Análisis de datos de la energía a nivel mundial aplicando herramientas de visualización y algoritmos de ML.

Ignacio, Bértola; Francisco, Franco; Ramiro Sclerandi**

UNRaf, Universidad Nacional de Rafaela, Santa Fe, Argentina.
Aprendizaje Automático y Grandes Datos (IC)

(**) Ramiro Sclerandi: ramirosclerandi1@gmail.com

Resumen:

Se llevó a cabo un estudio en dos fases sobre un conjunto de datos de energía con el fin de abordar los procesos involucrados en el aprendizaje automático. En la primera fase, se realizó un análisis exploratorio de datos que incluyó una descripción detallada del dataset y una presentación gráfica de las variables mediante histogramas y gráficos de tortas. Este enfoque inicial dio lugar a objetivos descriptivos relacionados con la comprensión de patrones y tendencias en los datos.

En la segunda fase, se formularon dos objetivos, uno que involucra la clasificación y otro la predicción. Se implementaron múltiples algoritmos de aprendizaje automático para cumplir con estos objetivos. Además, se realizará la comparación y evaluación de los algoritmos, utilizando técnicas y métricas específicas para cada uno de ellos. Esta fase permitió la identificación de modelos efectivos y precisos para clasificar y predecir variables relevantes en el contexto de la energía.

Palabras clave: Aprendizaje automático, conjunto de datos de energía, análisis exploratorio de datos, clasificación, regresión, métricas.

1. Introducción

El proyecto se contextualiza en un dataset integral que abarca indicadores de energía sostenible y otros factores relevantes en todos los países durante el período de 2000 a 2020. Este conjunto de datos proporciona información esencial sobre aspectos como el acceso a la electricidad, la energía renovable, las emisiones de carbono, la intensidad energética, flujos financieros y el crecimiento económico, entre otros aspectos. Su riqueza en datos lo convierte en una fuente valiosa para analizar y comprender el panorama global de consumo de energía a lo largo del tiempo.

La elección de este dataset se basó en su atractivo por dos razones principales. En primer lugar, el tema de la energía sostenible resultó altamente interesante para el grupo de investigación, dada su relevancia actual en la agenda global. En segundo lugar, el conjunto de datos ofrece una amplia variedad de tipos de información, lo que permite obtener una visión panorámica de la situación energética de la mayoría de los países del mundo.

El trabajo se divide en dos partes distintas pero complementarias. En la primera etapa, se realizó un análisis exploratorio de datos para comprender en profundidad el dataset, seguido de la formulación de objetivos descriptivos destinados a revelar patrones y tendencias en los datos.

Posteriormente, se enfocará en una segunda fase que incluirá objetivos del tipo predictivo y de clasificación. En esta etapa, se aplicarán diversos algoritmos de aprendizaje automático, incluyendo,

Estos algoritmos fueron escogidos cuidadosamente para abordar los desafíos específicos relacionados con el análisis de datos de energía sostenible y para ofrecer un enfoque integral en la resolución de problemas.

La evaluación de estos algoritmos se llevará a cabo utilizando una variedad de métricas apropiadas para cada uno, lo que permitirá una comparación exhaustiva del rendimiento de cada método. Métricas como la precisión, la sensibilidad, la especificidad, el valor F1, el error cuadrático medio y el coeficiente de silueta se utilizarán para medir la eficacia y la idoneidad de los algoritmos en función de los objetivos específicos de clasificación y predicción.

La combinación de algoritmos diversificados y métricas exhaustivas proporcionará una evaluación rigurosa, lo que permitirá determinar cuál de ellos se ajusta mejor a los objetivos planteados en este estudio.

Durante la búsqueda de trabajos relacionados, se encontró el estudio [1], que se enfoca en un análisis exploratorio centrado en la identificación de los parámetros esenciales en la contabilidad y la analítica de datos contables. En esta investigación, se plantean objetivos descriptivos con la finalidad de analizar los datos recopilados y extraer información valiosa sobre el rendimiento empresarial. Se llevaron a cabo múltiples análisis descriptivos, como la representación gráfica de la ubicación de las empresas, la comparación de los ingresos totales y los gastos totales, la identificación de los ratios financieros más relevantes y la detección de valores atípicos. Además,

se realizaron análisis de series temporales para identificar tendencias en los ingresos, gastos y ratios financieros a lo largo del tiempo.

En un enfoque similar, en el trabajo [2], se planteó el objetivo principal de explorar cómo las percepciones de los estudiantes pueden predecir su nivel de competencia en educación abierta. Para alcanzar este propósito, se aplicaron técnicas de aprendizaje automático supervisado para pronosticar el nivel de competencia de los estudiantes en educación abierta. Se emplearon varios algoritmos, como árboles de decisión, regresión logística, redes neuronales y máquinas de vectores de soporte, que posteriormente se compararon. Los resultados indicaron que los árboles de decisión y las máquinas de vectores de soporte se destacaron como los algoritmos más efectivos para predecir el nivel de competencia de los estudiantes en educación abierta.

Finalmente, en el estudio [3], se estableció el objetivo principal de definir una tipología de pacientes diabéticos mediante la construcción de modelos de aprendizaje automático a partir de información clínica registrada, medicación, herramientas de diagnóstico complementarias, datos terapéuticos y de monitoreo, y datos de medicación registrados. Para lograr este propósito, se emplearon algoritmos de agrupamiento, como K-means y DBSCAN, así como algoritmos de clasificación, como Random Forest y Support Vector Machine (SVM). El estudio utilizó una base de datos de pacientes diabéticos para entrenar y validar los modelos de aprendizaje automático. Los resultados resaltan la capacidad de estos modelos para identificar perfiles de pacientes diabéticos, lo que puede mejorar la atención médica y la gestión de la enfermedad.

En este estudio, se llevará a cabo un análisis exploratorio de datos, similar al trabajo [1], para examinar en profundidad el conjunto de datos en su totalidad. También se aplicarán algoritmos que brinden la posibilidad de predecir valores, similares a los utilizados en el trabajo [2] y otros algoritmos similares a los utilizados en el trabajo [3], que permitan realizar agrupaciones de datos.

Se utilizaron herramientas de análisis de datos, como la biblioteca Pandas de Python [4] y ydata_profilling [5] que permitieron obtener información valiosa sobre el conjunto de datos y sus variables. Además, se usaron librerías para visualizar los datos como lo son Matplotlib [6] y Seaborn [7].

Con el propósito de abordar de manera integral los datos, se establecieron una serie de objetivos de tipo descriptivo. Se planeó establecer un total de 12 objetivos, que incluyen desde preguntas básicas de visualización destinadas a proporcionar una visión general de diversos aspectos del conjunto de datos, hasta objetivos más complejos que involucran el análisis de tendencias. Este último tipo de objetivos se basará en el uso de series temporales, gráficos de dispersión e histogramas, lo que permitirá una exploración más profunda de las relaciones entre dos o tres variables clave.

Luego de completar la fase de los objetivos descriptivos, se avanzará hacia la formulación de objetivos orientados a la predicción y clasificación. Estos objetivos involucrarán tanto la predicción de variables basadas en relaciones con específicos como la agrupación de datos en categorías relevantes.

Para resolver estas cuestiones además de las librerías mencionadas anteriormente, se empleará principalmente la librería scikit-learn de Python [8].

En el caso de la predicción, se utilizarán enfoques de aprendizaje supervisado, haciendo uso de datos etiquetados para entrenar modelos que sean capaces de predecir valores de atributos específicos. Esto permitirá anticipar comportamientos o resultados en función de las relaciones entre las variables del conjunto de datos.

Además, se explorarán métodos de agrupación que no dependen de datos etiquetados. Estos algoritmos agruparán automáticamente datos similares en categorías, lo que puede proporcionar una comprensión más profunda de la estructura inherente de los datos.

Este enfoque diversificado de objetivos descriptivos, predictivos y de agrupación se traducirá en un análisis exhaustivo que arrojará luz sobre diferentes aspectos del conjunto de datos y ayudará a comprender mejor su contenido y contexto.

Los objetivos de tipo descriptivo dispararon las siguientes preguntas que se buscaron responder:

- ¿Cuáles son los países con mayor porcentaje de energía renovable consumida con respecto a la energía total consumida en un año específico?
- ¿Cómo varía el consumo de energía per cápita en Sudamérica, entre 2000 y 2020?
- ¿Cómo es la distribución de las fuentes de energía eléctrica en función de la energía total producida en Sudamérica?
- ¿Cómo contribuye cada una de las fuentes de energía (renovables y no renovables) a la matriz energética Argentina? y ¿Cómo varían estos aportes a lo largo del tiempo?
- ¿Qué tendencia marca la evolución de la generación de energía renovable en Sudamérica?
- ¿Cómo cambia el porcentaje de la energía renovable consumida con respecto al total de energía eléctrica consumida en Sudamérica?

- ¿Existe una relación directa entre los flujos de dinero recibidos y el porcentaje de energía renovable consumida sobre el porcentaje total de energía consumida en Sudamérica?
- ¿El dinero que ingresa tiene repercusión real en proyectos de energía renovable?
- ¿Cómo varía el consumo de energía, expresado en variación porcentual, en comparación con el crecimiento del PIB?
- ¿Cómo varía la producción de energías renovables (%), expresado en variación porcentual, en comparación con el crecimiento del PIB?
- ¿Qué países tienen las mayores emisiones y cuáles las menores? Tanto en Sudamérica y a nivel mundial.
- ¿La generación de energía renovable per cápita contribuye a las emisiones de dióxido de carbono?

Luego, los objetivos de predicción y de agrupación que se buscaron responder fueron los siguientes:

- Predecir consumos de energía per cápita anuales a nivel mundial.
- Agrupar países según su producción de electricidad baja en carbono y su participación de energía renovable en el consumo total de energía final.

2. Metodología

En esta sección se explayará sobre el análisis del dataset original. Se explicarán los procedimientos de limpieza, preprocesamiento y las herramientas analíticas aplicadas justificando su elección para lograr los objetivos planteados en la sección número uno.

Como primera medida se tomó el conjunto de datos original, obtenido desde la web Kaggle [9], y se le realizó un reporte utilizando la biblioteca ydata_profiling [5].

Los resultados obtenidos de este reporte se pueden ver en la figura 1. Se proporcionan datos como el número de variables, el número de registros y la cantidad de celdas con valores faltantes.

Este último dato supone un problema grave que tiene este dataset, sobre todo a la hora de responder los objetivos descriptivos, porque estos abarcan una gran cantidad de variables del dataset, y faltan datos en todas las columnas.

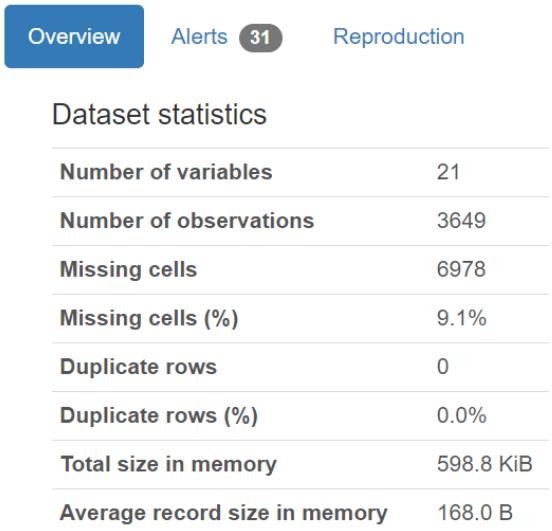

Figura 1. Reporte del dataset original.

2.1 Preprocesamiento de datos

Para solucionar el problema de los datos faltantes se pueden aplicar cuatro hay cuatro opciones principales:

"Como se muestra en la siguiente lista, existen cuatro enfoques diferentes para abordar el problema.

valores faltantes:

• Manténgalos como están.

• Elimine los objetos de datos (filas) con valores faltantes.

• Elimine los atributos (columnas) con valores faltantes.

• Estimar e imputar un valor.

Cada una de las estrategias anteriores podría ser la mejor estrategia en diferentes circunstancias.

De todos modos, cuando tratamos con valores faltantes, tenemos los dos objetivos siguientes:

• Mantener tantos datos e información como sea posible

• Introducir la menor cantidad posible de sesgo en nuestro análisis." [10]

Como se mencionó anteriormente los objetivos descriptivos planteados involucran una gran cantidad de variables, por lo tanto, la mejor opción para resolver el problema es estimar e imputar los valores faltantes.

De este modo se cumple con el objetivo de mantener la mayor cantidad de información posible. Pero, por contraparte, se introduce cierto nivel de sesgo al estimar los valores que faltan.

2.2 Algoritmo estimador e imputador

Para realizar las estimaciones de los valores faltantes se optó por utilizar el algoritmo de bosques aleatorios. Se eligió este modelo por sobre otros como la regresión lineal debido a su versatilidad y buen desempeño en una amplia variedad de aplicaciones.

"Uno de los grandes aspectos de los árboles de decisión es su flexibilidad. Dado que son conceptualmente sencillos (buscar regiones con resultados similares, etiquetar todo en esa región de alguna manera) se pueden adaptar fácilmente a otras

tareas.” [11].

“Los bosques aleatorios se basan en árboles de decisión. En lugar de utilizar un solo árbol, un bosque utiliza una colección de ellos (como habrás adivinado por el nombre).

Cada árbol se construye con una muestra de los datos que obtenemos del arranque o muestreo con reemplazo. También subconjuntos aleatorios de las características en cada división, como vimos que es posible con max_features para los árboles de decisión de sklearn.

Los bosques aleatorios también se pueden paralelizar entre procesadores u computadoras, ya que podemos construir cada árbol de decisión por separado porque son independientes entre sí.” [12].

En los bosques aleatorios, cada árbol del conjunto se construye a partir de una muestra extraída con reemplazo, es decir una muestra de arranque, del conjunto de entrenamiento. Además, al dividir cada nodo durante la construcción de un árbol, la mejor división se encuentra entre todas las características de entrada.

Los bosques aleatorios logran una variación reducida al combinar diversos árboles, a veces a costa de un ligero aumento en el sesgo. En la práctica, la reducción de la varianza suele ser significativa, por lo que se obtiene un mejor modelo en general.

La implementación de scikit-learn combina clasificadores promediando su predicción probabilística, en lugar de permitir que cada clasificador vote por una única clase.

Se procederá a realizar una explicación del código empleado, que se encuentra detallado en la sección 1.2. del notebook llamado “ProyectoEndToEnd” que se entregó en conjunto con el presente escrito y está en un repositorio en Github [13].

Primero se define una lista de nombres de columnas que se utilizarán para predecir los valores faltantes en otras columnas. En otras palabras, estas son las columnas que se considerarán para estimar los valores que faltan en el conjunto de datos.

Luego se crea un objeto imputador de datos. Este se utilizará para estimar y rellenar los valores faltantes en el conjunto de datos. El hiperparámetro “estimator” se configura con el algoritmo de bosques aleatorios de regresión para estimar los valores faltantes. También se establece el hiperparámetro de iteraciones en 15, lo que significa que se realizarán 15 repeticiones para imputar los valores faltantes. Y el hiperparámetro “random_state” se establece en 0 para asegurar la reproducibilidad de los resultados.

Después se utiliza el objeto creado anteriormente para ajustar el modelo a los datos en las columnas especificadas en anteriormente. Luego, se transforma el conjunto de datos rellenando los valores faltantes en esas columnas.

Finalmente, el dataset contendrá los datos originales de con los valores faltantes reemplazados por estimaciones calculadas por el modelo de regresión de bosque aleatorio.

2.3. Objetivos descriptivos

Los primeros seis objetivos descriptivos que se plantearon aportan conocimiento del conjunto de datos, ya que solo muestran distintas variables o evoluciones de estas a lo largo del tiempo. Por esto, se utilizaron gráficos sencillos para ilustrar estas cuestiones.

Para realizar estos gráficos, en cada caso, se filtró el conjunto de datos para trabajar sólo sobre un subconjunto relevante y graficar las variables pertinentes del objetivo planteado. Estos filtrados pueden ser por un país específico o un conjunto de estos, también por años específicos, y principalmente variables específicas del objetivo o un conjunto de estas.

A continuación, se explicará la metodología que se siguió para elaborar cada objetivo con su correspondiente visualización. Los códigos se encuentran detallados en el mencionado notebook, a partir de la sección 2.1. hasta la 2.6.

En la primera cuestión, se busca averiguar qué países tienen mayor porcentaje de energía renovable consumida. Se filtró el conjunto de datos con la variable de consumo de energías renovables respecto a la energía total consumida y se utilizó el año 2020. Para visualizar gráficamente los resultados, se utilizó un histograma que muestra los 30 países con mayor porcentaje de energía renovable consumida.

El segundo objetivo, se quiere conocer cómo evoluciona la variación del consumo de energía per cápita en Sudamérica. Se utilizaron un filtro para obtener los países que pertenecen a Sudamérica que aparecen en el conjunto de datos y otro filtro que obtenga la variable del consumo de energía eléctrica por persona. Para graficar esto, se empleó un gráfico de múltiples líneas donde se muestra una línea por país y estas muestran las tendencias de la variable seleccionada a lo largo de los 20 años que hay datos.

El tercer planteo busca averiguar cómo está distribuida la generación de energía eléctrica, es decir, cómo está constituida la matriz energética de los países de Sudamérica. Se emplearon dos filtros nuevamente, uno para seleccionar los países de Sudamérica y otro para seleccionar las variables que representen las fuentes de generación de energía eléctrica. Luego, para representar el porcentaje de cada fuente, primero una se debió confeccionar una nueva columna, la cual tenga la suma de las distintas fuentes de generación energía eléctrica. A continuación, se calculó el porcentaje que representaba cada una de las fuentes con respecto al total y se agregaron en columnas del dataset filtrado. Finalmente teniendo todos estos datos ya fue posible

poder graficar la comparación, la cual empleó un gráfico de barras apiladas para país.

El cuarto objetivo busca conocer también cómo evolucionan las distintas fuentes de producción de energía eléctrica pero solamente en Argentina. Por lo tanto, se realizó un filtro para obtener los valores correspondientes a Argentina y no Sudamérica, como los objetivos anteriores. Luego se tomaron las columnas de las diversas fuentes de generación de energía eléctrica y se graficaron los resultados. Para poder visualizarlo de una manera entendible, se utilizó un gráfico de áreas apiladas.

En el quinto planteo, se desea conocer cómo evoluciona temporalmente la generación de energía renovable per cápita en los países de Sudamérica. En este caso, se vuelve a aplicar el filtro de los países de Sudamérica y se filtra por la variable de generación de energía renovable per cápita. Para visualizar los resultados se empleó un gráfico de múltiples líneas, una para cada país que muestran la tendencia de la variable a lo largo de los años.

El sexto objetivo busca saber cómo fue la evolución del porcentaje de energías renovables consumidas en función de la energía total consumida en Sudamérica. Otra vez se aplica el filtro de los países sudamericanos presentes en el dataset, en conjunto con el de la variable que indica el consumo de energías renovables en función de la energía total consumida. Para visualizar los resultados se utilizó un gráfico de múltiples líneas donde cada país posee una línea y se permite ver la evolución a lo largo de los años.

Luego, el segundo grupo de seis objetivos descriptivos consta de preguntas que para responderse gráficamente utilizan dos variables, por lo que los gráficos contendrán dos variables y en algunos casos tres, incluyendo la evolución temporal.

Al igual que los objetivos descritos anteriormente, para poder llevar a cabo una visualización mediante gráficos, se aplicarán diversos filtros, incluyendo los de años específicos, conjunto de variables y/o registros específicos. Los códigos correspondientes a cada objetivo se encuentran detallados en el mencionado notebook, a partir de la sección 2.7. hasta la 2.12.

El primer objetivo de este grupo plantea la búsqueda de alguna relación directa entre los flujos de dinero recibidos y el porcentaje de energía renovable consumida, en los países de Sudamérica. Para resolver esta cuestión se aplican tres filtros, el primero es el de los países de Sudamérica, el segundo busca las variables relacionadas con el objetivo y el tercero es un año específico, en este caso se utilizó el 2019. Para poder visualizar correctamente los resultados se utilizó un gráfico de burbujas. En el eje de las abscisas se colocó el porcentaje de energía renovable consumida y en el eje de las ordenadas la variable de los flujos de dinero recibidos. La primera variable tiene una escala porcentual pero la otra variable posee valores muy grandes, por lo tanto, se utiliza una escala logarítmica.

El segundo planteamiento está relacionado con el anterior y busca encontrar si existe una correspondencia directa entre los flujos financieros recibidos y la generación de energía eléctrica por persona. Para llevar a cabo esta comparación se realizará sobre Argentina, por lo tanto, se aplica un filtro al dataset para seleccionar los datos correspondientes a Argentina. Luego, se indican las variables que se utilizarán en el gráfico realizando un filtrado, incluida la variable que contiene los años. Se utilizó un gráfico de dos líneas donde cada una se corresponde con las variables y se ve la evolución de estas a lo largo de los años. Además, cada variable tiene su escala, la capacidad de generación de electricidad tiene una escala natural, mientras que los flujos financieros, como el objetivo anterior, tiene una escala logarítmica.

La tercera cuestión plantea analizar si existe una conexión directa entre el consumo de energía eléctrica y la variación del PBI[1] de un año al siguiente en Argentina. Para poder realizar esto se debió tener las dos variables en una magnitud similar, como lo es la variación de un año a otro. Para resolver esta cuestión se calculó la variación anual del consumo de energía eléctrica utilizando un método de la librería Pandas. Posteriormente se agregaron todos los valores calculados en una nueva columna en el dataset. Finalmente, para visualizar estos valores, se empleó un gráfico de dos líneas que muestran ambas variables cómo evolucionan a lo largo de los años.

La cuarta cuestión plantea la existencia de una relación entre el porcentaje de energías renovables consumidas y la variación anual del PBI de Argentina. Primero se filtró el dataset para ver los datos de Argentina y seleccionar las variables involucradas. Luego, para realizar la comparación se utiliza la misma estrategia descrita en el objetivo anterior, calcular la variación anual de la variable de energías renovables consumidas y agregar los valores al conjunto de datos filtrado. Finalmente se grafica los resultados utilizando un gráfico de dos líneas que muestren la evolución de las variables.

El quinto objetivo busca mostrar un panorama general sobre los países con mayores emisiones de dióxido de carbono. Luego, mostrar la situación en Sudamérica y específicamente la evaluación temporal de Argentina y Brasil. Esto se realiza para ver si un alto nivel de industrialización tiene un impacto en las emisiones de CO^{22}. Para representar estas cuestiones se realizan tres gráficos que están relacionados entre sí. Para el primero se ordena el dataset por la columna

[1] Producto interno bruto

[2] Dióxido de Carbono

de emisiones de dióxido de carbono, se toman los primero 30 países y se grafica utilizando un histograma horizontal. Para el segundo, se filtra por los países de Sudamérica, también por la variable de emisiones y se grafican con un histograma vertical. Finalmente, para el tercero se buscan los datos de Argentina y Brasil, la columna de las emisiones y se grafican utilizando un gráfico con dos líneas que marque la evolución de la variable.

El sexto planteamiento busca una tendencia entre las emisiones de dióxido de carbono y la generación de energías renovables. Para resolver esta cuestión, se aplicaron filtros de las dos variables y los datos correspondientes a Argentina. Luego, se realiza el gráfico de las emisiones de CO^2 en función de la generación de energía renovables y ,finalmente, se confeccionó una regresión lineal sobre los datos para ver si los datos reflejan tal tendencia lineal.

2.4. Objetivo de agrupación

En este objetivo se buscó explorar y entender las dinámicas globales en base a la producción de electricidad baja en carbono y la participación de energía renovable en el consumo total de energía. La relevancia del análisis radica en su capacidad para identificar patrones y tendencias en las políticas energéticas y compromisos ambientales qué diferentes países toman sobre el desafío existente.

Para comprender dichas dinámicas y poder realizar un buen análisis se aplicaron diversas técnicas de agrupamiento para identificar dichos patrones de similitud entre los países en relación con el porcentaje de electricidad generada a partir de fuentes bajas en emisiones de dióxido de carbono y el consumo de energías renovables. Además, para cada algoritmo se utilizaron diferentes configuraciones de los hiperparámetros particulares, para conseguir el mejor desempeño de cada algoritmo.

Los códigos correspondientes a la agrupación comienzan en la sección 3.1. del notebook. Como primer paso se confeccionó un subconjunto de datos que contenga los datos de todos los países, pero sólo de los últimos cinco años, para tener una tendencia actualizada. Luego se calculó la media de las variables que se utilizarán para realizar el agrupamiento, dando como resultado un registro por país.

Luego, antes de aplicar los distintos algoritmos, se debe obtener el número óptimo de clusters debido a que algunos algoritmos lo necesitan tener este dato de antemano para funcionar correctamente.

Para realizar esto se utilizó la gráfica del codo. Primero se seleccionaron las características que se utilizarán para agrupar, luego se escalaron los valores utilizando la estandarización y finalmente se aplicó K-Means con diversas cantidades de clusters para obtener la inercia o la suma de las distancias cuadradas[3] de cada cluster. Esta magnitud va disminuyendo a medida que aumente el número de clusters, y la cantidad exacta se encuentra en la parte de la gráfica que se suaviza la pendiente.

En la figura 2 se puede ver el gráfico del codo. El número óptimo de clusters sería cuatro.

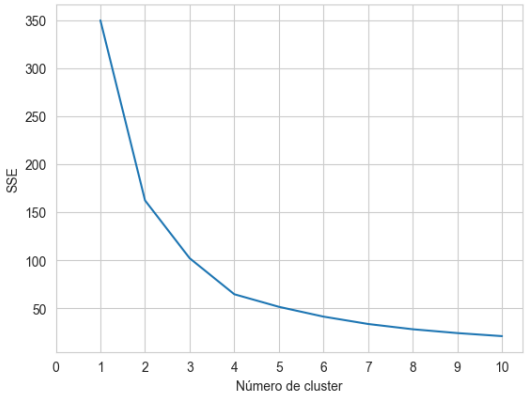

Figura 2. *Gráfica del codo para encontrar número óptimo de clusters.*

Como último paso previo a la aplicación de los algoritmos se buscaron las métricas que posibiliten evaluar todos los algoritmos que se implementarán. Se utilizó el coeficiente de silueta que evalúa cuán bien se ha asignado cada punto a su cluster, comparando la distancia promedio con los puntos dentro de su propio cluster y la distancia promedio con los puntos en los otros clusters. Este varía entre -1 y 1, donde un valor cercano a 1 indica que los puntos están muy cerca de los otros puntos en su propio cluster y lejos de los puntos en los otros clusters, lo cual es deseable.

También se empleó el índice de Davies-Bouldin, el cual evalúa la media de las similitudes entre cada cluster y su cluster más similar, donde la similitud es una medida que compara la distancia entre los clusters y el tamaño de los clusters. Un valor más bajo del índice indica un mejor rendimiento del modelo.

Una vez obtenido el número clusters y explicadas las métricas que se emplearán, se comenzó a probar los distintos algoritmos. El primero que se probó es K-Means, que, según scikit-learn, es un algoritmo que agrupa datos intentando separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique el número de clústeres. Se adapta bien a una gran cantidad de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes-

En el notebook, a partir de la sección 3.1.1. comienza el algoritmo de K-Means. Se define el

número de clusters, necesario de antemano, luego se inicia el modelo con los hiperparámetros que garanticen la reproducibilidad del modelo y la cantidad de veces que se utilizarán distintos puntos como centroides diferentes. Seguido a esto, se utilizan las variables con las que se realiza la agrupación, se las escala y se aplica el método que ajusta estos valores al modelo y ejecuta el modelo.

A continuación, se agrega al conjunto de datos una nueva columna llamada “Clusters” la cual indica a qué cluster pertenece cada registro. Esta columna permite graficar visualmente cada cluster utilizando un gráfico de dispersión, lo que se hace después. En dicho gráfico también se identifican los centroides, que son los puntos centrales de cada grupo.

Finalmente se calculan las métricas mencionadas anteriormente. Este proceso se replicó varias veces con distinto número de clusters para comparar el funcionamiento del modelo.

Para una interpretación que brinde más información sobre los resultados, aparte del gráfico de dispersión de los clusters, se realizó un gráfico de barras el cual contiene los valores promedio de las variables de agrupación por cada cluster.

Para realizar esto se debió filtrar los datos obtenidos por cluster para calcular las medias de las dos variables de agrupación y luego confeccionar un dataset con estos valores para poder graficarlo.

Como último paso también se imprimieron cada uno de los países que componen cada cluster, para identificar características sociales y económicas de cada grupo de países.

Además de K-Means, se utilizó el agrupamiento jerárquico para agrupar países en función de su comportamiento en las variables mencionadas. De acuerdo con, scikit-learn, la agrupación jerárquica es una familia general de algoritmos de agrupación que construyen agrupaciones anidadas fusionándolas o dividiéndolas sucesivamente. Esta jerarquía de conglomerados se representa como un árbol o dendrograma. La raíz del árbol es el único racimo que reúne todas las muestras, siendo las hojas los racimos con una sola muestra.

Los códigos correspondientes a este algoritmo comienzan en la sección 3.1.2. del notebook. Se siguió un procedimiento muy similar al detallado anteriormente para K-Means.

Primero se filtra el dataset por las características de agrupación y se estandarizan los valores. Luego, se aplica el clustering jerárquico en tus datos, y se emplea el método de Ward para calcular las distancias entre los clusters, el cual minimiza la varianza total dentro de los clusters.

Es necesario establecer el número de clusters. Esto se hace para que quede un número de grupos racional, debido a que, si no se indica, este algoritmo podría clasificar los registro hasta se encuentre un registro en cada nodo hoja.

Posteriormente, se agrega una columna que alberga el número de cluster en cada registro. Se calculan los centroides, se realiza el gráfico de dispersión para visualizar cada cluster y se calculan las métricas para distintos valores de cantidad de clusters.

Finalmente se realizó el mismo gráfico de barras que contiene las medias de las variables de agrupación para cada cluster y se imprimieron los países de cada grupo.

El algoritmo DBSCAN también se implementó en este estudio. De acuerdo con scikit-learn, este algoritmo ve los grupos como áreas de alta densidad separadas por áreas de baja densidad. Debido a esta visión bastante genérica, los grupos encontrados por DBSCAN pueden tener cualquier forma, a diferencia de K-Means, que supone que los grupos tienen forma convexa. El componente central del DBSCAN es el concepto de muestras centrales, que son muestras que se encuentran en áreas de alta densidad.

Es particularmente útil para identificar grupos de densidades variables en el espacio de características. En este caso no es necesario indicar con antelación el número óptimo de clusters, pero si necesita otros parámetros, los cuales son epsilon y el número mínimo de muestras.

Epsilon indica la máxima distancia entre dos muestras para que una sea considerada en la vecindad o región densa de la otra, mientras que el número mínimo de muestras hace referencia el número de muestras en una vecindad para que un punto sea considerado como un punto central.

Luego también se empleó el algoritmo HDBSCAN. Según scikit-learn, este puede verse como una extensión de DBSCAN y OPTICS. Este algoritmo alivia el funcionamiento de DBSCAN en grupo de datos donde la densidad no es homogénea, lo hace explorando todas las escalas de densidad posibles mediante la construcción de una representación alternativa del problema de agrupación.

Este algoritmo tampoco necesita el número óptimo de clusters, en su defecto precisa valores de los hiperparámetros correspondientes al tamaño mínimo que un cluster puede tener. Los grupos que tienen menos puntos que este valor se considerarán ruido. Y el otro, es el que indica el número mínimo de vecinos que un punto debe tener para ser considerado un punto central.

Los códigos correspondientes a estos algoritmos se encuentran en la sección 3.1.3. del notebook. Se probaron ambos algoritmos siguiendo la misma metodología de los algoritmos anteriores. Primero se filtró el conjunto de datos por las variables de agrupación y se estandarizaron los valores. Luego se

pasó a configurar los hiperparámetros propios de cada algoritmo.

En el caso de DBSCAN se configuraron epsilon y el número mínimo de muestras para optimizar la calidad de los clusters. Lo mismo se realizó para HDBSCAN con el tamaño del cluster y la cantidad mínima de vecinos para considerar un punto como punto central.

Posteriormente, se crean los modelos con dichos hiperparámetros, se ajusta y entrena el modelo. Se añade la columna “Cluster” que permite realizar el gráfico de dispersión y al final se calculan las métricas.

Se probaron con varias combinaciones de estos valores, pero no se llegó a un resultado útil. Por lo tanto, no se prosiguió con el gráfico de las medias y la impresión de los países.

Finalmente, se empleó la propagación de afinidad. Este algoritmo, según scikit-learn, crea clústeres enviando mensajes entre pares de muestras hasta la convergencia. Luego describe un conjunto de datos utilizando una pequeña cantidad de ejemplos, que se identifican como los más representativos de otras muestras.

Los mensajes enviados entre pares representan la idoneidad de una muestra para ser modelo de la otra, que se actualiza en respuesta a los valores de otros pares. Esta actualización ocurre de manera iterativa hasta la convergencia, momento en el cual se eligen los ejemplos finales y, por lo tanto, se proporciona la agrupación final.

Este algoritmo tampoco necesita que se le indique la cantidad óptima de clusters, ya que la identifica automáticamente. Pero si se le deben configurar dos hiperparámetros. El primero de estos corresponde al suavizado, que es el grado en que se mantiene el valor actual en relación con los valores entrantes. Esto se utiliza con el fin de evitar oscilaciones numéricas al actualizar estos valores (mensajes). El segundo es un número que corresponde al número máximo de iteraciones que realizará el algoritmo hasta entregar un resultado final.

Los códigos correspondientes al este algoritmo se encuentran en el notebook en la sección 3.1.4. Se repite el mismo procedimiento aplicados a los demás algoritmos. Primero se filtra el conjunto de datos por las variables de interés y se escalan los datos.

Luego se inicia el algoritmo con valores correspondientes a los hiperparámetros descritos en el párrafo anterior, se ajusta el modelo y se ejecuta el algoritmo. Se añade la columna “Cluster” al dataset para poder realizar el gráfico de dispersión. Posterior a graficar, se calculan las mismas métricas que los demás algoritmos y se imprime la cantidad de clusters.

Además del mencionado gráfico, se calculan las medias de las variables de agrupación por cluster para poder graficar los valores en un histograma de dos barras. Finalmente se imprimen los países que forman parte de cada grupo para poder realizar un análisis contextualizado.

2.5. Objetivo predictivo

Para poner en práctica distintos algoritmos de predicción de valores al conjunto de datos se aplicó el objetivo planteado en la introducción, el cual busca predecir consumos futuros de energía per cápita anuales a nivel mundial. Este valor puede ser importante para inferir en decisiones de políticas energéticas a nivel país.

Antes de comenzar con el desarrollo del objetivo predictivo en sí, es importante aclarar que en este caso se utilizará el dataset original obtenido de Kaggle [9]. Esto es debido a que el dataset completo posee predicciones de valores que pueden influir en la resolución del objetivo predictivo planteado.

A continuación, se realizará una explicación de los pasos que se siguieron para responder al objetivo. El código empleado se puede ver en la sección 4.2. del notebook.

Una vez aclarado esto, el primer paso que se realizó una exploración de correlaciones entre las variables del conjunto de datos para comprender las interrelaciones y posibles patrones. Para visualizar esto, se utilizó un mapa de calor y un gráfico de barras para visualizar las correlaciones, centrándose especialmente en la variable de interés, "Consumo primario de energía per cápita."

Se creó un conjunto de datos sin las columnas que contienen en nombre del país y el año de los datos, ya que estas no aportan información relevante para el análisis de correlación. Luego, se calculó la matriz de correlación utilizando el método propio de Pandas, la cual se puede ver en la figura 3.

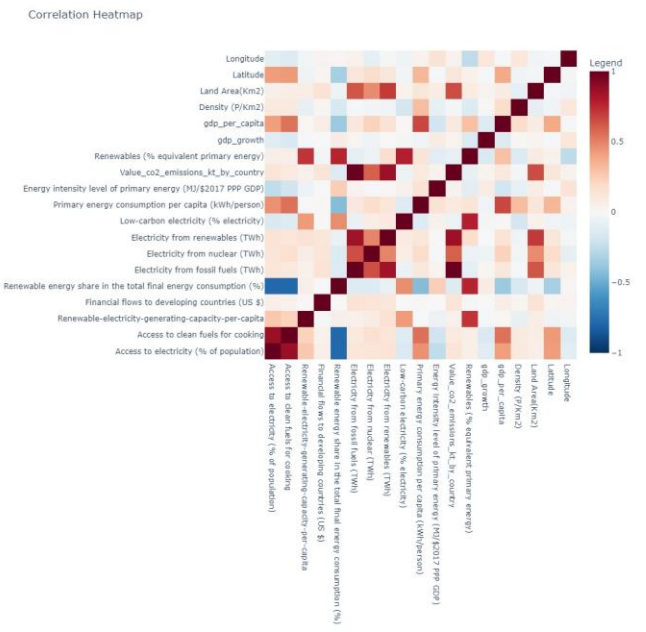

Figura 3. *Matriz de correlación del dataset original.*

Luego, con el objetivo de destacar las relaciones de la variable de interés con las demás, se seleccionó esta variable y se generó un gráfico de barras para visualizar las correlaciones. Este gráfico se visualiza en la figura 4.

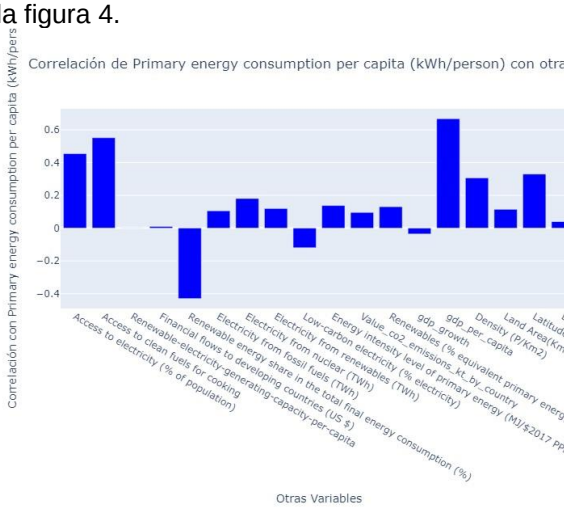

Figura 4. *Correlación de la variable de consumo de electricidad per cápita.*

Este proporciona una visión clara de cómo se relaciona la variable de interés con otras variables del conjunto de datos. Los valores más cercanos a 1 indican una correlación positiva, mientras que los cercanos a -1 indican una correlación negativa. Este análisis preliminar orienta las elecciones de características que utilizará el modelo de regresión.

Como segundo paso se define la variable objetivo (variable a predecir su valor) la que indica el consumo de electricidad consumida por persona y se seleccionan las variables características, que son relevantes para construir el modelo predictivo. Se utilizaron las variables que mayor valor de correlación tenían con la variable objetivo.

Luego, para asegurar la calidad de los datos, se elimina cualquier fila en la que falte el valor del objetivo. Esto garantiza que el modelo se entrene y evalúe solo en instancias completas.

Como tercer paso se procede a dividir el conjunto de datos en dos partes fundamentales: las características (x) y el objetivo (y). Para abordar valores faltantes en las características, se realiza una imputación simple de valores, utilizando la media. Este enfoque ayuda a mantener la integridad de los datos.

A continuación, se dividen los subconjuntos de datos obtenidos anteriormente en conjuntos de entrenamiento y prueba, utilizando el método de la librería scikit-learn que separa conjuntos de datos en dos subconjuntos: uno de entrenamiento y uno de prueba. En este caso, el 60% de los datos se utiliza para entrenar los modelos, y el 40% se reserva para evaluar su rendimiento.

Luego, se realiza un escalado de las características. Esto es esencial para garantizar que todas las variables tengan la misma escala. Se utiliza un escalador estándar para estandarizar las características, lo que ayuda a mejorar la convergencia y el rendimiento de los modelos.

Este proceso es necesario debido a que los algoritmos seleccionados se basan en distancias euclidianas o en umbrales de decisión, lo que los lleva a ser sensibles antes características con escalas distintas.

El paso siguiente consiste en entrenar los modelos de algoritmos que se utilizarán. En este caso se emplearon el algoritmo de regresión lineal, árboles de decisión, bosques aleatorios y aumento del gradiente[4].

En el caso de los árboles de decisión los hiperparámetros que se deben definir son, la profundidad máxima, el número mínimo de muestras requeridas para dividir un nodo y el número mínimo de muestras requeridas para ser consideradas como una hoja.

Para seleccionar los hiperparámetros óptimos para cada caso se utilizó la estrategia de búsqueda en grilla, utilizando el método de la librería scikit-learn. Este enfoque sistemático implica probar diversas combinaciones de valores, seleccionados previamente, para los hiperparámetros, permitiendo así identificar la configuración óptima. Este método selecciona la combinación en base a minimizar el error cuadrático medio negativo del algoritmo.

Los hiperparámetros que se deben configurar en el caso del bosque aleatorio son el número de estimadores o cantidad de árboles del bosque, la profundidad máxima o la longitud de la raíz hasta las hojas y el número mínimo de muestras necesarias para dividir un nodo.

Los hiperparámetros del algoritmo mejora del gradiente que se debe configurar son el número de estimadores o la cantidad de etapas de impulsos que produce el algoritmo, la profundidad máxima y la tasa de aprendizaje o cuánto se ajustan los pesos de cada etapa de impulso.

Finalmente, para el algoritmo de regresión lineal, no es necesario realizar una búsqueda de hiperparámetros. Esto se debe a que los coeficientes se determinan directamente a partir de los datos durante el proceso de entrenamiento. Esto permite ahorrar tiempo computacional y facilitar su implementación.

Una vez con los hiperparámetros optimizados obtenidos en el paso anterior, se procede a la inicialización de los modelos y su entrenamiento utilizando el conjunto de datos de entrenamiento.

Para los casos de bosques aleatorios, arboles de decisión y aumento del gradiente, se fija el hiperparámetro “random_state” en 42. Ese valor hace

referencia a la semilla con la que se inicializa un modelo, y dicho valor garantiza la reproducibilidad del mismo.

Los modelos se entrenan utilizando el conjunto de datos de entrenamiento escalados. Este proceso ajusta los parámetros internos de cada modelo para que se adapten mejor a los datos proporcionados. La regresión lineal, al ser un modelo más simple, no requiere optimización de hiperparámetros y se entrena directamente.

Finalmente se produce el entrenamiento propiamente dicho de los modelos. Se procede a realizar predicciones sobre el conjunto de datos de prueba. Estas predicciones se obtienen para cada modelo y con estas es posible obtener métricas que permiten estimar cual fue el rendimiento de cada uno de los algoritmos, y así elegir el más conveniente para ser aplicado en el conjunto de datos seleccionado.

Se emplearon dos métricas para evaluar y comparar los algoritmos. La primera es el error cuadrático medio. Esta métrica calcula la diferencia al cuadrado entre las predicciones y los valores reales y se promedia. Para obtenerlo se utiliza el método de scikit-learn, que da como resultado el MSE[5] del modelo, y luego se le aplica la raíz cuadrada. Este error proporciona una medida del error en las mismas unidades que la variable de respuesta, lo que resulta útil para interpretar la magnitud de los errores en el contexto del problema específico.

La segunda métrica es el coeficiente de determinación[6] mide la proporción de la variabilidad en la variable dependiente que es predecible a partir de las variables independientes. La ventaja de este valor respecto al RMSE[7] es que no es necesario comparar el valor de la métrica con los valores obtenidos. Un valor cercano a 1 indica que el modelo ajusta bien los datos, mientras que así está cercano a 0 sugiere que el modelo no está capturando la variabilidad de la variable dependiente.

3. Resultados

3.1. Visualización de objetivos descriptivos

En esta sección se presentarán los resultados gráficos de los objetivos descriptivos. En la figura 5 se observan los resultados obtenidos de los primeros tres objetivos descriptivos.

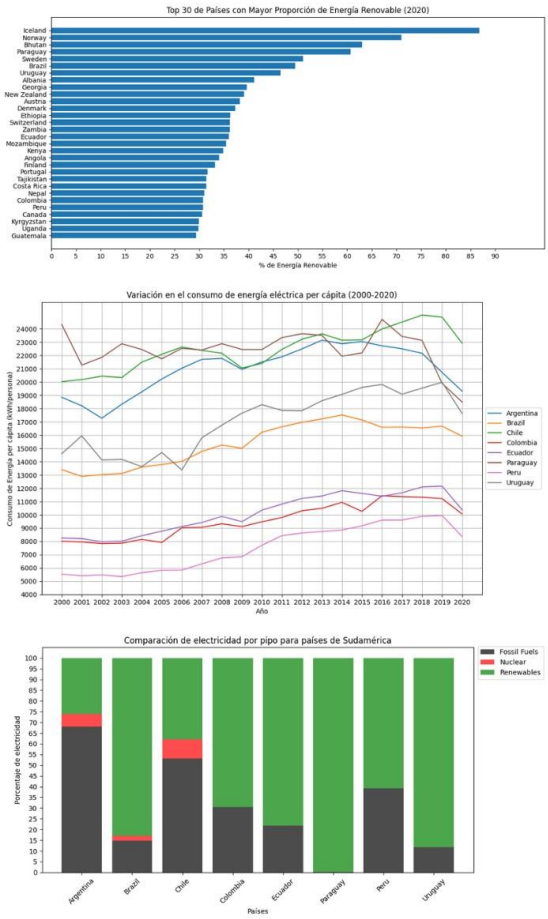

Figura 5. *Gráficos de los primeros tres objetivos descriptivos.*

En la parte superior de la figura se visualiza en histograma que presenta los países con mayor porcentaje de energías renovables consumidas con respecto a la energía total consumida. En el centro de la figura se puede ver un gráfico que presenta la evolución del consumo de energía por persona en los países de Sudamérica. Finalmente, en la parte inferior de la figura se observa un gráfico de barras apiladas donde se presentan los porcentajes de las diversas fuentes de generación de energía eléctrica. Estas fuentes son combustibles fósiles, las diferentes fuentes renovables agrupadas y de origen nuclear.

Luego, en la figura 6 se encuentran los objetivos descriptivos número 4, 5 y 6 respectivamente. En el primer gráfico se puede visualizar como evolucionaron las diversas fuentes para la generación de energía eléctrica en Argentina utilizando un gráfico de áreas apiladas. En el centro de la figura se presenta un gráfico de tendencias lineales sobre la capacidad de generación de energía renovables por persona. Y en la parte inferior de a figura se puede ver un gráfico de múltiples líneas las cuales representan la evolución de energías renovables consumidas de cada país de Sudamérica.

[5] Mean Square Error
[6] También conocido como coeficiente R^2
[7] Root Mean Square Error

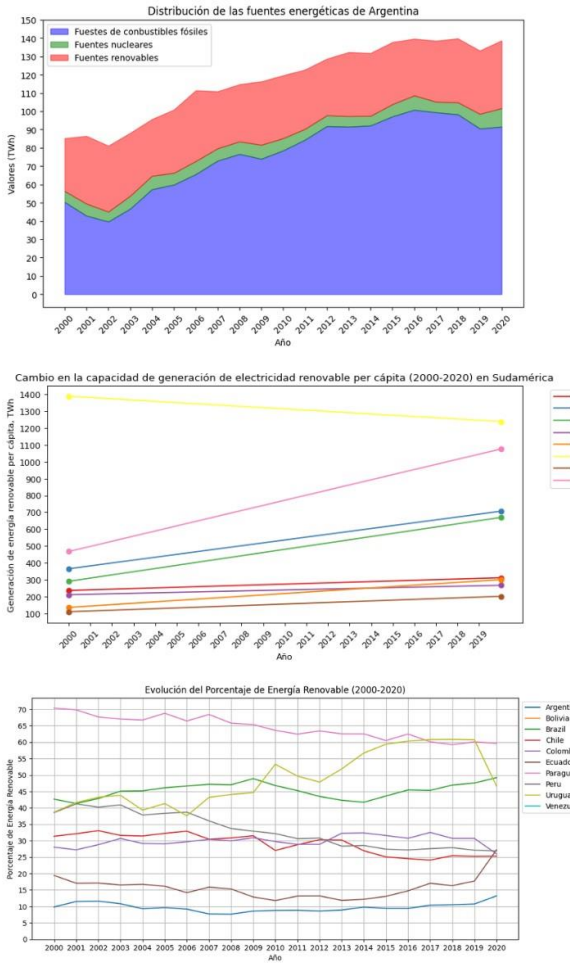

Figura 6. *Gráficos de los objetivos descriptivos número 4, 5 y 6.*

Continuando con la visualización, en la figura 7 se encuentran los objetivos número 7 y 8 respectivamente.

En la parte superior de la figura se grafican los flujos financieros de países desarrollados hacia los demás países para proyectos de energías renovables en función del porcentaje de energía renovable consumida en el año 2019.

En la parte inferior se encuentra un gráfico de dos líneas. La primera línea grafica la evaluación en la capacidad de generación de energía renovable por persona. Mientras que, la segunda línea esboza la evolución de los flujos financieros a lo largo del tiempo.

En la figura 8 se pueden observar dos gráficos de líneas que muestran evoluciones a lo largo del tiempo de variables. En la parte superior se ve la evolución de la variación porcentual año a año del PBI en una línea y luego en otra la misma variación del consumo de energía. En la parte inferior el gráfico muestra la variación del PBI y la variación del porcentaje de energías renovables consumidas.

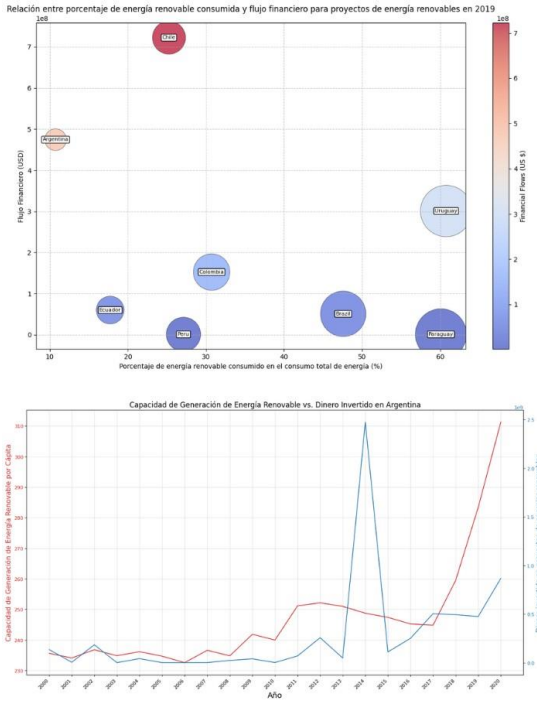

Figura 7. *Gráficos de los objetivos descriptivos número 7 y 8.*

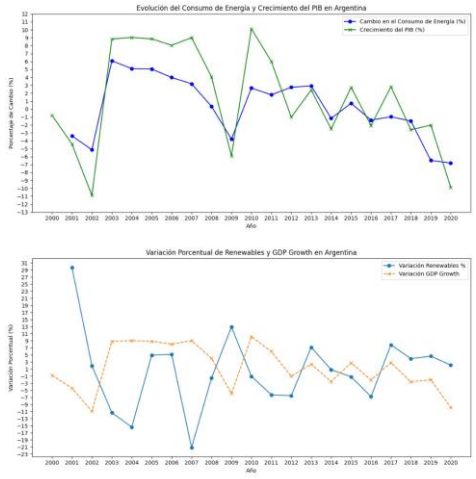

Figura 8. *Gráficos de los objetivos descriptivos número 9 y 10.*

Continuando, en la figura 9, se presentan tres gráficos. El primero, ubicado en la parte superior de la figura, se ordenan los países por emisiones de carbono, en forma de histograma horizontal. El segundo gráfico, ubicado en el centro de la figura, se profundiza en las emisiones de carbono en Sudamérica, en forma de histograma vertical. Y el tercero ubicado en la parte inferior, muestra la evolución de las emisiones a lo largo de los años específicamente en Argentina y Brasil, utilizando un gráfico de líneas.

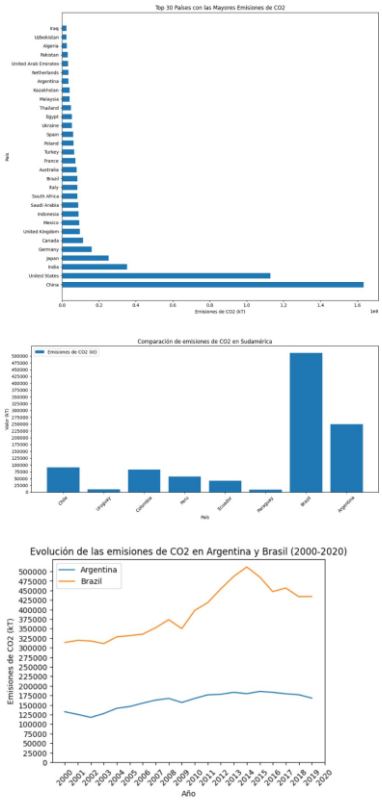

Figura 9. *Gráfico del objetivo descriptivo número 11.*

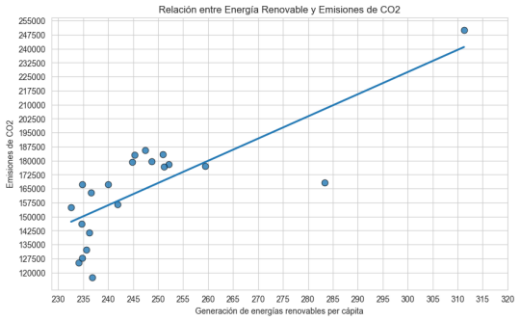

Figura 10. *Gráfico del objetivo descriptivo número 12.*

Para finalizar en la figura 10 se encuentran graficadas las emisiones de CO^2 en función de la generación de energía por persona de Argentina, utilizando un gráfico de dispersión . Además, se traza una línea que representa la tendencia que siguen lo puntos.

3.2. Resultado de objetivo de clustering

Como se mencionó anteriormente se realizaron múltiples pruebas modificando los hiperparámetros de los algoritmos aplicados, para obtener los valores de las métricas.

Se comienza con los resultados que se obtuvieron con el algoritmo de K-Means. En la figura 11 se pueden ver los valores de las métricas, donde se varió el hiperparámetro "k", el cual indica la cantidad de clusters óptimos.

Cantidad de clusters	Puntuación de silueta	Índice de Davies-Bouldin
2	0.484	0.835
3	0.488	0.835
4	0.521	0.683
5	0.457	0.773
6	0.461	0.802
7	0.490	0.702

Figura 11. *Métricas para diversos valores del hiperparámetro k.*

La aplicación de K-Means con cuatro clusters resultó tener el mejor desempeño, con una puntuación de silueta de 0,521, indicando una razonable asignación de puntos a los clusters, y un valor del índice de Davies-Bouldin de 0,6725, sugiriendo una buena separación entre los clusters.

En la figura 12 se ve el gráfico de dispersión confeccionado utilizando este algoritmo.

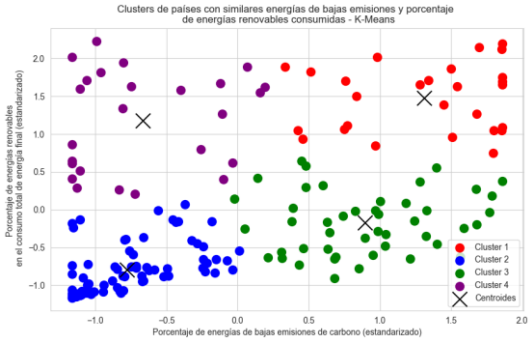

Figura 12. *Clusters obtenidos con el algoritmo K-Means.*

Finalmente, en la figura 13 se puede ver los valores promedios de las variables de agrupación, electricidad baja en emisiones de CO^2. Esto se utilizó para realizar conclusiones que se encuentran al final de la sección 3.1.1.

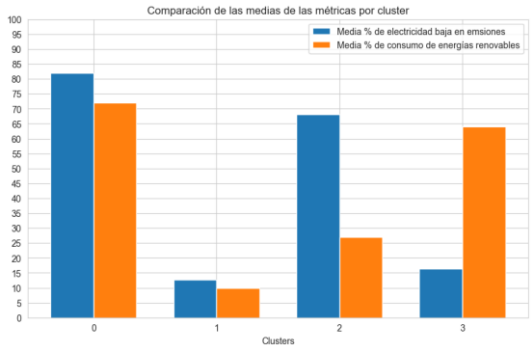

Figura 13. *Valores promedio de las variables de agrupación de cada cluster con el algoritmo K-Means.*

Continuando con los resultados de clustering jerárquico, en la figura 14 se pueden visualizar las métricas de las pruebas efectuadas, donde se varió el hiperparámetro número de clusters.

Cantidad de clusters	Puntuación de silueta	Índice de Davies-Bouldin
2	0.442	0.833
3	0.464	0.976
4	0.481	0.757
5	0.497	0.828
6	0.421	0.844
7	0.434	0.746

Figura 14. *Métricas para diversos valores del hiperparámetro k.*

El mejor resultado se obtuvo con cuatro clusters, donde obtuvo una puntuación de silueta de 0,481 y un índice de Davies-Bouldin de 0,757.

En la figura 15 se puede ver gráficamente los clusters obtenidos utilizando el algoritmo de clustering jerárquico.

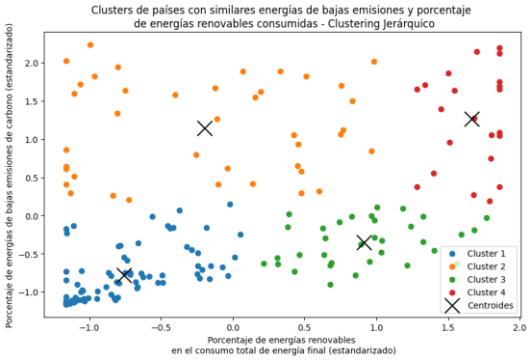

Figura 15. *Clusters obtenidos con el algoritmo de clustering jerárquico.*

Finalmente se graficaron las medias de las variables de los distintos clusters, lo que se muestra en la figura 16. Además, se obtuvieron los distintos países correspondientes a cada cluster.

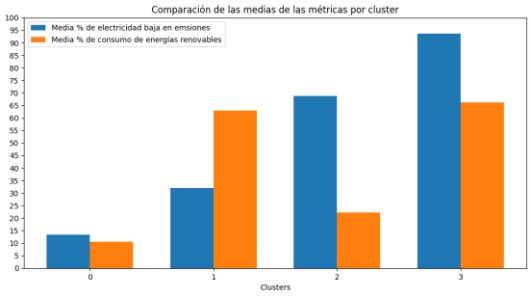

Figura 16. *Valores promedio de las variables de agrupación de cada cluster con el algoritmo de clustering jerárquico.*

Continuando con el análisis de los resultados, se presentan los obtenidos para los algoritmos DBSCAN y HDBSCAN en la figura 17.

Como ya se anticipó en la sección de metodología, estos algoritmos presentaron un resultado bastante alejado a lo esperado, por lo cual no se prosiguió a graficar las distintas medias de las variables de agrupación para cada cluster. Esto se explayará en la sección de discusión.

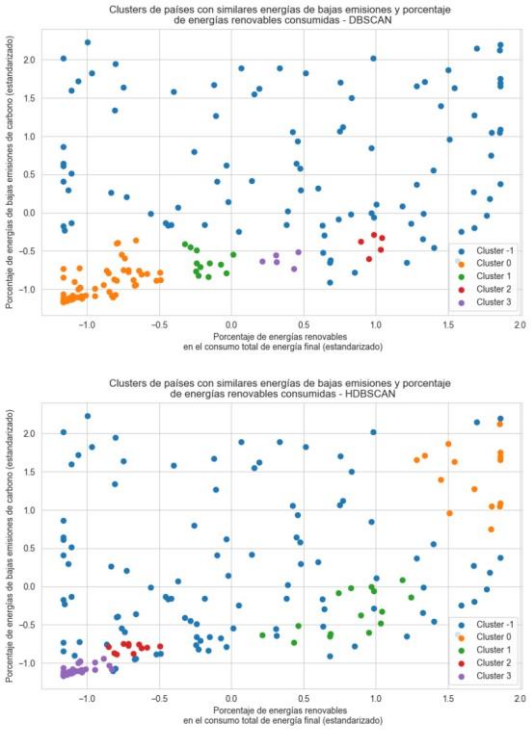

Figura 17. *Clusters obtenidos con los algoritmos DBSCAN y HDBSCAN respectivamente.*

Finalmente, se prosiguen con los resultados obtenidos para el algoritmo de propagación por afinidad. En la figura 18 se presentan los valores de las métricas para distintos valores de los hiperparámetros de amortiguación o “damping” y el número máximo de iteraciones.

Cantidad de clusters	Damping	Iteraciones máximas	Puntuación de silueta	Índice de Davies-Bouldin
11	0.5	200	0.429	0.743
11	0.6	200	0.426	0.759
11	0.7	200	0.432	0.745
11	0.8	200	0.432	0.745
11	0.9	200	0.432	0.745
11	0.5	500	0.429	0.743
11	0.6	500	0.426	0.759
11	0.7	500	0.432	0.745
11	0.8	500	0.432	0.745
11	0.9	500	0.432	0.745

Figura 18. *Métricas para diversos valores del hiperparámetro damping y el número máximo de iteraciones.*

El valor óptimo del coeficiente de silueta es 0,432 y el correspondiente al índice de Davies-Bouldin es 0,745. Estos valores se repiten en varias combinaciones de hiperparámetros. Se tomó para realizar el gráfico de dispersión los valores de 0,7 en damping y 200 iteraciones máximas, para ahorrar costo computacional.

El gráfico obtenido con dichos hiperparámetros se puede visualizar en la figura 19.

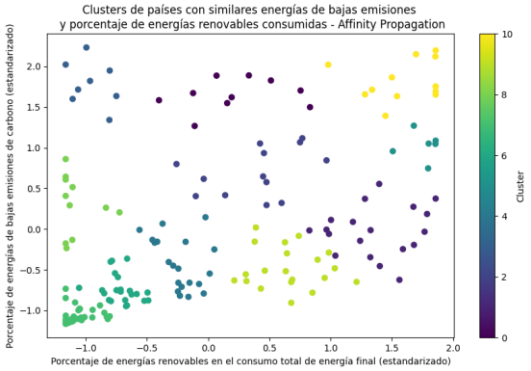

Figura 19. *Clusters obtenidos con el algoritmo de propagación por afinidad.*

Finalmente se realizó el mismo gráfico de dos barras para graficar los valores promedios de las variables de agrupación para cada cluster. Esto se puede observar en la figura 20. Y luego se obtuvieron los países pertenecientes a cada clusters.

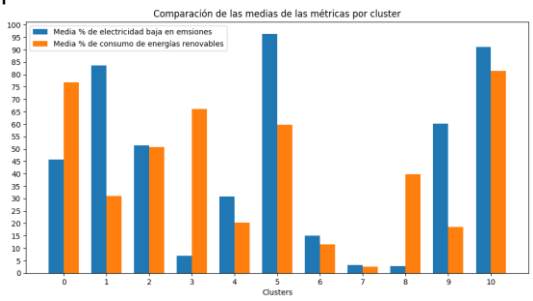

Figura 20. Valores promedio de las variables de agrupación de cada cluster con el algoritmo de propagación por afinidad.

3.3. Resultado de objetivo predictivo

Los primeros resultados que se obtuvieron fueron los valores óptimos de los hiperparámetros de cada modelo, obtenidos mediante el método de búsqueda de rejilla. Estos resultados se encuentran plasmados en tabla 1.

Modelo	Profundidad máxima	Hiperparámetro 2	Hiperparámetro 3
Bosque aleatorio	20	Número mínimo de división en un nodo: 5	Número de estimadores: 100
Aumento del gradiente	5	Tasa de aprendizaje: 0,01	Número de estimadores: 100
Árboles de decisión	30	Número mínimo de nodos hoja: 4	Número mínimo de división en un nodo: 5

Tabla 1. *Tabla de hiperparámetros óptimos para los algoritmos de predicción.*

Una vez obtenidos los valores de los hiperparámetros se entrenaron y realizaron predicciones con estos valores, para luego evaluarlos.

Como ya se mencionó, las métricas utilizadas para evaluar y comparar los algoritmos son RSME y R^2. Esto se puede ver en la figura 21.

Modelo	Predicciones	RMSE	R^2
Random Forest	1674.452982	11647.807277	0.891524
Linear Regression	2069.914435	26295.410479	0.447152
Gradient Boosting	1759.143408	12891.439128	0.867123
Decision Tree	1627.245775	14759.919078	0.825814

Figura 21. *Métricas y predicciones de los algoritmos.*

Las métricas dan como mejor algoritmo al bosque aleatorio seguido del aumento del gradiente, luego los árboles de decisión y lejos en cuanto a valores la regresión lineal.

Para corroborar esto, se procedió a probar estos algoritmos con instancias reales. Se escogió un registro al azar, se le extrajeron los valores de las características y se realizó la predicción para compararla con el valor real. Los resultados se pueden ver en la figura 22.

```
Esperado: 18210.12
random forest:  [18466.16527222]
gradient boosting:  [17942.2719045]
regresion lineal:  [21950.37974467]
decision tree:  [20365.16675]
```

Figura 22. *Predicciones en base a un registro existente utilizando los cuatro algoritmos.*

4. Conclusiones parciales

4.1 Interpretación del EDA

Como se trató en la sección de metodología, el análisis exploratorio de datos se centró en el principal problema que tenía el conjunto de datos, el cual era la gran cantidad de valores faltantes.

Se plantearon cuatro posibles soluciones las cuales eran eliminar columnas o filas con valores faltantes, imputar datos mediante algún algoritmo de predicción o mantener los datos como están.

Se optó por realizar una imputación de datos debido a que lo que los objetivos que se plantearon sobre el dataset requerían la totalidad de los datos. Además, es necesario para comprender los datos completamente y tener la posibilidad de generar gráficos, reportes y conclusiones que permitan genera decisiones en base a las tendencias que generan estos datos.

Luego el EDA también incluyó en algunas partes renombre de variables para que se facilite el procesamiento de los datos y, además, no se dificulte el entendimiento de los datos.

4.2 Interpretación de los gráficos de los objetivos descriptivos

Los distintos objetivos brindaron una visión general en diversos aspectos del conjunto de datos, la cual permitió realizar análisis estadísticos, que luego contrastados con información de los distintos países posibilitó arribar a conclusiones sobre los aspectos analizados.

Como se mencionó en la introducción, se utilizaron diversos gráficos, como histogramas, gráficos de líneas, gráficos de burbujas, entre otros. El propósito

principal de estos es mostrar de manera gráfica las variables seleccionadas para interpretar de manera sencilla los datos.

Además, la mayoría de los objetivos planteados se focalizaron en datos de Argentina o de los países pertenecientes a Sudamérica, debido a que estos comparten características sociales, culturales y económicas.

Todas las conclusiones correspondientes a los objetivos descriptivos se encuentran debajo de cada gráfico en el notebook, a partir de la sección 2.

4.3 Interpretación de las salidas del algoritmo de clustering

Los tres algoritmos que entregaron resultados positivos en base al objetico de agrupación planteado fueron K-Means, clustering jerárquico y propagación por afinidad, los dos restantes, DBSCAN y HDBSCAN, no produjeron resultados satisfactorios.

La principal razón de los resultados fallidos de DEBSCAN y HDBSCAN se debe a que estos algoritmos utilizan la noción de regiones densas para clasificar a los conjuntos de datos. En el caso del análisis planteado, el subconjunto de datos seleccionado para la agrupación no cumple con esta característica.

De hecho, dicho subconjunto tiene una distribución de datos, después de haberlos escalados, bastante dispersa y poco densa. Por lo tanto, estos algoritmos no son capaces de identificar grupos de datos. En su defecto, categorizan a la mayoría de los registros como ruido, etiquetados como “Cluster -1”.

Luego continuando con los otros tres algoritmos empleados, cada uno de estos, tiene sus propias fortalezas y debilidades, y la elección del método depende de las características específicas de los datos. En este estudio, se utilizó una combinación de estos métodos para obtener una comprensión más completa de los patrones en los datos.

La principal diferencia que se notó es que, tanto K-Means y clustering jerárquico, necesitan que se le indique por adelantado el número de clusters óptimo, lo cual, en ocasiones no puede ser posible. Además, este número resultó ser 4, por lo que generó 4 grupos, donde las características de agrupación, en promedio, adoptan solamente dos valores posibles: alto o bajo. Esto se puede ver reflejado en la figura 13 y 16, respectivamente.

Esto puede llegar a ser una limitante si se buscaba realizar un análisis más profundo del tema. Pero aún así estos algoritmos obtuvieron el mejor desempeño, en cuanto métricas, entre todos los probados, siendo el mejor K-Means y luego en segundo puesto clustering jerárquico.

K-Means produjo clusters bien definidos y separados con una distribución equilibrada y un agrupamiento coherente de los países según las características de interés. Esto que indica que es una metodología robusta y adecuada para este tipo de datos.

Luego, clustering jerárquico, a pesar de que los clusters identificados eran distinguibles, la calidad del clustering no fue tan alta como en K-Means, lo que se reflejó en las métricas obtenidas. Estos presentaron una menor cohesión y una mayor superposición, lo que sugiere que este método puede no ser tan efectivo para el conjunto de datos en estudio comparado con K-Means.

En contraparte, el algoritmo de propagación por afinidad, que no necesita el valor óptimo de clusters, resuelve este problema automáticamente. Esta resolución automática devolvió 11 clusters como los óptimos.

Esto permite que las medias de las variables de agrupación en cada cluster puedan asumir más de los valores altos y bajos, pudiendo ser intermedios. Esto se puede visualizar en la figura 20. Así se puede realizar un análisis mucho más detallado y profundo de las variables estudiadas. El tener una mayor agrupación conlleva un resultado un poco inferior en cuanto a métricas.

Por lo tanto, la conclusión final depende del problema que se busque resolver. En este caso particular, si se desea contar con una mayor cantidad de información y relegando un poco de calidad de agrupación se debe elegir la propagación por afinidad. Por el contrario, si con cuatro grupos que presentan información básica es suficiente entonces se puede optar por elegir K-Means.

4.4 Interpretación de las salidas del algoritmo de regresión

Para resolver el objetivo predictivo planteado se aplicaron cuatro algoritmos, árboles de decisión, bosques aleatorios, aumento del gradiente y regresión lineal. Todos brindaron resultados satisfactorios, como se puede ver en la figura 21, la cual contiene los resultados de las métricas utilizadas. Los mejores algoritmos resultaron ser en primer lugar el bosque aleatorio y luego el aumento del gradiente.

Para probar realmente el desempeño de estos algoritmos se tomó un registro al azar del conjunto de datos que se encontrara completo, y luego este fue pasado a los algoritmos para que realicen la predicción de la variable objetivo en base a las variables características.

Esto se puede observar en la figura 22 donde se encuentra el valor esperado o real de la variable objetivo y los valores de las predicciones de los distintos algoritmos.

Algoritmos	B.A.	A.G.	A.D.	R.L.
Valor real	18210,12	18210,12	18210,12	18210,12
Valor predicho	18466,16	17942,27	20365,16	21950,37
Diferencia	256,04	267,85	2155,04	3740,25
Diferencia %	+1,4	-1,47	+11,83	+20,54

Tabla 2. Tabla de las diferencias entre las predicciones de los distintos algoritmos.

En la tabla 2 se pueden ver las predicciones en comparación con el valor real del registro seleccionado aleatoriamente. La primera columna corresponde al algoritmo bosques aleatorios, la segunda al aumento del gradiente, la tercera árboles de decisión y la última es la regresión lineal.

Esto refleja que las métricas son acertadas y los mejores algoritmos, para este conjunto de datos, son bosque aleatorio y el aumento del gradiente.

5. Conclusiones generales del estudio

Este exhaustivo estudio integra a la perfección rigurosos análisis exploratorios de datos y técnicas de regresión y agrupación para obtener información matizada sobre el panorama mundial de la producción y el consumo de energía y energía sostenible.

Empleando un enfoque metodológico, se han descubierto patrones y tendencias, evaluando la eficacia de diversos algoritmos de aprendizaje automático aplicados a nuestro conjunto de datos. Más allá de proporcionar una instantánea del estado actual de la sostenibilidad energética, nuestra metodología nos permite anticipar y comprender posibles desarrollos futuros.

El análisis exploratorio de datos brindó la posibilidad de profundizar en los entresijos del conjunto de datos, desvelando patrones y relaciones inherentes.

El proceso de clustering surgió como un componente fundamental del análisis. A través de la agrupación, se ha podido delinear grupos significativos dentro del conjunto de datos, arrojando luz sobre los países que muestran características similares en términos de electricidad baja en carbono y consumo de energías renovables. Estos datos son muy valiosos para los responsables políticos y las partes interesadas que buscan estrategias informadas para fomentar prácticas energéticas sostenibles en todo el mundo.

El análisis de regresión, por su parte, proporcionó modelos fiables para predecir el consumo de energía, ofreciendo una valiosa herramienta para los responsables de la toma de decisiones en la búsqueda de prácticas energéticas sostenibles.

Finalmente, este estudio subraya la importancia fundamental de la analítica avanzada para la toma de decisiones y la elaboración de políticas bien fundadas en materia de energía sostenible y cambio climático. Más allá de las contribuciones individuales de EDA, agrupación y regresión, la sinergia de estas metodologías proporciona una comprensión sumamente profunda de variables en el ámbito de la sostenibilidad energética.

Este trabajo no sólo sirve como testimonio del poder de los enfoques basados en datos, sino que también establece un marco para la investigación y la formulación de políticas futuras en la búsqueda de un panorama energético mundial más sostenible en el largo plazo.

6. Referencias

[1] P. Chakri, P. Saurabh, Lakshay y K. G. Sanjeeb, «An exploratory data analysis approach for analyzing financial accounting,» *Decision Analytics Journal,* 29 Marzo 2023.

[2] I.-V. Gerardo, R.-M. María Soledad, B.-F. Mariana y O. Gustavo, «Predicting Open Education Competency Level: A Machine approach,» *Heliyon,,* 29 Septiembre 2023.

[3] G. João, L. João, G. Tiago y S. Manuel Filipe, «Identifying Diabetic Patient Profile Through Machine Learning- Based Clustering Analysis,» *Procedia Computer Cience* , 17 Marzo 2023.

[4] NumFOCUS, Inc., «Pandas,» 2023. [En línea]. Available: https://pandas.pydata.org/. [Último acceso: 3 Noviembre 2023].

[5] YData, «YData Profiling,» [En línea]. Available: https://docs.profiling.ydata.ai/4.6/. [Último acceso: 2023 Noviembre 3].

[6] Matplotlib development team, «Matplotlib documentation,» 2023. [En línea]. Available: https://matplotlib.org/devdocs/index.html. [Último acceso: 3 Noviembre 2023].

[7] M. L. Waskom, «seaborn: statistical data visualization,» *Journal of Open Source Software,* vol. 6, nº 60, p. 3021, 2021.

[8] scikit-learn team, «Scikit-learn Machine Learning in Python,» Octubre 2023. [En línea]. Available: https://scikit-learn.org/stable/index.html. [Último acceso: Noviembre 2023].

[9] A. Tanwar, «Global Data on Sustainable Energy (2000-2020),» Agosto 2023. [En línea]. Available: https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy. [Último acceso: Septiembre 2023].

[10] R. Jafari, Hands-On Data Preprocessing in Python, Birmingham: Packt, 2022.

[11] M. E. Fenner, Machine Learning with Python for Everyone, Pearson Addison-Wesley, 2020.

[12] N. George, Practical Data Science With Python, Birmingham: Packt, 2021.

[13] R. Sclerandi, «GitHub,» 14 Diciembre 2023. [En línea]. Available: https://github.com/RamiroSclerandi/ProyectoAnalisisEnergetico.