Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Matemáticas e Informática

Trabajo Fin de Grado

Análisis de datos de desarrollo humano global

Autor: Alejandro Baltasar Sanz
Tutor(a): Juan Antonio Fernández del Pozo

Madrid, Junio 2024

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Matemáticas e Informática

Título: Análisis de datos de desarrollo humano global

Junio 2024

Autor: Alejandro Baltasar Sanz
Tutor: Juan Antonio Fernández del Pozo
Inteligencia Artificial
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

Resumen

En la actualidad, algunas de las mayores preocupaciones a nivel mundial son la igualdad global, la mejora de la calidad de vida de las personas en países subdesarrollados y su desarrollo sostenible. Esto se recoge en los Objetivos de Desarrollo Sostenible (ODS) del Programa de las Naciones Unidas para el Desarrollo (PNUD).

El presente Trabajo de Fin de Grado pretende realizar un estudio del desarrollo humano a través de tres áreas claves de cada país: características demográficas de la población, características geográficas y desarrollo tecnológico.

Abstract

Currently, some of the biggest global concerns are global equity, the improvement of life quality in underdeveloped countries and their sustainable development. These issues are collected in the Sustainable Development Goals (SDG) of the United Nations Development Programme (UNPD).

This Final Degree Project aims to carry out an study of human development through three key areas of each country: demographic characteristics of the population, geographical features and technological development.

Tabla de contenidos

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Definición del trabajo

El desarrollo humano es un concepto que fue creado por el Programa de las Naciones Unidas para el Desarrollo (PNUD). Este concepto se basa en que el desarrollo no debe medirse únicamente en el factor económico, sino en la calidad de vida de la población. Este Trabajo de Fin de Grado se trata de un análisis y desarrollo de un clasificador del desarrollo global humano. Para esto se analizará un conjunto de datos con distintos indicadores de desarrollo humano recopilados a nivel global por el banco mundial[4]. Se utilizarán técnicas y herramientas avanzadas de análisis de datos, estudiando posibles tendencias, patrones o relaciones para una posterior clasificación de los datos y un mayor entendimiento de la evolución del desarrollo humano en las diferentes regiones del mundo. Este TFG estudiará indicadores de desarrollo tanto económicos, como de salud, tecnológicos y geográficos, para tener una visión global de la calidad de vida de las personas. En resumen, se pretende conseguir una perspectiva detallada del desarrollo global humano de los últimos años.

1.2. Motivación y alcance

Este Trabajo de Fin de Grado permite conocer el desarrollo humano de las distintas regiones del mundo para analizar y clasificar los distintos parámetros de desarrollo urbano y rural, agricultura, salud e infraestructura y cómo afectan a la calidad de vida de la población. Gracias a este conjunto de datos se puede visualizar claramente las desigualdades de población según genero, riqueza o situación geográfica. Conocer las tendencias de desarrollo pasadas resulta de gran utilidad a la hora de predecir posibles corrientes futuras y así contribuir a una posible toma de decisiones en base a datos económicos y de infraestructuras. Personalmente, me resulta de gran interés el poder analizar datos reales con un claro impacto en la sociedad.

1.3. Objetivos

Como ya se adelantó en el Plan de Trabajo, los objetivos generales del TFG son los siguientes:

1. Estudiar los datos y su contexto
2. Dominar los métodos y técnicas de análisis de datos
3. Preparación de los datos
4. Documentación de todo el proceso
5. Estudio y desarrollo de modelos para exploración de datos
6. Estudio y desarrollo de modelos de clasificación, predicción o simulación que aporte valor a los datos

En cuanto a los objetivos específicos, en este trabajo se propone implementar modelos de clasificación[5] que permitan agrupar los distintos países. Se van a implementar varios modelos para poder comparar cual es el que da mayor valor a los datos. A la hora de la exploración de los datos, se ha optado por una serie de histogramas durante todo el proceso de preparación y un estudio de clustering de varios tipos:

1. El algoritmo de clustering de densidad (DbScan)
2. El algoritmo de clustering particional (K-Medoids)

1.4. Descripción del contenido de la memoria

En este capítulo 1 se ha hecho una introducción del Trabajo de Fin de Grado, identificando los objetivos y un resumen del mismo. En el capítulo 2 se hará un estudio del Estado del Arte del conjunto de datos, donde se explicará su origen y definición, junto con un estudio de posibles técnicas y herramientas para resolver el problema. En el Capítulo 3 se realizará la propuesta de análisis, con una explicación del trabajo realizado. Finalmente, en el capítulo 4 se presentarán los resultados y conclusiones del análisis de los datos, incluyendo un apartado con las posibles líneas futuras y continuación del trabajo. Adicionalmente, se incluirá un anexo donde se mostrará el código empleado en R durante todo el análisis.

Capítulo 2

Estado del arte

En esta sección se abordarán distintos puntos del análisis de datos y su importancia. En primer lugar se introducirá el contexto de los datos, donde se explicará el concepto del desarrollo humano, su contexto histórico y los objetivos de este.
En segundo lugar, con un enfoque más técnico, se llevará a cabo un análisis de las herramientas y conceptos a utilizar durante el análisis, incluyendo tipos y técnicas de análisis de datos, clustering y distintos software donde se podría llevar a cabo este trabajo. También se incluirá una breve explicación de algunos paquetes de R utilizados.

2.1. Desarrollo global humano

El desarrollo[6] es un concepto que se ha observado durante todas las épocas de la sociedad y constituye un cambio o evolución. Hay múltiples tipos de desarrollo, como el humano, económico, etc. [7] El desarrollo global humano es un concepto fundamental que se refiere al progreso y bienestar de las personas. Sus principales competencias es la mejora de la calidad de vida y la igualdad de oportunidades en todas partes del mundo.
El objetivo es que todas las personas vivan una vida plena y digna, considerando aspectos más allá del económico, como son la educación, la salud y la igualdad. El propio PNUD (Programa de las Naciones Unidas para el Desarrollo) define este concepto como:
"la expansión de las libertades que tiene la gente para vivir vidas largas, saludables y creativas; para avanzar en otras metas que tienen razones para valorar; y para comprometerse activamente en modelar el desarrollo de forma equitativa y sostenible en un planeta compartido.

Figura 2.1: Gráfico resumen Desarrollo Humano

En la Figura 2.1 se pueden ver cuatro conceptos clave del desarrollo humano, como lo son la armonía con la naturaleza, el enfoque multidimensional, la promoción de la equidad y el enfoque de derechos.

2.1.1. Contexto histórico

Las bases del Desarrollo Humano Global se remontan a los años posteriores a la primera guerra mundial, cuando la comunidad internacional intentaba reconstruir las naciones destruidas y promover su desarrollo. Durante estos primeros años se enfocó dicho desarrollo principalmente en la economía. A esto se le llamó la Teoría Económica del Desarrollo[1] (1945-1957). Durante las décadas posteriores, este desarrollo global fue evolucionando hacia un enfoque mucho más amplio, empezando a tener en cuenta dimensiones más humanas enfocadas en el bienestar. Fue aquí cuando nació el Índice de Desarrollo Humano (IDH), fruto del Programa de las Naciones Unidas para el Desarrollo (PNUD), en la década de los 90. Este fue un hito muy importante, que a día de hoy se sigue utilizando como medida del desarrollo humano.
En este enfoque del desarrollo, se centra a las personas como riqueza de la sociedad, constituyendo su objetivo principal. Es por eso por lo que se busca potenciar las capacidades humanas, para poder optar a más oportunidades y así mejorar la calidad de vida.

Teorías y Enfoques de la Economía del Desarrollo					
Teorías / Décadas	TEORIA DE LA MODERNIZACION	TEORIA ESTRUCTURALISTA	TEORIA NEOMARXISTA	TEORIA NEOLIBERAL	TEORIAS ALTERNATIVAS
'50					
'60					Necesidades Básicas
'70					Desarrollo Multidimensional
'80					Desarrollo Autónomo
'90					Desarrollo Humano

Figura 2.2: Distintas teorías de desarrollo durante los años [1]

En la Figura 2.2 se pueden ver las distintas teorías relacionadas con el desarrollo humano desde las post-guerra (años 50), hasta alcanzar la tendencia actual en los años 90.

2.1.2. Programa de las Naciones Unidas para el Desarrollo (PNUD)

[8]El PNUD es una agencia de las Naciones Unidas enfocada en el desarrollo internacional, que trabaja en 170 países y territorios buscando la igualdad y erradicar la pobreza. El objetivo de esta división es ayudar a los países donde trabajan a lograr los Objetivos de Desarrollo Sostenible (ODS). Sus 3 principales competencias son:

1. Desarrollo sostenible
2. Gobernanza democrática y consolidación de la paz
3. Resiliencia climática y ante desastres

El PNUD nace de la combinación del Programa Ampliado de Asistencia Técnica de las Naciones Unidas (1949) y el Fondo Especial de las Naciones Unidas (1958). Se estableció en 1965 por la Asamblea General de las Naciones Unidas. (web PNUD)

Objetivos de Desarrollo Sostenible (ODS)

[2]Los Objetivos de Desarrollo Sostenible u Objetivos Globales nacen en 2015 en las Naciones Unidas con el objetivo de poner fin a la pobreza, proteger el planeta, con la gran meta de que, para el 2030 todas las personas disfruten de paz y prosperidad (web ODS). Los ODS pretenden acabar con la pobreza, el sida, el hambre y la discriminación de mujeres y niñas, por medio de la creatividad, el conocimiento, la tecnología y recursos financieros.

Figura 2.3: Objetivos de desarrollo sostenible (ODS) [2]

Como se puede ver en la Figura 2.3, existen 17 Objetivos de Desarrollo Sostenible:

1. Fin de la pobreza
2. Hambre cero
3. Salud y bienestar
4. Educación de calidad
5. Igualdad de género
6. Agua limpia y saneamiento
7. Energía asequible y no contaminante
8. Trabajo decente y crecimiento económico
9. Industria, innovación e infraestructura
10. Reducción de las desigualdades
11. Ciudades y comunidades sostenibles
12. Producción y consumo responsables
13. Acción por el clima
14. Vida submarina
15. Vida de ecosistemas terrestres
16. Paz, justicia e instituciones sólidas
17. Alianza para lograr los objetivos

2.1.3. Índice de Desarrollo Humano (IDH)

Como se ha explicado anteriormente, el IDH es el baremo de referencia frente al desarrollo humano y refleja las desigualdades entre regiones. Este índice permite crear un ranking anual de los países, donde se puede conocer tanto el desarrollo de un país como comparar con el resto de países, permitiendo a los gobiernos comprender sus opciones de crecimiento y orientando a los órganos globales cómo proceder con las ayudas internacionales [9]. El IDH nace en 1990, cuando nace también la publicación anual del Informe sobre Desarrollo Humano. Las principales variables con las cuales se cálcula el IDH son:

1. **Salud:** se evalúa la esperanza de vida.
2. **Educación:** mide la media de los años de escolarización de los adultos y la esperanza de vida escolar de los niños.
3. **Economía:** se mide a través del Ingreso Nacional Bruto (INB) per cápita.

El cálculo del IDH se ejecuta a través de la media geométrica de estas tres consignas, devolviendo un valor entre 0 y 1. Este cálculo arroja la siguiente división entre países:

1. **Muy alto:** IDH >0,8
2. **Alto:** 0,7 <IDH <0,8
3. **Medio:** 0,55 <IDH <0,7
4. **Bajo:** IDH <0,55

Sin embargo, dado el amplio significado del Desarrollo Humano Global, este índice no es suficiente para medir el desarrollo de cada país, por lo que existen más índices para estudiar este campo con más exactitud:

1. **IDH ajustado por desigualdad (IDH-D):** donde el IDH reflejaría lo que se podría alcanzar en caso de no haber desigualdad.
2. **Índice de desigualdad de género (IDG):** donde se refleja la salud reproductiva, empoderamiento y participación en el empleo de la mujer.
3. **Índice de Pobreza Multidimensional:** donde también se tiene en cuenta las carencias de los hogares y sus habitantes en diversos ámbitos como la salud, el nivel de vida o la educación.

En conclusión, el Índice de Desarrollo Humano es el baremo utilizado por las Naciones Unidas para cuantificar el desarrollo en cada país, siendo el objetivo aumentar este índice.

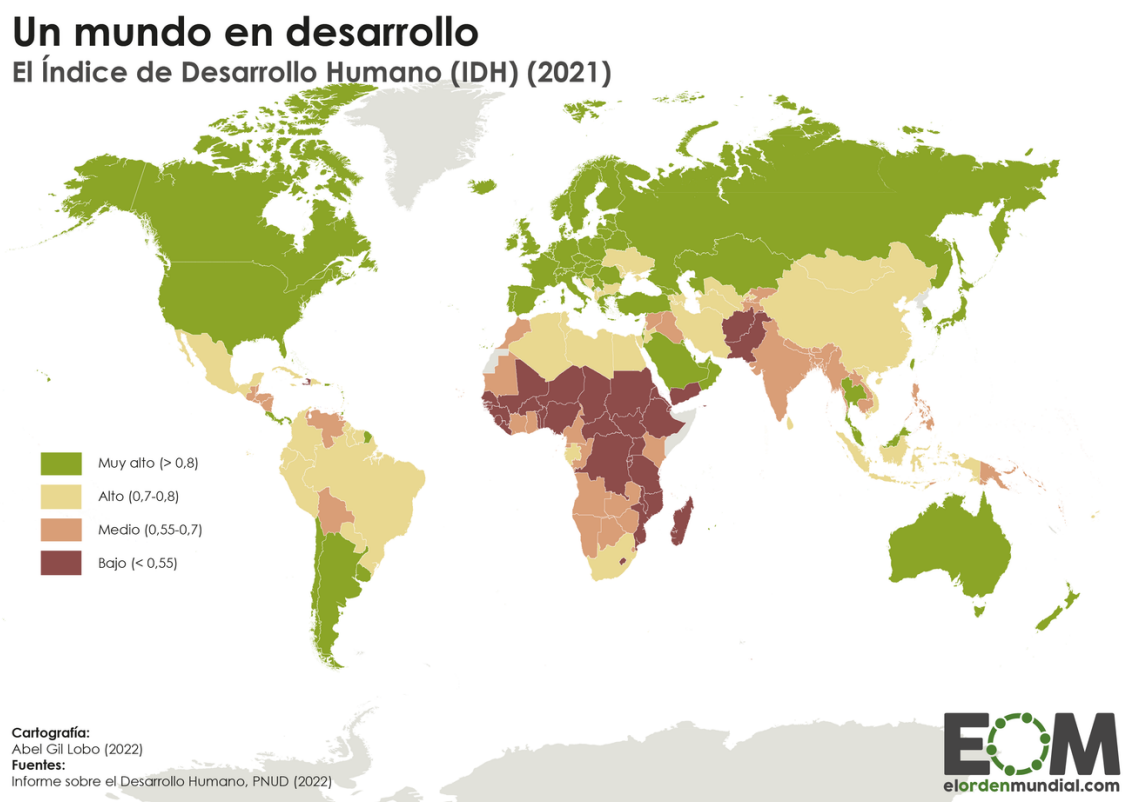

Figura 2.4: Clasificación países según IDH en 2021 [3]

En la Figura 2.4 aparece un mapa del mundo donde los países están pintados según su IDH, los registros más bajos se encuentran mayoritariamente en África.

2.1.4. Informes sobre Desarrollo Humano

Los informes sobre desarrollo humano son elaborados por el PNUD desde 1990, anualmente, con el objetivo de reflejar la situación mundial con respecto al desarrollo humano. Más allá de recoger el IDH de cada país, el informe también presenta un análisis para desarrollar posibles iniciativas y cuestiona las actividades actuales para conseguir el objetivo de desarrollo. Estos informes sirven como sustento y tienen una gran influencia en los debates de desarrollo global.[10].

Situación actual y último informe sobre Desarrollo Humano

Desde hace unos años atrás, se observa una tendencia en los informes donde se produce una polarización del IDH y hay una creencia que afirma que ahora más que nunca, la actividad humana tiene más peso en el porvenir del planeta [10]. Reafirmándose en el último informe presentado por el PNUD del año 2023-2024 donde la gran conclusión es que los países ricos han aumentado su IDH, mientras que los pobres están retrocediendo, debido a la gran polarización política y a la desconfianza. [11]

El IDH descendió en 2020 y 2021 por primera vez en su historia, en gran parte debido a la pandemia del COVID-19. Sin embargo, en 2023 se ha recuperado, alcanzando máximos históricos, pero con una creciente polarización entre los países con el IDH muy alto y muy bajo.

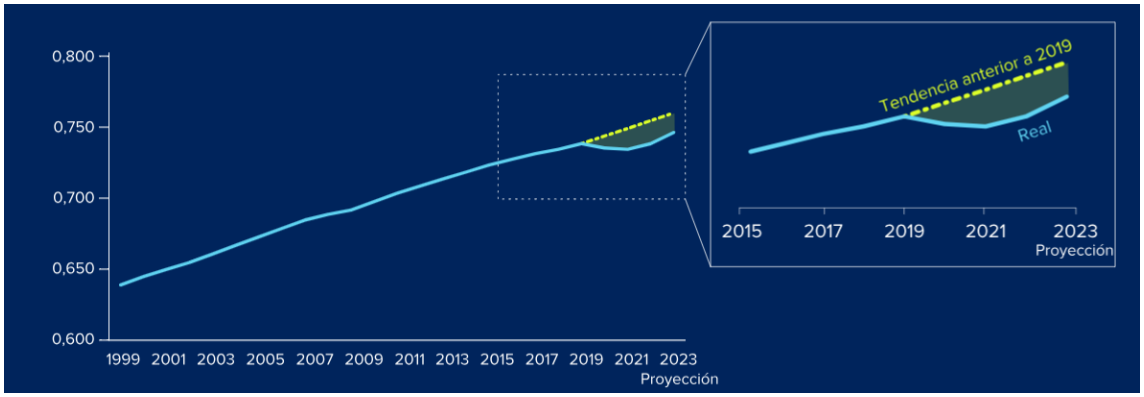

Figura 2.5: Evolución del IDH mundial durante el siglo XXI

En la Figura 2.5 se puede observar como a partir del 2019, se ha perdido la tendencia anterior de los últimos 20 años, recuperándose a partir del 2022.
Se proponen en el informe distintas soluciones para contrarrestar este estancamiento:

1. **Construir una arquitectura del siglo XXI para los bienes públicos mundiales:** el objetivo de esta solución es poder hacer llegar a los países más pobres los bienes públicos mundiales de los que todos dependemos. Esta vía de cooperación se uniría a las otras dos ya existentes: la asistencia para el desarrollo centrada en los países más pobres y la ayuda humanitaria centrada en las emergencias.

2. **Reducir la crispación y hacer retroceder la polarización:** las dos principales maneras de reducir la polarización son proporcionar bienes públicos y fomentar la confianza interpersonal, que podría lograrse gracias a espacios deliberativos como asambleas ciudadanas. La cooperación internacional es clave en este problema.

3. **Reducir las brechas de la capacidad de actuación:** para lograr esto se requiere un enfoque centrado en las personas y su desarrollo. Siguiendo la línea anterior, se precisa la cooperación para poder enfocarse en lo que podemos lograr juntos.

2.2. Análisis teórico

En esta segunda parte del estado del arte, se analizarán en profundidad los distintos métodos y técnicas del análisis de datos.

2.2.1. Preparación de los datos

La preparación de los datos es el primer paso antes de realizar su análisis. Es vital trabajar con unos datos limpios y bien estructurados para obtener unos resultados más relevantes. Pese al gran costo de tiempo que supone la preparación, permite llevar acabo un análisis estadístico más eficaz y con mejores interpretaciones de los resultados. Los datos en bruto pueden contener valores nulos o incorrectos, lo que puede mermar la obtención de los resultados. Aplicar técnicas como la normalización o discretización puede ser también interesante para obtener un conjunto de datos con valores más visibles y que su manipulación sea más sencilla.
Las técnicas y análisis a aplicar dependerá tanto del conjunto de datos como del objetivo principal de su análisis.

2.2.2. Análisis de datos

A continuación se explicarán distintas técnicas del análisis de datos.

2.2.3. Análisis descriptivo

Se pretende examinar y describir las características del conjunto de datos, para explorar posibles relaciones entre variables y entender los datos gracias a conceptos como las medidas de tendencia central o de dispersión.

2.2.4. Análisis exploratorio

El análisis exploratorio [12] es la fase inicial del análisis de datos donde se pretende examinar a fondo los datos para su mejor comprensión, lo que es de vital importancia antes de realizar análisis más complejos. Las técnicas utilizadas en este análisis suelen ser métodos de visualización de datos y técnicas de agrupación. Estas técnicas darán información sobre qué clase de estudio será el más efectivo para los datos en cuestión y sobre cómo abordarlo.
En cuanto a las técnicas de agrupación, en este trabajo se emplearán tres métodos de clustering [13], que son DbScan, K-Medoids y Jerárquico. Se realizará un estudio de resultados de los tres para concluir cuál es el que mejor funciona con los datos y hacer un estudio de su estructura interna.

2.2.5. Análisis inferencial

Este análisis se basa en extrapolar la información de una pequeña muestra de los datos al conjunto entero de los mismos con métodos como estimación por intervalos de confianza, contraste de hipótesis o análisis de la varianza.

2.2.6. Clasificación, predicción y simulación

En el análisis inteligente de datos y aprendizaje automático existen 3 enfoques para obtener conocimiento de los datos:

1. Clasificación supervisada

2. Predicción y regresión
3. Simulación da datos según la distribución conjunta y sus patrones de evolución espacio temporal

En primer lugar, para explorar los datos se consideran técnicas de clustering que ayudan a extraer información de la estructura interna de los datos.

Clasificación y predicción

Con esta técnica se pretende emplear algoritmos de aprendizaje automático para realizar una clasificación de los datos. Para desarrollar esta técnica se emplean datos previamente clasificados para entrenar el modelo y después poder aplicar los mismos patrones a los datos objetivo. Algunos de los algoritmos más comunes son el de regresión logística, árboles de decisión o Redes neuronales. Con esto se consigue predecir futuros valores en base a patrones o tendencias pasadas.

Simulación

Esta técnica pretende simular el comportamiento futuro de un sistema basándose en condiciones iniciales o pasadas. Es especialmente interesante a la hora de investigar posibles sucesos en base a distintos escenarios, por lo que es una herramienta muy útil a la hora de la toma de decisiones.

2.3. Técnicas de análisis

En esta sección se expondrán algunas técnicas y medidas de análisis que se utilizarán en el trabajo.

2.3.1. Medidas de análisis

Media

La media es una medida estadística que refleja el valor numérico central de un conjunto de valores. Se trata de una medida de tendencia central y se calcula sumando todos los valores y dividiendo el resultado entre el número de valores: $\frac{X_1+X_2+...+X_n}{n}$
Es una medida muy útil y popular en el análisis de datos. En este trabajo se utilizará para la normalización de los datos, explicada más adelante.

Mediana

La mediana es otra medida de tendencia central que refleja el valor numérico y ordinal que se encuentra en el centro de un conjunto de valores. Al contrario que la media, la mediana no es sensible a valores extremos o atípicos, por lo que se convierte en una medida muy útil si tienes valores muy extremos.

Desviación típica

La desviación típica o desviación estándar nos da información acerca de la variabilidad y dispersión de un conjunto de datos. Cuanto mayor sea la desviación típica, mayor será la dispersión de los datos en relación a la media. Por el contrario, cuanto más similares sean los datos, estarán más cerca de la media y por lo tanto la desviación típica será más cercana a 0, siendo 0 cuando todos los valores sean iguales.

Correlación lineal

La correlación es una medida estadística que mide la dependencia lineal de una variable respecto a otra, es decir, mide el nivel de relación de dos variables. Los valores de la correlación oscilan entre -1 y 1, siendo -1 una correlación negativa perfecta, 1 una correlación positiva perfecta y el 0 indica que no hay correlación. La correlación positiva indica que cuando una variable aumenta, la otra también tiende a aumentar y viceversa. Sin embargo, la correlación no implica que si una variable cambia, la otra tenga que cambiar necesariamente, únicamente mide la relación estadística entre ellas. La correlación requiere variables numéricas. Una medida más general de dependencia es la información mutua[14], que se explicará más adelante.

Normalización

A la hora de comparar variables o interpretar los datos, la normalización es una herramienta fundamental. El propósito de ésta es ajustar los valores a una escala común para un uso más sencillo de los datos y para prevenir la dominancia de variables que pueden opacar a otras por las escalas utilizadas. En concreto, en este trabajo se utilizará la estandarización, que se logra tras aplicar la siguiente operación matemática: $z = \frac{X-\mu}{\sigma}$, siendo:

1. z el valor obtenido estandarizado.
2. X el valor original.
3. μ la media de la variable.
4. σ la desviación típica estándar de la variable.

Con esta operación logramos una media de 0 y desviación típica estándar 1, realizando una transformación lineal a los datos que no altera la distribución de probabilidad y permite que la descriptiva sea eficaz, las gráficas son mejores y el análisis cluster define mejor los grupos.

2.3.2. Técnicas de análisis

Discretización

La discretización es el proceso por el que se convierten las variables de continuas a discretas. Esto es útil para simplificar los datos y así hacer que su análisis sea mas sencillo. La discretización consiste en dividir el conjunto de datos en un

número n de intervalos y asignar todos los valores a esos intervalos. Como la discretización supone pérdida de alguna información, hay que elegir bien los intervalos que dividirán los datos para minimizarla. En este trabajo se empleará esta técnica para hacer operaciones posteriores como clustering o el cálculo de la entropía.

Naturaleza de los datos

En el análisis de datos es esencial trabajar con un conjunto de datos manejable. Para ello hay que elegir bien los datos con los que vas a trabajar. En este trabajo en concreto, se considerará hacer los datos atemporales y mantener un único registro por país, lo que hará que el conjunto de datos sea tratable, simplificando su análisis, aunque se pierda la información de dependencias temporales y espaciales que los datos de desarrollo humano recogen.

Selección de variables

A la hora de realizar algoritmos en el contexto del aprendizaje automático o clasificación, se puede producir overfitting debido a la excesiva complejidad de los datos. Este problema surge a raíz de una adaptación demasiado buena a los datos de entrenamiento, donde no se captan únicamente las tendencias generales, sino también el ruido, imposibilitando una buena adaptación a datos nuevos. Esto puede suceder por tener demasiados parámetros en los datos.

Otro problema asociado a esto puede ser la complejidad computacional, que puede suponer tiempos de ejecución demasiado altos. Para solventar estos problemas se puede emplear la selección de variables[15], que supone una disminución del número de variables, donde el problema reside en la correcta elección de las variables más relevantes para así disminuir en gran medida los problemas causados por la alta complejidad, pero al mismo tiempo, perder la menor cantidad de información posible.

Los principales métodos de selección de variables son los siguientes:

Métodos de filtro

Estos métodos se basan en elegir variables en función de características estadísticas de la variable en relación a la variable objetivo. Algunas de estas características pueden ser:

1. Correlación
2. Información mutua
3. Análisis de varianza

Métodos de Wrapper

Consisten en entrenar y evaluar el modelo para varios subconjuntos de variables hasta alcanzar el que mejor resultados obtiene. Esta variación de subconjuntos

se consigue tras añadir/eliminar variables del conjunto constantemente. Este tipo de método de selección de variables suele lograr resultados más precisos pese a que tienen un coste computacional significativamente más alto.

Análisis de Componentes Principales

El análisis de componentes principales o PCA [16] es una técnica de aprendizaje no supervisado. Una de sus aplicaciones más populares es la reducción de dimensionalidad, donde se busca disminuir el número de variables perdiendo la menor cantidad de información posible (varianza).

Cada variable generada por el PCA será una combinación lineal de variables originales y tendrán la menor correlación posible entre sí. Es una herramienta muy útil para la visualización de los datos, ya que se pueden reducir las variables a dos componentes del PCA para poder graficar los datos.

2.4. Herramientas del análisis

A continuación se explicarán algunas de las opciones de herramientas de análisis de datos que se han planteado para desarrollar este trabajo.

2.4.1. Python

Python [17] es un lenguaje de alto nivel y bastante intuitivo. Se utiliza para todos los campos de la programación como la orientada a objetos o funcional. Python es uno de los lenguajes mas versátiles y populares para el análisis de datos. Su sintaxis clara, una inmensa variedad de bibliotecas y paquetes y su facilidad de aprendizaje hacen que Python sea una excelente opción para este propósito. Algunas de las bibliotecas más populares para los analistas de datos son Pandas[18], NumPy [19] o Matplotlib [20].

2.4.2. R

R [5] es uno de los lenguajes más populares para análisis de datos, gracias a su amplia oferta de paquetes y librerías y enfoque específico en la manipulación y visualización de datos. Este lenguaje abierto se desmarca del resto de lenguajes por la gran variedad a la hora de graficar los datos, pudiendo visualizar datos multivariables con total personalización visual. Aunque tiene algunas desventajas respecto a Python, como puede ser su sintaxis algo más compleja y su menor rendimiento, este es el lenguaje que se ha elegido para realizar este trabajo.

2.4.3. Microsoft Power BI

Microsoft Power BI [21]es un software simple con el que se puede convertir todo tipo de información no relacionada en información coherente y fácilmente visualizable. Esta herramienta es muy flexible en cuanto al origen de datos, ya que acepta tanto tablas de Excel, como archivos de texto, entre otros. Una de las

ventajas de este software es la capacidad de visualización de los datos, a través de variedad de gráficos, como la posibilidad de crear un gráfico personalizado a través de bibliotecas.

2.4.4. Paquetes de R

A continuación se van a mencionar algunos paquetes utilizados en este trabajo con una breve explicación de cada uno:

1. **ggplot2 (3.4.4)** Es una potente librería enfocada en la creación de gráficos. Tiene una amplia variedad de gráficas como histogramas o diagramas de dispersión y una gran opción de personalización.[22]
2. **dplyr (1.1.3)** Es una biblioteca diseñada para facilitar la manipulación y transformación de los datos. Proporciona funciones intuitivas para tratar los datos y realizar operaciones como filtrar, agrupar o resumir. En específico se ha usado la función summarize() para realizar una agrupación de los datos.[23]
3. **gridExtra (2.3)** Este paquete facilita la creación y manipulación de gráficos para organizarlos, por ejemplo, en cuadrículas. Se suele combinar con el paquete ggplot2. Específicamente en este trabajo se utiliza la función grid.arrange() para organizar las gráficas producidas por ggplot2 en cuadrículas para un estudio más sencillo y eficiente.[24]
4. **fpc (2.2.10)** Este paquete contiene funciones para evaluar y estudiar la precisión de la clasificación o clustering. Contiene funciones que proporcionan índices que dan valor al clustering previamente realizado. También se utiliza para realizar el clustering, ya que tiene funciones que seleccionan el número óptimo de clusters en los que dividir los datos.[25]
5. **infotheo (1.2.0.1)** Esta librería implementa varias medidas de información basadas en la entropía. En concreto se ha utilizado la función entropy de este paquete.[26]

Capítulo 3

Análisis Exploratorio

En este tercer capítulo se llevará a cabo una explicación de la base de datos que se ha utilizado para el análisis, así como su diccionario de datos. También se desarrollará la preparación de datos empleada y un análisis exploratorio con clustering y gráficos.

3.1. El conjunto de datos

El conjunto de datos contiene información respecto al desarrollo humano entre los años 1980 y 2013. Gracias a la base de datos se puede seguir el desarrollo de un total de 146 países y 43 agrupaciones de países, donde encontramos agrupaciones geográficas de países ("Arab World", "World", "Euro Area", "Caribben Small States"), agrupaciones de países pertenecientes a otros grupos ("OECD Members", "European Union") y agrupaciones por niveles económicos ("Low and middle income"," High income", "Least developed countries: UN classification").

La base de datos consta de 25 variables, entre las que encontramos el país, el año y otras 23, referidas a diferentes puntos de vista del desarrollo del país, relativas a datos tanto geográficos como demográficos o tecnológicos. Como se ha explicado en la motivación del trabajo, el análisis de estos datos es de gran importancia para conocer el impacto de la situación geográfica, economía y otras casuísticas en la población. Con un análisis exitoso se podría predecir las futuras tendencias de cada región e incluso definir posibles soluciones para cambiar dichas situaciones.

Este conjunto de datos se encuentra en la plataforma kaggle [27], que es una plataforma abierta donde podemos encontrar una gran variedad de conjuntos de datos de un amplio abanico de temas.

3.1.1. Origen del conjunto de datos

Estos datos han sido recogidos por el Banco Mundial[4], por lo que podemos hallar en su plataforma todo tipo de datos acerca de nuestro dataset y otros

de ramas parecidas [4] . Concretamente podemos encontrar un glosario [28] de variables con toda la información de las utilizadas.

3.1.2. Diccionario de datos

A continuación, se realizará una explicación de cada variable con su correspondiente formato, unidades de medida y un ejemplo:

1. **Country**: (Cadena de caracteres) País o agrupación de países (Çanada").
2. **Year**: (Entero) Año (1980).
3. **Data.Health.Birth Rate**: (Decimal) Número de nacimientos durante el año por cada 1000 habitantes (medidos a mitad de año) (15.4).
4. **Data.Health.Death Rate**: (Decimal) Número de muertes durante el año por cada 1000 habitantes (medidos a mitad de año) (7.0).
5. **Data.Health.Fertility Rate**: (Decimal) Número de hijos que tendría cada mujer si viviese durante todo su periodo fértil basado en la tasa de fertilidad por edad del año específico (1.754).
6. **Data.Health.Life Expectancy at Birth.Female**: (Decimal) Número de años que una mujer recién nacida viviría si los patrones de mortalidad del momento en que nació se prolongasen durante toda su vida (78.59).
7. **Data.Health.Life Expectancy at Birth.Male**: (Decimal) Número de años que un hombre recién nacido viviría si los patrones de mortalidad del momento en que nació se prolongasen durante toda su vida (71.32).
8. **Data.Health.Life Expectancy at Birth.Total**: (Decimal) Número de años que un recién nacido viviría si los patrones de mortalidad del momento en que nació se prolongasen durante toda su vida independientemente de su género (74.86).
9. **Data.Health.Population Growth**: (Decimal, Porcentaje). Tasa exponencial de crecimiento de la población a mediados de año desde el año pasado. Se entiende como población a todos los residentes independientemente de su estatus legal o ciudadanía (0.997).
10. **Data.Health.Total Population**: (Decimal) Número total de residentes ese año (24277000.0).
11. **Data.Infrastructure.Mobile.Cellular Subscriptions**: (Decimal) Número de suscripciones telefónicas a servicios telefónicos públicos (12000).
12. **Data.Infrastructure.Mobile Cellular Subscriptions per 100 People**: (Decimal) Número de suscripciones telefónicas a servicios telefónicos públicos por cada 100 habitantes (0.27337385).
13. **Data.Infrastructure.Telephone.Lines**: Variable eliminada del conjunto de datos por redundancia .

14. **Data.Infrastructure.Telephone.Lines.per.100.People**: Variable eliminada del conjunto de datos por redundancia .

15. **Data.Rural Development.Agricultural.Land**: Kilómetros cuadrados de tierra que es cultivable, con cultivos y pastos permanentes. También se incluyen los cultivos temporales, prados para la siega o para pasto, áreas destinadas al mercado como huertas y tierras en barbecho. Se entiende como cultivos permanentes a aquella tierra que está cultivada durante largos periodos de tiempo sin necesidad de ser replantada después de cada cosecha, como por ejemplo, el cacao o el café. Los pastos permanentes son tierras utilizadas durante al menos 5 años como forraje, incluyendo cultivos naturales y cultivados (669030.0).

16. **Data.Rural Development.Agricultural.Land.Percent**: (Decimal, Porcentaje) Porcentaje de Land Area de tierra que es cultivable, tanto con cultivos permanentes como temporales (7.35722509789949).

17. **Data.Rural Development.Arable.Land**: (Decimal) Hectáreas persona de tierra cultivable con cultivos temporales. Se incluyen los cultivos temporales, prados para la siega o para pasto, áreas destinadas al mercado como huertas y tierras en barbecho. Las tierras abandonadas a causa del cultivo itinerante quedan excluidas (1.82782057091074).

18. **Data.Rural Development.Arable.Land.Percent**: (Decimal) Porcentaje de Land Area de tierra cultivable con cultivos pertenecientes a Arable Land (4.8797439052687).

19. **Data.Rural Development.Land.Area**: (Decimal) Kilómetros cuadrados de tierra del país, excluyendo zonas bajo masas de agua en interior (riós, lagos), zonas nacionales de plataforma continental y zonas exclusivamente económicas (9093510.0).

20. **Data.Rural Development.Rural.Population**: (Decimal) Número de personas que viven en zonas rurales (5918004).

21. **Data.Rural Development.Rural.Population.Growth**: (Decimal) Porcentaje de crecimiento anual de personas que viven en zonas rurales (0.833711883207287).

22. **Data.Rural Development.Surface.Area**: (Decimal) Kilómetros cuadrados de tierra del país incluyendo zonas de agua de interior (ríos, lagos) y algunas zonas de la costa (9984670.0).

23. **Data.Urban Development.Population.Density**: (Decimal) Número de personas por kilómetro cuadrado de Land Area. Se calcula dividiendo la población a mitad de año entre Land Area (2.66970619705702).

24. **Data.Urban Development.Urban.Population.Percent**: (Decimal) Porcentaje de personas que viven en zonas urbanas respecto al total de la población (75.623).

25. **Data.Urban Development.Urban.Population.Percent.Growth**: (Decimal) Porcentaje de crecimiento de población urbana durante un año (1.05057823382459).

3.2. Preparación de los datos

El análisis del conjunto de datos comenzó con su preparación para una manipulación más sencilla y efectiva. Como ya se ha explicado previamente, los datos se encuentran originalmente agrupados por país y año. De cada registro hay distintas variables recogidas en 4 temáticas distintas:

1. **Data Health**. Datos sobre la salud.
2. **Data Infrastructure**. Desarrollo tecnológico.
3. **Data Rural Development**. Desarrollo del territorio rural.
4. **Data Urban Development**. Desarrollo del territorio urbano.

3.2.1. Simplificación de los datos

En primer lugar, se decidió prescindir de las variables

1. Data.Infrastructure.Telephone Lines
2. Data.Infrastructure.Telephone Lines per 100 People

Debido a que resultaban redundantes respecto a las mismas variables con **cellular subscriptions** y no aportaban más información.

Después se cambió el nombre de las variables para acortar su longitud, cambiando Data.Health por **H** Data.Infrastructure por **I** Data.Rural.Development por **RD** y Data.Urban.Development por **UD**. Obteniendo la siguiente lista de variables, que serán las que se utilizará en el trabajo:

1. H.Birth.Rate
2. H.Death.Rate
3. H.Fertility.Rate
4. H.Life.Expectancy.at.Birth.Female
5. H.Life.Expectancy.at.Birth.Male
6. H.Life.Expectancy.at.Birth.Total
7. H.Population.Growth
8. H.Total.Population
9. I.Mobile.Cellular.Subscriptions
10. I.Mobile.Cellular.Subscriptions.per.100.People
11. RD.Agricultural.Land
12. RD.Agricultural.Land.Percent
13. RD.Arable.Land
14. RD.Arable.Land.Percent

15. RD.Land.Area
16. RD.Rural.Population
17. RD.Rural.Population.Growth
18. RD.Surface.Area
19. UD.Population.Density
20. UD.Urban.Population.Percent
21. UD.Urban.Population.Percent.Growth

Como se ha explicado anteriormente, aunque se han eliminado los registros, el conjunto de datos no solo contenía registros de países, sino conjuntos de los mismo, como por ejemplo:

1. Arab World
2. Low & middle income
3. South Asia
4. Caribbean small states

Los datos abarcan una línea temporal desde 1980 hasta 2013, es decir, 34 años. Esta condición podría no ser óptima a la hora de realizar el análisis clasificatorio debido a la posible disparidad de datos durante esos 34 años. Las variables más afectadas podrían ser las de infraestructura tecnológica y desarrollo de núcleos urbanos y rurales. Debido a esto, se tomó la decisión de limitar los datos a los correspondientes a la última década, es decir, desde 2004 hasta 2013. Con esto ganamos robustez en los datos y un análisis más claro y preciso de la situación de cada país a día de hoy, reflejando con mayor precisión las condiciones y realidades actuales.

A su vez, con el paso de los años, el desarrollo tecnológico permite una mayor capacidad para recopilar datos, por lo que podemos asumir que los actuales serán más precisos que aquellos tomados hace tres décadas. Como es evidente, otra razón por la que se tomó esta decisión fue por una cuestión de manejabilidad de los datos, ya que se redujo notablemente su volumen, permitiendo así un manejo más efectivo para un análisis más riguroso.

Una vez con un conjunto de datos relativo a los diez últimos años, se decidió transformar el conjunto a uno atemporal,es decir, eliminar la variable *Year*, relativa al año, por diversas razones. La principal, por el modelo de análisis elegido, uno clasificatorio para el que no necesitamos la temporalidad de los datos. Esta sería necesaria para una posible predicción futura, que podría ser una posible línea futura de este Trabajo de Fin de Grado, que se detallará en un próximo capítulo. Para realizar este ajuste se ha utilizado el estadístico media. Ya que todas las variables son numéricas, se ha aplicado la media a los diez valores existentes por país obteniendo así un único registro por cada uno de ellos.

3.2.2. Estandarización

El último paso previo a obtener el dataset final con el que se trabajará es la estandarización. Este tratamiento es vital por varias razones: en primer lugar, para conseguir datos adimensionales y en una escala común, ya que se facilitarán las comparaciones y permite un mayor rendimiento en las técnicas de clasificación y regresión.
En segundo lugar, la estandarización transforma los datos para que sean mas manejables, por lo que facilita el análisis y mejora la calidad y eficiencia de los resultados. Para obtener este dataset final, se ha aplicado la fórmula (presentada anteriormente en este trabajo) a todas las variables, obteniendo en las mismas, media 0 y desviación típica 1.

3.2.3. Discretización

La discretización de los datos es una técnica muy empleada en análisis de datos, que consiste en transformar el valor de las variables de continuas a discretas. Las ventajas de este proceso son principalmente la reducción de complejidad, el mayor entendimiento de los datos y, como es el caso de este trabajo, la preparación para algoritmos posteriores que precisan de este tipo de datos.
A la hora de realizar la discretización, se ha utilizado la función cut[29] junto a hist[30], que proporciona el argumento *breaks* que define el número de niveles de la versión discreta de las variables.

3.2.4. Gráficos

Debido a que todas las variables del dataset son numéricas, se ha decidido utilizar histogramas y diagramas de cajas en este análisis exploratorio, por varias razones:

1. Distribución de los datos: permite ver cómo se distribuyen los datos, si hay sesgos, simetría y si hay concentraciones de valores. También se mostrarán claramente posibles patrones que pueden proporcionar información importante de las variables.
2. Detección de valores atípicos: gracias a los diagramas de caja se pueden identificar fácilmente valores atípicos en la distribución de los datos.

Los gráficos nos muestran, en el eje horizontal, los distintos valores de la variable y en el vertical, el número de observaciones de dichos valores. A continuación se van a exponer varios ejemplos de histogramas de distintas variables del dataset.

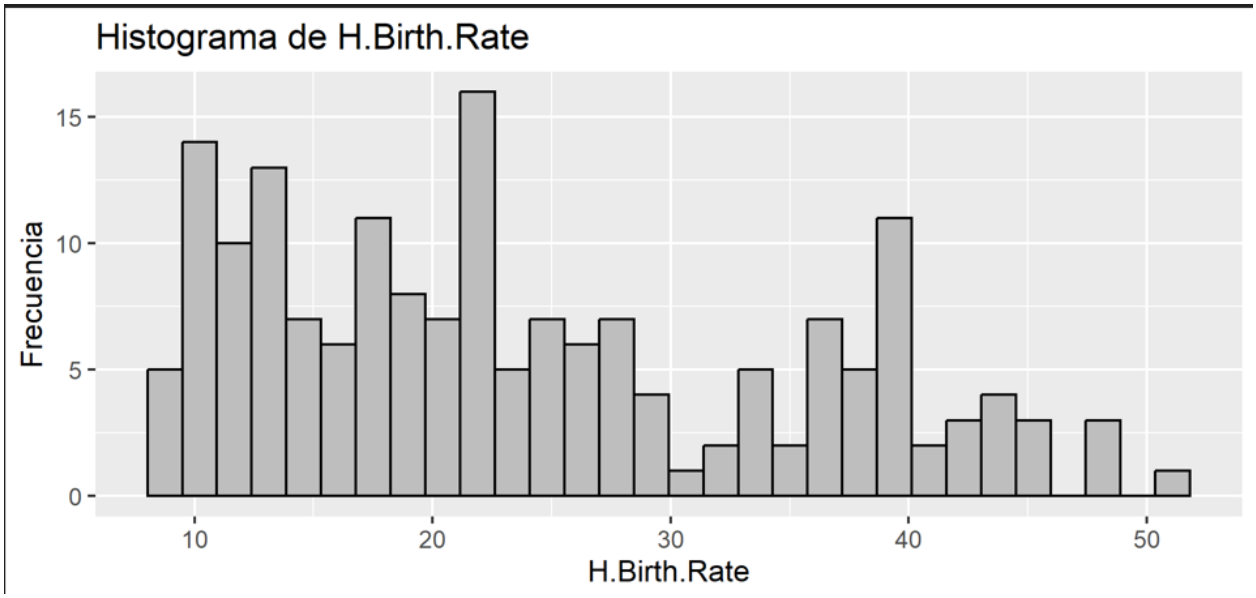

Figura 3.1: Histograma Data.Health.Birth.Rate

Como se ha explicado anteriormente, la variable **Data.Healt.Birth.Rate** refleja en la Figura 3.1 el número de nacimientos por cada 1000 habitantes medido a mitad del año. Podemos observar que el rango de valores va desde 0 hasta algo más de 50 y que esta sesgado hacia la izquierda, es decir, son más comunes los valores más pequeños. Siendo el valor más típico entre los 21-22 nacimientos por cada 1000 habitantes. Se puede interpretar en base a este histograma, que son pocos los países donde hay mas de 30 nacimientos por cada 1000 habitantes, habiendo un repunte al rededor de los 40 nacimientos.

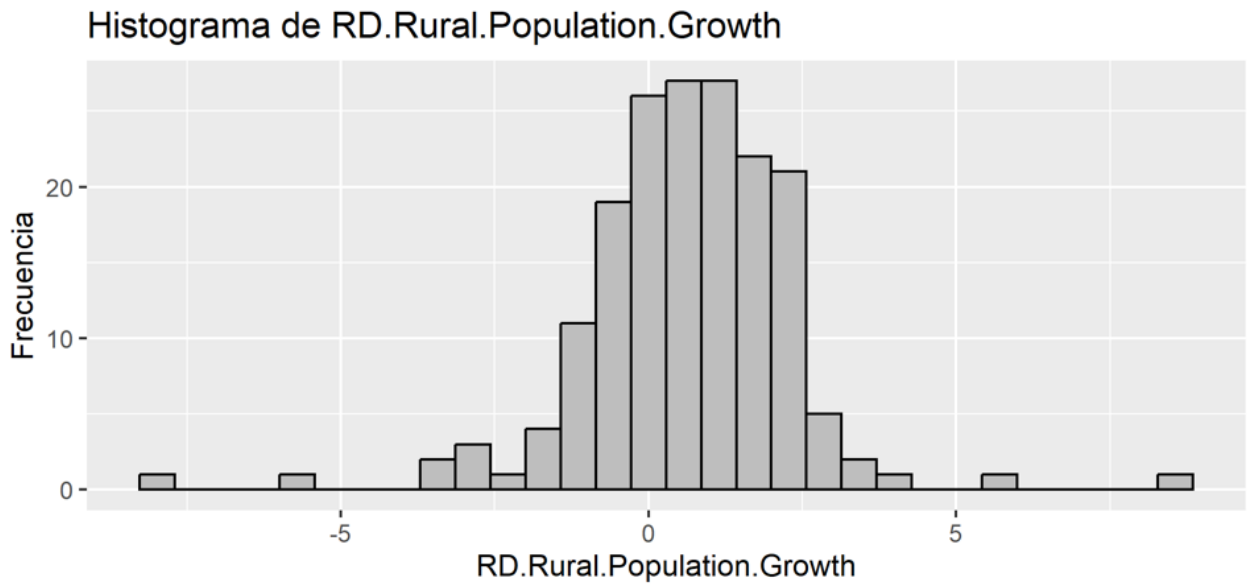

Figura 3.2: Histograma Data.Rural.Development.Rural.Population.Growth

La variable **Data.Data.Rural.Development.Rural.Population.Growth** refleja el porcentaje de crecimiento anual de la población rural. En este caso se puede ver en la Figura 3.2 una distribución similar a la normal, centrada al rededor del valor 1. Está ligeramente sesgada a la derecha, siendo 0 el valor central. Se pueden observar ciertos valores atípicos a partir del valor +-5, siendo los valores ciertamente simétricos. En cuanto al valor teórico de este histograma, se puede interpretar que el crecimiento de población rural es, en general, ligeramente positivo, con el valor más típico sobre el 1 %, aunque muy equiparado con el 0 %.

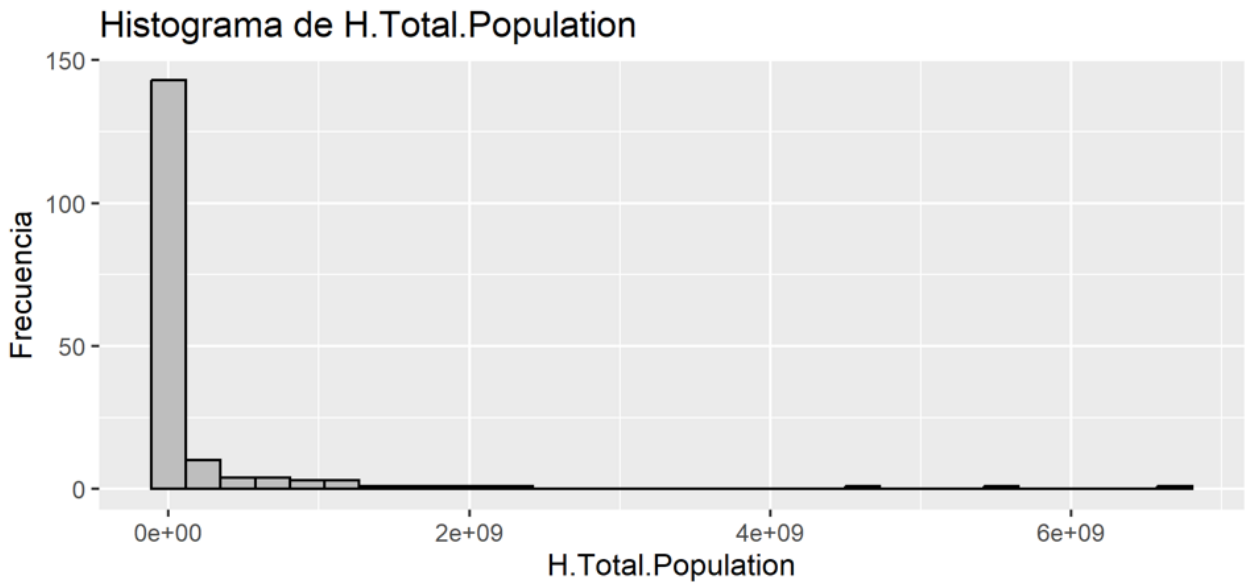

Figura 3.3: Histograma Data.Health.Total.Population

En este caso, la Figura 3.3 presenta el histograma de **Data.Health.Total.Population** (Población total del país), que está completamente sesgado a la izquierda, con un valor pequeño y prácticamente el resto de valores son atípicos. Esto es teóricamente lógico, ya que de todos los países que hay, la gran mayoría son pequeños y con poca población, al contrario de los países grandes con mucha población, que son prácticamente residuales.

Transformación de datos

La transformación de datos es un proceso del análisis de datos que se basa en la modificación de los datos originales para establecer ciertos objetivos. En el caso de este TFG se va a aplicar esta técnica en las variables que están extremadamente sesgadas hacia un extremo, con el objetivo de conseguir una distribución más parecida a la normal. Con esto se conseguirá más facilidad del análisis y de la interpretación de los datos.
Las variables a transformar son las siguientes:

1. H.Total.Population
2. I.Mobile.Cellular.Subscriptions
3. RD.Agricultural.Land
4. RD.Arable.Land
5. RD.Land.Area
6. RD.Rural.Population
7. RD.Surface.Area
8. UD.Population.Density

Todas las variables mencionadas presentan una distribución completamente sesgada a la izquierda, hacia los valores bajos, por lo que se aplicará la siguiente transformación: $\log(1+x)$
Aplicando esté cálculo a las variables anteriormente nombradas, conseguimos reducir el impacto de los valores muy bajos y que de ese modo sea más fácil su tratamiento y su interpretación. A continuación se va a incluir una comparación de la figura 3.3, que se puede ver arriba su completa asimetría hacia la izquierda.

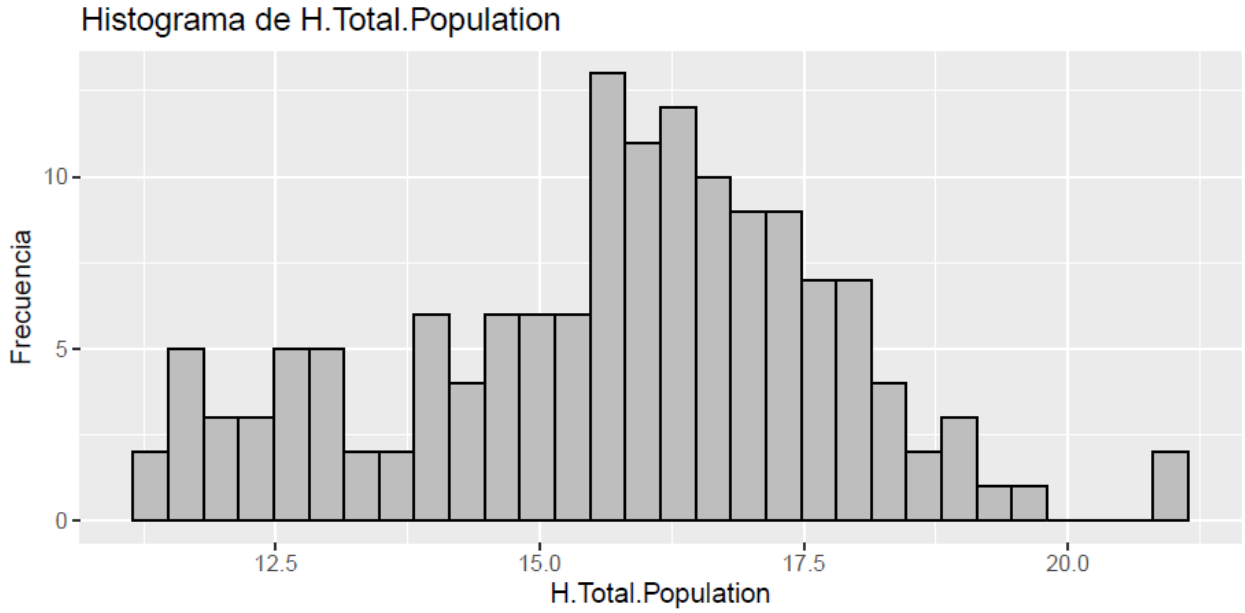

Figura 3.4: Histograma Data.Health.Total.Population transformada

En la Figura 3.4, en comparación con la figura 3.3, se puede observar su distribución más simétrica.

Comparación histogramas datos en bruto contra datos preparados

Tras la preparación de los datos, se han vuelto a obtener histogramas de las variables con el objetivo de evaluar la calidad de la preparación. Se han comparado ambos histogramas para cuantificar la cantidad de información perdida y ver si se han transformado mucho los datos.
A continuación, se incluirán algunos ejemplos de comparación entre ambos histogramas.

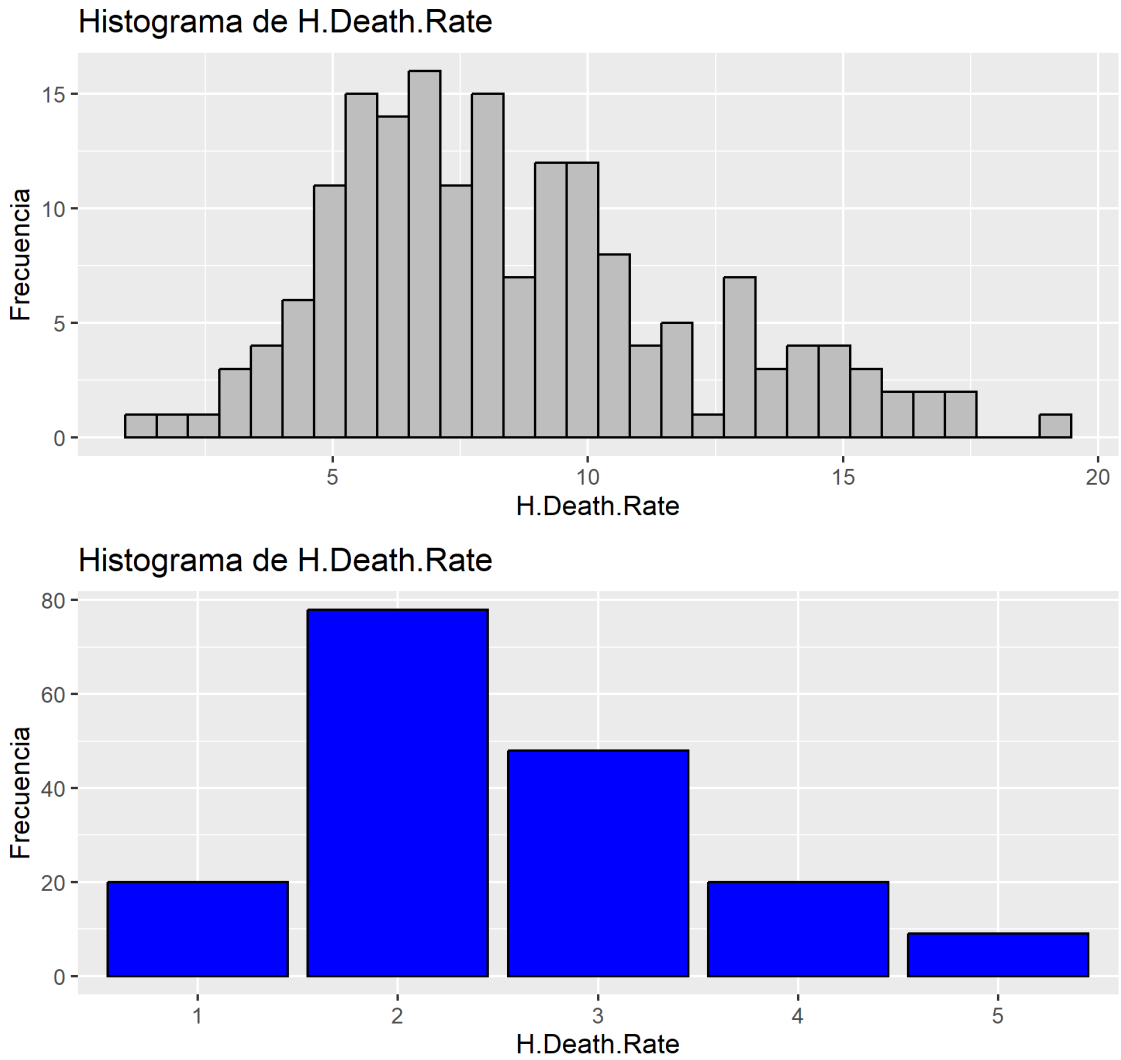

Figura 3.5: Comparación histogramas Data.Health.Death.rate

Se puede observar en la Figura 3.5 que, tras la preparación, se han agrupado los posibles valores de la variable **Data.Health.Death.Rate** (mortalidad por cada 1000 habitantes) a 5. El nuevo histograma sigue una distribución similar a los datos previos a la preparación.

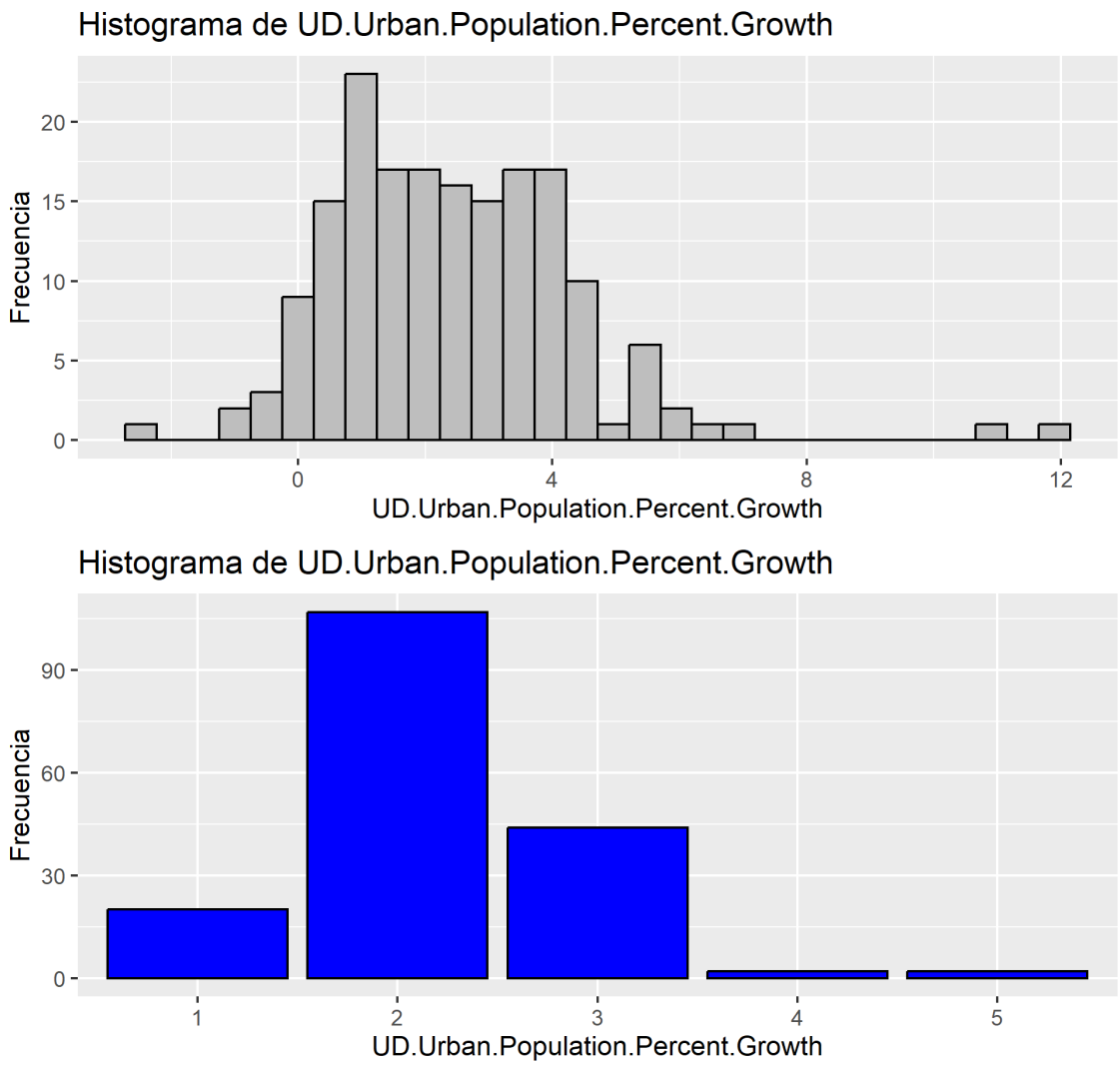

Figura 3.6: Comparación histogramas Data.Urban.Development.Urban.Population.-Percent.Growth

De igual manera que en la Figura 3.5, la Figura 3.6 muestra que la variable **Data.Urban.Development.Urban.-Population.Percent.Growth** (porcentaje de crecimiento de población urbana medido a mediados de año) se divide también en 5 clases y se observa una distribución similar a los datos en bruto.

3.2.5. Clustering

Una vez obtenido este último dataset, donde tenemos una única fila por cada país, sin años y con los valores de las variables estandarizados, se ha procedido a realizar una exploración más a fondo de los datos. En concreto se han obtenido gráficas de las variables y se ha realizado un proceso de clustering. El clustering ayudará a la hora de visualizar los datos y detectar posibles tendencias y comprender mejor la estructura de los datos.

El análisis de grupos o clustering [31] es una técnica muy empleada en el análisis exploratorio, que consiste en crear clases agrupando los elementos por similitud.

Previo a realizar el clustering, se requiere la estandarización y normalización de las variables, ya que los algoritmos de clusters agrupan datos en base a su similitud, por lo que tener datos normalizados va a ayudar en el correcto desempeño del clustering.

Adicionalmente, puede ser conveniente una selección de variables y/o la reducción de la dimensión. En este caso se va a realizar la selección de variables: un algoritmo se realizará con las variables relativas a la salud y el otro con una variable de cada contexto, se detallará más adelante.

Para evaluar los resultados del clustering, a parte del análisis gráfico, se va a utilizar el $Silhouettescore$[32], que mide lo cerca que está un punto de los puntos de su mismo cluster y lo lejos que está de los puntos de otros clusters. Esta métrica puede tomar valores desde -1 a 1:

1. Un valor cercano a -1 indica que los puntos pueden haber sido asignados al cluster erróneo.
2. Un valor cercano a 1 indica que los puntos están bien asignados a su cluster.

En general, un valor a partir de 0.5 se considera un buen resultado.

K-Medoids

De manera parecida al algoritmo k-means, el algoritmo **k-medoids**[33] agrupa los datos minimizando la distancia entre puntos. Concretamente, este algoritmo escoge un punto como el centro del cluster (grupo) y el resto de datos en torno al centro. El algoritmo utilizado para realizar el k-medoids ha sido el algoritmo PAM (Partición Alrededor de Medoids) que trabaja de la siguiente forma:

1. Se seleccionan k (número de clusters) de los puntos como los medios (medoids)
2. Se asocia cada punto restante al medoid mas cercano
3. Se intercambian los medoids con otro punto que no sea medoid y se calcula su costo total (suma de distancias de puntos a sus medoids). Si el costo aumenta, se deshace el cambio.

Antes de realizar el clustering, se han estandarizado los datos. Para este algoritmo se han decidido utilizar todas las variables de salud, para estudiar como se agrupan los países en base a dichos datos.

Para elegir el número de clusters se ha elegido el método del codo (Elbow Method) [34],en el que se utiliza la distancia media de las obervaciones a su centro, con el objetivo de minimizar la suma dentro del cluster. El valor de k a elegir es aquel

donde se produzca la forma de codo en la gráfica, es decir, donde un aumento de K no suponga un gran cambio en la suma de distancias. En este caso se

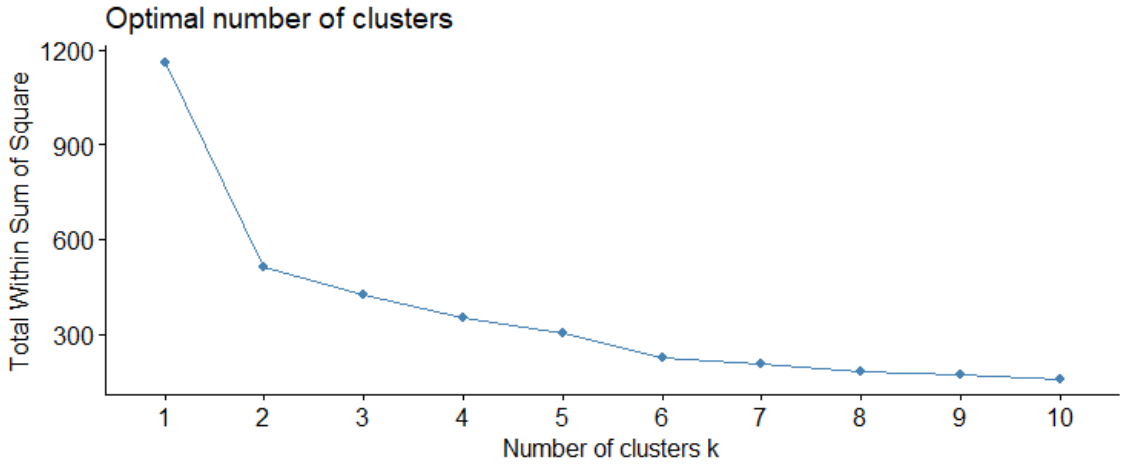

Figura 3.7: Método del codo

ha tomado $K = 2$ como el número óptimo de clusters, ya que es el valor con el cambio más significativo de suma de distancias en la Figura 3.7.

A continuación se van a presentar las gráficas de este algoritmo.

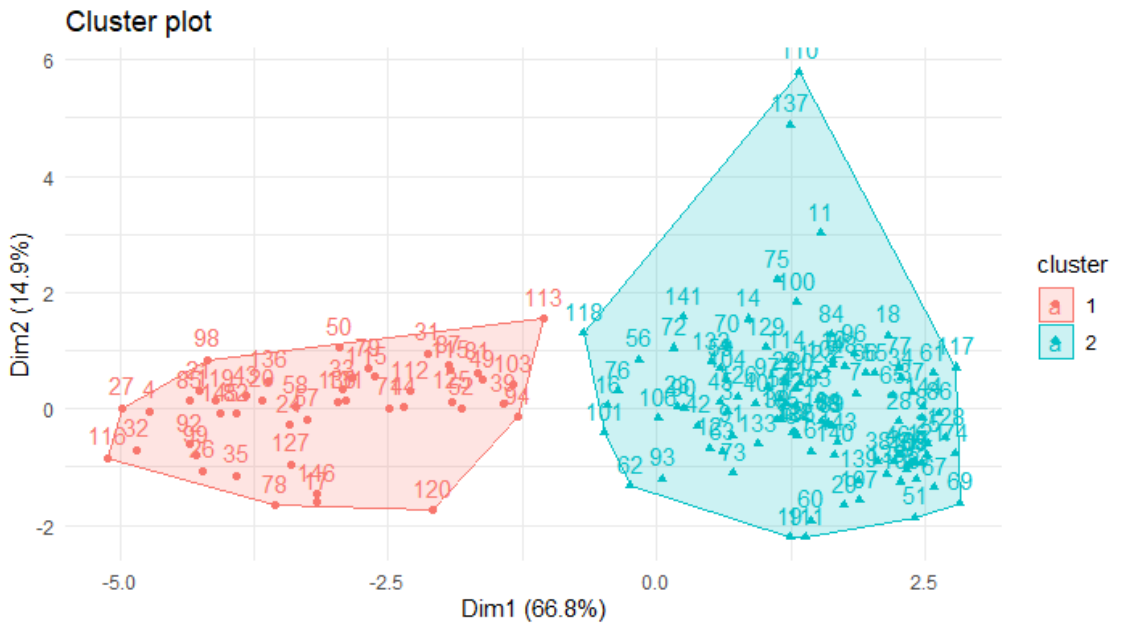

Figura 3.8: Gráfica del cluster PAM

La figura 3.8 representa dos clusters claramente separados, donde el rojo representa el primer cluster y el azul el segundo. Se puede ver claramente que el cluster 2 tiene una mayor concentración y número de puntos, lo que puede

significar que los datos de esta agrupación tienen una alta homogeneidad y son más similares entre sí. En este contexto de salud, se puede afirmar que en el cluster 2, las condiciones sanitarias son similares entre países. No obstante, en este mismo cluster hay más valores atípicos. Respecto al análisis PCA que se ha usado para poder graficar el cluster, se tiene que la dimensión 1 explica el 66.8 % de la variación total, mientras que la dimensión 2 explica el 14.9 %, por lo que tenemos que en conjunto explican el 81.7 %, lo cual es un valor elevado, indicativo de que los resultados del clustering son buenos.

La gran diferencia de porcentaje entre clusters puede deberse a que la Dim1 representa las variables más significativas, mientras que la Dim2 las menos significativas y por ello, con menos varianza. En general, el cluster 1 (Rojo) agrupa los países con tasas sanitarias más desfavorables.

Como se puede ver en el Cuadro A.1, dicho cluster recoge los países que a priori están menos desarrollados, lo que influye directamente en las condiciones sanitarias, mientras que el segundo cluster recoge los países más desarrollados, aunque también presenta países menos desarrollados, pero con buen índice de salud.

A continuación se van a presentar medias de algunas variables significativas por cada cluster.

Cuadro 3.1: Medias de Variables por Cluster K-Medoids

Variable	**Cluster 1**	**Cluster 2**
H.Birth.Rate	38.3	17.6
H.Life.Expectancy.at.Birth.Total	54.6	74.6
H.Death.Rate	12.4	6.58
H.Population.Growth	2.52	1.37
UD.Urban.Population.Percent	36.7	63.8

Se puede observar en el Cuadro 3.1 que, en general, los índices sanitarios son más bajos en el cluster 1, ya que la esperanza de vida es menor y la tasa de muerte es mayor, casi el doble que en cluster 2. La tasa de nacimientos es más alta en el cluster 1, lo cual es común en los países menos desarrollados, debido a la distinta cultura, el poco acceso a métodos anticonceptivos, etc, lo cual explica que la tasa de crecimiento sea prácticamente el doble que en el cluster 2. Esta casuística tiene repercusión en los bajos índices sanitarios de los países poco desarrollados, donde si los recursos son pocos, se produce una mayor escasez al tener una alta tasa de nacimiento.

Se ha calculado la media también de un indicativo demográfico, como lo es el porcentaje de población urbana, que en el cluster 2 es el doble que en el primero, que confirma que en los países menos desarrollados (que suelen tene menos area urbana), los índices sanitarios son menores.

Finalmente se han hecho gráficas BoxPlot de algunas de las variables para reforzar los datos anteriores y poder observar su distribución en función del cluster.

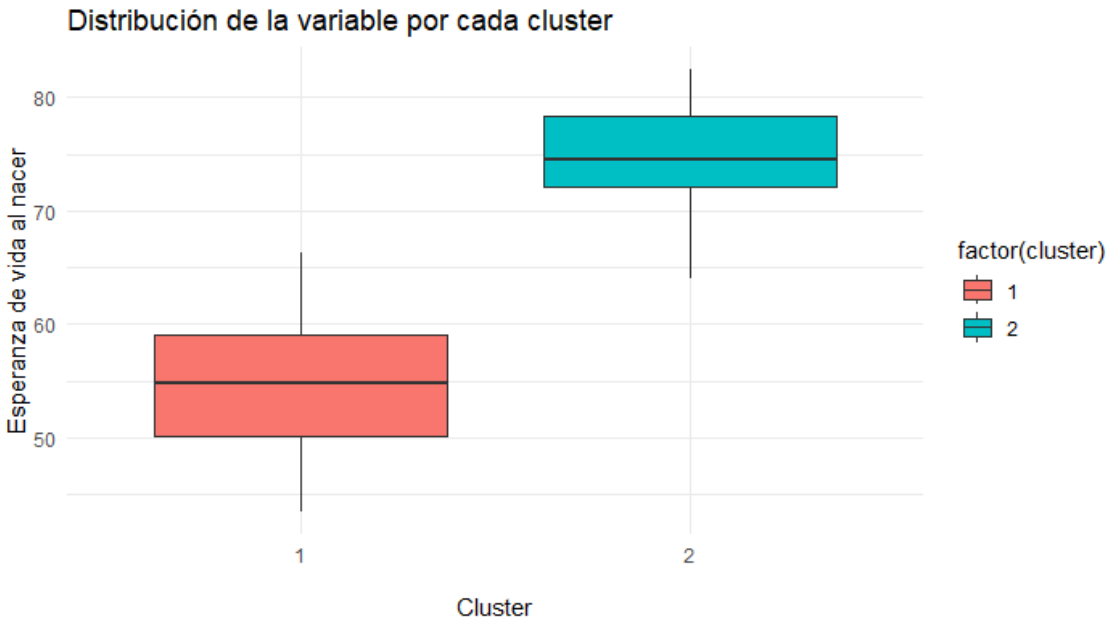

Figura 3.9: Distribución de H.Life.Expectancy.at.Birth.Total por cluster K-Medoids

Como se puede ver en la figura 3.9, la diferencia de las medianas coincide con la calculada anteriormente en la media. El cluster 1 presenta una mayor variabilidad en los datos debido a la mayor anchura de la caja, mientras que los datos en el cluster 2 están más cerca de la mediana. No se observan datos atípicos significativos. En general, se reafirman las conclusiones anteriores.

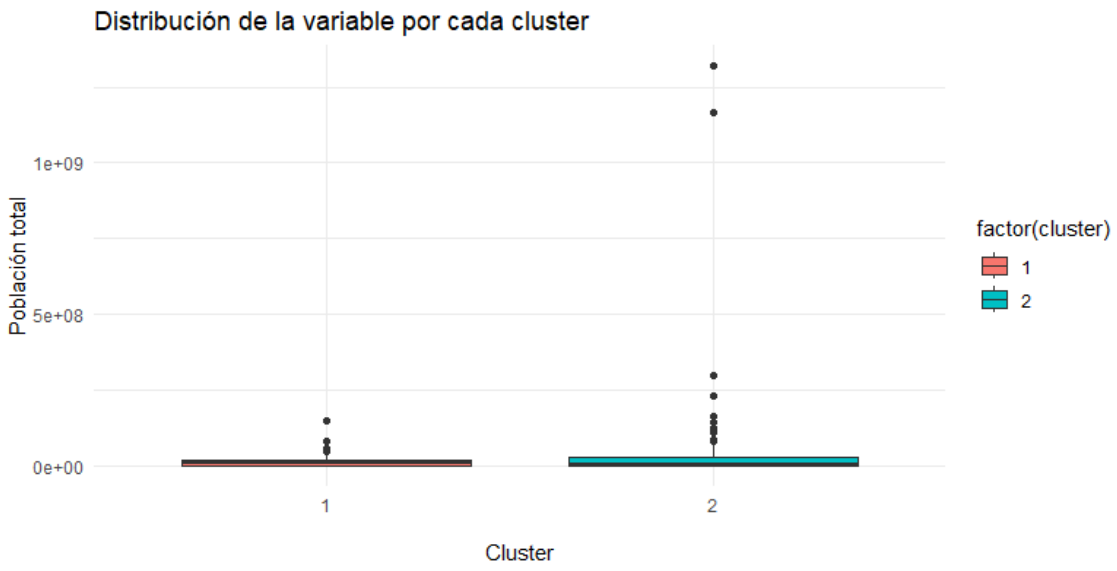

Figura 3.10: Distribución de H.Total.Population por cluster K-Medoids

En este caso se observa en la Figura 3.10 la distribución de población total por cluster. Se puede observar que la población total no es un buen indicativo para conocer los índices sanitarios de un país. Debido a la gran variabilidad de población en el mundo, donde hay muchos países pequeños y pocos grandes, la gráfica no es de gran ayuda. Sin embargo, gracias a los valores atípicos, sobretodo presentes en el cluster 2, se puede afirmar que los países que cuentan con una gran población, están mas desarrollados, como podría ser Estados Unidos, China o Rusia.
Como se ha comentado al inicio de este apartado, la forma de evaluación de los resultados del clustering es el índice de Silueta.

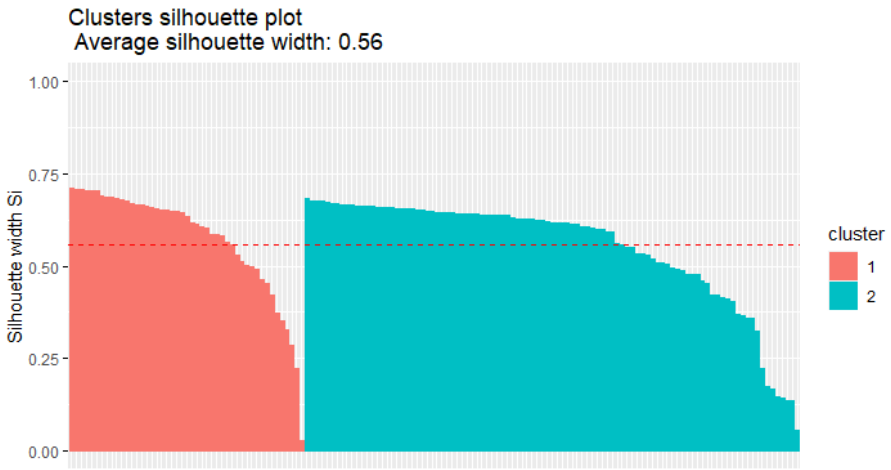

Figura 3.11: Índice de Silueta K-Medoids

Se observa, en la Figura 3.11, un índice de silueta medio igual a 0.56, lo cual indica que el clustering es bueno. Adicionalmente se puede observar que el número de países involucrados en el cluster 2 (azul)es significativamente más alto que en el cluster 1 (rojo), concretamente el cluster 2 (azul) cuenta con 99 observaciones y un índice de silueta de 0.55, mientras que el cluster 1 (rojo), 47 y 0.57 respectivamente. Se observa que el índice de silueta es prácticamente idéntico en ambas agrupaciones, lo que significa que la distribución de datos en ambos grupos es muy similar.

DbScan

Density-based spatial clustering of applications with noise (DbScan)[35] es un método de clustering basado en la densidad entre puntos para agrupar los datos, debiendo fijarse un numero mínimo de puntos en cada cluster. Adicionalmente, se debe seleccionar la distancia que se tomará como radio en los clusters (eps). Este algoritmo etiqueta los puntos atípicos en el cluster 0; dichos puntos no son considerados por el algoritmo como válidos para pertenecer a ningún cluster.

En este caso se ha decidido tomar una variable representativa de cada categoría para ver cómo se agrupan los datos con una visión más global:

1. En el contexto de la salud, se ha elegido la variable H.Life.Expectancy.at.Birth.Total, ya que es la más representativa.
2. Para el desarrollo tecnológico, se ha elegido I.Mobile.Cellular.Subscriptions.per.100.People, por la misma razón que la anterior.
3. Como representación de características de la población, se ha tomado UD.Population.Density.

Para los 3 casos se han elegido variables que expresen una proporción, ya que son las mejor comparables entre países, debido a la amplia variedad de sus tamaños y de sus poblaciones.

Debido a las dimensiones del dataset del proyecto, se ha realizado una iteración con distintos valores de ambos argumentos y se ha elegido la pareja de argumentos con mejores resultados a la hora de realizar el clustering. Los posibles valores de puntos mínimos por cluster han sido desde 5 hasta 30 y de eps se ha iterado desde 0.1 hasta 3 en saltos de 0.1. Los valores con mejores resultados han sido 13 puntos como mínimo por cluster, con una distancia de 0.6. Se ha añadido, como cláusula adicional al algoritmo, que el número de clusters resultantes sea mayor que 2, ya que el cluster 0 refleja el ruido. El número de clusters efectivos resultante han sido 2, confirmando la elección tomada anteriormente para el cluster K-Medoids.

A continuación se van a presentar los resultados obtenidos.

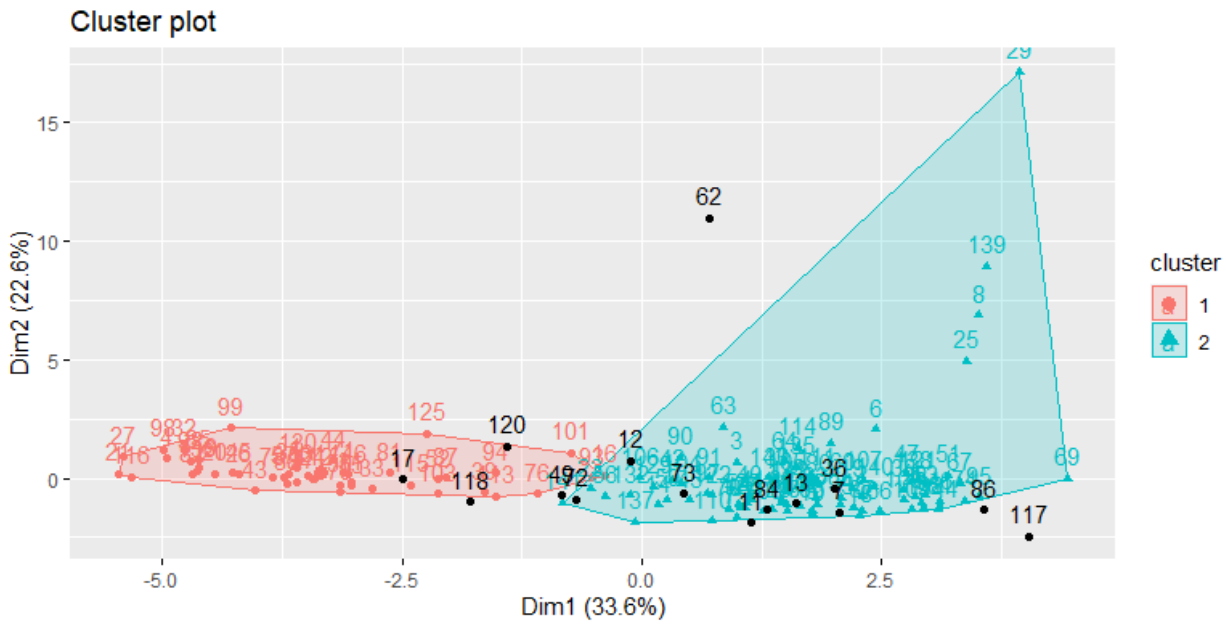

Figura 3.12: Gráfica del cluster DbScan

En este caso, de manera similar al clustering anterior, se diferencian claramente 2 clusters en la Figura 3.12, el rojo y el azul. Además, presentan una distribución similar. Los puntos negros son clasificados por el algoritmo como ruido, podrían pertenecer a países con características particulares que no sean compartidas por los demás países, por lo que no pueden agruparse en clusters. En este caso, la diferenciación entre clusters no es tan clara como en el clustering de K-Medoids, lo cual parece lógico ya que se han añadido variables referentes a distintos contextos. En general, las conclusiones de este análisis son similares al anterior, con una diferenciación entre países subdesarrollados y otros más desarrollados.

Respecto al PCA, en este caso la Dim1 explica un 33.6 % de la variación total mientras que la Dim2 un 22.6 %, siendo la combinación de ambas un 56.2 % del total, lo cual es notablemente inferior al 81.7 % presente en el anterior clustering. Esto se explica también, como se ha dicho anteriormente, debido a la mayor dificultad de agrupar los países teniendo en cuenta el rango de características que se han elegido para el clustering. Esto, combinado a un índice de silueta de un 0.47 mostrado en la Figura 3.13, confirma unos peores resultados que los anteriores. En este caso, se han agrupado 47 observaciones en el cluster 1, 84 en el cluster 2 y 15 países han sido etiquetados como ruido.

Se pueden consultar qué países pertenecen a cada agrupación en el Cuadro A.2.

Figura 3.13: Índice de Silueta DbScan

Como es lógico, en la Figura 3.13, el cluster 0 cuenta con un índice de silueta negativo ya que los puntos no pertenecen a ningún cluster. También se observa un valor negativo en algunas observaciones del cluster 2, denotando valores atípicos dentro de este.
Como anteriormente, se calculan las medias de algunas de las variables del dataset según su cluster.

Cuadro 3.2: Medias de Variables por Cluster DbScan

Variable	**Cluster 0**	**Cluster 1**	**Cluster 2**
H.Birth.Rate	19.4	38.0	17.5
H.Life.Expectancy.at.Birth.Total	69.0	55.4	75.2
H.Death.Rate	8.0	12.0	6.5
H.Population.Growth	1.6	2.5	1.3
UD.Urban.Population.Percent	57.0	35.5	65.7
I.Mobile.Cellular.Subscriptions.per.100.People	65.4	27.4	81.5
UD.Population.Density	913.0	76.7	132.9

A grandes rasgos, se puede observar en el Cuadro 3.2, que las medias calculadas son muy similares a los resultados del clustering K-Medoids, como se puede ver en el cuadro 3.1, lo que refuerza los resultados obtenidos.

En cuanto al ruido, podemos ver que en la mayoría de las variables calculadas, los valores se encuentran entre ambos clusters, presentando valores "normales". Sin embargo, la variable UD.Population.Density presenta una media muy elevada en el cluster 0 de 913.0, muy por encima de 76.7 en el cluster 1 o 132.9 en el cluster 2. Esto podría explicar parte del ruido, por lo que se puede asumir que

el cluster 0 cuenta con países muy densos poblacionalmente; en el cuadro A.2 se puede observar que países como Bangladesh o Maldivas se clasifican como ruido.
A continuación se mostrarán las gráficas de distribución por cluster de las variables empleadas en el algoritmo DbScan.

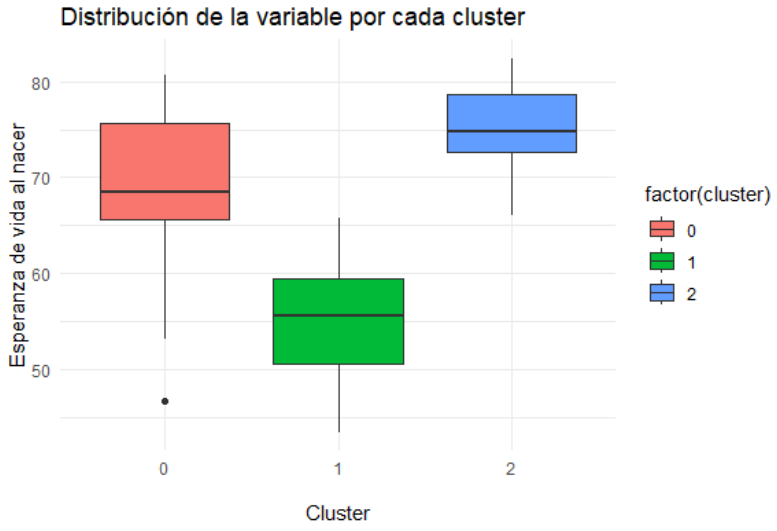

Figura 3.14: Distribución de H.Total.Population por cluster DbScan

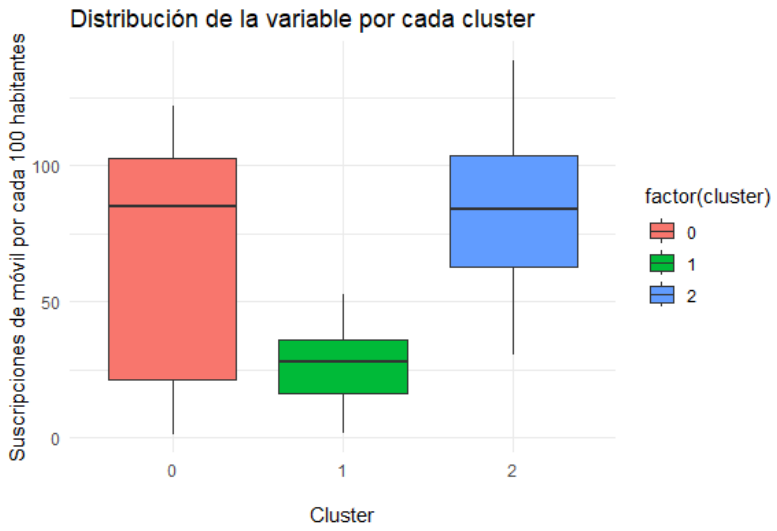

Figura 3.15: Distribución de I.Mobile.Cellular.Subscriptions.per.100.People por cluster DbScan

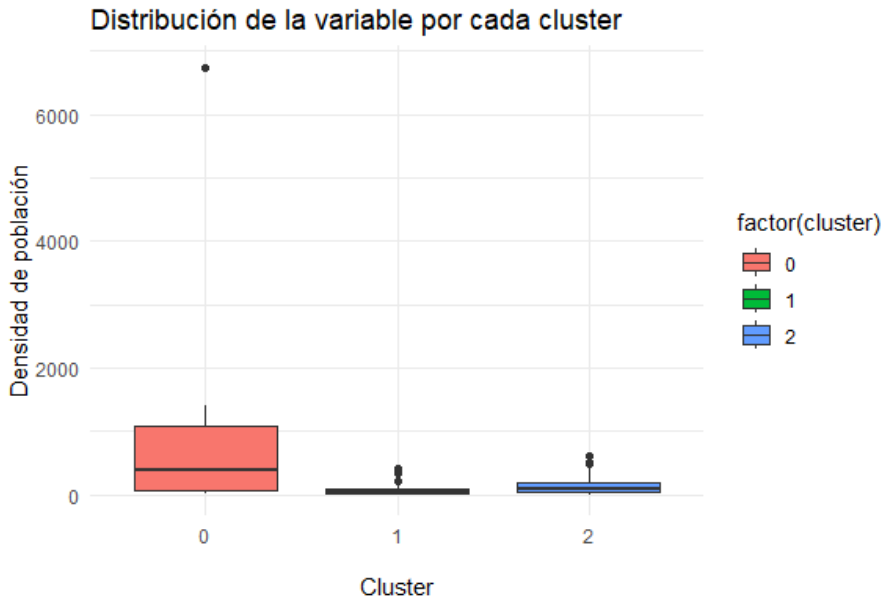

Figura 3.16: Distribución de UD.Population.Density por cluster DbScan

En general, En las Figuras 3.14, 3.15, 3.16, los registros calificados como ruido tienen una mayor variabilidad, como es lógico, ya que no se agrupan. En cuanto a la esperanza de vida 3.14, los resultados son muy similares. En el número de suscripciones de teléfonos móviles 3.15, vemos también una clara diferenciación entre clusters, con mucha más variabilidad en el cluster 2 que en el cluster 1. Finalmente, se puede observar que la densidad poblacional 3.16, no es un buen indicativo para separar los países, ya que la distribución de ambos clusters es muy parecida, siendo ligeramente mayor en el cluster 2 (azul) que en el cluster 1 (verde). Esto sugiere que la densidad de población no influye en el desarrollo de un país.

Conclusión del clustering

A modo de conclusión, se observan unos resultados bastante parecidos aplicando ambos algoritmos, lo que refuerza las conclusiones obtenidas. No obstante, en el primer caso se obtiene una mejor distinción entre los países más y menos desarrollados, que puede deberse a que solo se ha agrupado con variables de salud. Cuando se ha ampliado a variables que recogen otro tipo de datos en el algoritmo DbScan, aunque los resultados no son tan claros, se diferencian sin problema lo que podría clasificarse como países desarrollados y subdesarrollados.

Capítulo 4

Análisis de Datos

Una vez realizado el análisis exploratorio, donde se han comprendido mejor los datos, sus estructuras y tendencias, se procede a realizar un análisis clasificatorio de los datos.

El fin del UNPD, como se ha comentado en el Capítulo 2, es mejorar la calidad de vida de las personas de todo el mundo. Es por esto que las variables de la sección de la salud van a tomar una mayor importancia en esta base de datos. Se puede resumir, para simplificar la tarea, que la variable con más peso en este objetivo será H.Life.Expectancy.at.Birth.Total, que se refiere a la esperanza de vida que tiene un individuo al nacer. Esta decisión se ha tomado en base a que, cuanta más calidad de vida tenga un individuo, es más probable que viva más años, por lo que se puede extrapolar el objetivo de las Naciones Unidas a que la esperanza de vida sea la mayor posible.

A modo de recordatorio, el análisis clasificatorio clasifica la variable objetivo en base a ciertas variables. Este modelo es interesante en este caso, ya que podría tener una importante aplicación a la hora de predecir la esperanza de vida en cada país en los años venideros según ciertos parámetros.

En este cuarto capítulo se va a desarrollar, en primer lugar, una explicación del método empleado para la elección de variables a utilizar en el análisis, después se estudiará el análisis y por último, se analizará el rendimiento del análisis empleando indicadores que nos ayudarán a cuantificar su eficacia.

4.1. K-vecinos más cercanos (KNN)

El algoritmo KNN [36] es un método importante de clasificación supervisada no paramétrico. Este algoritmo se basa en la proximidad de las observaciones para realizar la clasificación de la variable clasificatoria. El funcionamiento del algoritmo es sencillo: para cada observación a clasificar, divide los datos de entrenamiento en grupos de k vecinos más cercanos donde la clase predominante será la que se le dará a la observación.

Para medir la distancia entre observaciones, se utiliza la siguiente fórmula de distancia:

$$d(P,Q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

4.1.1. Elección del parámetro K

Una buena elección del parámetro k es fundamental para el desempeño de este algoritmo, ya que el cambio del número de vecinos que se incluirá en cada grupo cambiará los resultados del clasificador. En general, la elección de un parámetro k alto reduce el efecto del ruido pero puede separar observaciones muy similares.

En este trabajo se ha realizado una iteración del clasificador para valores de k entre 1 y 20 y se ha elegido el número que mejores resultados obtenía, que ha sido 3. Por lo tanto, a la hora de realizar el algoritmo KNN, se ha elegido $k = 3$.

4.2. Árboles de clasificación

Los árboles de clasificación[37] son una herramienta utilizada para la predicción. Se trata de modelos de aprendizaje supervisado empleados para clasificar datos en grupos definidos. Su funcionamiento se basa en la elección de una variable objetivo, en la cual se clasificarán los datos según una serie de variables predictoras. Esta técnica se conoce como CART: Classification And Regression Trees. Dentro de esta técnica, se usará regresión si la variable objetivo es continua y clasificación cuando es discreta. Por eso se utiliza clasificación en este trabajo.

Particularmente, se empleará el algoritmo RPART(Recursive Partioning And Regression Tree). A continuación se detallará su funcionamiento.

Recursive Partioning And Regression Tree

Este algoritmo parte del nodo raíz que se corresponde a todo el conjunto de datos. A partir de ahí, el algoritmo busca la variable predictora que mejor separa los datos en las categorías de la variable objetivo, realizando una partición de los datos y creando dos nuevos nodos. Una vez en estos dos nuevos nodos, se repite la operación de partición de datos recursivamente. El algoritmo acaba cuando no se encuentra una partición de los datos que mejore la del nodo anterior, los cuales pasan a ser los nodos terminales u hojas. Para medir el rendimiento de cada nodo o partición se utiliza el Índice de Gini.

Índice de Gini

El Índice de Gini[38] es una medida de homogeneidad (impureza) que varía entre el 0 y el 1. Si el índice es igual a 0, significa que hay una perfecta homogeneidad donde la partición tiene el mismo número de datos en todas las categorías de

la variable objetivo. Sin embargo, si el índice se aproxima a 1, significa que la repartición entre las categorías de la variable objetivo es desigual. Una vez conocido el funcionamiento del Índice de Gini, se puede explicar el método de validación del algoritmo CART a la hora de crear nuevos nodos. Cuando se hace una nueva partición a raíz de un nodo, se calcula el índice de Gini: si este mejora a la iteración anterior (más cercano a 0), se crea el nuevo nodo, en otro caso, el nodo actual se convierte en hoja y acaba el algoritmo.

4.2.1. Selección de variables

Como se ha explicado en Capítulo 2 del trabajo, existen dos tipos de métodos para realizar la selección de variables:

1. Métodos de filtro
2. Método de Wrapper

En este trabajo se va a utilizar principalmente un método de filtro, aunque también se empleará Wrapper para hacer algunas comprobaciones.

En primer lugar se han de elegir las variables que servirán para predecir la variable objetivo, que como se ha dicho, en este caso será *H.Life.Expectancy.at-.Birth.Total*. Para elegir las variables se va a utilizar la información mutua e información mutua condicional, por lo que se ha usado la discretización para poder calcular dichos parámetros.

Se ha dividido este proceso en dos etapas:

1. Información Mutua entre variables y la variable objetivo
2. Información Mutua Condicionada con la variable objetivo

Adicionalmente, se ha tomado la decisión de eliminar previamente a esta selección las variables de *Life.Expectancy.At.Birth.Male* y *Life.Expectancy.At.Birth.Female*, ya que nos dan prácticamente la misma información que la variable objetivo y en caso de querer predecir dicha variable, normalmente no se dispondrá de las dos mencionadas anteriormente, ya que sería demasiado sencillo.

Información Mutua entre variables y la variable objetivo

La información mutua [14] entre dos variables dice cuánta información proporciona una variable A respecto a otra variable B. Es decir, cuanta más información mutua, más conocemos de B sabiendo A, o de A sabiendo B. La forma de calcular la información mutua entre dos variables es la siguiente:

$$I(A;B) = \sum_{b \in B} \sum_{a \in A} p(a,b) \cdot \log \left(\frac{p(a,b)}{p(a)p(b)} \right)$$

Donde:

1. $p(a,b)$ es la función conjunta de distribución de probabilidad de A y B.

2. $p(a)$ es la función marginal de distribución de probabilidad de A.
3. $p(b)$ es la función marginal de distribución de probabilidad de B.

Una vez comprendida la información mutua, se explicará la forma de emplearla. En esta primera etapa de la selección de variables se ha calculado la información mutua de cada variable con la variable objetivo. A mayor información mutua de cada variable, mayor peso tendrá conocerla para poder estimar la variable objetivo, por lo que se ha cogido la mitad de las variables con la información mutua más alta, empleando la mediana de todos los valores.

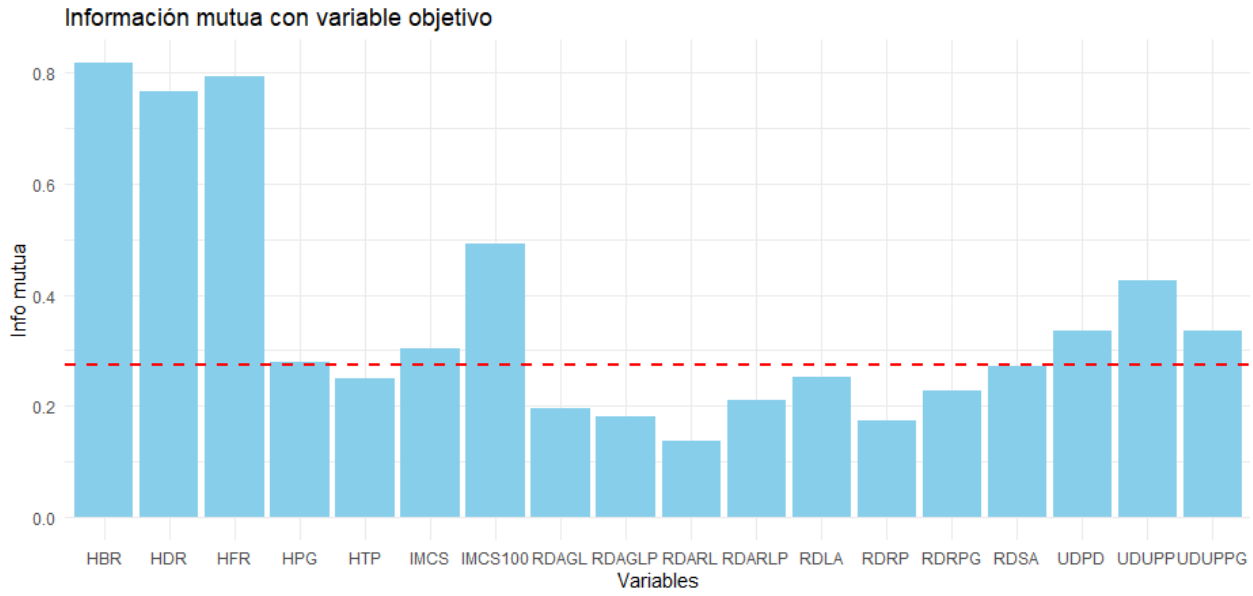

Figura 4.1: Información Mutua de las variables con la variable objetivo

En la Figura 4.1 se puede ver el valor de la información mutua de todas las variables con la variable objetivo y, en rojo, podemos ver la mediana de todos los valores. Por lo tanto, las siguientes variables serán las que pasarán el primer filtro de la selección de variables:

1. **(HBR)**
2. **(HDR)**
3. **(HFR)**
4. **(HPG)**
5. **(IMCS)**
6. **(IMCS100)**
7. **(UPDP)**
8. **(UDUPP)**

9. **(UDUPPG)**

Las anteriores simplificaciones de variables se pueden consultar en el Cuadro A.3 del Anexo A.

Información Mutua Condicional entre parejas de variables y la variable objetivo

La información mutua condicional[39] comparte el mismo concepto que la información mutua, es decir, la cantidad de información que podemos conocer de una variable sabiendo otra. En este caso se incluye una tercera variable. Más precisamente, la información mutua condicional entre dos variables A y B, dada una tercera variable C, se define como la reducción de la incertidumbre sobre A sabiendo B, dada la variable C. Matemáticamente se define de la siguiente forma:

$$I(a;b|c) = \sum_{a\in A}\sum_{b\in B}\sum_{c\in C} p(a,b,c)\log\left(\frac{p(a,b|c)}{p(a|c)p(b|c)}\right)$$

Donde:

1. $p(a,b,c)$ es la función conjunta de distribución de probabilidad de a, b y c.
2. $p(a|c)$ es la función probabilidad condicional de a dado c.
3. $p(b|c)$ es la función probabilidad condicional de b dado c.
4. $p(a,b|c)$ es la función conjunta de distribución de probabilidad de a y b dado c.

Para este trabajo, se va a utilizar la Información Mutua entre una pareja de variable condicionada a la variable objetivo (Life.Expectancy.At.Birth.Total). De esta manera, se cuantifica la cantidad de información que se obtiene de la variable objetivo, conociendo la pareja de variables. Cuanto mayor sea la IMC, más información proporcionará la pareja de variables.

Una vez escogidas las 9 variables con mayor información mutua con la variable objetivo, se agrupan en parejas para obtener la información mutua condicionada respecto a la variable objetivo y así realizar el segundo filtro de la selección de variables.

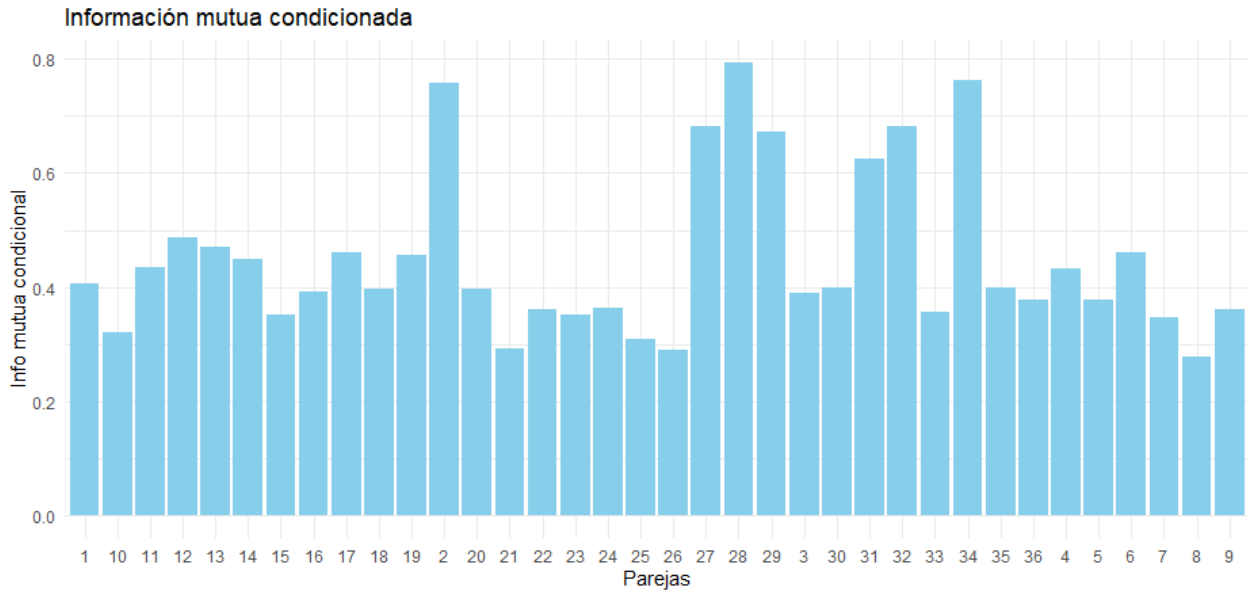

Figura 4.2: Información mutua condicionada de todas las parejas de variables con la variable objetivo

En la Figura 4.2 se puede ver el valor de la información mutua condicionada a la variable objetivo de cada pareja de variables elegidas en el primer filtro de la selección de variables. El segundo filtro se basará en una elección según dicho valor.

Hay 7 valores de IMC que destacan por encima del resto, por lo que se han escogido esas 7 parejas de variables, que son las siguientes:

1. IMCS - UDPD
2. UDPD - UDUPP
3. HBR - HFR
4. IMCS100 - UDUPP
5. IMCS - IMCS100
6. IMCS - UDUPP
7. IMCS100 - UDUP

Como se puede ver, se repiten variables, lo cual es positivo, ya que asegura que la mayoría de variables entre sí tiene un IMC alto con la variable objetivo, lo que puede suponer mejores resultados.

Estas 7 parejas aportan las siguientes variables individuales:

1. IMCS
2. UDPD
3. HBR
4. IMCS100
5. UDUPP
6. HFR

En este punto, se ha tomado la decisión de coger la octava pareja con mayor IMC, ya que las siete primeras solo aportan seis variables distintas y la octava aporta una nueva variable:

1. HDR

(H.Birth.Rate), la cual, a priori, puede parecer lógico que tenga cierta importancia en la esperanza de vida de las personas.

Por lo que las variables finalmente seleccionadas han sido:

1. IMCS
2. UDPD
3. HBR
4. IMCS100
5. UDUPP
6. HFR
7. HDR

Una vez realizado el árbol clasificatorio, se ha empleado el método Wrapper para una posible mejora de los resultados. Para ello se ha iterado el algoritmo del árbol clasificatorio variando las variables utilizadas. No se ha mejorado el resultado tras aplicar este algoritmo, por lo que se han mantenido las 7 variables elegidas inicialmente.

4.3. Evaluación de los clasificadores

La evaluación de los clasificadores [40] es fundamental para conocer los resultados de estos, entender su funcionamiento y así conocer su capacidad de predicción. Se van a utilizar distintos métodos y métricas para evaluar cada clasificador.

4.3.1. Métodos y métricas de evaluación

Matriz de confusión

Esta herramienta es la que mejor permite ver el funcionamiento del clasificador. Por un lado, las columnas representan las observaciones reales de los datos divididas en clases y las filas las predichas por el clasificador. Se pueden visualizar el número de aciertos y errores del clasificador, lo cual hace muy sencillo conocer su desempeño.

	Observación		
	Bajo	Medio	Alto
Predicción	V Bajo	F Medio	F Alto
	F Bajo	V Medio	F Alto
	F Bajo	F Medio	V Alto

Cuadro 4.1: Matriz de confusión donde V=Verdadero y F=Falso

En el Cuadro 4.1 se puede ver la estructura de la matriz de confusión, donde en las columnas se tienen los valores reales de las observaciones y en las filas, las predicciones hechas por el modelo.

Exactitud

Cuantifica el número de aciertos del clasificador

$$\text{Exactitud} = \frac{V_{Total}}{Total}$$

Tasa de error

Cuantifica el número de errores del clasificador

$$\text{Tasa de error} = \frac{F_Total}{Total}$$

Precisión

Indica la tasa de acierto entre las predichas positivamente

$$\text{Precisión} = \frac{V_x}{Total_{xPredicho}}$$

Siendo x cada clase a predecir. Haciendo la media de estos 3 valores se obtiene la precisión global.

Sensibilidad

Indica la tasa de aciertos correctamente predichos por el modelo

$$\text{Sensibilidad} = \frac{V_x}{Total_{xActual}}$$

Siendo x cada clase a predecir. Haciendo la media de estos 3 valores se obtiene la sensibilidad global.

F1-Score

Esta métrica propone un resultado combinado entre la sensibilidad y la precisión, realizando una media de ambas para tener una visión más global del desempeño del modelo.

$$\text{F1-Score} = \frac{2 * (Precision * Sensibilidad)}{Precision + Sensibilidad}$$

Una vez presentadas las principales métricas de evaluación de los métodos a tener en cuenta, se va a hacer un análisis de los resultados.

4.3.2. Análisis de los resultados

Modelo KNN

Para este modelo se obtiene la siguiente matriz de confusión a partir de la que se van a sacar las métricas.
Las métricas asociadas son:

	Observación		
	Bajo	Medio	Alto
Predicción	14	1	0
	1	8	7
	0	2	13

Cuadro 4.2: Matriz de confusión modelo Knn

1. **Exactitud** = 0.760
2. **Tasa de error** = 0.239
3. **Precisión** = 0.767
4. **Sensibilidad** = 0.770
5. **F1-Score** = 0.768

A partir de estos resultados y el Cuadro 4.2, se puede ver que el modelo clasifica correctamente el 76 por ciento de las ocasiones, lo cual es un valor decente pero no alto. La precisión y la sensibilidad son bastante similares, lo que indica que el modelo tiene una capacidad similar tanto para clasificar bien cada clase como para evitar clasificar erróneamente. Esto se refleja en la métrica **F1-Score**, ya que es un valor muy similar a los otros dos. Se puede ver que la clase más errada resulta ser la clase **Alto**, donde se clasifican 7 observaciones en **Medio** cuando deberían ser **Alto**. Sin embargo, la clase más precisa es **Bajo**, con una precisión y sensibilidad de $0{,}933$. Esta clase es la que más interesa en el análisis, ya que el objetivo global de las ODS se centra en los países más desfavorecidos y, por tanto, con la esperanza de vida más baja.

Árbol clasificatorio

En primer lugar, se incluye el gráfico del árbol clasificatorio mostrado por el modelo.

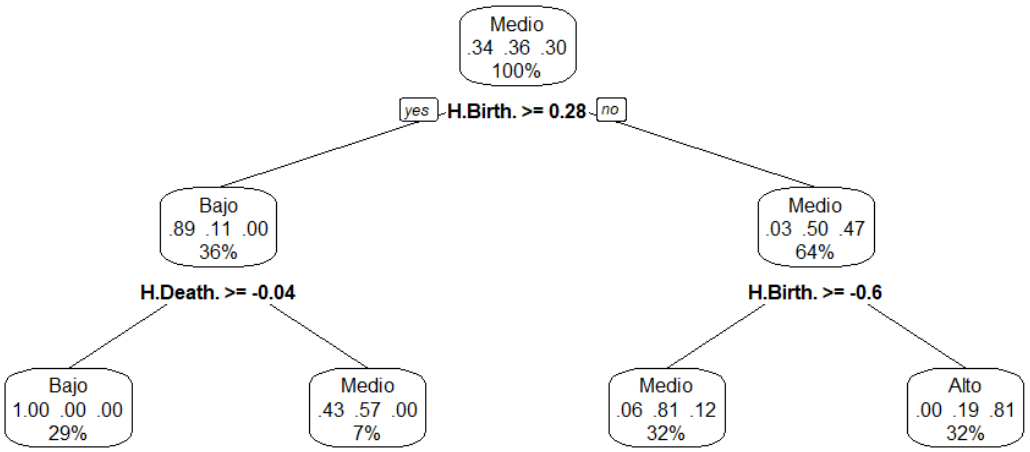

Figura 4.3: Árbol Clasificatorio

Como se puede ver en la Figura 4.3, la variable que se ha utilizado como nodo raíz ha sido **H.Birth.Rate**, cuyo criterio de división es si el valor de dicha variable es mayor o igual que 0.28. En caso positivo, el árbol clasifica en la categoría **Bajo** con una distribución de 89, 11 y 0 porciento para **Bajo**, **Medio** y **Alto** respectivamente. Posteriormente, se utiliza la variable **H.Death.Rate**. En caso de ser mayor o igual que -0.04, el árbol lo clasificará como **Bajo** con una distribucion de 100 % **Bajo**, es decir, con total precisión. En caso de que **H.Death.Rate** sea menor que -0.04, se clasificará como **Medio**, con una precisión del 57 %, el 43 % restante se etiqueta en **Bajo**.
Siguiendo la otra rama, en caso de que **H.Birth.Rate** sea menor que 0.28, el árbol clasificará el dato como **Medio**, con una distribución de 3, 50 y 47 por ciento respectivamente. En este nodo se vuelve a tomar la decisión con **H.Birth.Rate**, con -0.6 como valor diferencial. En caso de ser mayor, el dato se clasificará como **Medio**, con una distribución igual a 6, 81 y 12 por ciento. En caso contrario, se clasificará como **Alto**, con una precisión de 81 %, donde el 19 % restante se clasificará como **Medio**.

A continuación se inserta la matriz de confusión de donde se sacarán las métricas.

1. **Exactitud** = 0.674
2. **Tasa de error** = 0.326
3. **Precisión** = 0.706
4. **Sensibilidad** = 0.672
5. **F1-Score** = 0.688

	Observación		
	Bajo	Medio	Alto
Predicción	12	0	0
	3	7	7
	0	5	12

Cuadro 4.3: Matriz de confusión modelo árbol clasificatorio

En este caso, la estructura de los resultados y el Cuadro 4.3 es similar, es decir, la **precisión**, **sensibilidad** y **F1-Score** son bastante similares, pero el nivel de acierto es inferior al algoritmo Knn, lo que sugiere que el árbol clasificatorio tiene un rendimiento peor. Adicionalmente, se clasifican varios países en la clase **Medio**, cuando deberían clasificarse en **Bajo**, lo cual puede ser especialmente peligroso en este contexto, ya que se podría a tratar algunos países como más desarrollados de lo que son y no aplicarles las medidas apropiadas.

Capítulo 5

Conclusiones y Líneas futuras

5.1. Conclusiones

A modo de recordatorio, al comienzo de este proyecto se establecieron los siguientes objetivos, que se procede a evaluar:

1. Estudiar los datos y su contexto
2. Dominar los métodos y técnicas de análisis de datos
3. Preparación de los datos
4. Documentación de todo el proceso
5. Estudio y desarrollo de modelos para exploración de datos
6. Estudio y desarrollo de modelos de clasificación, predicción o simulación que aporte valor a los datos

En cuanto al primer punto, como se ha plasmado en el Estado del Arte, gracias a la web del Banco Mundial [4], se ha comprendido el origen y estructura de todos los datos de la base de datos, así como su contexto e importancia en las Naciones Unidas para el Desarrollo [8].

Los puntos 3 y 4 se pueden resumir en el Capítulo 3, donde se ha realizado toda la preparación de los datos y su análisis exploratorio. En cuanto a la preparación de datos, ha sido un proceso costoso, pero crucial para una correcta obtención de resultados. En cuanto a la exploración de los datos, tal como se propuso, se han implementado dos técnicas de clustering: DbScan y K-Medoids, que han sido de vital importancia para comprender la estructura de los datos y mejorar su comprensión.

En cuanto al punto 6, se han estudiado e implementado varios modelos clasificatorio-predictivos como son los árboles clasificatorios y el algoritmo knn. Posteriormente, se ha evaluado su rendimiento, siendo mejor el de knn. No obstante, como se va a ampliar en la siguiente sección, se pueden implementar algunas técnicas para mejorar estos resultados.

De manera general, durante todo el proyecto se han ido desarrollando los dos puntos restantes, el 2 y el 4. El dominio de métodos y técnicas de análisis de datos ha sido el más costoso durante el proyecto ya que mi conocimiento de análisis de datos y R era muy básico. En cuanto a la documentación del proceso, se ve reflejada en esta memoria.

5.2. Líneas Futuras

En esta sección se van a comentar las conclusiones de este trabajo, valorando el cumplimiento de los objetivos establecidos y contextualizando los resultados. Adicionalmente, se van a proponer posibles lineas futuras de este trabajo, que se podrían realizar a partir del análisis realizado.

Profundización en preparación y análisis de los datos

En primer lugar, como se ha visto en el Capítulo 3, se ha eliminado la variable $Year$ (año), de la base de datos original. Una posible ampliación podría ser mantener esta información y realizar una clasificación a nivel temporal, a diferentes periodos de tiempo como años o décadas.

Otra posible mejora relacionada con la anterior sería tener en cuenta las agrupaciones de países eliminadas al inicio y realizar unas similares con los resultados obtenidos para así poder compararlas con las agrupaciones originales.

A nivel de análisis, se podrían incluir nuevos modelos de clasificación y regresión así como nuevas técnicas de clustering en busca de mejor entendimiento y resultados del análisis. Entre estos modelos podría estar la clasificación Bayesiana o bosques aleatorios. También, una mejora de los parámetros utilizados en los modelos podría servir como mejora del análisis, incluyendo la selección de variables, donde se podría probar un modelo de Wrapper con las variables iniciales, lo cual sería computacionalmente muy pesado pero podría resultar en mejores conclusiones.

Ampliación del conjunto de datos

A modo de recordatorio, el conjunto de datos se ha sacado de la web Kaggle[27], pero la información ha sido recogida por el Banco Mundial[4] por lo que en su web se pueden encontrar una amplia gama de variables adicionales clasificadas por país y año. De esta manera, se podrían elegir nuevas variables relevantes en el análisis de Desarrollo Humano Global que puedan mejorar el modelo. Algunos campos a tener en cuenta podrían ser:

1. **Variables socieconómicas:** información sobre educación, empleo o renta per cápita.
2. **Variables medioambientales:** como calidad del aire o recursos y desastres naturales.

Análisis político

En una línea similar al anterior punto, incluir variables relacionadas con la politica (educación, sanidad...) para posteriormente evaluar las decisiones políticas y su nivel de impacto en el desarrollo humano. Tras este análisis se podrían elaborar una serie de propuestas para mejorar los países cuya esperanza de vida es más baja.

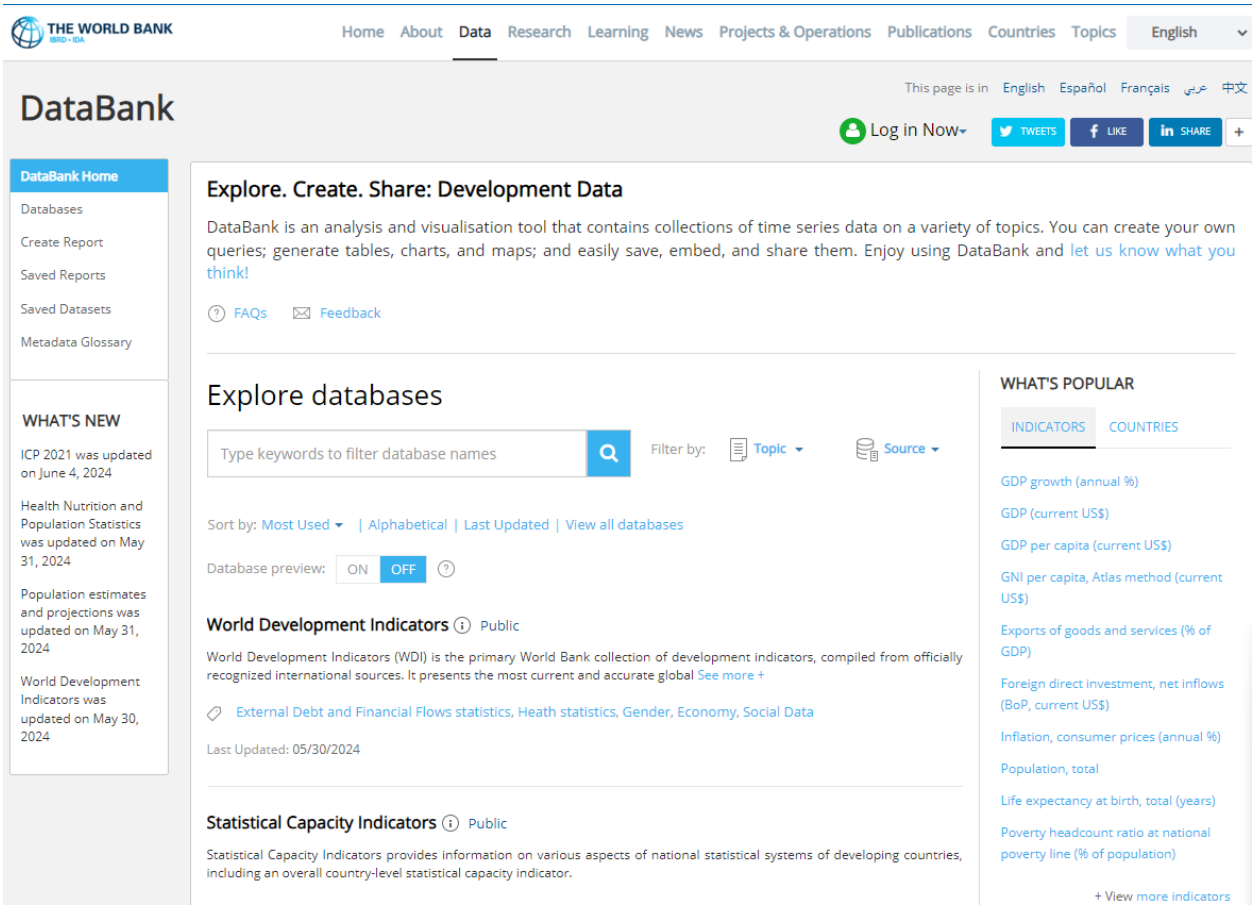

Figura 5.1: Página Web del Banco Mundial [4] mencionada en la ampliación del conjunto de datos de las líneas futuras

La Figura 5.1 pertenece a la web del Banco Mundial [4], más concreto, en este apartado se pueden consultar todas las bases de datos disponibles y sus variables. De aquí se han sacado las descripciones de cada variable de este proyecto.

Bibliografía

[1] K. Griffin. (2001) DESARROLLO HUMANO: Origen, evolución e impacto. [Online]. Available: https://www.yorksj.ac.uk/media/content-assets/social-economy/documents/GriffinDesarrolloHumano.pdf

[2] PNUD. (2016) Los ODS en acción. [Online]. Available: https://www.undp.org/es/sustainable-development-goals

[3] (2021) El orden mundial. [Online]. Available: https://elordenmundial.com/mapas-y-graficos/

[4] WorldBankGroup. (2023) Databank worlbank. [Online]. Available: https://databank.worldbank.org/

[5] Rstudio. (2006) Rstudio. [Online]. Available: https://www.rdatamining.com/

[6] Wikipedia. (2024) Desarrollo Humano. [Online]. Available: https://es.wikipedia.org/wiki/Desarrollo_humano

[7] . P. M. F. A. B. y Prof. Mgter. (2021) EL DESARROLLO HUMANO CONCEPTO E INDICADORES. [Online]. Available: https://hum.unne.edu.ar/revistas/geohoy/contenidos/geohoy03/peclasbonf_3.pdf

[8] PNUD. (2012) PROGRAMA DE LAS NACIONES UNIDAS PARA EL DESARROLLO. [Online]. Available: https://www.undp.org/es/sustainable-development-goals

[9] Iberdrola. (2020) Índice de Desarrollo Humano. [Online]. Available: https://www.iberdrola.com/compromiso-social/indice-desarrollo-humano#:~:text=Se%20eval%C3%BAa%20a%20trav%C3%A9s%20de,y%20una%20m%C3%A1xima%20de%2085.

[10] G. de España. (2017) Informes sobre Desarrollo Humano. [Online]. Available: https://www.miteco.gob.es/es/ceneam/recursos/pag-web/informes-ambientales/onu.html

[11] Programa de las Naciones Unidas para el Desarrollo. (2023) Informe sobre desarrollo humano 2023/24. [Online]. Available: https://report.hdr.undp.org/?_gl=1*11uwp03*_ga*MTY3MDQwOTc1MC4xNjk5MDI0MDQw*_ga_3W7LPK0WP1*MTcxMDQ1Nzg5Mi4xNTQuMC4xNzEwNDU3ODk0LjU4LjAuMA..#

[12] IBM. (2021) Análisis de datos exploratorio. [Online]. Available: https://www.ibm.com/es-es/topics/exploratory-data-analysis

[13] Y. Zhao. (2011) R and data mining. [Online]. Available: https://www.rdatamining.com/

[14] J. G. . L. N. Daniel McClure, Daniel Wheeler. (2023) Correlación e información mutua. [Online]. Available: https://espanol.libretexts.org/Ingenieria/Ingenier%C3%ADa_Industrial_y_de_Sistemas/Libro%3A_Din%C3%A1mica_y_Controles_de_Procesos_Qu%C3%ADmicos_(Woolf)/13%3A_Estad%C3%ADsticas_y_antecedentes_probabil%C3%ADsticos/13.13%3A_Correlaci%C3%B3n_e_informaci%C3%B3n_mutua

[15] K. P.Murphy. (2012) Murphy - machine learning probabilistic perspective.

[16] C. G. Martínez. (2018) AnÁlisis de componentes principales (pca). [Online]. Available: https://rpubs.com/Cristina_Gil/PCA

[17] P. S. Foundation. (2014) Python. [Online]. Available: https://www.python.org/

[18] Varios. (2023) Pandas. [Online]. Available: https://pandas.pydata.org/

[19] P. Comunity. (2023) Numpy. [Online]. Available: https://numpy.org/

[20] T. M. D. Team. (2021) Matplotlib. [Online]. Available: https://matplotlib.org/

[21] Microsoft. (2024) ¿Qué es Power BI? [Online]. Available: https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview

[22] L. H. T. L. P. K. T. C. W. K. W. H. Y. D. D. Hadley Wickham, Winston Chang. (2016) ggplot2. [Online]. Available: https://ggplot2.tidyverse.org/

[23] L. H. K. M. D. V. Hadley Wickham, Romain François. (2023) Package 'dplyr'. [Online]. Available: https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

[24] A. A. c. Baptiste Auguie [aut, cre]. (2017) gridextra: Miscellaneous functions for "grid"graphics. [Online]. Available: https://cran.r-project.org/web/packages/gridExtra/index.html

[25] C. Hennig. (2023) fpc: Flexible procedures for clustering. [Online]. Available: https://cran.r-project.org/web/packages/fpc/index.html

[26] P. E. Meyer. (2022) Infotheo: Information-theoretic measures. [Online]. Available: https://cran.r-project.org/web/packages/infotheo/index.html

[27] R. Whitcomb. (2016) Global Development CSV File. [Online]. Available: https://corgis-edu.github.io/corgis/csv/global_development/

[28] W. B. Group. (2019) Metadata glossary. [Online]. Available: https://databank.worldbank.org/metadataglossary/World-Development-Indicators/series

[29] Rdocumentation. (2023) cut: Convert numeric to factor. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut

[30] Rdocumentation. (2023) hist: Histograms. [Online]. Available: https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/hist

[31] Wikipedia. (2021) Análisis de grupos. [Online]. Available: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_grupos

[32] J. Pele. (2020) Dbscan. [Online]. Available: https://search.r-project.org/CRAN/refmans/bios2mds/html/sil.score.html

[33] Wikipedia. (2023) K-medoids. [Online]. Available: https://es.wikipedia.org/wiki/K-medoids

[34] J. A. A. Godoy. (2024) Métodos de particionamiento. [Online]. Available: https://rpubs.com/JairoAyala/MP

[35] J. Soto. (2020) Dbscan. [Online]. Available: https://rpubs.com/elias_jurgen/605966

[36] J. A. A. Godoy. (2024) K-vecinos más cercanos. [Online]. Available: https://rpubs.com/JairoAyala/KNN

[37] J. B. M. Vega. (2018) Arboles de decisión con r - clasificación. [Online]. Available: https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

[38] Wikipedia. (2024) Coeficiente de gini. [Online]. Available: https://es.wikipedia.org/wiki/Coeficiente_de_Gini

[39] ——. (2023) Conditional mutual information. [Online]. Available: https://en.wikipedia.org/wiki/Conditional_mutual_information

[40] C. Zelada. (2017) Evaluación de modelos de clasificación. [Online]. Available: https://rpubs.com/chzelada/275494

Anexos

Apéndice A

Anexo

A.1. Primer Anexo: Cuadros

En este anexo se adjuntaran algunos cuadros a los que se ha hecho referencia en el TFG.

Cuadro A.1: Distribución de Países por Clusters K-Medoids

Country	Cluster	Country	Cluster
Afghanistan	1	Malawi	1
Angola	1	Mali	1
Benin	1	Mauritania	1
Botswana	1	Mozambique	1
Burkina Faso	1	Namibia	1
Burundi	1	Niger	1
Cameroon	1	Nigeria	1
Central African Republic	1	Papua New Guinea	1
Chad	1	Rwanda	1
Comoros	1	Sao Tome and Principe	1
Congo, Dem. Rep.	1	Senegal	1
Congo, Rep.	1	Sierra Leone	1
Cote d'Ivoire	1	Solomon Islands	1
Djibouti	1	Somalia	1
Equatorial Guinea	1	South Africa	1
Ethiopia	1	Sudan	1
Gabon	1	Swaziland	1
Gambia, The	1	Tanzania	1
Ghana	1	Togo	1
Guinea	1	Uganda	1
Guinea-Bissau	1	Zambia	1
Kenya	1	Zimbabwe	1

Continuación en la siguiente página

Cuadro A.1 – *Continuación de la anterior página*

Country	**Cluster**	**Country**	**Cluster**
Lesotho	1	Liberia	1
Madagascar	1		
Albania	2	Malaysia	2
Algeria	2	Maldives	2
Antigua and Barbuda	2	Malta	2
Argentina	2	Mauritius	2
Aruba	2	Mexico	2
Australia	2	Mongolia	2
Austria	2	Morocco	2
Bahamas, The	2	Myanmar	2
Bahrain	2	Netherlands	2
Bangladesh	2	New Caledonia	2
Barbados	2	Nicaragua	2
Belize	2	Oman	2
Bolivia	2	Pakistan	2
Brunei Darussalam	2	Panama	2
Bulgaria	2	Paraguay	2
Cabo Verde	2	Peru	2
Cambodia	2	Philippines	2
Canada	2	Poland	2
Chile	2	Portugal	2
China	2	Puerto Rico	2
Colombia	2	Qatar	2
Costa Rica	2	Romania	2
Cuba	2	Saudi Arabia	2
Cyprus	2	Singapore	2
Denmark	2	Spain	2
Dominican Republic	2	Sri Lanka	2
Ecuador	2	St. Lucia	2
Egypt, Arab Rep.	2	St. Vincent and the Grenadines	2
Fiji	2	Suriname	2
Finland	2	Switzerland	2
France	2	Syrian Arab Republic	2
French Polynesia	2	Tonga	2
Germany	2	Trinidad and Tobago	2
Greece	2	Tunisia	2
Grenada	2	Turkey	2
Guam	2	United Arab Emirates	2
Guatemala	2	United Kingdom	2
Honduras	2	United States	2
Hungary	2	Uruguay	2
Iceland	2	Vanuatu	2
India	2	Venezuela, RB	2

Continuación en la siguiente página

Cuadro A.1 – *Continuación de la anterior página*

Country	**Cluster**	**Country**	**Cluster**
Indonesia	2	Vietnam	2
Iran, Islamic Rep.	2	Virgin Islands (U.S.)	2
Ireland	2	Israel	2
Italy	2	Jamaica	2
Japan	2	Jordan	2
Kiribati	2	Korea, Dem. Rep.	2
Korea, Rep.	2	Kuwait	2
Lao PDR	2	Lebanon	2
Libya	2		

Cuadro A.2: Distribución de Países por Clusters DbScan

Country	Cluster	Country	Cluster
Aruba	0	Bahrain	0
Bangladesh	0	Barbados	0
Botswana	0	Cuba	0
Gabon	0	India	0
Kiribati	0	Korea, Dem. Rep.	0
Maldives	0	Malta	0
Singapore	0	Solomon Islands	0
South Africa	0	Afghanistan	1
Angola	1	Benin	1
Bolivia	1	Burkina Faso	1
Burundi	1	Cameroon	1
Central African Republic	1	Chad	1
Comoros	1	Congo, Dem. Rep.	1
Congo, Rep.	1	Cote d'Ivoire	1
Djibouti	1	Equatorial Guinea	1
Ethiopia	1	Gambia, The	1
Ghana	1	Guinea	1
Guinea-Bissau	1	Kenya	1
Lao PDR	1	Lesotho	1
Liberia	1	Madagascar	1
Malawi	1	Mali	1
Mauritania	1	Mozambique	1
Myanmar	1	Namibia	1
Niger	1	Nigeria	1
Pakistan	1	Papua New Guinea	1
Rwanda	1	Sao Tome and Principe	1
Senegal	1	Sierra Leone	1
Somalia	1	Sudan	1
Swaziland	1	Tanzania	1
Togo	1	Uganda	1
Zambia	1	Zimbabwe	1
Albania	2	Algeria	2
Antigua and Barbuda	2	Argentina	2
Australia	2	Austria	2
Bahamas	2	Belize	2
Brunei Darussalam	2	Bulgaria	2
Cabo Verde	2	Cambodia	2
Canada	2	Chile	2
China	2	Colombia	2
Costa Rica	2	Cyprus	2
Denmark	2	Dominican Republic	2

Continuación en la siguiente página

Cuadro A.2 – *Continuación de la anterior página*

Country	**Cluster**	**Country**	**Cluster**
Ecuador	2	Egypt, Arab Rep.	2
Fiji	2	Finland	2
France	2	French Polynesia	2
Germany	2	Greece	2
Grenada	2	Guam	2
Guatemala	2	Honduras	2
Hungary	2	Iceland	2
Indonesia	2	Iran, Islamic Rep.	2
Ireland	2	Israel	2
Italy	2	Jamaica	2
Japan	2	Jordan	2
Korea, Rep.	2	Kuwait	2
Lebanon	2	Libya	2
Malaysia	2	Mauritius	2
Mexico	2	Mongolia	2
Morocco	2	Netherlands	2
New Caledonia	2	Nicaragua	2
Oman	2	Panama	2
Paraguay	2	Peru	2
Philippines	2	Poland	2
Portugal	2	Puerto Rico	2
Qatar	2	Romania	2
Saudi Arabia	2	Spain	2
Sri Lanka	2	St. Lucia	2
St. Vincent and the Grenadines	2	Suriname	2
Switzerland	2	Syrian Arab Republic	2
Tonga	2	Trinidad and Tobago	2
Tunisia	2	Turkey	2
United Arab Emirates	2	United Kingdom	2
United States	2	Uruguay	2
Vanuatu	2	Venezuela, RB	2
Vietnam	2	Virgin Islands (U.S.)	2

Cuadro A.3: Variables - diminutivos

Variable	Diminutivo
Country	C
Year	Y
Data.Health.Birth Rate	DHBR
Data.Health.Death Rate	DHDR
Data.Health.Fertility Rate	DHFR
Data.Health.Life Expectancy at Birth.Female	DHLEBF
Data.Health.Life Expectancy at Birth.Male	DHLEBM
Data.Health.Life Expectancy at Birth.Total	DHLEBT
Data.Health.Population Growth	DHPG
Data.Health.Total Population	DHTP
Data.Infrastructure.Mobile.Cellular Subscriptions	DIMCS
Data.Infrastructure.Mobile Cellular Subscriptions per 100 People	DIMCSPHP
Data.Infrastructure.Telephone.Lines	DITL
Data.Infrastructure.Telephone.Lines.per.100.People	DITLPHP
Data.Rural Development.Agricultural.Land	DRDAL
Data.Rural Development.Agricultural.Land.Percent	DRDALP
Data.Rural Development.Arable.Land	DRDAL
Data.Rural Development.Arable.Land.Percent	DRDALP
Data.Rural Development.Land.Area	DRDLA
Data.Rural Development.Rural.Population	DRDRP
Data.Rural Development.Rural.Population.Growth	DRDRPG
Data.Rural Development.Surface.Area	DRDSA
Data.Urban Development.Population.Density	DUDPD
Data.Urban Development.Urban.Population.Percent	DUDUPP
Data.Urban Development.Urban.Population.Percent.Growth	DUDUPPG

A.2. Segundo anexo: Código en R

En este anexo se adjuntará el código de R empleado para realizar el análisis.

Funciones

En primer lugar se van a adjuntar las funciones propias empleadas.

Listing A.1: Script funciones

```
##SE SACAN LAS 12 VARIABLES CON MENOS VARIANZA DE ENTRE LOS
   PARES MAS CORELACIONADOS

selec_variables <- function(matrix){

cor_matrix <- cor(matrix)
pares_ordenados <- which(cor_matrix != 1 &
   upper.tri(cor_matrix), arr.ind = TRUE)
```

```
top_pares_correlacionadas <-
   pares_ordenados[order(abs(cor_matrix[pares_ordenados]),
   decreasing = TRUE)[1:100], ]

variables <- data.frame(variable1 = character(), variable2=
   character(),stringsAsFactors = FALSE)
eliminar <- NULL
for(i in 1:nrow(top_pares_correlacionadas)){
  variable1 <-
     rownames(cor_matrix)[top_pares_correlacionadas[i,1]]
  variable2 <-
     colnames(cor_matrix)[top_pares_correlacionadas[i,2]]
  variables <- rbind(variables, data.frame(variable1,
     variable2))
}

for(j in 1:nrow(variables)){
  aux <- variables[j,1]
  aux2 <- variables[j,2]

  var1 <- var(matrix[[aux]])
  var2 <- var(matrix[[aux2]])
  if(var1 > var2 & !any(eliminar == variables[j,2]) &
     length(eliminar) < 12){
    eliminar <- c(eliminar,variables[j,2])
  } else if (!any(eliminar == variables[j,1]) &
     length(eliminar) < 12){
    eliminar <- c(eliminar,variables[j,1])
  }
}
return(eliminar)
}


#FUNCION PARA ESTANDARIZAR un valor
stand<- function(df){
  df<- (df-mean(df))/sd(df)
  return (df)
}

#FUNCION PARA TRANSFORMACION DE DATOS LOG(1+X)
transf <- function(var){
  var <- (log(var+1))
  return(var)
}

# -  Se estandarizan las variables
```

```
# -  Se discretizan las variables
# -  Se sacan histogramas de las variables originales y
   discretizadas y se guardan en un documento
source("C:/Users/Alejandro/Documents/TFG/R/func_stand.R")

func_histograms <- function(prep){
  stan<-subset(prep, select = -c(Country))
  disc<-stan
  pdf("./GRAFICAS/histogramas.pdf")
  for(col in names(columnas)){
    h<-ggplot(prep,aes(x=prep[[col]])) +
       geom_histogram(fill="grey",color="black") +
       labs(title=paste("Histograma de",col),x =
       col,y="Frecuencia")
    stan[[col]]<-stand(prep[[col]])
    his<-hist(x=stan[[col]])

    cuantiles <-
       quantile(stan[[col]],probs=c(0,0.20,0.4,0.6,0.8,1))

    disc[[col]] <- discrete(disc[[col]],stan[[col]],his$breaks)

    # disc[[col]]<-cut(stan[[col]],breaks = his$breaks,
       include.lowest = TRUE,ordered_result = TRUE,
       include_min=TRUE, labels = FALSE)

    hprep<-ggplot(disc,aes(x=disc[[col]])) +
       geom_bar(fill="blue",color="black") +
       labs(title=paste("Histograma de",col),x =
       col,y="Frecuencia")
    hists<-grid.arrange(h,hprep,ncol = 1)


    print(hists)


  }
  dev.off()
  return(disc)
}

#FUNCION PARA ESTANDARIZAR EL DATASET
estandarizar <- function(matrix){
  stan<-subset(prep, select = -c(Country))
  for(col in names(columnas)){
    stan[[col]]<-stand(prep[[col]])
  }
```

```
83
84    return(stan)
85  }
86
87  #FUNCION PARA DISCRETIZAR SEGÚN BREAKS
88  discrete <- function(df,df2,breaks){
89    df <- cut(df2,breaks,include.lowest=TRUE, ordered_result =
          TRUE, include_min=TRUE, labels = FALSE)
90    return(df)
91  }
92
93  # -  SE QUITA EL AGNO
94  # -  SE HACE LA MEDIA DE LAS VARIABLES NUMERICAS
95  # -  POR ULTIMO SE HACE UN DISTINCT PARA SACAR UNA UNICA FILA
       POR PAIS
96  library("dplyr")
97  vars_nums <- function(dataset_s21){
98    dataset_resum <- dataset_s21 %>%
99      group_by(Country) %>%
100     summarize(
101       Country=Country,
102       H.Birth.Rate=mean(H.Birth.Rate),
103       H.Death.Rate=mean(H.Death.Rate),
104       H.Fertility.Rate=mean(H.Fertility.Rate),
105       H.Life.Expectancy.at.Birth.Female=
106         mean(H.Life.Expectancy.at.Birth.Female),
107       H.Life.Expectancy.at.Birth.Male=
108         mean(H.Life.Expectancy.at.Birth.Male),
109       H.Life.Expectancy.at.Birth.Total=
110         mean(H.Life.Expectancy.at.Birth.Total),
111       H.Population.Growth=mean(H.Population.Growth),
112       H.Total.Population=mean(H.Total.Population),
113       I.Mobile.Cellular.Subscriptions=
114         mean(I.Mobile.Cellular.Subscriptions),
115       I.Mobile.Cellular.Subscriptions.per.100.People=
116         mean(I.Mobile.Cellular.Subscriptions.per.100.People),
117       RD.Agricultural.Land=mean(RD.Agricultural.Land),
118       RD.Agricultural.Land.Percent=
119         mean(RD.Agricultural.Land.Percent),
120       RD.Arable.Land=mean(RD.Arable.Land),
121       RD.Arable.Land.Percent=mean(RD.Arable.Land.Percent),
122       RD.Land.Area =mean(RD.Land.Area),
123       RD.Rural.Population=mean(RD.Rural.Population),
124       RD.Rural.Population.Growth=mean(RD.Rural.Population.Growth),
125       RD.Surface.Area=mean(RD.Surface.Area),
126       UD.Population.Density=mean(UD.Population.Density),
127       UD.Urban.Population.Percent
```

```
          =mean(UD.Urban.Population.Percent),
128       UD.Urban.Population.Percent.Growth=
129         mean(UD.Urban.Population.Percent.Growth)
130     )
131
132   return(distinct(dataset_resum))
133 }
134
135 #FUNCION PARA APLICAR INFORMACION CONDICIONAL MUTUA
136 CALCULAR_CMI <- function(var1, var2){
137   condinformation(disc_prep[[var1]],disc_prep[[var2]],
138     disc_prep[["H.Life.Expectancy.at.Birth.Total"]])
139 }
140
141 #FUNCION PARA EVALUAR MODELO CLASIFICATORIO
142 evaluacion <- function(predicciones, train) {
143   confusion_matrix <- table(Predicted = predicciones, Actual =
        train$H.Life.Expectancy.at.Birth.Total)
144   accuracy <- sum(diag(confusion_matrix)) /
        sum(confusion_matrix)
145   precision <- diag(confusion_matrix) /
        rowSums(confusion_matrix)
146   recall <- diag(confusion_matrix) / colSums(confusion_matrix)
147   f1_score <- 2 * (mean(precision) * mean(recall)) /
        (mean(precision) + mean(recall))
148   metrics <- c(precisión = accuracy, tasa_error = 1-accuracy,
        precisión = mean(precision), recall = mean(recall),
        f1_score = mean(f1_score))
149   return(metrics)
150 }
```

Resto de código del análisis

Listing A.2: Bibliotecas e importación de dataset

```
1  #Se importan librerias que se utilizaran a lo largo de todo el
       trabajo
2  library("entropy")
3  library("infotheo")
4  library("data.table")
5  library("dplyr")
6  library("ggplot2")
7  library("gridExtra")
8  library("fpc")
9  library("caret")
10 library("dbscan")
11 library(cluster)
12 library("class")
```

```
library("rpart")
library("rpart.plot")
library("ROCR")
source("R/algoritmo.R")
library("factoextra")

#Se importa el dataset
dataset <- read.csv('C:/Users/Alejandro/Documents/TRABAJO FIN GRADO/global_development.csv', header = TRUE, sep = ",")
attach(dataset)

#Se crea una carpeta donde se almacenaran los histogramas
if(!dir.exists("GRAFICAS")){
  dir.create("GRAFICAS")
}
```

Listing A.3: Preparación de datos

```
#se crean acortamientos de las variables
acortamientos_de_variables=c('HBR','HDR','HFR','HPG','HTP','IMCS',
    'IMCS100','RDAGL','RDAGLP','RDARL','RDARLP','RDLA','RDRP',
    'RDRPG','RDSA', 'UDPD','UDUPP','UDUPPG')

#Se renombran las variables
names(dataset)[3] = 'H.Birth.Rate'
names(dataset)[4] = 'H.Death.Rate'
names(dataset)[5] = 'H.Fertility.Rate'
names(dataset)[6] = 'H.Life.Expectancy.at.Birth.Female'
names(dataset)[7] = 'H.Life.Expectancy.at.Birth.Male'
names(dataset)[8] = 'H.Life.Expectancy.at.Birth.Total'
names(dataset)[9] = 'H.Population.Growth'
names(dataset)[10]= 'H.Total.Population'
names(dataset)[11]= 'I.Mobile.Cellular.Subscriptions'
names(dataset)[12]=
   'I.Mobile.Cellular.Subscriptions.per.100.People'
names(dataset)[13]= 'I.Telephone.Lines'
names(dataset)[14]= 'I.Telephone.Lines.per.100.People'
names(dataset)[15]= 'RD.Agricultural.Land'
names(dataset)[16]= 'RD.Agricultural.Land.Percent'
names(dataset)[17]= 'RD.Arable.Land'
names(dataset)[18]= 'RD.Arable.Land.Percent'
names(dataset)[19]= 'RD.Land.Area'
names(dataset)[20]= 'RD.Rural.Population'
names(dataset)[21]= 'RD.Rural.Population.Growth'
names(dataset)[22]= 'RD.Surface.Area'
names(dataset)[23]= 'UD.Population.Density'
names(dataset)[24]= 'UD.Urban.Population.Percent'
names(dataset)[25]= 'UD.Urban.Population.Percent.Growth'
```

```
30
31 #Variables a utilizar ya que se eliminan
      I.Mobile.Cellular.Subscriptions y
      I.Mobile.Cellular.Subscriptions.per.100.People
32 columnas <-
      dataset[,c("H.Birth.Rate","H.Death.Rate","H.Fertility.Rate",
33 "H.Life.Expectancy.at.Birth.Female",
34 "H.Life.Expectancy.at.Birth.Male",
35 "H.Life.Expectancy.at.Birth.Total","H.Population.Growth",
36 "H.Total.Population","I.Mobile.Cellular.Subscriptions",
37 "I.Mobile.Cellular.Subscriptions.per.100.People",
38 "RD.Agricultural.Land","RD.Agricultural.Land.Percent",
39 "RD.Arable.Land","RD.Arable.Land.Percent","RD.Land.Area",
40 "RD.Rural.Population","RD.Rural.Population.Growth",
41 "RD.Surface.Area","UD.Population.Density",
42 "UD.Urban.Population.Percent",
43 "UD.Urban.Population.Percent.Growth")]
44
45
46 #Se eliminan las agrupaciones de paises
47 agrupaciones_paises <-c("Other small states", "Middle East &
      North Africa (developing only)", "Central Europe and the
      Baltics",
48 "Arab World", "Middle East & North Africa (all income levels)",
      "High income: nonOECD", "North America", "Euro area",
49 "Latin America & Caribbean (all income levels)", "Sub-Saharan
      Africa (all income levels)", "Least developed countries: UN
      classification",
50 "Upper middle income", "High income: OECD", "OECD members",
      "High income  ", "Lower middle income", "Middle income",
51 "Fragile and conflict affected situations", "Heavily indebted
      poor countries (HIPC)", "Latin America & Caribbean
      (developing only)",
52 "Sub-Saharan Africa (developing only)", "Low income", "European
      Union", "South Asia", "East Asia & Pacific (developing
      only)",
53 "East Asia & Pacific (all income levels)", "Low & middle
      income", "Caribbean small states", "Pacific island small
      states","World","High income", "Small states")
54
55 #Se cogen los datos de la ultima decada y se quitan las
      agrupaciones de paises
56
57 dataset_s21 <-subset(dataset, dataset$Year>=2004
58                       & !dataset$Country %in%
                             agrupaciones_paises)
59
```

```
#SE SACAN LOS DATOS CON UNA uNICA FILA POR PAIS (SE QUITA EL AGNO)
prep <- vars_nums(dataset_s21)

#SE REALIZA LA TRANSFORMACIoN DE DATOS: LOG(1+X) PARA CIERTAS VARIABLES
prep$H.Total.Population <- transf(prep$H.Total.Population)
prep$I.Mobile.Cellular.Subscriptions <- transf(prep$I.Mobile.Cellular.Subscriptions)
prep$RD.Agricultural.Land <- transf(prep$RD.Agricultural.Land)
prep$RD.Arable.Land <- transf(prep$RD.Arable.Land)
prep$RD.Rural.Population <- transf(prep$RD.Rural.Population)
prep$RD.Surface.Area <- transf(prep$RD.Surface.Area)
prep$UD.Population.Density <- transf(prep$UD.Population.Density)
prep$RD.Land.Area <- transf(prep$RD.Land.Area)


#Se sacan los histogramas de las variables iniciales y discretizadas y se almacena en disc el dataframe discretizado
disc <- func_histograms(prep)

#Se estandariza el df
standard <- estandarizar(prep)
```

Listing A.4: Clustering

```
##############CLUSTERING K-MEDOIDS#################

#Variables que se utilizaran en el clustering K-Medoids
df_clust<-standard[,c("H.Birth.Rate","H.Death.Rate",
"H.Fertility.Rate","H.Life.Expectancy.at.Birth.Female"
,"H.Life.Expectancy.at.Birth.Male",
"H.Life.Expectancy.at.Birth.Total"
,"H.Population.Growth","H.Total.Population")]


#ELBOW METHOD PARA NUMERO DE CLUSTERS K-Medoids
n_cluster<-fviz_nbclust(df_clust,pam,method = "wss")

#cluster K-Medoids
pamk.result<-pam(df_clust,k=2)

#PCA
datos_pca <- prcomp(standard,scale=TRUE)
pca_results <- data.frame(datos_pca$x[, 1:2])
pca_results$cluster<-pamk.result$clustering
pca_results$Country<-prep$Country

```

```
#Se meten los resultados del clustering en el dataset y tambien en el dataset con paises
df_clust_pam<-subset(prep, select = -c(Country))
df_clust_pam$cluster<-pamk.result$clustering
df_clust_pam_paises<-prep
df_clust_pam_paises$Cluster<-pamk.result$clustering

#visualizacion del cluster
cluster_plot_2 <- fviz_cluster(pamk.result, data = df_clust)+
  theme_minimal()
#Grafica silhouette
cluster_plot_3 <- fviz_silhouette(silhouette(pamk.result))
#visualizacion datos con clusters
cluster_plot_4 <-ggplot(pca_results, aes(x = PC1, y = PC2, color = as.factor(
  cluster))) +
   geom_point() +
   theme_minimal() +
   labs(color = "Cluster")

# Crear un boxplot para comparar la distribucion de H.Life.Expectancy.at.Birth.Total entre los clusters
cluster_plot_5<-ggplot(df_clust_pam, aes(x = factor(cluster), y = H.Life.Expectancy.at.Birth.Total, fill = factor(cluster))) +
  geom_boxplot() +
   labs(title = "Distribucion de la variable por cada cluster", x = "
Cluster", y = "Esperanza de vida al nacer") +
   theme_minimal()

# Crear un boxplot para comparar la distribucion de H.Total.Population entre los clusters
cluster_plot_6<-ggplot(df_clust_pam, aes(x = factor(cluster), y = H.Total.Population, fill = factor(cluster))) +
  geom_boxplot() +
  labs(title = "Distribucion de la variable por cada cluster", x = "
Cluster", y = "Poblacion total") +
  theme_minimal()

#Lista de paises y su correspondiente cluster de K-Medoids
pais_cluster<- pca_results%>%
  select(Country,cluster)%>%
  arrange(cluster,Country)

#SE CALCULAN MEDIAS DE VARIAS VARIABLES SIGNIFICATIVAS PARA
```

```r
   CADA CLUSTER K-MEDOIDS
medias_por_cluster_PAM <- df_clust_pam_paises %>%
  group_by(Cluster) %>%
  summarize(
    #MEDIA TASA DE NACIMIENTO CLUSTER PAM
    media_H_Birth_Rate_pam = mean(H.Birth.Rate, na.rm = TRUE),
    #MEDIA ESPERANZA DE VIDA CLUSTER PAM
    media_H_Life_Expectancy_at_Birth_Total_pam =
       mean(H.Life.Expectancy.at.Birth.Total, na.rm = TRUE),
    #MEDIA ESPERANZA DE VIDA CLUSTER PAM
    media_H_Death_Rate_pam = mean(H.Death.Rate, na.rm = TRUE),
    #MEDIA CRECIMIENTO POBLACION CLUSTER PAM
    media_H_Population_Growth_pam = mean(H.Population.Growth,
       na.rm = TRUE),
    #MEDIA URBAN POPUL CLUSTER PAM
    media_UD_Urban_Population_Percent_pam =
       mean(UD.Urban.Population.Percent, na.rm = TRUE)
  )
print(medias_por_cluster_PAM)


####################CLUSTERING DBSCAN#######################


#Variables que se utilizaran en el clustering DbScan
df_clust_D<-standard[,c(
  "H.Life.Expectancy.at.Birth.Total",
  "I.Mobile.Cellular.Subscriptions.per.100.People",
  "UD.Population.Density")]


#ALGORITMO PARA SACAR MEJOR EPS Y MINPTS
minPts_values <- 5:15
eps_values <- seq(0.1, 3, by = 0.1)


results <- list()
silhouette_scores <- numeric(length(minPts_values) *
   length(eps_values))
set.seed(123)
# Iterar sobre todas las combinaciones de minPts y eps
for (minPts in minPts_values) {
  for (eps in eps_values) {
    # Aplicar DBSCAN
    dbscan_result <- dbscan(df_clust_D, eps = eps, minPts =
       minPts)

```

```
102     # Guardar los resultados en la lista
103     results[[paste("minPts", minPts, "eps", eps)]] <-
            dbscan_result
104
105     #comprobar que hay mas de dos cluster
106     if(length(unique(dbscan_result$cluster)) >2){
107
108     # Calcular la Silueta para cada combinacion y guardarla
109     sil <- silhouette(dbscan_result$cluster, dist(df_clust_D))
110     sil_score<-mean(sil[,3])
111     }else{
112       sil_score<-NA
113     }
114     silhouette_scores[[paste("minPts", minPts, "eps", eps)]] <-
            sil_score
115   }
116 }
117 #Se calcula la mejor combinacion de minPts y eps y su valor de
        Silhouette
118 best_combination <-
        names(silhouette_scores)[which.max(silhouette_scores)]
119 best_silhouette <- silhouette_scores[best_combination]
120 print(best_combination)
121 print(best_silhouette)
122
123 #UNA VEZ CON MEJOR MINPTS Y EPS SE REALIZA EL CLUSTERING DbScan
124 set.seed(123)
125 dbscan_resultados <- dbscan(df_clust_D, eps = 0.6, minPts = 13)
126
127 #Silhouette
128 db_plot_1 <-
        fviz_silhouette(silhouette(dbscan_resultados$cluster,
        dist(df_clust_D)))
129
130 #Grafica cluster DbScan
131 db_plot_2 <- fviz_cluster(dbscan_resultados, data = standard)
132
133 #Se meten los resultados del clustering en el dataset y tambien
        en el dataset con paises
134 df_clust_scan<-subset(prep, select = -c(Country))
135 df_clust_scan$cluster<-dbscan_resultados$cluster
136 df_clust_scan_paises<-prep
137 df_clust_scan_paises$Cluster<-dbscan_resultados$cluster
138
139 #Lista de paises con su correspondiente cluster DbScan
140 pais_cluster_Db<- df_clust_scan_paises%>%
141   select(Country,Cluster)%>%
```

```
142   arrange(Cluster,Country)
143
144 #SE CALCULAN MEDIAS DE VARIAS VARIABLES SIGNIFICATIVAS PARA
      CADA CLUSTER DBSCAN
145 medias_por_cluster_DBSCAN <- df_clust_scan_paises %>%
146   group_by(Cluster) %>%
147   summarize(
148     #MEDIA TASA DE NACIMIENTO CLUSTER PAM
149     media_H_Birth_Rate_dbscan = mean(H.Birth.Rate, na.rm =
          TRUE),
150
151     #MEDIA ESPERANZA DE VIDA CLUSTER PAM
152     media_H_Life_Expectancy_at_Birth_Total_dbscan =
          mean(H.Life.Expectancy.at.Birth.Total, na.rm = TRUE),
153
154     #MEDIA ESPERANZA DE VIDA CLUSTER PAM
155     media_H_Death_Rate_dbscan = mean(H.Death.Rate, na.rm =
          TRUE),
156
157     #MEDIA CRECIMIENTO POBLACION CLUSTER PAM
158     media_H_Population_Growth_dbscan =
          mean(H.Population.Growth, na.rm = TRUE),
159
160     #MEDIA URBAN POPUL CLUSTER PAM
161     media_UD_Urban_Population_Percent_dbscan =
          mean(UD.Urban.Population.Percent, na.rm = TRUE),
162
163     #MEDIA moviles por cada 100 personas CLUSTER dbscan
164     media_I_Mobile_Per100_dbscan =
          mean(I.Mobile.Cellular.Subscriptions.per.100.People,
          na.rm = TRUE),
165
166     #MEDIA DENSIDAD DE POBLACION CLUSTER PAM
167     media_UD_Pop_Density_dbscan = mean(UD.Population.Density,
          na.rm = TRUE)
168   )
169 print(medias_por_cluster_DBSCAN)
170
171 # Crear un boxplot para comparar la distribucion de
      H.Life.Expectancy.at.Birth.Total entre los clusters DbScan
172 cluster_plot_7<-ggplot(df_clust_scan, aes(x = factor(cluster),
      y = H.Life.Expectancy.at.Birth.Total, fill =
      factor(cluster))) +
173   geom_boxplot() +
174   labs(title = "Distribucion de la variable por cada cluster",
        x = "
175 Cluster", y = "Esperanza de vida al nacer") +
```

```
176   theme_minimal()
177
178 # Crear un boxplot para comparar la distribucion de
      I.Mobile.Cellular.Subscriptions.per.100.People entre los
      clusters DbScan
179 cluster_plot_8<-ggplot(df_clust_scan, aes(x = factor(cluster),
      y = I.Mobile.Cellular.Subscriptions.per.100.People, fill =
      factor(cluster))) +
180   geom_boxplot() +
181   labs(title = "Distribucion de la variable por cada cluster",
        x = "
182 Cluster", y = "Suscripciones de movil por cada 100 habitantes")
      +
183   theme_minimal()
184
185 # Crear un boxplot para comparar la distribucion de
      UD.Population.Density entre los clusters DbScan
186 cluster_plot_9<-ggplot(df_clust_scan, aes(x = factor(cluster),
      y = UD.Population.Density, fill = factor(cluster))) +
187   geom_boxplot() +
188   labs(title = "Distribucion de la variable por cada cluster",
        x = "
189 Cluster", y = "Densidad de poblacion") +
190   theme_minimal()
```

Listing A.5: Análisis clasificatorio y predictivo

```
1  ###############ANALISIS CLASIFICATORIO###################
2  #SE ELIMINAN LAS VARIABLES H.Life.Expectancy.at.Birth.Female y
     H.Life.Expectancy.at.Birth.Male
3  disc_prep<- select(disc, -c("H.Life.Expectancy.at.Birth.Female",
4  "H.Life.Expectancy.at.Birth.Male"))
5
6  # Informacion mutua de cada variable predictora con la variable
     objetivo
7  mut_obj <- sapply(disc_prep[, colnames(disc_prep) !=
     "H.Life.Expectancy.at.Birth.Total"], function(x) {
8    mutinformation(disc_prep[["H.Life.Expectancy.at.Birth.Total"]],
       x)
9  })
10
11 #se convierte en un dataframe para hacer un grafico
12 mut_obj_df<-data.frame(mut_obj)
13
14 #se acortan los nombres a sus iniciales para ver mejor el
     grafico
15 rownames(mut_obj_df)<-acortamientos_de_variables
16
```

```
#Se almacenan las variables y su informacion mutua correspondiente
variables<-rownames(mut_obj_df)
valores<-mut_obj_df$mut_obj

#Se realiza grafico con cada valor de la informacion mutua con variable objetivo
graf_mut<-ggplot(mut_obj_df,aes(x=variables,y=valores)) + geom_bar(stat='identity',fill='skyblue')+ggtitle("Informacion mutua con variable objetivo")+xlab("Variables") + ylab("Info mutua") + theme_minimal()
graf_mut + geom_hline(yintercept=0.2754715,color="red",linetype="dashed", size=1)

# Se coge la mediana de los valores elegidos anteriormente y se eligen las variables por encima de este umbral
mut_obj_median <- median(mut_obj)
mut_obj_altas <- names(mut_obj[mut_obj > mut_obj_median])

# Calcula la informacion mutua condicionada entre pares de variables elegidas
parejas <- combn(mut_obj_altas, 2)

#SE CALCULA CMI Y SE METE EN UN DATAFRAME CON SU PAREJA CORRSPONDIENTE Y SE ORDENA DE MAYOR A MENOR
valores_cmi<-mapply(CALCULAR_CMI,parejas[1,],parejas[2,])
mut_comb_df<-data.frame(Var1=parejas[1,],Var2=parejas[2,],
                        CMI=valores_cmi)
mut_comb_df<-mut_comb_df[order(-mut_comb_df$CMI),]

#Se almacenan las variables y su informacion mutua condicional correspondiente
variables_2<-rownames(mut_comb_df)
valores_2<-mut_comb_df$CMI

#Se realiza grafico con cada valor de la informacion mutua con variable objetivo
ggplot(mut_comb_df,aes(x=variables_2,y=valores_2)) + geom_bar(stat='identity',fill='skyblue')+ggtitle("Informacion mutua condicionada")+xlab("Parejas") + ylab("Info mutua condicional") + theme_minimal()

# Selecciona los pares con la menor informacion mutua condicionada entre las variables elegidas
pareja_selecc_bajo <- tail(mut_comb_df,5)
pareja_selecc_bajo
```

```
49
50 #Se seleccionan los pares de variables con mayor informacion
      mutua condicionada
51 pareja_selecc_alto <-head(mut_comb_df,8)
52 pareja_selecc_alto
53
54 #Se almacenan en una lista todos los nombres de las parejas de
      las 5 parejas con CMI mas alta
55 nombres_parejas_cmi<-c(pareja_selecc_alto$Var1,pareja_selecc_alto$Var2)
56 nombres_parejas_cmi<-unique(nombres_parejas_cmi)
57
58
59
60 #SE COMPRUEBAN LOS CMI DE TODAS LAS VARIABLES ELEGIDAS PARA
      COMPROBAR QUE NO SEA BAJA NINGUNA
61 parejas_var_selecc<-combn(nombres_parejas_cmi, 2)
62 valores_cmi_var_sel<-mapply(CALCULAR_CMI,parejas_var_selecc[1,],
63                             parejas_var_selecc[2,])
64
65 mut_comb_varsel<-data.frame(Var1=parejas_var_selecc[1,],
66             Var2=parejas_var_selecc[2,],CMI=valores_cmi_var_sel)
67
68 mut_comb_varsel<-mut_comb_varsel[order(-mut_comb_varsel$CMI),]
69
70
71
72 #se meten las variables objetivo y predictoras en una formula
      para no meterlo a mano en rpart
73 #Se itera quitando 1 variable en cada momento hasta quedarnos
      con 2
74 #variables para conseguir la mejor combinacion
75 set.seed(123)
76 for(i in 0:5){
77   subvars<-nombres_parejas_cmi[1:(length(nombres_parejas_cmi)-i)]
78   var_predict<-paste(subvars,collapse=" + ")
79   vars_rpart<-as.formula(paste("H.Life.Expectancy.at.Birth.Total
        ~",var_predict))
80   clasif_arbol <- rpart(vars_rpart, data = entrenamiento,
        method = "class")
81   prediccion <- predict(clasif_arbol, prueba, type = "class")
82   matriz_conf_arbol <- table(Predicted = prediccion, Actual =
        prueba$H.Life.Expectancy.at.Birth.Total)
83   exactitud_arbol <- sum(diag(matriz_conf_arbol)) /
        sum(matriz_conf_arbol)
84   print(subvars);
85   print(exactitud_arbol);
86 }
```

```
87
88 #Se itera quitando 1 variable distinta cada vez para conseguir
      la mejor combinacion de 1 variable menos
89 set.seed(123)
90 for( variable in nombres_parejas_cmi){
91   subvars<-nombres_parejas_cmi[nombres_parejas_cmi!=variable]
92   var_predict<-paste(subvars,collapse=" + ")
93   vars_rpart<-as.formula(paste("H.Life.Expectancy.at.Birth.Total
        ~",var_predict))
94   clasif_arbol <- rpart(vars_rpart, data = entrenamiento,
        method = "class")
95   prediccion <- predict(clasif_arbol, prueba, type = "class")
96   matriz_conf_arbol <- table(Predicted = prediccion, Actual =
        prueba$H.Life.Expectancy.at.Birth.Total)
97   exactitud_arbol <- sum(diag(matriz_conf_arbol)) /
        sum(matriz_conf_arbol)
98   print(subvars);
99   print(exactitud_arbol);
100 }
101
102 #Se concatenan la variable objetivo con la variable predictora
103 var_predict<-paste(nombres_parejas_cmi,collapse=" + ")
104 vars_rpart<-as.formula(paste("H.Life.Expectancy.at.Birth.Total
      ~",var_predict))
105
106 #Se obtiene el dataset sobre el que se realizara el analisis
107 #se obtienenen las clases de la variable objetivo
108 df_clasif <-standard
109 df_clasif$H.Life.Expectancy.at.Birth.Total <-
      cut(df_clasif$H.Life.Expectancy.at.Birth.Total,
110 breaks = quantile(df_clasif$H.Life.Expectancy.at.Birth.Total,
      probs = c(0, 1/3, 2/3, 1)), labels = c("Bajo", "Medio",
      "Alto"), include.lowest = TRUE)
111
112 #Se obtienen los datos de entrenamiento y prueba
113 set.seed(123)
114 muestra<- sample(1:146,100)
115 entrenamiento <-df_clasif[muestra,]
116 prueba <- df_clasif[-muestra,]
117
118 #Se aplica el algoritmo rpart
119 set.seed(123)
120 clasif_arbol <- rpart(vars_rpart, data = entrenamiento, method
      = "class")
121 prp(clasif_arbol,type=2,extra="auto")
122
123 #Prediccion del modelo
```

```
prediccion <- predict(clasif_arbol, prueba, type = "class")

#Se evalua el modelo
arbol_eval <- evaluacion(prediccion,prueba)

#ANALISIS CLASIFICATORIO KNN

#Se elige el mejor k (numero de vecinos mas cercanos) para
   realizar el analisis
set.seed(123)
res<-list(0)
for(i in 1:20){
clasif_knn <- knn(train = entrenamiento[,-6],test=prueba[,-6],
    cl=entrenamiento$H.Life.Expectancy.at.Birth.Total,k=i)

matriz_conf_knn <-
   table(prueba$H.Life.Expectancy.at.Birth.Total,clasif_knn,
    dnn=c("Esperado","Predicho"))

exactitud_knn <- sum(diag(matriz_conf_knn)) /
   sum(matriz_conf_knn)
  if(exactitud_knn>res[1]){
    res[1]<-exactitud_knn
    k<-i
  }
}
print(k)

#Se aplica el algoritmo knn
set.seed(123)
clasif_knn <- knn(train = entrenamiento[,-6],test=prueba[,-6],
    cl=entrenamiento$H.Life.Expectancy.at.Birth.Total,k=3)

#Se evalua el modelo knn
knn_eval<- evaluacion(clasif_knn,prueba)
```

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Sun Jun 30 16:43:56 CEST 2024
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)