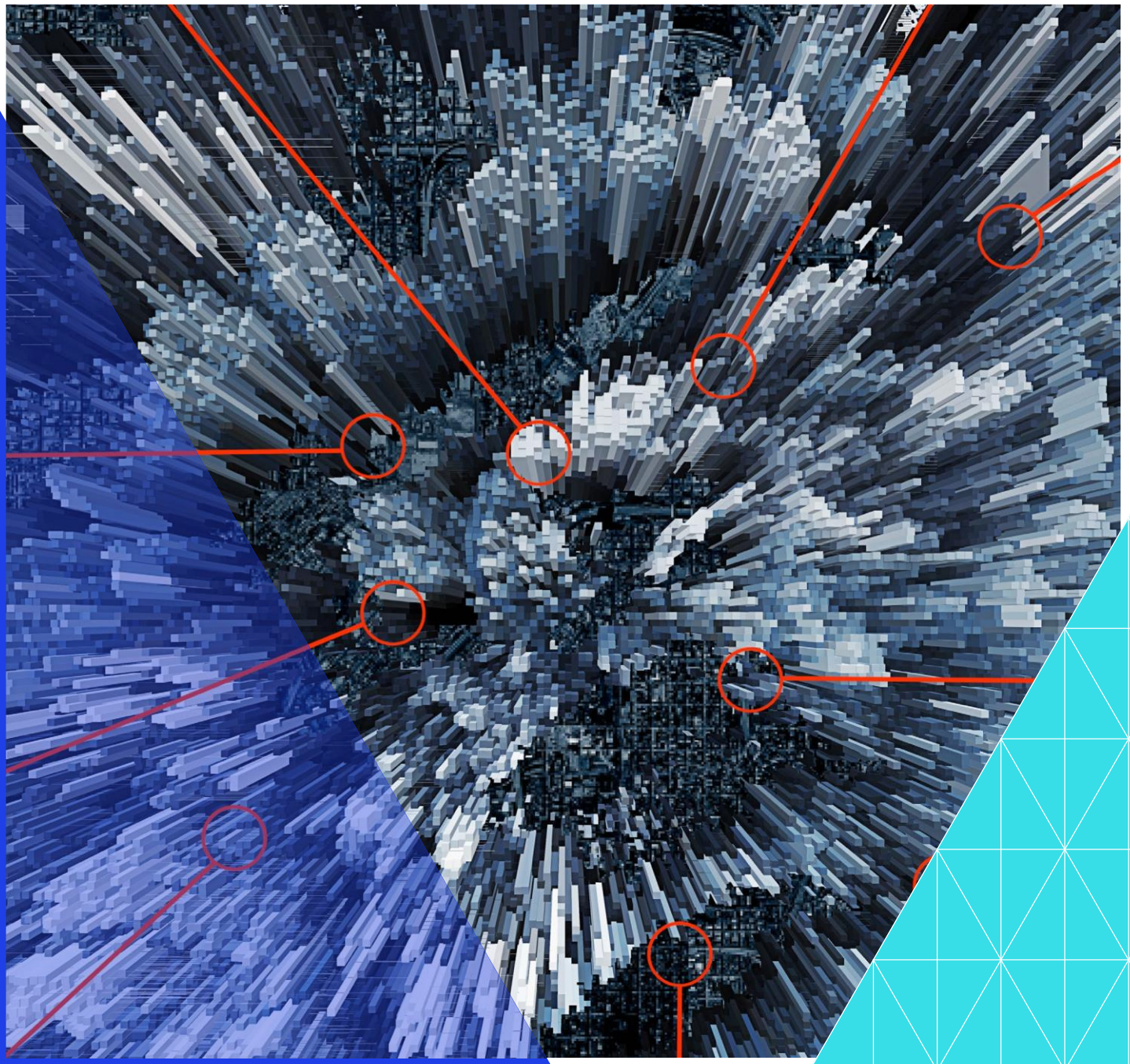


MEMORIA PROYECTO ESPERANZA DE VIDA: DATOS Y RESULTADOS

La Esperanza de Vida desde una Perspectiva Analítica

Autores:
Óscar Marín Esteban
Álvaro Enol Alonso Ortega

Memoria Proyecto Esperanza de Vida: Datos y Resultados

Introducción a el objetivo:

Esta memoria se centrará en seleccionar el modelo estadístico que mejor pueda predecir nuestros datos. Comenzaremos con la limpieza y preparación de los datos, seguida de la verificación de varias hipótesis estadísticas esenciales, como la normalidad, homocedasticidad, independencia o multicolinealidad, entre otras. Estas pruebas son cruciales para establecer un modelo de bondad y facilitar la estimación en nuestros modelos predictivos.

En la siguiente fase, nos enfocaremos en el análisis y comparación de los distintos modelos predictivos, en los que se incluyen la Regresión Lineal Múltiple y Logística, métodos de regularización como Lasso, Ridge y Elastic Net. Y, por último, modelos de reducción de dimensionalidad, como PLS y PCR.

1. Descripción de los datos:

El presente conjunto de datos incluye información recopilada con el propósito de examinar y entender los factores que influyen en la esperanza de vida a nivel mundial. La procedencia de los datos se divide en la OMS, la página web de las Naciones Unidas, datos del Banco Mundial y el proyecto "Our World in Data" proyecto de la Universidad de Oxford. Se recopiló todo en el siguiente conjunto de datos.

Este análisis se basa en 2.864 registros y 21 variables, cada una representando una combinación única de país y año, con un ámbito territorial que se extiende a 179 países y un ámbito temporal que comprende desde el año 2000 hasta el 2015.

Dentro de las variables, tenemos mediciones de varios tipos. La *mortandad en bebés, adolescentes y adultos* se mide por fallecimiento entre cada mil habitantes. El *consumo de alcohol* se mide en litros puros consumidos por mayores de 15 años. Las variables relativas a enfermedades representan el porcentaje de menores de 1 año inmunizados contra ellas, y el *VIH* representa la extensión de dicha enfermedad por cada mil habitantes entre 15 y 49 años. El *BMI* es el índice de masa corporal. El *GDP* es el PIB per cápita. La *población* se mide en millones de habitantes. La *delgadez en menores* se mide a través del BMI. La *tasa de escolarización* se mide por la media de años que los habitantes mayores de 25 de un determinado país pasaron estudiando.

Finalmente, tenemos dos indicadores categóricos binarios que distinguen entre países *desarrollados* y *en desarrollo*, sacados de la clasificación socioeconómica de las naciones según la Organización Mundial del Comercio y ajustados por el Producto Nacional Bruto (PNB) Per Cápita.

La variable respuesta del conjunto de datos es la *esperanza de vida*, una medida continua, resultado de las diferentes variables. Al ser continua, nos permite analizar y aplicar modelos para evaluar las relaciones entre la esperanza de vida y las variables predictoras.

En resumen, este conjunto de datos nos da la oportunidad de analizar los factores que determinan la esperanza de vida a nivel global, con información verificada de diferentes fuentes muy relevantes y fiables.

2. Formulación de preguntas a analizar:

Cuando tenemos un conjunto de datos a analizar, siempre tenemos que formularnos diferentes preguntas para saber qué tenemos que hacer con los datos.

Primero debemos empezar por preguntas sobre la naturaleza de los propios datos: ¿Cuál es nuestra variable respuesta? ¿Las variables presentan una distribución normal? ¿Cómo es la correlación entre las diversas variables, y entre ellas y la variable respuesta? A través de estas preguntas, analizaremos ámbitos tales como la normalidad, la independencia o homocedasticidad, multicolinealidad, el ECM (error cuadrático medio), qué variables son buenas para el modelo, etc.

Una vez hayamos hecho un análisis exhaustivo de nuestro conjunto, llegamos al objetivo final del trabajo, el entrenamiento de modelos. Para cada modelo tenemos unos objetivos o preguntas a responder

- ¿Podemos hacer un buen modelo predictivo de regresión lineal para predecir la esperanza de vida? ¿Cuál es el error de este modelo? ¿Qué podemos hacer para mejorarlo si el error es muy grande? ¿Va nuestro modelo a cumplir normalidad, independencia y homocedasticidad tras pasarlo por validación cruzada? Si no es el caso, ¿qué podemos hacer para que estas estadísticas mejoren?
- ¿Cuál es el ECM en un modelo de regresión múltiple? ¿Cumple con los métodos de regularización más comunes (Ridge, Lasso, etc.)?
- ¿Podemos convertir nuestra variable respuesta en un factor y aplicar un modelo de regresión logística? ¿Qué ECM tiene este modelo? ¿Cumple con los métodos de regularización previos?

Finalmente, nos queda la pregunta principal: ¿Cuál de estos modelos será el mejor para predecir nuestra variable de expectativa de vida?

Todas estas preguntas las vamos a resolver en los distintos puntos que vienen a continuación.

3. Descripción de la depuración/validación de los conjuntos de datos empleados:

Para empezar con la **depuración** de los datos, tenemos que analizarlos y entenderlos. Lo primero que haremos con este propósito será mostrar el encabezado con head y un resumen de cada columna con **summary**.

Lo que podemos ver con esto no es sólo el tipo de los datos contenidos en cada columna, sino también algunos datos generales como el menor y mayor valor de cada una, **los cuartiles, la media y la mediana**. Esto nos permite tener una visión general sobre la naturaleza de los datos. Ahora, veremos si R reconoce cada columna por su tipo de datos correcto.

Todos los datos son reconocidos correctamente, salvo dos, *Economy_status_Developed* y *Economy_status_Developing*. Este dato es de tipo binario, por lo que utiliza 0 y 1 como True o False para definir si un país es un país desarrollado o se encuentra en vías de desarrollo. Sin embargo, en la lectura de los datos se han transformado a tipo entero de forma incorrecta, lo que puede después causar problemas al realizar transformaciones o gráficas. Por tanto, convertimos estas dos columnas a factores utilizando **as.factor**.

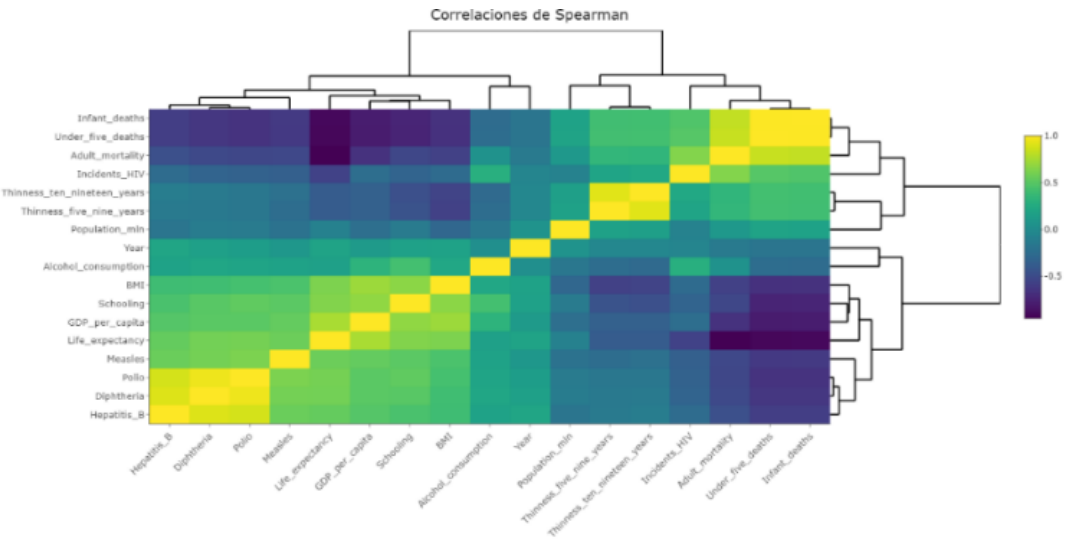
Observación: Tras diversos análisis, nos damos cuenta de un detalle muy importante, estas columnas separan países cuya esperanza de vida se rige según diferentes modelos. Las diferencias entre estos tipos de países hacen que no sea correcto usar el mismo modelo predictivo para calcular la esperanza de vida en países desarrollados y países en vías de desarrollo. Por tanto, vamos a separar el conjunto en dos subconjuntos. Tomaremos el subconjunto de países en vías de desarrollo: no solo contiene más valores que los desarrollados, sino que puede ser más interesante.

Vamos a asegurarnos de que no haya ningún valor NA, eliminando estos datos con **na.omit**. Vamos a analizar ahora la simetría de nuestras variables respuesta con **skewness**, Los valores no simétricos, fuera del intervalo (-2, 2), no presentan una distribución normal. Por lo tanto, no podemos calcular los **outliers** de dichas variables con un boxplot. Vamos a aplicar una transformación **boxcox** a nuestras tres variables asimétricas: *Incidents_HIV, GDP_per_capita* y *Population_mln*.

Finalmente, vamos a **eliminar** los outliers. Los valores atípicos se consideran influyentes a partir de un rango de 3, así que vamos a eliminar estos. Para ello aplicamos un bucle que se repita hasta que no queden outliers a cada columna numérica. Finalmente, podemos aplicar este bucle una vez más a las siguientes variables en este orden: *Hepatitis_B, Diphteria, Polio*. Así, hemos eliminado todos los datos atípicos influyentes de todas nuestras variables.

4. Análisis realizado para responder a cada pregunta formulada:

Tras la depuración y validación de los datos, podemos pasar a analizar nuestro objetivo ya por fin, iremos paso a paso, empezamos con el análisis de las **correlaciones**, que nos puede quitar alguna de las variables ya si está relacionada con otras que no sean la respuesta, nuestra predicción podría verse afectada, utilizaremos **Spearman** porque no nos da normalidad ninguna de las variables, descartando así Pearson.

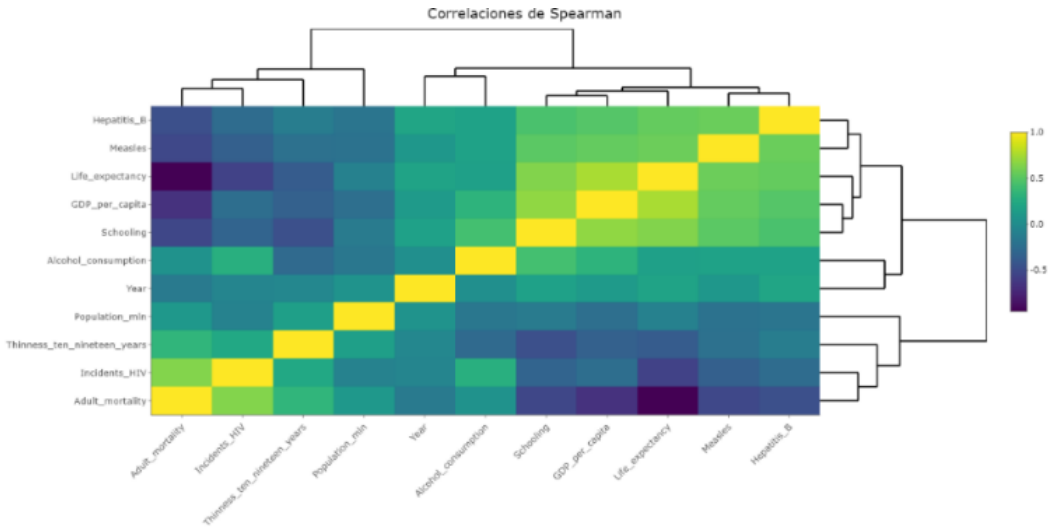


Vemos que hay variables que están altamente relacionadas, las vamos a ir analizando con los tests para saber si es influyente la correlación realmente.

Observación: Al elegir cuál de las dos variables eliminar, se ha ido quitando en base a un balance entre las variables que realmente va a utilizar el modelo, probando las diferentes combinaciones y de qué manera se consiguen quitar menos variables.

Empezaremos comprobando la relación entre las variables *Infant_deaths* y *Under_five_deaths*, en la gráfica ha salido muy alta, pero vamos a aplicarle el test (**cor.test**) con su nivel de significancia. El coeficiente es significativo, ($p - value < 2.2e - 16$), se muestra una correlación alta entre las 2 variables, con valores positivos, por lo que se puede concluir que existe una correlación positiva significativa entre *Infant_deaths* y la *Under_five_deaths*, lo cual tiene sentido, pues se trata de algo muy parecido, por lo que quitamos *Infant_deaths*, en este caso da igual cuál quitar porque cualquiera de las dos las vamos a quitar luego también.

Se han aplicado la misma mecánica a los diferentes **pares** que salen correlacionados, obteniendo así resultados similares: *Adult_mortality* y *Under_five_deaths*, donde quitamos *Under_five_deaths*, *Polio* y *Diphtheria*, quitamos *Diphtheria*, *Polio* y *Hepatitis_B*, quitamos *Polio*, *Thinness_five_nine_years* y *Thinness_ten_nineteen_years*, un caso muy similar al primero que se analizó, se quita *Thinness_five_nine_years*, y, por último, *GPD_per_capita* y *BMI*, se quita *BMI*, comprobamos ahora el **heatmap**.



Sale un gráfico mucho más limpio, en el que vemos una fuerte correlación entre la variable respuesta y *Adult_mortality*, que posteriormente nos ayudará a hacer una mejor predicción, también hemos conseguido **reducir** el número de variables, pero sin perder información importante, ya que al estar correlacionadas una de las variables tiene la variabilidad de la otra.

Vamos a continuar ajustando un modelo con todas las variables válidas que haya, para ver cuáles son válidas para el modelo.

Observación: La finalidad de este apartado es sacar un modelo de regresión lineal múltiple, cuando hablamos de esto, nos referimos a un modelo lineal en el que el valor de la variable dependiente o respuesta, en nuestro caso la esperanza de vida (*Life_expectancy*) se determina a partir de un conjunto de variables independientes llamadas predictores.

Vemos que la variable *Country* y *Region* son de tipo caracter, y no van a tener un impacto significativo en la variable respuesta, ya que el objetivo del proyecto no requiere usarlas, por lo que las podemos quitar, a su misma vez, la variable *Year* no nos va a aportar tampoco nada significativo, no la tendremos en cuenta, pero no la eliminaremos en este caso (veremos más adelante la razón).

Hacemos un **summary** con este primer modelo con todo, nos da un **p-value** en todas menores que 0.05, lo que indica que el modelo está bastante bien, pero siempre se puede mejorar, también nos da una **r cuadrado ajustada** de: 0.9779, lo que es muy bueno, ya que podemos decir que aproximadamente, un 97.79% se puede explicar por el modelo, pero tenemos aquí el problema de **overfitting**, ya que, al tener tantas variables, cuando hagamos un modelo con menos variables bajará.

Para la selección del mejor modelo, haremos un **regSubSet**, con estas variables, e identificaremos qué variables vamos a tener en este primer modelo, las variables incluidas son las siguientes: *Adult_mortality*, *Alcohol_consumption*, *Hepatitis_B, Incidents_HIV*, *GDP_per_capita*, *Population_mln* y *Schooling*, una vez seleccionadas las variables, dividimos el conjunto entre **entrenamiento** (70% en este caso, un total de 1566 registros) y **test** (30%, 672 registros) con una semilla de 5.

Una propiedad que es interesante comprobar, es la **multicolinealidad**, esto ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores, lo que hace que si hay multicolinealidad pequeños cambios en los datos hacen que varíen mucho las estimaciones de coeficientes, por lo que hacemos un **VIF** para comprobarlo, tras su comprobación, vemos que no es causa de preocupación, por lo que continuamos analizando.

Adult_mortality	Alcohol_consumption	Hepatitis_B	Incidents_HIV	GDP_per_capita	Population_mln	Schooling
3.029447	1.416022	1.327378	2.143936	2.628676	1.100965	2.301154

Vamos a pasar **validación cruzada**, que es como probar el modelo de predicción varias veces usando diferentes grupos de datos (**folds**) del mismo tamaño, en el que se van ajustando los folds como un conjunto de validación y estima el **ECM** del conjunto de prueba (test) y comprobamos normalidad, independencia y homocedasticidad, vemos que a priori solo nos da independencia, procedemos entonces a quitar los **outliers** con el método **Leverage** uno a uno, y en por cada uno que quitamos volvemos a pasar validación cruzada, hacemos este proceso y no logramos normalidad ni homocedasticidad, a pesar de quitar los 17 outliers existentes.

Posteriormente, procedemos a hacerle una transformación **boxcox** a la variable respuesta, pasamos validación cruzada de nuevo, y en este caso, la mejora de la normalidad de esta variable con el boxcox es indudable, ahora nos da un **p-value** de 0.43, frente al valor 6.659×10^{-6} que salía al principio, sin embargo, la homocedasticidad no da, ni siquiera ha movido el valor. Por lo que concluimos aquí la comprobación de las propiedades, habiendo hecho todo lo posible por mejorar las condiciones.

Por comentar las demás condiciones, como la **parsimonia**, que hemos venido cuidando desde el principio de este apartado, tratando de explicar con mayor precisión la variable respuesta utilizando el menor número de predictores.

Observación: En esta sección solo se explicará el proceso que se sigue para calcular cada uno, su utilidad y su error, conclusiones que se puedan sacar acerca de los diferentes modelos se harán en el siguiente apartado.

Ahora queremos calcular el Error Cuadrático Medio (ECM) del **modelo de regresión lineal múltiple**, para esto tenemos que **transformar** la variable respuesta de nuevo a la forma original, por así decirlo, quitarle el boxcox, después de hacerlo, tenemos un **ECM** con valor de: 1.306823, pero esto nos presenta el mismo problema que la cuasivarianza, y es que ninguno presenta una medida igual a la cantidad que se intenta estimar, por eso, intentando encontrar similitud con la desviación estándar, tomamos la raíz cuadrada del ECM, para obtener el **RMSE**, que tiene las mismas unidades que la cantidad que se estiman, este valor da: 1.143164.

Una vez hecho este primero, vamos ahora a comprobar los **métodos de regularización,** este enfoque ajusta un modelo con todos los predictores, reduciendo sus coeficientes hacia cero para disminuir la varianza, lo que puede resultar en algunos coeficientes exactamente cero, permitiendo así la selección de variables. Empezamos creando las matrices, que producen una matriz correspondiente a todos nuestros predictores y las transforma en dicotómicas.

Vamos con **Ridge**, con Alpha 0, que minimiza una suma penalizada de residuos y cuadrados de coeficientes, variando los coeficientes según el parámetro lambda. Esto permite un balance entre ajuste y complejidad del modelo, como ya tenemos separado el conjunto y demás, lo único que tenemos que hacer es calcular un modelo y hacer la predicción con la mejor lambda, al predecir nos sale un **RMSE** de: 1.377616.

Aplicamos **Lasso**, mantendremos la estructura anterior, pero con Alpha 1, en este caso penaliza la magnitud de los coeficientes, llevando algunos a cero, y facilita tanto la mejora del modelo como la selección de variables relevantes mediante validación cruzada para elegir la mejor lambda, en este caso nos da un **RMSE** de: 1.204829, un poco mejor que Ridge.

En nuestro caso, **Lasso** nos conviene más, ya que tenemos muchos predictores, y no todos son igual de importantes, por ejemplo, en el caso de *Adult_mortality*, es una variable muy importante para el modelo, además Ridge sabemos que actúa peor en este caso porque hemos quitado las variables que están muy correlacionadas entre ellas.

Para finalizar con regularización, vamos a emplear **Elastic Net**, con un Alpha 0.9, lo que busca este modelo es encontrar un equilibrio entre los dos métodos mencionados, combinando ambas estrategias para seleccionar lo mejor de cada uno, en este caso nos da un **RMSE** de: 1.205443.

Tras hacer estos modelos, vamos a pasar a la **reducción de dimensionalidades**, este enfoque proyecta los predictores en un subespacio de menor dimensión, utilizando combinaciones lineales de las variables para formar nuevos predictores, sobre los cuales se ajusta un modelo de regresión lineal, vamos a realizar los métodos **PCR** y **PLS**.

Empezamos seleccionando las **componentes principales**, que básicamente busca resumir la información contenida en las variables originales, en un conjunto menor que tenga la mayor cantidad posible de la variación de los datos originales.

Ahora pasamos con el **PCR**, este modelo combina PCA y regresión lineal, utilizando las componentes principales como predictores y haciendo un modelo, nos sale un **RMSE** de: 1.126047.

Para terminar con la reducción de dimensionalidades, vamos con el **PLS**, que es bastante parecido al PCR, pero utiliza la variable respuesta para guiar la selección de características, enfocándose en aquellas que son relevantes tanto para los predictores como para la respuesta, nos da un **RMSE** de: 0.6482499.

Por último, veamos que tan bien clasifica nuestro modelo, con la **regresión logística**, la regresión logística es mejor que la lineal para variables binarias porque sus predicciones siempre están entre 0 y 1, rango adecuado para modelar probabilidades, mientras que la regresión lineal puede dar valores fuera de este rango, esto nos va a llevar un poco más de trabajo, ya que tenemos que preparar una variable diferente específicamente para ello, en nuestro caso, vamos a **dividir** la variable respuesta por la mediana, de tal manera que los valores que salgan mayor que la mediana los consideraremos más desarrollados, dentro de los subdesarrollados, con valor 1 y los que salgan menor, serán entonces los que peor desarrollo tienen de todos, con valor 0.

Observación: Podríamos haber limpiado los que tienen *Economy_status_Developed* igual a 1, y ver así los desarrollados y los no desarrollados, de hecho lo intentamos, pero ocasionaba muchos problemas, por lo que decidimos esta medida, que es incluso más interesante, pues si quisiéramos podríamos ver qué países están cerca de estar desarrollados según el año (por esta razón no la eliminamos antes), y hacer comparativas de medidas políticas/económicas/militares que han afectado en subir o mantener la esperanza de vida, aunque se sale de los objetivos del proyecto.

Vamos a tratar la variable como factor, y a dividirla entre **entrenamiento** (70%) y **test** (30%) y creamos un modelo logístico, primero que nada, con **glm**, con familia binomial y con los datos de entrenamiento. Hacemos un **summary** del modelo de nuevo para ver cómo se está comportando. Creamos un **trainControl** con método de **validación cruzada**, que tenga 10 capas, posteriormente entrenamos ese modelo con el trainControl que hemos creado anteriormente, y usando el método de glm.

Ya podemos ver los resultados, para empezar, nos da un **accuracy** del 94.89%, esto significa que, en promedio, el modelo predice correctamente la clase (0 o 1) aproximadamente un 95% de las veces, la **sensibilidad** (o Tasa de Verdaderos Positivos) da 94.33%. Esto significa que el modelo es capaz de identificar correctamente el 94.33% de los casos positivos (clase '0').

La **especificidad** (o Tasa de Verdaderos Negativos) da 94.66%. Esto indica que el modelo identifica correctamente el 94.66% de los casos negativos (clase '1').

5. Conclusiones:

Casualmente vemos que se escogen las mismas variables en todos los modelos, esto puede significar que las variables tienen una relación sólida con la variable de respuesta, no existe mucha multicolinealidad, la penalización en métodos como Lasso o Ridge es leve, y los datos son consistentes.

MODELO	ADULT MORTALITY	ALCOHOL CONSUMPTION	HEPATITIS B	MEASLES	INCIDENTS HIV	GPD PER CAPITA	POPULATION MLN	THINESTEN NINETEEN YEARS	SCHOOLING	RMSE
RLM	✓	✓	✓		✓	✓	✓		✓	1.14316
RIDGE	✓	✓	✓		✓	✓	✓		✓	1.37761
LASSO	✓	✓	✓		✓	✓	✓		✓	1.2048
E-NET	✓	✓	✓		✓	✓	✓		✓	1.2054
PCR	✓	✓	✓		✓	✓	✓		✓	1.12604
PLS	✓	✓	✓		✓	✓	✓		✓	0.6482

El modelo PLS ha demostrado ser superior en el análisis, ofreciendo las predicciones más precisas con el menor RMSE.