Facultad de Ciencias Económicas y Empresariales

Análisis predictivo multidisciplinar del precio del hidrógeno verde

Autor: Pablo Martínez Galindo

Director: Isabel Catalina Figuerola Ferretti

Madrid | Junio de 2024

Resumen

Este trabajo de fin de grado (TFG) está centrado en el análisis cualitativo y cuantitativo del hidrógeno verde, su precio de referencia en el mercado, y el estudio de la relación entre este y el de otras materias primas de referencia a nivel mundial: el petróleo americano (*US crude oil WTI*), el petróleo crudo (*Brent crude ICE*), y el gas natural TTF (*Dutch TTF gas*).

La información cualitativa ha sido extraída de artículos académicos y publicaciones de referentes en la industria energética, y los datos cuantitativos han sido tomados de fuentes de análisis de datos y precios globalmente utilizadas y reconocidas. En primer lugar, se realiza un análisis cualitativo llevando a cabo un estudio de cada una de las materias primas, analizando cuestiones como su producción, influencia en precios de materiales sustitutivos, volatilidad, y evolución de precios históricos, entre otros. En segundo lugar, se analiza cuantitativamente cada una de las materias primas mencionadas, estudiando el comportamiento de su precio de referencia y evolución histórica. Por último, se lleva a cabo la parte de análisis y modelaje de precios, con un estudio de covarianza y correlación mediante diferentes métodos, y se observan las similitudes entre las materias primas. Con ello se crea un modelo de regresión lineal predictor de precio de futuros del hidrógeno verde en base a los de la materia prima más parecida, y un algoritmo de KNN predictor del precio histórico del hidrógeno verde en base al precio histórico de todas las materias primas de referencia.

El objetivo fundamental es dar luz al papel del hidrógeno verde como materia prima de referencia en el mercado energético mundial, y avanzar en la determinación de factores que afectan a su precio y mercado. Este trabajo crea una serie de modelos de predicción de precios del hidrógeno verde y aporta novedades significativas que resaltan la importancia del mismo como materia prima de referencia.

Los modelos generados han permitido ver que se puede estimar de manera efectiva el precio histórico y futuro de una materia prima en base a otras. La repercusión a nivel global de poder predecir correctamente el precio del hidrógeno verde es inmensa, ya que la industria en la que se ha desarrollado el estudio es uno de los principales motores de la economía y el desarrollo mundial, y cualquier factor que provoque un cambio o acontecimiento relevante en ella afectará directa o indirectamente a la mayoría de las industrias globales.

Palabras clave: hidrógeno verde, petróleo, gas natural, regresión lineal, algoritmo KNN, análisis cuantitativo, modelo predictivo de precios, materias primas, análisis de datos.

Abstract

This Final Degree Project (TFG) is focused on the qualitative and quantitative analysis of green hydrogen, its reference price in the market, and the study of the relationship between it and other reference raw materials at a global level: crude oil, American oil, and TTF natural gas (European). This Final Degree Project (TFG) is focused on the qualitative and quantitative analysis of green hydrogen, its reference price in the market, and the study of the relationship between this and that of other reference raw materials worldwide: American oil (US crude oil WTI), crude oil (Brent crude ICE), and natural gas TTF (Dutch TTF gas).

The qualitative information has been extracted from academic articles and energy industry benchmark publications, and the quantitative data has been taken from globally used and recognised data and price analysis sources. Firstly, a qualitative analysis is carried out by conducting a study of each of the raw materials, analysing issues such as their production, influence on prices of substitute materials, volatility, and evolution of historical prices, among others. Secondly, each of the raw materials is analysed quantitatively, studying the behaviour of its reference price and historical evolution. Finally, the price analysis and modelling part is carried out, with a covariance and correlation study using different methods, and the similarities between the commodities are observed. With this, a linear regression model is created to predict the green hydrogen futures price based on those of the most similar raw material, and a KNN algorithm to predict the historical price of green hydrogen based on the historical price of all the reference raw materials.

The key objective is to shed light on the role of green hydrogen as a reference raw material in the world energy market, and to make progress in determining the factors that affect its price and market. This work creates a series of models for predicting green hydrogen prices and provides significant novelties that highlight the importance of green hydrogen as a reference raw material.

The models generated have shown that the historical and future price of a raw material can be effectively estimated based on other raw materials. The global repercussion of being able to correctly predict the price of green hydrogen is immense, as the industry in which the study has been developed is one of the main drivers of the world economy and development, and any factor that causes a change or relevant event in it will directly or indirectly affect most global industries.

Key words: green hydrogen, oil, natural gas, linear regression, KNN algorithm, quantitative analysis, predictive pricing model, commodities, data analysis.

1. Introducción

1.1. Objetivo y novedades

El objetivo último de este trabajo de fin de grado es conseguir estimar de manera precisa los precios históricos y futuros del hidrógeno verde en base a la valoración de mercado de materias primas relacionadas con él. Se busca también, como se menciona en el resumen, estudiar de manera exhaustiva esta materia prima y evaluar su importancia y potencial como elemento de referencia en el mercado energético mundial.

Se investigarán y analizarán los diferentes factores económicos, regulatorios y tecnológicos que afectan a su coste de producción y comercialización, así como las barreras que entorpecen la adopción masiva de esta materia prima y las oportunidades que facilitan su posicionamiento y consolidación en el mercado de la energía. Se proporcionarán finalmente diferentes recomendaciones y directrices enfocadas en futuros estudios, investigaciones y proyectos que fomenten la adopción de esta materia.

Este trabajo de fin de grado aporta novedades significativas que resaltan la importancia del hidrógeno verde como materia prima y herramienta de referencia del mercado energético mundial. No sólo se identifican obstáculos y oportunidades para su integración, sino que se proponen soluciones específicas muy variadas mediante un enfoque proactivo para sobrepasar las barreras existentes y consolidar las ventajas económicas a nivel global que ello supondría. Se utiliza también un enfoque novedoso basado en la comparación con diferentes materias primas de referencia, estudiando cada una de ellas, ofreciendo una perspectiva más clara y basada sobre el impacto económico-financiero del hidrógeno verde.

1.2. Historia y actualidad del hidrógeno verde

El hidrógeno verde o renovable es un elemento químico producido de forma sostenible utilizado como combustible en gran cantidad de industrias. El proceso a través del cual se obtiene el hidrógeno se conoce como electrólisis del agua, y para que este sea considerado verde, el proceso de descomposición del agua en oxígeno e hidrógeno debe ser llevado a cabo mediante el uso de fuentes renovables, como la energía solar, la eólica o la hidroeléctrica, fundamentalmente, evitando así la generación de gases de efecto invernadero. El hidrógeno obtenido es utilizado como fuente de energía en todo tipo de ámbitos, y desde el comienzo de su producción en la década de 1990, se ha popularizado. Las cuestiones que han provocado el éxito de dicho elemento han sido, entre otras, su sostenibilidad, ya que se genera a través de fuentes de energía renovable, evitando la generación de emisiones y la consiguiente contaminación; un segundo factor clave

para su éxito es la facilidad de almacenamiento, haciendo posible la recolección y guardado de energía renovable para su posterior uso; la tercera razón ha sido su aporte a la diversificación energética, evitando la dependencia de otras fuentes como los combustibles fósiles; y finalmente, también han sido elementos importantes del éxito su factor innovador, el hecho de que permita una reducción de costes y, por supuesto, la facilidad de producción derivada de décadas de investigación en la materia.

La primera planta generadora de hidrógeno verde entró en funcionamiento durante la década de 1990 en el sur de Alemania, se trataba de un laboratorio experimental que comenzó a separar las partículas del agua mediante un proceso de electrólisis completamente renovable. Durante dicha década comienzan a inaugurarse las primeras plantas de hidrógeno renovable en España, que se convirtió en uno de los países pioneros a nivel mundial en esta etapa de transición energética que marcó un antes y un después en este ámbito. Prueba de ello es que la mayor planta productora de hidrógeno verde industrial a nivel mundial se encuentra en Puertollano (Ciudad Real, España). En la actualidad, numerosos países de la Unión Europea contribuyen al desarrollo y la transición energética a través de la construcción de plantas de producción de hidrógeno verde (entre otras fuentes de energía renovables) y de centros de investigación y desarrollo, siendo para muchos el foco catalizador fundamental para la evolución energética mundial (Iberdrola, 2022).

El hidrógeno verde ha ganado popularidad de manera sorprendentemente rápida por sus cualidades y capacidades. En primer lugar, el hidrógeno verde es un vector energético con métodos de almacenaje y posterior generación de energía respetuosos con el medio ambiente: uno de los hitos fundamentales del hidrógeno radica en la capacidad de almacenar energía renovable transformada en esta materia prima, la mayor limitación de esta energía desde sus comienzos; el hecho de lograr almacenar energía verde de manera indefinida ha supuesto un hito que abre un amplio campo de aplicaciones y oportunidades. En segundo lugar, el desarrollo del hidrógeno verde tiene potencial de generar beneficios a nivel económico y técnico, limitando el monopolio de los grandes productores de fuentes de energía tradicionales, punto de controversia a lo largo de la historia. La producción de hidrógeno verde permite a muchos territorios no sólo disponer de otra fuente de energía útil, aumentando su poder frente a los proveedores existentes, sino tener el potencial de convertirse (dependiendo de las características de cada región) en un productor autónomo. En tercer lugar, la producción de hidrógeno verde impulsa la cooperación entre países no sólo a nivel económico, sino a nivel técnico, cultural y científico; ha fomentado y se prevé que continuará fomentando el apoyo entre estados productores para la descarbonización progresiva de la energía, permitiendo la creación de sinergias, relaciones político-económicas y organizaciones entre países, potencialmente beneficiosas en cualquier otro ámbito a medio y largo plazo (Antoranz, 2022).

Las aplicaciones del hidrógeno verde son amplias debido a la versatilidad del elemento en cuestión: el hidrógeno es una de las fuentes de energía más versátiles gracias a su facilidad de producción, transporte, almacenaje y transformación. Las aplicaciones más populares del hidrógeno verde incluyen el transporte, para todo tipo de modalidades, desde vehículos pesados como camiones y trenes, hasta los vehículos puramente propulsados por hidrógeno; la industria, incluyendo todo tipo de procesos de fabricación, producción y desarrollo; y la residencial, incluyendo sistemas de calefacción y refrigerado.

En definitiva, el hidrógeno verde es una fuente novedosa con un enorme potencial de desarrollo gracias a sus cualidades, su composición, su generación, su potencial y sus características. La relativa novedad de la materia prima explica la limitada cantidad de estudios basados en series temporales sobre el esta, que permitirían observar su comportamiento a lo largo del tiempo cualitativa y cuantitativamente, observando tendencias y pudiendo predecir más fácilmente su comportamiento. La mayoría de los estudios disponibles se encuentran dentro del ámbito de la ingeniería, centrándose fundamentalmente en aspectos técnicos y de eficiencia.

1.3. Elementos comparativos

Para llevar a cabo el estudio y predicciones del precio futuro del hidrógeno verde se utilizan diferentes materias primas de referencia en el sector de la energía, de las que existe gran cantidad de información histórica, sobre todo en términos de precio e instrumentos financieros relacionados que pueden influir directa e indirectamente en su valoración. Se toman como materias primas comparativas el petróleo americano (*US crude oil West Texas International spot "WTI"*), el petróleo crudo (*Brent crude Intercontinental Exchange spot "ICE"*), y el gas natural TTF (*Dutch Title Transfer Facility gas spot "TTF"*, cotizado en el ICE)

- Petróleo americano (*US crude oil*): el petróleo americano es una clase de crudo proveniente de yacimientos sitos en Estados Unidos. Este crudo es extraído de ubicaciones marinas y ubicaciones en tierra firme alrededor de toda su geografía. Este componente es un elemento clave en la industria petrolera mundial, ya que, dada su versatilidad y abundancia histórica, ha sido y es utilizado para la producción de todo tipo de bienes, incluyendo la gasolina, el diésel y muchos otros productos petroquímicos. Tiene una gran relevancia dentro del mercado energético mundial, ya que es una de las principales referencias utilizadas para la fijación de precios de numerosos bienes derivados e instrumentos financieros, se considera un indicador clave de la salud económico-financiera americana y de la estabilidad geopolítica de la región. Esta materia

prima tiene una fuerte influencia en el precio del hidrógeno verde por el impacto directo e indirecto en los costes de producción y distribución de este. Es por ello por lo que los cambios en el precio del petróleo inciden en la valoración y coste del hidrógeno verde. La volatilidad del petróleo, especialmente del americano es media-baja en comparación con el resto de materias primas utilizadas por razones como la diversidad de suministro (hay numerosas regiones productoras de petróleo, reduciendo la dependencia directa de una fuente determinada, estabilizando así el precio del activo), la facilidad relativa de almacenamiento y transporte de larga distancia (amplia red de infraestructura existente a nivel mundial facilitando dichas actividades, equilibrando los niveles de oferta y demanda, fomentando así la estabilidad del precio), y el control por parte de autoridades (al ser un activo tan sumamente reconocido y utilizado, hay una fuerte intervención de gobiernos, que llevan a cabo políticas de estabilización de precio para evitar desajustes de oferta y demanda). En la sección 2.2 del trabajo se ahonda más en detalle en cuestiones de volatilidad, tratando diferentes causas y reacciones del precio del activo.

- Petróleo crudo (*Brent crude*): el brent crude es una clase de petróleo obtenido de reservas sitas en la cuenca del Mar del Norte. Es uno de los recursos con mayor volumen de negociación a nivel mundial a causa de su calidad (bajo contenido en azufre) y su versatilidad. Es también un clave indicador clave en el mercado de la oferta y la demanda mundial de materias primas: se utiliza como precio de referencia para gran cantidad de elementos, incluyendo la mayor parte del petróleo negociado a nivel mundial, y es también un indicador fiable de la salud del sector petrolífero global. El precio de este activo influye también en el precio del hidrógeno verde por las mismas razones que el petróleo americano, a causa del fuerte impacto en los precios de producción y distribución de este. Al igual que el petróleo americano, el crudo tiene una volatilidad media-baja por las mismas razones: la diversidad de suministro, la facilidad de almacenamiento y transporte de larga distancia, y el control de autoridades y gobiernos en el precio del activo. Se tratarán más en detalle en la sección 2.2 los factores de volatilidad del activo. Afecta por tanto de manera directa e indirecta al precio del hidrógeno verde y del resto de materias primas a nivel global, en general.

- Gas natural TTF (*Dutch TTF gas*): el gas natural TTF es un indicador de referencia del precio del gas natural cotizado en el *Title Transfer Facility*, mercado virtual ubicado en Holanda en el que cotizan diversos contratos de gas natural, fijando el precio del dicho elemento en toda Europa. Es un mercado altamente líquido en el que se negocian gran cantidad de instrumentos financieros relacionados con la materia prima en cuestión. Este activo referencia es fundamental para la fijación de precios del gas natural en todo el

mercado europeo, y sirve como punto de referencia para medir los niveles de oferta y demanda de la materia y establecer el precio en cada momento (Equipo Singular Bank, 2022). La relación del activo con el hidrógeno verde es muy alta, no sólo por el hecho de que también influye en los costes directos de producción y distribución del activo, sino por la estrecha correlación entre el hidrógeno verde y el hidrógeno azul (materia prima producida a través de un proceso de hidrólisis mediante gas natural): el precio del gas natural incide de manera directa y total en el precio del hidrógeno azul, que a su vez, a causa de la similitud de composición y uso, incide fuertemente en el precio del hidrógeno verde. Se observa más adelante, en la sección 2.3.1 la fuerte correlación entre los dos activos. En cuanto a la volatilidad, el gas natural TTF presenta unos niveles relativamente más altos que los del petróleo crudo y el americano. Los movimientos en el precio de este activo son mucho más bruscos por cuestiones como la dependencia regional (un elevado porcentaje de la industria europea depende directamente de dicho activo, y las fuentes de suministro son más escasas que las de petróleo), la relativa dificultad de transporte (la dependencia de infraestructura específica para el transporte del elemento, a través de gasoductos, por ejemplo, hace que el precio oscile de forma más brusca), y los factores geopolíticos (aunque el petróleo haya estado sometido a grandes fuentes de incertidumbre, la producción de gas natural en Europa es actualmente más sensible por la cercanía de los conflictos de oriente medio, los controversiales cambios políticos y regulatorios de cada región productora, y, sobre todo, por la dependencia europea del gas ruso). En la sección 2.2 se estudia más a fondo la volatilidad del gas natural TTF y sus causas.

Estos tres activos son considerados fundamentales para determinación del precio de muchas otras materias primas, ya que afectan de manera directa o indirecta a la producción, almacenaje, distribución y uso de todas ellas, y es por ello por lo que tiene sentido utilizarlas para predecir el precio del hidrógeno verde. Por todo ello y por la disponibilidad de información histórica a nivel financiero y la abundancia de instrumentos bursátiles, se estudiará la relación de cada uno con el precio del hidrógeno verde para obtener conclusiones que más adelante se comentarán.

2. Análisis cuantitativo

2.1. Extracción de datos

En este proceso se ha llevado a cabo una búsqueda y recolección de los precios de las materias primas mencionadas anteriormente. Las dos fuentes utilizadas para la recolección de datos han sido S&P Global Commodity Insights (para la obtención de los precios históricos del hidrógeno verde en euros a través de su extensión de Microsoft Excel, la cual, al introducir el código específico del hidrógeno "HYNWB00", devuelve los precios de dicha materia prima para un rango de tiempo establecido), y Factset (para la obtención de los precios históricos y de los futuros en euros del resto de materias primas).

Para facilitar el análisis, se han tomado los precios desde el año 2014 para todas las materias primas, excepto para el hidrógeno verde, ya que de este solamente se tienen datos a partir de finales del año 2021. Es necesario tener en cuenta que, debido a que el hidrógeno verde todavía no cotiza en ningún mercado de referencia, los datos utilizados son valoraciones de precios establecidas en función del coste de producción del hidrógeno verde.

La hoja final a importar para llevar a cabo el análisis de datos está compuesta por una primera columna con las fechas diarias de cotización y cuatro columnas adyacentes con los precios diarios en euros del hidrógeno, el petróleo americano, el petróleo crudo y el gas natural TTF, para cada uno de los días. Es evidente que, como cada materia prima cotiza en un mercado diferente, se dan ocasiones en las que alguno de estos está cerrado y no hay precio de referencia disponible; para estos casos, por motivos de consolidación, se ha optado por tomar el precio más reciente disponible y extrapolarlo al día correspondiente.

Una vez extraídos, consolidados y agrupados todos los precios históricos, se ha obtenido una hoja de Excel final que posteriormente será importada en el programa de análisis de datos.

2.2. Importación y análisis cualitativo

Los datos del documento de Excel anteriormente mencionado se han almacenado en un paquete de datos (*dataset* en inglés) llamado "datos". Dicho *dataset* es inicial y superficialmente analizado utilizando las funciones *summary*(), que indica para cada materia prima el precio mínimo, el primer cuartil, la mediana, la media, el tercer cuartil y el precio máximo; *view*(), que crea una tabla visual con los datos del Excel importado; y *str*(), que devuelve la composición del paquete de datos importado, así como el formato de datos de cada una de las columnas y su tamaño.

Posteriormente, se almacena el precio de cada una de las materias primas en paquetes de datos secundarios, indicando para cada uno de ellos el número de columna del *dataset* primario al que se quiere hacer referencia.

Para llevar a cabo la visualización preliminar de los datos de precios, se ha creado un gráfico de caja y bigotes (*boxplot* en inglés), que muestra la distribución de precios de manera conjunta, aportando información sobre los datos máximo y mínimo, la mediana y los demás cuartiles para cada materia prima, y un gráfico de línea, que muestra la evolución histórica del precio diario de cada materia prima. Complementariamente, se ha calculado la media, desviación típica y varianza de cada una de las variables.

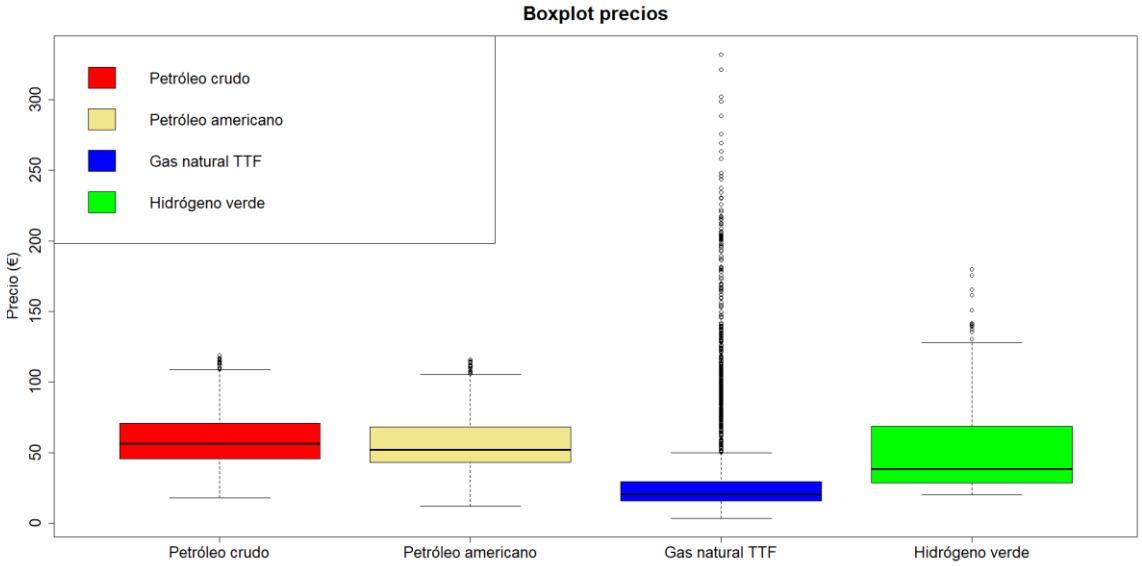

Fig 1. Boxplot de niveles de precios de los petróleos, el gas natural y el hidrógeno verde

El gráfico de caja y bigotes representado nos da una visión de los niveles de precios en los que se mueven las materias primas analizadas a lo largo del tiempo estudiado. El petróleo americano spot y el petróleo crudo spot mantienen niveles de precio relativamente similares, con medianas muy parecidas en torno a los 50 euros, con una dispersión relativamente baja, y con pocos valores atípicos (*outliers* en inglés), ello se debe a que los valores de cotización de estas dos materias primas se han movido de manera similar y estable a lo largo del periodo comprendido entre los años 2014 y 2024 dada su naturaleza poco volátil (desarrollada en la sección 1.2 “Elementos comparativos”).

Por otro lado, el gas natural TTF presenta una tendencia bastante diferente: en primer lugar, la caja es bastante reducida debido a que, como se observa posteriormente en el gráfico de línea, durante el periodo 2014 – 2021, la volatilidad de precios es relativamente baja, situándose estos

constantemente entre los 20 y los 30 euros aproximadamente; en segundo lugar, se observa una cantidad de *outliers* sorprendentemente alta, debido a que durante el periodo 2022 – 2024 la cotización ha sido anómalamente volátil.

Por último, el hidrógeno muestra una tendencia volátil, como se puede observar dada la altura de la caja, con ciertos *outliers*, debido a que su precio de referencia, desde 2021, ha tenido bruscas subidas y bajadas por razones como el inicio de la crisis energética, la recuperación económica tras la pandemia y el consecuente aumento exponencial de la demanda, los conflictos geopolíticos en regiones productoras de energía, y los problemas de producción derivados de la crisis COVID, entre otras. Posteriormente se desarrollarán todas estas cuestiones y sus consecuencias en el precio de la energía.

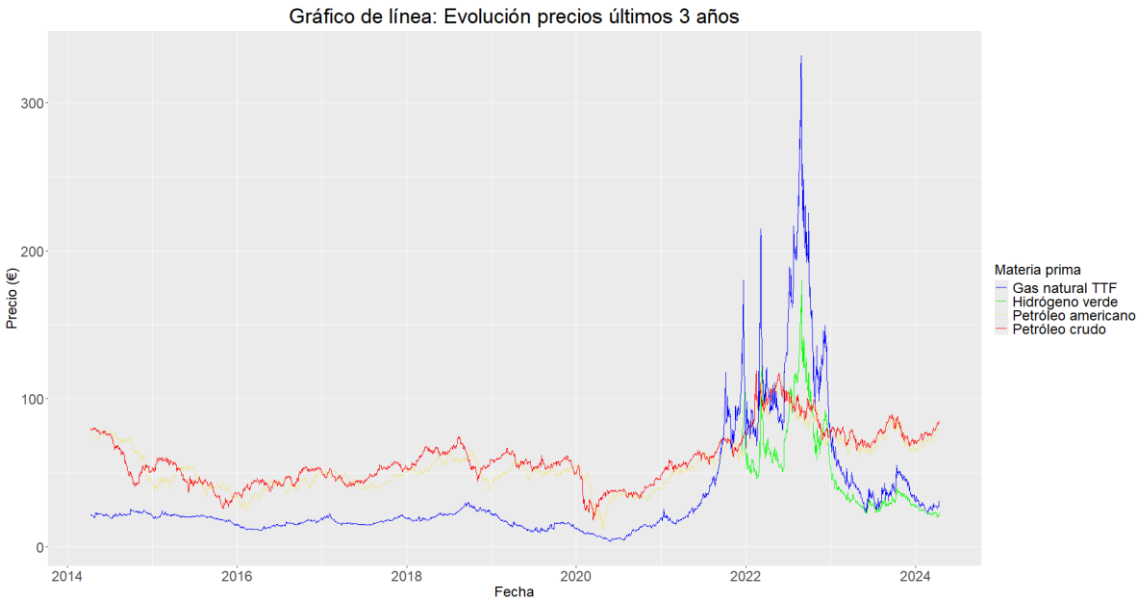

Fig 2. Gráfico de línea de niveles de precios de los petróleos, el gas natural y el hidrógeno verde

Con el apoyo del gráfico de línea se ha llevado a cabo un estudio más profundo de las causas que explican la estabilidad y la volatilidad de cada uno de los activos en el plazo de tiempo estudiado:

- Petróleo americano WTI (*US Crude oil WTI*): la correlación y similitud de movimiento de precios entre el petróleo americano y el petróleo crudo ha sido bastante alta: la tendencia durante todo el periodo mencionado ha sido la misma, pronunciadas fluctuaciones en el periodo 2014-2020 por las razones anteriormente mencionadas, caída durante los primeros meses de la crisis COVID, y recuperación y crecimiento durante el resto del periodo estudiado. Aunque el movimiento haya sido muy parecido, la fluctuación de precios ha sido ligeramente menor en el caso del petróleo americano,

debido fundamentalmente a que los acontecimientos que han afectado al precio de la energía recientemente han incidido de manera más intensa en Europa. Tras la caída post-COVID, el petróleo americano ha crecido de manera ligeramente menos brusca que el *brent* debido a que el conflicto europeo recientemente acaecido no ha afectado apenas a Estados Unidos, uno de los principales consumidores de esta materia prima que, por razones económico-políticas, no depende prácticamente de otras economías para abastecerse, especialmente a nivel energético. Es por ello por lo que no se ha observado ninguna fuerte reducción en la oferta que haya podido afectar de manera brusca y en ese sentido al precio de la materia prima en cuestión. Como en el caso del *brent*, el precio del petróleo americano se ha visto también fuertemente afectado por la inflación mundial, que ha provocado que, a partir del año 2022 sobre todo, el precio de la energía suba de manera drástica. Es necesario mencionar que el petróleo (tanto el *brent* como el americano) ha sido una de las materias primas menos afectadas en cuanto a precio por la inflación y los acontecimientos geopolíticos mencionados, ya que se trata de un elemento poco volátil del que dependen gran cantidad de industrias a nivel mundial y cuya producción no se ve tan afectada por factores externos, provocando que los niveles de oferta y demanda sean razonablemente estables.

- Petróleo crudo (*Brent crude ICE*): el movimiento de precio de esta materia prima ha sido relativamente estable en comparación con el resto. Se puede observar cierta fluctuación en el precio durante el periodo comprendido entre los años 2014 y 2020, causada principalmente por factores como la Crisis del Petróleo comprendida entre los años 2014 y 2016, que provocó volatilidad en la trayectoria del precio (Cervera & Figuerola Ferretti, 2024). Este último año se observa fuerte una caída en el precio a causa de la crisis COVID, que provocó una reducción de la demanda a causa del desplome de la actividad económica. La demanda de petróleo se reactivó rápidamente al tratarse de una materia prima fundamental en muchas industrias, recuperando niveles de precio normales en cuestión de meses (Banco Mundial, 2020). Tras dicha recuperación se observa un crecimiento progresivo en el precio del petróleo crudo, que se mantiene en niveles máximos durante casi dos años. La subida en el precio de dicha materia prima se debe a varias causas, como la reactivación de la economía mundial tras la pandemia (incrementando considerablemente la demanda y provocando movimientos bruscos en el precio), el conflicto entre Rusia, uno de los principales proveedores de energía mundiales, y Ucrania (provocando un aumento indirecto en el precio de esta materia prima entre otras), los fallos en las cadenas de suministro de petróleo derivadas de la crisis COVID (que afectaron gravemente a la industria petrolera), o los desastres naturales y condiciones adversas en zonas productoras de petróleo (afectando a la producción y por consiguiente,

al precio). La subida del precio del *brent* va también ligada al aumento de la inflación vivido a nivel mundial: la subida de precios generalizada que se lleva gestando desde la recuperación de la crisis de 2008 ha tocado máximos históricos, siendo el incremento del precio de la energía en general una de las causas fundamentales (Cohen, 2022).

- Gas natural TTF (*Dutch TTF gas)*: esta materia prima presenta una tendencia muy estable durante el periodo 2014 – 2021, y una muy fuerte fluctuación durante el periodo 2021 – 2024. La estabilidad en el precio del gas natural durante los años comprendidos entre 2014 y 2021 se debe, entre otras razones, a su estabilidad en cuanto a demanda: el gas natural es una materia prima utilizada para la generación de electricidad y las aplicaciones industriales (entre otras), sectores que presentan una gran estabilidad incluso en momentos de crisis (Soler, 2023). La situación cambia drásticamente durante el año 2021 debido a fuertes cambios tanto en la oferta como en la demanda. La demanda de gas natural se ha disparado recientemente debido a factores como la progresiva reducción de consumo de carbón en Europa (impulsada por organismos reguladores que buscan reducir emisiones) y por ciclos bajistas en la producción de energía eólica, provocando un crecimiento en el consumo de fuentes alternativas, especialmente gas natural. La oferta se ha visto afectada por los problemas derivados de la falta de mantenimiento y cuidado, y de una menor inversión de los yacimientos de gas durante la etapa COVID, resultando en una producción menor en cantidad y calidad (Euro News, 2021). Las subidas y bajadas de precio de esta materia prima se han visto especialmente afectadas por el conflicto entre Rusia y Ucrania, ya que la primera ha sido el mayor importador de gas natural en Europa durante los años 2019 y 2020, y la falta de suministro a dicha región durante la guerra, las sucesivas negociaciones entre países, y los altibajos del conflicto han provocado movimientos drásticos en el precio del gas natural TTF, directamente afectado por la situación. Todo ello, sumado a la fuerte inflación vivida y a las medidas estatales para atajarla, han provocado unas fluctuaciones nunca vistas en el precio del gas natural, moviéndose en rangos superiores a los 200 y 300 euros, cuando históricamente se ha situado entre los 20 y los 30 euros (Chadwick, 2021).

- Hidrógeno verde: las primeras referencias de precios datan de finales de 2021, y desde entonces, los movimientos de precio han sido relativamente fuertes: se observan tendencias similares a las del gas natural TTF, pero al no haber referencias de precios históricos más allá del año 2021, no se puede afirmar si dicha volatilidad es algo normal en el precio de esta materia prima. Hay muchas causas directamente relacionadas con dichas fluctuaciones: en primer lugar, los fuertes movimientos en el precio del gas natural (el coste de producción del hidrógeno verde está influenciado en gran parte por el precio

de la electricidad, vinculada notablemente al precio del gas natural, aunque mucha electricidad se produzca a partir de fuentes renovables; por lo tanto, todas las variables que afectan al precio del gas natural afectan, directa o indirectamente al precio del hidrógeno verde); en segundo lugar, la dependencia de energía renovable (para producir el hidrógeno verde mediante electrólisis es necesario el uso de energía eólica o solar, principalmente, por lo tanto, los recientes cambios en el precio de estas derivados, entre otros, de condiciones climáticas no óptimas sobre todo en Europa, han provocado cambios bruscos en el precio del hidrógeno verde); y en tercer lugar, la inestabilidad geopolítica causada por el conflicto entre Rusia y Ucrania (la falta de abastecimiento por parte de Rusia y las sucesivas negociaciones han generado altibajos en la demanda, y por tanto en el precio de cotización del hidrógeno, fuente sustitutiva). Hay muchos otros factores que han provocado fuertes movimientos en el precio del hidrógeno, pero los mencionados son los principales (Scheibe, A. & Poudineh, R, 2023). Adicionalmente, es necesario recalcar que, como se menciona en las razones que explican la volatilidad del hidrógeno verde, el precio de esta materia prima está fuertemente condicionado por el precio del gas natural TTF. Esta relación de precios radica en la correlación entre los mercados del gas natural, la electricidad y el hidrógeno verde: la producción del hidrógeno verde depende directamente del precio de la electricidad (el proceso de electrólisis requiere energía renovable para llevarse a cabo). Esta última está a su vez influenciada fuertemente por el precio del gas natural a causa de la estructura del mercado eléctrico actual (Sharma & Escobari, 2018). Por ello, como se observa en el gráfico, el precio del hidrógeno verde está en parte relacionado con el del gas natural TTF.

Una vez analizadas todas las materias primas, sus movimientos de precio y las causas, a primera vista se puede afirmar que, por numerosas y diversas razones, el gas natural TTF es la que mayor relación guarda con el hidrógeno verde a nivel cualitativo, ya que están afectadas prácticamente por los mismos acontecimientos geopolíticos, económicos y regulatorios.

En las siguientes secciones del trabajo se llevarán a cabo análisis numéricos que permitirán determinar qué materia prima guarda una mayor relación cuantitativa con el hidrógeno verde, para posteriormente ver cuál de ellas se utilizará para desarrollar los modelos predictivos.

2.3. Análisis y modelaje

Una vez importados los precios de las materias primas y analizadas las causas fundamentales que explican los movimientos en su cotización, se procede a realizar el análisis que permitirá conocer qué materia prima tiene una mayor relación a nivel cuantitativo con el hidrógeno verde, para

posteriormente realizar los modelos predictivos en base a ella. En primer lugar, se lleva a cabo un análisis de correlaciones y covarianzas de cada una de las materias primas de referencia (el petróleo crudo, el petróleo americano y el gas natural TTF) con el hidrógeno, y, en segundo lugar, una vez decidida la materia prima a tener en cuenta, se crea un bosque aleatorio para reafirmar que la elección ha sido correcta, y que se trata de la materia prima indicada para crear los modelos predictivos.

2.3.1. Análisis de covarianza y correlación

Para el análisis de covarianza y correlación se han importado datos dimensionados de las materias primas de referencia, es decir, tomando todos los precios de cotización de materias primas desde el día de inicio de cotización del hidrógeno verde estudiado, descargadas de un segundo fichero Excel utilizado para el estudio (la composición de este fichero es la misma que la del primer fichero importado, con la única diferencia de que solamente contiene datos desde el 13 de diciembre de 2021 hasta el 12 de abril de 2024 para cada una de las materias primas).

Posteriormente, se han normalizado los precios de las materias primas con el objetivo de estandarizar las variables fomentando una comparación más precisa de estas, ya que cotizan en índices diferentes, se mueven en tendencias y rangos diversos, y los valores de precio son diferentes unos de otros. Para normalizar estas variables se ha utilizado la función *scale*() de R Studio, que escala cada uno de los valores del conjunto de datos restando la media y dividiendo el resultado por la desviación típica del mismo. Antes de estudiar la relación numérica entre variables, se han representado de manera visual las variables normalizadas, para ver la composición y evolución de los precios normalizados entre los años 2021 y 2024.

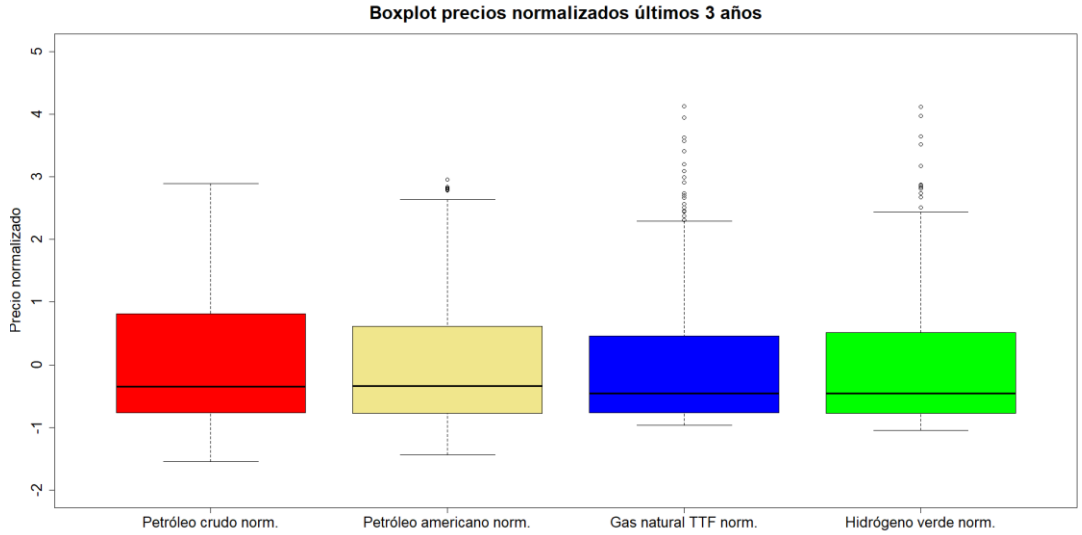

Fig 3. Boxplot de precios normalizados de los petróleos, el gas natural y el hidrógeno verde

Como se puede observar, al normalizar las variables los conjuntos de datos se comportan de manera similar, siendo el gas natural TTF y el hidrógeno verde las materias primas con mayor dispersión, como se puede observar gracias a la cantidad de *outliers*.

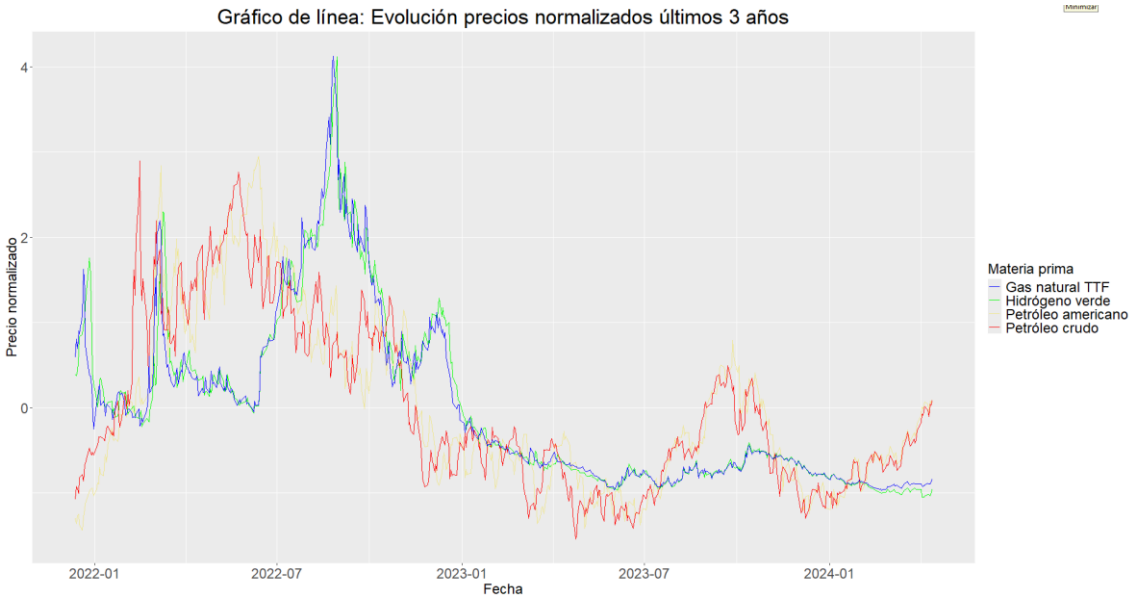

Fig 4. Gráfico de línea de precios normalizados de los petróleos, el gas natural y el hidrógeno verde

En el gráfico de línea con los precios normalizados se observa de manera mucho más clara cómo la tendencia del gas natural TTF es la más parecida a la del hidrógeno verde con diferencia. Es necesario comentar que los precios graficados están normalizados, por lo que la evolución histórica de cada una de las materias primas únicamente sirve para ver la evolución a lo largo del tiempo sin conocer las diferencias de niveles en una unidad específica. Por otro lado, se observa como la correlación entre los dos petróleos también es elevada.

Una vez graficados de nuevo los precios, esta vez normalizados, se lleva a cabo el estudio de covarianzas y correlaciones. Para este estudio se han implementado funciones básicas en R Studio que permiten analizar covarianzas y correlaciones entre las materias primas de referencia y el hidrógeno, y se han graficado para hacer el estudio más visual

- Covarianza: la función de R Studio cov() permite conocer la covarianza entre dos conjuntos de datos introducidos. En este caso, se han estudiado las siguientes covarianzas:
    - Covarianza entre los precios del petróleo crudo y el hidrógeno verde: 0,55
    - Covarianza entre los precios del petróleo americano y el hidrógeno verde: 0,54
    - Covarianza entre los precios del gas natural TTF y el hidrógeno verde: 0,97

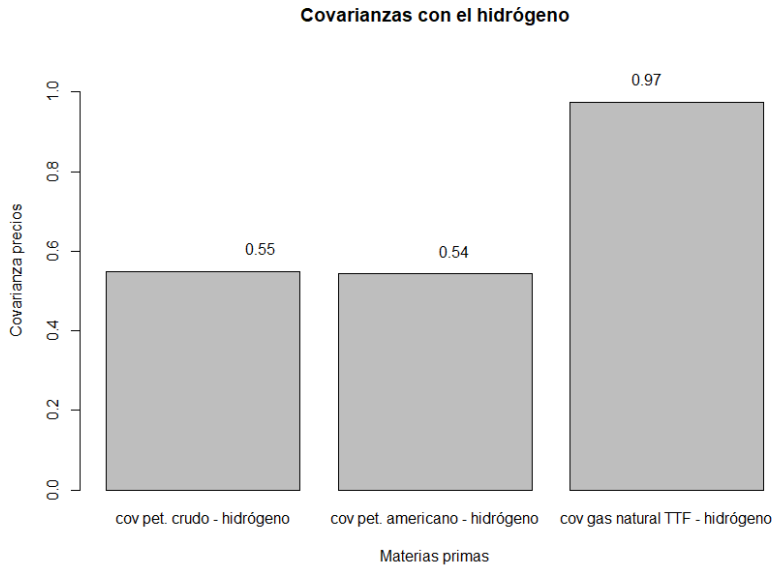

Fig 5. Gráfico de barras de la covarianza entre los petróleos y el gas natural con el hidrógeno verde

Se puede observar como la covarianza más alta del conjunto de datos estudiado es la del gas natural TTF con el hidrógeno verde.

- Correlación: la función de R Studio corr() permite conocer la correlación entre dos conjuntos de datos introducidos. En este caso se han estudiado las siguientes correlaciones:
    - Correlación entre el petróleo crudo y el hidrógeno verde: 0,55
    - Correlación entre el petróleo americano y el hidrógeno verde: 0,54
    - Correlación entre el gas natural TTF y el hidrógeno verde: 0,97

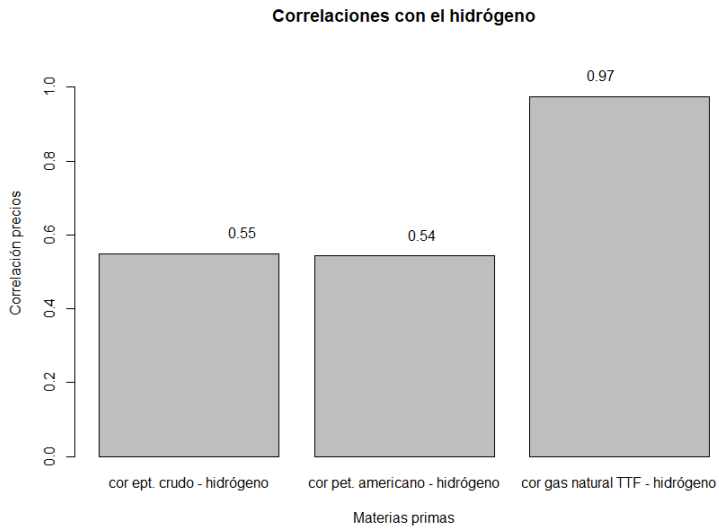

Fig 6. Gráfico de barras de la correlación entre los petróleos y el gas natural con el hidrógeno verde

Se puede observar como la correlación más alta del conjunto de datos estudiado es la del gas natural TTF con el hidrógeno verde.

El valor de la covarianza de cada una de las materias primas de referencia con el hidrógeno verde es la misma que la respectiva correlación, debido a que, cuando las variables están normalizadas, la única diferencia que radica entre estas dos métricas, la escala, desaparece, haciendo que los resultados sean los mismos.

Como conclusión principal se observa que la materia prima que mayor covarianza y correlación guarda con el hidrógeno verde es el gas natural TTF, reafirmando por tanto las conclusiones cualitativas obtenidas observando los gráficos de secciones previas. De esta manera, al ser el gas natural TTF la fuente de energía que mayor relación guarda con el hidrógeno verde, tanto a nivel cualitativo como cuantitativo, se plantea como variable a utilizar para la creación de las predicciones.

2.3.2. Bosque aleatorio

Una vez decidida la materia prima a utilizar, se lleva a cabo la creación de un bosque aleatorio para asegurarse de que se trata de la indicada para realizar predicciones. El bosque aleatorio es un algoritmo de aprendizaje automatizado usado para problemas de clasificación y regresión. El bosque aleatorio o *random forest* en inglés, está formado por un conjunto de árboles de decisión (método de toma de decisiones en base a una serie de condiciones y/o preguntas específicas).

El bosque aleatorio permite tomar decisiones basadas en grandes cantidades de información, incorporando numerosas variables y teniendo muchos escenarios y condiciones en cuenta (Daniel, 2023). Para llevar a cabo la creación del bosque se ha establecido una semilla aleatoria con la función de R Studio *set.seed*(), fomentando la reproducibilidad de resultados cada vez que se ejecute el código, al tratarse de un proceso aleatorio que genera un resultado diferente cada vez que se pone en marcha.

Después se esperaría llevar a cabo una partición aleatoria de los datos de los precios, pero en este caso, como el bosque aleatorio es utilizado únicamente para ver cuál de las variables es más efectiva para predecir el precio del hidrógeno, se utiliza el 100% de los datos disponibles para entrenar el modelo, ya que no será puesto a prueba, y de esta manera, el modelo será más efectivo al haber utilizado la totalidad de información disponible.

El bosque aleatorio se calcula con la función *randomForest*(), que crea un conjunto de árboles de decisión en base a los precios de las materias primas aportados para entrenar el modelo mediante un proceso conocido como *Bootstrap*, que realiza un muestreo con reemplazo de datos para hacer el modelo lo más efectivo posible.

Una vez creados los árboles y agrupados en bosques, se promedian las predicciones de cada uno de los árboles individuales del bosque, y se ponen en común para lograr el modelo más eficiente y preciso posible. Una vez realizadas las sucesivas combinaciones y creado el bosque aleatorio, se imprime un resumen para obtener información acerca del número de árboles y la importancia de cada una de las variables (ver Código 1– *Random forest* en el anexo).

El modelo ha calculado un bosque aleatorio ajustado de tipo regresión, ya que el modelo se quiere para predecir la efectividad de variables continuas, y no categóricas, y ha utilizado 500 árboles aleatorios para ajustar el modelo, tomando una única variable predictora en cada nodo de división de los árboles (*No. of variables tried at each Split*).

Se calcula también el error cuadrático medio (*Mean of squared error/residuals*), que indica la media de las diferencias al cuadrado entre las predicciones realizadas por el modelo y los valores reales de las variables utilizadas, en este caso, los precios históricos de las materias primas a comparar.

Por último, el porcentaje de varianza explicada (*% var explained*) indica el nivel de variabilidad explicado por el modelo calculado, es decir, el porcentaje de variabilidad del precio del hidrógeno explicado por el resto de variables o materias primas utilizadas, un 94.68% en este caso, indicando

que se trata de un modelo bastante bueno, con un buen ajuste de los datos, siendo capaz de capturar casi la totalidad de la varianza de los datos estudiados.

Después de esto, se calcula la importancia de las variables usadas para generar el modelo de bosque aleatorio creado mediante la función *importance*(), que mide la relevancia relativa de cada variable predictora del modelo mediante el cálculo de la reducción de impurezas por parte de cada una de ellas, siendo la más efectiva aquella capaz de eliminar el mayor número (ver Código 2 – Importancia *random forest* en el anexo).

La variable *IncNodePurity* indica la contribución de cada variable a eliminar las impurezas generadas en cada uno de los nodos del árbol. Como se puede observar, el gas natural TTF (conjunto de precios dimensionado) es la variable que más purezas elimina, prácticamente triplicando las impurezas eliminadas por el resto de las variables. Se han representado en un gráfico las impurezas eliminadas por cada variable para hacerlo más visual con la función *varImpPlot*():

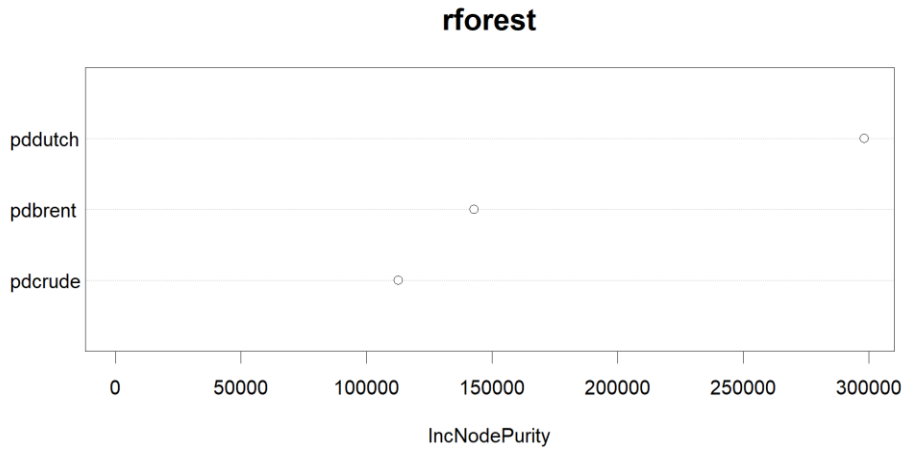

Fig 7. Gráfico de puntos de las impurezas eliminadas por el gas natural dimensionado, y los petróleos dimensionados

Gracias a este modelo de bosque aleatorio se puede afirmar con aún más certeza que la variable más adecuada a utilizar para predecir los precios del hidrógeno verde es el gas natural TTF, ya que los datos incluidos dentro de esta variable son los más parecidos a los del hidrógeno verde, siendo también los que más impurezas eliminan. Es necesario tener en cuenta la estrecha relación existente entre el precio del gas natural, la electricidad, y el hidrógeno verde, comentada en los apartados 2.2 y 3.1 del trabajo.

2.4. Evaluación de modelos: regresión lineal y KNN

Una vez seleccionada la variable más representativa, se llevará a cabo el estudio de los precios del hidrógeno y las predicciones con ella, tomando el gas natural TTF como referencia por similitud y reducción de impurezas en el modelo. Previamente, se representan ambas variables juntas para ver de manera clara cómo los precios de estas mantienen una tendencia similar.

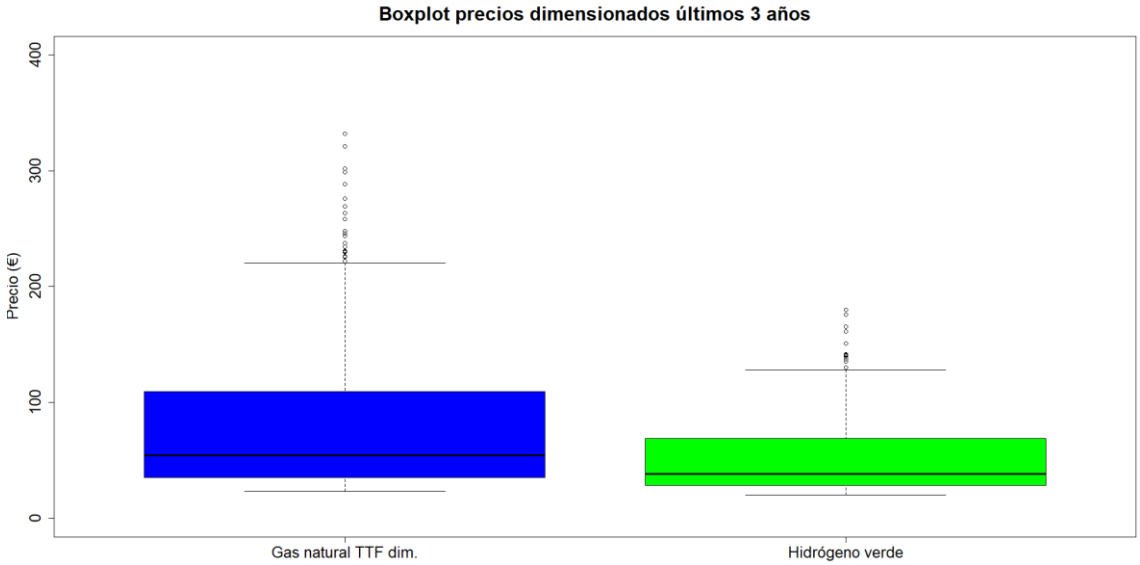

Fig 8. Boxplot de precios dimensionados del gas natural y el hidrógeno verde

Se puede observar una mayor dispersión en los precios del gas natural TTF, con una cantidad de *outliers* ligeramente superior.

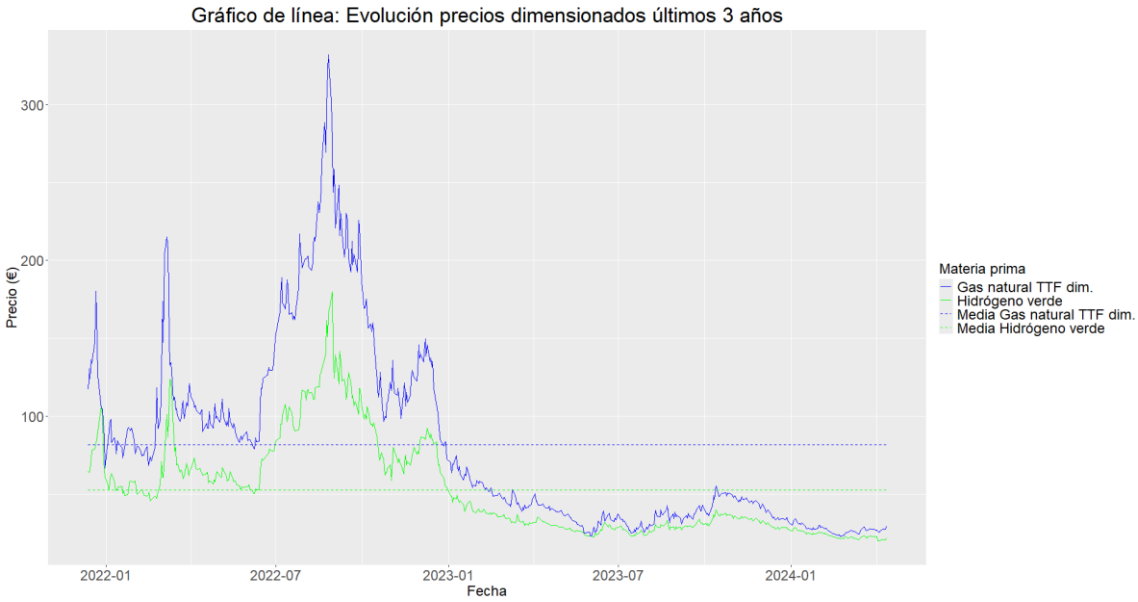

Fig 9. Gráfico de línea de precios dimensionados del gas natural y el hidrógeno verde

Con el gráfico de línea se percibe una tendencia bastante similar a lo largo del periodo estudiado, observando picos y valles en el precio de las materias primas en los mismos momentos del tiempo (causas de las que se hablan en la sección 2.2). Se representan también las medias de precios de cada una de las materias primas para facilitar su comparación.

2.4.1. Regresión lineal

El primer método utilizado para la predicción de los precios futuros del hidrógeno verde se basa en el cálculo de una regresión lineal, método estadístico utilizado para el estudio de la relación entre una variable dependiente (en este caso el precio del hidrógeno verde), y una o más variables independientes (en este caso el gas natural TTF) con la idea de establecer una relación entre estas que permita la obtención de conclusiones y la estimación de predicciones (Peláez, 2016).

Se busca obtener un coeficiente derivado de la regresión que nos permita, en base a los precios reales actuales de los futuros del gas natural TTF, estimar el precio esperado para el hidrógeno verde en esas fechas, calculando unos futuros ficticios para esta materia prima.

En primer lugar, se calcula la relación lineal entre los precios del hidrógeno verde y los precios dimensionados del gas natural TTF mediante la formula *lm*(), con la idea de ajustar un modelo que permita conocer la respuesta de la variable dependiente (los precios del hidrógeno verde) frente a la independiente (los precios del gas natural TTF). Con la función *summary*() se obtiene un resumen detallado del modelo de regresión lineal previamente creado (ver Código 3 – Resumen regresión lineal en el anexo)

- *Call*: indica que los resultados tratados son los obtenidos de la función *lm*().

- *Residuals*: devuelve un resumen de los residuos generados por el modelo, es decir, la diferencia entre los valores reales y los valores predichos por el modelo generado. Se puede observar que los valores mínimo y máximo estimados difieren en 39 y 46 unidades respectivamente de su valor real, pero el primer cuartil, la mediana, y el tercer cuartil (donde se concentran la mayoría de los datos) tienen una dispersión bastante moderada, cercana a las cero unidades, lo que sugiere que el modelo ha logrado capturar adecuadamente la variabilidad de las respuestas.

- *Coefficients*: dentro de esta métrica encontramos el intercepto (*intercept*) y el precio dimensionado del gas natural (pddutch). En este caso, el estudio se centra específicamente en el estudio de pddutch, ya que la interpretación del intercepto suele ser compleja y generalmente no tan significativa

    - *Estimate*: indica cuánto se estima que cambie el precio del hidrógeno por cada unidad de cambio en el precio del gas natural TTF. En este caso, se puede esperar que, cuando el precio del gas natural TTF aumente en una unidad, el precio del hidrógeno verde lo hará en aproximadamente 0,5.
    - *Standard error*: mide la precisión de la estimación, indica cuánto cambiarán las estimaciones si se utilizan conjuntos de datos diferentes. En este caso, el error estándar es bastante bajo, 0,005 aproximadamente, indicando que la estimación es muy precisa y que si se cambia el conjunto de datos estudiados funcionará de manera efectiva.
    - *T value*: nos indica cuánto mayor es la cifra estimada con respecto a su error estándar. En este caso se trata de un número elevado, aproximadamente 104, sugiriendo que se trata de una variable altamente significativa para el modelo.
    - Pr(>|t|): también conocido como valor p, mide la probabilidad de obtener un valor t mayor al observado (en este caso, aproximadamente 104). El valor p obtenido es de <2e-16, siendo un valor muy cercano a cero, indicando que se trata de una variable muy relevante y significativa para el modelo.

- *Residual standard error*: o error estándar residual, que mide la dispersión de los residuos con respecto a la media del modelo. En este caso se observa que los residuos se desvían de la media del modelo en aproximadamente 7 unidades. Se trata de un número relativamente bajo dada la media del modelo, por lo que el modelo está bien ajustado. Los 584 grados de libertad se obtienen restando el número de coeficientes a las observaciones del modelo.

- *Multiple R-squared*: mide el grado de variabilidad explicado por el modelo de regresión. En este caso es de aproximadamente 0,95, indicando que el modelo generado explica prácticamente el total de la variabilidad del modelo (aproximadamente el 95%).

- *Adjusted R-squared*: se trata de la misma métrica que el *multiple R-squared* pero ajustada al número de variables predictoras del modelo, en este caso una, por lo tanto el resultado y la interpretación es la misma.

- *F-statistic*: indica si el modelo generado funcionaría mejor que un modelo que no incluyese ninguna variable predictora. En este caso el valor de F es muy bajo, indicando que el modelo es significativo para predecir variables.

- *P-value*: indica lo mismo que el anteriormente explicado, pero para el *F-statistic*. Se ha obtenido un valor cercano a cero, indicando que se trata de un modelo muy significativo.

Tras haber estudiado el resumen detallado de la regresión lineal generada, se procede a realizar el test de Dickey-Fuller a los residuos del modelo y a los precios del hidrógeno verde, ya que, al haber realizado la regresión en niveles, es decir, con precios absolutos, es necesario que los residuos generados por el modelo sean estacionarios.

El objetivo es conocer si hay una tendencia significativa en las series temporales analizadas mediante un contraste de hipótesis de raíz unitaria. El rechazo de la hipótesis de raíz unitaria indica un ajuste adecuado del modelo (que captura correctamente las relaciones entre variables estudiadas) con errores predecibles (las desviaciones de precios frente a la media tienden a regresar a ella). Si esto se cumple, las predicciones del modelo serán más fiables y el modelo de regresión con precios estará ajustado y estimará de manera adecuada (Rodó, 2021).

Se ha realizado el test para los residuos del modelo y para los precios del hidrógeno verde con la función *adf.test*(), a través de la cual se analizará el p-valor de los residuos generados por la regresión lineal.

De esta manera se obtiene información sobre su estacionariedad, ayudando a rechazar la hipótesis nula de que la serie tiene una raíz unitaria, sugiriendo por tanto que es estacionaria y consiguientemente adecuada.

El p-valor de los residuos es 0,01, menor que el nivel de significancia común de 0,05, por lo que se rechaza la hipótesis nula de que hay raíces unitarias. Gracias a ello, se puede afirmar con confianza que los residuos son estacionarios, confirmando que el modelo captura correctamente las relaciones entre datos y por tanto puede predecir adecuadamente.

Tras todo ello, se utiliza la función *coef*() para obtener el coeficiente del modelo de regresión lineal ajustado. El coeficiente es de aproximadamente 0,49, indicando, como se explica en *Coefficients: Estimate*, que se espera que el precio del hidrógeno verde aumente aproximadamente en media unidad por cada aumento unitario en el precio del gas natural TTF.

Una vez estudiado el modelo, se importan los precios de los futuros del gas natural TTF, también obtenidos de Factset, y se analizan para comprobar que el estado de estos es el adecuado para el estudio. La importación y análisis de estos datos es igual que la realizada inicialmente con los datos de los precios, por lo que no se considera necesario repetir el proceso llevado a cabo.

Una vez analizados, se procede a estimar el precio de los futuros ficticios del hidrógeno verde mediante el producto entre el coeficiente estimado por el modelo de regresión lineal y cada uno de los precios de los futuros del gas natural TTF en cada una de sus respectivas fechas.

Se ilustran posteriormente los resultados obtenidos, observando por un lado la evolución de los precios reales de los futuros del gas natural TTF, y, por otro lado, los precios futuros estimados para el hidrógeno verde en dichas fechas.

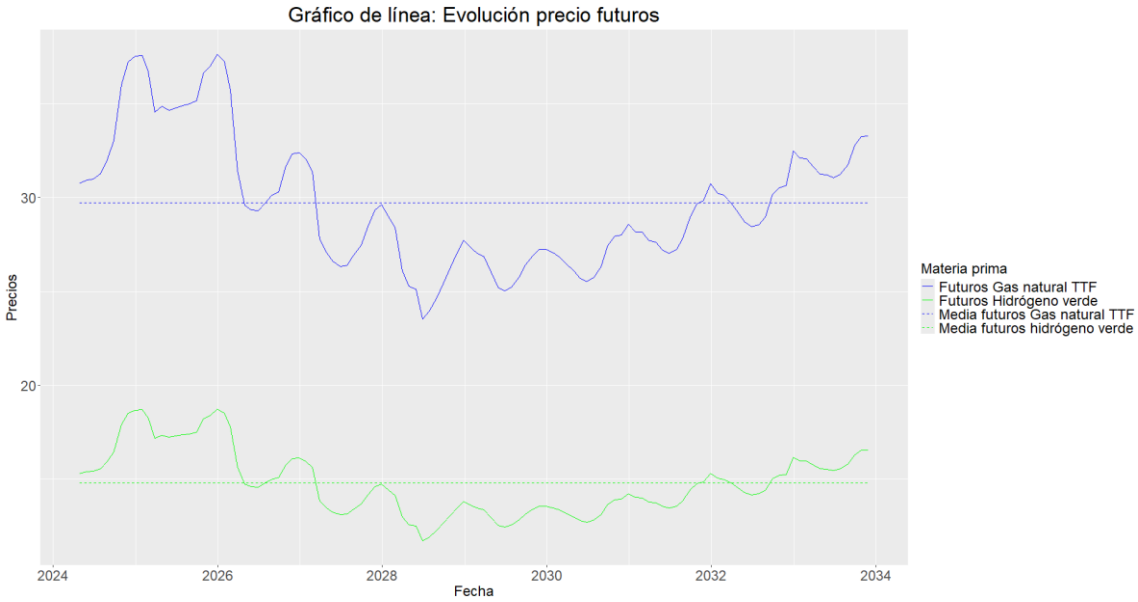

Fig 10. Gráfico de línea de precios reales de los futuros del gas natural y estimaciones de precios de los futuros del hidrógeno verde

Como se puede observar en el gráfico, la tendencia es la misma, ya que se ha multiplicado el precio de cada futuro por el coeficiente calculado, pero volatilidad de los futuros del hidrógeno verde es evidentemente menor, ya que, al ser números de menor magnitud, no están tan afectados como los del gas natural por un cambio de igual magnitud.

Se puede ver patrón relativamente similar dentro de cada año en los precios de los futuros del gas natural, debido a la estacionalidad de la demanda provocada por cuestiones como el aumento de uso de la calefacción, los patrones de producción y almacenamiento de productores y consumidores, o las limitaciones de transporte durante el invierno. Estas cuestiones provocan que,

como se observa, el precio aumente durante los meses de invierno y caiga el resto del año, siguiendo una tendencia parecida, pero con una evolución global irregular, como es de esperar.
En base a todo el exhaustivo análisis previamente realizado para la selección de variables y la predicción de precios, se puede esperar que los precios de los supuestos futuros del hidrógeno se moverían en dicha tendencia.

Al no existir instrumentos financieros derivados del hidrógeno, como los futuros, no se puede probar la efectividad real del modelo generado más allá del análisis realizado con la función *summary*(), en la que se han estudiado todos los componentes del modelo, reafirmando la alta efectividad de este.

2.4.2. KNN

El segundo método utilizado para la predicción de los precios del hidrógeno verde es el método KNN o K-vecinos más cercanos (*K-Nearest Neighbours* en inglés), un algoritmo de aprendizaje supervisado que permite calcular regresiones.

El algoritmo almacena datos históricos de otras variables aportadas (en este caso, el resto de las materias primas usadas para predecir el precio del hidrógeno verde) recopilando información que utilizará posteriormente a la hora de realizar estimaciones. Llegado el momento, clasificará (o predecirá) la variable en base a la similitud guardada con sus vecinos más cercanos (Orea, Vargas & Alonso, 2005).

En este caso, se ha entrenado el algoritmo para predecir, en base a los precios (o referencias de precios) históricos dimensionados (desde diciembre 2021 hasta abril 2024) del petróleo americano, el petróleo crudo, y el gas natural TTF para predecir el precio histórico del hidrógeno verde. Al ser predicciones realizadas para precios históricos, en este caso sí se tendrá información real para comprobar la efectividad del modelo.

En primer lugar, se ha establecido una semilla aleatoria para facilitar la reproducibilidad de resultados, ya que, como en el caso del *random forest*, se ponen en marcha procesos aleatorios sucesivos que hacen variar los resultados en cada iteración.

En segundo lugar, se lleva a cabo una partición de los datos del 80%/20%, con la idea de utilizar el 80% de los datos de precios históricos de las materias primas seleccionadas (petróleo americano, petróleo crudo y gas natural TTF) para entrenar el modelo KNN, y el 20% para ponerlo a prueba y comprobar su efectividad.

Una vez realizada la partición, se crea el algoritmo utilizando la función de RStudio knn() para predecir los precios de la segunda columna del conjunto de datos global (el hidrógeno verde), asignando un total de 5 vecinos utilizados para realizar las estimaciones (no se quieren asignar pocos vecinos para evitar el ruido, es decir, la sensibilidad a valores atípicos, ni muchos para evitar el sobreajuste, obteniendo un modelo poco representativo). En el proceso de entrenamiento, el algoritmo memoriza los datos aportados y estudia las relaciones de cercanía entre cada una de las observaciones. Posteriormente tiene lugar la predicción, en la cual, el modelo estudia los datos de entrenamiento más cercanos al punto que se desea predecir, y calcula la media de los K vecinos para asignar un valor numérico a ese punto que se desea predecir.

Una vez hecha la predicción de los precios del hidrógeno para cada una de las fechas con el algoritmo KNN, se almacenan en un mismo paquete o conjunto de datos las fechas dimensionadas (desde diciembre de 2021 hasta abril de 2024), los precios reales del hidrógeno, y los precios del hidrógeno predichos por el KNN. Posteriormente se representan mediante un gráfico de línea para percibir de manera visual la eficacia del modelo.

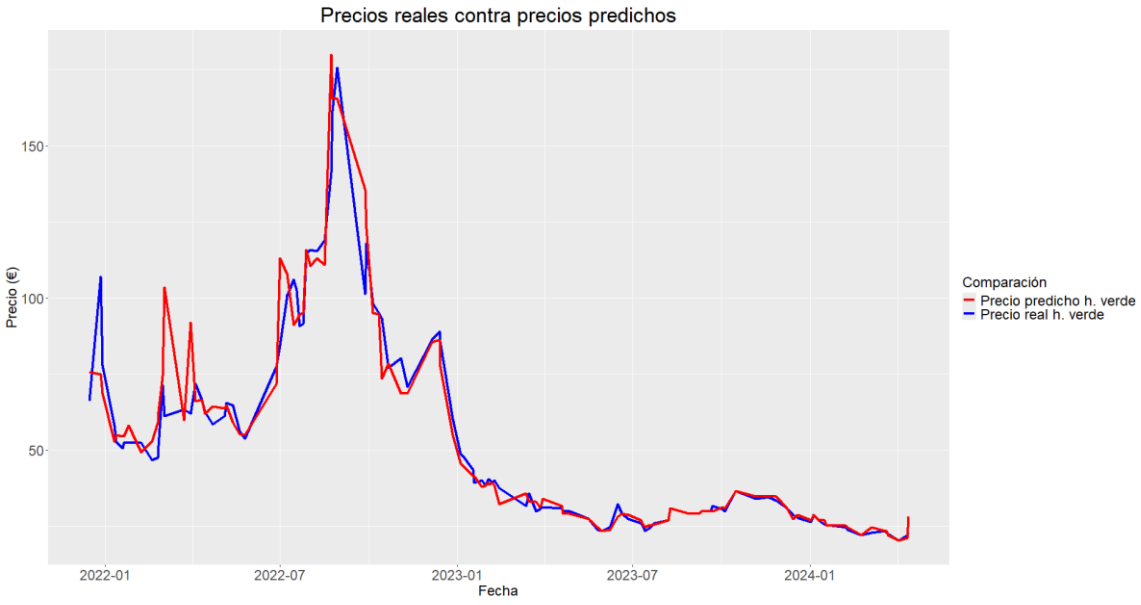

Fig 11. Gráfico de línea de los precios históricos reales y predichos del hidrógeno verde

Se observa como, por lo general, la tendencia es bastante similar, habiendo mayores niveles de dispersión en momentos de alta volatilidad en los cuales, los vecinos seleccionados por el algoritmo presentaban valores atípicos que han afectado al modelo. Se puede ver que en momentos de baja volatilidad (momentos del tiempo no afectados por crisis económicas o geopolíticas graves) las diferencias entre precios son mínimas. Se puede apreciar como el algoritmo KNN también está altamente influenciado por los valores de precios del gas natural

TTF, ya que no se observan apenas momentos de alta volatilidad para esta materia prima en los cuales el precio predicho del hidrógeno verde no acompañe dicha tendencia.

Posteriormente se lleva a cabo un estudio de la efectividad técnica del modelo KNN mediante el cálculo de las siguientes métricas (Ruf & Wang, 2019):

- MSE o error cuadrático medio (*mean squared error* en inglés): mide la calidad de la estimación realizada, midiendo la media de las diferencias al cuadrado entre los valores reales y los valores predichos por el algoritmo. Se obtiene un error cuadrático medio de aproximadamente 80 unidades, un número que indica que el modelo no es completamente preciso, y hay más variables de diferente naturaleza no estudiadas de además de los precios de las materias primas comparables que afectan al precio del hidrógeno verde. Se puede pensar que se trata de un número elevado, pero se deben tener en cuenta la magnitud de las variables tratadas, así como el hecho de que las diferencias están elevadas al cuadrado, haciendo que la cifra aumente.

- RMSE o raíz cuadrada del error cuadrático medio (*root mean square error* en inglés): lleva a cabo el mismo estudio que el MSE, pero calculando la raíz cuadrada del error medio. En este caso, el resultado obtenido ha sido de aproximadamente 9 unidades, indicando también que el modelo, evidentemente, no es completamente preciso.

- Validación cruzada: método adicional utilizado para medir la efectividad del modelo. En este caso, se ha dividido el conjunto de datos en 10 grupos diferentes que se utilizan tanto para entrenar como para testear el modelo. En cada iteración el modelo se entrena con 9 de los 10 grupos de datos creados y se prueba con el grupo restante. El proceso iterativo se realiza durante 10 rondas y se calcula un promedio de los resultados obtenidos en cada una de ellas para medir el rendimiento del modelo (Murillo & Arcedalia, 2019). A través de este proceso se calcula un nuevo RMSE y la desviación típica de este (RMSED) para la creación de un intervalo de confianza que medirá la efectividad del modelo. A través del cálculo de la validación cruzada se han obtenido un RMSE y un RMSED de aproximadamente 8 y 3 respectivamente. El intervalo de confianza típicamente utilizado es del 95% por su amplia aceptación, ya que el valor crítico correspondiente de esta es el de 1,96 (valor que en las distribuciones normales de media 0 y desviación típica 1 deja un 2,5% de probabilidad en cada cola de la distribución). Tomando el valor crítico de 1,96 el intervalo de confianza quedaría como RMSE +/- 1,96* RMSED; sustituyendo por los valores reales, el intervalo final quedaría como 8,21 +/- 1,96*3,31, es decir, un intervalo de [1,71, 14,70]. Dicho intervalo nos dice que hay un 95% de probabilidad de

que el RMSE real del modelo esté dentro de ese rango calculado. Para medir la eficacia y magnitud del rango es necesario tener en cuenta ciertos factores: en primer lugar, la dispersión de los datos estudiados: en las representaciones se observan picos y valles abruptos que generan inconsistencias en los datos y hacen que la amplitud sea mayor; en segundo lugar, el hecho de que se está buscando una confianza del 95%; y en tercer lugar, que hay muchos otros factores de diversa naturaleza que afectan a los precios de las materias primas, como se menciona en apartados anteriores. Todo ello genera inconsistencias en los datos (valores atípicos), y dificultad de captar en el modelo todos los factores que afectan al precio de estas materias. Por tanto, habiendo observado la dispersión del modelo y de las variables durante el estudio de la desviación de los datos y la representación de los mismos en gráficos, se puede afirmar que se trata de un intervalo adecuado, correcto y contenido en cuanto a tamaño, teniendo en cuenta la volatilidad de todas las variables estudiadas, y los resultados obtenidos en cuanto a RMSE y RMSED indican que se trata de un modelo consistente y no sobre ajustado.

- MAE: media de los errores absolutos (*mean absolute error* en inglés), mide las desviaciones del modelo frente a los valores reales de manera absoluta, y se calcula dividiendo el sumatorio del valor absoluto de las diferencias entre datos calculados y datos reales entre el número de observaciones. En este caso, se obtiene un MAE de 4,34, indicando que las predicciones se desvían aproximadamente en 4 unidades de los datos reales, de media. La efectividad del MAE se debe medir teniendo en cuenta la desviación típica de las materias primas a estudiar, siendo la del petróleo crudo aproximadamente 18, la del petróleo americano aproximadamente 18, la del gas natural TTF aproximadamente 41, y la del hidrógeno aproximadamente 31. Si se relativiza el MAE respecto a la desviación de precios del conjunto de datos del hidrógeno, se observa que, en promedio, el error medio del modelo es mucho menor que la variabilidad de precios reales del hidrógeno verde, dando a entender que la media de los errores absolutos del modelo es relativamente baja, por lo que se entiende que el modelo está haciendo un buen trabajo en cuanto a precisión y calidad, indicando que este tiene la capacidad de predecir con bastante precisión.

- MAPE: media porcentual de los errores absolutos (*mean absolute percentage error* en inglés), mide las desviaciones del modelo frente a los valores reales de forma porcentual, y se calcula dividiendo el sumatorio de la desviación de cada una de las predicciones frente al dato real entre el número de observaciones. En este caso, el MAPE obtenido es del 6,63%, indicando que las predicciones se desvían aproximadamente un 7% de los datos reales, de media. Al no tener métricas en términos porcentuales y no tener modelos

comparativos similares a este, la relativización del MAPE no tiene sentido. El MAPE nos aporta la misma conclusión que el MAE, en tanto por ciento, por lo que no es necesario compararlo con ningún otro resultado.

Las métricas estudiadas nos indican que el algoritmo ha predicho relativamente bien los precios del hidrógeno verde utilizando el resto de las materias primas como referencia, si bien es evidente que hay momentos de volatilidad a lo largo del periodo estudiado que han provocado que la magnitud de las diferencias calculadas aumente por la dispersión de precios, pero los resultados obtenidos indican que la variabilidad generada por el modelo es relativamente baja en comparación con la de los datos iniciales utilizados.

Gracias a los cálculos hechos, se puede afirmar que el algoritmo calculado es preciso y efectivo a la hora de predecir el precio del hidrógeno verde basándose en otras variables (materias primas). Los errores absolutos y relativos calculados son bajos y la consistencia es alta, por lo tanto, se puede concluir con que el modelo rinde de manera adecuada y es confiable.

3. Conclusiones e implicaciones

Tras el exhaustivo análisis realizado en las secciones anteriores, se han podido tomar una serie de conclusiones a cerca de los resultados obtenidos:

3.1. Similitud relativa del hidrógeno verde

Se puede afirmar que, tanto cualitativa como cuantitativamente, la materia prima más similar al hidrógeno verde es el gas natural TTF, por las siguientes cuestiones

- Aspectos cualitativos – en primer lugar, por la fuerte relación que guardan los mercados de ambas materias primas: como se comenta anteriormente, la producción de hidrógeno verde depende directamente del precio de la electricidad, a su vez influenciada por el precio del gas natural TTF, dada la estructura actual del mercado eléctrico mundial; en segundo lugar, por la relación geopolítica de ambas: las dos materias mencionadas cotizan en mercados muy similares, afectados prácticamente por los mismos acontecimientos de carácter económico, geopolítico, social y regulatorio, haciendo que la relación entre ambos sea considerablemente elevada; y en tercer lugar, por la dispersión de los datos: al normalizar las variables, ambas materias primas mostraban niveles de dispersión muy superiores al resto de variables, mostrando una vez más un comportamiento similar.

- Aspectos cuantitativos – en primer lugar, por la elevada correlación y covarianza entre variables: al estudiar estas métricas, el nivel de similitud entre las dos materias primas en base a estos aspectos es muy elevado (0,97), especialmente relativizándolo con el del resto de variables (0,55 con el petróleo crudo y 0,54 con el petróleo americano); en segundo lugar, por la elevada reducción de impurezas: el gas natural TTF es la variable del bosque aleatorio estudiado que más contribuye a la reducción de impurezas del modelo, incluso triplicando la efectividad en comparación con otras variables, siendo la materia prima más representativa para predecir los precios del hidrógeno verde; en tercer lugar, por las tendencias históricas similares: los gráficos muestran visual y numéricamente que los movimientos de precios históricos de ambas variables son muy similares, más aun comparándolas con el resto de materias primas; en cuarto lugar, por la obtención de una regresión lineal muy significativa entre ambas materias primas: el modelo generado usando el gas natural TTF como variable independiente muestra un coeficiente muy significativo, indicando que el nivel de relación entre ambas variables es elevado; y en quinto lugar, por los resultados obtenidos del bosque aleatorio: el modelo generado nos reafirma que la variable más adecuada para llevar a cabo predicciones sobre

el hidrógeno verde es el gas natural TTF, ya que es la que capta de manera más efectiva la variabilidad de precios de este.

3.2. Efectividad de los modelos de regresión lineal y KNN

Tras la creación e implementación de los modelos creados a lo largo del trabajo, se puede afirmar que ambos han sido efectivos a la hora de realizar predicciones sobre el precio del hidrógeno

- Efectividad del modelo de regresión lineal – el modelo creado se entrenó utilizando los precios dimensionados del gas natural TTF como variable independiente, con la idea de predecir en base a este los precios del hidrógeno verde. Los resultados obtenidos fueron significativos a nivel estadístico, reafirmando la estrecha relación entre ambas variables. Tras ello, se estudiaron diferentes métricas para medir la precisión y la efectividad del modelo, relativamente alta. Es evidente que el modelo estudiado, además de predecir con alta efectividad, tiene ligeras limitaciones, en parte por el hecho de haber calculado intencionadamente la regresión con precios absolutos y no con rentabilidades, con el objetivo de fomentar la interpretabilidad de resultados y coeficientes al tratarse de precios en unidades originales, sin perder información y facilitando la interpretación. Para afirmar la eficacia del modelo de precios se ha realizado el test de Dickey-Fuller, obteniendo residuos estacionarios que afirman que el modelo captura correctamente tendencias en los precios y predice adecuadamente de manera consistente.

- Efectividad del modelo de KNN – el algoritmo generado se creó para predecir el precio del hidrógeno verde en base a la cercanía de puntos en el espacio multidimensional generado, utilizando como referencia los precios del resto de materias primas: el petróleo crudo, el petróleo americano, y el gas natural TTF. El modelo creado demostró su efectividad a la hora de predecir los precios del hidrógeno verde, tanto visual (gracias a la representación gráfica de precios predichos y precios reales históricos) como numéricamente (las métricas de rendimiento estándar evaluadas, como el MSE, la validación cruzada o el MAE, mostraban que las predicciones presentaban un alto nivel de precisión). Es evidente que el modelo estudiado, además de predecir con alta efectividad, puede presentar ligeras limitaciones por la falta de normalización de datos para predecir, no realizada para evitar la pérdida de información e interpretabilidad del modelo.

Ambos modelos presentan resultados positivos a la hora de medir su efectividad como se ha podido observar a lo largo del desarrollo de cada uno de ellos en la sección 2.4. No conviene

llevar a cabo una comparación entre ambos, ya que cada uno funciona de una forma distinta, tratando los datos de manera diferente.

3.3. Implicaciones y recomendaciones

Los resultados y conclusiones extraídas del análisis realizado tienen una serie de implicaciones a nivel económico-financiero: la capacidad de predecir el precio del hidrógeno verde tomando como referencia el precio histórico de otras materias primas relacionadas influiría de manera significativa en todas las medidas, inversiones y proyectos de numerosas industrias. Algunas de las principales implicaciones y recomendaciones a tener en cuenta son las siguientes:

- Estabilidad del mercado energético: las estimaciones de precios pueden permitir a empresas y gobiernos adelantarse a bruscos cambios en el precio de esta fuente de energía, mediante la implementación de estrategias que estabilicen el mercado y eviten el desabastecimiento o el fuerte encarecimiento del hidrógeno verde o fuentes relacionadas.

- Optimización de costes y planificación presupuestaria: la capacidad de las empresas de adelantarse y estimar en cierta manera el precio del hidrógeno verde les permite planificar su producción y administrar el consumo de la materia prima basándose en el coste que esta supondrá.

- Planificación de inversiones: las predicciones de precio, así como la identificación de ciclos en el valor del hidrógeno verde pueden influir en las decisiones de inversión en tecnología, infraestructura o almacenamiento. Las predicciones pueden determinar los momentos adecuados para invertir en infraestructura, incrementar el almacenamiento o desarrollar nuevas tecnologías, fomentando el desarrollo de la industria en general.

- Incentivos fiscales y ayudas gubernamentales: las predicciones de precios del hidrógeno verde pueden ayudar a gobiernos, asociaciones y organizaciones gubernamentales a diseñar y establecer programas de incentivos fiscales más efectivos y adaptados, evitando sobre incentivar a las empresas cuando el ciclo es alcista o desatenderlas en momentos de crisis. La captación de tendencias y predicciones puede permitir a los gobiernos desarrollar políticas que faciliten la producción inicial y el desarrollo tecnológico de empresas mediante subsidios o créditos fiscales, el reparto de ayudas o ventajas al inicio de los ciclos bajistas para evitar reacciones tardías y consecuencias más graves, o fomentando la transición hacia una economía verde.

- Incremento de la regulación: la capacidad de predecir con cierta fiabilidad los precios del hidrógeno verde puede provocar un incremento de la regulación vigente, aplicando un control mayor sobre el mercado, afectando a todos los actores directa e indirectamente relacionados con este. Es necesario tener esto en cuenta, ya que el marco regulatorio de la Unión Europea, como se ha podido ver durante la crisis energética, que favorece el desarrollo de un mercado eléctrico integrado (con un modelo de precios zonales marginales), ha expuesto a numerosas empresas a costes y riesgos elevados a causa de la fuerte dependencia en la producción electricidad de gas natural. Por tanto, tal y como aconsejan los expertos, es necesario reevaluar y modificar el marco regulatorio europeo en términos del mercado eléctrico, para reducir la exposición de empresas y consumidores a la fuerte volatilidad de precios y mejorar la estabilidad financiera de todos los agentes involucrados en el mercado (Segarra, Atanasova, Figuerola-Ferretti, 2024).

Estas son algunas de las ventajas que puede tener el hecho de utilizar el modelo creado en este trabajo para predecir el precio del hidrógeno verde en base a datos históricos. Como se puede observar, las repercusiones positivas a nivel económico, financiero, social, político, regulatorio y medioambiental son inmensas.

Adicionalmente, se considera necesario comentar que el conflicto entre Rusia y Ucrania, mencionado en secciones anteriores, afectó considerable y directamente a la estabilidad del mercado energético europeo debido a la falta de abastecimiento de gas, contribuyendo al desarrollo de una inflación sin precedentes.

La Comisión Europea, como organismo regulador, propuso numerosas medidas regulatorias para atajar la crisis y frenar la caída del mercado energético, como la conservación de mercados energéticos cortoplacistas (asegurando la eficiencia productiva y controlando la volatilidad), el apoyo de los contratos de largo plazo (con el objetivo de establecer señales de precio duraderas y el fomento de la inversión en renovables y electrificación), o la obligación de contratos de precio fijo (para proteger al consumidor final y reducir la volatilidad del precio) (Fabra, 2023).

3.4. Conclusiones finales

El estudio realizado ha permitido la extracción de conclusiones, implicaciones e ideas fundamentales a cerca de la predicción de precios del hidrógeno verde, así como los factores más relevantes a la hora de llevar a cabo su estudio cualitativo y cuantitativo.

Se ha estudiado cualitativamente cada una de las variables utilizadas para las predicciones, tratando cuestiones como su procedencia, su evolución histórica, y factores que influyen en su precio y volatilidad. Se ha identificado que el gas natural TTF es, de entre las variables estudiadas, la materia prima más relacionada numérica y conceptualmente con el hidrógeno verde, y la más indicada para predecir el precio de este.

Los modelos generados han permitido ver que se puede estimar de manera efectiva el precio histórico y futuro de una materia prima en base a otras directa o indirectamente relacionadas, siempre que la elección de las mismas esté fundamentada.

La repercusión a nivel global de poder predecir correctamente y con cierto grado de fiabilidad el precio del hidrógeno verde es inmensa, ya que la industria en la que se ha desarrollado el estudio es uno de los principales motores de la economía y el desarrollo mundial, y cualquier factor que provoque un cambio o acontecimiento relevante en ella afectará directa o indirectamente a la mayoría de las industrias globales.

El estudio realizado también contribuye en la creación de un camino hacia una economía sostenible, apoyando la transición energética hacia el uso de fuentes renovables, limpias y económicamente viables, repercutiendo de manera positiva en el ámbito económico y social.

4. Declaración de uso de herramientas de inteligencia artificial generativa

Por la presente, yo, Pablo Martínez Galindo, estudiante de Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado " Análisis predictivo multidisciplinar del precio del hidrógeno verde", 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. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Interpretador de código**: Para realizar análisis de datos preliminares.
4. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
5. **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: 20 de junio de 2024
Firma: Pablo Martínez Galindo

5. Referencias bibliográficas

Antoranz, J. L. (2022). El hidrógeno verde en la Unión Europea: una vía necesaria para la transición energética. Revista Española de Desarrollo y Cooperación, 48, 13-33. https://doi.org/10.5209/redc.81174

Banco Mundial. (2020, 22 octubre). El impacto de la COVID-19 sobre los mercados de productos básicos se hace notar principalmente en los precios de la energía; es probable que la demanda de petróleo se siga contrayendo después de 2021. World Bank. https://www.bancomundial.org/es/news/press-release/2020/10/22/impact-of-covid-19-on-commodity-markets-heaviest-on-energy-prices-lower-oil-demand-likely-to-persist-beyond-2021

Cervera, I., & Figuerola-Ferretti, I. (2024). Credit risk and bubble behavior of credit default swaps in the corporate energy sector. International Review Of Economics & Finance, 89, 702-731. https://doi.org/10.1016/j.iref.2023.07.033

Chadwick, L. (2021, 29 octubre). Crisis energética: ¿Por qué se disparan los precios del gas natural y cómo afectará a los europeos? Euronews. https://es.euronews.com/2021/10/21/crisis-energetica-por-que-se-disparan-los-precios-del-gas-natural-y-como-afectara-a-los-eu

Cohen, P. (2022, 2 julio). 'Podría tener consecuencias devastadoras': el alza del combustible amenaza la estabilidad social en varias regiones del mundo. The New York Times. https://www.nytimes.com/es/2022/07/02/espanol/precio-combustible.html

Daniel. (2023, 30 octubre). Random Forest: Bosque aleatorio. Definición y funcionamiento. Formación En Ciencia de Datos | DataScientest.com. https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento

Fabra, N. (2023). Reforming European electricity markets: Lessons from the energy crisis. Energy Economics, 126, 106963. https://doi.org/10.1016/j.eneco.2023.106963

Iberdrola. (2023). Breve (y eterna) historia del Hidrógeno. Iberdrola. Recuperado de https://www.iberdrola.es/blog/sostenibilidad/historia-hidrogeno-verde

Equipo Singular Bank. (2022, 23 febrero). TTF: precio de referencia del gas en Europa y en el mundo | Blog Singular. El Blog de SelfBank by Singular Bank. https://blog.selfbank.es/inflacion-ttf-precio-de-referencia-del-gas-en-europa-y-en-el-mundo/

Murillo, R., & Arcedalia, N. (2019). Análisis de validación cruzada bajo diferentes condiciones de ruido. http://51.143.95.221/bitstream/TecNM/810/1/Natalia%20Arcedalia%20Rodr%c3%adguez%20Murillo.pdf

Orea, S. V., Vargas, A. S., & Alonso, M. G. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Ene, 779(73), 33. https://www.academia.edu/download/34203825/e1.pdf

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. *Revista Seden*, *14*, 195-214. https://www.revistaseden.org/files/14-cap%2014.pdf

Rodó, P. (2021, 19 febrero). Contraste de Dickey-Fuller. Economipedia. https://economipedia.com/definiciones/contraste-de-dickey-fuller.html

Ruf, J., & Wang, W. (2019). Neural networks for option pricing and hedging: a literature review. arXiv preprint arXiv:1911.05620. https://arxiv.org/pdf/1911.05620

Scheibe, A., & Poudineh, R. (2023). Regulating the future European hydrogen supply industry: A balancing act between liberalization, sustainability, and security of supply? The Oxford Institute for Energy Studies, ET26. https://www.oxfordenergy.org/wpcms/wp-content/uploads/2023/10/ET26-Regulating-the-future-European-hydrogen-supply-industry-with-Exec-Summary.pdf

Segarra, I., Atanasova, C., & Figuerola-Ferretti, I. (2024). Electricity markets regulations: The financial impact of the global energy crisis. Journal Of International Financial Markets, Institutions & Money, 93, 102008. https://doi.org/10.1016/j.intfin.2024.102008

Sharma, S., & Escobari, D. (2018). Identifying price bubble periods in the energy sector. Energy Economics, 69, 418-429. https://doi.org/10.1016/j.eneco.2017.12.007

Soler, L. N. (2023, 24 julio). En las entrañas de una central de gas natural: así funciona el ciclo combinado para generar electricidad. Newtral. https://www.newtral.es/gas-natural-energia/20220627/

6. Anexo

Código 1 – *Random forest*: *print(rforest)*

```
> print(rforest)

Call:
 randomForest(formula = pdhid ~ pdbrent + pdcrude + pddutch, data = datosdim)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 50.75106
                    % Var explained: 94.68
```

Código 2 – Importancia *random forest: importance(rforest)*

```
> importance(rforest)
         IncNodePurity
pdbrent       134711.6
pdcrude       108834.6
pddutch       308672.2
```

Código 3 – Resumen regresión lineal: *summary(rlin)*

```
> rlin <- lm(pdhid ~ pddutch, data = datosdim)
> summary(rlin)

Call:
lm(formula = pdhid ~ pddutch, data = datosdim)

Residuals:
    Min      1Q  Median      3Q     Max
-39.068  -1.826  -0.626   1.115  45.560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.998901   0.487970   24.59   <2e-16 ***
pddutch      0.496849   0.004793  103.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.026 on 584 degrees of freedom
Multiple R-squared:  0.9485,    Adjusted R-squared:  0.9484
F-statistic: 1.075e+04 on 1 and 584 DF,  p-value: < 2.2e-16
```

Código completo

```
datos <- read_excel("excelprecios2.xlsx")
summary(datos)
View(datos)
str(datos)
fechas <- datos[[1]]
precioshid <- datos[[2]]
precioscrude <- datos[[3]]
preciosbrent <- datos[[4]]
preciosdutch <- datos[[5]]
#Analisis comparativo de precios: gráfico de caja y bigotes, gráfico de líneas, promedios,
desviaciones típicas, varianza, covarianza, correlaciones
#visualizacion de las variables
```

```
#boxplot
boxplot(preciosbrent, precioscrude, preciosdutch, precioshid,
        names = c("Petróleo crudo", "Petróleo americano", "Gas natural TTF", "Hidrógeno
verde"),
        main = "Boxplot precios",
        ylab = "Precio (€)",
        col = c("red", "khaki", "blue", "green"),
        outline = TRUE,
        cex.axis = 1.6,
        cex.lab = 1.6,
        cex.main = 2.0)
  legend("topleft", legend = c("Petróleo crudo", "Petróleo americano", "Gas natural TTF",
"Hidrógeno verde"),
       fill = c("red", "khaki", "blue", "green"),
       cex = 1.6)
#gráfico de línea
graficolinea <- data.frame(fechas = fechas, phid = precioshid, pcru = precioscrude,
                           pbre = preciosbrent, pdut = preciosdutch)

ggplot(graficolinea, aes(x = fechas)) +
  geom_line(aes(y = phid, color = "Hidrógeno verde")) +
  geom_line(aes(y = pcru, color = "Petróleo americano")) +
  geom_line(aes(y = pbre, color = "Petróleo crudo")) +
  geom_line(aes(y = pdut, color = "Gas natural TTF")) +
  labs(x = "Fecha", y = "Precio (€)", title = "Gráfico de línea: Evolución precios últimos 3 años")
+
  scale_color_manual(name = "Materia prima",
                     values = c("Hidrógeno verde" = "green", "Petróleo americano" = "khaki",
"Petróleo crudo" = "red", "Gas natural TTF" = "blue")) +
  theme(axis.text = element_text(size = 18),
        legend.text = element_text(size = 18),
        axis.title = element_text(size = 18),
        legend.title = element_text(size = 18),
        plot.title = element_text(size = 26, hjust = 0.5),
        legend.position = "right")
#Media, desviación típica, varianza, covarianza, correlación
mediabrent <- mean(preciosbrent, na.rm = TRUE)
```

```
mediacrude <- mean(precioscrude, na.rm = TRUE)
mediadutch <- mean(preciosdutch, na.rm = TRUE)
mediahid <- mean(precioshid, na.rm = TRUE)
desvbrent <- sd(preciosbrent, na.rm = TRUE)
desvcrude <- sd(precioscrude, na.rm = TRUE)
desvdutch <- sd(preciosdutch, na.rm = TRUE)
desvhid <- sd(precioshid, na.rm = TRUE)
varbrent <- var(preciosbrent, na.rm = TRUE)
varcrude <- var(precioscrude, na.rm = TRUE)
vardutch <- var(preciosdutch, na.rm = TRUE)
varhid <- var(precioshid, na.rm = TRUE)
datosdim <- read_excel("excelpreciosdimensionados.xlsx")
summary(datosdim)
View(datosdim)
str(datosdim)
fechasdim <- datosdim[[1]]
pdhid <- datosdim[[2]]
pdcrude <- datosdim[[3]]
pdbrent <- datosdim[[4]]
pddutch <- datosdim[[5]]
rdhid <- datosdim[[7]]
rddutch <- datosdim[[10]]



pnhid <- scale(pdhid)
pncrude <- scale(pdcrude)
pnbrent <- scale(pdbrent)
pndutch <- scale(pddutch)
#graficamos los datos normalizados para ver su aspecto...
#boxplot
boxplot(list(pnbrent, pncrude, pndutch, pnhid),
        names = c("Petróleo crudo norm.", "Petróleo americano norm.", "Gas natural TTF norm.",
"Hidrógeno verde norm."),
        main = "Boxplot precios normalizados últimos 3 años",
        ylab = "Precio normalizado",
        col = c("red", "khaki", "blue", "green"),
        outline = TRUE,
```

```
      cex.axis = 1.6,
      cex.lab = 1.6,
      cex.main = 2.0,
      ylim = c(-2, 5))
#gráfico de línea
graficolinea <- data.frame(fechasdim = fechasdim, pnhid = pnhid, pncrude = pncrude,
                           pnbrent = pnbrent, pndutch = pndutch)
ggplot(graficolinea, aes(x = fechasdim)) +
  geom_line(aes(y = pnhid, color = "Hidrógeno verde")) +
  geom_line(aes(y = pncrude, color = "Petróleo americano")) +
  geom_line(aes(y = pnbrent, color = "Petróleo crudo")) +
  geom_line(aes(y = pndutch, color = "Gas natural TTF")) +
  labs(x = "Fecha", y = "Precio normalizado", title = "Gráfico de línea: Evolución precios
normalizados últimos 3 años") +
  scale_color_manual(name = "Materia prima",
                     values = c("Hidrógeno verde" = "green", "Petróleo americano" = "khaki",
"Petróleo crudo" = "red", "Gas natural TTF" = "blue")) +
  theme(axis.text = element_text(size = 18),
      legend.text = element_text(size = 18),
      axis.title = element_text(size = 18),
      legend.title = element_text(size = 18),
      plot.title = element_text(size = 26, hjust = 0.5),
      legend.position = "right")
# se observa una fuerte correlación entre el hidrógeno y el Dutch TTF
#estudiamos covarianzas
covbrenthid <- cov(pnbrent, pnhid)
covcrudehid <- cov(pncrude, pnhid)
covdutchhid <- cov(pndutch, pnhid)
#barplot para ver las covarianzas
covarianzas <- c(covbrenthid, covcrudehid, covdutchhid)
nombres_covarianzas <- c("cov pet. crudo - hidrógeno", "cov pet. americano - hidrógeno", "cov
gas natural TTF - hidrógeno")
barplot(covarianzas, names.arg = nombres_covarianzas,
      main = "Covarianzas con el hidrógeno", xlab = "Materias primas", ylab = "Covarianza
precios",
      ylim = c(0, 1.1)) +
  text(x = 1:3, y = covarianzas + 0.02, labels = round(covarianzas, 2), pos = 3)
```

```
#ahora vemos cuales son las variables con menor covarianza con respecto al hidrógeno, para ver la más similar y por tanto mas adecuada para utilizar como base para predecir el precio futuro del hidrógeno
corbrenthid <- cor(pnbrent, pnhid)
corcrudehid <- cor(pncrude, pnhid)
cordutchhid <- cor(pndutch, pnhid)
#barplot para ver las correlaciones
correlaciones <- c(corbrenthid, corcrudehid, cordutchhid)
nombres_correlaciones <- c("cor ept. crudo - hidrógeno", "cor pet. americano - hidrógeno", "cor gas natural TTF - hidrógeno")
barplot(correlaciones, names.arg = nombres_correlaciones,
        main = "Correlaciones con el hidrógeno", xlab = "Materias primas", ylab = "Correlación precios",
        ylim = c(0, 1.1)) +
  text(x = 1:3, y = correlaciones + 0.02, labels = round(correlaciones, 2), pos = 3)


set.seed(123)
rforest <- randomForest(pdhid ~ pdbrent + pdcrude + pddutch , data = datosdim)
print(rforest)
importance(rforest)
varImpPlot(rforest, cex = 2.5, cex.main = 1.5)
#una vez seleccionada la utility a comparar para la prediccion ploteamos otra vez con las medias
#boxplot
boxplot(list(pddutch, pdhid),
        names = c("Gas natural TTF dim.", "Hidrógeno verde"),
        main = "Boxplot precios dimensionados últimos 3 años",
        ylab = "Precio (€)",
        col = c("blue", "green"),
        outline = TRUE,
        cex.axis = 1.6,
        cex.lab = 1.6,
        cex.main = 2.0,
        ylim = c(0, 400))
#gráfico de línea
graficolinea3 <- data.frame(fechasdim = fechasdim, pdhid = pdhid, pddutch = pddutch)
mediadimhid <- mean(pdhid)
```

```
mediadimdutch <- mean(pddutch)
ggplot(graficolinea3, aes(x = fechasdim)) +
  geom_line(aes(y = pdhid, color = "Hidrógeno verde")) +
  geom_line(aes(y = pddutch, color = "Gas natural TTF dim.")) +
  geom_line(aes(y = mediadimhid, color = "Media Hidrógeno"), linetype = "dashed") +
  geom_line(aes(y = mediadimdutch, color = "Media Gas natural TTF dim."), linetype =
"dashed") +
  labs(x = "Fecha", y = "Precio (€)", title = "Gráfico de línea: Evolución precios dimensionados
últimos 3 años") +
  scale_color_manual(name = "Materia prima",
                     values = c("Hidrógeno verde" = "green", "Gas natural TTF dim." = "blue", "Media
Gas natural TTF dim." = "blue", "Media Hidrógeno" = "green"))+
  theme(axis.text = element_text(size = 18),
        legend.text = element_text(size = 18),
        axis.title = element_text(size = 18),
        legend.title = element_text(size = 18),
        plot.title = element_text(size = 26, hjust = 0.5),
        legend.position = "right")
#regresión lineal
rlin <- lm(pdhid ~ pddutch, data = datosdim)
summary(rlin)
#hacemos el test de Dickey-Fuller a residuos y precios para analizar si los residuos son
estacionarios y el análisis es bueno
residuos <- rlin$residuals
adf_residuos <- adf.test(residuos)
print(adf_residuos)
adf_pdhid <- adf.test(pdhid)
print(adf_pdhid)
coefdutch <- coef(rlin)["pddutch"]
#importamos el precio de los futuros del Dutch TTF Gas
datosfuturos <- read_excel("futurosdutch.xlsx")
summary(datosfuturos)
View(datosfuturos)
str(datosfuturos)
fechasfut <- datosfuturos[[1]]
pfutdutch <- datosfuturos[[2]]
```

```
#sacamos los coeficientes dutch, es decir, las magnitudes de cambio esperadas en el precio del hidrógeno por cada unidad de cambio en el precio del dutch (que es la variable predictora).
pfuthid <- coefdutch*pfutdutch
#lo ploteamos en un gráfico para ver el resultado de las predicciones...
graficolinea4 <- data.frame(fechasfut = fechasfut, pfuthid = pfuthid, pfutdutch = pfutdutch)
mediafuthid <- mean(pfuthid)
mediafutdutch <- mean(pfutdutch)
ggplot(graficolinea4, aes(x = fechasfut)) +
  geom_line(aes(y = pfuthid, color = "Futuros Hidrógeno verde")) +
  geom_line(aes(y = pfutdutch, color = "Futuros Gas natural TTF")) +
  geom_line(aes(y = mediafuthid, color = "Media futuros hidrógeno verde"), linetype =
"dashed") +
  geom_line(aes(y = mediafutdutch, color = "Media futuros Gas natural TTF"), linetype =
"dashed") +
  labs(x = "Fecha", y = "Precios", title = "Gráfico de línea: Evolución precio futuros") +
  scale_color_manual(name = "Materia prima", values = c("Futuros Hidrógeno verde" =
"green", "Futuros Gas natural TTF" = "blue", "Media futuros Gas natural TTF" = "blue",
"Media futuros hidrógeno verde" = "green"))+
  theme(axis.text = element_text(size = 18),
        legend.text = element_text(size = 18),
        axis.title = element_text(size = 18),
        legend.title = element_text(size = 18),
        plot.title = element_text(size = 26, hjust = 0.5),
        legend.position = "right")
#KNN
set.seed(123)
index <- sample(1:nrow(datosdim), 0.8 * nrow(datosdim))
traindata <- datosdim[index, ]
testdata <- datosdim[-index, ]
modelknn <- knn(train = traindata[, 3:5],
                test = testdata[, 3:5],
                cl = traindata[[2]], k = 5
modelknn <- as.numeric(as.character(modelknn))
fechas_prediccion <- datosdim[-index, 1]
#preparo el dataframe fechas y precios con las predicciones del modelo KNN
prediction2 <- data.frame(fecha = fechas_prediccion, predprecio = modelknn)
prediction2[[1]] <- as.Date(prediction2[[1]])
```

```
datosdim[[1]] <- as.Date(datosdim[[1]])
#preparo el dataframe con los precios reales del hidrógeno para las fechas aleatoriamente
predichas por el modelo KNN indexando filas
indices <- which(datosdim[[1]] %in% prediction2[[1]])
precios_reales_hidrogeno <- datosdim[[2]][indices]
resultados <- data.frame(fechas = datosdim[[1]][indices], precio_h2 =
precios_reales_hidrogeno)

df_combinado <- data.frame(
  fechas = resultados[, 1],
  precio_real = resultados[, 2],
  precio_predicho = prediction2[, 2]
)
print(df_combinado)
#ploteamos el dataframe combinado para ver las diferencias entre el precio real y el precio
predicho
ggplot(data = df_combinado, aes(x = fechas)) +
  geom_line(aes(y = precio_real, color = "Precio Real"), size = 1) +
  geom_line(aes(y = precio_predicho, color = "Precio Predicho"), size = 1) +
  labs(x = "Fecha", y = "Precio", title = "Precios Reales vs. Predichos") +
  scale_color_manual(values = c("Precio Real" = "blue", "Precio Predicho" = "red"),
                     name = "Tipo de Precio") +
  theme_minimal() +
  theme(legend.position = "top")
ggplot(data = df_combinado, aes(x = fechas)) +
  geom_line(aes(y = precio_real, color = "Precio real h. verde"), size = 1.5) +
  geom_line(aes(y = precio_predicho, color = "Precio predicho h. verde"), size = 1.5) +
  labs(x = "Fecha", y = "Precio (€)", title = "Precios reales contra precios predichos") +
  scale_color_manual(name = "Comparación", values = c("Precio real h. verde" = "blue",
"Precio predicho h. verde" = "red"))+
  theme(axis.text = element_text(size = 18),
        legend.text = element_text(size = 18),
        axis.title = element_text(size = 18),
        legend.title = element_text(size = 18),
        plot.title = element_text(size = 26, hjust = 0.5),
        legend.position = "right")
#Cálculo técnico de la efectividad del modelo KNN
```

```
#MSE: Mean Squared Error:
mse <- mean((df_combinado[, 2] - df_combinado[, 3])^2)
rmse <- sqrt(mean((df_combinado[, 2] - df_combinado[, 3])^2))
#hacemos cross validation 10 veces
set.seed(123)
cv <- train(
  precio_real ~ precio_predicho,
  data = df_combinado,
  method = "lm",  # Usamos regresión lineal para calcular el RMSE
  trControl = trainControl(method = "repeatedcv", number = 10, repeats = 5)
)
intervalo_confianza <- cv$results[c("RMSE", "RMSESD")] #aqui el RMSE es diferente, no se
toma el calculado antes sino que la cross validación calcula el suyo propio
# Imprimir los resultados
print("Intervalo de confianza del 95% para el RMSE:")
print(intervalo_confianza)
#MAE Y MAPE
mae <- mean(abs(df_combinado$precio_real - df_combinado$precio_predicho))
mape <- mean(abs((df_combinado$precio_real - df_combinado$precio_predicho) /
df_combinado$precio_real)) * 100
print(paste("MAE:", round(mae, 2)))
print(paste("MAPE:", round(mape, 2)))
```