciones de cada variable con cada componente. Por ejemplo, *REP_Corruption* tiene una correlación de -0.4 con la componente 1, 0.405 con la componente 2, 0.226 con la componente 3, -0.039 con la componente 4, 0.242 con la componente 5, -0.393 con la componente 6 -0.147 con la componente 7 y -0.023 con la componente 8.

En verde está, para cada variable, la dimensión con la que está más correlacionada mientras que en rojo con la que menos.

Cosenos^2	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
REP_Corruption	0.16	0.164	0.051	0.001	0.059	0.154	0.022	0.001
REP_Desempleo	0.003	0.346	0.057	0.01	0.001	0.125	0.087	0.005
REP_Esperanza_vida	0.841	0.001	0.03	0.004	0.002	0.002	0.012	0.006
REP_Generosity	0.03	0.286	0.061	0.014	0	0	0.162	0.212
REP_Healthy_life_expectancy	0.852	0.004	0.022	0	0.003	0	0.015	0.01
REP_Horas_sol	0.218	0.011	0.073	0	0.098	0.165	0.002	0.001
REP_IDH	0.915	0.002	0.012	0	0	0.007	0.007	0.003
REP_Ln_PIB_per_capita_	0.845	0.001	0.018	0.004	0.001	0.013	0.008	0.002
REP_Social_support	0.665	0	0.003	0.02	0.001	0.001	0.004	0.031
IMP_REP_Freedom	0.243	0.375	0.001	0.016	0.002	0.006	0.022	0.016
IMP_REP_Prevalencia	0.479	0.167	0.002	0.026	0.015	0	0.009	0.023
0	0.286	0.142	0.152	0.017	0.082	0.064	0.002	0.137
1	0.286	0.142	0.152	0.017	0.082	0.064	0.002	0.137
Calido	0.875	0.032	0	0.001	0.026	0.001	0.005	0.001
Parte Frio	0.232	0.032	0.132	0.224	0.148	0.099	0.012	0.016
Templado	0.723	0.125	0.02	0.06	0.003	0.013	0.01	0.002
Variado	0.011	0.043	0.59	0.007	0.051	0.018	0.138	0.008
M constitucional o parlamentaria	0.413	0.1	0.003	0.004	0.228	0.071	0.003	0.002
Otro	0.106	0.099	0.225	0.159	0.084	0.094	0.017	0.109
R parlamentaria	0.428	0.131	0.052	0.098	0.09	0.002	0.095	0
R presidencialista o semip	0.609	0.001	0	0.215	0.007	0.065	0.027	0.01
Africa	0.862	0.038	0.006	0.007	0.037	0.023	0	0.001
America y ANZ	0.057	0.033	0.411	0.287	0.073	0.067	0.013	0.001
Asia	0.007	0.477	0.041	0.187	0.135	0.005	0.038	0.01
CEI	0.017	0.005	0.091	0.387	0.267	0.045	0.059	0.007
Europa	0.828	0.041	0.057	0.035	0	0.006	0.011	0.004

Tabla 8-3. Cosenos al cuadrado. 8 componentes

En la *tabla 8-3*. se tienen los cosenos al cuadrado que representan la proporción de varianza de cada variable que es explicada por cada componente. Si el valor está próximo a 1, ello significa que esa variable está bien representada en dicha dimensión mientras que si es cercana a 0 entonces es sinónimo de mala representación.

Por ejemplo, la proporción de varianza que es explicada por Europa en la componente 1 es 0.828. Además, como el valor es próximo a 1 entonces la representación en dicha dimensión es excelente. Por el contrario, en esta misma variable, su proporción es muy cercana 0 para la quinta dimensión

Contribuciones	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
REP_Corruption	2.366	7.427	3.031	0.099	4.271	13.466	2.332	0.066
REP_Desempleo	0.039	15.661	3.42	0.674	0.069	10.866	9.29	0.593
REP_Esperanza_vida	12.439	0.062	1.765	0.295	0.167	0.137	1.323	0.698
REP_Generosity	0.443	12.975	3.629	0.949	0.015	0	17.364	25.712
REP_Healthy_life_expectancy	12.6	0.18	1.328	0.029	0.208	0	1.652	1.156
REP_Horas_sol	3.225	0.497	4.357	0	7.152	14.39	0.19	0.182
REP_IDH	13.522	0.088	0.707	0	0.003	0.625	0.732	0.342
REP_Ln_PIB_per_capita_	12.489	0.046	1.065	0.271	0.071	1.123	0.907	0.238
REP_Social_support	9.834	0.022	0.16	1.353	0.057	0.112	0.392	3.785
IMP_REP_Freedom	3.586	16.968	0.069	1.067	0.14	0.54	2.405	1.936
IMP_REP_Prevalencia	7.085	7.542	0.101	1.712	1.112	0.024	1.017	2.736
0	0.127	0.591	1.092	0.155	0.884	0.981	0.043	4.096
1	0.818	3.811	7.044	1	5.703	6.326	0.275	26.419
Calido	3.568	1.222	0	0.042	2.537	0.15	1.06	0.271
Parte Frio	0.677	0.865	6.241	13.192	10.503	10.063	1.839	3.211
Templado	3.108	5.026	1.382	5.157	0.264	1.888	2.194	0.639
Variado	0.028	1.021	23.958	0.347	3.089	1.609	18.209	1.338
M constitucional o parlamentaria	1.303	2.962	0.128	0.273	17.508	7.794	0.543	0.371
Otro	0.279	2.443	9.578	8.43	5.398	8.552	2.373	19.324
R parlamentaria	1.206	3.472	2.369	5.586	6.209	0.164	14.142	0.062
R presidencialista o semip	1.44	0.014	0.01	10.239	0.401	5.329	3.303	1.58
Africa	4.64	1.941	0.559	0.772	4.814	4.395	0.112	0.483
America y ANZ	0.146	0.806	17.134	14.88	4.591	5.971	1.729	0.22
Asia	0.019	11.96	1.755	10.088	8.829	0.451	5.389	1.854
CEI	0.042	0.113	3.615	19.114	15.96	3.796	7.651	1.223
Europa	4.971	2.284	5.504	4.278	0.045	1.249	3.534	1.464

Tabla 8-4. Contribuciones. 8 dimensiones

Las contribuciones constituyen el porcentaje de varianza de la dimensión explicada por la variable. Por ejemplo, un 13.52% de la varianza de la dimensión 1 es explicada por la variable *REP_IDH.*

Las variables que más contribuyen a cada dimensión están en verde mientras que las que menos porcentaje aportan son de color rojo.

Las tablas de la *8-2.* a la *8-4.* serán utilizadas para la representación tanto de las variables cualitativas como cuantitativas, y de ambas en conjunto (Puntos 8.1, 8.2. y 8.3).

8.1. Representación de variables cuantitativas

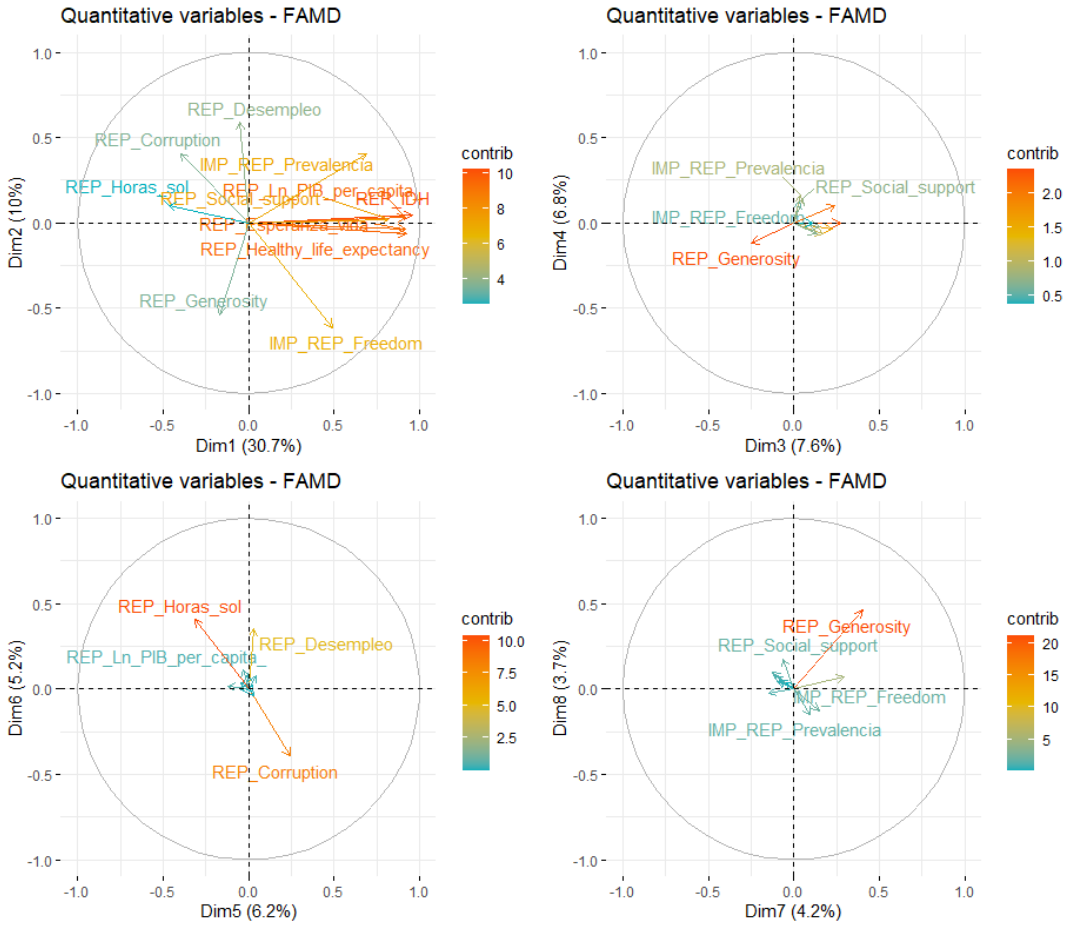


Gráfico 7. Variables cuantitativas. Dimensiones (1,2), (3,4), (5,6) y (7,8)

A la vista de los gráficos y su contribución marcada con un tono más rojo, se observa que:

- Los porcentajes que aparecen son la variabilidad explicada por cada unas de las 8 dimensiones. Por ejemplo, la primera dimensión explica un 30.7% mientras que la 7 un 4.2%.

- **Dimensión (1,2):**
  - Correlación inversa entre la percepción de la corrupción y la libertad individual, es decir, cuanta más libertad en cuestión de decisiones con sus vidas privadas menor es la corrupción percibida por los ciudadanos. Lo mismo ocurriría para el caso de *IMP_REP_Freedom* con *REP_Desempleo*. Además, países con baja prevalencia poseen valores altos en generosidad (*REP_Generosity*, *IMP_REP_Prevalencia*). Otras relaciones indirectas son horas de sol con *REP_IDH*, *REP_Ln_PIB_per_capita, REP_Esperanza_vida* y *REP_Healthy_life_expectancy*.
  - Correlación directa: *REP_Corrupcion* con *REP_Desempleo*, es decir, son los países con alta tasa de desempleo en los que se suele tener una mayor percepción de la corrupción. Además, se le podría añadir *REP_Horas_sol* lo que querría decir que suelen ser países con una luz

de sol al año mayores al resto. Estas tres últimas variables representarían a la segunda dimensión mientras que la primera estaría formada por *IMP_REP_Prevalencia, REP_Social_support, REP_IDH*, *REP_Ln_PIB_per_capita, REP_Esperanza_vida* y *REP_Healthy_life_expectancy* que son las que tienen una correlación más alta en dicha dimensión.

- **Dimensión (3,4):**
  - Correlación inversa: *REP_Generosity, REP_Desempleo*, es decir, altos valores en generosidad suponen baja tasa de desempleo.
  - Correlación directa: *REP_Esperanza_vida* con *REP_LIB_PIB_per_capita, REP_Horas_sol, REP_IDH*, y *REP_Desempleo*, serán países con características similares en cuanto a los valores de estas variables y que pertenecerán a la tercera dimensión. También, la correlación entre el desempleo y la ayuda social o prevalencia de depresión.
- **Dimensión (5,6):** en general, contribuciones bajas
  - Correlación inversa: claramente *REP_Corruption* y *REP_Horas_sol*. Luego, países con valores positivos en corrupción en la quinta dimensión serán aquellos con pocas horas de sol al año en la secta dimensión.
  - Correlación directa: *REP_Horas_sol* con *REP_Desempleo*.
- **Dimensión (7,8):**
  - Correlación inversa: el apoyo social (8ª dimensión) con la libertad individual y la tasa de prevalencia (ambas 7ª dimensión).
  - Correlación directa**:** *REP_Generosity* entre *IMP_REP_Freedom* e *IMP_REP_Prevalencia* que son aquellas variables que representan al octavo factor (sobretodo la generosidad).

8.2. Representación de variables cualitativas

A continuación, se pasa a la representación de las categorías que pertenecen a las 4 variables cualitativas.

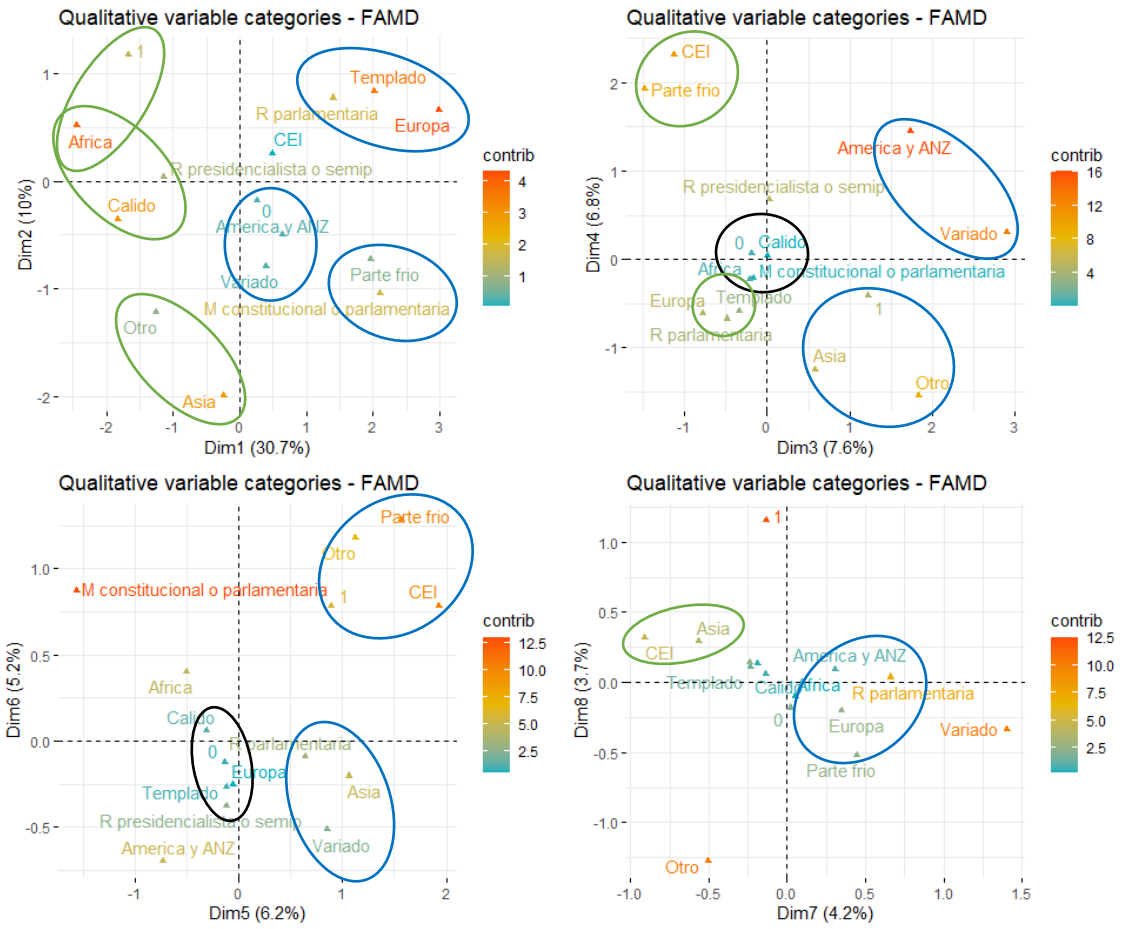


Gráfico 8. Variables cualitativas. Dimensiones (1,2), (3,4), (5,6) y (7,8)

- **Dimensión (1,2):**
  - Relación directa: por un lado, en la dimensión 1 se tiene que los países europeos (aunque puede encontrarse alguno de la CEI) están muy relacionados con clima templado y cuyo tipo de gobierno es una república parlamentaria. También los climas con parte frío y una monarquía constitucional o parlamentaria. Relación positiva además entre *America y ANZ* con *Variado* y *0* (peligroso). Por otro lado, en la dimensión 2 destacan los países africanos peligrosos y los asiáticos con *Otro* gobierno.
  - Relación inversa: casi todo lo que representa a la primera dimensión tienen relación indirecta con la dimensión 2. Por ejemplo, los países africanos peligrosos con todo lo que está en el cuarto cuadrante (abajo a la derecha) y que representa muy bien al primero.
  - En la parte de izquierda se tendrían a los países africanos o asiáticos de clima cálido, peligrosos y con otro tipo de gobierno mientras que en la derecha se observa al resto de continentes junto con los demás tipos de clima y de gobierno.
- **Dimensión (3,4):**
  - Relación directa: para la tercera dimensión claramente entre *America y ANZ* con *Variado*. Además, países asiáticos con otro tipo de gobierno y peligrosos. Para la cuarta, asociación entre Europa, climas templados y

con república parlamentaria. También, climas fríos con Estados independientes. Éste último posee altas contribuciones en ambas dimensiones, es decir, cuando en la dimensión 3 toma un valor positivo entonces será negativo en la cuarta (y viceversa).

- Relación inversa: se tienen el grupo del primer cuadrante contra el del tercero, es decir, países americanos de clima variado frente a los europeos, templados y con repúblicas parlamentaria. Por el contrario, los CEI de clima tropical o cálido con parte del territorio de clima frío, versus los asiáticos peligrosos y cuya categoría en tipo de gobierno es *Otro*.
- Las categorías dentro del óvalo negro no tienen relación con ninguna de estas dos dimensiones. Las repúblicas presidencialistas o semipresidencialistas si tiene un poco de asociación con la 4ª dimensión pero no con la tercera.

❖ **Dimensión (5,6):**

- Relación directa: como nuevo conjunto de asociaciones de países asiáticos parlamentarios con clima variado en la quinta dimensión. Para la sexta se tendrían en su mayoría países con *monarquía constitucional o parlamentaria* y otro con la región de *America y ANZ*. Éstos poseen contribuciones media altas en ambas dimensiones.
- Relación inversa: la región de America y ANZ entre los CEI con Otro gobierno, peligrosos y con parte del su clima frío. También, las monarquías de África constitucionales o parlamentarias y el conjunto de *R Parlamentaria, Asia* y *Variado*.
- No tienen relación ni directa ni inversa las categorías dentro del óvalo negro con ninguna de estas dos dimensiones.

❖ **Dimensión (7,8):**

- Relación directa**:** climas en su mayoría variado en la dimensión 7 y como conjunto el rodead en azul en el gráfico. Por el contrario, en la dimensión 8 destacan países asiáticos o de la CEI.
- Relación inversa: en este caso sería la dimensión 7 y la 8 ya que está prácticamente diferenciada entre el segundo y cuarto cuadrante.
- Altas contribuciones en ambas dimensiones de: *Peligro=1, Gobierno=Otro* y *Clima=Variado.*

8.3. Representación de variables cuantitativas y cualitativas

Toda esta información nos va a permitir representar todas estas métricas en distintos tipos de gráficos y, de esta manera, poder analizar más detalladamente cada dimensión. A la izquierda las variables (tanto cuantitativas como cualitativas) representadas en los planos factoriales en el cual el color indicará el coseno al cuadrado. Además, para las variables categóricas se está utilizando la ratio de la correlación al cuadrado entre la dimensión i y la variable cualitativa mientras que para las variables continuas se usa el coeficiente de correlación al cuadrado entre la dimensión y la variable cuantitativa.

Por el contrario, a la derecha aparecerá la contribución a la dimensión de cada variable en la dimensión correspondiente.

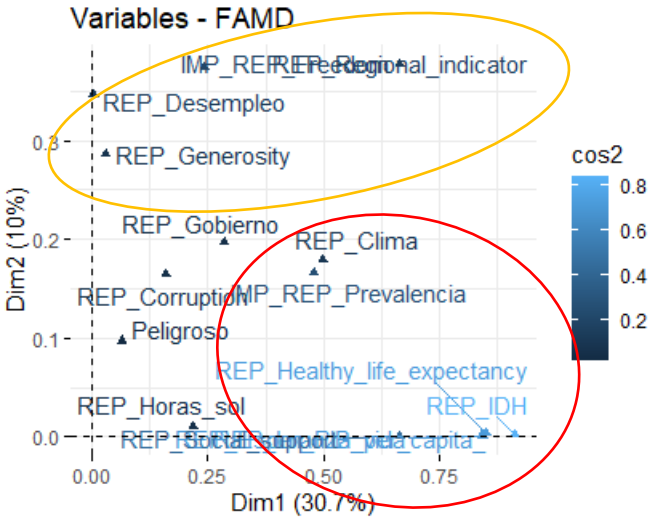


Gráfico 9. Variables cuantitativas y cualitativas. Dimensión (1,2)

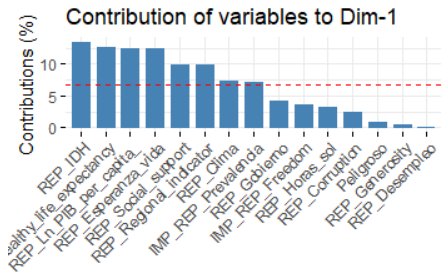


Gráfico contribuciones 1. Dimensión 1

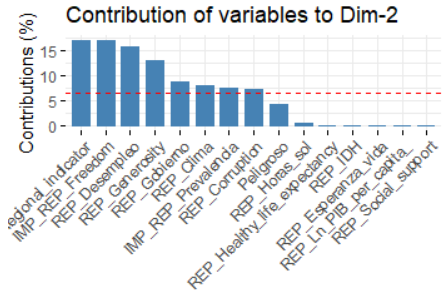


Gráfico contribuciones 2. Dimensión 2

La primera dimensión supone el 30.7% de la variabilidad de los datos mientras que la segunda un 10% como ya se vio en la tabla de los autovalores.

Los valores utilizados para las contribuciones son los de la tabla 8-4.

Las variables que se sitúan más a la derecha son las que más correlación (próxima a 1) tienen con la **dimensión 1**, es decir, *REP_Regional_Indicator, REP_IDH, REP_Healthy_life_Expectancy, REP_Ln_PIB_per_capita_ , REP_Esperanza_vida, REP_Social_Support, REP_Clima y REP_Prevalencia*. Además, todas estas (excepto *REP_Regional_indicator*) tienen poca relación con el factor 2 puesto que están cerca del eje Y=0.

Las variables que están representadas en la **dimensión 2** son: *REP_Regional_Indicator, IMP_REP_Freedom, REP_Desempleo* y *REP_Generosity.*

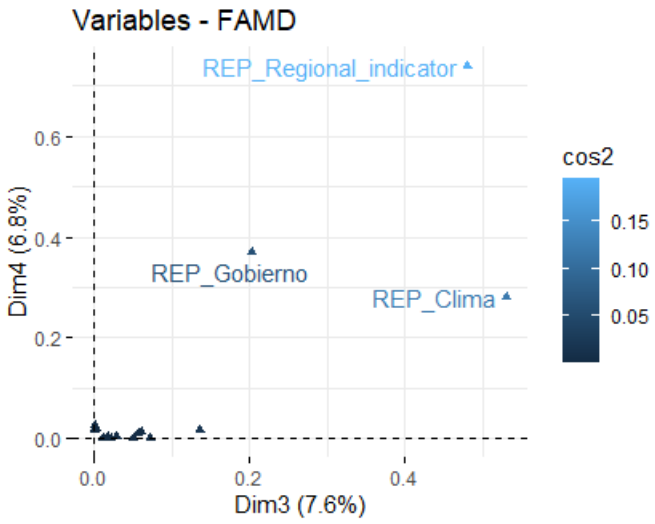


Gráfico 10. Variables cuantitativas y cualitativas. Dimensión (3,4)

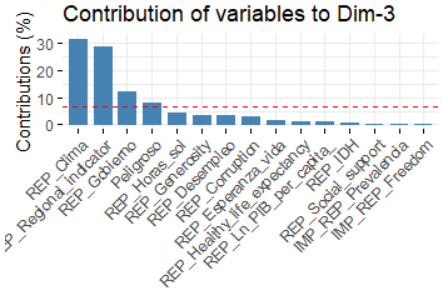


Gráfico contribuciones 3. Dimensión 3

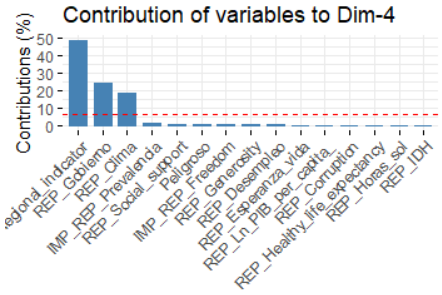


Gráfico contribuciones 4. Dimensión 4

Con respecto a las **dimensiones 3 y 4**, cuyo porcentaje de variabilidad explicada es 7.6% y 6.8%, respectivamente, se tiene claramente que para ambas son igual de importantes las mismas variables: *REP_Regional_Indicator*, *REP_Gobierno* y *REP_Clima* (más carga en la tercera) aunque esta primera es la que tiene más peso en la cuarta dimensión. Las demás variables, toman valores muy proximos a 0. Además, ninguna de las variables son cuantitativas por lo que veremos más adelante en qué dimensiones está cada categoría.

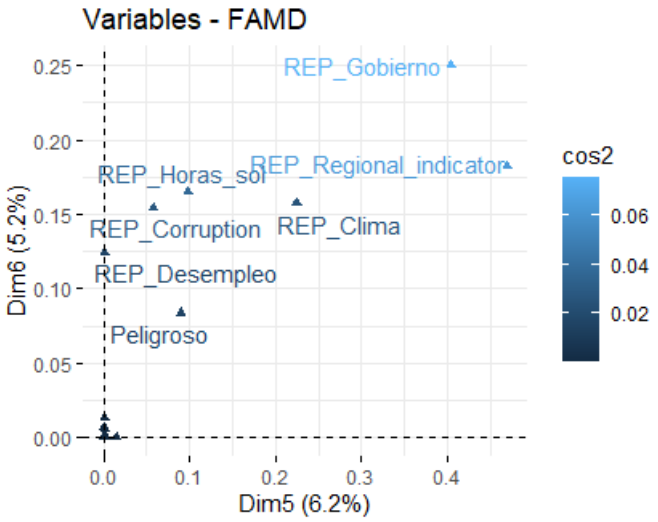


Gráfico 11. Variables cuantitativas y cualitativas. Dimensión (5,6)

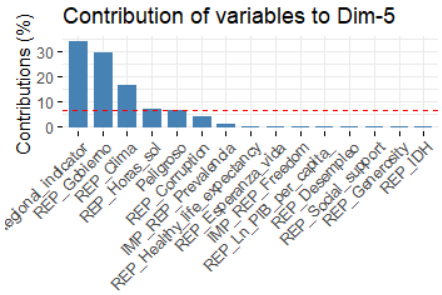


Gráfico contribuciones 5. Dimensión 5

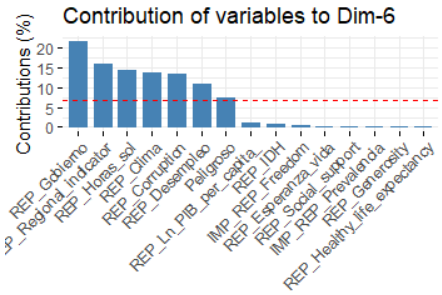


Gráfico contribuciones 6. Dimensión 6

A partir de estas dimensiones las cargas comienzan a tomar valores bajos.

La **dimensión 5** explica un 6.2% de la variabilidad y las variables con las que se encuentra más correlacionado son: *REP_Regional_indicator* y *REP_Gobierno.*

La **dimensión 6** posee un 5.2% de la variabilidad de los datos. Las mayores contribuciones: *REP_Gobierno*, *REP_Regional_Indicator, REP_horas_sol, REP_Clima, REP_Corruption, REP_Desempleo y REP_Peligroso*. Por tanto, puede ser que sean países con varias horas de luz al año, que no se consideren seguros, tengan un gobierno estricto o en el cual haya mucha corrupción y que haya una tasa de desempleo alta. Es decir, seguramente sean países en donde la escala de la felicidad no sea muy alta.

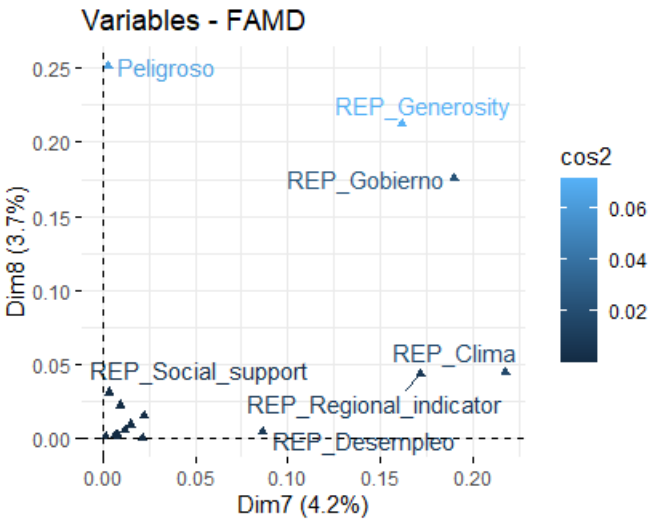


Gráfico 12. Variables cuantitativas y cualitativas. Dimensión (7,8)

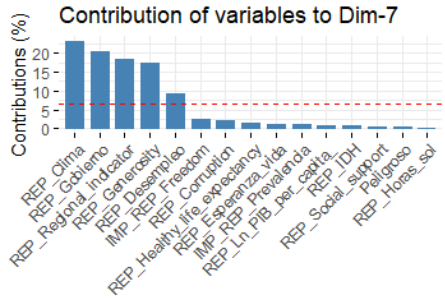


Gráfico contribuciones 7. Dimensión 7

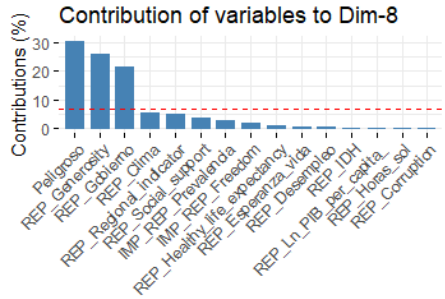


Gráfico contribuciones 8. Dimensión 8

Por último, las dimensiones 7 y 8 corresponden con el último 8% de los 74.69% de la variabilidad explicada con 8 factores.

La **dimensión 7** estaría representada en *REP_Clima, REP_gobierno, REP_Regional_indicator, REP_Generosity* y *REP_Desempleo*. Mientras, por otro lado, en la **dimensión 8** se encontrarían *Peligroso, REP_Generosity* y *REP_Gobierno*.

Destacar en este análisis que la variable *REP_Regional_Indicator* está representada en las dimensiones (excepto a la 8). Como ya se vio en el mapa, se considera una variable importante puesto que nos ayuda a predecir mejor nuestra variable objetivo.

8.4. Representación de los individuos

En R también ha sido posible representar los individuos en el plano de las componentes. Estas representaciones nos pueden ayudar a obtener información adicional sobre el comportamiento de los individuos en cada dimensión. Para ello, se puso a *Happiness* como variable suplementaria en el AFDM.

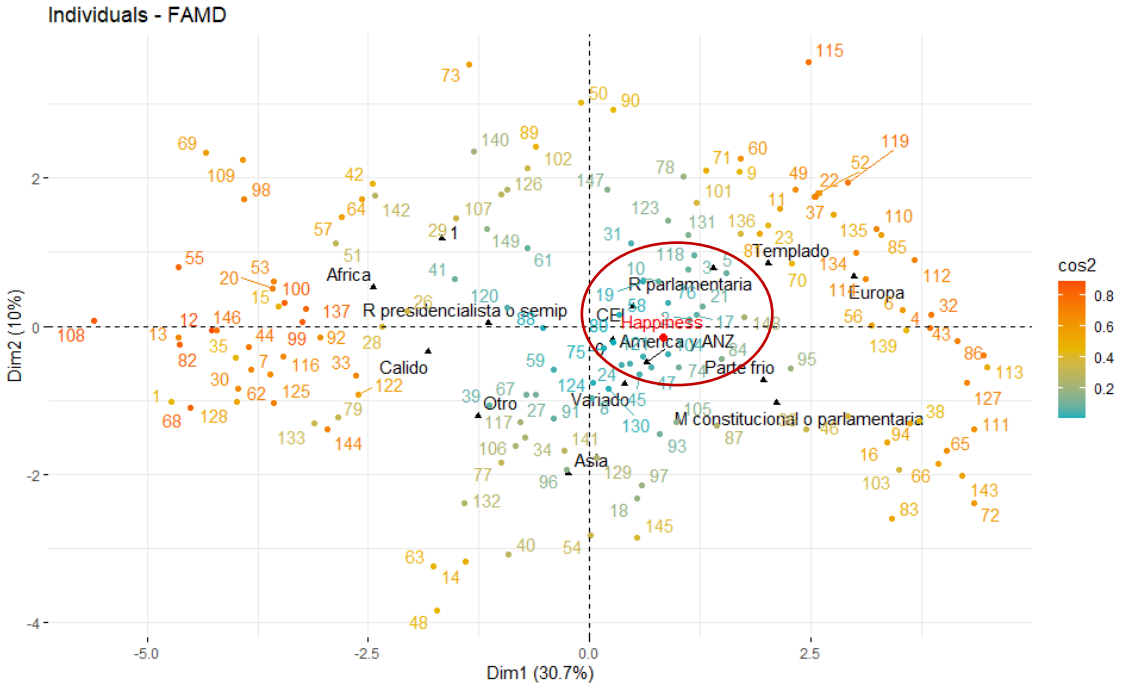


Gráfico 13. Individuos. Dimensión (1,2)

Se observa que los países más felices son prácticamente de la primera dimensión debido a que la proporción de varianza explicada en la segunda dimensión es prácticamente nula (cercana a 0). Las observaciones más próximas a *Happiness* son aquellas que están más relacionadas.

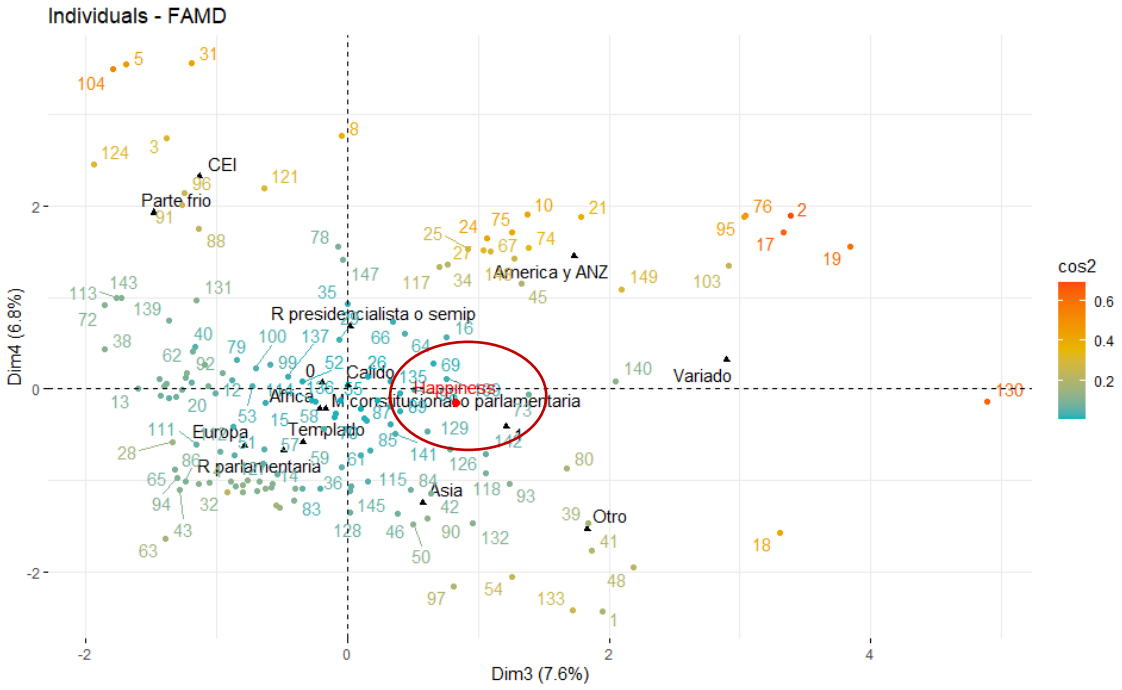


Gráfico 14. Individuos. Dimensión (3,4)

Ocurre algo parecido que el gráfico anterior, los países más relacionados con la felicidad están en la tercera dimensión y cuyo gobierno es una monarquía constitucional o parlamentaria.

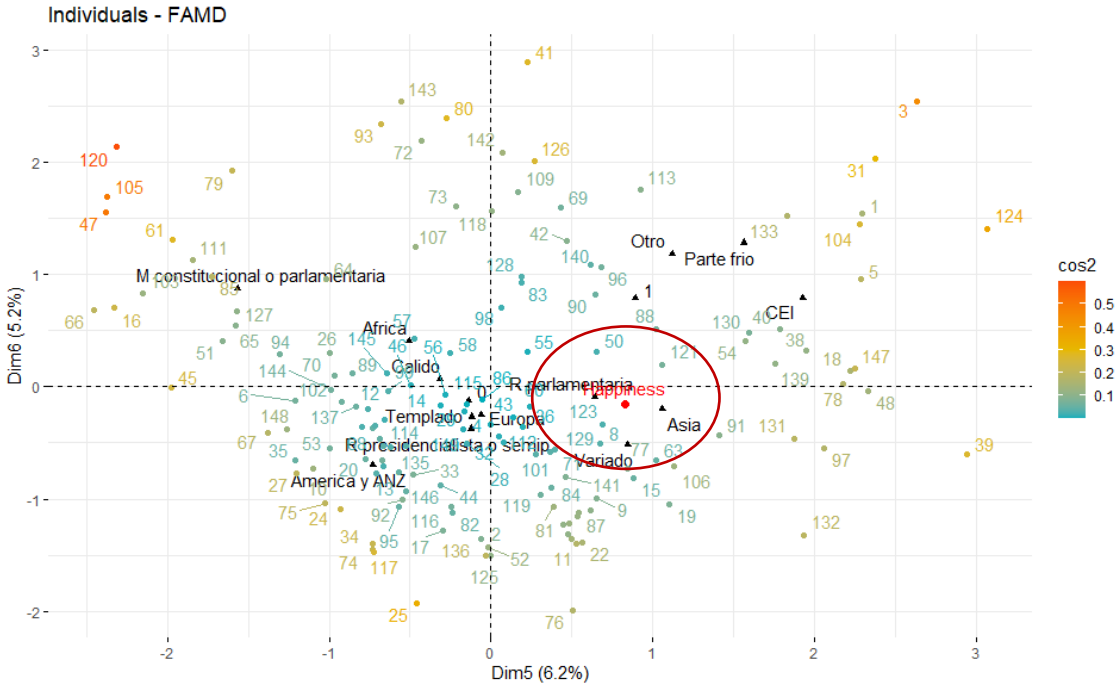


Gráfico 15. Individuos. Dimensión (5,6)

En este caso se observa que los países se encuentran más dispersos y alejados de la variable objetivo. Como “cercanos” podríamos considerar a los países de Asia, los de clima variado y los que son repúblicas parlamentarias.

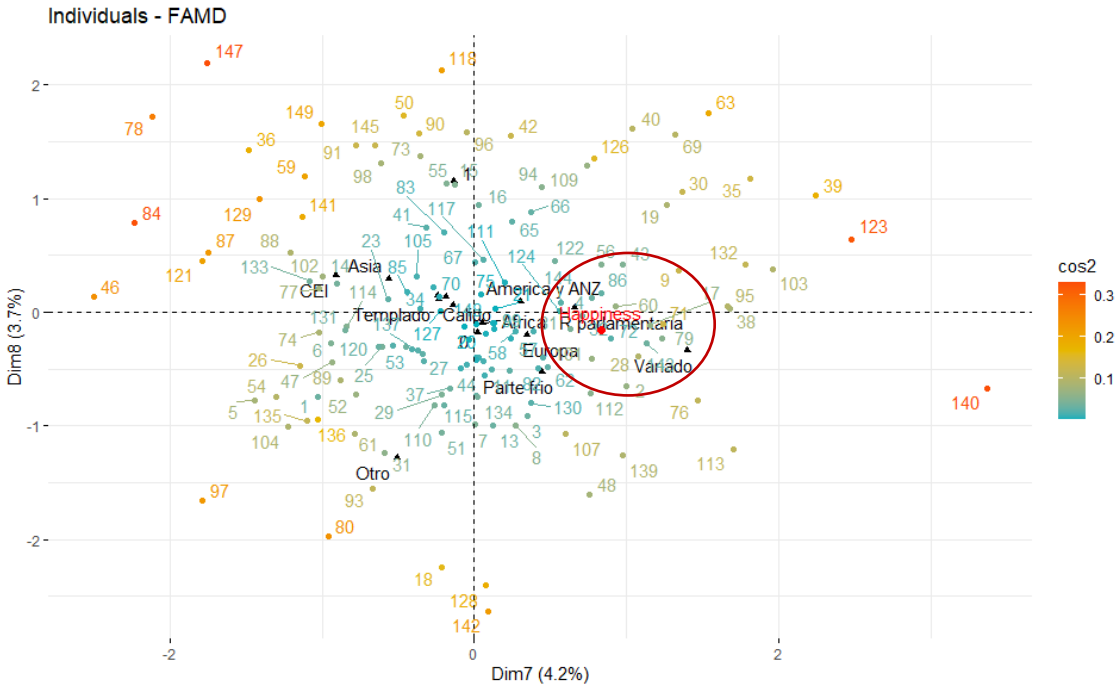


Gráfico 16. Individuos. Dimensión (7,8)

Las categorías que se concentran alrededor de *Happiness* son *America y ANZ, R parlamentaria* y *Variado.* En cuanto a países, algunos de ellos son Bosnia y Herzegovina, Suiza, Montegro o Bierorrusia, entre otros.

8.5. Autovectores

A continuación, se presentará una tabla con los coeficientes de cada variable en cada componente:

AUTOVECTORES	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
REP_Corruption	-0.154	0.273	0.174	-0.031	0.207	-0.367	-0.153	-0.026
REP_Desempleo	-0.020	0.396	0.185	0.082	0.026	0.330	0.305	0.077
REP_Esperanza_vida	0.353	-0.025	0.133	-0.054	-0.041	0.037	-0.115	0.084
REP_Generosity	-0.067	-0.360	-0.191	-0.097	-0.012	-0.002	0.417	0.507
REP_Healthy_life_expectancy	0.355	-0.043	0.115	-0.017	-0.046	-0.001	-0.129	0.108
REP_Horas_sol	-0.180	0.071	0.209	-0.002	-0.267	0.379	-0.044	0.043
REP_IDH	0.368	0.030	0.084	-0.002	-0.006	0.079	-0.086	0.059
REP_Ln_PIB_per_capita_	0.353	0.021	0.103	-0.052	-0.027	0.106	-0.095	0.049
REP_Social_support	0.314	0.015	0.040	0.116	0.024	-0.034	-0.063	0.195
IMP_REP_Freedom	0.189	-0.412	0.026	0.103	0.037	0.074	0.155	-0.139
IMP_REP_Prevalencia	0.266	0.275	0.032	0.131	-0.105	0.015	0.101	-0.165
0	0.036	-0.077	-0.105	0.039	-0.094	-0.099	0.021	-0.202
1	-0.091	0.195	0.265	-0.100	0.239	0.252	-0.052	0.514
Calido	-0.189	-0.111	0.002	0.021	-0.159	0.039	-0.103	0.052
Parte Frio	0.082	-0.093	-0.250	0.363	0.324	0.317	0.136	-0.179
Templado	0.176	0.224	-0.118	-0.227	-0.051	-0.137	-0.148	0.080
Variado	0.017	-0.101	0.489	0.059	0.176	-0.127	0.427	-0.116
M constitucional o parlamentaria	0.114	-0.172	-0.036	-0.052	-0.418	0.279	-0.074	0.061
Otro	-0.053	-0.156	0.309	-0.290	0.232	0.292	-0.154	-0.440
R parlamentaria	0.110	0.186	-0.154	-0.236	0.249	-0.040	0.376	0.025
R presidencialista o semip	-0.120	0.012	0.010	0.320	-0.063	-0.231	-0.182	0.126
Africa	-0.215	0.139	-0.075	-0.088	-0.219	0.210	0.034	-0.070
America y ANZ	0.038	-0.090	0.414	0.386	-0.214	-0.244	0.132	0.047
Asia	-0.014	-0.346	0.132	-0.318	0.297	-0.067	-0.232	0.136
CEI	0.021	0.034	-0.190	0.437	0.400	0.195	-0.277	0.111
Europa	0.223	0.151	-0.235	-0.207	-0.021	-0.112	0.188	-0.121

Tabla 8-5. Autovectores 8 componentes

Los autovectores asociados a los autovalores nos dan los coeficientes. Éstos son combinación lineal de las variables originales con la que se construyen los componentes (tabla 8-5.).

9. APRENDIZAJE SUPERVISADO – Regresión y clasificación

Como técnicas en esta rama de Machine Learning (ML) serán llevadas a cabo la regresión lineal y la logística ya que en el dataset se tiene la variable objetivo como continua y categórica. Además, los factores obtenidos en el AFDM se emplearán para intentar predecir la felicidad.

9.1. Regresión lineal

9.1.1. Selección de variables

Se ha llevado a cabo una selección previa de las variables con el **PROC GLMSELECT** mediante el método de Stepwise y cuyo criterio de parada fue AIC y BIC.

Por un lado, el set de variables que se obtuvo fue en AIC:

`Intercept REP_Regional_indicator REP_Corruption REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia`

Por otro lado, en el caso de BIC salió el mismo conjunto pero sin la variable `IMP_REP_Prevalencia`.

NOTA: La creación de dummies en las variables de las variables categóricas no fue realizada porque en la depuración se hizo la recategorización uniendo aquellas categorías que no eran muy representativas. Además, en el caso del software SAS, esto lo hace internamente.

A continuación, se prueba a hacer el método stepwise en submuestras utilizando la macro ***%randomselect*** que fue aportada en clase por el profesor Javier Portela fijando como semilla de inicio 2000 y como semilla final 2200. Lo que va a hacer esta macro es, con la semilla inicial puesta, sorteará el 80% de los datos, hará selección de variables (mediante AIC o BIC) y creará un modelo.

En la salida, el ranking de los modelos con una frecuencia mayor o igual a 6 ha sido:

Modelo AIC	Freq	%	Nº
Intercept REP_Regional_indicator REP_Corruption REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom	14	6.965	7
Intercept REP_Regional_indicator REP_Corruption REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia	9	3.478	8
Intercept REP_Regional_indicator REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom	8	3.98	6
Intercept REP_Regional_indicator REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia	8	3.98	7
Intercept REP_Regional_indicator REP_Corruption REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia	7	3.483	7
Modelo BIC			
Intercept REP_Regional_indicator REP_Corruption REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom	27	13.433	6
Intercept REP_Regional_indicator REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom	22	10.945	6
Intercept REP_Regional_indicator REP_Desempleo REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia	15	7.463	6
Intercept REP_Regional_indicator REP_Generosity REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom	14	6.965	6
Intercept REP_Regional_indicator REP_Corruption REP_Horas_sol REP_Ln_PIB_per_capita_ REP_Social_support IMP_REP_Freedom	12	5.97	6

Tabla 9-1. Ranking mejores modelos con AIC y BIC

Al parecer, el modelo AIC es menos restrictivo en la en la selección y, por eso, posee conjuntos con mayor número de variables que BIC, cuyos sets elegidos más frecuentes poseen únicamente 6 variables. Ambos coinciden en el que está en negrita siendo en BIC el segundo que más se ha repetido. Además, se observa que la frecuencia de los modelos en AIC no ha sido tan grande como ha ocurrido en BIC, es decir, hay discrepancias. Por ejemplo, el primero en AIC ha aparecido casi el 7% de las 201 semillas mientras que el primero del BIC 13.4%. No obstante, como método de selección de variables, se probarán todos estos.

Con respecto al efecto que ha tenido cada variable se muestra a la izquierda la salida con AIC y a la derecha con BIC:

Obs	efecto	COUNT	PERCENT
1	IMP_REP_Freedom	201	13.1803
2	REP_Regional_indicator	201	13.1803
3	REP_Social_support	198	12.9836
4	REP_Horas_sol	187	12.2623
5	REP_IDH	124	8.1311
6	REP_Desempleo	108	7.0820
7	REP_Corruption	101	6.6230
8	REP_Ln_PIB_per_capita_	87	5.7049
9	IMP_REP_Prevalencia	82	5.3770
10	REP_Generosity	75	4.9180
11	REP_Esperanza_vida	64	4.1967
12	Peligroso	51	3.3443
13	REP_Gobierno	34	2.2295
14	REP_Healthy_life_expectancy	12	0.7869

Tabla 9-2. Efecto variables AIC

Obs	efecto	COUNT	PERCENT
1	IMP_REP_Freedom	201	16.2885
2	REP_Regional_indicator	198	16.0454
3	REP_Social_support	196	15.8833
4	REP_Horas_sol	153	12.3987
5	REP_IDH	151	12.2366
6	REP_Desempleo	89	7.2123
7	REP_Corruption	79	6.4019
8	REP_Ln_PIB_per_capita_	48	3.8898
9	IMP_REP_Prevalencia	46	3.7277
10	REP_Generosity	37	2.9984
11	Peligroso	18	1.4587
12	REP_Esperanza_vida	11	0.8914
13	REP_Gobierno	5	0.4052
14	REP_Healthy_life_expectancy	2	0.1621

Tabla 9-3. Efecto variables BIC

Estos han sido prácticamente iguales con ambos criterios a excepción de *Peligroso* y *REP_Esperanza_vida* que se intercambian posiciones. Como se puede apreciar, en AIC la aparición de las variables es mayor que con respecto a BIC que es más selectivo.

Ambos coinciden en que las 3 variables más frecuentes son: *IMP_REP_Freedom*, *IMP_REP_Regional_Indicator* y *REP_Social_Support* por lo que se puede añadir de prueba como décimo modelo.

Otra propuesta sería escoger las variables que han aparecido más de la mitad de las veces. En AIC las 7 primeras mientras que en BIC las 5 primeras (undécimo). En este caso, las variables del primer criterio son coincidentes con la más frecuente de la tabla anterior por lo que se podría intentar con las 4 más frecuentes (duodécimo).

Por último, se propone hacer un decimotercer modelo con las 5 variables más valiosas a partir del gráfico de importancia del MINER (apartado Medición de relación entre variables).

9.1.2. RL con variables seleccionadas

Para que pueda ser seleccionado el mejor modelo de regresión lineal (RL) con nuestra variable objetivo *Happiness* se utilizará la macro `%cruzada` que consiste en realizar VCR fijada a 5 grupos y para 201 semillas. Se realizará para los 12 modelos mencionados anteriormente. Por último, para poder hacer la comparación en sesgo y varianza, se saca un gráfico de cajas y bigotes.

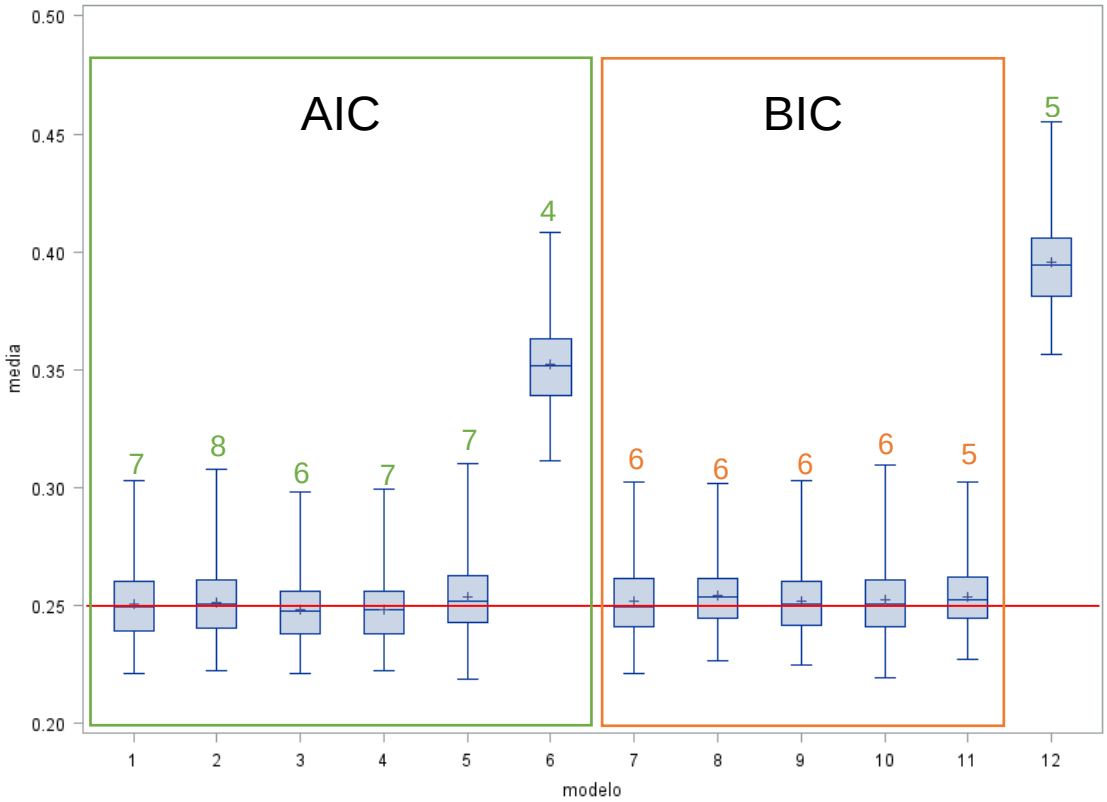


Gráfico VCR 1. Candidatos

Los modelos 6 y 12 son con diferencia los que poseen un sesgo medio más alto. Además, destacar que los demás en variabilidad y error medio son prácticamente iguales. Por tanto, para tener una mejor visibilidad, quitamos éstos y,

- El 8 porque es el que más variables posee y hay otros sets que son mejores en sesgo y varianza con menos variables.
- El 5 porque tiene la varianza más alta de los de 7 variables.
- El 7 y 10 porque también tienen varianzas altas.

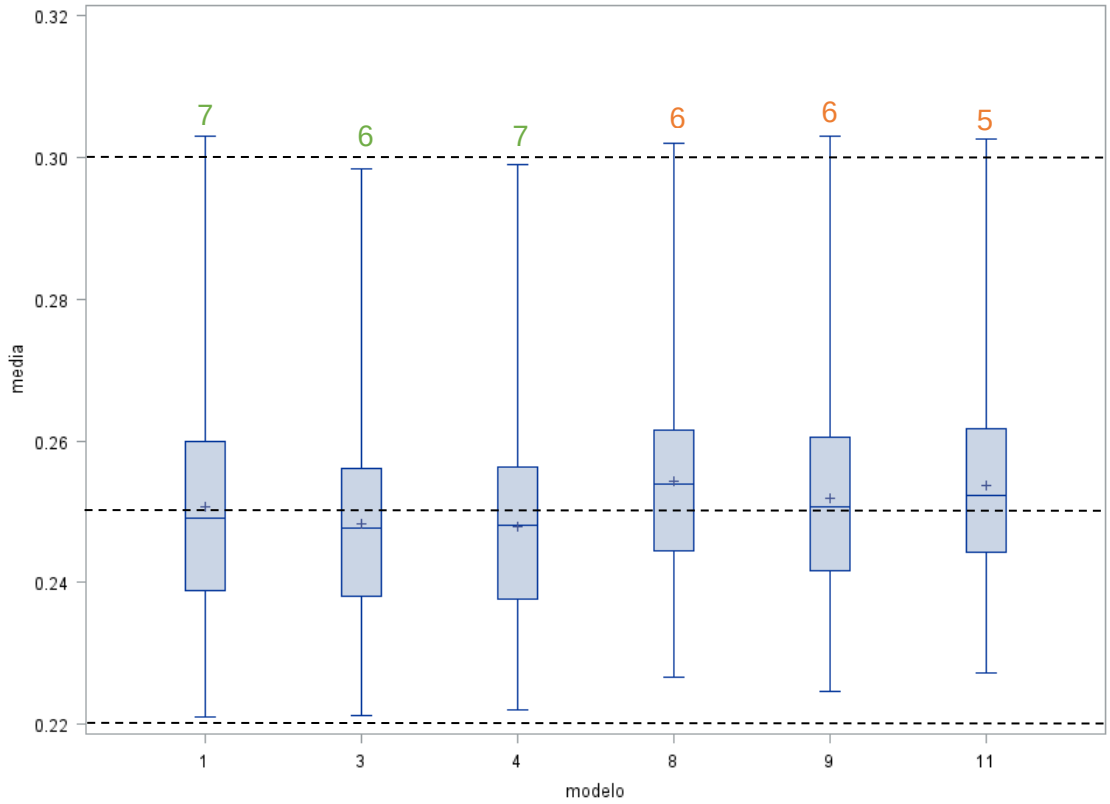


Gráfico VCR 2. Mejores candidatos

	Mean	Median	Variance	Range	Minimum	Maximum
Global	**0.25117**	**0.25013**	**0.00020**	**0.08210**	**0.22098**	**0.30308**
1	0.25071	0.24919	0.00025	0.08210	0.22098	0.30308
3	0.24835	0.24764	0.00018	0.07711	0.22123	0.29833
4	0.24791	0.24815	0.00018	0.07698	0.22211	0.29910
8	0.25433	0.25388	0.00017	0.07537	0.22664	0.30201
9	0.25195	0.25072	0.00021	0.07833	0.22470	0.30303
11	0.25375	0.25237	0.00019	0.07540	0.22720	0.30260

Tabla 9-4. Estadísticos descriptivos mejores candidatos

Se han puesto unas líneas intermitentes los valores 0.22, 0.25 (media) y 0.3 además de sus estadísticos descriptivos en una tabla para poder tener más precisión a la hora de seleccionar un modelo. Luego, sin duda el primero, con 7 variables es el peor de todos estos debido a que posee alta varianza y alto rango. También, a pesar de que su varianza es la más baja, el número 8 de 6 variables es el pero en cuanto a media y mediana.

Para poder tomar un mejor decisión, ya que los que quedan están por debajo del global, vamos a sacar el R^2 de cada uno, es decir, qué porcentaje de la variabilidad de los datos explica el modelo. En la salida del **PROC GLM** se han tenido 0.826682, 0.832048 y 0.818445, respectivamente. A pesar de que el R^2 máximo es el del modelo 4, nos vamos a quedar con el 3 puesto que sus diferencias son ínfimas y, además, es el que tiene la mediana más pequeña de los 12 modelos.

No obstante, nótese que la elección de cualquiera de estos modelos habría sido válida puesto que las discrepancias entre éstos son de milésimas y, también, dependerá si se quieren más o menos variables. En este caso, 6 variables era lo más frecuente.

Por tanto, **el mejor modelo de regresión lineal** que se ha encontrado **para predecir la escala de la felicidad con las variables seleccionadas (mediante criterio AIC) ha sido el 3** compuesto de 6 variables y con $R^2 = 0.832$.

9.1.3. Análisis del modelo ganador

Para sacar los coeficientes estimados se utilizará el **PROC GENMOD** el cual nos va a permitir incluir variables categóricas.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	90	19.6347	0.2182
Scaled Deviance	90	100.0000	1.1111
Pearson Chi-Square	90	19.6347	0.2182
Scaled Pearson X2	90	100.0000	1.1111
Log Likelihood		-60.5003	
Full Log Likelihood		-60.5003	
AIC (smaller is better)		143.0007	
AICC (smaller is better)		146.0007	
BIC (smaller is better)		171.6576	

Tabla 9-5. Criterios para evaluar la bondad de ajuste

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2306	0.6250	-1.4555	0.9944	0.14	0.7122
REP_Regional_indicat	Africa	1	0.0101	0.1797	-0.3421	0.3622	0.00	0.9554
REP_Regional_indicat	America y ANZ	1	0.1694	0.1511	-0.1268	0.4656	1.26	0.2624
REP_Regional_indicat	Asia	1	-0.6111	0.1634	-0.9313	-0.2909	13.99	0.0002
REP_Regional_indicat	CEI	1	-0.4106	0.2005	-0.8036	-0.0177	4.20	0.0405
REP_Regional_indicat	Europa	0	0.0000	0.0000	0.0000	0.0000	.	.
REP_Desempleo		1	-0.0177	0.0081	-0.0336	-0.0018	4.75	0.0293
REP_Horas_sol		1	-0.0002	0.0001	-0.0004	-0.0000	4.69	0.0303
REP_IDH		1	3.4114	0.5003	2.4309	4.3919	46.50	<.0001
REP_Social_support		1	2.1089	0.6141	0.9054	3.3125	11.79	0.0006
IMP_REP_Freedom		1	2.8053	0.5029	1.8196	3.7911	31.11	<.0001

Tabla 9-6. Parámetros estimados por máxima verosimilitud

El modelo posee 6 variables y 10 parámetros.

Se observa que la constante no es significativa pero es importante incluirla en el modelo debido a las dificultades que supondría en el cálculo de la suma de cuadrados para un modelo sin constante. También, aunque las regiones de *África*, y *America y ANZ* son significativamente distintas de 0, al tratarse de categorías de una variable no pueden ser eliminadas.

Además, que los países procedan de Asia y de los CEI hace que Y disminuya. Asimismo, cuando crece la tasa de desempleo y las horas de sol.

Las demás variables tienen un efecto positivo sobre la escala de la felicidad.

El IDH, el apoyo social y la libertad individual son las variables con más peso en *Happiness*.

Interpretación coeficientes:

- *REP_Regional_Indicator = Africa*: si el indicador regional es África entonces la escala de la felicidad aumenta un 0.01 frente a cuando es Europa.
- *REP_Regional_indicator = America y ANZ*: la felicidad se incrementa 0.17 cuando el indicador regional es *America y ANZ* frente a cuando es Europa.
- *REP_Regional_Indicator = Asia*: cuando la región es Asia, la felicidad disminuye 0.6 en comparación a cuando es Europa.
- *REP_Regional_Indicator = CEI*: si el indicador regional es CEI entonces la escala de la felicidad decrece 0.41 con respecto a cuando la región es Europa.
- *REP_Desempleo*: cuando la tasa de desempleo en ese país aumenta un 1% entonces la felicidad disminuye en 0.017 unidades.
- *REP_Horas_sol*: si las horas medias de sol al año crecen una hora, la escala de la felicidad decrece 0.0002 unidades.
- *REP_IDH*: la felicidad acrecienta 3.41 cuando el índice de desarrollo humano aumenta un 1%.
- *REP_Social_Support*: cuando el apoyo social incrementa una unidad, la felicidad lo hace en 2.11 unidades.
- *IMP_REP_Freedom*: cuando la media del valor de la libertad individual en el país aumenta en una unidad entonces la felicidad crece 2.81 unidades.

9.2. Regresión logística

En este apartado realizaremos una regresión logística (RLog) convirtiendo la variable *Happiness* en binaria (*Happiness_bin*) siendo 1 cuando el país es feliz y 0 cuando no lo es. El punto de corte establecido fue en el valor 5.

9.2.1. Selección de variables

Para la selección inicial de variables se ha utilizado el **`PROC LOGISTIC`**. Éstas fueron según el método:

- Stepwise: `REP_IDH IMP_REP_Freedom`
- Forward: `REP_Desempleo REP_IDH IMP_REP_Freedom`
- Backward: `REP_Ln_PIB_per_capi IMP_REP_Freedom`

A continuación, se prueba a hacer el método stepwise en submuestras utilizando la macro ***`%randomselectlog`*** que fue aportada en clase por el profesor Javier Portela fijando como semilla de inicio 2000 y como semilla final 2200. Lo que va a hacer esta macro es, con la semilla inicial puesta, sorteará el 80% de los datos, hará selección de variables cuyo criterio es AIC y creará un modelo.

En la salida, el ranking de los modelos con una frecuencia mayor o igual a 25 ha sido:

Modelo AIC	Freq	%	Nº
REP_IDH IMP_REP_Freedom	44	22.222	2
REP_Ln_PIB_per_capi IMP_REP_Freedom	36	18.182	2
IMP_REP_Freedom	27	13.636	1
REP_Social_support IMP_REP_Freedom	27	13.636	2

Tabla 9-7. Ranking mejores modelos con AIC

Obs	efecto	COUNT	PERCENT
1	IMP_REP_Freedom	154	40.9574
2	REP_IDH	79	21.0106
3	REP_Ln_PIB_per_capi	49	13.0319
4	REP_Horas_sol	32	8.5106
5	REP_Social_support	31	8.2447
6	REP_Desempleo	13	3.4574
7	REP_Healthy_life_ex	9	2.3936
8	REP_Generosity	8	2.1277
9	REP_Esperanza_vida	1	0.2660

Tabla 9-8. Efecto variables AIC

Con variable objetivo binaria se tuvo mucho más claro cuáles son los modelos predominantes debido a que los demás modelos salieron menores del 5% (frecuencia<9). El set de variables del stepwise y del forward fueron el primero y segundo más frecuentes, respectivamente.

Con respecto al efecto de las variables, no hay duda alguna de que *IMP_REP_Freedom* es muy importante para predecir Y. Destacar, además, que ninguna variable categórica tiene efecto importante sobre la felicidad en caso binario, no como ocurría en el apartado anterior. Es decir, en un 13.03% de las 201 veces basta con saber el promedio de la libertad no es necesaria la región para saber simplemente si un país es feliz o no.

La lista de **modelos propuestos**:

1) 1º más frecuente: `REP_IDH IMP_REP_Freedom`
2) 2º más frecuente: `REP_Ln_PIB_per_capi IMP_REP_Freedom`
3) 3º más frecuente: `IMP_REP_Freedom`
4) 4º más frecuente: `REP_Social_support IMP_REP_Freedom`
5) Set de variables forward: `REP_Desempleo REP_IDH IMP_REP_Freedom`
6) 3 variables más frecuentes: `IMP_REP_Freedom REP_IDH REP_Ln_PIB_per_capi`
7) 5 variables más frecuentes: `IMP_REP_Freedom REP_IDH REP_Ln_PIB_per_capi REP_Horas_sol REP_Social_support`

9.2.2. RLog Variables seleccionadas

Como ocurría en la regresión lineal, para que pueda ser elegido el mejor modelo para predecir la variable objetivo *Happiness_bin* se utilizará la macro `%cruzadalogistica` y cuyas características van a ser las mismas que en el caso continuo (5 grupos y 201 semillas). Se realizará para los 7 modelos mencionados anteriormente. Para poder hacer la comparación en sesgo y varianza, se saca un gráfico de cajas y bigotes.

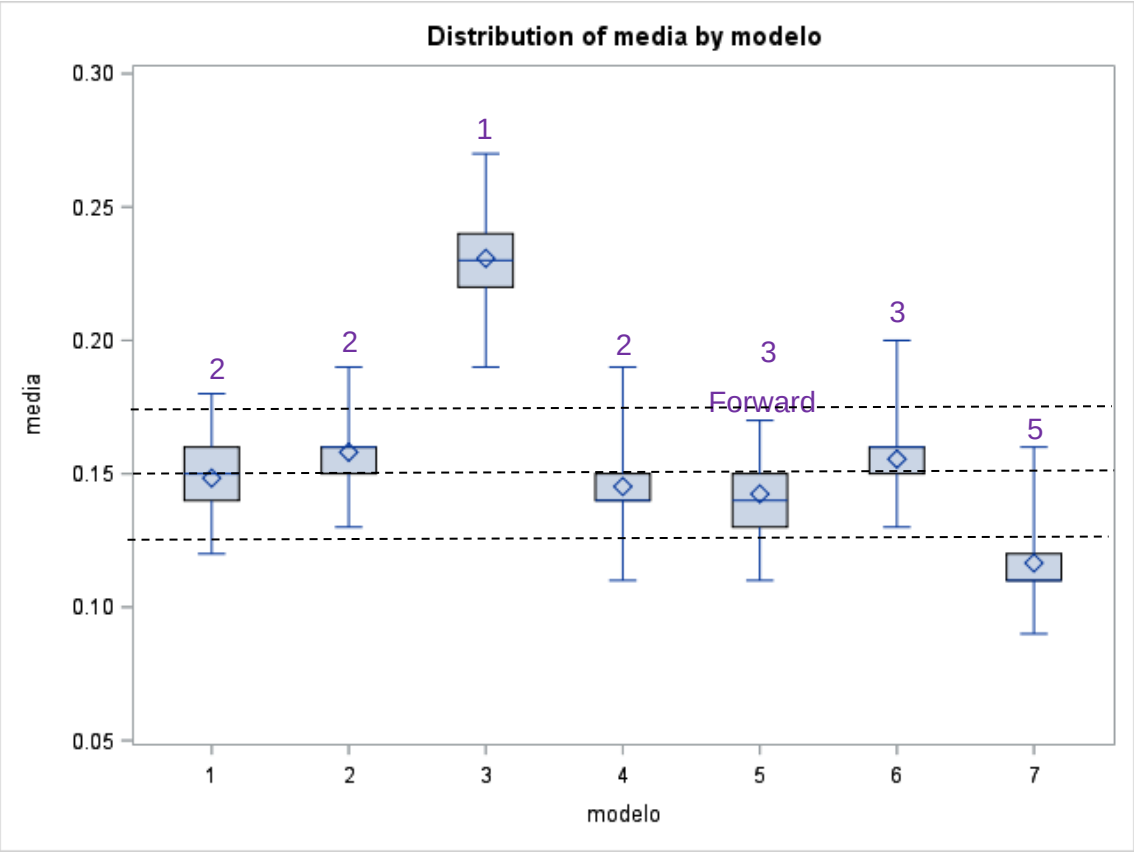


Gráfico VCR 3. Candidatos

	Mean	Median	Variance	Range	Minimum	Maximum
Global	**0.1567**	**0.15**	**0.00125**	**0.18**	**0.09**	**0.27**
1	0.1484	0.15	0.00014	0.06	0.12	0.18
3	0.1581	0.16	0.00014	0.06	0.13	0.19
4	0.1424	0.14	0.00016	0.06	0.11	0.17
8	0.1165	0.11	0.00019	0.07	0.09	0.16

Gráfico VCR 4. Estadísticos descriptivos mejores candidatos

Es más que evidente que con una sola variable se tiene un sesgo más alto en comparación con las demás en el modelo 3. Sin embargo, hay que destacar qu7e la diferencia en media con los demás no es demasiado y no se estaría cometiendo mucho error si se diera el caso.

El caso del 4 es el que posee un rango mayor frente al resto mientras que el 5 es el que menor rango tiene.

Para mayor precisión se ha sacado la tabla … sin el 3 ni el 4 ni el 6 donde se observan peores métricas. El "*Global*" hace referencia a los 7 modelos por lo que estos 4 seleccionados se encuentran por debajo de la media. En rojo está el peor valor de entre estos 4 conjuntos de variables mientras que en verde el mejor.

El set de 5 variables tanto en mediana como en media se sitúa por debajo de los demás y, las discrepancias en rango y rango intercuartílico con 2 o 3 variables no es mucha. Son bastante similares. Además, de la tabla vemos que posee las mejores métricas en

media, mediana, mínimo y máximo. A pesar de que no ocurre lo mismo en varianza y rango, aún así, se encuentran por debajo del global.

Por tanto, **el mejor modelo de regresión logística** que se ha encontrado **para predecir la felicidad o no felicidad de un país ha sido con set 7** compuesto de las 5 variables más frecuentes.

9.2.3. Análisis del modelo ganador

En este caso, se volverá a utilizar el **`PROC LOGISTIC`**.

Response Profile		
Ordered Value	happiness_bin	Total Frequency
1	0	32
2	1	68

Tabla de frecuencias 6. Happiness_bin

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	127.374	67.720
SC	129.979	83.351
-2 Log L	125.374	55.720

Tabla 9-9. Estadísticos de ajuste

En el fichero train se tienen 68 países que son felices (1) y el 32 restante no son felices (0). Y el AIC conseguido fue 127.374

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
IMP_REP_Freedom	<0.001	<0.001	0.017
REP_IDH	0.200	<0.001	>999.999
REP_Ln_PIB_per_capit	0.505	0.050	5.108
REP_Horas_sol	1.001	1.000	1.002
REP_Social_support	<0.001	<0.001	1.720

Tabla 9-10. Estimadores odds ratio

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	18.7348	7.1688	6.8297	0.0090
IMP_REP_Freedom	1	-11.2694	3.6684	9.4375	0.0021
REP_IDH	1	-1.6076	8.8424	0.0331	0.8557
REP_Ln_PIB_per_capit	1	-0.6832	1.1807	0.3349	0.5628
REP_Horas_sol	1	0.00122	0.000649	3.5248	0.0605
REP_Social_support	1	-8.3181	4.5206	3.3857	0.0658

Tabla 9-11. Parámetros estimados por máxima verosimilitud

P(Happiness_bin = 1) = $\frac{1}{1+e^{-z}}$***, donde*** $Z = 18.735 - 11.269 \cdot IMP\_REP\_Freedom - 1.608 \cdot REP\_IDH - 0.683 \cdot REP\_Ln\_PIB\_per\_capita\_ + 0.001 \cdot REP\_Horas\_sol - 8.318 \cdot REP\_Social\_Support$

Notar que hay variables que no son significativas. El mejor modelo se ha elegido con el menor ASE y este criterio no tiene en cuenta si las variables son significativas o no. Y es bastante común que ocurra este tipo de situaciones.

Interpretación coeficientes:

- *IMP_IDH:* es 5 veces menos probable que un país no sea feliz cuando el Índice de desarrollo humano aumenta un 1%.
- *REP_Ln_PIB_per_capit:* cuando el logaritmo neperiano del país aumenta un 1% es la mitad de probable que el país no sea feliz.
- *REP_Horas_sol:* por cada hora de sol media al año las posibilidades de que ese país sea feliz aumentan un 0.05%.
- Todas las variables excepto las horas de sol poseen asociación negativa (odds ratio<1).

9.3. KNN

El modelo de vecino más próximo (KNN) es un método de clasificación supervisada cuya idea principal estriba en que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus k vecinos (observaciones) más próximos. De esta forma, se obtiene el valor de predicción como la media o la moda de los valores de la variable objetivo según sea esta de intervalo o de clase. En este trabajo se utilizará KNN no sólo para conseguir una mejor predicción si no para ver cuáles son los vecinos más cercanos que escoge este algoritmo para predecir la felicidad y, de esta manera, sea más comprensible ver qué países son similares.

Previamente, se realizará una partición de los datos de entrenamiento (70-30) para obtener el de Validación. Esto es debido a que en KNN lo habitual es elegir el modelo que menor error produzca en el conjunto de datos de validación pues hacerlo sobre Train o Test conllevan problemas.

Además, para que todas las variables input tengan el mismo peso a la hora de medir su distancia, previamente hay que transformarlas: las de intervalo deben ser estandarizadas (mediante tipificación o rango) y las de clase convertirlas en dummies. Por ello, para la elección del mejor modelo KNN serán llevado a cabo distintos caminos probando diferentes tipos de transformaciones en la base de datos además de vecinos.

9.3.1. Y continua

- **Camino 1**: variables sin seleccionar (SS), tipificación, K=5 (1.1), 7 (1.2), 9 (1.3) y 11 (1.4).
- **Camino 2**: variables sin seleccionar, rango, K=5 (2.1), 7 (2.2), 9 (2.3) y 11 (2.4).
- **Camino 3**: variables seleccionadas MINER (nodo: *Selección De Variables*), tipificación, K=5 (3.1), 7 (3.2), 9 (3.3) y 11 (3.4).
- **Camino 4**: variables seleccionadas MINER, rango, K=5 (4.1), 7 (4.2), 9 (4.3) y 11 (4.4).
- **Camino 5**: variables seleccionadas RL, tipificación, K=5 (5.1), 7 (5.2), 9 (5.3) y 11 (5.4).
- **Camino 6**: variables seleccionadas RL, rango, K=5 (6.1), 7 (6.2), 9 (6.3) y 11 (6.4).

El criterio de selección será el que tenga un menor valor en el ASE (Average Square Error). Sin embargo, para poder comparar no se puede hacer con los datos del fichero de entrenamiento y se utilizó el VASE (Valid ASE). En los **resultados** se obtuvo:

- Entre los caminos sin selección (1 y 2) previa de variables, el mejor método es el del rango siendo el de 5 vecinos el que posee el VASE más bajo (0.2855). Del mismo camino, con 7 o 9 vecinos también se ha obtenido un VASE<0.3. Se

observa que a medida que aumentamos K, el VASE es mayor. Ocurre lo contrario con el método de tipificación.

Estadísticos de ajuste

Modelo seleccionado	Nodo predecesor	Descripción del modelo	Criterio de selección: Valid: Average Squared Error
Y	MBR5	K=5, ss, rango	0.285547
	MBR6	K=7, ss, rango	0.292459
	MBR7	K=9, ss, rango	0.299551
	MBR8	K=11, ss, rango	0.301355
	MBR4	K=11, ss, std	0.336882
	MBR2	K=7, ss, std	0.338154
	MBR3	K=9, ss, std	0.343191
	MBR	K=5, ss, std	0.349228

Tabla 9-12. Comparación VASE camino 1 y 2

- Entre los caminos con selección realizada con MINER (3 y 4), por un lado, las variables cualitativas se han recategorizado de la siguiente forma:

Variables de grupo

Nombre ▲	Grupo	Variable	Nivel
G_REP_Clima	0	REP_Clima	CALIDO
G_REP_Clima	1	REP_Clima	PARTE FRIO
G_REP_Clima	1	REP_Clima	TEMPLADO
G_REP_Clima	1	REP_Clima	VARIADO
G_REP_Gobierno	1	REP_Gobierno	M CONSTITUCIONAL O PARLAMENTA...
G_REP_Gobierno	0	REP_Gobierno	OTRO
G_REP_Gobierno	1	REP_Gobierno	R PARLAMENTARIA
G_REP_Gobierno	0	REP_Gobierno	R PRESIDENCIALISTA O SEMIP
G_REP_Regional_indicator	0	REP_Regional_indicator	AFRICA
G_REP_Regional_indicator	2	REP_Regional_indicator	AMERICA Y ANZ
G_REP_Regional_indicator	1	REP_Regional_indicator	ASIA
G_REP_Regional_indicator	1	REP_Regional_indicator	CEI
G_REP_Regional_indicator	3	REP_Regional_indicator	EUROPA

Tabla 9-13. Recategorización variables camino 3 y 4

Y fueron rechazadas por tener un R^2 pequeño: *IMP_REP_Prevalencia*, *REP_Generosity* y *REP_Healthy_life_Expectancy*.

Estadísticos de ajuste

Modelo seleccionado	Nodo predecesor	Descripción del modelo	Criterio de selección: Valid: Average Squared Error
Y	MBR16	K=11, cs MINER, rango	0.256529
	MBR15	K=9, cs MINER, rango	0.261127
	MBR14	K=7, cs MINER, rango	0.287191
	MBR11	K=9, cs MINER, std	0.296301
	MBR12	K=11, cs MINER, std	0.304322
	MBR13	K=5, cs MINER, rango	0.307319
	MBR10	K=7, cs MINER, std	0.315944
	MBR9	K=5, cs MINER, std	0.330949

Tabla 9-14. Comparación VASE caminos 3 y 4

Por otro lado, el modelo ganador fue K=11 (VASE=0.2565) mediante el método del rango. En este caso, a medida que se aumenta el número de vecinos el VASE va disminuyendo por lo que probamos aumentando dicho valor a 13,15 y 21. En los resultados se obtuvo que con 13 vecinos se logra disminuir el VASE (0.2475) y, a partir de 15 comienza a aumentar (0.2764 y 0.3422).

En este camino, la estandarización tampoco da buenos resultados en el VASE.

- Entre el camino con selección de la RL (5 y 6), los mejores fueron K=5 (0.2510, ganador) con tipificación y K=9 con rango (0.2561). En esta ocasión, no hay distinción entre un método y otro como ha ocurrido con los otros caminos.

Comparación de los modelos a ser candidatos a ganadores

Estadísticos de ajuste

Modelo seleccionado	Nodo del modelo	Descripción del modelo	Criterio de selección: Valid: Average Squared Error
Y	MBR17	K=5, cs RL, std	0.250971
	MBR23	K=9, cs RL, rango	0.256099
	MBR24	K=11, cs RL, rango	0.27348
	MBR22	K=7, cs RL, rango	0.284518
	MBR20	K=11, cs RL, std	0.285174
	MBR19	K=9, cs RL, std	0.289144
	MBR18	K=7, cs RL, std	0.290822
	MBR21	K=5, cs RL, rango	0.32592

Tabla 9-15. Comparación VASE mejores modelos candidatos

De los mejores caminos, el VASE y R^2 correspondientes son:

Camino	Variables	K	VASE	R2
K=5, ss, rango	15	5	0.285547	0.75059219
K=7, ss, rango	15	7	0.292459	0.74455498
K=11, cs MINER, rango	12	11	0.256529	0.77593764
K=9, cs MINER, rango	12	9	0.261127	0.77192157
K=13, cs MINER, rango	12	13	0.247533	0.78379509
K=5, cs RL, std	6	5	0.250971	0.78079221
K=9, cs RL, rango	6	9	0.256099	0.77631322

Tabla 9-16. Datos modelos candidatos

El que da peores resultados en ambas métricas es el camino con todas las variables mientras que haciendo selección se explica más del 77% de la variabilidad.

El camino con los valores más altos es el 4.5 (*K=13, cs MINER, rango*). No obstante, con el 6.1 (*K=5, cs RL, std*) apenas existen discrepancias y se utilizan menos variables y menos vecinos.

Luego, para evitar los problemas de aleatoriedad y poder hacer un estudio del sesgo y de la varianza, se realizará validación cruzada repetida representada en un gráfico de cajas y bigotes. Los caminos con los que se llevará a cabo son los que poseen un ASE

menor a 0.26 (*K=11, cs MINER, rango; K=13, cs MINER, rango; K=5, cs RL, std*; y *K=9, cs RL, rango*).

Agrupar índice	Modelo seleccionado ▼	Nodo predecesor	Descripción del modelo	Criterio de selección: Valid: Average Squared Error
1Y	MBR31	K=5, cs RL, std		0.423708
2Y	MBR31	K=5, cs RL, std		0.392259
3Y	MBR31	K=5, cs RL, std		0.415868
4Y	MBR31	K=5, cs RL, std		0.321304
5Y	MBR31	K=5, cs RL, std		0.200909
6Y	MBR31	K=5, cs RL, std		0.44212
7Y	MBR31	K=5, cs RL, std		0.292359
8Y	MBR31	K=5, cs RL, std		0.281088
9Y	MBR31	K=5, cs RL, std		0.252532
10Y	MBR31	K=5, cs RL, std		0.295968

Tabla 9-17. VASE VCR mejores candidatos Happiness

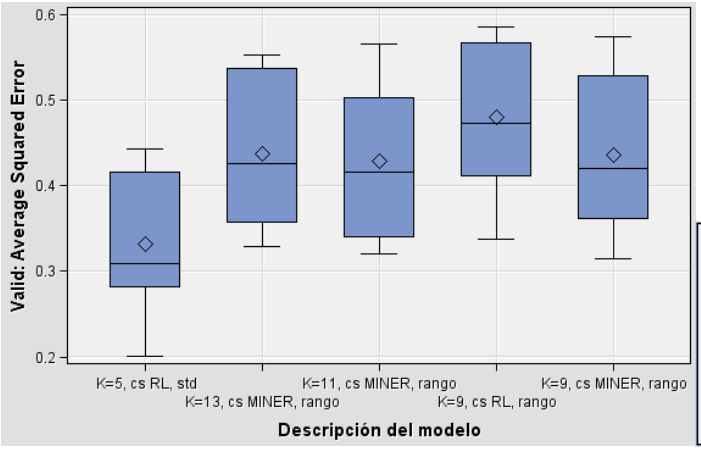


Gráfico VCR 5. Mejores candidatos Happiness

Descripción del modelo =	K=5, cs RL, std
Whisker mínimo =	0.200909
Primer cuartil =	0.281088
Mediana =	0.308636
Tercer cuartil =	0.415868
Whisker máximo =	0.44212
Media =	0.331812

Ilustración 5. Descripción modelo camino 5.1.

En las 10 repeticiones realizadas ha salido como modelo ganador el camino 4.5 y, a la vista del gráfico, no cabe duda de que es el mejor tanto en sesgo como en varianza.

Luego, **el mejor modelo en KNN cuando Y es continua es aquel con las variables seleccionadas de la regresión lineal, estandarización de variables cuantitativas y 5 vecinos**.

Teniendo σ^2 = 1.1449, $ASE_{entrenamiento}$=0.226716, $ASE_{validación}$=0.311287 y ASE_{prueba}=0.416191.

Con el $\mathbf{R^2_{entrenamiento}}$ se explica un 80.20% de la variabilidad de los datos.

Con el $\mathbf{R^2_{validación}}$ se explica un 72.81% de la variabilidad de los datos.

Con el $\mathbf{R^2_{prueba}}$ se explica un 63.48% de la variabilidad de los datos.

9.3.2. Y nominal

Los caminos que vamos a realizar van a ser los mismos que en el caso anterior con previa partición de datos de entrenamiento 70-30 utilizando a la variable objetivo como nominal en 5 categorías (*Happiness_cat*):

- **Camino 1**: variables sin seleccionar (SS), tipificación, K=5 (1.1), 7 (1.2), 9 (1.3) y 11 (1.4).
- **Camino 2**: variables sin seleccionar, rango, K=5 (2.1), 7 (2.2), 9 (2.3) y 11 (2.4).
- **Camino 3**: variables seleccionadas MINER (nodo: *Selección De Variables*), tipificación, K=5 (3.1), 7 (3.2), 9 (3.3) y 11 (3.4).
- **Camino 4**: variables seleccionadas MINER, rango, K=5 (4.1), 7 (4.2), 9 (4.3) y 11 (4.4).
- **Camino 5**: variables seleccionadas RL, tipificación, K=5 (5.1), 7 (5.2), 9 (5.3) y 11 (5.4).
- **Camino 6**: variables seleccionadas RL, rango, K=5 (6.1), 7 (6.2), 9 (6.3) y 11 (6.4).
- **Camino 7**: variables seleccionadas RLog, tipificación, K=5 (7.1), 7 (7.2), 9 (7.3) y 11 (7.4).
- **Camino 8**: variables seleccionadas RLog, rango, K=5 (8.1), 7 (8.2), 9 (8.3) y 11 (8.4).

Comparando todos estos caminos, los resultados son los siguientes:

- En el nodo *Selección de variables*, se han rechazado *IMP_REP_Freedom, IMP_REP_Prevalencia, REP_Gobierno, REP_Horas_sol*, *REP_IDH* y *REP_Ln_PIB_per_capita*.

USE	MODEL	MODELDESCRIPTION	_CRITERION_
Y	MBR16	K=5, cs MINER, std	0.068889
	MBR40	K=5, cs RLog, rango	0.073778
	MBR35	K=5, cs RLog, std	0.073778
	MBR9	K=7, cs MINER, std	0.074376
	MBR36	K=7, cs RLog, std	0.074603
	MBR12	K=5, cs MINER, rango	0.075111
	MBR41	K=7, cs RLog, rango	0.077324
	MBR3	K=9, ss, std	0.078601
	MBR13	K=7, cs MINER, rango	0.078685
	MBR4	K=11, ss, std	0.080073
	MBR6	K=7, ss, rango	0.080726
	MBR2	K=7, ss, std	0.081179
	MBR	K=5, ss, std	0.081333
	MBR11	K=11, cs MINER, std	0.081635
	MBR37	K=9, cs RLog, std	0.082305
	MBR10	K=9, cs MINER, std	0.082305
	MBR5	K=5, ss, rango	0.082667
	MBR42	K=9, cs RLog, rango	0.083265
	MBR14	K=9, cs MINER, rango	0.085597
	MBR18	K=7, cs RL, std	0.085941
	MBR17	K=5, cs RL, std	0.086222
	MBR15	K=11, cs MINER, rango	0.086501
	MBR21	K=5, cs RL, rango	0.086667
	MBR19	K=9, cs RL, std	0.087243
	MBR7	K=9, ss, rango	0.08738
	MBR22	K=7, cs RL, rango	0.087755
	MBR43	K=11, cs RLog, rango	0.087879
	MBR39	K=11, cs RLog, std	0.088981
	MBR8	K=11, ss, rango	0.089715
	MBR23	K=9, cs RL, rango	0.090123
	MBR20	K=11, cs RL, std	0.090725
	MBR24	K=11, cs RL, rango	0.094399

Tabla 9-18. Comparación VASE todos los caminos Happiness_cat

- En este caso, hay variedad en estandarizar o aplicar el rango. Además, en el ranking de los que tienen un VASE menor, se encuentran variables seleccionadas ya sea por MINER o por logística y con valores de K no muy grandes (5 o 7). En este último, se observa que es preferible con pocos vecinos frente a incrementar dicho número. El camino 5 y 6 no ha generado buenos resultados, se encuentra entre los peores.

  Como ganador de todos los caminos realizados se tiene *K=5, cs MINER, std* con un VASE=0.0689.

La **vcr** se realizará con los modelos recuadrados en rojo cuyo $ASE_{validación}$ es menor que 0.075.

LOOP	USE ▼	MODELDESCRIPTION	_CRITERION_
1Y		K=5, cs MINER, std	0.082963
2Y		K=7, cs MINER, std	0.074587
3Y		K=7, cs MINER, std	0.090379
4Y		K=7, cs RLog, std	0.095742
5Y		K=5, cs RLog, std	0.063908
6Y		K=7, cs MINER, std	0.117493
7Y		K=7, cs MINER, std	0.091912
8Y		K=7, cs MINER, std	0.068571
9Y		K=7, cs RLog, std	0.072814
10Y		K=5, cs MINER, std	0.071905

Tabla 9-19. VCR mejores candidatos Happiness_cat

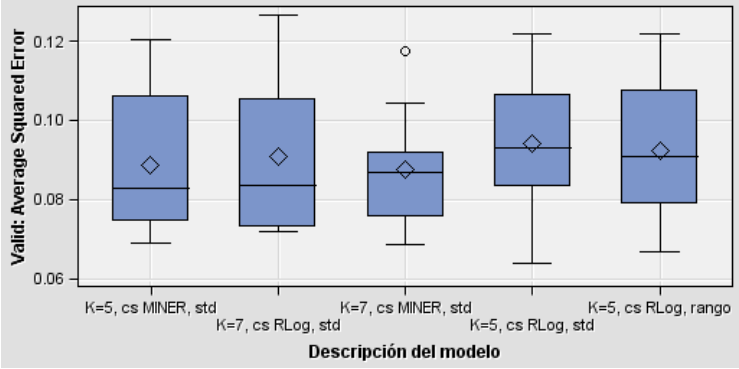


Gráfico VCR 6. Mejores candidatos Happiness_cat

Los modelos en los que ha sido llevado a cabo el método del rango no han aparecido como de los mejores en ninguna ocasión. Además, los 5 modelos se encuentran en la misma media y mediana.

7 de las 10 repeticiones ha salido como vencedor *K*=7 mientras que los 3 restantes son de *K=5.*

El set más frecuente ha sido K=7*, cs MINER, std* siendo el que menor rango intercuartílico posee.

Luego, **el mejor modelo en KNN cuando Y es categórica es aquel con las variables seleccionadas en MINER, estandarización de variables cuantitativas y 7 vecinos**. Los valores de ASE en los distintos ficheros son: $ASE_{entrenamiento}$=0.0673, $ASE_{validación}$=0.0907 y ASE_{prueba}=0.0839

10. CONCLUSIONES

A lo largo de este TFM ha sido posible un estudio paralelo según el tipo de variable objetivo que se tenía indagando por varias técnicas de machine learning tanto de aprendizaje no supervisado como supervisado los cuales han hecho posible el cumplimiento de los objetivos y cuyas conclusiones fueron las siguientes:

En primer lugar, se ha visto en el **análisis descriptivo** que la mayor parte de los países en el mundo se encuentran en una escala de la felicidad de 5 a 6. Además, a través del mapa nos hemos dado cuenta que los continentes donde existe una mayor felicidad son América y Europa mientras que en África se sitúan los países menos felices. Gracias a ello, nos hemos dado cuenta de que la variable *REP_Regional_indicator* iba a ser clave para poder cumplir con el objetivo 2.

En segundo lugar, gracias a la **depuración de datos** fue posible corregir los errores, detectar los atípicos y llevar a cabo una imputación de los datos faltantes. Además, independientemente de si la variable objetivo estaba en forma numérica o categórica, las 5 variables más importantes según el gráfico de valor eran la esperanza de vida saludable, el PIB per cápita, el índice de desarrollo del país, la esperanza de vida y el apoyo social.

En tercer lugar, debido a que muchas de las variables de la base de datos estaban altamente correlacionadas con *Happiness* se realizó una **reducción de la dimensión** a través de FAMD el cual nos permitía trabajar con datos mixtos. Se pasó de explicar la felicidad con 15 variables a explicar casi el 75% de la variabilidad con 8 componentes. De esta manera, fue posible visualizar las relaciones entre las variables tanto numéricas como cualitativas y, a su vez, representar la relación de estos grupos de variables con nuestra variable objetivo la felicidad.

En cuarto lugar, como métodos de aprendizaje supervisado y, lo que ha permitido el cumplimiento de los objetivos 2 y 3, fue llevado a cabo un estudio más profundo para poder sacar un modelo de predicción para *Happiness* y *Happiness_bin* mediante una previa selección de variables y utilizando VCR.

Por un lado, el **mejor modelo de regresión lineal** que se ha encontrado para predecir la escala de la felicidad ha sido el compuesto por el indicador regional, la tasa de desempleo, las horas medias de sol al año, el índice de desarrollo humano, el apoyo social y la libertad individual. Éstas tres últimas fueron las que más peso tenían sobre Y. Su R^2 fue de 0.832.

Por otro lado, el **mejor modelo de regresión logística** que se ha encontrado para predecir la felicidad o no felicidad de un país ha sido el formado por la libertad individual, el índice de desarrollo humano, el PIB, las horas medias de sol al año y el apoyo social. Destacar que todas las variables excepto las horas de sol poseían asociación negativa (odds ratio<1).

Por último, en cuanto al **KNN**, a pesar de que cuando se aplicaba VCR parecía que el método del rango para las variables cuantitativas iba a salir como ganador, se veía en el gráfico de cajas y bigotes que no eran muy estables en varianza (outliers). En consecuencia, tanto para *Happiness* como para *Happiness_cat*, nos hemos quedado con la estandarización. En el primer caso, el mejor modelo en KNN es aquel con las variables seleccionadas de la regresión lineal y 5 vecinos. En el segundo caso, con las variables seleccionadas del MINER y nos quedamos con 7 vecinos.

Otras aplicaciones que se podrían haber elaborado hubieran sido:

- Incluir más variables para ver si pueden resultar significativas para predecir la felicidad en el mundo.
- Probar con más conjuntos de variables tanto en regresión lineal como logística.

- Utilizar la regresión logística multinomial para predecir *Happiness_cat*.
- Llevar a cabo un análisis clúster para buscar países que se parezca entre sí y agruparlos.
- Realización de un árbol, ya sea de clasificación o de predicción.

11. BIBLIOGRAFÍA

Bécue-Bertaut, M., & Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, *52*(6), 3255-3268. https://doi.org/10.1016/j.csda.2007.09.023

Carlos Cobos. (2018, marzo 20). *Proceso KDD y CRISP-DM*. https://www.youtube.com/watch?v=HTWcPMIGOiU

Depression Rates By Country 2021. (s. f.). Recuperado 8 de junio de 2021, de https://worldpopulationreview.com/country-rankings/depression-rates-by-country

FAMD - Factor Analysis of Mixed Data in R: Essentials - Articles - STHDA. (s. f.). Recuperado 21 de junio de 2021, de http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/

FAMD in R Using FactoMineR: Quick Scripts and Videos - Articles - STHDA. (s. f.). Recuperado 22 de junio de 2021, de http://www.sthda.com/english/articles/22-principal-component-methods-videos/72-famd-in-r-using-factominer-quick-scripts-and-videos/

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857-871. https://doi.org/10.2307/2528823

Harvard Second Generation Study. (s. f.). Harvardstudy. Recuperado 28 de junio de 2021, de https://www.adultdevelopmentstudy.org

Hdr_2020_overview_spanish.pdf. (s. f.). Recuperado 9 de junio de 2021, de http://hdr.undp.org/sites/default/files/hdr_2020_overview_spanish.pdf

Life Expectancy by Country 2021. (s. f.). Recuperado 12 de junio de 2021, de https://worldpopulationreview.com/countries/life-expectancy

Maya, J. S. C. (s. f.). *Los 4 Factores Determinantes De La Felicidad*. Desarrollo Personal Por Juan Sebastián Celis Maya. Recuperado 8 de junio de 2021, de https://www.sebascelis.com/los-4-factores-determinantes-de-la-felicidad/

Montuschi, L. (s. f.). *CRECIMIENTO ECONOMICO, PROGRESO SOCIAL Y FELICIDAD*. 30.

Most Dangerous Cities In The World. (s. f.). Recuperado 8 de junio de 2021, de https://worldpopulationreview.com/world-city-rankings/most-dangerous-cities-in-the-world

The GENMOD Procedure. (s. f.). 221.

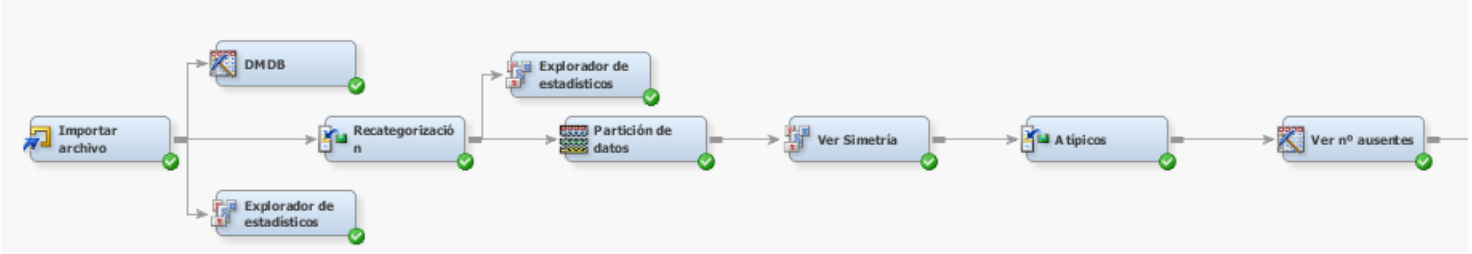
Viens, A. (2019, octubre 26). World Cities Ranked by Average Annual Sunshine Hours. *Visual Capitalist*. https://www.visualcapitalist.com/world-cities-ranked-by-average-annual-sunshine-hours/

WHR+21.pdf. (s. f.). Recuperado 8 de junio de 2021, de https://happiness-report.s3.amazonaws.com/2021/WHR+21.pdf

12.ANEXO

SAS MINER

Diagrama depuración de los datos



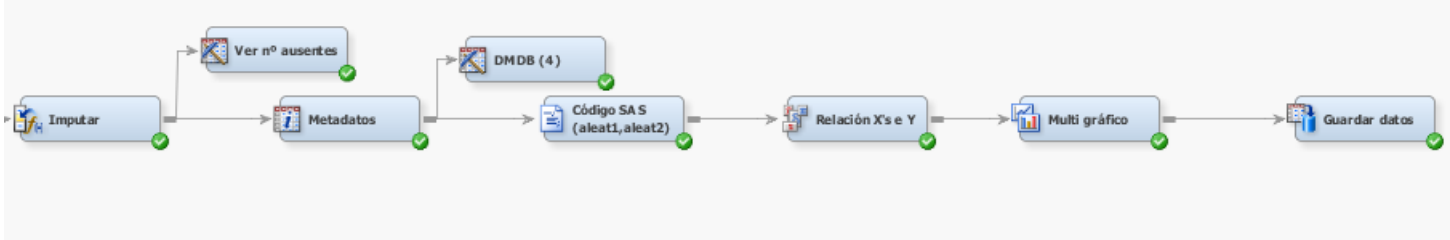
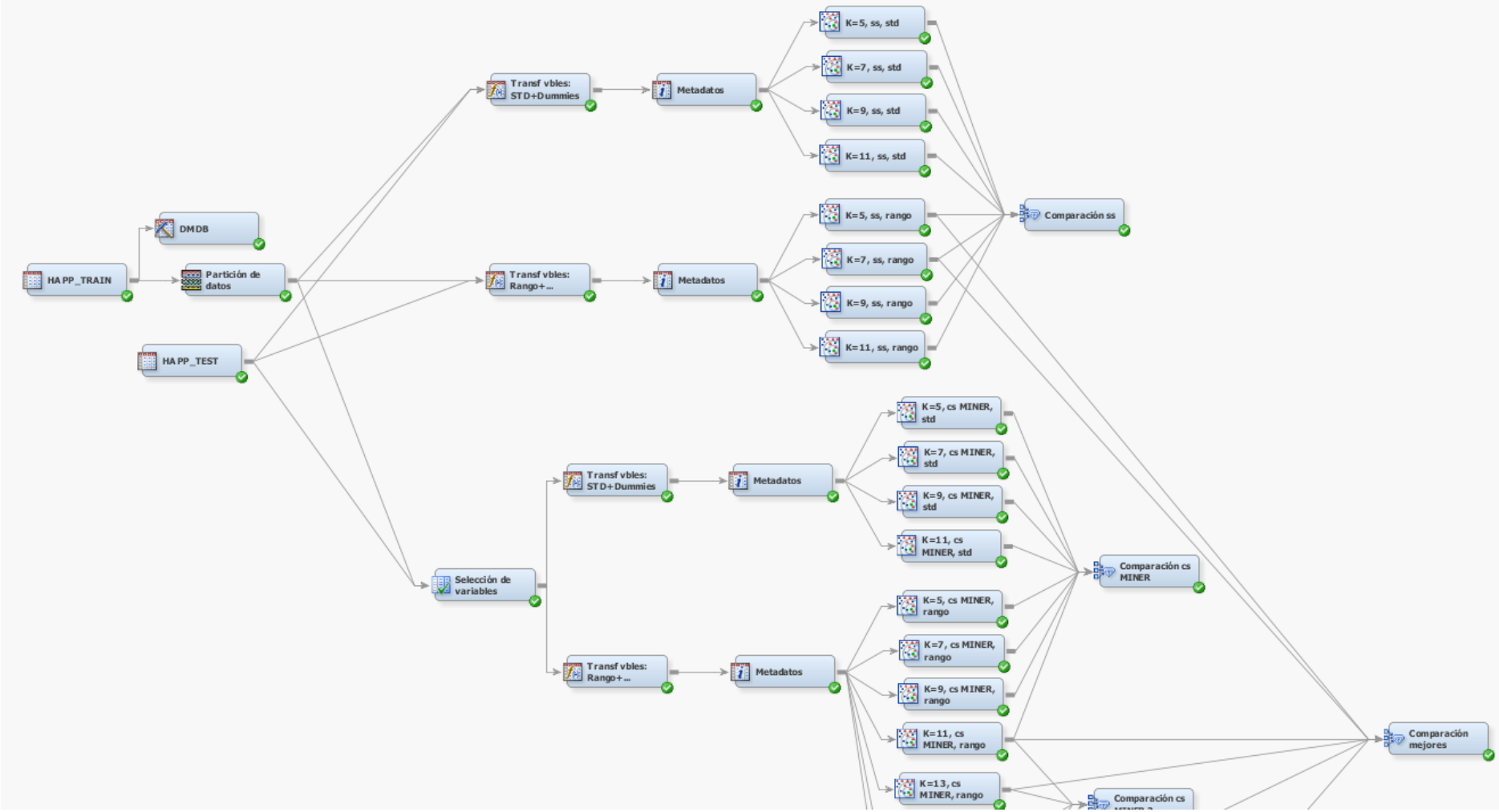
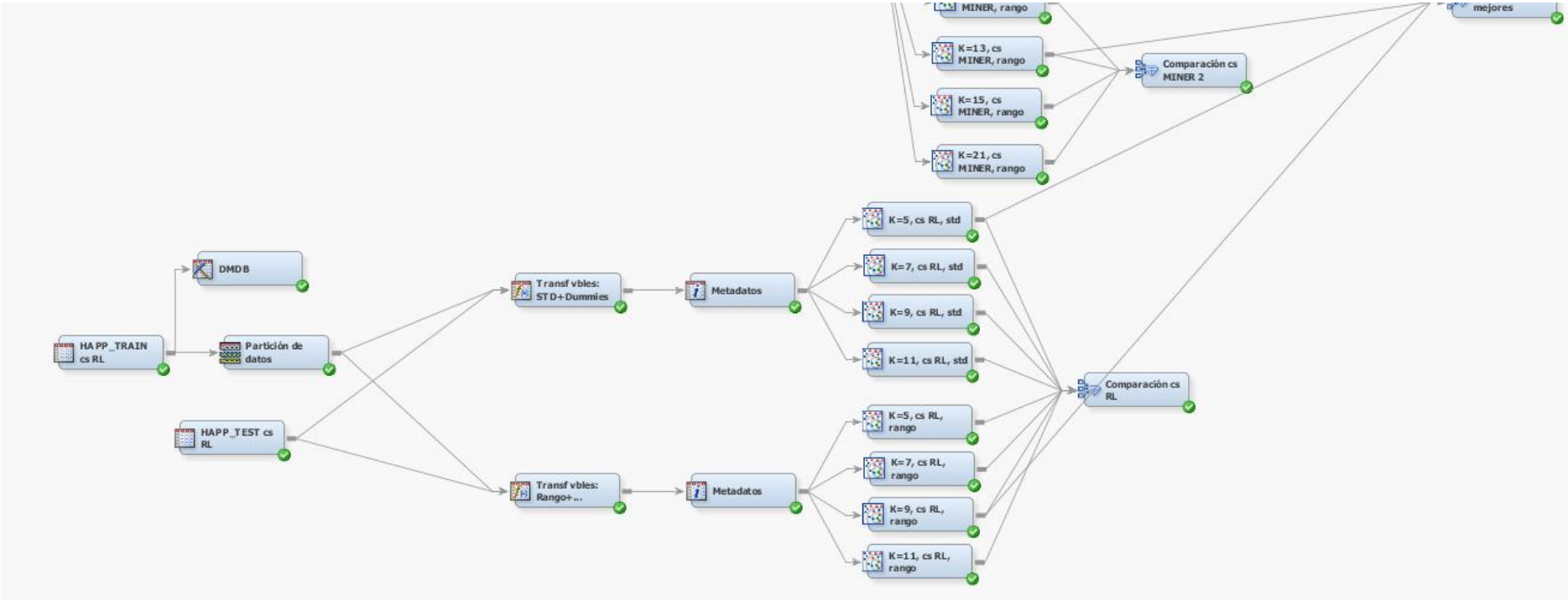


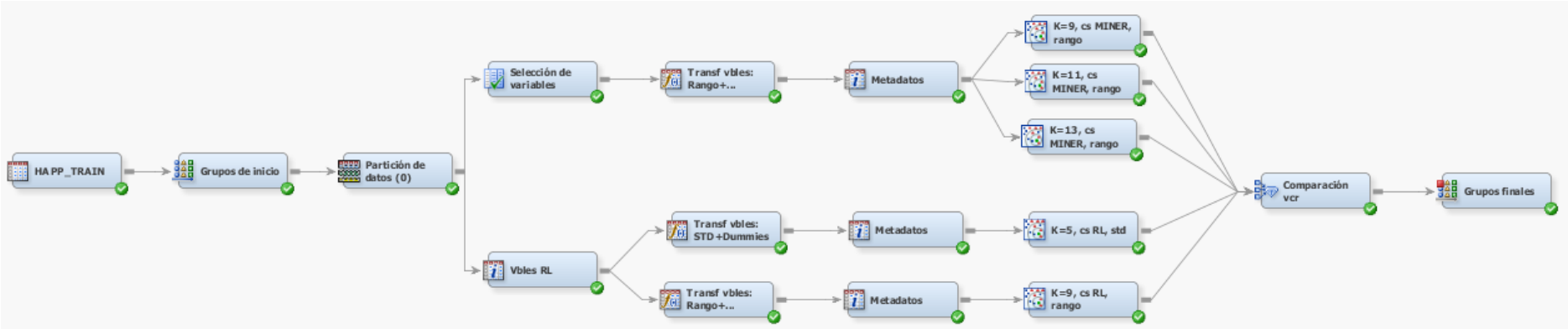
Diagrama KNN con Y continua

Caminos para la obtención del mejor K




MINER, rango
mejores
K=13, cs MINER, rango
K=15, cs MINER, rango
K=21, cs MINER, rango
Comparación cs MINER 2
K=5, cs RL, std
K=7, cs RL, std
K=9, cs RL, std
K=11, cs RL, std
DMDB
Transf vbles: STD+Dummies
Metadatos
HAPP_TRAIN cs RL
Partición de datos
HAPP_TEST cs RL
Comparación cs RL
K=5, cs RL, rango
K=7, cs RL, rango
K=9, cs RL, rango
K=11, cs RL, rango
Transf vbles: Rango+...
Metadatos

Bucles inicio-fin (VCR)



Datos del mejor K

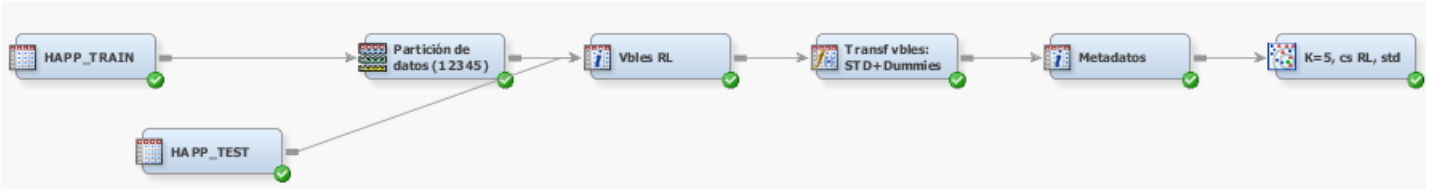
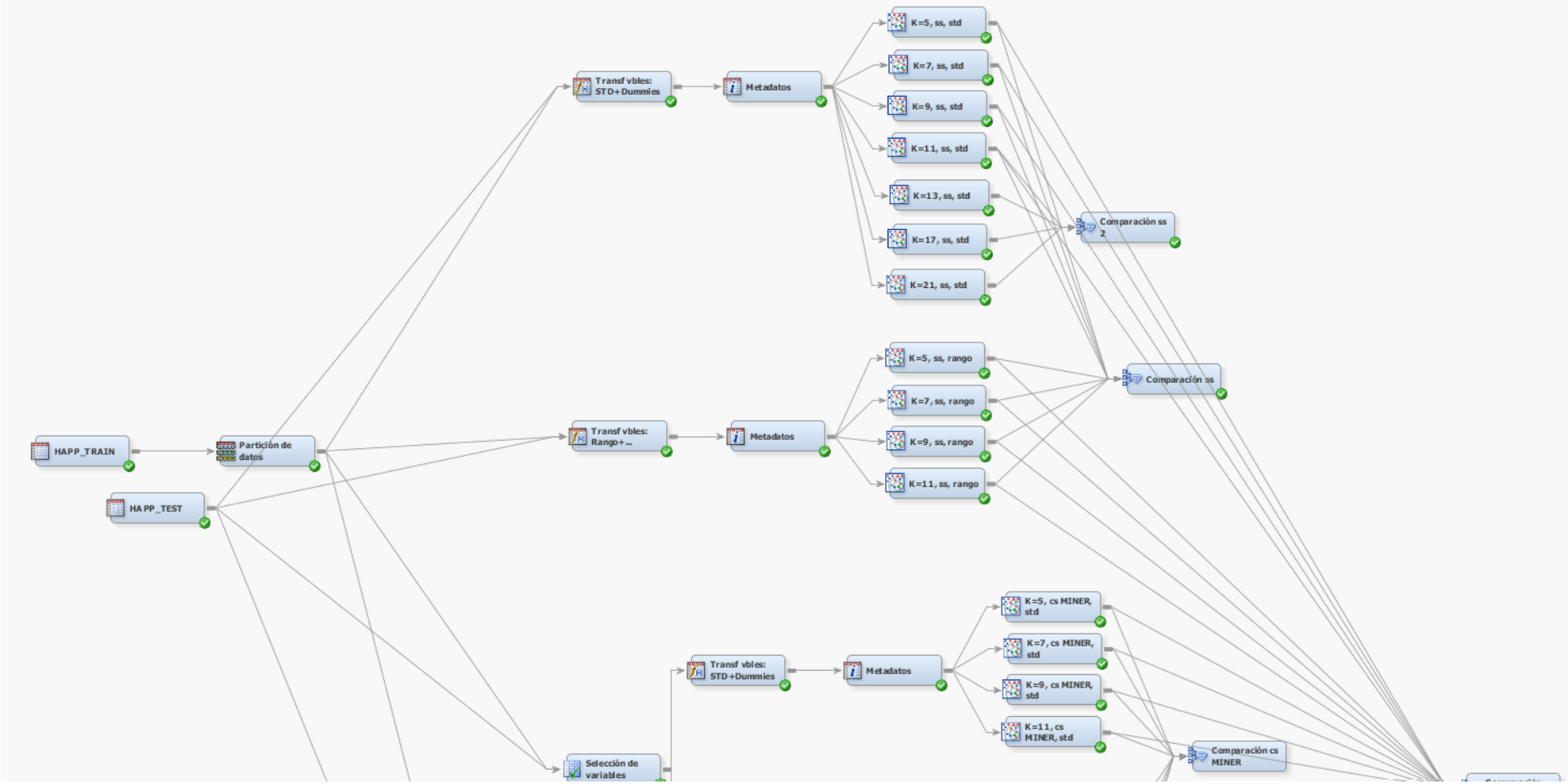
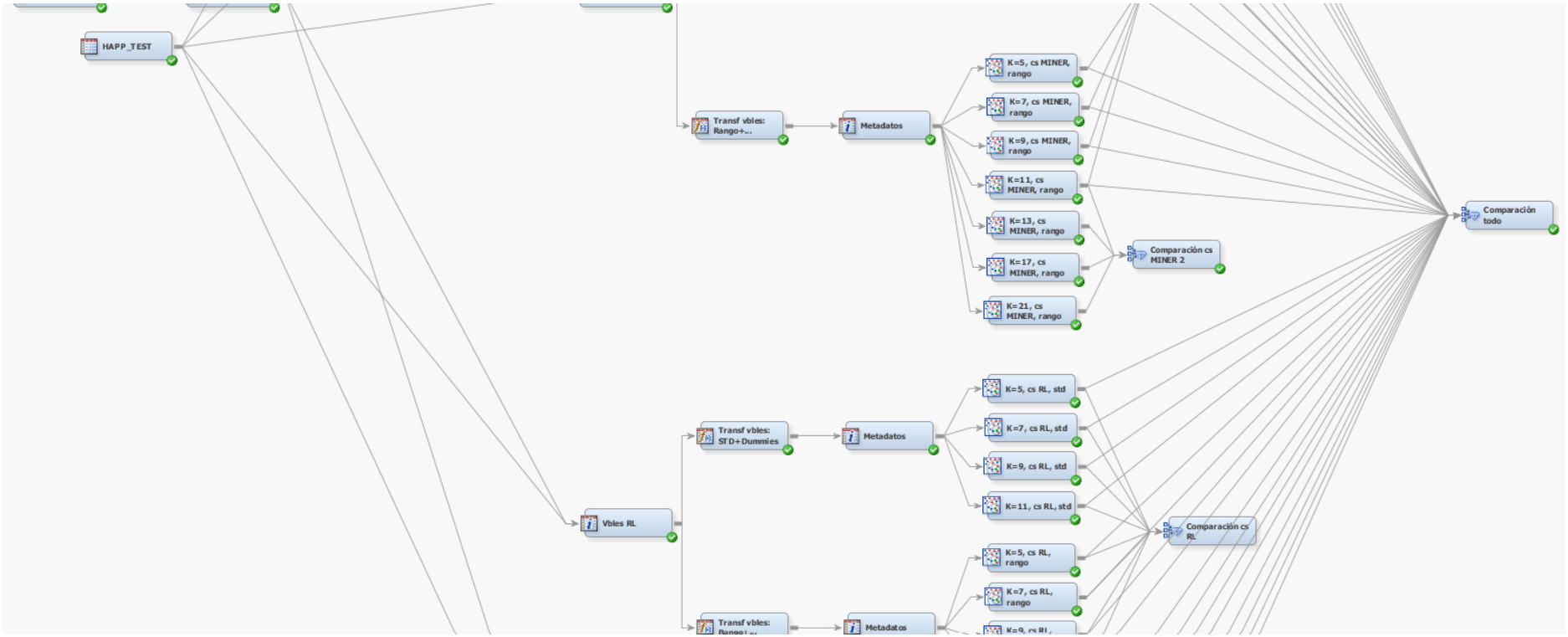
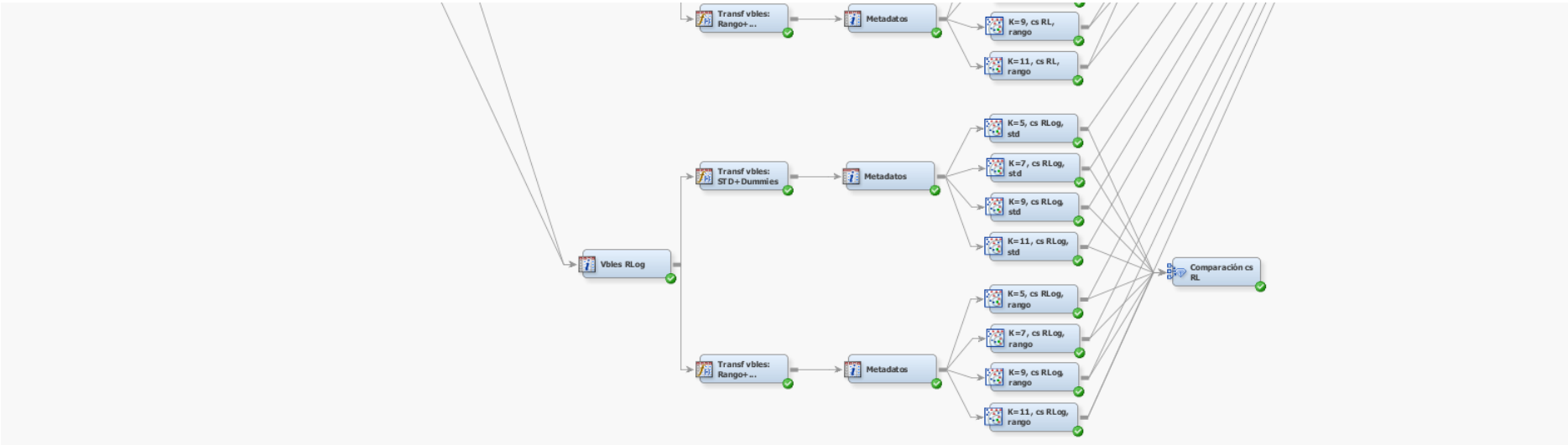


Diagrama KNN con Y categórica

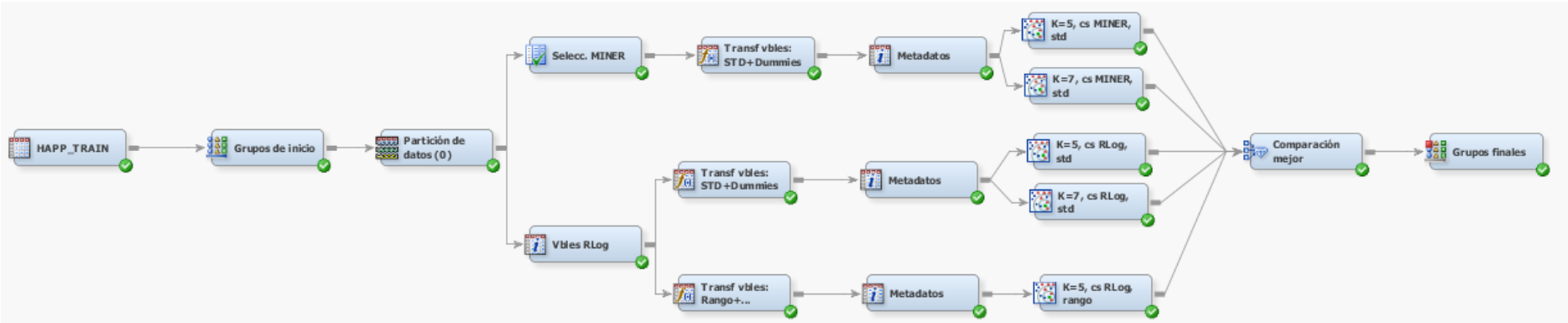
Caminos para la obtención del mejor K



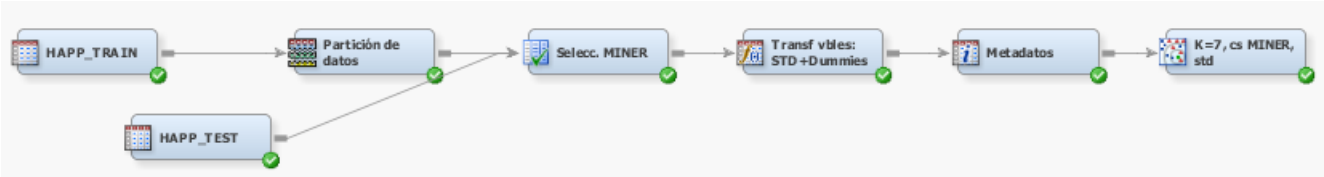

HAPP_TEST
Transf vbles: Rango+...
Metadatos
K=5, cs MINER, rango
K=7, cs MINER, rango
K=9, cs MINER, rango
K=11, cs MINER, rango
K=13, cs MINER, rango
K=17, cs MINER, rango
K=21, cs MINER, rango
Comparación cs MINER 2
Comparación todo
Vbles RL
Transf vbles: STD+Dummies
Metadatos
K=5, cs RL, std
K=7, cs RL, std
K=9, cs RL, std
K=11, cs RL, std
Comparación cs RL
K=5, cs RL, rango
K=7, cs RL, rango
Transf vbles:
Metadatos



Bucle inicio-fin (VCR)



Datos del mejor K



R

Paquetes/Librerías

```
install.packages(c("FactoMineR", "factoextra"))

library("FactoMineR")

library("factoextra")

library(visualpred)

library(sas7bdat)
```

Lectura y preparación fichero

```
setwd('C:/Users/belen/Documents/UCM/MASTER/TFM')

data <- read.sas7bdat("depu.sas7bdat")

listconti <-
c("REP_Corruption","REP_Desempleo","REP_Esperanza_vida","REP_Generosity",
               "REP_Healthy_life_expectancy","REP_Horas_sol","REP_IDH",
               "REP_Ln_PIB_per_capita","REP_Social_Support", "REP_Freedom",
               "IMP_REP_Prevalencia")

listcat <- c('Rep_Regional_indicator', 'Peligroso', 'REP_Clima',
'REP_Gobierno')


data$Peligroso <- as.factor(data$Peligroso)

df <- data[,c(2, 4:18)]
```

FAMD

```
# Con todas las variables #

AFMD1 <- FAMD (df, ncp=30, graph = FALSE) #hasta 22

get_eigenvalue(AFMD1)

## Visualize

library(gridExtra)

library(grid)

a <- fviz_eig(AFMD1,
              choice='eigenvalue',
              geom='line')

grid.arrange(a)

  # con 8 autovalores es suficiente.


fviz_screeplot(AFMD1)

fviz_screeplot(FAMD (df, ncp=8, graph = FALSE))


AFMD8 <- FAMD (df, ncp=8, graph = FALSE, sup.var=2)

    #AFM de 8 con vble suplementoria

# # Coordinates of variables

correlaciones <-
rbind(AFMD8[["quanti.var"]][["coord"]],AFMD8[["quali.var"]][["coord"]])

correlaciones
```

```
# # Cos2: quality of representation on the factore map
cos2 <- rbind(AFMD8[["quanti.var"]][["cos2"]],AFMD8[["quali.var"]][["cos2"]])
round(cos2,digits=3)
  # proporción de varianza de cada variable que es explicada por cada
componente

# # Contributions to the dimensions
# head(var$contrib)
contribuciones <-
rbind(AFMD8[["quanti.var"]][["contrib"]],AFMD8[["quali.var"]][["contrib"]])
round(contribuciones,digits=3)

# AUTOVECTORES: coeficientes para construir cada componente.
options(digits=3)
autovectores <- AFMD8$svd$V

# Plot of variables cuali + cuanti
par(cex.axis=0.5,mfrow = c(4, 2))
# dim 1 y 2
fviz_famd_var(AFMD8, repel = TRUE, col.var = 'cos2')
# dim 3 y 4
fviz_famd_var(AFMD8, repel = TRUE, axes = c(3,4), col.var = 'cos2')
# dim 5 y 6
fviz_famd_var(AFMD8, repel = TRUE, axes = c(5,6),col.var = 'cos2')
# dim 7 y 8
fviz_famd_var(AFMD8, repel = TRUE, axes = c(7,8), col.var = 'cos2')
# Plot contributions in each dimensión
ggarrange(ncol=2,nrow=4,
          plot(fviz_contrib(AFMD8, "var", axes = 1)),
          plot(fviz_contrib(AFMD8, "var", axes = 2)),
          plot(fviz_contrib(AFMD8, "var", axes = 3)),
          plot(fviz_contrib(AFMD8, "var", axes = 4)),
          plot(fviz_contrib(AFMD8, "var", axes = 5)),
          plot(fviz_contrib(AFMD8, "var", axes = 6)),
          plot(fviz_contrib(AFMD8, "var", axes = 7)),
          plot(fviz_contrib(AFMD8, "var", axes = 8)))

# Quantitative variables
require(egg)
library(visualpred)
ggarrange(ncol =2,nrow=2,
fviz_famd_var(AFMD8, axes=c(1,2), "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
```

```
              repel = TRUE),
fviz_famd_var(AFMD8, axes=c(3,4), "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE),
fviz_famd_var(AFMD8, axes=c(5,6), "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE),
fviz_famd_var(AFMD8, axes=c(7,8), "quanti.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE))
# Qualitive variables
ggarrange(ncol =2,nrow=2,
fviz_famd_var(AFMD8, axes = c(1,2), "quali.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel=TRUE),
fviz_famd_var(AFMD8, axes = c(3,4), "quali.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel=TRUE),
fviz_famd_var(AFMD8, axes = c(5,6), "quali.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel=TRUE),
fviz_famd_var(AFMD8, axes = c(7,8), "quali.var", col.var = "contrib",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel=TRUE))

# graph of individuals cuali + cuanti (categorias)
p <-fviz_famd_ind(AFMD8, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE, axes = c(1,2))
fviz_add(p, AFMD8$quanti.sup$coord , color ="red")

p <-fviz_famd_ind(AFMD8, col.ind = "cos2",
                  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                  repel = TRUE, axes = c(3,4))
fviz_add(p, AFMD8$quanti.sup$coord , color ="red")

p <-fviz_famd_ind(AFMD8, col.ind = "cos2",
                  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                  repel = TRUE, axes = c(5,6))
fviz_add(p, AFMD8$quanti.sup$coord , color ="red")

p <-fviz_famd_ind(AFMD8, col.ind = "cos2",
                  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
```

```
                repel = TRUE, axes = c(7,8))
fviz_add(p, AFMD8$quanti.sup$coord , color ="red")
# matriz de correlaciones en cada dimensión
corrplot(cargas, is.corr= FALSE)

# Puntuaciones: valores de los indiviuos en cada componente
puntuaciones <- get_famd_ind(AFMD8)$coord
```

SAS BASE

```
LIBNAME tfm 'C:\Users\belen\Documents\UCM\MASTER\TFM';
PROC CONTENTS data=tfm.happ;
RUN;
```

Análisis exploratorio

```
/* 1 - ANÁLISIS EXPLORATORIO DE LOS DATOS */
      /* 1.1. Variables cualitativas */
            /* happiness_cat */
PROC FREQ data=tfm.happ;
       TABLES happiness_cat;
RUN;
PROC SGPLOT DATA = tfm.happ; ** variable ordinal;
VBAR happiness_cat;
RUN;
            /* regional_indicator */
PROC FREQ data=tfm.happ;
       TABLES regional_indicator;
RUN;
PROC GCHART data=tfm.happ;
       PIE regional_indicator;
RUN;
            /* clima */
PROC FREQ data=tfm.happ;
       TABLES clima;
RUN;
PROC GCHART data=tfm.happ;
       PIE clima;
RUN;
            /* gobierno */
PROC FREQ data=tfm.happ;
       TABLES gobierno;
RUN;
PROC GCHART data=tfm.happ;
       PIE gobierno;
RUN;
            /* peligroso */
PROC FREQ data=tfm.happ;
       TABLES peligroso;
RUN;
PROC GCHART data=tfm.happ;
       PIE peligroso;
RUN;
      /* 1.2. Variables cuantitativas */
PROC MEANS DATA=tfm.happ fw=8 maxdec=3 nmiss mean median std cv kurtosis skew
min max range qrange;
      VAR Prevalencia IDH Desempleo Esperanza_vida Happiness
Ln_PIB_per_capita_ Social_support
            Healthy_life_expectancy Freedom Generosity Corruption;
RUN;

/* MAPA FELICIDAD */
goptions reset=all border;
TITLE 'Escala de la felicidad en el mundo';
```

```
DATA mapa;
      MERGE tfm.happ tfm.mapsid ;
RUN;
PROC GMAP map=mapsgfk.world DATA=mapa all;
      ID ISO;
      CHORO happiness / levels=5 cdefault=cxfcfad2;
RUN;
```

Regresión lineal

```
/* 2. REGRESIÓN LINEAL */
LIBNAME tfm 'C:\Users\belen\Documents\UCM\MASTER\TFM';
      /* 2.1. RL con Y */
            /* 2.1.1. SELECCIÓN DE VBLES */
            /* para la selección de variables utilizaremos la macro de
Portela de ML*/
                  /* 2.1.1.1. AIC */
ODS OUTPUT SelectedEffects = efectos;
PROC GLMSELECT data=tfm.happ_train noprint;
      CLASS Peligroso REP_Clima REP_Gobierno REP_Regional_Indicator;** vbles
categóricas;
    MODEL Happiness = Peligroso REP_Clima--IMP_REP_Prevalencia
                  / selection=stepwise(select=AIC choose=AIC);
RUN;
PROC PRINT data=efectos; RUN;
DATA; set efectos; put effects; RUN;
      ** 8 vbles:
      Intercept REP_Regional_indicator REP_Corruption REP_Desempleo
REP_Horas_sol REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia;
                  /* 2.1.1.2. BIC */
ODS OUTPUT SelectedEffects = efectos;
PROC GLMSELECT data=tfm.happ_train;
      CLASS Peligroso REP_Clima REP_Gobierno REP_Regional_Indicator;** vbles
categóricas;
    MODEL Happiness = Peligroso REP_Clima--IMP_REP_Prevalencia
                  / selection=stepwise(select=BIC choose=BIC);
RUN;
PROC PRINT data=efectos; RUN;
DATA; set efectos; put effects; RUN;
      ** 7 vbles: Mismo pero sin IMP_REP_PREVALENCIA;
            /* 2.1.2. REPETICIÓN STEPWISE EN SUBMUESTRAS (VC) */
/* NO OLVIDAR Quitar el output HTML!!!! */
ODS GRAPHICS ON;
%randomselect(data = tfm.happ_train, /* todas macros regresion 4.0.sas */
                        listclass = Peligroso REP_Clima REP_Gobierno
REP_Regional_Indicator,
                        vardepen = happiness,
                        modelo = Peligroso REP_Clima--IMP_REP_Prevalencia,
                        criterio = AIC,
                        sinicio = 2000,
                        sfinal = 2200,
                        fracciontrain=0.8,
                        directorio = C:\Users\belen\Documents\UCM\MASTER\TFM
                        );
ODS GRAPHICS ON;
%randomselect(data = tfm.happ_train,
                        listclass = Peligroso REP_Clima REP_Gobierno
REP_Regional_Indicator,
                        vardepen = happiness,
                        modelo = Peligroso REP_Clima--IMP_REP_Prevalencia,
                        criterio = BIC,
                        sinicio = 2000,
                        sfinal = 2200,
                        fracciontrain=0.8,
                        directorio = C:\Users\belen\Documents\UCM\MASTER\TFM
                        );
            /* 2.1.3. VCR */
                  /* modelo 1: AIC +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
```

```
            conti = REP_Corruption REP_Desempleo REP_Horas_sol REP_IDH
REP_Social_support IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final1; SET final; modelo=1; RUN;
                  /* modelo 2: AIC 2° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Corruption REP_Desempleo REP_Horas_sol REP_IDH
REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final2; SET final; modelo=2; RUN;
                  /* modelo 3: AIC 3° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support
IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final3; SET final; modelo=3; RUN;
                  /* modelo 4: AIC 4° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Desempleo REP_Horas_sol REP_IDH REP_Social_support
IMP_REP_Freedom IMP_REP_Prevalencia,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final4; SET final; modelo=4; RUN;
                  /* modelo 5: AIC 5° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Corruption REP_Horas_sol REP_IDH REP_Social_support
IMP_REP_Freedom IMP_REP_Prevalencia,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final5; SET final; modelo=5; RUN;
                  /* modelo 6: AIC 4 vbles +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Horas_sol REP_Social_support IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final6; SET final; modelo=6; RUN;
                  /* modelo 7: BIC 1° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Corruption REP_Horas_sol REP_IDH REP_Social_support
IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final7; SET final; modelo=7; RUN;
                  /* modelo 8: BIC 3° +freq  */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Desempleo REP_IDH REP_Social_support IMP_REP_Freedom
IMP_REP_Prevalencia,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final8; SET final; modelo=8; RUN;
                  /* modelo 9: BIC 4° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti = REP_Generosity REP_Horas_sol REP_IDH REP_Social_support
IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final9; SET final; modelo=9; RUN;
                  /* modelo 10: BIC 5° +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
            conti =  REP_Corruption REP_Horas_sol REP_Ln_PIB_per_capita_
REP_Social_support IMP_REP_Freedom,
            categor = REP_Regional_indicator,
            ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final10; SET final; modelo=10; RUN;
```

```
                    /* modelo 11: BIC 5 vbles +freq */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
             conti = REP_Horas_sol REP_Social_support IMP_REP_Freedom REP_IDH,
             categor = REP_Regional_indicator,
             ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final11; SET final; modelo=11; RUN;
                    /* modelo 12: +imp en MINER */
%cruzada(archivo = tfm.happ_train, vardepen = happiness,
             conti = REP_Healthy_life_expectancy REP_Ln_PIB_per_capita_
REP_IDH REP_Esperanza_vida REP_Social_support,
             categor = ,
             ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final12; SET final; modelo=12; RUN;

DATA TFM.union_rl; set tfm.final1 tfm.final2 tfm.final3 tfm.final4 tfm.final5
tfm.final6
                                    tfm.final7 tfm.final8 tfm.final9
tfm.final10 tfm.final11 tfm.final12;
RUN;
PROC SORT data=tfm.union_rl; BY modelo; run;
PROC BOXPLOT data=tfm.union_rl;
       PLOT media*modelo;
RUN;
/* para mayor visibilidad, quitamos los de sesgo más alto y otros que son
peores con el mismo nº de vbles */
DATA union_rl; set tfm.final1 tfm.final3 tfm.final4
                                  tfm.final8 tfm.final9 tfm.final11;
RUN;
PROC BOXPLOT data=union_rl;
       PLOT media*modelo;
RUN;
PROC MEANS data=union_rl mean median var range min max;
       var media;
       BY modelo;
RUN;
PROC MEANS data=union_rl mean median var range min max;
       var media;
RUN;
/* duda entre modelos 3, 4 y 11 */
PROC GLM data=tfm.happ_train;
       CLASS REP_Regional_Indicator;
    MODEL Happiness = REP_Regional_Indicator REP_Desempleo REP_Horas_sol
REP_IDH REP_Social_support IMP_REP_Freedom;
RUN;
PROC GLM data=tfm.happ_train; ** modelo ganador;
       CLASS REP_Regional_Indicator;
    MODEL Happiness = REP_Regional_Indicator REP_Desempleo REP_Horas_sol
REP_IDH REP_Social_support IMP_REP_Freedom IMP_REP_Prevalencia;
RUN;
PROC GLM data=tfm.happ_train;
       CLASS REP_Regional_Indicator;
    MODEL Happiness = REP_Regional_Indicator REP_Horas_sol REP_Social_support
IMP_REP_Freedom REP_IDH;
RUN;

PROC GENMOD data=tfm.happ_train;
       CLASS REP_Regional_Indicator;
    MODEL Happiness = REP_Regional_Indicator REP_Desempleo REP_Horas_sol
REP_IDH REP_Social_support IMP_REP_Freedom
                                        /dist=NOR;
RUN;
```

Regresión logística

```
/* 3 - RLog con Y_bin */
LIBNAME tfm 'C:\Users\belen\Documents\UCM\MASTER\TFM';
       /* 3.1. Creación de la Y binaria */
DATA tfm.happ;
```

```
       SET tfm.happ;
       IF happiness < 5 THEN happiness_bin = 0; ** 0=no feliz;
       ELSE happiness_bin = 1; ** 1=feliz;
RUN;
DATA tfm.happ_train;
       SET tfm.happ_train;
       IF happiness < 5 THEN happiness_bin = 0;
       ELSE happiness_bin = 1;
RUN;
DATA tfm.happ_test;
       SET tfm.happ_test;
       IF happiness < 5 THEN happiness_bin = 0;
       ELSE happiness_bin = 1;
RUN;
       /* 3.2. SELECCIÓN DE VBLES */
              /* stepwise */
ODS OUTPUT type3=parametros;
PROC LOGISTIC data=tfm.happ_train namelen=20 descending ;
       CLASS Peligroso REP_Clima REP_Gobierno REP_Regional_Indicator;
       MODEL happiness_bin = Peligroso REP_Clima--IMP_REP_Prevalencia
                     /selection=stepwise link=glogit;
RUN;QUIT;
DATA mode;length effect $20. modelo $ 20000;retain modelo " ";set parametros
end=fin;effect=cat(' ',effect);
       if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then output;
RUN;
data;set mode;put modelo;run;
       **REP_IDH IMP_REP_Freedom;

              /* forward */
ODS OUTPUT type3=parametros;
PROC LOGISTIC data=tfm.happ_train namelen=20 descending ;
       CLASS Peligroso REP_Clima REP_Gobierno REP_Regional_Indicator;
       MODEL happiness_bin = Peligroso REP_Clima--IMP_REP_Prevalencia
                     /selection=forward link=glogit;
RUN;QUIT;
DATA mode;length effect $20. modelo $ 20000;retain modelo " ";set parametros
end=fin;effect=cat(' ',effect);
       if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then output;
RUN;
data;set mode;put modelo;run;
       **REP_Desempleo REP_IDH IMP_REP_Freedom;

              /* backward */
ODS OUTPUT type3=parametros;
PROC LOGISTIC data=tfm.happ_train namelen=20 descending ;
       CLASS Peligroso REP_Clima REP_Gobierno REP_Regional_Indicator;
       MODEL happiness_bin = Peligroso REP_Clima--IMP_REP_Prevalencia
                     /selection=backward link=glogit;
RUN;QUIT;
DATA mode;length effect $20. modelo $ 20000;retain modelo " ";set parametros
end=fin;effect=cat(' ',effect);
       if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then output;
RUN;
DATA;set mode;put modelo;run;
       **REP_Ln_PIB_per_capi IMP_REP_Freedom;
ODS GRAPHICS ON;
%randomselectlog(data = tfm.happ_train, /* todas macros logistica 6.0.sas */
                        listclass = Peligroso REP_Clima REP_Gobierno
REP_Regional_Indicator,
                        vardepen = happiness_bin,
                        modelo = Peligroso REP_Clima--IMP_REP_Prevalencia,
                        sinicio = 2000,
                        sfinal = 2200,
                        fracciontrain=0.8,
                        directorio = C:\Users\belen\Documents\UCM\MASTER\TFM
                        );
```

```
      /* 3.3 VCR */
            /* modelo RLog 1: 1° +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = REP_IDH IMP_REP_Freedom,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog1; SET final; modelo=1; RUN;
            /* modelo RLog 2: 2° +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = REP_Ln_PIB_per_capita_ IMP_REP_Freedom,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog2; SET final; modelo=2; RUN;
            /* modelo RLog 3: 3° +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = IMP_REP_Freedom,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog3; SET final; modelo=3; RUN;
            /* modelo RLog 4: 4° +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = REP_Social_support IMP_REP_Freedom,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog4; SET final; modelo=4; RUN;
            /* modelo RLog 5: forward */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = REP_Desempleo REP_IDH IMP_REP_Freedom,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog5; SET final; modelo=5; RUN;
            /* modelo RLog 6: 3 vbles +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = IMP_REP_Freedom REP_IDH
REP_Ln_PIB_per_capita_,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog6; SET final; modelo=6; RUN;
            /* modelo RLog 7: 5 vbles +freq */
%cruzadalogistica(archivo = tfm.happ_train,
                          vardepen = happiness_bin,
                          conti = IMP_REP_Freedom REP_IDH
REP_Ln_PIB_per_capita_ REP_Horas_sol REP_Social_support,
                          categor =,
                          ngrupos = 5, sinicio = 2000, sfinal = 2200);
DATA tfm.final_rlog7; SET final; modelo=7; RUN;

DATA TFM.union_rlog; set tfm.final_rlog1 tfm.final_rlog2 tfm.final_rlog3
                                    tfm.final_rlog4 tfm.final_rlog5
tfm.final_rlog6
                                    tfm.final_rlog7;

RUN;
PROC SORT data=tfm.union_rlog; BY modelo; run;
PROC BOXPLOT data=tfm.union_rlog;
      PLOT media*modelo;
RUN;

DATA union_rlog; set tfm.final_rlog1 tfm.final_rlog2
                                    tfm.final_rlog5
```

```
                                        tfm.final_rlog7;

RUN;
PROC SORT data=union_rlog; BY modelo; run;
PROC BOXPLOT data=union_rlog;
       PLOT media*modelo;
RUN;

PROC MEANS data=tfm.union_rlog mean median var range min max;
       var media;
RUN;
PROC MEANS data=union_rlog mean median var range min max;
       BY modelo;
       var media;
RUN;

PROC LOGISTIC data=tfm.happ_train;
       MODEL happiness_bin = IMP_REP_Freedom REP_IDH REP_Ln_PIB_per_capita_
REP_Horas_sol REP_Social_support;
RUN;
```