COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Empresariales

Análisis del sector energético y el crecimiento de las energías renovables.

Autor: Ana María Chacón Rueda
Director: Lourdes Fernández Rodríguez

MADRID | Junio 2024

RESUMEN

El presente Trabajo de Fin de Grado pretende analizar el sector energético con un enfoque en las fuentes renovables. En los últimos años, las energías alternativas han ganado cada vez más presencia como respuesta al cambio climático y al alto coste de las energías convencionales. El objetivo principal es estudiar la evolución de la energía renovable a escala mundial y predecir su desarrollo en los próximos años.

Se ha realizado una reducción de dimensionalidad de los datos a través de Análisis de Componentes Principales para eliminar información redundante. Además, se han aplicado modelos de redes neuronales para predecir nuevos datos y obtener la proyección de la proporción de generación y uso de energía renovable frente al total.

Las predicciones obtenidas indican que la evolución de la cuota de energía renovable variará según el país: en algunos la tendencia es positiva mientras que en otros se mantendrá más estable. Sin embargo, estos resultados no son representativos a nivel global ya que solo se han obtenido las proyecciones para un número limitado, por lo que no se puede concluir que este comportamiento ocurra de manera generalizada para todos los países. A pesar de que los resultados son coherentes, se ha identificado la necesidad de ampliar el conjunto de datos o emplear modelos de regresión más sencillos a los utilizados para conseguir predicciones más precisas y evitar incurrir en *overfitting*.

Palabras clave: energías renovables, machine learning, predicción de energías renovables, evolución de energías renovables

SUMMARY

This project aims to analyze the energy sector with a focus on renewable energy. In recent years, alternative energies have gained increasing prominence in response to climate change and the high costs of conventional energy sources. The main objective is to study the evolution of renewable energy on a global scale and predict its development in next years.

Principal Component Analysis (PCA) was used to reduce data dimensionality and eliminate redundant information. Additionally, neural network models were employed to predict new data and project the proportion of renewable energy generation and usage relative to the total energy.

Predictions indicate that the evolution of the renewable energy share will vary by country: in some, the trend is positive, while in others, it will remain more stable. However, these results are not globally representative, as projections were only made for a limited number of countries, so it cannot be concluded that this behavior is generalized worldwide. Although the results are coherent, the need to expand the dataset or use simpler regression models has been identified to achieve more accurate predictions and avoid overfitting.

Keywords: renewable energy, machine learning, renewable energy prediction, renewable energy evolution

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1. Estudio de la cuestión

En los últimos años, la energía renovable ha experimentado un crecimiento a escala global. Esto se debe a la tendencia de disminuir las emisiones de carbono, frenar el cambio climático, y conseguir un precio y rendimiento similar al de las fuentes de energía convencionales, donde el precio de la energía solar ha disminuido en un 85% en solo diez años, desde 2010 a 2020. Sin embargo, todavía muchos países siguen siendo importadores de energías fósiles pudiendo ser ellos mismos los generadores de energía renovable.

En el presente trabajo se estudiará el sector energético con un enfoque particular en el crecimiento de la generación de las energías renovables a escala global y regional. Se examinará cómo las diferentes energías sostenibles como la solar, eólica, geotérmica, hidroeléctrica y oceánica han obtenido más presencia en el mercado energético y cuál es el futuro de estas.

A lo largo del trabajo se realizará un análisis exhaustivo de una base de datos (Data on Sustainable Energy (2000-2020)) que recopila información sobre indicadores energéticos para conocer y predecir la evolución de las energías renovables y qué variables son las más determinantes para conseguir una transición a energías verdes y sostenibles.

2. Marco Teórico

La energía es uno de los principales motores de la economía ya que contribuye al funcionamiento de las industrias y al desarrollo de innovaciones, que tiene como consecuencia un crecimiento económico. Sin embargo, el sector energético está sufriendo grandes cambios ya que las principales economías están invirtiendo para conseguir una transición energética y que las energías renovables sean la principal fuente de electricidad.

Según la Agencia Internacional de la Energías, se prevé que en el año 2030 el 65% de la electricidad mundial será generada por energías renovables, y que en el año 2050 estas llegarán a suministrar cerca del 90% de la electricidad global (Naciones Unidas, 2023). Por el momento, las energías eólica y solar han generado en 2021 un 10,3% de la electricidad mundial en comparación con el 4,6% en 2015 (Ember, 2022).

Otros estudios centrados en el sector energéticos han observado tendencias positivas en la transición hacia fuentes renovables a partir de técnicas de *machine learning*. Para ello, se ha definido un nuevo indicador compuesto Rastreador de Emisiones de Generación Eléctrica (EGE) que compara el rendimiento de sostenibilidad en la creación eléctrica de diferentes países. Esta investigación refleja de manera general una reducción en las emisiones, indicando que esta tendencia varía en función del país (Portela et al., 2023)

3. Objetivos

El presente Trabajo de Fin de Grado (TFG) tiene como objetivos:

- Estudiar el sector energético y determinar las fuentes de energía que están más explotadas actualmente y también en el futuro.
- Comparar la presencia y el crecimiento de las energías renovables entre diferentes regiones.
- Examinar y predecir la generación de energías renovables de manera global y por regiones.
- Conocer las variables que influyen en la generación de energías renovables.

4. Metodología

Para llevar a cabo el estudio del sector energético, se ha seleccionado la base de datos “Data on Sustainable Energy (2000-2020)” de la fuente Kaggle. Esta base de datos cuenta con más de 3600 observaciones y 21 variables que recopilan indicadores de energía sostenible a escala global desde el año 2000 hasta el año 2020.

En primer lugar, se ha realizado un tratamiento de la base de datos donde se ha llevado a cabo una limpieza de la información contenida junto con una descripción y un análisis descriptivo de las variables para extraer las características más representativas de los datos cuantitativos. Se ha hecho uso de medidas de tendencia central, dispersión, asimetría, percentiles y correlación de las variables. Además, se ha identificado la composición de variables que más afecta a la generación de energía renovable a través de un Análisis de Componentes Principales (PCA, por sus siglas en inglés), que sirve para disminuir la dimensionalidad y la probabilidad de incurrir en errores de multicolinealidad. Por último, se ha desarrollado un análisis de las redes neuronales para predecir la generación de energías renovables en el futuro.

2. BASE DE DATOS

1. Tratamiento y limpieza de datos

Como ya se ha mencionado anteriormente, la base de datos que se ha seleccionado es "Data on Sustainable Energy (2000-2020)" que cuenta con más de 3600 observaciones y 21 variables. Recoge información de indicadores energéticos y otros factores desde el año 2000 hasta el año 2020. Antes de comenzar con el análisis de la base de datos, se ha realizado una limpieza para poder identificar la información contenida que pueda ser incorrecta o incompleta. Esta limpieza nos dará la seguridad de que el análisis del sector energético será lo más preciso, coherente y válido posible.

En primer lugar, se ha modificado el nombre de las variables del *dataset* original y se ha observado el número de valores "*Not Available*" (NA en adelante) que existe en cada uno de los campos, como se representa la Tabla 1.

Tabla 1. Valores NA (*Not Available)* de cada una de las variables del conjunto original.

Variable	Valores NA
País (Country)	0
Año (Year)	105
Acceso Electricidad (A_Electricty_PercP)	115
Acceso Fuentes Renovables (A_CleanFuelsCooking_PercP)	274
Capacidad de Generar Energía Renovable (Ren_Gen_Cap_PerCap)	1015
Flujos Financieros (Fin_Flows_USD)	2120
Cuota de Energía Renovable (RenShare_TotEnergyConsump)	294
Electricidad de Combustibles Fósiles (Electricity_FossilFuels_TWh)	126
Electricidad Nuclear (Electricity_Nuclear_TWh)	231
Electricidad Renovable (Electricity_Ren_TWh)	126
Porcentaje Electricidad Baja Emisión (Electricity_LowCarbon_Perc)	147
Consumo Energía Primaria Per Cápita (Energy_Consump_KWhPC)	105
Nivel de Intensidad Energética (Energy_IntLevel_PrimE)	307
Emisiones CO2 Per Cápita (CO2Emissions_TonsPerCap)	528
Proporción Energía Renovable de Primaria (Ren_PercEquiv_PrimE)	2179
Crecimiento anual del PIB (GDPGrowth_AnPerc)	422
PIB Per Cápita (GDP_PerCap)	387
Densidad de Población (Pop_Density)	106
Superficie (Land_Area_Km2)	106
Latitud (Latitude)	106
Longitud (Longitude)	

Fuente: Elaboración propia.

A pesar de que la mayoría de variables no superan los 300 valores NA, existen algunas donde este número es muy elevado y puede afectar la calidad de los resultados del análisis. Para evitar esto, se ha llevado a cabo un análisis comparativo entre dos conjuntos de datos diferentes. A lo largo de todo el proyecto, se hará una comparación entre ambos conjuntos de datos y se referirá al primer conjunto como "conjunto original" donde se incluyen todas las variables y al segundo como "conjunto seleccionado" que está compuesto por 18 variables al haberse excluido las tres variables con más valores NA: Proporción Energía Renovable de Primaria (Ren_PercEquiv_PrimE), Flujos Financieros (Fin_Flows_USD) y Capacidad de Generar Energía Renovable (Ren_Gen_Cap_PerCap). A la hora de omitir los valores NA y eliminar todas aquellas filas con algún valor faltante, se obtiene que el conjunto de datos original presenta una reducción más significativa en el número de observaciones, manteniendo un total de 324 observaciones como consecuencia de contar con más valores NA. En el otro conjunto de datos que contiene menos variables y menos valores NA, existen un total de 2768 observaciones.

2. Descripción de las variables

Las variables que forman parte de esta base de datos cubren aspectos como la generación de energía a partir de fuentes renovables y no renovables, indicadores económicos clave y otros factores geográficos para comprender el sector energético y su evolución hacia una alternativa más sostenible. La base de datos cuenta con 21 variables donde se encuentran las siguientes:

- País (country): variable de tipo carácter (chr) que contiene el nombre del país.
- Año (year): variable de tipo entero (int) que indica el año.
- Acceso Electricidad (A_Electricty_PercP): variable de tipo numérico (num) que refleja el porcentaje de la población que tiene acceso a la electricidad.
- Acceso Fuentes Renovables (A_CleanFuelsCooking_PercP): variable de tipo numérico (num) que representa el porcentaje de la población que utiliza combustibles de fuentes renovables únicamente para cocinar.

- Capacidad de Generar Energía Renovable (Ren_Gen_Cap_PerCap): variable de tipo numérico (num) que indica la capacidad de generar energía renovable por persona medido en vatios per capita (W/persona).
- Flujos Financieros (Fin_Flows_USD): variable de tipo numérico (num) que muestra el total de flujos financieros destinados a países en desarrollo para proyectos de energía renovable, medido en dólares.
- Cuota de Energía Renovable (RenShare_TotEnergyConsump): variable de tipo numérico (num) que indica el porcentaje de energías renovables en el consumo total de energía.
- Electricidad de Combustibles Fósiles (Electricity_FossilFuels_TWh): variable de tipo numérico (num) que muestra la cantidad de electricidad generada de combustibles fósiles, donde se incluyen carbón, petróleo y gas, medido en teravatios-hora.
- Electricidad Nuclear (Electricity_Nuclear_TWh): variable de tipo numérico (num) que recoge información sobre la cantidad de electricidad generada a partir de energía nuclear medido en teravatios-hora.
- Electricidad Renovable (Electricity_Ren_TWh): variable de tipo numérico (num) que representa la cantidad de electricidad generada a partir de fuentes renovables, donde se incluyen hidráulica, solar, eólica, geotérmica, hidroeléctrica y oceánica, medido en teravatios-hora.
- Porcentaje Electricidad Baja Emisión (Electricity_LowCarbon_Perc): variable de tipo numérico (num) que representa el porcentaje de electricidad proveniente de fuentes de energía bajas en emisión de carbono, donde encontramos las fuentes de energía nuclear y renovables.
- Consumo Energía Primaria Per Cápita (Energy_Consump_KWhPC): variable de tipo numérico (num) que indica el consumo total de energía primaria por persona, medido en kilovatios-hora.
- Nivel de Intensidad Energética (Energy_IntLevel_PrimE): variable de tipo numérico (num) que muestra el uso de energía por unidad del Producto Interior Bruto (PIB en adelante) medido en paridad de poder adquisitivo (USD 2011).

- Emisiones CO2 Per Cápita (CO2Emissions_TonsPerCap): variable de tipo numérico (num) que representa la cantidad de emisiones de carbono por persona medido en toneladas métricas.
- Proporción Energía Renovable de Primaria (Ren_PercEquiv_PrimE): variable de tipo numérico (num) que recopila información sobre la proporción de energía producida a partir de fuentes renovables en relación con la cantidad total de la energía primaria utilizada.
- Crecimiento anual del PIB (GDPGrowth_AnPerc): variable de tipo numérico (num) que muestra la tasa de crecimiento anual del PIB medido en porcentaje y basado en la moneda local.
- PIB Per Cápita (GDP_PerCap): variable de tipo numérico (num) que representa el PIB por persona.
- Densidad de Población (Pop_Density): variable de tipo entero (int) que refleja la densidad de población medido en personas por kilómetro cuadrado.
- Superficie (Land_Area_Km2): variable de tipo entero (int) que refleja la superficie terrestre total medido en kilómetros cuadrados.
- Latitud (Latitude): variable de tipo numérico (num) que indica la latitud del país en grados decimales.
- Longitud (Longitude): variable de tipo numérico (num) que representa la longitud del país en grados decimales.

3. Análisis descriptivo de las variables

La estadística descriptiva consiste en describir, examinar y resumir la información recopilada para obtener características relevantes de un conjunto de datos. El objetivo principal de llevar a cabo este análisis descriptivo es obtener una visión general explorando la base de datos y obtener información relevante sobre las variables. Este análisis se ha realizado para ambos conjuntos de datos excluyendo las variables Country y Year del estudio al proporcionar información ya conocida sin necesidad de llevar a cabo un estudio más exhaustivo.

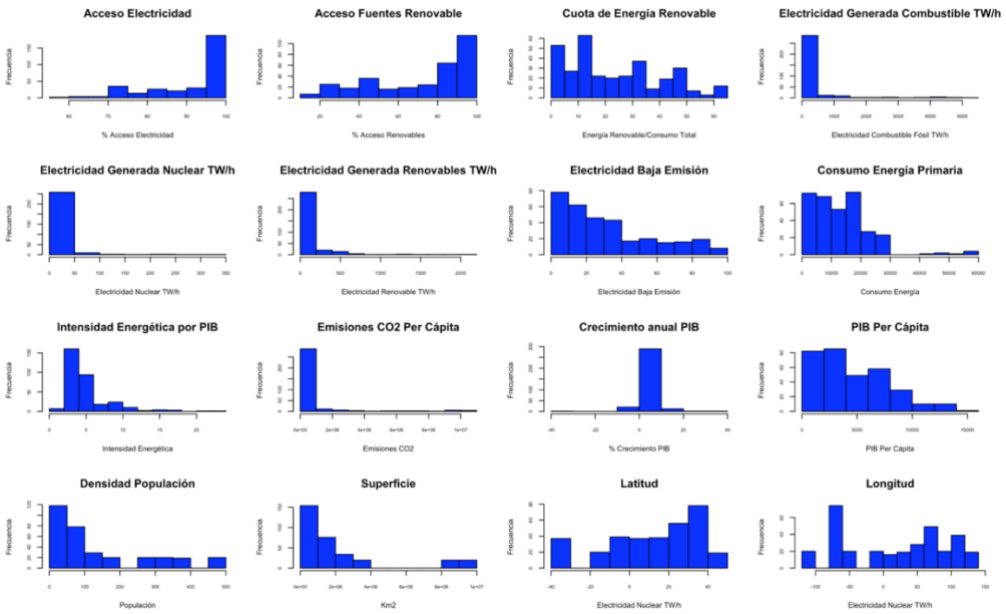
De manera general, se ha llevado a cabo el resumen de las distintas variables para conocer medidas como la tendencia central, la posición relativa y los valores máximos y

mínimos que estas pueden tomar. La herramienta utilizada para hacer el análisis ha sido R Studio. La diferencia más significativa entre ambas síntesis, la del conjunto original y la del conjunto seleccionado de datos, es la disparidad entre los resultados que se obtienen. Esto se debe a la diferencia entre el número de observaciones que hay en cada uno de ellos. Dicho con otras palabras, se justifica por la variación en la cantidad de registros que forman cada conjunto, haciendo que las medidas estadísticas se vean influenciadas.

Sin embargo, a pesar de las diferencias ya mencionadas, se puede observar una consistencia y coherencia entre ellos en la distribución de los datos. La disposición relativa de los datos y la distribución se mantienen uniformes en ambos conjuntos de datos *(Anexo 2 y Anexo 3).* Cuando una variable tiene una distribución asimétrica en el primer conjunto, también la tendrá en el segundo. Por ejemplo, la variable "Electricity_LowCarbon_Perc" tiene una asimetría positiva en ambos conjuntos de datos.

Comparando los histogramas de ambos conjuntos en la Figura 1 y la Figura 2, se puede confirmar la consistencia de los datos en términos de distribución. La mayoría de las variables se distribuyen asimétricamente, lo que indica que las observaciones se concentran más en los extremos, bien sean los inferiores o superiores dependiendo del tipo de asimetría. Esto difiere de la distribución normal o gaussiana, caracterizada por una frecuencia menor en las colas o extremos del histograma y mayor en el centro o la media de los datos. Además, la moda, la mediana y la media tienen valores iguales. La variable Crecimiento Anual PIB (GDPGrowth_AnPerc) se distribuye como una normal, donde la media y la mediana son prácticamente iguales, y la representación gráfica en el histograma respalda esta similitud.

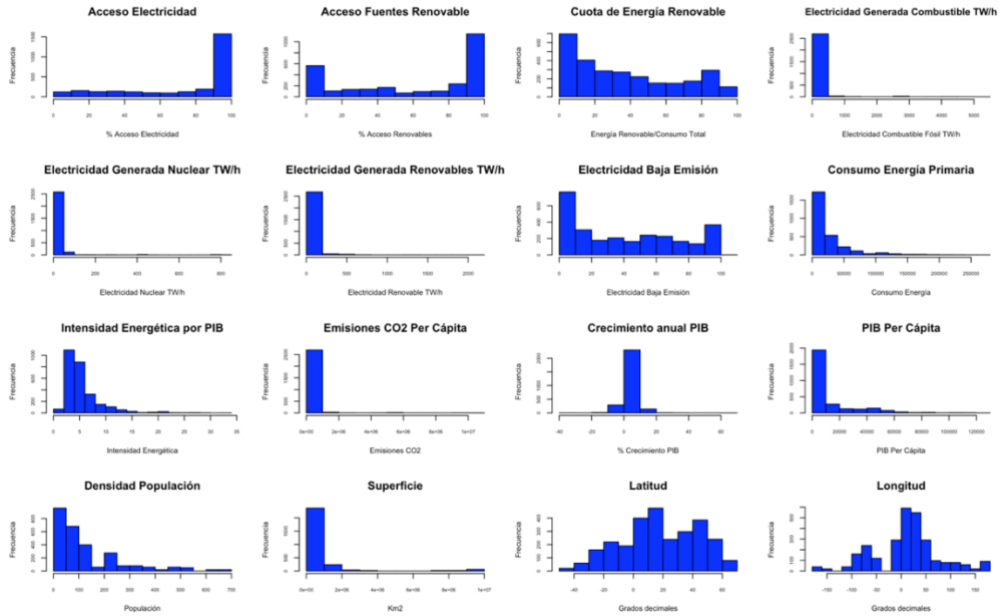
Figura 1. Histogramas de variables del conjunto original de datos



Fuente: Elaboración propia.

Las variables de Electricidad generada a partir de combustibles fósiles (Electricity_FossilFuels_TWh) y Electricidad generada a partir de renovables (Electricity_Ren_TWh) y Electricidad generada a partir de nuclear (Electricity_Nuclear_TWh) son un ejemplo claro de asimetría puesto que la media es significativamente mayor que la mediana indicando que la asimetría es positiva. Variables como Electricidad Baja Emisión (Electricity_Low_Carbon_Perc), PIB Per Cápita (GDP_Per_Cap) e Intensidad Energética por PIB (Energy IntLevel_PrimE) también presentan una distribución de estas características. La Figura 1 y la Figura 2 ilustran que existe una asimetría pronunciada hacia la derecha en esas variables. Al contrario, otras variables como Acceso a Electricidad (A_Electricity_PercP) y Acceso Fuentes Renovables (A_CleanFuelsCooking_PercP) presentan una distribución asimétrica negativa donde la mayoría de los valores toman valores altos, haciendo que la media sea menor que la mediana.

Figura 2. Histogramas de variables del conjunto seleccionado de datos



Fuente: Elaboración propia.

Con el objetivo de complementar la información obtenida a través de los histogramas, se han representado diagramas de caja o boxplots que muestran gráficamente información sobre la dispersión y los valores atípicos, más conocidos como *outliers*, y pueden definirse como un subconjunto de observaciones cuyos valores se consideran inconsistentes o inusuales con el resto de las observaciones. Debe existir un conjunto de observaciones que se considere regular o típico donde poder identificar aquellos valores que se desvían de la norma (Sim, Gan, & Chang, 2005).

La mediana representa el valor central de todo el conjunto de datos y lo divide en dos partes iguales en términos de cantidad de observaciones. Las variables Acceso Electricidad, Acceso Fuentes Renovables y Longitud presentan una mediana cercana a la parte superior de la caja, lo que indica que la mayoría de los datos se encuentran en el extremo superior siendo una distribución asimétrica hacia la izquierda, confirmando así la información extraída de los histogramas. Del mismo modo, la variable Crecimiento Anual PIB posee una distribución normal y el resto de las variables una distribución asimétrica hacia la izquierda *(Anexo 4 y Anexo 5)*.

Algunas dimensiones presentan valores atípicos u outliers que se encuentran por encima de los límites representados por lo bigotes del diagrama de caja. Al encontrarse por encima de los valores máximos o mínimos, se consideran desviados del resto de datos. Es el caso de las variables Electricidad generada a partir de combustibles fósiles (Electricity_FossilFuels_TWh) y Electricidad generada a partir de renovables (Electricity_Ren_TWh) y Electricidad generada a partir de nuclear (Electricity_Nuclear_TWh) donde presentan valores atípicos en el extremo superior. Estos *outliers* pueden ser la causa por la que la distribución de estas dimensiones es altamente positiva. Otros casos similares a estos son las variables Electricidad Baja Emisión (Electricity_Low_Carbon_Perc), PIB Per Cápita (GDP_Per_Cap) e Intensidad Energética por PIB (Energy IntLevel_PrimE). A pesar de que en muchos casos se decide eliminar los *outliers*, esta omisión no representaría la realidad del sector energético. Por eso mismo, los *outliers* se mantendrán y se tendrán en cuenta para el análisis posterior.

Además, se ha analizado la relación que existe entre las variables. La correlación indica la dirección y la fuerza de asociación o relación lineal entre las variables. Si la correlación es positiva, las variables se desplazan en la misma dirección. Es decir, si una aumenta, la otra también tiende a aumentar y viceversa. Por el contrario, si la correlación es negativa, las variables se dirigen a direcciones contrarias.

Existen algunas diferencias en la correlación entre los dos conjuntos de datos. Las correlaciones entre las variables del conjunto de datos originales son más fuertes y negativas que las del otro conjunto. Esta diferencia afectará al Análisis de las Componentes Principales ya que cuanta más correlación exista entre las variables, más información o varianza se explicará *(Anexo 6 y Anexo 7)*.

3. ANÁLISIS DE COMPONENTES PRINCIPALES

1. Reducción de dimensionalidad: motivos y finalidad

En conjuntos de datos con muchas variables, la cantidad de dimensiones puede dificultar el análisis, aumentar las necesidades computacionales y complicar la visualización de los datos. Además, muchas de esas variables pueden estar altamente correlacionadas entre ellas o aportar información redundante al análisis. Reduciendo este tipo de campos no solo se preserva la información relevante y se mantiene la varianza, sino que se consigue superar la maldición de la dimensionalidad. Existen múltiples métodos de reducción de dimensionalidad como mapas de difusión, Análisis Factorial método de Incrustación Localmente Lineal (LLE, por sus siglas en inglés) y Análisis de Componentes Principales (PCA, por sus siglas en inglés) además de sus variantes como el Análisis de Componentes Simples, entre otros. La base de estos métodos es la reducción de la estructura inicial con n dimensiones a otra estructura más pequeña a partir de esas variables originales con m dimensiones, siendo $m<n$ (Rousson & Gasser, 2004).

El objetivo principal de realizar PCA es la reducción de información redundante al haber variables que están muy correlacionadas entre ellas llegando a valores máximos, positivos y negativos. Por ejemplo, las variables "Electricidad de Combustibles Fósiles" (Electricity_FossilFuels_TWh) y "Emisiones CO2 Per Cápita" (CO2Emissions_TonsPerCap) para ambos *dataset* están altamente correlaciones ya que ambas indican el nivel de energía no renovable *(Anexo 6 y Anexo 7)*.

Para la elección del método de reducción de dimensionalidad, se ha realizado una comparativa entre tres técnicas comunes que son el Análisis de Componentes Principales, Análisis Factorial y t-SNE (por su nombre en inglés *t-distributed Stochastic Neighbor Embedding*), tal y como se muestra en la Tabla 2.

Tabla 2. Comparativa entre diferentes métodos de reducción de dimensionalidad

Método de	Definición	Ventajas	Limitaciones
Análisis de Componentes Principales (PCA)	Permite crear variables sintéticas a partir de combinaciones lineales de las originales.	- Mantiene las características de las variables - Rápida ejecución - La información o varianza que se retiene es alta	- En algunos casos los resultados pueden ser poco satisfactorios en modelos de regresión y clasificación - Dificultad de interpretación de las variables sintéticas - Los outliers pueden modificar el resultado -
Análisis factorial	Permite conocer las relaciones intrínsecas entre las diferentes variables a partir de correlaciones para identificar aquellas que explican la mayor parte de los datos	- Información simplificada - Fácil interpretación	- Pérdida de varianza no compartida entre las variables al centrarse en la común, perdiendo así información relevante
t-SNE	Permite visualizar muchas dimensiones de datos en mapas bidimensionales o tridimensionales	- Los outlier no afectan en el modelo	- Gran esfuerzo computacional - La reducción está limitada a 2 o 3 dimensiones

Fuente: Adaptado de Morris & Opazo (s.f.).

Tras la comparación, se ha escogido el Análisis de Componentes Principales como el método de reducción de dimensionalidad por sus ventajas, a destacar la fácil ejecución y poco esfuerzo computacional. Además, porque explica gran parte de la información del conjunto de datos inicial al mantener las relaciones entre las variables. Por otro lado, el análisis factorial es una técnica que identifica las correlaciones entre variables por lo que gran parte de la información puede ser perdida y, por el contrario, t-SNE es una técnica compleja con gran esfuerzo computacional y limitada en el número de dimensiones.

El Análisis de Componentes Principales es una técnica de aprendizaje no supervisado que permite reducir el número de variables, que están correlacionadas, y crear combinaciones lineales normalizadas de las variables originales independientes entre sí, generando un conjunto de componentes principales que mantiene la mayor información o varianza posible. Sin embargo, además de las mencionadas previamente,

esta técnica también presenta otras limitaciones que se deben de tener en cuenta antes de realizar el análisis. Los análisis clásicos de reducción de dimensionalidad, como PCA, se basan en priorizar distancias entre puntos más separados en el espacio, es decir, en distancias más grandes. Por lo tanto, no se capturan las relaciones de todos los puntos en su totalidad especialmente aquellos que se encuentran cerca en el espacio de datos, preservando la estructura global de estos frente a la estructura local (Park & Zhao, 2019). Estas limitaciones pueden tener un impacto negativo en la representación del nuevo espacio de datos y afectar al posterior análisis.

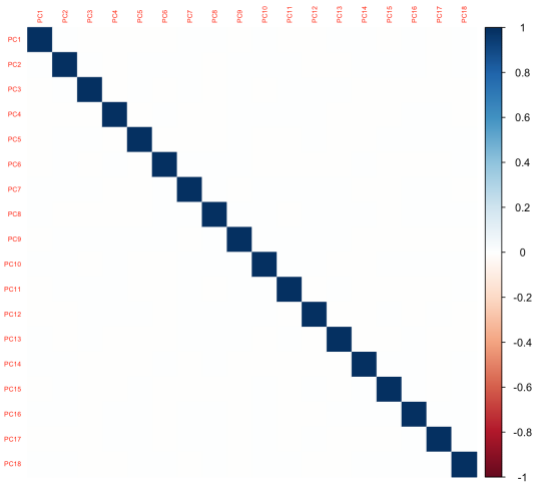
2. Proceso de aplicación del Análisis de Componentes Principales

El análisis se ha realizado sobre los diferentes conjuntos de datos con la herramienta R Studio. Antes de aplicar PCA**,** es necesario centrar y estandarizar las variables para que estén medidas en las mismas unidades y tengan la misma importancia en términos de variabilidad. De no hacerlo, las variables con escalas de datos más altas en comparación con el resto tendrían más peso, como puede ser el caso de Flujos Financieros (Fin_Flows_USD). Esta técnica solo es aplicable a variables numéricas por lo que, de nuevo, se excluyen las variables Year y Country del análisis. Además, también se ha excluido la variable Cuota de Energía Renovable (RenShare_TotEnergyConsump) ya que es la variable *target* u objetivo que se pretende predecir.

En primer lugar, se ha aplicado PCA en el conjunto original donde se han obtenido 18 componentes principales, uno para cada variable original excluyendo las tres variables previamente indicadas. En la matriz de los *scores (Anexo 8),* cada columna corresponde a un componente principal y las filas a las variables originales, que se obtienen a partir de la combinación lineal de los datos escalados y la matriz de vectores de carga (*loading vectors* en inglés). La matriz de los *scores* indica la proyección de las variables originales en las componentes principales. La matriz de *loading vectors (Anexo 9),* se basa en los autovectores de la matriz de covarianza de los datos originales, cuyos valores representan el peso o dirección de cada variable original en la formación de las componentes principales. A su vez, los autovalores asociados a estos autovectores muestran la varianza retenida por cada variable sintética. Estos valores están ordenados de tal manera que la primera componente principal es aquella cuyo autovalor o dirección recoge la mayor

varianza. Además, estas direcciones son ortogonales o perpendiculares entre sí lo que indica que la siguiente componente principal (PC2) recoge la siguiente mayor varianza que no está correlacionada con la primera componente principal (Gil Martínez, 2018). A continuación, se ha examinado si los datos obtenidos son los correctos a partir de tres características de PCA: las componentes principales son independientes entre sí, es decir, existe una correlación nula entre las variables sintéticas (Figura 3), los *loading vectors* son ortogonales obteniendo que el producto escalar de las dos primeras componentes principales es cero, y la suma de las varianzas de las nuevas variables es igual a la suma de las varianzas de las variables originales estandarizadas.

Figura 3. Matriz de correlación de las componentes principales del conjunto de datos originales



Fuente: Elaboración propia.

El número óptimo de componentes principales que se deben retener debe ser menor al número total de variables originales. Para establecer cuantas componentes principales se deben escoger, se analiza la proporción de varianza explicada de cada uno de estos componentes principales. Para las primeras variables sintéticas, la proporción de varianza explicada es mayor que para el resto de las componentes principales, representado en la Tabla 3, ya que los *loading vectors* están ordenados de tal manera que la primera componente principal recoja la mayor varianza.

Tabla 3. Proporción de varianza explicada de las componentes principales del conjunto de datos original

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0,2668	0,2427	0,1897	0,0646	0,0569	0,0469	0,0358	0,0306	0,0196
PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
0,0161	0,0107	0,0069	0,0048	0,0043	0,0021	0,0007	0,0005	0,0001

Fuente: Elaboración propia.

Sin embargo, la proporción de la varianza explicada no es muy elevada en las variables sintéticas, donde el mayor valor es de 0,2668. Esto indica que la primera componente principal explica el 26,68% de la información. Además, tal y como se ha indicado una de las limitaciones del análisis PCA es que es sensible a *outliers*. Cuanta más alta es la correlación, más información redundante existe entre las variables y la combinación de ellas explicaría gran parte de la varianza. En cuanto a la proporción de la varianza explicada, se obtiene la cantidad de información acumulada por las primeras componentes principales. Se obtiene a partir de los autovalores. Las tres primeras componentes principales explican casi el 70% de la información del conjunto, tal y como representa la Tabla 4.

Tabla 4. Proporción de varianza explicada acumulada de las componentes principales del conjunto de datos original

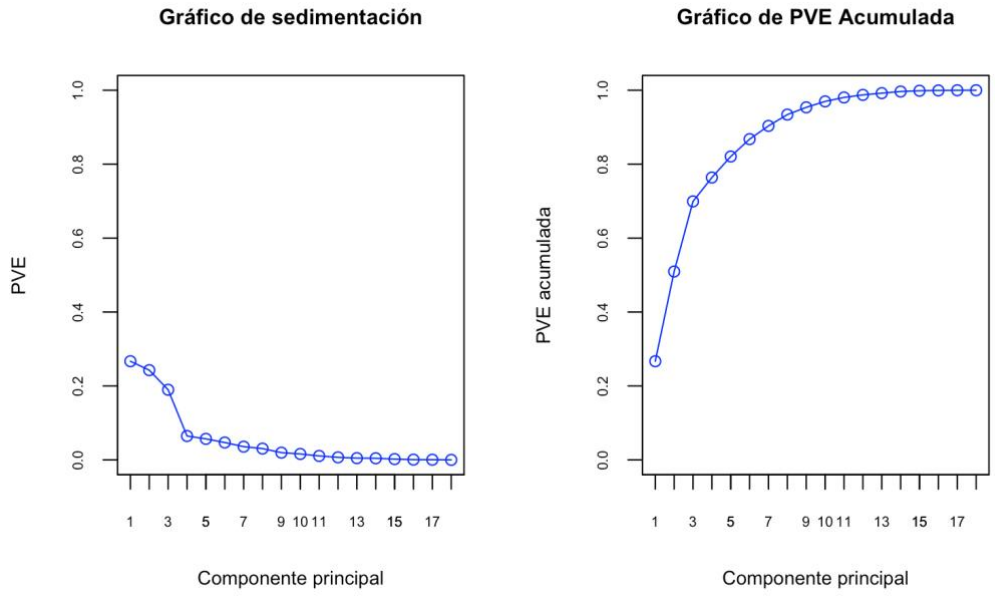
PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0,2668	0,5096	0,6993	0,7639	0,8208	0,8678	0,9036	0,9342	0,9538
PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
0,9698	0,9805	0,9874	0,9922	0,9965	0,9987	0,9994	0,9999	1,00

Fuente: Elaboración propia.

Para determinar el número óptimo de componentes principales, se ha escogido el método de la Regla del codo por su fácil interpretación y poco gasto computacional al no tener que añadir parámetros en comparación con otros métodos como Validación Cruzada. Esta regla consiste en encontrar el punto en el que la proporción de varianza explicada de las dimensiones disminuye, creando así una forma de "codo", de donde proviene el nombre. En el gráfico de sedimentación de la Figura 4 se observa que hay una

disminución significativa en la pendiente de la varianza explicada en la cuarta componente principal. Siguiendo este método, se debería escoger las tres primeras componentes principales. Esto recogería casi el 70% de la información explicada. Además, en el gráfico de Proporción de Varianza Explicada (PVE) Acumulada de la Figura 4 también representa como el crecimiento de la proporción de varianza explicada se ralentiza y se vuelve más estable a partir de la tercera componente principal.

Figura 4. Sedimentación y proporción de varianza explicada acumulada de las componentes principales del conjunto de datos original.



Fuente: Elaboración propia.

De las 18 variables iniciales, se escogen 3 componentes principales ya que estas dimensiones recogen la gran mayoría de las características de los datos. Para conocer cómo se componen las componentes principales, se observan los *loading vectors* de la Tabla 5, que indica el peso de las variables en cada una de las dimensiones. Los *loading vectors* negativos indican que hay una relación inversa entre la variable y la componente principal, mientras que los valores positivos reflejan una correlación directa.

Tabla 5. Loading vectors para las tres primeras componentes principales del conjunto de datos original.

Variable	PC 1	PC 2	PC 3
Acceso Electricidad (A_Electricty_PercP)	-0,12	-0,34	0,03
Acceso Fuentes Renovables (A_CleanFuelsCooking_PercP)	-0,01	-0,43	0,08
Capacidad de Generar Energía Renovable (Ren_Gen_Cap_PerCap)	-0,04	0,07	-0,04
Flujos Financieros (Fin_Flows_USD)	0,12	0,30	-0,34
Electricidad de Combustibles Fósiles (Electricity_FossilFuels_TWh)	-0,43	0,07	-0,09
Electricidad Nuclear (Electricity_Nuclear_TWh)	-0,40	0,04	-0,14
Electricidad Renovable (Electricity_Ren_TWh)	-0,40	0,07	-0,23
Porcentaje Electricidad Baja Emisión (Electricity_LowCarbon_Perc)	0,12	-0,06	-0,48
Consumo Energía Primaria Per Cápita (Energy_Consump_KWhPC)	-0,20	-0,29	0,19
Nivel de Intensidad Energética (Energy_IntLevel_PrimE)	-0,21	-0,02	0,29
Emisiones CO2 Per Cápita (CO2Emissions_TonsPerCap)	-0,43	0,07	-0,10
Proporción Energía Renovable de Primaria (Ren_PercEquiv_PrimE)	0,10	-0,06	-0,48
Crecimiento anual del PIB (GDPGrowth_AnPerc)	-0,12	0,13	0,06
PIB Per Cápita (GDP_PerCap)	-0,08	-0,37	-0,12
Densidad de Población (Pop_Density)	-0,00	0,43	-0,01
Superficie (Land_Area_Km2)	-0,31	-0,05	-0,27
Latitud (Latitude)	-0,14	0,17	0,25
Longitud (Longitude)	-0,16	0,34	0,19

Fuente: Elaboración propia.

En cuanto a la primera componente principal, los *loading vectors* con mayor valor absoluto son negativos y se encuentran las variables “Electricidad de Combustibles Fósiles”, “Electricidad Renovable” y “Emisiones de CO2 Per Cápita”, lo cual puede indicar una dimensión que captura el desarrollo energético al integrar electricidad renovable y no renovable. A pesar de que la interpretación es algo contradictoria al tener las variables la misma dirección de los autovectores, se podría decir que países con un alto valor de PC 1 representan mayor dependencia de energías no renovables.

Para la segunda componente principal, destaca “Densidad de Población” con un vector de carga mayor a 0,40, y valores negativos en las variables “Acceso a Electricidad” y “Consumo de Energía Primaria Per Cápita”. Esta componente principal indica que el acceso a la energía es menor cuando existe mayor densidad de población. Por lo que,

países con un PC 2 menor serán aquellos que tengan mayor acceso y consumo de electricidad.

Por último, los *loading vectors* más relevantes para PC 3 son "Latitud" y "Consumo de Energía Primaria" con un valor positivo y "Electricidad Renovable" con una relación negativa. Por lo que esta dimensión captura la relación entre factores geográficos y consumo de energía. Los países con valores altos de PC 3 pueden mostrar un mayor consumo de energía causado por dimensiones geográficas como latitud. Por el contrario, los países con un menor valor en esta componente principal indican que existe más electricidad renovable.

En relación con el conjunto de datos seleccionado, se ha aplicado el Análisis de Componentes Principales siguiendo la misma metodología que en el caso anterior. De igual manera, se han excluido las variables categóricas por condición del análisis y, dado que este conjunto de datos contiene un menor número de variables en comparación con el original, se obtendrán menos componentes principales, una para cada atributo.

Se ha obtenido la matriz de los *scores (Anexo 10)* y la matriz de *loading vectors (Anexo 11)* que describen las características de las componentes principales. Por un lado, las proyecciones de los datos originales o combinaciones lineales de los *loading vectors* y los datos originales y, por otro lado, esos vectores de carga que representan la dirección donde las componentes principales explican una mayor varianza.

En cuanto a la proporción de varianza explicada de las componentes principales, se observa en la Tabla 6, que la primera componente principal explica el 28,3% de los datos y la segunda el 19%. Sin embargo, esta explicación de la varianza disminuye significativamente en la tercera componente principal con un valor de 8,9%.

Tabla 6. Proporción de varianza explicada de las componentes principales del conjunto de datos seleccionado

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0,2827	0,1909	0,0893	0,0854	0,0642	0,0605	0,0541	0,0470
PC9	PC10	PC11	PC12	PC13	PC14	PC15	
0,0418	0,0333	0,0231	0,0122	0,0081	0,0071	0,0005	

Fuente: Elaboración propia.

Es por eso por lo que la proporción de varianza explicada acumulada recoge el 47,3% de la varianza total en las dos primeras componentes principales, tal y como se muestra en la Tabla 7. La varianza explicada de este conjunto de datos seleccionado es menor en comparación con el original, es decir, se necesitan más dimensiones para poder capturar la misma información. Esto ocurre porque las variables están más correlacionadas en los datos originales lo que hace que las direcciones en las que se mueven las variables sean similares, compartiendo así gran cantidad de información entre ellas.

Tabla 7. Proporción de varianza explicada acumulada de las componentes principales del conjunto de datos seleccionado

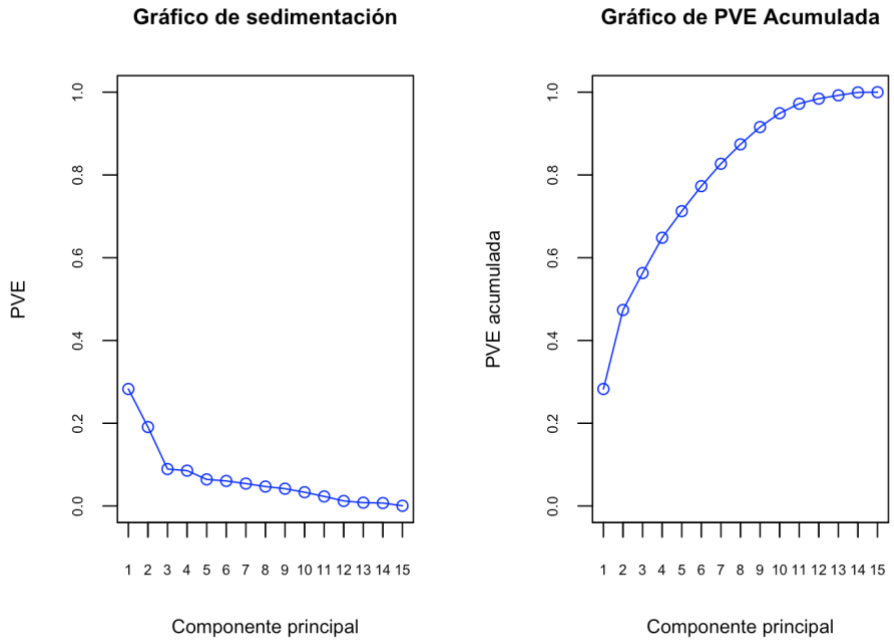
PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0,2827	0,4736	0,5623	0,6482	0,7125	0,7729	0,8271	0,8741
PC9	PC10	PC11	PC12	PC13	PC14	PC15	
0,9158	0,9491	0,9722	0,9843	0,9924	0,9995	1,00	

Fuente: Elaboración propia

En cuanto al gráfico de sedimentación se observa como hay una caída entre la segunda y la tercera componente principal después de la cual la curva empieza a aplanarse. En relación con el gráfico de PVE Acumulada, se observa como a partir del tercer punto el incremento por la varianza explicada es menor. Se ha decidido escoger las tres primeras componentes principales a pesar de que la regla del codo indique escoger únicamente los dos primeros. Esto es porque a pesar de que la varianza explicada por la tercera componente principal sea baja, aporta más información del conjunto de datos, alcanzando así el 56% de PVE Acumulada. Además, no se pretende obtener un modelo

muy simple sino un modelo completo para realizar posteriormente análisis de modelos supervisados.

Figura 5. Sedimentación y PVE Acumulada de las componentes principales del conjunto de datos seleccionado



Fuente: Elaboración propia

Del mismo modo que en el conjunto de datos original, se han analizado los *loading vectors* de las tres componentes principales para identificar de qué variables se componen, y comprender lo que indican cada una, tal y como se muestra en la Tabla 8.

Tabla 8. *Loading vectors* para las tres primeras componentes principales del conjunto de datos seleccionado

Variable	PC 1	PC 2	PC 3
Acceso Electricidad (A_Electricty_PercP)	0,25	-0,39	0,14
Acceso Fuentes Renovables (A_CleanFuelsCooking_PercP)	0,26	-0,42	0,07
Electricidad de Combustibles Fósiles (Electricity_FossilFuels_TWh)	0,40	0,28	0,03
Electricidad Nuclear (Electricity_Nuclear_TWh)	0,33	0,13	0,06
Electricidad Renovable (Electricity_Ren_TWh)	0,39	0,25	0,04
Porcentaje Electricidad Baja Emisión (Electricity_LowCarbon_Perc)	-0,03	0,08	-0,04
Consumo Energía Primaria Per Cápita (Energy_Consump_KWhPC)	0,25	-0,32	-0,35
Nivel de Intensidad Energética (Energy_IntLevel_PrimE)	-0,03	0,15	-0,59
Emisiones CO2 Per Cápita (CO2Emissions_TonsPerCap)	0,39	0,29	0,02
Crecimiento anual del PIB (GDPGrowth_AnPerc)	-0,04	0,12	-0,38
PIB Per Cápita (GDP_PerCap)	0,26	-0,34	-0,16
Densidad de Población (Pop_Density)	0,01	-0,09	0,09
Superficie (Land_Area_Km2)	0,33	0,26	0,03
Latitud (Latitude)	0,20	-0,29	-0,21
Longitud (Longitude)	-0,03	0,07	-0,52

Fuente: Elaboración propia

En relación con la primera componente principal, parece indicar el nivel de consumo de energía y de electricidad, tanto de fuentes fósiles como renovables junto con medidas de desarrollo económico al haber variables como PIB con loadings vectors altos.

La segunda componente principal se compone de la variable "Emisiones de CO2 Per Cápita", "Electricidad de Combustibles Fósiles" y "Electricidad Renovable" con *loading vectors* positivos, y de variables con vectores de carga negativos como "Acceso a Electricidad" y "PIB Per Cápita". Este componente principal indica la relación entre el nivel de electricidad con el desarrollo económico que permite acceder a fuentes de energía.

La última componente principal se compone principalmente de variables con loading vectors negativos "Nivel de Intensidad energética", "Consumo Energía Primaria Per Cápita" y "Longitud". La variable con un valor positivo remarcable es "Acceso a la

Electricidad". Esta relación puede indicar el nivel de uso energético con la disponibilidad de la electricidad.

Para ambos conjuntos de datos se seleccionan las tres componentes principales ya que capturan la mayoría de las características de las variables. Sin embargo, la proporción de varianza explicada es mayor para el conjunto de datos original que para el seleccionado con una diferencia del 13%. Esto indica que las tres componentes principales del conjunto de datos original explican más la información de la totalidad de las variables que lo que hace el *dataset* seleccionado. La mayoría de las dimensiones tienen un *loading vector* en valor absoluto elevado en alguna de las tres componentes principales por lo que es complicado identificar cuáles son las que más influyen en variable de Cuota de Energía Renovable. Para ello, se realizará un modelo de regresión múltiple a lo largo del presente Trabajo de Fin de Grado.

4. ANÁLISIS DE REDES NEURONALES

1. Motivos y finalidad

Existen diferentes modelos de aprendizaje automático para resolver problemas de regresión. Algunos de estos modelos predictivos son los árboles de regresión y los modelos de regresión. Dentro de las aplicaciones de las redes neuronales se encuentra la capacidad de predecir datos. Además, las redes neuronales ofrecen ventajas como la capacidad de captar relaciones no lineales en los datos y ser más flexibles. Sin embargo, el tiempo de aprendizaje es elevado y no son fáciles de interpretar lo que aprende la red, simplemente los resultados (Rivera, 2003; Basogain, s.f.). Entre los diferentes modelos de predicción existen los arboles de regresión que consisten en dividir los datos mediante un criterio de partición en grupos homogéneos hasta obtener uno más pequeño (Serna Pineda, 2009). Por otro lado, también existen los modelos de regresión que pretenden explicar las relaciones entre variables o estimar valores futuros que puede tomar una variable en función de otra, es decir, predecir. Son modelos fáciles de aplicar e interpretar, sin embargo, solo identifican relaciones lineales entre las observaciones (Moral, 2006).

La finalidad principal del uso de redes neuronales es conseguir uno de los objetivos principales del presente Trabajo de Fin de Grado, que es la predicción de los niveles de energía renovable en el futuro, de acuerdo con datos históricos. Se ha escogido predecir la variable "Cuota de Energía Renovable" que mide el porcentaje de energías verdes frente al total de energías. No solo se busca evaluar la generación de electricidad renovable, sino el uso real de la misma que como contrapartida evita que se utilicen energías provenientes de combustibles fósiles.

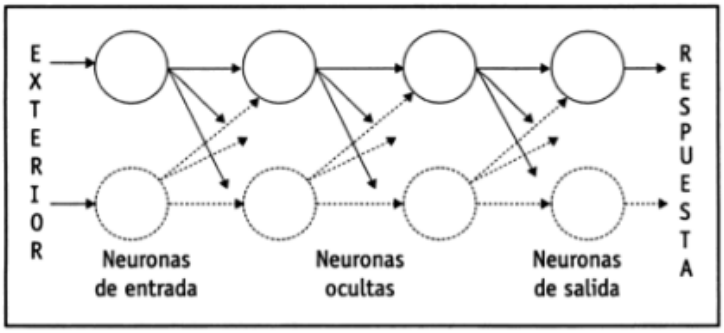
2. Conceptos básicos

"Una red neuronal artificial es un sistema de procesamiento de información que tiene ciertas características de desempeño en común con las redes neuronales biológicas". Se caracterizan por la arquitectura que conecta las neuronas, por el ajuste de los pesos de

las conexiones a través de logaritmos de entrenamiento y por la función de activación. (Fausett, 1994:3)

El nodo o neurona es la menor unidad dentro de la arquitectura de la red neuronal artificial, pudiendo ser de entrada, de salida u oculta dependiendo de sus conexiones. En cuanto a las neuronas de entrada, son las que reciben señales de datos provenientes del entorno; las de salida transmiten su información fuera del sistema; y las ocultas se encuentran entre las dos anteriores, recibiendo entrada de datos y emitiendo señales a las siguientes capas de neuronas (Flórez López y Fernández Fernández, 2008).

Figura 6. Arquitectura de una red neuronal



Fuente: Flórez López y Fernández Fernández (2008). Tipos de neuronas

Según IBM (s.f.), los modelos de *machine learning*, están formados por capas de neuronas que se conectan entre sí. Cada nodo es similar a una regresión lineal, que procesa información y transmite señales a otros nodos. A su vez, estos nodos están formados por los siguientes elementos principales:

- Entradas *(Inputs)*: es toda la información de entrada que recibe cada nodo del nodo anterior (salidas de otro nodo), o en caso de ser la primera capa, los datos de entrada.
- Pesos: a cada entrada se le asigna un peso que es la importancia o influencia de esa entrada en la salida del nodo. Estos pesos son ajustables y se suelen modificar en el entrenamiento de la red.
- *Bias*: es un valor adicional a los valores de entrada y que se incluye en la suma ponderada que hará la neurona.

- Función de suma: el nodo o neurona calcula una suma ponderada de las entradas, en función de los pesos e incluyendo el sesgo.
- Función de Activación: es una función matemática, normalmente no lineal, de la suma de ponderaciones, que va a determinar la salida.
- Salida *(Outputs):* es el resultado de la función de activación. Si pasa el umbral definido, el nodo se activa y será la entrada para los nodos posteriores o la salida final si el nodo está en la última capa.

Una de las características esenciales de las redes neuronales es el aprendizaje, que "puede definirse como el proceso por el cual una red neuronal crea, modifica o destruye sus conexiones en respuesta a una información de entrada" (Flórez López y Fernández Fernández, 2008:31). En el momento de formación de la red neuronal, los pesos se establecen aleatoriamente, por lo que es necesario entrenar la red y ajustar los pesos en función del objeto de estudio minimizando al máximo funciones de error como el Error cuadrático medio, más conocido en inglés como *Mean Squared Error* (MSE), o la Raíz del error cuadrático medio (*Root Mean Squared Error*, RMSE) (Flórez López y Fernández Fernández, 2008). Dentro de los tipos de aprendizaje automático, existen dos grupos principales que son el aprendizaje supervisado y no supervisado. Por un lado, el supervisado es aquel que tiene información etiquetada como datos de entrada y de salida, por lo que el objetivo es clasificar o predecir datos. Por otra parte, el aprendizaje no supervisado tiene como objetivo identificar patrones ocultos de conjuntos de datos sin etiquetar (Delua, 2021).

Existen diferentes tipos de redes neuronales según la finalidad y la estructura de cada una de ellas. Se encuentran las redes monocapa que no cuentan con capas ocultas, sino que solo tienen una capa de entrada y otra de salida, como las redes de perceptrón simple. De estas se derivan las redes multicapa, o de perceptrón múltiple, que cuentan con capas ocultas donde los nodos procesan la información y se conectan con las de la capa siguiente. Además, existen las redes convolucionales donde al contrario que en los casos anteriores, solo algunas neuronas se relacionan con las de la siguiente, denominándose nodos especializados. Y, por último, las redes concurrentes que se

caracterizan por tener neuronas organizadas entre sí sin tener una organización en capas (UNIR, 2021).

3. Proceso de aplicación de redes MLP

Dentro de los diferentes tipos de redes neuronales que existen, se ha entrenado la red de perceptrón multicapa (MLP), un modelo predictivo de una o más variables objetivo dependientes, que aprende de los valores proporcionados por las variables predictoras (IBM, 2023). Estas redes se caracterizan por tener una capa de entrada, una oculta y otra salida, tal y como se ha indicado. Además, todas las neuronas de la capa están conectadas entre sí, teniendo así una conectividad completa, y la información de una capa a otra se transmite hacia delante de una manera unidireccional, por lo que se denominan redes *feedforward.* El aprendizaje de este tipo de redes neuronales *feedforward* es supervisado, donde el algoritmo más común para el entrenamiento es el de retropropagación o *backpropagation*. Es un algoritmo que permite propagar los errores identificados en las capas de salida a las capas ocultas, con el objetivo de modificar los pesos y el *bias* de las conexiones para reducir la función de error (Camejo et al, 2020; Fausett, 1994).

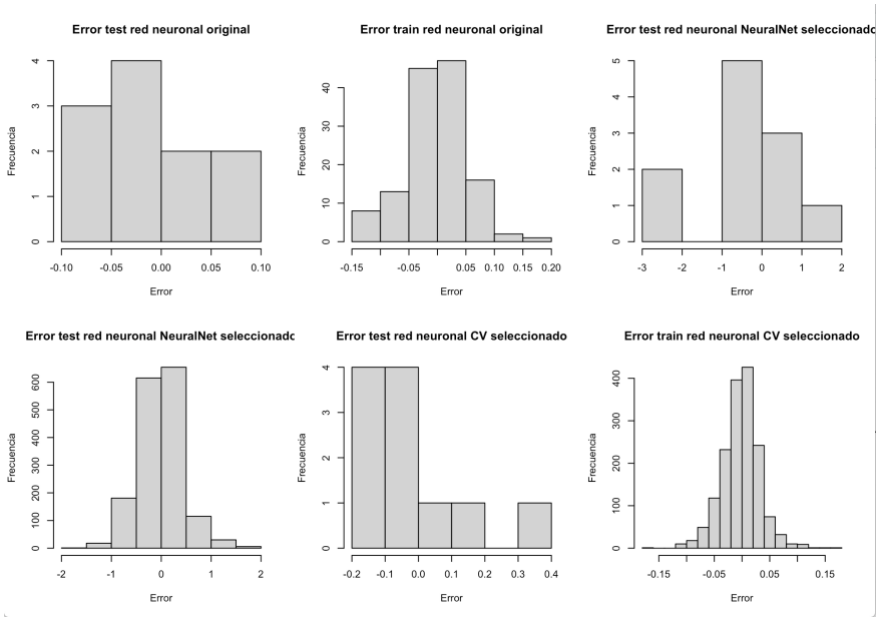
Tal y como se ha comentado previamente, el objetivo final de emplear redes neuronales es predecir la evolución de la energía renovable a escala mundial, por lo que se ha escogido la red neuronal MLP donde la variable *target* es Cuota de Energía Renovable (RenShare_TotEnergyConsump) y las variables predictoras son los tres componentes principales que se han obtenido al realizar PCA para ambos conjuntos de datos. Del mismo modo que se ha ido realizando a lo largo del Trabajo de Fin de Grado, se han llevado a cabo las ejecuciones para cada conjunto de datos, el original y el denominado seleccionado.

La herramienta empleada para realizar las predicciones a través de redes neuronales ha sido R Studio, con el paquete neuralnet que entrena la red utilizando el algoritmo de retropropagación, y permite realizar predicciones.

La aplicación para ambos conjuntos de datos es similar, sin embargo, difieren en los resultados. Para cada caso, el conjunto de datos que se ha utilizado para entrenar la red de perceptrón multicapa está formados por las variables Country, Year, PC 1, PC 2, PC 3 y la variable *target* Cuota de Energía Renovable. Antes del entrenamiento del modelo, se ha llevado a cabo un preprocesamiento de los datos. Por un lado, se ha estandarizado la variable objetivo para que todos los datos estén en la misma escala, dado que las componentes principales ya lo estaban al haber realizado PCA previamente. Además, se han filtrado aquellos países con información para los años entre 2005 y 2018, excluyendo al mismo tiempo los datos de otros años. De esta manera, el modelo es más consistente al incluir un mismo periodo temporal para todas las regiones. Se han añadido retardos a las variables predictoras por país para capturar la evolución temporal de las componentes principales en cada región y mejorar la predicción. Estos retardos reflejan el valor inmediatamente anterior de las observaciones. Dado que solo se tienen observaciones a partir del año 2005, y estas no tienen retardos, se eliminan para evitar valores NA.

En cuanto al entrenamiento de la red neuronal, se ha realizado una partición de los datos, en datos de *train* y *test*. Esta división es necesaria para estimar el error de predicción. Para mantener la temporalidad de los datos, se han incluido en el conjunto de entrenamiento todos menos los dos últimos años para todos los países, que forman parte del *dataset test*. El error de predicción se ha estimado tanto para el conjunto de *train* como de prueba para evaluar la precisión y el ajuste del modelo a los datos, explicado a continuación. La métrica para calcular el error es el error medio cuadrático (RMSE, por su nombre en inglés *Root Mean Squared Error*) que es una medida estadística utilizada para evaluar modelos. El motivo de escoger este tipo de error frente a otros como MSE o MAE (error absoluto medio) es porque RMSE se emplea cuando la distribución de los errores es de tipo normal o gaussiana, mientras que MSE cuando los errores siguen una distribución laplaciana (Hodson, 2022). De esta manera, se han evaluado las distribuciones de las redes neuronales donde todas ellas seguían una distribución gaussiana tal y como se puede observar en la Figura 7, escogiendo entonces el RMSE para evaluar el modelo. La frecuencia de los errores difiere ya que el número de observaciones no es la misma para los diferentes conjuntos de datos, ni para *test* y *train*.

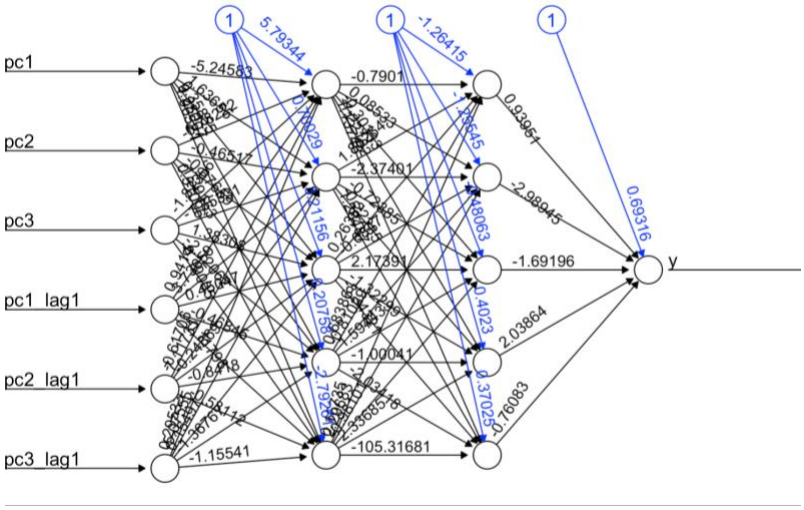
Figura 7. Histogramas de errores train y test



Fuente: Elaboración propia

En relación con la arquitectura, se han entrenado tres diferentes redes neuronales. Estos modelos no lineales cuentan con seis variables de entrada que son las componentes principales y los retardos asociados a cada una de ellas. En el caso del conjunto de datos original se ha entrenado una sola red neuronal dado que los resultados obtenidos son buenos, comentados a continuación. La red cuenta con dos capas ocultas de neuronas, cada una de cinco neuronas, e interconectadas entre sí con un peso asociado a cada conexión. En cuanto a la capa de salida, solo existe una neurona que predice el valor objetivo, tal y como se muestra en la Figura 8. Además, se han fijado únicamente dos hiperparámetros de *threshold* que indican por debajo de qué error el modelo debe dejar de entrenar, con un valor de 0,01, y *stepmax* que marca el número de pasos máximos de entrenamiento 1e+08. Ambos parámetros sirven para evitar que el modelo tarde más de lo necesario en entrenar.

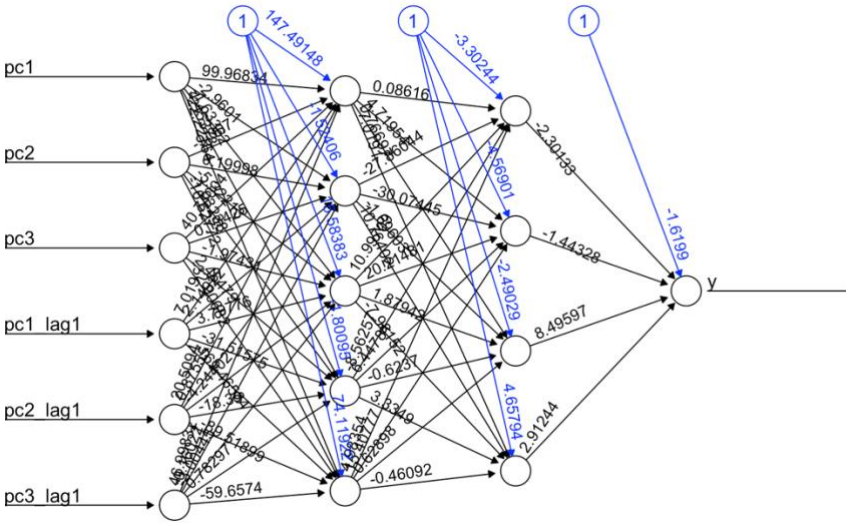
Figura 8. Arquitectura red neuronal conjunto original



Fuente: Elaboración propia

En el caso del conjunto de datos seleccionado, se han ejecutado dos redes neuronales diferentes. La primera de ellas se ha realizado con el paquete neuralnet y tiene dos capas ocultas, de cinco y cuatro neuronas respectivamente. El ajuste de los parámetros e hiperparámetros se ha realizado de manera manualmente en función de los resultados obtenidos y del coste computacional, sin embargo, el error de entrenamiento y de prueba es muy alto y, además, se incurre en *overfitting* lo que significa que el modelo se ajusta demasiado a los datos de entrada y no es capaz de generalizar. Dado que los resultados no han sido los óptimos, se ha realizado una validación cruzada de *grid search* o búsqueda de cuadrícula, una técnica utilizada para encontrar las mejores combinaciones de hiperparámetros para minimizar el error de entrenamiento. Realiza un entrenamiento de diferentes modelos con hiperparámetros asociados a cada uno, escogiendo aquel cuyo desempeño sea el mejor. Se ha empleado esta técnica a la función "nnet" incorporada en R Studio, que entrena redes neuronales con una única capa oculta, dado que "neuralnet" no tiene una función incorporada de validación cruzada de *grid search.* Los hiperparámetros que se han obtenido con la validación cruzada de *grid search* son *size*=5 que indicia el número de neuronas en la capa oculta y *decay*= 0,1 que es el hiperparámetro que controla la regularización, y esta a su vez, se emplea para controlar el sobreajuste y subajuste (Kolluri et al., 2020; Amat Rodrigo, 2018; Adnan et al, 2022).

Figura 9. Arquitectura red neuronal conjunto seleccionado



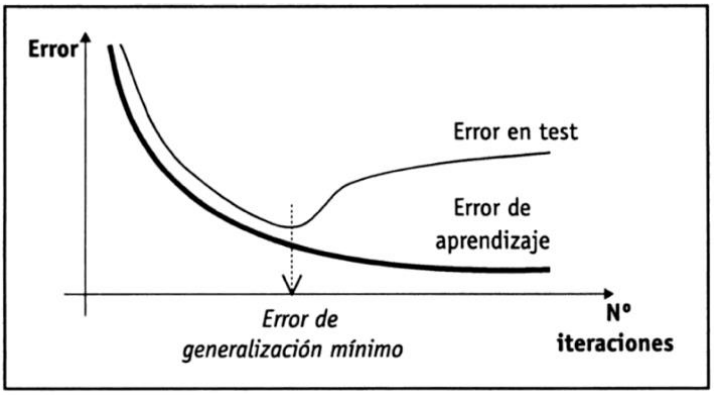
Fuente: Elaboración propia

5. RESULTADOS

En este capítulo se analizarán los resultados obtenidos de la aplicación de las redes neuronales. Este análisis incluirá una evaluación de precisión y de los resultados de cada uno de los conjuntos de datos, además de una comparativa entre ambos. Por otro lado, también se realizará un modelo de regresión logística para identificar las variables que más afectan a la objetivo, Cuota de energía renovable.

Para conocer la precisión de las predicciones, evaluar la red neuronal y saber si existe sobreajuste o subajuste, se han estimado los errores de *test* y de entrenamiento. Por un lado, el sobreajuste u *overfitting*, ocurre cuando un modelo aprende características irrelevantes de los datos en lugar de aprender la relación general entre la variable objetivo y las variables predictoras. Por otro lado, el subajuste o *underfitting*, ocurre cuando un modelo no es capaz de aprender las relaciones entre las variables (Bashir et al., 2020). A la hora de evaluar una red neuronal, es preferible que el modelo generalice correctamente en lugar de intentar reducir su error al máximo. En otras palabras, es mejor que el error de aprendizaje sea algo elevado antes de que incurra en *overfitting*. El sobreaprendizaje u *overfitting* ocurre cuando el error de *test* es mayor que el error de aprendizaje o el error de entrenamiento tal y como se observa en la Figura 10 (Flórez López y Fernández Fernández, 2008).

Figura 10. Ilustrativo del problema de sobreajuste



Fuente: Flórez López y Fernández Fernández (2008). Análisis del fenómeno de "sobreaprendizaje de la red"

En cuanto a la red neuronal entrenada para el conjunto original, se puede observar en la Tabla 9 cómo el error de *test* es ligeramente mayor que el error de entrenamiento, sin embargo, este sobreajuste es mayor cuanto mayor sea la diferencia entre ambos errores, que en este caso es de 0,13. Por lo tanto, se puede indicar que la red generaliza bien.

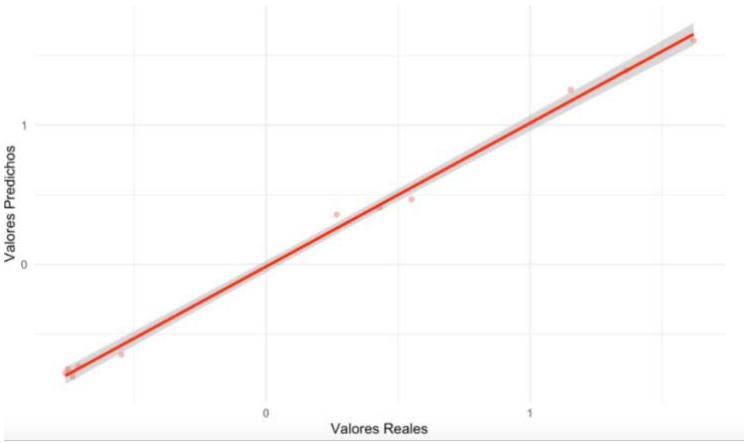
Tabla 9. Errores de la red neuronal del conjunto original

	Red conjunto original
Error de prueba o *test*	0,0702
Error de entrenamiento	0,0576

Fuente: Elaboración propia

La Figura 11 muestra la comparación entre los valores reales y los valores predichos. La relación es positiva, lo que indica que los valores predichos se ajustan a los reales. Además, las observaciones representadas por puntos están muy cerca de la línea indicando un buen ajuste del modelo e incluso se podría considerar que existe algo de sobreajuste.

Figura 11. Gráfico de dispersión de los valores predichos de la red neuronal del conjunto original



Fuente: Elaboración propia

En cuanto a las redes neuronales del conjunto seleccionado, presentan unos errores de prueba y entrenamiento muy dispares entre ellos, indicando en ambos casos *overfitting*. La red neuronal de dos capas ocultas tiene un error de prueba considerablemente mayor que el error de entrenamiento. La diferencia de estos errores es

excesivamente alta reflejando un *overfitting* en el entrenamiento del modelo. Del mismo modo y a pesar de haber empleado técnicas de optimización de hiperparámetros, la segunda red neuronal también incurre en *overfitting*. El entrenamiento generado por estos modelos no es correcto a diferencia del caso anterior.

Tabla 10. Errores de las redes neuronales del conjunto seleccionado

	Red 1 (2 capas)	**Red 2 (Validación Cruzada)**
Error de prueba o *test*	1,0770	0,1426
Error de entrenamiento	0,4409	0,0337

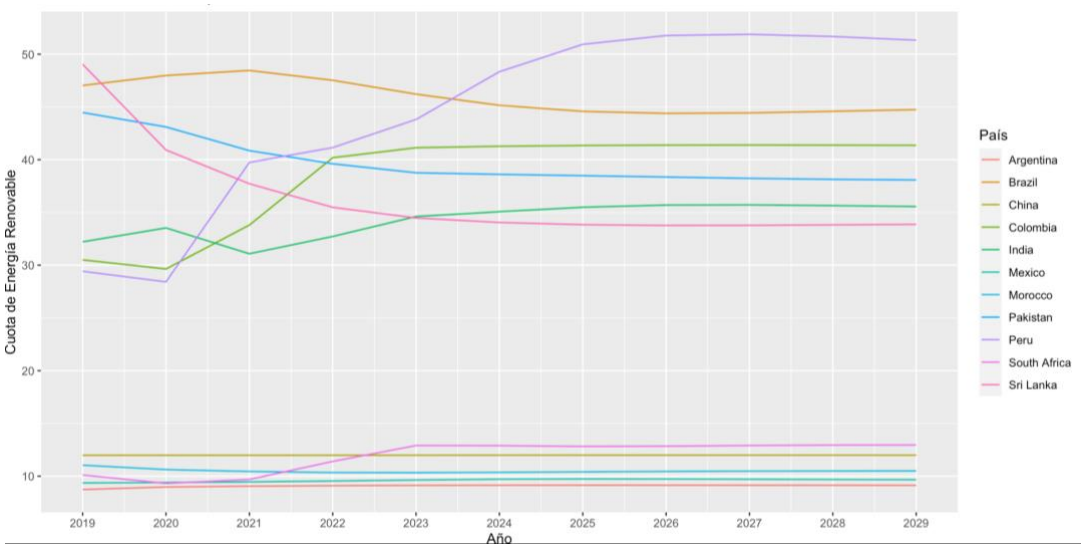
Fuente: Elaboración propia

Uno de los motivos por los que el conjunto original aprende mejor que el conjunto seleccionado y no incurre en sobreajuste es la información contenida en cada *dataset*. Al realizar el Análisis de Componentes Principales (PCA), la proporción de varianza explicada por el conjunto original es mayor que la del conjunto seleccionado, con una diferencia de 0,13. Esto indica que se recoge más información en el conjunto original, permitiendo a las redes neuronales captar mejor las relaciones entre las variables. De este modo, la red neuronal aprende mejor y sus predicciones son más precisas.

Dado que las predicciones de la red neuronal del conjunto original tienen un error más tolerable y no hay un sobreajuste excesivamente marcado, se han obtenido las proyecciones para los años futuros. Ya que tampoco se tienen los datos futuros de las variables predictoras, y son necesarias para conseguir la variable objetivo, se han estimado utilizando modelos de regresión lineal múltiple que pronostican en función de las relaciones con el resto de las variables regresoras (Hyndman, y Athanasopoulos, 2021). De esta manera, se obtienen valores de las variables predictoras (PC1, PC2 y PC3) y así, predecir la "Cuota de Energía Renovable" en años futuros desde el 2019 al 2029.

De esta manera, se puede observar en la Figura 12, la evolución de la variable objetivo "Cuota de Energía Renovable" para cada región. Además, se ha comparado con los valores de estas regiones en los años previos a las predicciones (*Anexo 12*) para confirmar que los niveles de cuota de energía renovable obtenidos son coherentes con los existentes.

Figura 12. Proyecciones de la Cuota de Energía Renovable



Fuente: Elaboración propia

La variable objetivo mide el porcentaje de energía renovable que existe en cada país, representado en el eje de ordenadas. A pesar de que los países predichos no representan el total de estos, se pueden observar algunas tendencias en la evolución de la cuota de energía renovable. A excepción de países como Sri Lanka y Paquistán que sufren una caída de la cuota de energía sostenible y de otros que mantienen unos valores constantes a lo largo del tiempo, se estima que haya un crecimiento de la variable en algunos países en los próximos diez años. Estos son Perú, con un valor esperado superior al 50% en 5 años y, Colombia e India con una cuota cercana al 40%.

Por otro lado, también se han identificado las variables que más influyen en la Cuota de Energía Renovable, mediante un modelo de regresión múltiple, considerando todas las variables del conjunto de datos sin reducir la dimensionalidad. Las variables más significativas son las siguientes:

- Acceso Electricidad: tiene una relación negativa, por lo que un aumento en el acceso de electricidad produce una disminución en la cuota de energía renovable.
- Acceso a Fuentes Renovables: también se relaciona contrariamente con la variable dependiente, que es aquella que se pretende predecir, con un coeficiente alto. Esto parece contradictorio por lo que cabe recordar que solo se consideran las fuentes

renovables empleadas para cocinar. Existiendo otras fuentes de energía renovable, puede que las destinadas a la energía del hogar como la biomasa no tengan tanta relevancia en el total de la cuota de energía limpia.

- Nivel de Intensidad Energética: tiene un valor cercano al -1, indicando una fuerte relación negativa entre las variables. A medida que aumenta la intensidad energética, la cuota de energía renovable disminuye significativamente.

6. CONCLUSIONES

El presente Trabajo de Fin de Grado (TFG) tiene como objetivo principal estudiar la evolución del sector energético a partir de variables que engloban factores económicos, energéticos y demográficos en diferentes regiones.

A lo largo del trabajo se ha obtenido un análisis exploratorio de los datos iniciales que han permitido identificar las diferentes fuentes de energía y su presencia a escala mundial. Además, se ha logrado identificar las variables más relevantes y su relación con la variable objetivo que refleja la proporción de energía renovable, indicando así su influencia en la generación y el uso de energías renovables. Por otro lado, también se ha proyectado la proporción estimada de energía renovable para los próximos cinco años. Se ha examinado la evolución de las energías renovables como la proporción de energía renovable sobre el total de energía para no solo identificar su generación, sino también para conocer su uso real en comparación con otras. Los hallazgos obtenidos a partir del trabajo, considerando tanto los resultados como las áreas a mejorar, pueden considerarse como punto de partida para futuros análisis del campo energético con enfoque en la evolución de renovables.

Los resultados obtenidos a partir de los modelos predictivos indican un crecimiento estable de la cuota de energía renovable en los próximos 5 años. Sin embargo, estos resultados solo se han obtenido para un número limitado de países como Sri Lanka o Perú que no representan la totalidad. Por lo tanto, no se pueden generalizar estos resultados a nivel global, sino solo para las regiones observadas.

Algunas de las debilidades observadas a lo largo del Trabajo de Fin de Grado son la limitación en la cantidad de datos y el sobreajuste observado en el entrenamiento de las redes neuronales. Como consecuencia, los resultados obtenidos no han sido los óptimos. Con la finalidad de mejorar las predicciones y poder profundizar más en el análisis, se debería acceder a un mayor número de datos para que el modelo de redes neuronales sea capaz de captar todas las relaciones entre las variables sin aprender de ruido e incurrir en sobreajuste. Por otro lado, también se podrían emplear otros modelos predictivos más

simples como regresión logística o árboles de decisión para cumplir con los objetivos. De esta manera, se obtendrían proyecciones más realistas y sobre más regiones para poder tener una imagen representativa del total de los países.

En resumen, este Trabajo de Fin de Grado (TFG) proporciona una base inicial para futuros análisis del sector energético. Los resultados obtenidos indican de manera general una evolución estable para los países estudiados. Además, también indican la necesidad de más datos y diferentes modelos predictivos para mejorar la precisión de las predicciones. De esta manera, se podrá profundizar sobre el análisis y obtener resultados más representativos.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Ana María Chacón Rueda, estudiante de E2 - Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis del sector energético y el crecimiento de las energías renovables", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
4. **Interpretador de código:** Para realizar análisis de datos preliminares.
5. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
6. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
7. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
8. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
9. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
10. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 21 de junio 2024

Firma:

7. BIBLIOGRAFÍA

- Adnan, M., Alarood, A. A. S., Uddin, M. I., & ur Rehman, I. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science.* Recuperado el día 17 de mayo de 2024 de PeerJ Computer Science: https://peerj.com/articles/cs-803/
- Amat Rodrigo, J. (2018). Machine Learning con R y caret. Recuperado el día 8 de mayo de 2024 de RPubs: https://rpubs.com/Joaquin_AR/383283
- Amat Rodrigo, J. (2021). Redes neuronales con R. Recuperado el día 7 de mayo de 2024 de RPubs: https://cienciadedatos.net/documentos/68-redes-neuronales-r
- APPA Renovables. (s.f.). Tipos de fuentes de energía renovable. Recuperado el día 28 de mayo de 2024 de: https://www.appa.es/energias-renovables/renovables-tipos-y-ventajas/tipos-de-fuentes-de-energia-renovable/
- Bashir, D., Montañez, G. D., Sehra, S., Sandoval Segura, P., & Lauw, J. (2020). An information-theoretic perspective on overfitting and underfitting. *AMISTAD Lab, Department of Computer Science, Harvey Mudd College*, CA, USA. Recuperado el día 10 de mayo de 2024: https://www.researchgate.net/publication/347202090_An_Information-Theoretic_Perspective_on_Overfitting_and_Underfitting
- Basogain Olabe, X. (s.f.). Redes neuronales artificiales y sus aplicaciones. *Universidad del País Vasco (UPV/EHU).* Recuperado el día 17 de junio de 2024 de: https://ocw.ehu.eus/file.php/102/redes_neuro/contenidos/pdf/libro-del-curso.pdf
- Bernal García, J. J., Martínez María-Dolores, S. M., & Sánchez García, J. F. (s.f). Modelización de los factores más importantes que caracterizan un sitio en la red. Recuperado el día 8 de enero de 2024 de: https://www.um.es/asepuma04/comunica/bernal_martinez_sanchez.pdf
- Camejo Corona, J., Gonzalez Diez, H. R., & Morell, C. (2020). Un estudio empírico del modelo de red neuronal MLP para problemas de predicción con salidas múltiples. Recuperado el día 30 de abril de 2024 de Dialnet: https://dialnet.unirioja.es/servlet/articulo?codigo=8590275

- Charte, F. (s.f.). Redes neuronales con R. Torre de Babel. Recuperado el día 3 de mayo de 2024 de Torre de Babel: https://fcharte.com/tutoriales/20160203-R-RedesNeuronales/
- Contento, M. R. (2019). Estadística con aplicaciones en R. Recuperado el día 17 de diciembre de: http://hdl.handle.net/20.500.12010/21660
- Deloitte (2018). Tendencias globales de las energías renovables. Recuperado el día 22 de octubre de 2023 de Deloitte: https://www2.deloitte.com/content/dam/Deloitte/es/Documents/energia/Deloitte-ES-tendencias-globales-energias-renovables.pdf
- Delua, J. (2021). Supervised vs. unsupervised learning: What's the difference? *IBM*. Recuperado el día 28 de abril de 2024 de IBM: https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning
- Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers*, *103*(1), 106–128. Recuperado el día 26 de diciembre de 2023 de JSTOR: http://www.jstor.org/stable/23485230
- Ember (2022). Global Electricity Review 2022. Recuperado el día 22 de octubre de 2023 de Ember Climate: https://ember-climate.org/app/uploads/2022/03/SP_Report-GER22.pdf
- Fausett, L. (1994). Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall. Recuperado el día 25 de abril de 2024 de Academia: https://www.academia.edu/24476782/Laurene_Fausett_Fundamentals_of_Neural_Networks
- Flórez López, R., & Fernández Fernández, J. M., (2008). Las redes neuronales artificiales. Netbiblo. Recuperado el día 25 de abril de 2024 de Google Académico: https://books.google.es/books?hl=es&lr=&id=X0uLwi1Ap4QC&oi=fnd&pg=PA11&dq=tipos+de+redes+neuronales+artificiales&ots=gPIBhqlpWi&sig=tVIdjK2-LnShR2CnGPoRt8u0wX8#v=onepage&q=tipos%20de%20redes%20neuronales%20artificiales&f=false

- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician, 43(1), 50–54*. Recuperado el día 15 de diciembre de JSTOR: https://doi.org/10.2307/2685173
- García García, F. (2014). Estadística descriptiva con R. Recuperado el día 17 de diciembre de 2023 de Github: https://biocosas.github.io/R/030_estadistica_descriptiva.html
- Gil Martínez, C. (2018). Análisis de Componentes Principales (PCA) en R. Recuperado el día 27 de diciembre de 2023 de RPUBS: https://rpubs.com/Cristina_Gil/PCA
- Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *U.S. Geological Survey Central Midwest Water Science Center*. Recuperado el día 3 de mayo de 2024 de: https://gmd.copernicus.org/articles/15/5481/2022/gmd-15-5481-2022.pdf
- Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts. Recuperado el día 11 de mayo de 2024 de OTexts: OTexts.com/fpp3
- IBM. (2023). Función LAG. *IBM*. Recuperado el día 10 de mayo de 2024 de IBM: https://www.ibm.com/docs/es/spss-statistics/saas?topic=expressions-lag-function
- IBM. (2023). Perceptrón multicapa. *IBM*. Recuperado el día 28 de abril de 2024 de IBM: https://www.ibm.com/docs/es/spss-statistics/saas?topic=networks-multilayer-perceptron
- IBM. (s.f.). ¿Qué es una red neuronal? *IBM*. Recuperado el día 26 de abril de 2024 de IBM: https://www.ibm.com/es-es/topics/neural-networks
- Kaggle (2023). Global Data on Sustainable Energy (2000-2020). Recuperado el día 22 de octubre de 2023 de Kaggle: https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy
- Kolluri, J., Kotte, K., Phridviraj, M. S. B., & Razia, S. (2020). Reducing overfitting problem in machine learning using novel L1/4 regularization method. Recuperado el día 7 de mayo de 2024 de Research Gate: https://www.researchgate.net/publication/343034965_Reducing_Overfitting_Problem_in_Machine_Learning_Using_Novel_L14_Regularization_Method

- Llinás Solano, H., & Rojas Álvarez, C. (2006). Estadística descriptiva. Ediciones Uninorte. Recuperado el día 12 de diciembre de Google Books: https://books.google.es/books?hl=es&lr=&id=3Tkb8HJ5toUC&oi=fnd&pg=PR11&dq=estad%C3%ADstica+descriptiva&ots=lU8SO55AUI&sig=NzsWQYET7ruQFTdEwRYBoROIpFM#v=onepage&q=estad%C3%ADstica%20descriptiva&f=false
- Martínez-Izquierdo, M., Molina-Sánchez, I., & Morillo-Balsera, M., (2019). Efficient dimensionality reduction using principal component analysis for image change detection. *IEEE Latin America Transactions, 17(4), 540-547*. Recuperado el día 26 de diciembre de 2023 de IEEE: https://latamt.ieeer9.org/index.php/transactions/article/view/1442/177
- Mavrou, I. (2015). Análisis factorial exploratorio: cuestiones conceptuales y metodológicas. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*. Recuperado el día 21 de mayo de 2024 de Revista Nebrija: https://revistas.nebrija.com/revista-linguistica/article/view/283/248
- Moral Pelaez, I. (2006). Capítulo 14. Modelos de regresión: lineal simple y regresión logística. Recuperado el día 18 de mayo de 2024 de: https://es.scribd.com/document/207066002/Modelos-de-regresion
- Morris & Opazo. (s.f.). Técnicas de reducción de dimensionalidad. *Morris & Opazo Blog.* Recuperado el día 21 de mayo de 2024 de Morris & Opazo Blog: https://blog.morrisopazo.com/blog/tecnicas-de-reduccion-de-dimensionalidad/
- Nielsen, M. (2015) Neural Networks and Deep Learning. Recuperado el día 10 de mayo de 2024: de https://www.ise.ncsu.edu/fuzzy-neural/wp-content/uploads/sites/9/2022/08/neuralnetworksanddeeplearning.pdf
- Park, S., & Zhao, H. (2019). Sparse principal component analysis with missing observations. *The Annals of Applied Statistics, 13(2), 1016–1042*. Recuperado el día 26 de diciembre de 2023 de JSTOR: https://www.jstor.org/stable/26754180
- Portela, J., Roch-Dupré, D., Figueroa-Ferretti Garrigues, I., Yéboles, C., & Salazar, A. (2023). Monitoring the green transition in the power sector with the electricity generation emissions (EGE) tracker. *Journal of Renewable Energy Research*. Recuperado el día 20 de junio de 2024 de: https://www.sciencedirect.com/science/article/pii/S2211467X23001864

- R Project Org. (s.f.). Grid Search CV. Recuperado el día 13 de mayo de 2024 de R Project Org: https://search.r-project.org/CRAN/refmans/superml/html/GridSearchCV.html
- R Project Org. (s.f.). Lag a Time Series. *R Project Org*. Recuperado el día 10 de mayo de 2024 de R Project Org: https://search.r-project.org/CRAN/refmans/tis/html/lags.html
- Rivera, E. (2003). Introducción a las Redes Neuronales Artificiales. *Grupo De Inteligencia Artificial.* Recuperado el día 5 de mayo de 2024 de Academia: https://www.academia.edu/101610213/Introducci%C3%B3n_a_las_Redes_Neuronales_Artificiales?uc-sb-sw=38321345
- Rousson, V., & Gasser, T. (2004). Simple Component Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 53(4), 539–555*. Recuperado el día 26 de diciembre de 2023 de JSTOR: http://www.jstor.org/stable/3592573
- Santana, A., & Hernández, C. N. (s.f.). Objetos en R: Series temporales. *Departamento de Matemáticas, ULPGC*. Recuperado el día 3 de mayo de 2024 de: https://estadistica-dma.ulpgc.es/cursoR4ULPGC/14-seriesTemporales.html
- Serna Pineda, S. C. (2009). Comparación de árboles de regresión y clasificación y regresión logística. *Universidad Nacional de Colombia*. Recuperado el día 18 de mayo de 2024 de: https://repositorio.unal.edu.co/bitstream/handle/unal/2421/42694070_2009.pdf?sequence=1&isAllowed=y
- Sim, C. H., Gan, F. F., & Chang, T. C. (2005). Outlier Labeling with Boxplot Procedures. *Journal of the American Statistical Association, 100(470), 642–652*. Recuperado el día 15 de diciembre de JSTOR: http://www.jstor.org/stable/27590584
- UNIR. (2021). Redes neuronales artificiales: qué son y cuáles son sus usos. Recuperado el día 28 de abril de 2024 de UNIR: https://www.unir.net/ingenieria/revista/redes-neuronales-artificiales/
- United Nations (2023). Energías renovables: energías para un futuro más seguro. Recuperado el día 22 de octubre de 2022 de Naciones Unidas: https://www.un.org/es/climatechange/raising-ambition/renewable-energy

- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9.* Recuperado el día 21 de mayo de 2024 de Journal of Machine Learning Research: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

8. ANEXOS

Anexo 1. Script del código en R

```
datos_originales<- read.csv("global-data-on-sustainable-energy.csv")

#Carga de las librerías
library(dplyr)
library(tidyr)
library(readxl)
library(ggplot2)
library(corrplot)
library(ggfortify)
library(plotly)
library("MASS")
library(stats)
library(caret)
library(factoextra)
library(neuralnet)
library(forecast)
library(nnfor)
library(RSNNS)

#####LIMPIEZA Y TRATAMIENTO DE DATOS####
head(datos_originales)
str(datos_originales)

#Cambio del nombre de variables
Country <-datos_originales [,1]
Year <- datos_originales[,2]
A_Electricty_PercP<- datos_originales [,3]
A_CleanFuelsCooking_PercP <- datos_originales[,4]
Ren_Gen_Cap_PerCap <- datos_originales [,5]
Fin_Flows_USD <- datos_originales[,6]
RenShare_TotEnergyConsump<- datos_originales[,7]
Electricity_FossilFuels_TWh <- datos_originales[,8]
Electricity_Nuclear_TWh <- datos_originales[,9]
Electricity_Ren_TWh <- datos_originales[,10]
Electricity_LowCarbon_Perc<- datos_originales[,11]
Energy_Consump_KWhPC <- datos_originales[,12]
Energy_IntLevel_PrimE <- datos_originales[,13]
CO2Emissions_TonsCap<- datos_originales[,14]
Ren_PercEquiv_PrimE <- datos_originales[,15]
GDPGrowth_AnPerc<- datos_originales[,16]
GDP_PerCap<- datos_originales[,17]
Pop_Density <- datos_originales [,18]
Land_Area_Km2 <- datos_originales [,19]
Latitude <- datos_originales [,20]
```

```
Longitude <- datos_originales [,21]
datos_originales<-data.frame(Country, Year, A_Electricty_PercP,
 A_CleanFuelsCooking_PercP, Ren_Gen_Cap_PerCap, Fin_Flows_USD,
 RenShare_TotEnergyConsump, Electricity_FossilFuels_TWh,
 Electricity_Nuclear_TWh, Electricity_Ren_TWh, Electricity_LowCarbon_Perc,
 Energy_Consump_KWhPC, Energy_IntLevel_PrimE, CO2Emissions_TonsCap,
 Ren_PercEquiv_PrimE,GDPGrowth_AnPerc, GDP_PerCap, Pop_Density,
 Land_Area_Km2, Latitude, Longitude)

#Número de NA que existen en cada variable
sum(is.na(datos_originales$Country))
sum(is.na(datos_originales$Year))
sum(is.na(datos_originales$A_Electricty_PercP))
sum(is.na(datos_originales$A_CleanFuelsCooking_PercP))
sum(is.na(datos_originales$Ren_Gen_Cap_PerCap))
sum(is.na(datos_originales$Fin_Flows_USD))
sum(is.na(datos_originales$RenShare_TotEnergyConsump))
sum(is.na(datos_originales$Electricity_FossilFuels_TWh))
sum(is.na(datos_originales$Electricity_Nuclear_TWh))
sum(is.na(datos_originales$Electricity_Ren_TWh))
sum(is.na(datos_originales$Electricity_LowCarbon_Perc))
sum(is.na(datos_originales$Energy_Consump_KWhPC))
sum(is.na(datos_originales$Energy_IntLevel_PrimE))
sum(is.na(datos_originales$CO2Emissions_TonsCap))
sum(is.na(datos_originales$Ren_PercEquiv_PrimE))
sum(is.na(datos_originales$GDPGrowth_AnPerc))
sum(is.na(datos_originales$GDP_PerCap))
sum(is.na(datos_originales$Pop_Density))
sum(is.na(datos_originales$Land_Area_Km2))
sum(is.na(datos_originales$Latitude))
sum(is.na(datos_originales$Longitude))

#Se crean dos datasets diferentes
#PRIMER DATA SET - Todas las variables
sum(is.na(datos_originales))
datos_originales_noNA<-na.omit(datos_originales)

#SEGUNDO DATA SET - Se han excluido las variables que más valores NA tienen:
 Renewable_Perc_Equiv_Primary_Energy, Financial_Flows_USD,
 Renewable_Generating_Capacity_PerCap
datos_selec<-data.frame(Country, Year, A_Electricty_PercP,
 A_CleanFuelsCooking_PercP, RenShare_TotEnergyConsump,
 Electricity_FossilFuels_TWh, Electricity_Nuclear_TWh, Electricity_Ren_TWh,
 Electricity_LowCarbon_Perc, Energy_Consump_KWhPC, Energy_IntLevel_PrimE,
 CO2Emissions_TonsCap,GDPGrowth_AnPerc, GDP_PerCap, Pop_Density,
 Land_Area_Km2, Latitude, Longitude)
sum(is.na(datos_selec))
datos_selec_noNA<-na.omit(datos_selec)
```

```
#ANÁLISIS DESCRIPTIVO

#Resumen de las variables del conjunto original y seleccionado
resumen_original <-summary(datos_originales_noNA[,-c(1,2)])
resumen_selec <-summary(datos_selec_noNA[,-c(1,2)])

#Histogramas Conjunto original
par(mfrow=c(4,4))
hist(x=datos_originales_noNA$A_Electricty_PercP, main ='Acceso Electricidad',
 xlab='% Acceso Electricidad', ylab= 'Frecuencia', col='blue', cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$A_CleanFuelsCooking_PercP, main ='Acceso Fuentes
 Renovable', xlab='% Acceso Renovables', ylab= 'Frecuencia', col='blue', cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$RenShare_TotEnergyConsump, main ='Cuota de
 Energía Renovable', xlab='Energía Renovable/Consumo Total', ylab= 'Frecuencia',
 col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Electricity_FossilFuels_TWh,  main ='Electricidad
 Generada Combustible TW/h', xlab='Electricidad Combustible Fósil TW/h', ylab=
 'Frecuencia', col='blue',cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Electricity_Nuclear_TWh, main ='Electricidad
 Generada Nuclear TW/h', xlab='Electricidad Nuclear TW/h', ylab= 'Frecuencia',
 col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Electricity_Ren_TWh,  main='Electricidad Generada
 Renovables TW/h',xlab='Electricidad Renovable TW/h', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Electricity_LowCarbon_Perc, main ='Electricidad Baja
 Emisión', xlab='Electricidad Baja Emisión', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Energy_Consump_KWhPC, main ='Consumo Energía
 Primaria', xlab='Consumo Energía', ylab= 'Frecuencia', col='blue',  cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Energy_IntLevel_PrimE, main ='Intensidad Energética
 por PIB', xlab='Intensidad Energética', ylab= 'Frecuencia', col='blue',cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$CO2Emissions_TonsCap, main ='Emisiones CO2 Per
 Cápita', xlab='Emisiones CO2', ylab= 'Frecuencia', col='blue', cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$GDPGrowth_AnPerc, main ='Crecimiento anual PIB',
 xlab='% Crecimiento PIB', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9,
 cex.axis=0.6)
hist(x=datos_originales_noNA$GDP_PerCap, main ='PIB Per Cápita', xlab='PIB Per
 Cápita', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Pop_Density, main ='Densidad Población',
 xlab='Población', ylab= 'Frecuencia', col='blue',cex.main=1.4, cex.lab=0.9,
 cex.axis=0.6)
hist(x=datos_originales_noNA$Land_Area_Km2, main ='Superficie', xlab='Km2',
 ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
```

```
hist(x=datos_originales_noNA$Latitude, main ='Latitud', xlab='Electricidad Nuclear
 TW/h', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_originales_noNA$Longitude, main ='Longitud', xlab='Electricidad
 Nuclear TW/h', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
par(mfrow=c(1,1))

#Histogramas Conjunto seleccionado
par(mfrow=c(4,4))
hist(x=datos_selec_noNA$A_Electricty_PercP, main ='Acceso Electricidad', xlab='%
 Acceso Electricidad', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9,
 cex.axis=0.6)
hist(x=datos_selec_noNA$A_CleanFuelsCooking_PercP, main ='Acceso Fuentes
 Renovable', xlab='% Acceso Renovables', ylab= 'Frecuencia', col='blue', cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$RenShare_TotEnergyConsump, main ='Cuota de Energía
 Renovable', xlab='Energía Renovable/Consumo Total', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Electricity_FossilFuels_TWh,  main ='Electricidad Generada
 Combustible TW/h', xlab='Electricidad Combustible Fósil TW/h', ylab= 'Frecuencia',
 col='blue',cex.main=1.2, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Electricity_Nuclear_TWh, main ='Electricidad Generada
 Nuclear TW/h', xlab='Electricidad Nuclear TW/h', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Electricity_Ren_TWh,  main='Electricidad Generada
 Renovables TW/h',xlab='Electricidad Renovable TW/h', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Electricity_LowCarbon_Perc, main ='Electricidad Baja
 Emisión', xlab='Electricidad Baja Emisión', ylab= 'Frecuencia', col='blue',
 cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Energy_Consump_KWhPC, main ='Consumo Energía
 Primaria', xlab='Consumo Energía', ylab= 'Frecuencia', col='blue',  cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Energy_IntLevel_PrimE, main ='Intensidad Energética por
 PIB', xlab='Intensidad Energética', ylab= 'Frecuencia', col='blue',cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$CO2Emissions_TonsCap, main ='Emisiones CO2 Per
 Cápita', xlab='Emisiones CO2', ylab= 'Frecuencia', col='blue', cex.main=1.4,
 cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$GDPGrowth_AnPerc, main ='Crecimiento anual PIB',
 xlab='% Crecimiento PIB', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9,
 cex.axis=0.6)
hist(x=datos_selec_noNA$GDP_PerCap, main ='PIB Per Cápita', xlab='PIB Per Cápita',
 ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Pop_Density, main ='Densidad Población',
 xlab='Población', ylab= 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9,
 cex.axis=0.6)
hist(x=datos_selec_noNA$Land_Area_Km2, main ='Superficie', xlab='Km2', ylab=
 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
```

```
hist(x=datos_selec_noNA$Latitude, main ='Latitud', xlab='Grados decimales', ylab=
 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
hist(x=datos_selec_noNA$Longitude, main ='Longitud', xlab='Grados decimales', ylab=
 'Frecuencia', col='blue', cex.main=1.4, cex.lab=0.9, cex.axis=0.6)
par(mfrow=c(1,1))

#Boxplot Conjunto Original
par(mfrow=c(4,4))
boxplot(datos_originales_noNA$A_Electricty_PercP, main ='Acceso Electricidad',
 col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$A_CleanFuelsCooking_PercP, main ='Acceso
 Fuentes Renovable', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$RenShare_TotEnergyConsump, main ='Cuota de
 Energía Renovable', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$Electricity_FossilFuels_TWh,  main ='Electricidad
 Generada Combustible TW/h', col='blue',cex.main=1.4)
boxplot(x=datos_originales_noNA$Electricity_Nuclear_TWh, main ='Electricidad
 Generada Nuclear TW/h', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$Electricity_Ren_TWh,  main='Electricidad
 Generada Renovables TW/h', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$Electricity_LowCarbon_Perc, main ='Electricidad
 Baja Emisión',  col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$Energy_Consump_KWhPC, main ='Consumo
 Energía Primaria',  col='blue',  cex.main=1.4)
boxplot(x=datos_originales_noNA$Energy_IntLevel_PrimE, main ='Intensidad
 Energética por PIB',  col='blue',cex.main=1.4)
boxplot(x=datos_originales_noNA$CO2Emissions_TonsCap, main ='Emisiones CO2
 Per Cápita',  col='blue', cex.main=1.4 )
boxplot(x=datos_originales_noNA$GDPGrowth_AnPerc, main ='Crecimiento anual
 PIB', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$GDP_PerCap, main ='PIB Per Cápita', col='blue',
 cex.main=1.4)
boxplot(x=datos_originales_noNA$Pop_Density, main ='Densidad Población',
 col='blue',cex.main=1.4)
boxplot(x=datos_originales_noNA$Land_Area_Km2, main ='Superficie', col='blue',
 cex.main=1.4)
boxplot(x=datos_originales_noNA$Latitude, main ='Latitud', col='blue', cex.main=1.4)
boxplot(x=datos_originales_noNA$Longitude, main ='Longitud', col='blue',
 cex.main=1.4)
par(mfrow=c(1,1))

#Boxplot Conjunto Seleccionado
par(mfrow=c(4,4))
boxplot(datos_selec_noNA$A_Electricty_PercP, main ='Acceso Electricidad',
 col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$A_CleanFuelsCooking_PercP, main ='Acceso Fuentes
 Renovable', col='blue', cex.main=1.4)
```

```
boxplot(x=datos_selec_noNA$RenShare_TotEnergyConsump, main ='Cuota de
 Energía Renovable', col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$Electricity_FossilFuels_TWh,  main ='Electricidad
 Generada Combustible TW/h', col='blue',cex.main=1.4)
boxplot(x=datos_selec_noNA$Electricity_Nuclear_TWh, main ='Electricidad Generada
 Nuclear TW/h', col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$Electricity_Ren_TWh,  main='Electricidad Generada
 Renovables TW/h', col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$Electricity_LowCarbon_Perc, main ='Electricidad Baja
 Emisión',  col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$Energy_Consump_KWhPC, main ='Consumo Energía
 Primaria',  col='blue',  cex.main=1.4)
boxplot(x=datos_selec_noNA$Energy_IntLevel_PrimE, main ='Intensidad Energética
 por PIB',  col='blue',cex.main=1.4)
boxplot(x=datos_selec_noNA$CO2Emissions_TonsCap, main ='Emisiones CO2 Per
 Cápita',  col='blue', cex.main=1.4 )
boxplot(x=datos_selec_noNA$GDPGrowth_AnPerc, main ='Crecimiento anual PIB',
 col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$GDP_PerCap, main ='PIB Per Cápita', col='blue',
 cex.main=1.4)
boxplot(x=datos_selec_noNA$Pop_Density, main ='Densidad Población',
 col='blue',cex.main=1.4)
boxplot(x=datos_selec_noNA$Land_Area_Km2, main ='Superficie', col='blue',
 cex.main=1.4)
boxplot(x=datos_selec_noNA$Latitude, main ='Latitud', col='blue', cex.main=1.4)
boxplot(x=datos_selec_noNA$Longitude, main ='Longitud', col='blue', cex.main=1.4)
par(mfrow=c(1,1))

#Correlación - datos originales
matriz_correlacion_datos_originales_noNA <-cor(datos_originales_noNA[,-c(1,2)])
corrplot(matriz_correlacion_datos_originales_noNA, tl.cex = 0.7,tl.col ="black",
 main="Matriz de Correlación del conjunto original")
#Correlación - datos seleccionados
matriz_correlacion_datos_selec_noNA<-cor(datos_selec_noNA[,-c(1,2)])
corrplot(matriz_correlacion_datos_selec_noNA, tl.cex = 0.7,tl.col ="black",
 main="Matriz de Correlación del conjunto seleccionado")

##########PRIMER DATA SET (DATOS ORIGINALES)#######################

#PCA
# Se aplica PCA
scale(datos_originales_noNA[, -c(1,2,7)]) #se quitan las variables categóricas y la
 objetivo
datos_originales_noNA_pca<-prcomp(datos_originales_noNA[,-c(1,2,5)],
 scale.=TRUE)
loadings_orig <-datos_originales_noNA_pca$rotation
scores_orig<-datos_originales_noNA_pca$x
```

```
#Análisis de Componentes Principales
cor(datos_originales_noNA_pca$x)
corrplot(cor(datos_originales_noNA_pca$x), method = "color", tl.cex=0.5)
sum(datos_originales_noNA_pca$rotation[,1]*datos_originales_noNA_pca$rotation[,2]
 )

##Comprobación de que las varianzas de las nuevas variables es igual a la suma de las
 varianzas de las variables originales estandarizadas
sum(apply(scale(datos_originales_noNA[, -c(1,2,7)]), MARGIN=2, var))
sum(apply(datos_originales_noNA_pca$x, MARGIN=2, var))

#Proporción de varianza explicada (PVE)
PVE_orig<-summary(datos_originales_noNA_pca)$importance[2,]
PVE_acum_orig<-summary(datos_originales_noNA_pca)$importance[3,]

#Número de componentes principales  - Gráfico de sedimentación y de PVE acumulada
par(mfrow=c(1,2))
plot(PVE_orig, xlab="Componente principal", ylab="PVE", ylim=c(0,1), type="o",
 col="blue", main="Gráfico de sedimentación", cex.main=0.9, cex.lab=0.8,
 cex.axis=0.6)
axis(side = 1, at = axTicks(1, axp = c(0, 19, 19)), labels = axTicks(1, axp = c(0, 19,
 19)), cex.axis=0.6)
plot(PVE_acum_orig, xlab="Componente principal", ylab="PVE acumulada",
 ylim=c(0,1), type="o", col="blue", main="Gráfico de PVE Acumulada", cex.main=0.9,
 cex.lab=0.8, cex.axis=0.6)
axis(side = 1, at = axTicks(1, axp = c(0, 19, 19)), labels = axTicks(1, axp = c(0, 19,
 19)),cex.axis=0.6)
par(mfrow=c(1,1))

#REDES NEURONALES
##RED CON EL PAQUETE NEURALNET
data_PCA_orig<-data.frame(Country=datos_originales_noNA$Country,
 Year=datos_originales_noNA$Year, pc1=scores_orig[,1], pc2=scores_orig[,2],
 pc3=scores_orig[,3], y=scale(datos_originales_noNA$RenShare_TotEnergyConsump))

#se guarda la media y variación estándar para posteriormente desentadarizar y volver a
 los valores originales
y_media_orig<-mean(datos_originales_noNA$RenShare_TotEnergyConsump)
y_sd_orig<-sd(datos_originales_noNA$RenShare_TotEnergyConsump)

#Se filtra el dataset para seleccionar el mismo periodo de tiempo
años_especificos <-
 c(2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018)
 #selección de los años

data_PCA_orig_filtro <- data.frame(
  Country = datos_originales_noNA$Country,
  Year = datos_originales_noNA$Year,
```

```
  pc1 = scores_orig[, 1],
  pc2 = scores_orig[, 2],
  pc3 = scores_orig[, 3],
  y = scale(datos_originales_noNA$RenShare_TotEnergyConsump)
)%>%
  filter(Year %in% años_especificos) #se filtra por los años especificados

data_PCA_orig_filtro<- data_PCA_orig_filtro %>%
  group_by(Country)%>%
  filter(all(años_especificos %in% Year) && length(unique(Year)) ==
 length(años_especificos)) %>%
  ungroup()

#Se añaden los retardos agrupado por país para mantener la temporalidad
datos_nnet_orig <- data_PCA_orig_filtro%>%
  arrange(Country, Year) %>%
  group_by(Country) %>%
  mutate(
    pc1_lag1 = lag(pc1, 1),
    pc2_lag1 = lag(pc2, 1),
    pc3_lag1 = lag(pc3, 1)
  ) %>%
  drop_na()

str(datos_nnet_orig)

#Partición de los datos: se divide el dataset en train y test (último año)
train_data_nnet_orig <- datos_nnet_orig %>%
  filter(Year <= max(Year) - 1)

test_data_nnet_orig <- datos_nnet_orig %>%
  filter(Year > max(Year) - 1)

#Se entrena la red con el paquete neuralnet
nn_model_orig <- neuralnet(
  y ~  pc1 + pc2 + pc3 + pc1_lag1 + pc2_lag1 + pc3_lag1,
  data = train_data_nnet_orig,
  hidden = c(5,5),
  linear.output = TRUE,
  threshold=0.01,
  stepmax=1e+08
)

plot(nn_model_orig, rep="best")

# Evaluación de las predicciones sobre el conjunto de prueba
test_predictions_nnet_orig <- compute(nn_model_orig, test_data_nnet_orig)
predicted_test_values_nnet_orig <- test_predictions_nnet_orig$net.result
```

```
# RMSE para el conjunto de test
test_actual_values_nnet_orig <- test_data_nnet_orig$y
test_errores_orig<-predicted_test_values_nnet_orig - test_actual_values_nnet_orig
hist_test_errores_orig<- hist(test_errores_orig, main="Error test red neuronal conjunto
 original", xlab="Error", ylab="Frecuencia")
test_rmse <- sqrt(mean((predicted_test_values_nnet_orig -
 test_actual_values_nnet_orig)^2))
print(test_rmse)

# Evaluación de las predicciones sobre el conjunto de prueba
train_predictions_nnet_orig <- compute(nn_model_orig, train_data_nnet_orig)
predicted_train_values_nnet_orig <- train_predictions_nnet_orig$net.result

# RMSE para el conjunto de train
train_actual_values_nnet_orig <- train_data_nnet_orig$y
train_errores_orig<-predicted_train_values_nnet_orig - train_actual_values_nnet_orig
hist(train_errores_orig, main="Error train red neuronal conjunto original", xlab="Error",
 ylab="Frecuencia")
train_rmse <- sqrt(mean((predicted_train_values_nnet_orig -
 train_actual_values_nnet_orig)^2))
print(train_rmse)

# Gráfico de dispersión para comparar los valores reales y predichos
ggplot(data = test_data_nnet_orig, aes(x = test_actual_values_nnet_orig, y =
 predicted_test_values_nnet_orig)) +
  geom_point(aes(color = "red"), alpha = 0.5) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Real vs. Predicho", x = "Valores Reales", y = "Valores Predichos") +
  theme_minimal()

#Predicciones de variables predictoras

forecast_pc1<-lm(pc1~y+pc2+pc3+pc2_lag1+pc3_lag1, data=datos_nnet_orig)
forecast_pc2<-lm(pc2~y+pc1+pc3+pc1_lag1+pc3_lag1, data=datos_nnet_orig)
forecast_pc3<-lm(pc3~y+pc2+pc1+pc1_lag1+pc2_lag1, data=datos_nnet_orig)

summary(forecast_pc1)$r.squared
summary(forecast_pc2)$r.squared
summary(forecast_pc3)$r.squared

ultimos_datos_por_pais <- datos_nnet_orig%>%
  group_by(Country) %>%
  slice(n()) %>%
  ungroup()

#Predicciones futuras
predicciones_futuras_orig<-data.frame()
```

```
for (pais in unique(ultimos_datos_por_pais$Country)) {
  ultimos_datos <- ultimos_datos_por_pais[ultimos_datos_por_pais$Country==pais, ]

  for (i in 1:11) {
    pred_nueva <- compute(nn_model_orig, ultimos_datos[, c("pc1", "pc2", "pc3",
 "pc1_lag1", "pc2_lag1", "pc3_lag1")])

    predicciones_futuras_orig <- rbind(predicciones_futuras_orig, data.frame(Country =
 pais, Year = as.numeric(ultimos_datos$Year) + i, Predicted_y =
 pred_nueva$net.result[1]))

    ultimos_datos$pc1_lag1 <- ultimos_datos$pc1
    ultimos_datos$pc2_lag1 <- ultimos_datos$pc2
    ultimos_datos$pc3_lag1 <- ultimos_datos$pc3

    ultimos_datos$pc1 <- predict(forecast_pc1, newdata = ultimos_datos)
    ultimos_datos$pc2 <- predict(forecast_pc2, newdata = ultimos_datos)
    ultimos_datos$pc3 <- predict(forecast_pc3, newdata = ultimos_datos)
  }
}

print(predicciones_futuras_orig)

# Predicciones
predicciones_futuras_orig <- predicciones_futuras_orig %>% arrange(Country, Year)
print(predicciones_futuras_orig)
predicciones_futuras_orig$Predicted_y<-predicciones_futuras_orig$Predicted_y *
 y_sd_orig + y_media_orig
print(predicciones_futuras_orig)

ggplot(predicciones_futuras_orig, aes(x = Year, y = Predicted_y, color = Country,
 group = Country)) +
  geom_line() +
  labs(title = "Predicciones Futuras por País", x = "Año", y = "Cuota de Energía
 Renovable") +
  scale_x_continuous(breaks = seq(min(predicciones_futuras_orig$Year),
 max(predicciones_futuras_orig$Year), by = 1)) +
  guides(color = guide_legend(title = "País"))

#Valores reales no escalados
data_PCA_orig_filtro_no_scale <- data.frame(
  Country = datos_originales_noNA$Country,
  Year = datos_originales_noNA$Year,
  pc1 = scores_orig[, 1],
  pc2 = scores_orig[, 2],
  pc3 = scores_orig[, 3],
  y = datos_originales_noNA$RenShare_TotEnergyConsump
)%>%
```

```
  filter(Year %in% años_especificos) #se filtra por los años especificados

data_PCA_orig_filtro_no_scale<- data_PCA_orig_filtro_no_scale %>%
  group_by(Country)%>%
  filter(all(años_especificos %in% Year) && length(unique(Year)) ==
 length(años_especificos)) %>%
  ungroup()

ggplot(data_PCA_orig_filtro_no_scale, aes(x = Year, y = y, color = Country, group =
 Country)) +
  geom_line() +
  labs(title = "Valores reales por país", x = "Año", y = "Cuota de Energía Renovable") +
  scale_x_continuous(breaks = seq(min(predicciones_futuras_orig$Year),
 max(predicciones_futuras_orig$Year), by = 1)) +
  guides(color = guide_legend(title = "País"))


################ SEGUNDO DATA SET - SELECCIONADO###################
#datos_selec_noNA

#PCA
# Se aplica PCA
scale(datos_selec_noNA[, -c(1,2,5)])
datos_selec_noNA_pca<-prcomp(datos_selec_noNA[,-c(1,2,5)], scale.=TRUE)
loadings_selec<-datos_selec_noNA_pca$rotation
scores_selec<-datos_selec_noNA_pca$x

#Análisis de Componentes Principales
cor(datos_selec_noNA_pca$x)
corrplot(cor(datos_selec_noNA_pca$x), method = "color", tl.cex=0.5)
sum(datos_selec_noNA_pca$rotation[,1]*datos_selec_noNA_pca$rotation[,2])
summary(datos_selec_noNA_pca)

##Comprobamos que las varianzas de las nuevas variables es igual a la suma de las
 varianzas de las variables originales estandarizadas
sum(apply(scale(datos_selec_noNA[, -c(1,2,5)]), MARGIN=2, var))
sum(apply(datos_selec_noNA_pca$x, MARGIN=2, var))

#Calculamos la proporción de varianza explicada
PVE_selec<-summary(datos_selec_noNA_pca)$importance[2,]
PVE_acum_selec<-summary(datos_selec_noNA_pca)$importance[3,]

#Número de componentes principales  - Gráfico de sedimentación y de PVE acumulada
par(mfrow=c(1,2))
plot(PVE_selec, xlab="Componente principal", ylab="PVE", ylim=c(0,1), type="o",
 col="blue", main="Gráfico de sedimentación", cex.main=0.9, cex.lab=0.8,
 cex.axis=0.6)
```

```
axis(side = 1, at = axTicks(1, axp = c(0, 19, 19)), labels = axTicks(1, axp = c(0, 19,
 19)), cex.axis=0.6)
plot(PVE_acum_selec, xlab="Componente principal", ylab="PVE acumulada",
 ylim=c(0,1), type="o", col="blue", main="Gráfico de PVE Acumulada", cex.main=0.9,
 cex.lab=0.8, cex.axis=0.6)
axis(side = 1, at = axTicks(1, axp = c(0, 19, 19)), labels = axTicks(1, axp = c(0, 19,
 19)),cex.axis=0.6)
par(mfrow=c(1,1))

#Explicación de componentes principales
loadings_selec[,1]
loadings_selec[,2]
loadings_selec[,3]

#REDES NEURONALES
##RED NEURONAL CON NNET Y NEURALNET
data_PCA_selec<-data.frame(Country=datos_selec_noNA$Country,
 Year=datos_selec_noNA$Year, pc1=scores_selec[,1], pc2=scores_selec[,2],
 pc3=scores_selec[,3], y=scale(datos_selec_noNA$RenShare_TotEnergyConsump))

#se guarda la media y variación estándar para posteriormente desentadarizar y volver a
 los valores originales
y_media_selec<-mean(datos_selec_noNA$RenShare_TotEnergyConsump)
y_sd_selec<-sd(datos_selec_noNA$RenShare_TotEnergyConsump)

#Se filtra el dataset para seleccionar el mismo periodo de tiempo
años_especificos <-
 c(2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018)
 #selección de los años

data_PCA_selec_filtro <- data.frame(
  Country = datos_selec_noNA$Country,
  Year = datos_selec_noNA$Year,
  pc1 = scores_selec[, 1],
  pc2 = scores_selec[, 2],
  pc3 = scores_selec[, 3],
  y = scale(datos_selec_noNA$RenShare_TotEnergyConsump)
)%>%
  filter(Year %in% años_especificos)

data_PCA_selec_filtro<- data_PCA_selec_filtro %>%
  group_by(Country)%>%
  filter(all(años_especificos %in% Year) && length(unique(Year)) ==
 length(años_especificos)) %>%
  ungroup()

#Una vez filtrado el dataset, se crean los retardos de las componentes principales
datos_nnet_selec <- data_PCA_selec_filtro%>%
```

```
  arrange(Country, Year) %>%
  group_by(Country) %>%
  mutate(
    pc1_lag1 = lag(pc1, 1),
    pc2_lag1 = lag(pc2, 1),
    pc3_lag1 = lag(pc3, 1)
  ) %>%
  drop_na()

str(datos_nnet_selec)

datos_nnet_selec <- data_PCA_selec_filtro%>%
  arrange(Country, Year) %>%
  group_by(Country) %>%
  mutate(
    pc1_lag1 = lag(pc1, 1),
    pc2_lag1 = lag(pc2, 1),
    pc3_lag1 = lag(pc3, 1)
  ) %>%
  drop_na()

#RED1 - NNET
#Se divide el dataset en train y test (último año)
train_data_nnet_selec <- datos_nnet_selec %>%
  filter(Year <= max(Year) - 1)

test_data_nnet_selec <- datos_nnet_orig %>%
  filter(Year > max(Year) - 1)

#Validación cruzada grid search para la red NNET
train_control <- trainControl(
  method = "cv",
  number = 10,
  search= "grid"
)
parametros <- expand.grid(size = c(1, 5, 10, 15, 20),
                          decay = c(0, 0.001, 0.01, 0.1, 1))

nnet_modelo_selec <- caret::train(x = train_data_nnet_selec,
                                  y = train_data_nnet_selec$y,
                                  method = "nnet",
                                  trControl = train_control,
                                  tuneGrid = parametros,
                                  linout = TRUE,
                                  trace = FALSE,
                                  maxit = 100
)
```

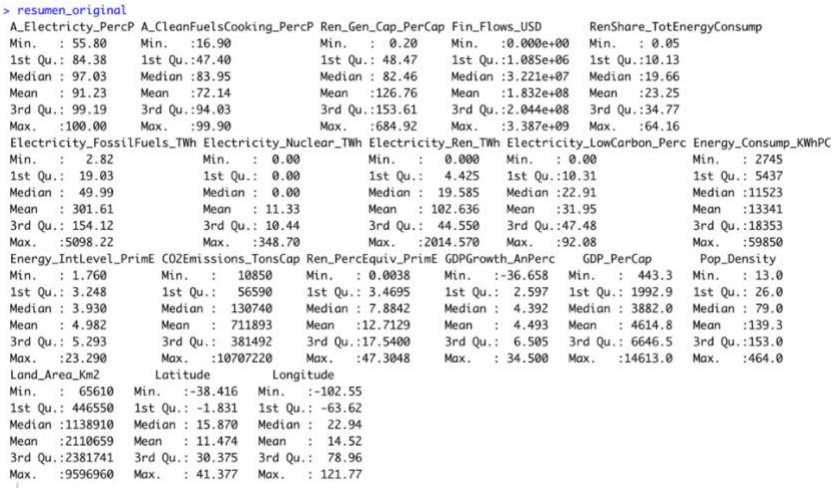
```
train_predictions_nnet_modelo_selec <- predict(nnet_modelo_selec, newdata =
 train_data_nnet_selec)

test_predictions_nnet_modelo_selec <- predict(nnet_modelo_selec, newdata =
 test_data_nnet_selec)

# Distribución de los errores
train_errores_nnet_modelo_selec<-train_predictions_nnet_modelo_selec -
 train_data_nnet_selec$y
hist(train_errores_nnet_modelo_selec, main="Error train red neuronal CV
 seleccionado", xlab="Error", ylab="Frecuencia")
test_errores_nnet_modelo_selec<-test_predictions_nnet_modelo_selec -
 test_data_nnet_selec$y
hist(test_errores_nnet_modelo_selec, main="Error test red neuronal CV seleccionado",
 xlab="Error", ylab="Frecuencia")

#RMSE de train y de test
train_rmse_nnet_modelo_selec <- RMSE(train_predictions_nnet_modelo_selec,
 train_data_nnet_selec$y)
test_rmse_nnet_modelo_selec <- RMSE(test_predictions_nnet_modelo_selec,
 test_data_nnet_selec$y)

#RED 2 - NEURALNET
neuralnet_modelo_selec <- neuralnet(
  y ~ pc1 + pc2 + pc3 + pc1_lag1 + pc2_lag1 + pc3_lag1,
  data = train_data_nnet_selec,
  hidden = c(5,4),
  linear.output = TRUE,
  threshold=0.03,
  stepmax=1e7
)

plot(neuralnet_modelo_selec, rep="best")

# Evaluación de las predicciones sobre el conjunto de test
test_predecir_neuralnet_modelo_selec <- compute(neuralnet_modelo_selec,
 test_data_nnet_selec)
predicciones_test_neuralnet_modelo_selec <-
 test_predecir_neuralnet_modelo_selec$net.result

# RMSE para el conjunto de prueba
test_valores_reales_neuralnet_modelo_selec <- test_data_nnet_selec$y
test_errores_neuralnet_modelo_selec<-predicciones_test_neuralnet_modelo_selec -
 test_valores_reales_neuralnet_modelo_selec
hist(test_errores_neuralnet_modelo_selec, main="Error test red neuronal NeuralNet
 seleccionado", xlab="Error", ylab="Frecuencia")
```

```
test_rmse_neuralnet_modelo_selec <-
 sqrt(mean((predicciones_test_neuralnet_modelo_selec -
 test_valores_reales_neuralnet_modelo_selec)^2))
print(test_rmse_neuralnet_modelo_selec)

# Evaluación de las predicciones sobre el conjunto de train
train_predecir_neuralnet_modelo_selec <- compute(neuralnet_modelo_selec,
 train_data_nnet_selec)
predicciones_train_neuralnet_modelo_selec <-
 train_predecir_neuralnet_modelo_selec$net.result

# RMSE para el conjunto de prueba
train_valores_reales_neuralnet_modelo_selec <- train_data_nnet_selec$y
train_errores_neuralnet_modelo_selec<-predicciones_train_neuralnet_modelo_selec -
 train_valores_reales_neuralnet_modelo_selec
hist(train_errores_neuralnet_modelo_selec, main="Error test red neuronal NeuralNet
 seleccionado", xlab="Error", ylab="Frecuencia")
train_rmse_neuralnet_modelo_selec <-
 sqrt(mean((predicciones_train_neuralnet_modelo_selec -
 train_valores_reales_neuralnet_modelo_selec)^2))
print(train_rmse_neuralnet_modelo_selec)

# Gráfico de dispersión - comparar los valores reales y predichos
ggplot(data = test_data_nnet_selec, aes(x = test_actual_values_nnet_selec, y =
 predicted_test_values_nnet_selec)) +
  geom_point(aes(color = "red"), alpha = 0.5) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Real vs. Predicho", x = "Valores Reales", y = "Valores Predichos") +
  theme_minimal()

#GRÁFICOS HISTOGRAMAS
par(mfrow=c(2,3))
hist(test_errores_orig, main="Error test red neuronal original", xlab="Error",
 ylab="Frecuencia")
hist(train_errores_orig, main="Error train red neuronal original", xlab="Error",
 ylab="Frecuencia")
hist(test_errores_neuralnet_modelo_selec, main="Error test red neuronal NeuralNet
 seleccionado", xlab="Error", ylab="Frecuencia")
hist(train_errores_neuralnet_modelo_selec, main="Error test red neuronal NeuralNet
 seleccionado", xlab="Error", ylab="Frecuencia")
hist(test_errores_nnet_modelo_selec, main="Error test red neuronal CV seleccionado",
 xlab="Error", ylab="Frecuencia")
hist(train_errores_nnet_modelo_selec, main="Error train red neuronal CV
 seleccionado", xlab="Error", ylab="Frecuencia")
par(mfrow=c(1,1))

#Comparación con modelo sin disminución de dimensión (PCA) y modelos de
 regresión múltiple
```

```
forecast_inicial<-lm(RenShare_TotEnergyConsump~. -Year - Country,
 data=datos_originales_noNA)
summary(forecast_inicial)
```

Anexo 2. Información estadística del conjunto original de datos

Tabla 11. Resumen estadístico del conjunto original de datos

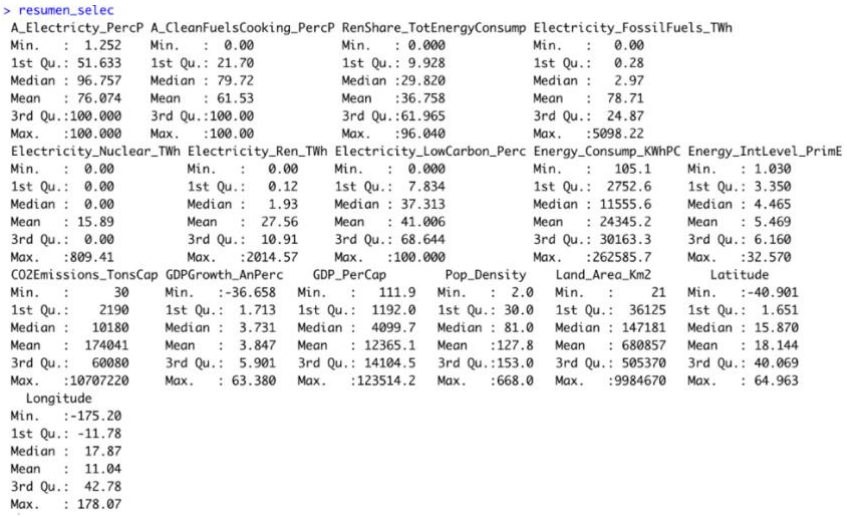


```
> resumen_original
 A_Electricty_PercP A_CleanFuelsCooking_PercP Ren_Gen_Cap_PerCap Fin_Flows_USD       RenShare_TotEnergyConsump
 Min.   : 55.80     Min.   :16.90             Min.   :  0.20     Min.   :0.000e+00   Min.   : 0.05
 1st Qu.: 84.38     1st Qu.:47.40             1st Qu.: 48.47     1st Qu.:1.085e+06   1st Qu.:10.13
 Median : 97.03     Median :83.95             Median : 82.46     Median :3.221e+07   Median :19.66
 Mean   : 91.23     Mean   :72.14             Mean   :126.76     Mean   :1.832e+08   Mean   :23.25
 3rd Qu.: 99.19     3rd Qu.:94.03             3rd Qu.:153.61     3rd Qu.:2.044e+08   3rd Qu.:34.77
 Max.   :100.00     Max.   :99.90             Max.   :684.92     Max.   :3.387e+09   Max.   :64.16
 Electricity_FossilFuels_TWh Electricity_Nuclear_TWh Electricity_Ren_TWh Electricity_LowCarbon_Perc Energy_Consump_KWhPC
 Min.   :   2.82             Min.   :  0.00         Min.   :   0.000    Min.   : 0.00              Min.   : 2745
 1st Qu.:  19.03             1st Qu.:  0.00         1st Qu.:   4.425    1st Qu.:10.31              1st Qu.: 5437
 Median :  49.99             Median :  0.00         Median :  19.585    Median :22.91              Median :11523
 Mean   : 301.61             Mean   : 11.33         Mean   : 102.636    Mean   :31.95              Mean   :13341
 3rd Qu.: 154.12             3rd Qu.: 10.44         3rd Qu.:  44.550    3rd Qu.:47.48              3rd Qu.:18353
 Max.   :5098.22             Max.   :348.70         Max.   :2014.570    Max.   :92.08              Max.   :59850
 Energy_IntLevel_PrimE CO2Emissions_TonsCap Ren_PercEquiv_PrimE GDPGrowth_AnPerc    GDP_PerCap       Pop_Density
 Min.   : 1.760        Min.   :   10850     Min.   : 0.0038     Min.   :-36.658   Min.   :  443.3   Min.   : 13.0
 1st Qu.: 3.248        1st Qu.:   56590     1st Qu.: 3.4695     1st Qu.:  2.597   1st Qu.: 1992.9   1st Qu.: 26.0
 Median : 3.930        Median :  130740     Median : 7.8842     Median :  4.392   Median : 3882.0   Median : 79.0
 Mean   : 4.982        Mean   :  711893     Mean   :12.7129     Mean   :  4.493   Mean   : 4614.8   Mean   :139.3
 3rd Qu.: 5.293        3rd Qu.:  381492     3rd Qu.:17.5400     3rd Qu.:  6.505   3rd Qu.: 6646.5   3rd Qu.:153.0
 Max.   :23.290        Max.   :10707220     Max.   :47.3048     Max.   : 34.500   Max.   :14613.0   Max.   :464.0
 Land_Area_Km2       Latitude         Longitude
 Min.   :  65610   Min.   :-38.416   Min.   :-102.55
 1st Qu.: 446550   1st Qu.: -1.831   1st Qu.: -63.62
 Median :1138910   Median : 15.870   Median :  22.94
 Mean   :2110659   Mean   : 11.474   Mean   :  14.52
 3rd Qu.:2381741   3rd Qu.: 30.375   3rd Qu.:  78.96
 Max.   :9596960   Max.   : 41.377   Max.   : 121.77
```

Fuente: Elaboración propia.

Anexo 3. Información estadística del conjunto seleccionado de datos

Tabla 12. Resumen estadístico del conjunto seleccionado de datos

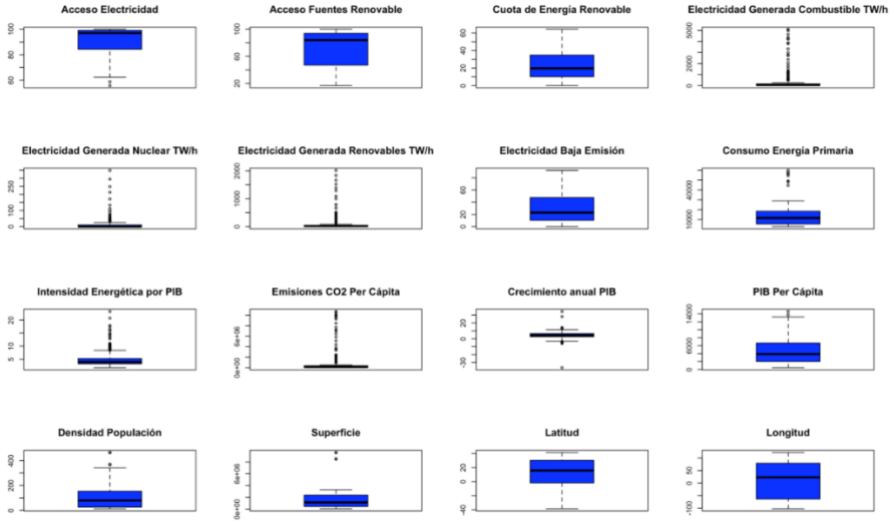


```
> resumen_selec
 A_Electricty_PercP A_CleanFuelsCooking_PercP RenShare_TotEnergyConsump Electricity_FossilFuels_TWh
 Min.   :  1.252    Min.   :  0.00            Min.   : 0.000            Min.   :   0.00
 1st Qu.: 51.633    1st Qu.: 21.70            1st Qu.: 9.928            1st Qu.:   0.28
 Median : 96.757    Median : 79.72            Median :29.820            Median :   2.97
 Mean   : 76.074    Mean   : 61.53            Mean   :36.758            Mean   :  78.71
 3rd Qu.:100.000    3rd Qu.:100.00            3rd Qu.:61.965            3rd Qu.:  24.87
 Max.   :100.000    Max.   :100.00            Max.   :96.040            Max.   :5098.22
 Electricity_Nuclear_TWh Electricity_Ren_TWh Electricity_LowCarbon_Perc Energy_Consump_KWhPC Energy_IntLevel_PrimE
 Min.   :  0.00          Min.   :   0.00     Min.   :  0.000           Min.   :   105.1     Min.   : 1.030
 1st Qu.:  0.00          1st Qu.:   0.12     1st Qu.:  7.834           1st Qu.:  2752.6     1st Qu.: 3.350
 Median :  0.00          Median :   1.93     Median : 37.313           Median : 11555.6     Median : 4.465
 Mean   : 15.89          Mean   :  27.56     Mean   : 41.006           Mean   : 24345.2     Mean   : 5.469
 3rd Qu.:  0.00          3rd Qu.:  10.91     3rd Qu.: 68.644           3rd Qu.: 30163.3     3rd Qu.: 6.160
 Max.   :809.41          Max.   :2014.57     Max.   :100.000           Max.   :262585.7     Max.   :32.570
 CO2Emissions_TonsCap GDPGrowth_AnPerc    GDP_PerCap       Pop_Density    Land_Area_Km2       Latitude
 Min.   :      30     Min.   :-36.658   Min.   :   111.9   Min.   :  2.0   Min.   :     21   Min.   :-40.901
 1st Qu.:    2190     1st Qu.:  1.713   1st Qu.:  1192.0   1st Qu.: 30.0   1st Qu.:  36125   1st Qu.:  1.651
 Median :   10180     Median :  3.731   Median :  4099.7   Median : 81.0   Median : 147181   Median : 15.870
 Mean   :  174041     Mean   :  3.847   Mean   : 12365.1   Mean   :127.8   Mean   : 680857   Mean   : 18.144
 3rd Qu.:   60080     3rd Qu.:  5.901   3rd Qu.: 14104.5   3rd Qu.:153.0   3rd Qu.: 505370   3rd Qu.: 40.069
 Max.   :10707220     Max.   : 63.380   Max.   :123514.2   Max.   :668.0   Max.   :9984670   Max.   : 64.963
   Longitude
 Min.   :-175.20
 1st Qu.: -11.78
 Median :  17.87
 Mean   :  11.04
 3rd Qu.:  42.78
 Max.   : 178.07
```

Fuente: Elaboración propia.

Anexo 4. Análisis de variables del conjunto original de datos

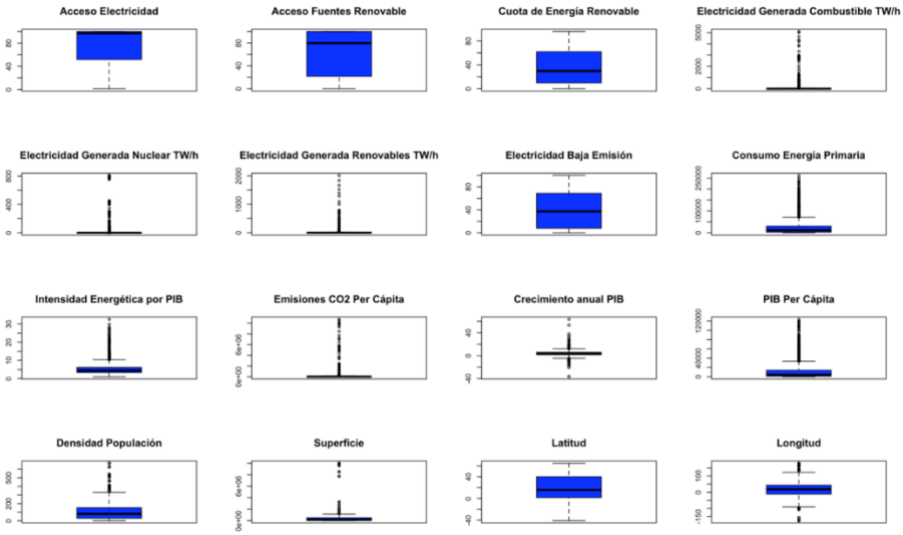
Figura 13. Diagramas de caja de las variables del conjunto de datos original



Fuente: Elaboración propia

Anexo 5. Análisis de variables del conjunto seleccionado de datos

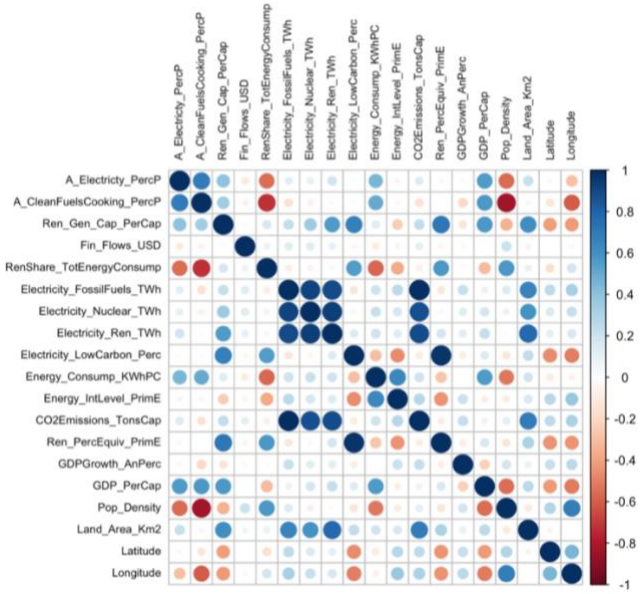
Figura 14. Diagramas de caja de variables del conjunto de datos seleccionado.



Fuente: Elaboración propia.

Anexo 6. Análisis de las relaciones entre las variables del conjunto original de datos

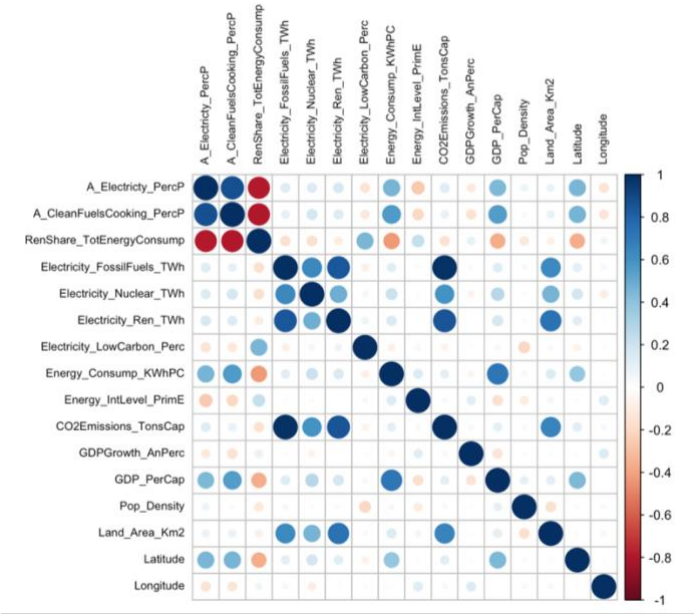
Figura 15. Matriz de Correlación de las variables del conjunto de datos originales.



Fuente: Elaboración propia.

Anexo 7. Análisis de las relaciones entre las variables del conjunto seleccionado de datos

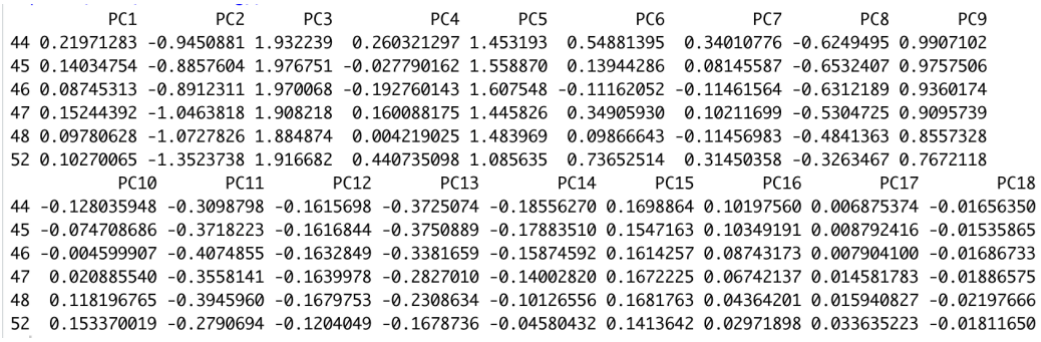
Figura 16. Matriz de Correlación de las variables del conjunto de datos originales.



Fuente: Elaboración propia.

Anexo 8. Matriz de scores de las componentes principales del conjunto de datos original.

Figura 17. Head matriz scores componentes principales del conjunto original.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
44	0.21971283	-0.9450881	1.932239	0.260321297	1.453193	0.54881395	0.34010776	-0.6249495	0.9907102
45	0.14034754	-0.8857604	1.976751	-0.027790162	1.558870	0.13944286	0.08145587	-0.6532407	0.9757506
46	0.08745313	-0.8912311	1.970068	-0.192760143	1.607548	-0.11162052	-0.11461564	-0.6312189	0.9360174
47	0.15244392	-1.0463818	1.908218	0.160088175	1.445826	0.34905930	0.10211699	-0.5304725	0.9095739
48	0.09780628	-1.0727826	1.884874	0.004219025	1.483969	0.09866643	-0.11456983	-0.4841363	0.8557328
52	0.10270065	-1.3523738	1.916682	0.440735098	1.085635	0.73652514	0.31450358	-0.3263467	0.7672118

	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
44	-0.128035948	-0.3098798	-0.1615698	-0.3725074	-0.18556270	0.1698864	0.10197560	0.006875374	-0.01656350
45	-0.074708686	-0.3718223	-0.1616844	-0.3750889	-0.17883510	0.1547163	0.10349191	0.008792416	-0.01535865
46	-0.004599907	-0.4074855	-0.1632849	-0.3381659	-0.15874592	0.1614257	0.08743173	0.007904100	-0.01686733
47	0.020885540	-0.3558141	-0.1639978	-0.2827010	-0.14002820	0.1672225	0.06742137	0.014581783	-0.01886575
48	0.118196765	-0.3945960	-0.1679753	-0.2308634	-0.10126556	0.1681763	0.04364201	0.015940827	-0.02197666
52	0.153370019	-0.2790694	-0.1204049	-0.1678736	-0.04580432	0.1413642	0.02971898	0.033635223	-0.01811650

Fuente: Elaboración propia

Anexo 9. Matriz de loading vectors de las componentes principales del conjunto de datos original.

Figura 18. Matriz de loading vectors de las componentes principales del conjunto original.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
A_Electricty_PercP	-0.123585054	-0.342259861	0.028690581	-0.03673502	0.36957341	-0.041839081	0.018520888	0.601271350	0.16001446
A_CleanFuelsCooking_PercP	-0.010280778	-0.435136287	0.084163526	0.06703856	0.21060046	-0.080455092	0.119946073	-0.097103149	0.04927887
Fin_Flows_USD	-0.048711203	0.075324514	-0.046938711	0.67659683	-0.12045933	-0.657851189	0.241416576	0.055948501	-0.01041226
RenShare_TotEnergyConsump	0.123324891	0.298900306	-0.341853551	-0.09504057	-0.17511484	0.004803529	0.066105989	0.244958323	-0.09636207
Electricity_FossilFuels_TWh	-0.430441857	0.075062287	-0.095608473	0.07072072	0.05340574	0.066855607	-0.095432284	-0.132783102	-0.02095613
Electricity_Nuclear_TWh	-0.404004046	0.036089530	-0.139439757	0.12688160	0.01278410	0.111600439	-0.147348583	-0.168928830	-0.33270040
Electricity_Ren_TWh	-0.400405784	0.007127103	-0.234137097	0.02601633	0.04492906	0.082058773	0.022322955	-0.015426657	-0.13248101
Electricity_LowCarbon_Perc	0.122191876	-0.060881971	-0.481142065	-0.18842532	0.04281927	-0.104848683	0.190477961	0.009647908	-0.24105041
Energy_Consump_KWhPC	-0.198699636	-0.291699101	0.188934940	-0.15527211	-0.43789631	-0.065109016	0.020165881	0.188725864	-0.28987232
Energy_IntLevel_PrimE	-0.213065209	-0.016750030	0.292415014	-0.32761677	-0.40881912	-0.149836871	0.461425984	-0.041473441	-0.08785319
CO2Emissions_TonsCap	-0.429288535	0.069708865	-0.101318986	0.02947037	0.06234629	0.056471403	-0.049240590	-0.120393325	0.07199628
Ren_PercEquiv_PrimE	0.106609528	-0.061347596	-0.483614817	-0.21204697	0.03202314	-0.076981611	0.227447903	0.121139776	-0.16234971
GDPGrowth_AnPerc	-0.117348520	0.131334497	0.060580213	-0.47898329	0.20692620	-0.688397763	-0.447123269	-0.058573166	-0.00946263
GDP_PerCap	-0.085023531	-0.368265233	-0.116060798	0.16850355	-0.19723821	0.025177536	-0.365708157	0.364785347	-0.18612757
Pop_Density	-0.002749056	0.434762168	-0.008409947	0.12808739	-0.04529251	0.018865475	-0.142663023	0.329293365	-0.01728480
Land_Area_Km2	-0.311556181	-0.052987233	-0.268993360	-0.11008039	-0.06344327	-0.035897300	0.243588776	0.030686252	0.67828636
Latitude	-0.141146585	0.178272989	0.253256665	-0.01705485	0.54877641	0.032226653	0.417362096	0.144303615	-0.37211235
Longitude	-0.157336293	0.342546858	0.197648733	-0.07842890	-0.11944626	0.106395782	0.000222744	0.431513849	0.13758949

	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
A_Electricty_PercP	-0.33141854	0.302132611	-0.020866854	-0.365047966	0.032401893	-0.0346617694	0.049158601	-0.010417172	-0.0028625847
A_CleanFuelsCooking_PercP	-0.09369338	-0.510193061	0.254939669	0.169070705	-0.115417617	-0.5575752921	-0.148360730	0.110412671	0.0288988414
Fin_Flows_USD	-0.07677196	0.005048466	-0.116337882	0.004363358	0.001075636	0.0589736140	0.013984214	-0.003574876	0.0027061004
RenShare_TotEnergyConsump	0.23248903	-0.092714391	-0.328645283	-0.381757540	-0.072323388	-0.5785093646	-0.046703070	0.080993739	0.0128817666
Electricity_FossilFuels_TWh	-0.09295994	0.275996745	-0.077865641	0.217294689	-0.091936138	-0.2557885529	0.061654913	0.156095501	-0.7213185272
Electricity_Nuclear_TWh	-0.16331079	-0.238478065	0.135662373	-0.470948750	0.125554396	0.2270262010	-0.298353839	0.374155327	0.0795516197
Electricity_Ren_TWh	-0.03445153	-0.312676236	0.035941406	-0.079848985	0.072464827	-0.0112544401	0.461875347	-0.656038188	0.0370657466
Electricity_LowCarbon_Perc	-0.19215825	0.147882047	0.096347909	0.185564789	-0.018277500	0.0802570041	-0.593606168	-0.370381334	-0.0906465674
Energy_Consump_KWhPC	0.05159904	-0.013643555	-0.075229882	-0.006839646	-0.686897878	0.1374379924	0.004074683	-0.013065902	0.0110304837
Energy_IntLevel_PrimE	-0.06137313	0.178891544	0.228458148	-0.082714576	0.467203200	-0.1877803808	0.030304738	0.016867176	-0.0004333673
CO2Emissions_TonsCap	-0.05724114	0.351865670	-0.164037935	0.306748495	-0.084344189	-0.2351798886	-0.073686734	0.067768570	0.6708236274
Ren_PercEquiv_PrimE	-0.14793273	-0.092989047	0.060674664	0.274111779	0.016395495	0.1749634760	0.474352699	0.487030586	0.0621864452
GDPGrowth_AnPerc	0.07989284	-0.093087813	-0.009050244	-0.011417693	0.020117543	-0.0007492809	0.008007444	-0.002495320	-0.0000986745
GDP_PerCap	0.45958466	0.011797580	0.015102125	0.264481393	0.429662355	-0.0073637153	-0.084630048	0.028451168	-0.0112100370
Pop_Density	0.02452098	0.090025747	0.770559809	0.063600294	-0.200917303	-0.1128426871	0.025998633	-0.004106928	0.0519316076
Land_Area_Km2	0.39764522	-0.123750010	0.114473020	-0.078654683	-0.099965496	0.2296572608	-0.162381056	0.064014114	-0.0637134390
Latitude	0.46385450	-0.017958736	-0.069890356	0.094119231	-0.044252858	0.1028871808	-0.067968273	0.037924701	-0.0190145712
Longitude	-0.35346834	-0.435804403	-0.278356638	0.342470755	0.125585711	0.1001593542	-0.210754706	0.024208798	-0.0390765519

Fuente: Elaboración propia

Anexo 10. Matriz de scores de las componentes principales del conjunto de datos seleccionado.

Figura 19. Head matriz de scores de las componentes principales del conjunto seleccionado.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
4	-1.447157	1.4961257	-0.44929004	0.170506305	-0.03269969	-1.797210	-0.001210644	0.6402372
5	-1.328007	1.2030353	0.22412500	0.381006429	-0.87927173	-1.114071	0.831350096	0.8644688
6	-1.359979	1.3902874	-0.59122731	-0.005831711	0.41059907	-1.986494	-0.179842021	0.5421496
7	-1.268547	1.1870855	-0.11321568	0.464920278	-0.38353061	-1.623615	0.344265090	0.7655604
8	-1.287732	1.3341216	-0.79136238	0.168694390	0.71672280	-2.434705	-0.538314746	0.4850375
9	-1.141770	0.9844529	-0.01159804	0.482384822	-0.42697682	-1.414065	0.538289102	0.8111773

	PC9	PC10	PC11	PC12	PC13	PC14	PC15
4	1.1036617	0.8704954	-0.3215738	-0.5515493	0.020101914	-0.6208859	-0.02001944
5	0.8977337	0.9352274	-0.3517090	-0.6442747	-0.042897350	-0.5157882	-0.01569123
6	1.0710530	0.7368495	-0.3206904	-0.5095141	-0.008685509	-0.4509065	-0.02312433
7	0.6816418	0.7235017	-0.3370161	-0.6072588	-0.121170798	-0.3788042	-0.02734294
8	0.8087329	0.5452418	-0.3082829	-0.5058429	-0.103487102	-0.3212121	-0.03518625
9	0.4775281	0.6115128	-0.3492412	-0.5871757	-0.186168061	-0.1724750	-0.02859044

Fuente: Elaboración propia

Anexo 11. Matriz de loading vectors de las componentes principales del conjunto de datos seleccionado.

Figura 20. Matriz de loading vectors de las componentes principales del conjunto seleccionado.

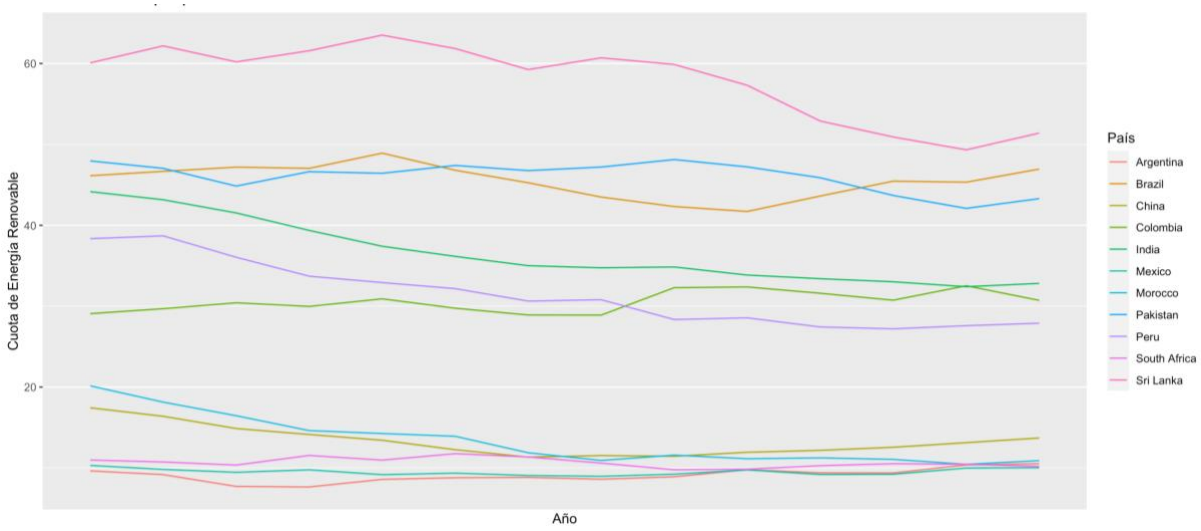
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
A_Electricty_PercP	0.31875515	-0.31097105	0.09302146	-0.082589527	0.188813200	-0.129221199	0.02284575	-0.006973199
A_CleanFuelsCooking_PercP	0.32752189	-0.33121842	-0.01282452	-0.081135891	0.163593462	-0.042588481	0.04528736	0.058275814
RenShare_TotEnergyConsump	-0.30391325	0.29796710	-0.28243907	-0.083023137	-0.228147523	-0.041907992	-0.11288600	-0.050773463
Electricity_FossilFuels_TWh	0.33219772	0.35555270	0.12555976	0.071196271	-0.055467546	-0.006456035	0.01634863	-0.101247225
Electricity_Nuclear_TWh	0.28934599	0.21060222	-0.05634527	-0.116631650	-0.146747578	0.051187167	-0.12673913	-0.213530105
Electricity_Ren_TWh	0.32106931	0.33713332	0.01207406	-0.048582326	-0.006634131	-0.071573464	0.02055979	0.055984415
Electricity_LowCarbon_Perc	-0.07403285	0.11912884	-0.52617930	-0.391717488	-0.165323223	-0.378691386	-0.06773128	0.001579570
Energy_Consump_KWhPC	0.27730059	-0.21224788	-0.38156489	0.185036714	-0.083455297	0.214440519	-0.10468229	0.355921578
Energy_IntLevel_PrimE	-0.06090984	0.14052906	-0.38478260	0.399794043	0.209948831	0.586759563	-0.16299495	-0.114977601
CO2Emissions_TonsCap	0.32824112	0.36155289	0.11981223	0.075018983	-0.030080042	-0.017640234	0.02159593	-0.089207709
GDPGrowth_AnPerc	-0.05233156	0.09316164	-0.03342385	0.436858819	0.380711668	-0.567092464	-0.52539593	0.171469532
GDP_PerCap	0.28269176	-0.21560224	-0.34323643	-0.001024982	-0.334516570	-0.064442448	-0.02956883	0.237141389
Pop_Density	0.03118557	-0.08538987	0.35189675	0.273157980	-0.680483250	0.014639297	-0.37838202	0.085775528
Land_Area_Km2	0.26423924	0.32307656	-0.03351884	-0.094037385	0.178154765	0.056961748	0.06359425	0.323942166
Latitude	0.23009059	-0.20130251	-0.21064609	0.143298409	-0.018609967	-0.136377525	-0.09943233	-0.767143507
Longitude	-0.03715960	0.05754399	-0.13856709	0.557393288	-0.172088524	-0.300254759	0.70095098	0.026334436

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
A_Electricty_PercP	0.341656101	-0.11300743	0.013750938	-0.153089012	0.596434053	-0.03227838	-0.472600057	-0.0069171343
A_CleanFuelsCooking_PercP	0.249593010	-0.15265420	-0.005035287	0.170883099	0.071215659	0.18143639	0.767439692	0.0217203562
RenShare_TotEnergyConsump	-0.056588835	0.20730850	0.034016208	-0.165037884	0.691218096	0.14292897	0.299707772	0.0211966792
Electricity_FossilFuels_TWh	0.020602544	-0.01907324	0.341120369	0.017025783	0.058806843	-0.30128693	0.113316895	-0.7101001124
Electricity_Nuclear_TWh	-0.393695877	-0.70732294	-0.201522781	-0.075891448	0.059084089	0.24074859	-0.049267394	0.0615123027
Electricity_Ren_TWh	0.209198805	0.32307363	0.093036717	-0.006705039	-0.172096405	0.74684936	-0.166580048	-0.0058583169
Electricity_LowCarbon_Perc	0.462878494	-0.21128173	0.005100145	-0.026882486	-0.271578292	-0.19123804	-0.064563991	-0.0234699569
Energy_Consump_KWhPC	-0.148138845	0.07025069	0.165196735	-0.656859185	-0.133908204	-0.04543361	0.044780269	-0.0012957871
Energy_IntLevel_PrimE	0.358773048	-0.15892455	-0.018283223	0.272502513	0.050111607	0.03012292	-0.092052986	-0.0060686115
CO2Emissions_TonsCap	0.074995315	0.05085272	0.322851021	0.012551828	0.025640417	-0.35053913	0.082576047	0.6992564109
GDPGrowth_AnPerc	-0.119974041	-0.07554530	0.017026393	0.054845338	0.004917335	0.02982711	0.017977892	-0.0024081175
GDP_PerCap	-0.323572568	0.16853633	0.105157216	0.626232676	0.124708127	-0.04359665	-0.169586212	0.0091025241
Pop_Density	0.345352939	-0.03239407	-0.238257776	-0.022389580	-0.020555851	-0.01338543	0.036038077	-0.0008509705
Land_Area_Km2	-0.004055784	0.20655206	-0.746426829	0.005702315	0.058470167	-0.25587714	0.045225867	-0.0326055135
Latitude	-0.076351100	0.36162501	-0.252353546	-0.099172345	-0.092093957	-0.06763335	0.029253522	-0.0160424828
Longitude	0.061931492	-0.17288031	-0.110295182	-0.043413214	0.045741336	0.06045198	0.009743384	0.0034775443

Fuente: Elaboración propia

Anexo 12. Valores reales de la cuota de energía renovable por país de conjunto de datos original

Figura 21. Cuota de energía renovable 2005-2018.



Fuente: Elaboración propia