COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas

¿Existe una prima verde (*greenium*) en el mercado de bonos corporativos de la U.E.? Evidencia empírica con modelos de Machine Learning

Autor: María Sánchez García

Tutor: María Coronado Vaca

MADRID | Junio 2025

RESUMEN

En un mundo donde la sostenibilidad está en el eje central de las preocupaciones sociales, ésta ha logrado constituirse como una guía en la integración de las prácticas empresariales. De esta manera, el beneficio económico ya no es el único objetivo, sino lograr un beneficio cuidando su impacto. En este contexto, en el año 2007 surge un nuevo instrumento financiero,los bonos verdes, diseñado para financiar exclusivamente los proyectos sostenibles y facilitar la movilidad de recursos de financiación hacia los proyectos de inversión verde. La idea original es que la rentabilidad de dichos bonos en la fecha de emisión sea menor que la de los bonos marrones (y por tanto su precio mayor que el de los bonos marrones), de manera que el coste de financiar proyectos verdes sea menor para el emisor. A dicha diferencia de rentabilidades se le denomina prima verde o greenium y por tanto, lo inicialmente previsto cuando se introdujeron los bonos verdes en los mercados financieros era que la greenium fuese negativa. Pero a día de hoy aún sigue abierto en la literatura académica el debate sobre si existe una prima verde (greenium) y, sobre el signo de esta. Se trata de un tema muy investigado pero para el cual la evidencia empírica no es unánime. Aunque la evidencia existente en la actualidad es mayoritaria para una greenium negativa, la diversidad de resultados sobre si la greenium es positiva o negativa depende en gran medida de las características de la muestra de bonos analizados en cuanto al emisor (soberanos, corporativos, etc.), geografía (EE. UU., China, Europa, etc.), periodo temporal, mercado (primario o secundario o ambos), etc. Por otro lado, la inmensa mayoría de los estudios publicados utilizan métodos econométricos tradicionales, siendo muy escasos los que utilizan modelos de ML. Por ello, nuestro estudio pretende contribuir a dicho debate aportando nueva evidencia empírica sobre el tema, en concreto sobre los bonos corporativos de la UE durante el periodo 2014-2025, aplicando técnicas de ML (random forest y gradient boosting). Para ello, analizaremos el comportamiento de los bonos corporativos en la UE, tanto verdes como no verdes, durante el periodo 2014-2025. Se obtiene una prima verde negativa, pequeña pero estadísticamente significativa, de 2.9927 puntos básicos en nuestra muestra. Este trabajo además de aportar mayor evidencia empírica que orienta el debate, también puede ser informativa para el inversor.

Palabras clave: Bonos verdes, greenium, prima verde, Yield Spread to Maturity, Random Forest, predicción, rentabilidad, Unión Europea (UE), ESG (Environmental, Social, Governance).

ABSTRACT

In a world where sustainability is at the core of social issues, it has become a key principle for incorporating business practices. Consequently, economic profit isn't the only goal anymore; instead, the focus is on achieving a return while considering its consequences. In this setting, a new financial tool—green bonds—was developed in 2007 specifically to fund sustainable initiatives and direct resources toward environmentally friendly investments. The initial concept was that the yield on these bonds at issuance would be less than that of "brown" bonds (and thus their price higher), resulting in reduced financing costs for green projects for the issuer. The disparity in yields became known as the "green premium," or greenium, and when green bonds were launched in financial markets, it was anticipated that the greenium would turn out to be negative. To this day, though, the scholarly literature still discusses whether a green premium truly exists and, if it does, what its value is. It is a well-explored subject, yet the empirical findings are not consistent. While most recent research indicates a negative greenium, the diverse results—whether positive or negative—are heavily influenced by the characteristics of the analyzed bond sample: type of issuer (sovereigns, corporates, etc.), location (U.S., China, Europe, etc.), time period, market (primary, secondary, or both), and more. Furthermore, most published research utilizes conventional econometric techniques; only a small number utilize machine-learning models. This research seeks to enhance this discussion by offering new empirical data—focused on EU corporate bonds from 2014 to 2025—employing ML methods (random forest and gradient boosting). We examine the performance of green and non-green EU corporate bonds from 2014 to 2025 and discover a minor yet statistically significant negative green premium of 2.9927 basis points in our analysis. In addition to supplying further empirical evidence to inform the debate, this work may also prove informative for investors.

Key Words: Green bonds, greenium, green yield, Random Forest, prediction, yield, European Union (EU), ESG (Environmental, Social, Governance).

Tabla de contenido

ÍNDICE DE GRÁFICOS E ILUSTRACIONES

1. INTRODUCCIÓN

Durante los últimos años, el paradigma global ha tomado unas nuevas direcciones. La falta de gobernanza ética, la importancia por la sostenibilidad y las desigualdades sociales, han hecho que a nivel social y empresarial haya una mayor preocupación por sus prácticas. Actualmente, las empresas no buscan solo el beneficio económico, sino que, ahora, la responsabilidad por generar impacto medioambiental es casi más importante que el recurso financiero (La Torre & Leo, 2024). Es decir, se ha creado una nueva conciencia y/o pensamiento en el que la empresa busca maximizar el bienestar. Ejemplo de ello son los compromisos que se adquieren en los marcos de Responsabilidad Social Corporativa (RSC) y Environmental Social and Governance (ESG) con las que las empresas guían sus prácticas.

El año 2007 es una fecha que marca el inicio de un comienzo sostenible. En este año, el Banco Europeo de Inversiones (BEI) emite el primer bono verde conocido hasta la fecha a partir del Climate Awareness Bond con la finalidad de invertir en recursos para respaldar proyectos energéticos. A este, le sigue el Banco Mundial en 2008 que, también emite su primer bono verde, buscando reducir los efectos del cambio climático . Aunque este nuevo tipo de bono no supera las emisiones de los bonos no verdes o marrones (según datos de mercado de Climate Bonds Initiative (2024), los bonos calificados como verdes según dicha taxonomía, representaron el 2,4% del total de las emisiones globales de bonos durante el tercer cuatrimestre de 2024), podemos decir que ha cobrado importancia a lo largo de los años. Reflejo de ello es su crecimiento en los mercados internacionales donde, mientras anteriormente eran casi inexistentes, el volumen acumulado de emisiones de bonos verdes (según la taxonomía de Climate Bonds Initiative) en todo el mundo desde sus inicios hasta noviembre de 2024 alcanza la cifra de 1,04 billones de USD (Climate Bonds Initiative, 2024).

Sin embargo, éste no representa un instrumento financiero más. No solo busca financiar proyectos sostenibles, sino que, es también el esfuerzo por integrar la sostenibilidad en los sectores económicos y de inversión. Además, su novedad nos hace plantearnos si estos bonos se comportan como los bonos tradicionales o marrones. Por el momento, la cuestión que guía este trabajo y que también se encuentra en debate, es la posibilidad de la existencia de la prima verde o *greenium* (en terminología anglosajona). Es decir, si existe un diferencial (spread) en la rentabilidad de los bonos verdes con respecto a la

deuda no verde de características similares, que llamaremos "prima verde" y si ese diferencial es positivo o negativo. En definitiva, si la financiación vía bonos verdes tiene menor coste para los emisores que la financiación a través de bonos marrones, como resultado de la menor rentabilidad que están dispuestos a obtener los inversores con interés en financiar el impacto positivo social y medioambiental.

En este contexto, el trabajo plantea una investigación con datos reales sobre los bonos verdes corporativos de la UE (mercado primario), con la intención de dar respuesta a la hipótesis de la existencia de la prima verde. Mediante el uso de técnicas de Machine Learning, consideraremos si esta prima es positiva o negativa para el inversor y el emisor.

Por último, constituye una doble perspectiva que aporta valor. En primer lugar y de manera más obvia, aporta conocimiento y posibles resultados para el debate que existe sobre los bonos verde, pero, además, puede ser un estudio que revele información importante para los inversores en este momento donde la sostenibilidad no es una opción.

1.1. OBJETIVOS

El objetivo principal del estudio es **determinar si existe una prima verde o greenium en los bonos corporativos de la Unión Europea (UE), mediante el uso de modelos de Machine Learning (ML).** Se trata por tanto de dar respuesta a dicha pregunta de investigación principal mediante la predicción, con ML, de las rentabilidades de ambos tipos de bonos, verdes y marrones y el análisis de sus diferencias.

Y, además, la estructura de análisis vendrá guiada por objetivos más específicos. Éstos dan respuesta a las siguientes preguntas:

- También este trabajo propone analizar cómo se comporta esa prima en el tiempo (en caso de haberla): ¿Es positiva o negativa? Mediante la selección de algunas variables de los bonos como el sector, la geografía y el rating, se podrá observar si alguno de estos factores influye en la aparición de esa prima verde y, por tanto, cuál de ellos es más influyente.
- El último objetivo tiene como finalidad estudiar qué factores son los que determinan esa supuesta prima verde. Para ello, usaremos tres modelos de ML, con una

combinación de variables, numéricas y cualitativas y escogeremos aquél que mejor ajuste y que mejor prediga las rentabilidades.

1.2. JUSTIFICACIÓN DEL TEMA

El tema de cómo financiar la transición hacia una economía descarbonizada es de gran relevancia si se quiere facilitar la consecución del objetivo de cero emisiones contaminantes, ya que dicha transición requiere de unas cantidades de financiación de enorme cuantía (ICMA, 2023; Chesini, 2024). Los bonos verdes nacieron como uno de los principales facilitadores a la hora movilizar los recursos financieros hacia inversiones limpias y sostenibles, pues la financiación obtenida mediante la emisión de bonos verdes solo puede ser destinada a proyectos o inversiones verdes. Pero su emisión lleva implícita una serie de costes indirectos, por ejemplo administrativos por la necesidad de obtener la calificación explicita de bono verde según alguna de las dos taxonomías principales existentes, como son el Climate Bond Standard emitido por la Climate Bonds Initiative (CBI, 2024) y los Green Bond Principles de la International Capital Market Association (ICMA, 2021). Si frente a ese incremento de costes de emisión no existieran otras ventajas añadidas en términos de coste de emisión para las empresas emisoras de bonos verdes, su emisión se vería entonces dificultada en comparación con la emisión de bonos parecidos pero no verdes (o marrones) y con ella la transición hacia una economía sostenible (Flammer, 2021).

La pregunta que se plantea es ¿están los inversores concienciados con la sostenibilidad del planeta dispuestos a obtener a cambio una menor rentabilidad al prestar su financiación a las empresas que la solicitan vía emisión de bonos, para emprender proyectos verdes? Si ese fuera el caso, el coste de financiación para las empresas emisoras de bonos verdes sería menor; y la emisión de bonos verdes constituiría una ventaja para sus emisores, al presentar una TIR (tasa interna de rentabilidad o YTM en terminología anglosajona) de emisión menor que la existente para sus bonos marrones equivalentes. Ese diferencial de rentabilidades entre bonos verdes y marrones es lo que se denomina prima verde o greenium y lo habitual en la literatura académica es definirla como la resta de la rentabilidad verde menos la rentabilidad marrón. Si bien podría definirse justo al revés, en este trabajo vamos a seguir esa misma línea. De esta manera, diremos que existe greenium positiva si los bonos verdes tienen más rentabilidad a fecha

de emisión que los marrones, en cuyo caso, la emisión de bonos verdes no representa una ventaja en términos de financiación para sus emisores frente a la emisión de bonos equivalente pero marrones. Y, cuando la rentabilidad a fecha de emisión de los bonos verdes sea menor que la de los marrones, diremos que la greenium es negativa, resultando entonces más conveniente para las empresas, emitir bonos verdes que no verdes, por su menor coste de financiación. Como vemos, denominarla positiva o negativa no implica que sea favorable o desfavorable (porque además dependerá para quién, el emisor o el inversor), sino que simplemente es el signo que resulta de la definición que hemos adoptado (rentabilidad verde menos rentabilidad marrón). Es decir, si decimos greenium positiva, nos referiremos a que la rentabilidad de los bonos verdes es mayor (y por tanto su precio menor) que la rentabilidad de los bonos marrones, y entonces una greenium positiva es favorable para el inversor porque obtiene más rentabilidad, pero desfavorable para el emisor, porque le supone un mayo coste de financiación.

Existe en la literatura académica un debate aún inconcluso sobre si existe una prima verde (greenium) y, sobre el signo de esta. Se trata de un tema muy investigado pero para el cual la evidencia empírica no es unánime, como se expondrá en el apartado de revisión de la literatura. Aunque la evidencia existente en la actualidad es mayoritaria para una greenium negativa (es decir, favorable para la emisión de bonos verdes por su menor coste de emisión para las instituciones emisoras), la diversidad de resultados sobre si la greenium es positiva o negativa depende en gran medida de las características de la muestra de bonos analizados en cuanto al emisor (soberanos, corporativos, que incluyan o no el sector financiero, etc.), geografía (EE.UU., China, Europa, etc.), periodo temporal, mercado (primario o secundario o ambos), etc. Por otro lado, la inmensa mayoría de los estudios publicados utilizan métodos econométricos tradicionales, siendo muy escasos los que utilizan modelos de ML. Por ello, nuestro estudio pretende contribuir a dicho debate aportando nueva evidencia empírica sobre el tema, en concreto sobre los bonos corporativos de la UE durante el periodo 2014-2025, aplicando técnicas de ML.

1.3. METODOLOGÍA GENERAL DEL TFG

La metodología de este trabajo tiene un enfoque cuantitativo, mediante la aplicación de técnicas de Business Analytics para el preprocesamiento y visualización de los datos obtenidos de la plataforma de información financiera LSEG (antes Refinitiv) y el uso de

tres modelos predictivos de ML para la rentabilidad de los bonos: regresión lineal, random forest y gradient boosting. El proceso metodológico resumido (pues se expondrá con detalle en el apartado 3.2) consta de los siguientes pasos: una vez obtenidos los datos, se realiza un preprocesamiento y análisis exploratorio minucioso de los mismos. A continuación, se separa la muestra total ya limpia y preprocesada en las dos submuestras de bonos verdes y bonos no verdes (marrones). Se entrenan los tres modelos mencionados para predecir la prima de riesgo de los bonos marrones y se evalúa su desempeño, tanto su capacidad explicativa (con las métricas R^2 y R^2 ajustado), como su capacidad predictiva (a través del MAE y RMSE). Ello permitirá seleccionar el mejor modelo que mejor predice de los tres, con el cual ya se realizará la predicción de la prima de riesgo de los bonos verdes. Se determina la prima verde como la resta de las medias de ambas rentabilidades predichas (verde menos marrón) y se testa la significatividad estadística de la greenium estimada mediante dos contrastes de diferencias de medias: uno sobre si las medias son iguales o no (testamos la existencia o no de la prima verde) y otro sobre cuál de ambas medias de rentabilidad es mayor (testamos el signo positivo o negativo de la prima verde).

1.4. ESTRUCTURA

El trabajo académico mezcla dos ámbitos: las finanzas sostenibles y la aplicación de algoritmos de Machine Learning. Por eso, el estudio académico está centrado en dar respuestas para ambos campos. En primer lugar, es necesario crear un pequeño contexto sobre los instrumentos financieros verdes ya que, es novedoso. Además, se complementa con una pequeña reflexión sobre el estado de la cuestión: ¿Hay algún estudio que haya explorado esta hipótesis?

En segundo lugar, damos respuestas a nuestros objetivos de manera empírica. La segunda parte del trabajo está dirigida a la creación del modelo de ML que sea capaz de darnos respuestas y crear nuevas hipótesis de investigación. Para ello, también se hace un trabajo previo de explicación sobre las variables seleccionadas para el estudio. Así, se comparte todo el proceso por el cual obtenemos unas conclusiones que, finalmente estudiaremos. Como todo estudio, tiene limitaciones, que se pueden convertir en futuras líneas de investigación y como tal las exponemos en ese mismo apartado de conclusiones.

2. MARCO TEÓRICO Y REVISIÓN DE LA LITERATURA

2.1. MARCO TEÓRICO

Actualmente existe una mayor preocupación por la sostenibilidad de la que había hace tan solo unos años. Esto, ha hecho que, tanto las prácticas sociales, como económicas, se transformen siguiendo como hilo conductor cumplir con los principios sostenibles (Gibson Brandon et al., 2022). En medio de esta situación, surge el concepto de Finanzas sostenibles, con la finalidad de estructurar el flujo de capitales que explicaremos a continuación (CNMV, s.f.). Este concepto surge en 1992 gracias a la Iniciativa de Finanzas del Programa de Naciones Unidas para el Medio Ambiente (PNUMA) que, tenía como objetivo dirigir la inversión privada en proyectos sostenibles (Naciones Unidas, 2018). Por otro lado, desde los Acuerdos de París de 2015, se han creado numerosas iniciativas de regulación para conseguir los objetivos de 2030 entre los que destacan, la emisión baja o neutral y los recursos sostenibles (CNMV, s.f.).

Las finanzas sostenibles, en realidad, carecen de una definición universal. Por ejemplo, la Comisión Europea ha expresado su preocupación, en tanto que, pide un concilio ante la gran multitud de definiciones. También, han generado debate por la dificultad de estandarizar sus criterios. Esto ha provocado que algunos gobiernos e instituciones financieras las definan según sus intereses generando así confusiones, tanto a nivel académico como a la hora de ponerlo en práctica (European Commission, 2020). No obstante, sí hay instrumentos financieros que se le asocian ejemplo de ello son los bonos verdes que pasaremos a definir a continuación.

Marco Migliorelli en su publicación: *What Do We Mean by Sustainable Finance?* (2021), tiene como idea la redefinición de las Finanzas Sostenibles como el instrumento que haga cumplir los principios de sostenibilidad y ESG (Environmental, Social, Governance), así como se intuía en la definición de la CNMV. Además, cree necesario que esta definición debe estar situada en torno a la identificación de los ámbitos más importantes de la sostenibilidad (por eso hace referencia a ESG y ODS-Objetivos de Desarrollo Sostenible-) y a la vez, recopilar información sobre las aportaciones de los sectores económicos a la sostenibilidad para poder mejorar sus prácticas.

Una de las herramientas o instrumentos más importantes de las finanzas sostenibles son los conocidos como bonos verdes (La Torre & Leo, 2024). Según la definición que se intuye de las ideas anteriores, los bonos verdes constituyen un papel muy importante en la estructuración del camino de las finanzas hacia la sostenibilidad, teniendo un papel de inversión importante para crear proyectos con huella social. Los bonos verdes son instrumentos de renta fija. Al igual que los semejantes bonos marrones, son técnicas que permiten financiar proyectos y acceder a los mercados financieros internacionales. La diferencia entre los bonos verdes y lo bonos marrones es la finalidad de uso. Mientras que en bonos marrones el emisor no debe especificar cuál es el fin de inversión de los recursos obtenidos con dicha emisión, los bonos verdes tienen como obligación un fin social, ambiental o de gobierno, es decir, que estén destinados a la financiación del beneficio sostenible (Manzano Romero et al., 2024).

Existen dos taxonomías o estándares principales para definir un bono como verde. Por un lado, la catalogación como tal que realiza la ICMA en su documento Green Bond Principles (ICMA, 2021) y por otro el estándar que establece la CBI (Climate Bonds Initiative, 2024). Ambos documentos se pueden observar tanto como una guía para el inversor en su proceso de comprar de bonos verdes como una guía para el emisor de bonos verdes. Ambos son de carácter voluntario.

Los *Green Bond Principles (GBP)* tienen como objetivos facilitar el camino de compra para los inversores, la transparencia y la integración del mercado de los bonos verdes que introduzca mayores proyectos verdes. Además para que sea posible la emisión de bonos verdes, establecen unos requisitos que se deben cumplir. En primer lugar, el fin de los ingresos como ya se ha mencionado y, por otro lado debe haber un procedimiento clave evaluación, seguimiento y de reporte que verifique el fin del bono (ICMA, 2021). Según ICMA (2021), existen cuatros tipos de bonos verdes por el momento. En primer lugar, el bono verde estándar. Este primer tipo de bono funciona como cualquier bono normal o marrón, es decir, el emisor se compromete a devolver la inversión mediante sus propios recursos, pero este debe ser destinado seguro los proyectos verdes. En segundo lugar, el bono de ingresos. Este tipo de bono tiene como peculiaridad que el pago depende de los ingresos o flujos de caja específicos de los que se destina la financiación. En caso de no generar ingresos, el inversor no recupera su dinero. En tercer lugar, el bono de proyecto verde, este dependerá de la rentabilidad del proyecto. De esta manera el riesgo del inversor dependerá del beneficio del proyecto. Por último, el bono verde garantizado. En

este se encuentran dos tipos de bonos: el colateral, el inversor recuperará el dinero invertido aunque el proyecto cree beneficios y, el estándar garantizado del que forma parte de una práctica de la empresa respaldado con otros proyectos verdes.

Por tanto, cada emisor deberá especificar el tipo de bono verde para proteger al inversor y debe clarificarse en toda la documentación del bono. Además, no podrán utilizar varios bonos para la financiación de los proyectos y por supuesto, deben seguir los principios GBP (ICMA, 2021).

En esta misma línea, es necesario exponer uno de los conceptos más importantes de los bonos verdes, la greenium, sobre cuya existencia y signo existe mucho debate, como expondremos en el apartado 2.2. Revisión de la literatura. El concepto de greenium es la combinación entre preemium y green. Como ya hemos adelantado en la introducción, el diferencial de rentabilidades entre bonos verdes y marrones es lo que se denomina prima verde o greenium y lo habitual en la literatura académica es definirla como la resta de la rentabilidad verde menos la rentabilidad marrón (MacAskill et al., 2021). Si bien podría definirse justo al revés, en este trabajo vamos a seguir esa misma línea. De esta manera, diremos que existe greenium positiva si los bonos verdes tienen más rentabilidad a fecha de emisión que los marrones, en cuyo caso, la emisión de bonos verdes no representa una ventaja en términos de financiación para sus emisores frente a la emisión de bonos equivalente pero marrones. Y, cuando la rentabilidad a fecha de emisión de los bonos verdes sea menor que la de los marrones, diremos que la greenium es negativa, resultando entonces más conveniente para las empresas, emitir bonos verdes que no verdes, por su menor coste de financiación. Como vemos, denominarla positiva o negativa no implica que sea favorable o desfavorable (porque además dependerá para quién, el emisor o el inversor), sino que simplemente es el signo que resulta de la definición que hemos adoptado (rentabilidad verde menos rentabilidad marrón). Es decir, si decimos greenium positiva, nos referiremos a que la rentabilidad de los bonos verdes es mayor (y por tanto su precio menor) que la rentabilidad de los bonos marrones, y entonces una greenium positiva es favorable para el inversor porque obtiene más rentabilidad, pero desfavorable para el emisor, porque le supone un mayo coste de financiación. Porque la idea original cuando se crearon los bonos verdes era reducir el coste de financiación del emisor (Di Tomasso et al., 2021), gracias a que el inversor concienciado con la sostenibilidad del planeta está dispuesto a obtener una menor rentabilidad por invertir en ellos (es decir, por

financiar proyectos de inversión verdes). Pero, existe controversia no solo sobre la existencia a o no de dicha greenium, sino también sobre su signo. Es decir, sobre si ésta es negativa (rentabilidad de los bonos verdes menor y por tanto su precio mayor) y de vedad se está consiguiendo el objetivo de minorar el coste de financiación de los emisores de bonos verdes.

Los tres conceptos que acabamos de exponer (finanzas sostenibles, bonos verdes y greenium) siguen una línea común y, es que, convergen en un mismo punto: salvaguardar la sostenibilidad y, por tanto hacer cumplir los Objetivos de Desarrollo Sostenible (ODS) según la agenda 2030 (Acuerdos de París) (Bhutta et al., 2022).

Esta estructura comienza desde la Responsabilidad Social Corporativa (RSC) que tiene como objetivo la aplicación de prácticas sostenibles para las empresas. Estas prácticas deben cumplirse respecto a los criterios ambientales, sociales, de gobierno, y en sus decisiones empresariales. Aunque es de carácter voluntario, debido a la preocupación por la sostenibilidad, muchas empresas han creado prácticas para integrar en sus objetivos estratégicos los beneficios sociales (Liang & Renneboog, 2020).

Debido a la falta de estructura clara en la línea de acción de RSC, surge un nuevo concepto: ESG. Estas siglas que recogen Environmental, Social and Governance, siguen los mismos ámbitos objetivo que tenía la RSC. Una de las diferencias clave de este concepto es que se comienza a establecer unas métricas de medición del impacto, que se han podido utilizar para comparar las empresas (Zervoudi et al., 2025).

Por último, en el marco teórico de este estudio de doble ámbito (finanzas sostenibles y ML), es necesario mencionar el papel del Machine Learning. Según definen Jordan & Mitchell (2015):

> "el ML se ocupa de diseñar y analizar algoritmos que aprenden a partir de los datos, buscando construir sistemas que mejoren su rendimiento en tareas específicas mediante la experiencia."

La evolución de la tecnología ha hecho que la capacidad las técnicas de ML sean cada vez más completas, capaces de crear grandes modelos beneficiosos para todos los ámbitos. En este estudio, consideramos la aplicación de un modelo predictivo de ML en el ámbito financiero.

2.2. REVISIÓN DE LA LITERATURA

La variable principal que guía la investigación, la greenium, es objeto de un gran debate actual y en la literatura previa existente. La multitud de estudios existente que analizan su existencia y sus signo, presentan una gran variabilidad de resultados tanto en cuanto a la existencia o no de la misma, como a su signo (positiva o negativa).

Los resultados de los estudios hasta 2019 están resumidos en la revisión sistemática de MacAskill et al., (2021), quienes concluyen que existe consenso sobre la existencia de prima verde en el 56% de los artículos que analizan el mercado primario y en el 70% de los trabajos que analizan el secundario, especialmente en el caso de bonos soberanos y con rating creditico en grado de inversión. Concluyen también que la cuantía de la prima (y su signo) varía muy ampliamente en el caso del mercado primario, si ben en el mercado secundario se puede acotar a un rango entre -1 y -9 puntos básicos (p.b.).

Con posterioridad a 2019, fecha límite de análisis de dicha revisión sistemática, se han seguido publicando también un número muy elevado de estudios que continúan investigando sobre el tema. Y de igual modo, los resultados son muy variables. La diversidad de resultados depende en gran medida de las características de la muestra de bonos analizados en cuanto a) el emisor: soberanos (Hinsche, 2021; Di Tomasso et al., 2024; Chesini, 2024, Ando et al., 2024), corporativos (Meyer & Henide, 2020), municipales (Larcker & Watts, 2020), que incluyan o no el sector financiero (Fatica et al., 2021; Apergis et al., 2024), etc.; b) de la geografía: EE.UU. (Caramichael & Rapp, 2024), China (Li et al., 2024), Europa (Sergei & Alesya, 2022; Hinsche, 2021; Chesini, 2024), UK (Tran, 2025), global internacional (Löffler et al., 2021; Fatica, et al. 2021), Colombia (Bonilla et al., 2023), etc.); c) mercado primario (Sergei & Alesya, 2022) o secundario (Meyer & Henide, 2020) o ambos (Löfffler et al., 2021); d) periodo temporal, etc. Por otro lado, todos los estudios publicados utilizan métodos eonométricos tradicionales, siendo muy escasos los documentos de trabajo sin publicar que utilizan modelos de ML.

La gran mayoría de los estudios, confirman la existencia de la greenium y que ésta es negativa. Por citar solo algunos: Zerbib (2019), de - 2 pb; Gianfrate & Peri (2019), de - 18 pb; Löffler et al. (2021), de - 0, 7 pb; de entre -15 y -20 pb; Hinsche (2021), de - 0,7 pb; Meyer & Henide (2020), de - 1,84 pb; Sergei & Alesya (2022), de – 4 pb; Barclays (2022), de entre – 5 y -8 pb. Karpf & Mandel (2018) encontraron que, si bien la prima era

negativa en su periodo considerado (2010-2016), ésta cambiaba de signo a lo largo de los años: la prima era positiva entre 2010-2014 y negativa en 2015 y 2016.

Aunque menos, también existen estudios que niegan la existencia de la greenium, como por ejemplo Larcker & Watss (2020), Hyun et al. (2020), Fatica et al. (2021).

Todos los estudios hasta ahora mencionados aplican técnicas econométricas tradicionales y solo hemos encontrado tres artículos que aplican ML (Fu et al., 2024; Kocaarslan, 2024; Kocaarslan & Soytas, 2023) para analizar la greenium. Fu et al. (2024), analizaron el impacto de las noticias financieras en los mercados en China. Para ello, usaron la técnica de análisis de sentimiento sobre 15.000 noticias desde 2008. Los hallazgos del estudio son muy interesantes. Los resultados mostraron que los días en los que las noticas detectaban palabras relacionadas con la sostenibilidad, la prima de los bonos verdes se reducía, mientras que, aquellos días en los que se detecta palabras negativas, la prima volvía a subir. La conclusión del estudio muestra como las variables más influyente en el mercado chino no depende de las características del bono, sino, de la calidad informativa. Aunque no tiene el mismo objetivo Kocaarslan (2024) utiliza árboles de decisión entrenados con los datos de los bonos verdes (en este caso estadounidenses). Este análisis ha predicho mejores resultados que los modelos más básicos (Regresión lineal) en los momentos donde hay picos de liquidez (. De manera más similar a nuestra investigación en el año 2023 se publica un estudio predictor de los bonos verdes. Esta investigación usa algoritmos de redes neuronales combinados con random forest, modelos robustos que alcanzaron a predecir un 75% del comportamiento de los bonos (Kocaarslan & Soytas, 2023).

2.2.1. Conclusión de la revisión de la literatura e identificación del gap de investigación

Podemos concluir, pues, que existe en la literatura académica un debate aún inconcluso sobre si existe una prima verde (greenium) y, sobre el signo de esta. Se trata de un tema muy investigado pero para el cual la evidencia empírica no es unánime. Aunque la evidencia existente en la actualidad es mayoritaria para una greenium negativa (es decir, favorable para la emisión de bonos verdes por su menor coste de emisión para las instituciones emisoras), la diversidad de resultados sobre si la greenium es positiva o negativa depende en gran medida de las características de la muestra de bonos analizados en cuanto al emisor (soberanos, corporativos, que incluyan o no el sector financiero, etc.),

geografía (EE.UU., China, Europa, etc.), periodo temporal, mercado (primario o secundario o ambos), etc. Por otro lado, la inmensa mayoría de los estudios publicados utilizan métodos eonométricos tradicionales, siendo muy escasos los que utilizan modelos de ML. Por ello, nuestro estudio pretende contribuir a dicho debate aportando nueva evidencia empírica sobre el tema, en concreto sobre los bonos corporativos de la UE durante el periodo 2014-2025, aplicando técnicas de ML.

2. ANÁLISIS EMPÍRICO DE LOS DATOS

En este apartado se desarrollan los procesos relativos a la creación del modelo predictivo. Desde la selección, el análisis y preprocesamiento de los datos, hasta la metodología que hemos seguido para construir el modelo. Además, en el último apartado, discutiremos los resultados del modelo construido.

Para el procesamiento de los datos y la creación del ML se ha utilizado Python y sus librerías de procesamiento y visualización. Este es el conjunto de las librerías utilizadas:

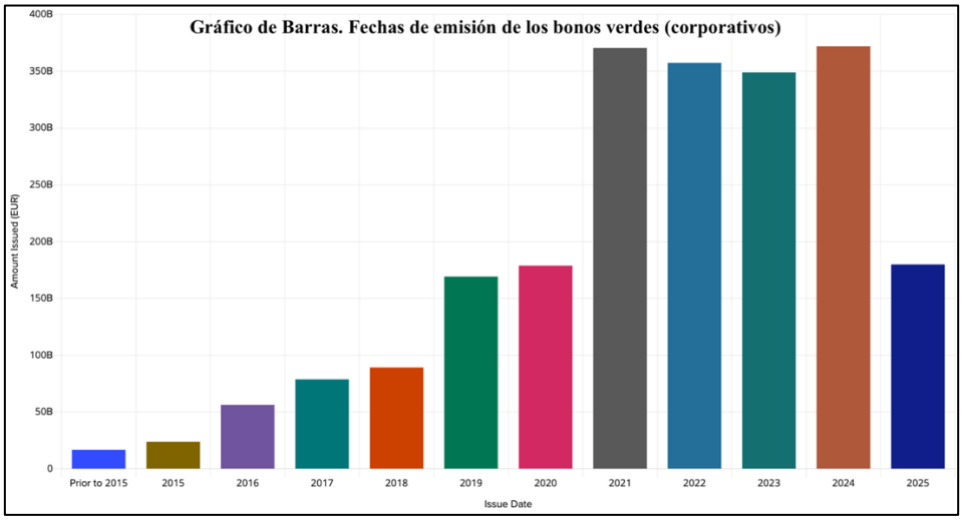
```
matplotlib. pyplot
seaborn
numpy
scipy. stats
sklearn. feature_selection
sklearn. ensemble
sklearn. pipeline
sklearn. compose
sklearn. preprocessing
sklearn. impute
sklearn. metrics
sklearn. model_selection
sklearn. linear_model
sklearn. ensemble
sklearn. compose
sklearn. preprocessing
sklearn. impute
scipy. stats
```

2.1. DATOS

Nuestros datos han sido obtenidos mediante la plataforma analítica de Refinitiv o LSEG. Esta plataforma es proveedora de datos y tecnologías respecto a mercados financieros. Así, hemos accedido a una base de datos que muestra la información sobre el instrumento de renta fija seleccionado junto a las métricas que nos ayudarán a guiar la investigación.

La investigación gira en torno al análisis de los bonos corporativos entre los años 2014 a 2025 para la geografía Unión Europea. ¿Por qué se seleccionan los datos a partir de 2014? Comentamos en la introducción que el primer bono verde emitido es en 2007. Sin embargo, no es hasta el 2008 cuando llega el segundo. Sabiendo que hay un periodo tan distante entre ellos, ha sido necesario analizar la fecha de emisión de los bonos siguientes. El gráfico nº1 muestra como antes de 2015, los bonos verdes emitidos eran casi inexistentes. Así, seleccionamos una muestra de los datos más significativa.

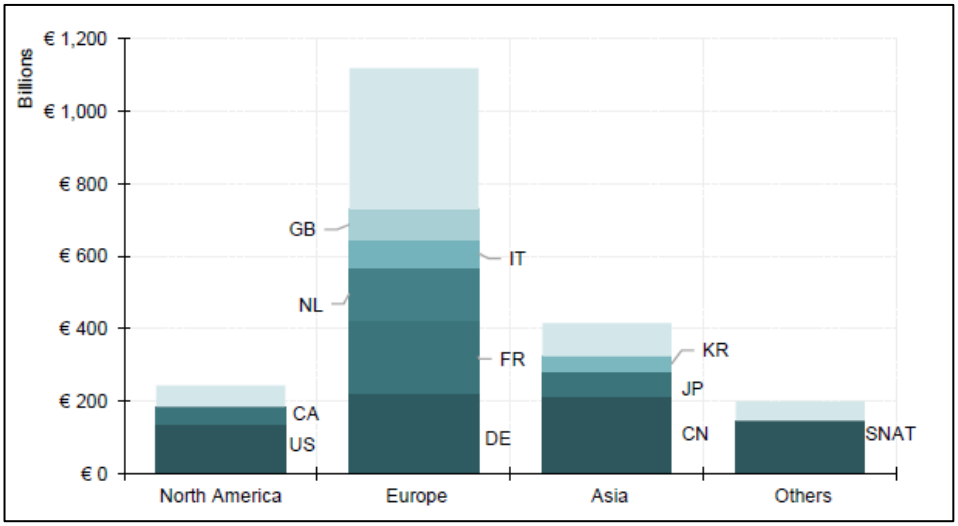
Gráfico 1: Fechas de emisión de los bonos verdes corporativo (años)



Fuente externa. Gráfico extraído de las muestras de Refinitiv (LSEG).

El gráfico nº 2 aporta evidencia sobre los grandes emisores de los bonos verdes. Categorizándolo en cuatro grupos de emisores, vemos que Europa es el mayor emisor de bonos verdes en comparación con el resto. Por tanto, es fácil señalar por qué hemos elegido la Unión Europea como geografía del estudio.

Gráfico 2: Emisiones totales (en €) de distintas geografías



Fuente externa. GSS bonds galore a market snapshot at the end of 2023, Societe Generale

Por otro lado, además de la elección del periodo temporal de emisión (2014-2025), hemos filtrado los datos por tipo de emisor (bonos coporativos) y por sector: dentro de los bonos corporativos, nos centramos en 4 sectores: energía, transporte, bancos (sector financiero) y telefonía. Hay una razón común en estas dos elecciones: el volumen de bonos verdes es mayor en el caso de las empresas que en el de los soberanos y también su volumen es mayor en esos 4 sectores. Además, un razón adicional para la elección de los bonos corporativos frente a los soberanos, en la UE, se deriva de la revisión de la literatura realizada, en la que hemos observado que, los estudios específicos sobre la UE se centran en su mayoría en el ámbito de los bonos verdes soberanos o de instituciones supranacionales, por lo que hemos querido aportar en el lado de los corporativos. Y más en concreto, nos parecían relevantes los sectores más contaminantes o que están inmersos en un proceso de transición ecológica importante, como son los 4 seleccionados, para los que, lógicamente, además, se cumple que emiten más bonos verdes que otros sectores.

Tras esta explicación sobre la selección de nuestros datos, podemos empezar a definir la composición de nuestro dataset. Los datos descargados de LSEG (Refinitiv) con esos tres filtros comentados anteriormente, nos proporcionan un total de las 88.935 bonos (filas). Además, las variables que seleccionamos son 19 aunque después del análisis del dataset, van fluctuando, así como pasará con las observaciones. Es importante mencionar que, el nombre que categoriza las variables viene dado por la plataforma de Refinitiv. Estas son:

- Green Bond: esta variable dicotómica identifica las observaciones en función de dos términos: “Yes” si son bonos verdes y “No” si no lo son.
- ISIN: se refiere al identificativo del bono en los mercados internacionales. No solo se usa en esta plataforma, también lo vemos en plataformas como Bloomberg.
- Issuer: esta variable categórica nominal contiene el nombre de los emisores. En este caso, el nombre de las empresas.
- Country of Issue: es también una variable categórica nominal referida al país emisor del bono.
- Coupon: es la variable numérica que contiene el valor del *coupon rate* (retorno de cupón respecto al precio del bono. Esta es una de las variables predicitivas de nuestra investigación.
- Maturity Date: es una variable *datetime*. Proporciona la fecha de vencimiento del bono.
- Issue Date: también es una variable *datetime.* Proporciona la fecha de emisión del bono.
- Issuer Type: es la variable categórica nominal que, además, de mostrar el tipo de emisor, nos refleja haber filtrado bien por “Corporate” (bonos corporativos)
- Coupon Type: variable categórica nominal. Aunque tiene 7 posibles valores (Variable, Floating, Discount, Fixed, Range, To be Priced y Ressetable Coupon), solo consideramos los de cupón fijo para el entrenamiento del modelo, ya que no sería directamente comparable la rentabilidad de los bonos que no sean “Plain Vanilla Fixed Coupon” fijos.
- Issue Price: es la variable numérica continua que contiene los valores de los precios de emisión para cada bono. Será también variable predictora para el modelo.
- Yield Spread (OTR) to Maturity: es una variable numérica. Se refiere a la diferencia entre el rendimiento del bono y el rendimiento de un bono del Tesoro con vencimientos similares. Esta será nuestra variable objetivo (target) para el modelo de predicción.
- Mac. Duration: variable numérica. Contiene los valores de la Duración de Macaulay para cada bono. Es un variable predictora para nuestro modelo.
- Moody's Long Term Issue Credit Rating: es una variable categórica que mide el rating crediticio del bono. Después, haremos una transformación de la variable

para poder tratarla en el modelo. Será también una variable predictora para el mismo.

- Industry Sector: Esta variable es categórica nominal. Contiene en sus valores los sectores de emisión de los bonos. Como ya hemos dicho, hemos filtrado por los sectores que mayor volumen de bonos han emitido. Haremos una transformación de la variable en 4 categorías.
- Face Issued Total: variable numérica que hace referencia al valor nominal total de la emisión del bono. Es una variable predictiva para el modelo de ML.
- Coupon Frequency: esta variable categórica ordinal hace referencia a la frecuencia del cupón en 3 categorías con valores 1, 2 y 4.
- Currency: Variable categórica referente a la divisa del bono emitido.
- Principal Currency: Esta variable es repetida, también hace referencia a la divisa de emisión. Como tenemos dos variables iguales, debemos elegir con cuál nos quedamos. En el preprocesamiento de los datos, vamos a ver con cuál de ellas debemos quedarnos.

Aunque por ahora, nuestro dataset tenga una dimensión de 88.935 x 19 variables, veremos a continuación que esto va a cambiar en ambos términos.

2.1.1. Pre-procesamiento de los datos

Aunque el dataset que hemos configurado puede parecer es muy extenso, en este apartado verificamos que, a través de la limpieza de los datos y transformación de las variables, se reducirá la extensión. Este pre-procesamiento tiene como finalidad obtener una mayor fiabilidad de nuestro dataset. Es muy importante que nuestras observaciones reflejen buena información para que nuestro modelo prediga los mejores resultados. En el dataset nos encontramos con algunas variables que pueden darnos error al ejecutar el modelo dado el origen de sus valores posibles. Es por eso por lo que, hay que tratarlos.

3.1.1.1. Transformación, creación y unificación de las variables

En primer lugar, decidimos filtrar la variable de "*Coupon type"* quedándonos solo con un único valor: **Vanilla Fixed Coupon**. Son bonos sencillos con cupón fijo y, por tanto, son instrumentos financieros más fáciles de comparar en términos de rentabilidad. Hemos eliminado, pues, aquellos bonos que no sean "Plain vanilla Fixed Coupon" y aquellos

cuya variable relevante de yield spread to maturity no sea “comparable” (por ejemplo los bonos que no tengan cupón fijo); es decir hemos eliminado: a) los que no tienen cupón fijo (lo tienen variable o floating); b) los que son cupón cero; y c) los que tienen opciones implícitas (convertibles (excheangeable), resetable/redeemable, putable, perpetual, fungible y preferred bonds). En concreto, dichos bonos no plian vanilla fijos, aparecen en LSEG bajo mútilples denominaciones o tipos (Annuity (perpetuos)Discount, Fixed margin over index, Fixed Ressetable, Fixed, then floating, Fixed, then zero coupon, Floating, then Fixed, Floating, then Ressetable, Floating, then variable, Floating, then zero, Multiple Payment Frequencies, Other/Complex Floating Rate, Pay at Maturity Floater, Range coupon, Step down-Margin over index, Step up/step down, Step up-Margin over index, To be priced coupon, Variable then float, Zero Coupon, Zero the fixed, Zero then floating, Zero then variable, Pay at Maturity Fixed). Por ello, se eliminan esas filas de la variable. Una vez que filtramos la variable, el dataset bajará a componerse de una muestra de **41.527** observaciones (bonos)

Resultado después de filtrar:

```
Filas restantes: 41527
Valores únicos en Coupon Type después del filtrado: ['Plain Vanilla Fixed Coupon']
```

Por otro lado, solo nos centraremos en los bonos a medio y largo plazo (vencimiento mayor o igual a 3 años), puesto que la rentabilidad de los bonos a corto plazo es mucho más volátil. Para ello, debemos crear una nueva variable o columna, que llamamos “**Vencimiento bonos (años)**”. Esta nueva columna recogerá, en la fecha de emisión, el tiempo hasta vencimiento, de estos bonos en años. Para crear esta nueva columna, utilizamos las columnas “Issue Date” y “Maturity Date”. A través de una diferencia de las columnas entre el valor de un año (365.25) obtenemos el vencimiento para cada bono, en años. Por tanto, hacemos una transformación de dos variables que eran *datetime* a una numérica.

El siguiente output del script muestra un ejemplo de las variables originales y la nueva variable creada:

	Maturity Date	Issue Date	Vencimiento bonos (años)
3	2044-09-11 22:00:00	2014-09-09 22:00:00	30.0
4	2114-01-22 23:00:00	2014-01-22 23:00:00	100.0
5	2044-03-04 23:00:00	2014-03-04 23:00:00	30.0
6	2044-07-28 22:00:00	2014-07-28 22:00:00	30.0
7	2039-01-30 23:00:00	2014-01-30 23:00:00	25.0

El tercer paso es la unificación de los sectores. Aunque hemos filtrado por 4 de los sectores con mayor volumen de emisión, todavía sigue habiendo valores que pueden unificarse aún más. Antes de la procesarlo, miraremos si hay algún valor nulo. En este caso, al imprimir el volumen de nulos, nos salen 2 observaciones. Por tanto, antes de arrastrar todo el error, lo eliminamos. Estos son los resultados:

```
Valores nulos en 'Industry Sector':
2
```

```
Valores nulos en 'Industry Sector' después de eliminar nulos:
0
```

Si observamos la composición de la variable "*Industry Sector",* son 7 posibles valores, pero, éstos tienen relación entre sí. Por ello, vamos a unificarlos en 4 categorías: Bancos, Energía, Transporte y Telefonía, en las que encontraremos:

- En la categoría Bancos nos encontramos con los valores "BANKS" y "OTHFINCL".
- Energía estará compuesta de "ENERGY", "GASDISTR" y "ELECTRIC".
- Transporte solo contiene ese mismo valor único
- El valor de Telefonía también corresponde con su mismo valor.

A continuación, se muestran los resultados obtenidos tras la transformación:

```
Los sectores unificados son :['Bancos' 'Energía' 'Transporte' 'Telefonía']
```

	Industry Sector	Sector Unificado
3	BANKS	Bancos
4	ELECTRIC	Energía
5	ELECTRIC	Energía
6	GASDISTR	Energía
7	BANKS	Bancos

Por último, al imprimir nuestro dataset podemos identificar que existen algunos valores en las variables que pueden entorpecer nuestro modelo después. Esto se debe a la presencia de "–", lo que significa que no existe un valor para esa observación en esa variable.

```
Valores '--' en 'Yield Spread (OTR) to Maturity' después de reemplazo: 21571
Valores '--' en 'Mac. Duration' después de reemplazo: 21565
Valores '--' en 'Issue Price' después de reemplazo: 1345
```

Aunque no son valores nulos, pueden provocar grandes errores en lo pasos siguientes. Por ejemplo, la identificación incorrecta del tipo de dato y por tanto, no coger bien las variables para crear el modelo.
Tras la aplicación de estos cuatro filtros, el número de observaciones que constituyen el dataset es **35.280**. Sigamos con los pasos que definen el número total.

3.1.1.2. Limpieza de los valores nulos

Este apartado se dedica a la eliminación de los valores vacíos que existe en cada variable, ya que, no aportan información y, además, no ayudan a predecir nuestro modelo.

La primera vez que se consultan los valores nulos, el resultado es una cantidad muy voluminosa de valores nulos, en las variables "*Yield Spread (OTR) to Maturity*", "*Mac. Duration"*, *"Moody's Long-term Issue Credit Rating", "Face Issued Total", "Issue Price" e "ISIN"*. Por tanto, antes de eliminar los valores nulos totales, debemos analizar las columnas (variables) para ver qué solución aplicar en cada caso. El resultado que obtendríamos si acabáramos con todos los valores nulos es de **7.940 observaciones**.

El siguiente output de nuestro script muestra los resultados impresos de los valores nulos existentes de nuestro dataset.

```
Yield Spread (OTR) to Maturity          21571
Mac. Duration                           21565
Moody's Long-term Issue Credit Rating   21066
Issue Price                              1345
Face Issued Total                         109
ISIN                                        7
Coupon                                      2
Ticker                                      0
Green Bond                                  0
Country of Issue                            0
Issuer                                      0
Issuer Type                                 0
Issue Date                                  0
Coupon Type                                 0
Maturity Date                               0
Industry Sector                             0
Coupon Frequency                            0
Principal Currency                          0
Currency                                    0
Vencimiento bonos (años)                    0
Sector Unificado                            0
dtype: int64
```

Vemos que la columna de ISIN no es una variable importante que nos vaya a ayudar a predecir nuestro modelo. Es por eso por lo que, eliminamos la variable entera del dataset. Además, hay unas variables repetidas: Currency y Principal Currency. Como ninguna de las dos tiene datos nulos, vamos a eliminar Principal Currency ya que, es más cómodo que sus valores sean los acrónimos identificativos de las monedas internacionales. Nos quedamos finalmente con 19 variables.

Para el resto de los valores nulos, no podremos eliminar directamente su variable, ya que, son condicionantes importantes para predecir la rentabilidad de los bonos. Es por eso por lo que, eliminamos los valores nulos existentes (filas) quedándonos con un total de **7.940** observaciones, como hemos explicado anteriormente.

```
Filas originales: 35280
Filas tras eliminar nulos: 7940
Porcentaje restante: 22.51%
```

3.1.1.3. Resumen general del dataset

Una vez que nuestro dataset no contiene valores nulos y está unificado, procedemos a verificar mediante un resumen de los datos, si nuestras variables son calificadas correctamente.

Al imprimir el tipo de datos que contiene nuestro dataset obtenemos este resumen:

```
Número de filas: 7940
Número de columnas: 19

Tipos de datos por columna:
Green Bond                                    object
Ticker                                        object
Issuer                                        object
Country of Issue                              object
Coupon                                       float64
Maturity Date                          datetime64[ns]
Issue Date                             datetime64[ns]
Issuer Type                                   object
Coupon Type                                   object
Mac. Duration                                float64
Issue Price                                  float64
Yield Spread (OTR) to Maturity               float64
Moody's Long-term Issue Credit Rating         object
Industry Sector                               object
Face Issued Total                            float64
Coupon Frequency                             float64
Currency                                      object
Vencimiento bonos (años)                     float64
Sector Unificado                              object
dtype: object
```

Identificamos errores en la calificación. Si no hubiéramos identificado previamente los "—", no calificaría bien las variables numéricas (nos las consideraba object, categóricas). Además, la variable "*Coupon Frecuency"* que es una variable categórica ordinal, está mal identificada, la reconoce como numérica ya que, sus valores, son número del pueden ser 1, 2 o 4. Al transformar estas variables según el tipo correcto de dato, el resultado es este:

```
Tipos de datos por columna:
Green Bond                                    object
Ticker                                        object
Issuer                                        object
Country of Issue                              object
Coupon                                       float64
Maturity Date                          datetime64[ns]
Issue Date                             datetime64[ns]
Issuer Type                                   object
Coupon Type                                   object
Mac. Duration                                float64
Issue Price                                  float64
Yield Spread (OTR) to Maturity               float64
Moody's Long-term Issue Credit Rating         object
Industry Sector                               object
Face Issued Total                            float64
Coupon Frequency                            category
Currency                                      object
Vencimiento bonos (años)                     float64
Sector Unificado                              object
dtype: object
```

Ahora si están todas la variables bien calificadas.

3.1.1.4. Análisis de Outliers

El último apartado relacionado con el preprocesamiento de los datos hace referencia al análisis de los outliers. El término de atípico corresponde con aquellos valores que en comparación con el resto de las observaciones destacan por ser demasiado altos o demasiado bajos, es decir, no corresponde con los valores normales de los datos.

Para medir la existencia de los outliers, se utiliza la técnica del rango intercuartílico. Esta técnica funciona identificando los valores extremos del dataset. Mediante la diferencia

entre sus dos extremos, el quartil 25% o extremo izquierdo y el quartil 75% o extremo derecho, identificaremos los valores que están fuera de la “media”.

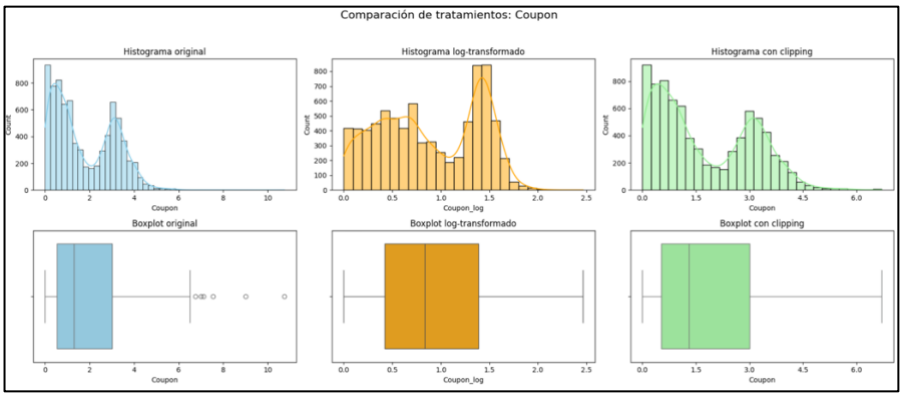
El análisis intercuartílico aplicado a nuestros datos identifica outliers en seis de las variables: *'Coupon', 'Mac. Duration', 'Issue Price', 'Yield Spread (OTR) to Maturity' "Face Issued Total", "Vencimiento bonos (años)".*

```
Coupon: 6 outliers detectados
Mac. Duration: 155 outliers detectados
Issue Price: 2077 outliers detectados
Yield Spread (OTR) to Maturity: 488 outliers detectados
Face Issued Total: 1013 outliers detectados
Vencimiento bonos (años): 295 outliers detectados
```

Para el tratamiento de los outliers utilizamos dos técnicas: la transformación logarítmica de las variables numéricas y la técnica clipping que permite limitar los datos en los valores extremos (en nuestro caso los hemos reemplazado con el valor de la variable en cuestión en su límite superior y/o inferior correspondiente al cuartil 75% más 1,5 veces el rango intercuartílico y al cuartil 25% menos 1,5 veces el rango intercuartílico, respectivamente: Q3+1,5xIQR y/o Q1-1,5xIQR. La comparativa entre estos modelos, nos ayuda a reconocer que con la aplicación correcta de clipping, eliminamos los outliers que se identifican. Las siguientes ilustraciones (nº 1, 2, 3, 4, 5, 6), muestran las distribuciones de propabilidad de nuestras variables numéricas con outliers, antes y después de su tratamiento (y éste con las dos técnicas).

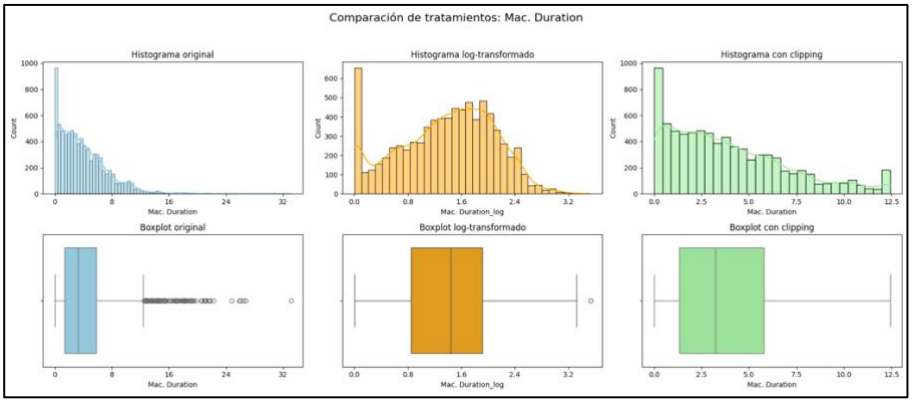
A continuación se muestran los gráficos que diferencian los dos métodos explicados. En primer lugar se muestra la distribución original, seguida de la logarítmica y la técnica clipping. En todos los casos podemos observar que no hay ningún outlier.

Ilustración 1:Resultado de la transformación de los outliers para Coupon



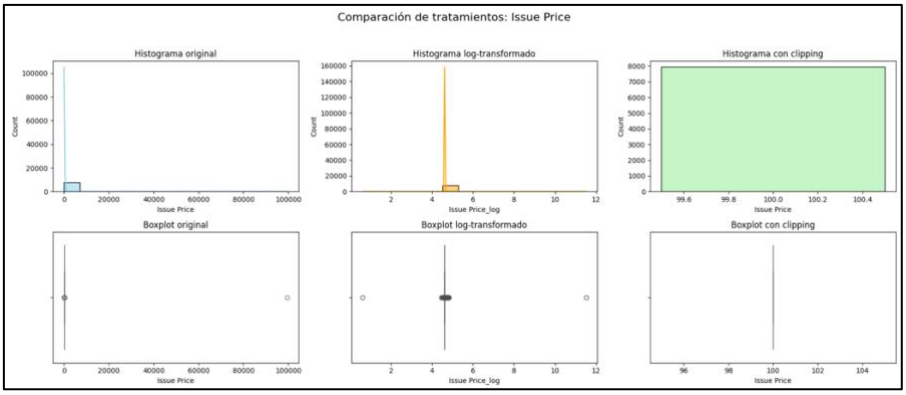
Fuente propia. Elaborado con la herramienta Python (resultados en script).

Ilustración 2: Resultado de la transformación de los outliers para Mac. Duration



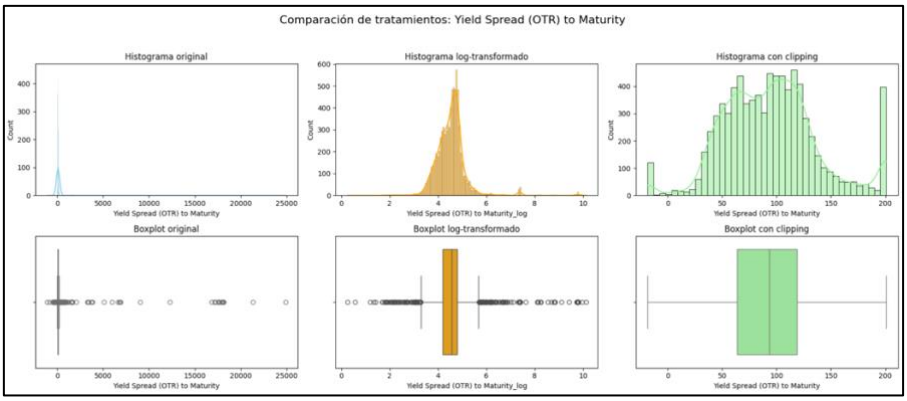
Fuente propia. Elaborado con la herramienta Python (resultados en script).

Ilustración 3: Resultado de la transformación de los outliers para Issue Price



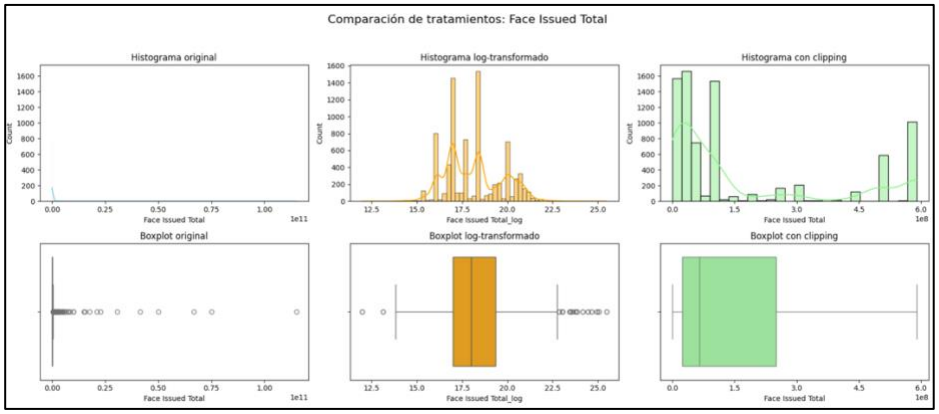
Fuente propia. Elaborado con la herramienta Python (resultados en script).

Ilustración 4: Resultado de la transformación de los outliers para Yield Spread (OTR) to Maturity



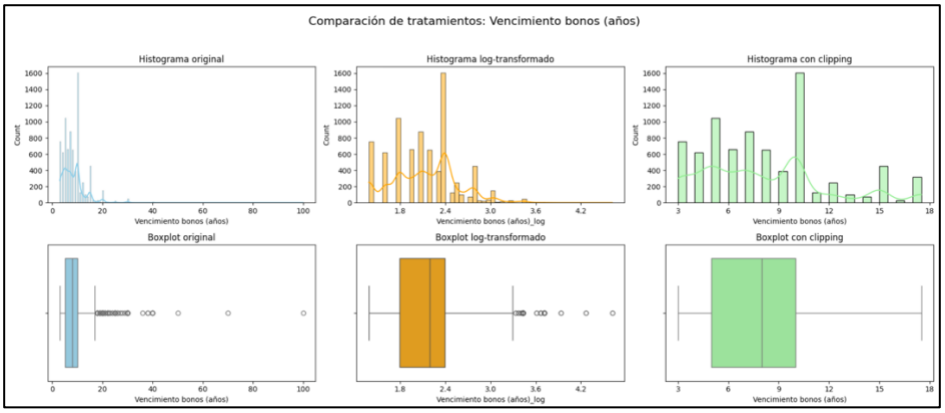
Fuente propia. Elaborado con la herramienta Python (resultados en script).

Ilustración 5: Resultado de la transformación de los outliers para Face Issued Total



Fuente propia. Elaborado con la herramienta Python (resultados en script).

Ilustración 6: Resultado de la transformación de los outliers para Vencimiento bonos (años)



Fuente propia. Elaborado con la herramienta Python (resultados en script).

Mientras que con la técnica clipping mitigamos los efectos de los outliers, la transformación logarítmica todavía los mantiene. El resultado que obtenemos cuando utilizamos clipping para nuestro dataset es:

```
Coupon: 0 outliers detectados
Mac. Duration: 0 outliers detectados
Issue Price: 0 outliers detectados
Yield Spread (OTR) to Maturity: 0 outliers detectados
Face Issued Total: 0 outliers detectados
Vencimiento bonos (años): 0 outliers detectados
```

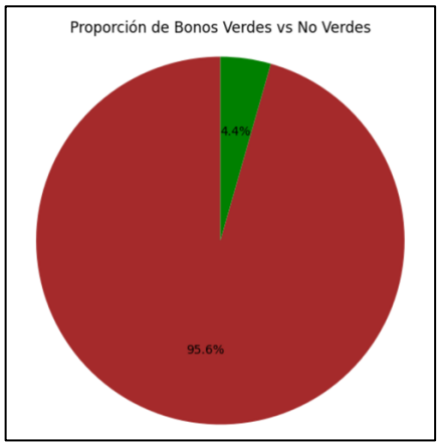
3.1.2. Análisis univariante y bivariante

Este apartado se dedica al estudio de las relaciones entre las variables. Se busca encontrar relaciones o información que aporte a nuestra investigación. Mediante la creación de gráficos de visualización, tenemos una primera idea de nuestras suposiciones. En algún

caso, los gráficos nos ayudan a visualizar nuestra estructura y, a su vez, nos explicarán porqué los resultados obtenidos llevan esa dirección.

Por ejemplo, en el gráfico nº3 podemos ver la proporción de los tipos de bonos de nuestro dataset. Obtenemos que un 4,4% de los bonos son verdes:

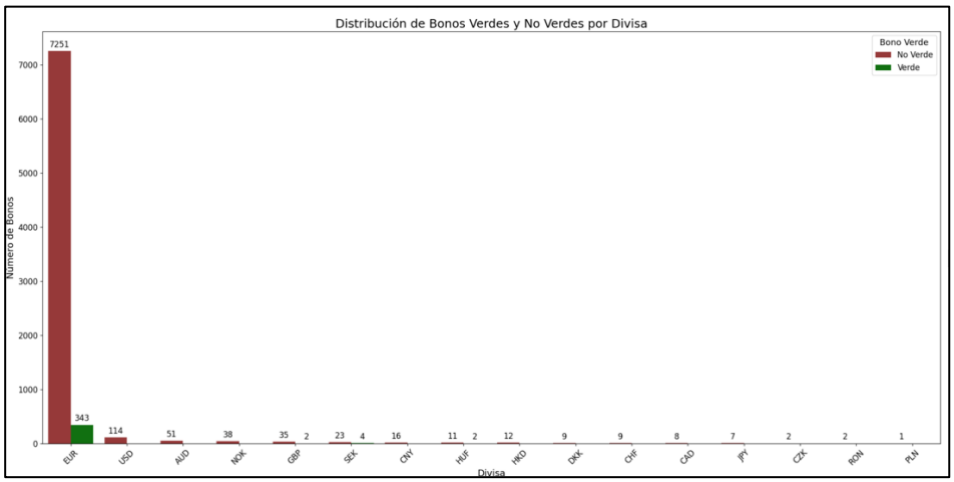
Gráfico 3: Proporción de los bonos verdes y no verdes en nuestra muestra



Fuente propia. Elaborado con la herramienta Python (resultados en script).

El gráfico nº 4 nos muestra la distribución de las divisas en la emisiones de los bonos. Teniendo en cuenta que nuestro ámbito geográfico es la Unión Europea, podemos destacar la presencia de divisas extranjeras como por ejemplo, los dólares estadounidenses y los yenes de Japón, si bien, el 95,64% de nuestros bonos están emitidos en euros, lógicamente:

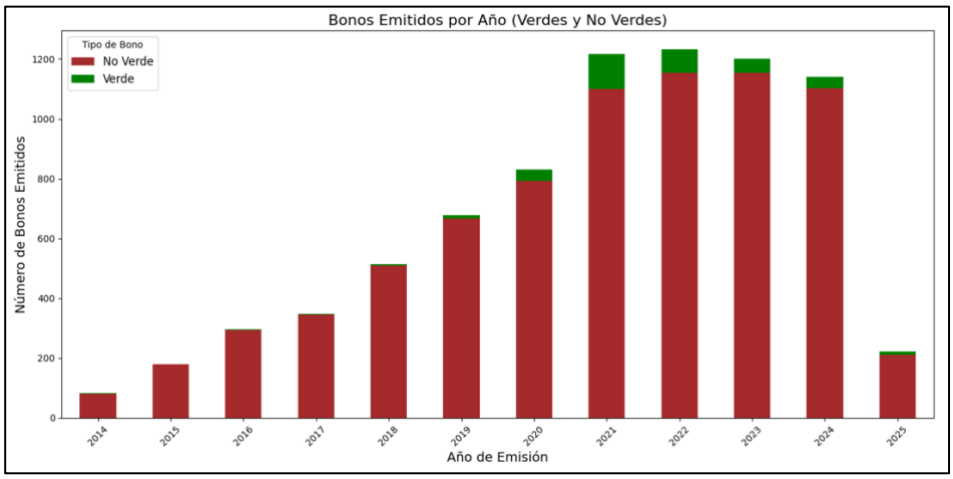
Gráfico 4: Distribución de los bonos verdes y no verdes por divisa respecto a nuestra muestra final (nº total de emisiones)



Fuente propia. Elaborado con la herramienta Python (resultados en script).

El gráfico nº5 representa el volumen de bonos emitidos a lo largo de los años del periodo considerado en nuestro estudio (2014-2025). Así, podemos ver que, a partir del año 2020, la emisión de los bonos verdes ya se empieza a notar. Aunque a partir de 2021 en que se obtuvo el máximo de bonos verdes corporativos emitidos por estos sectores, ha ido disminuyendo con los años. Puede deberse a la falta de confianza en los bonos verdes. El año 2025 contiene muy pocos datos. Esto se debe a que solo recoge los cuatro primeros meses del año.

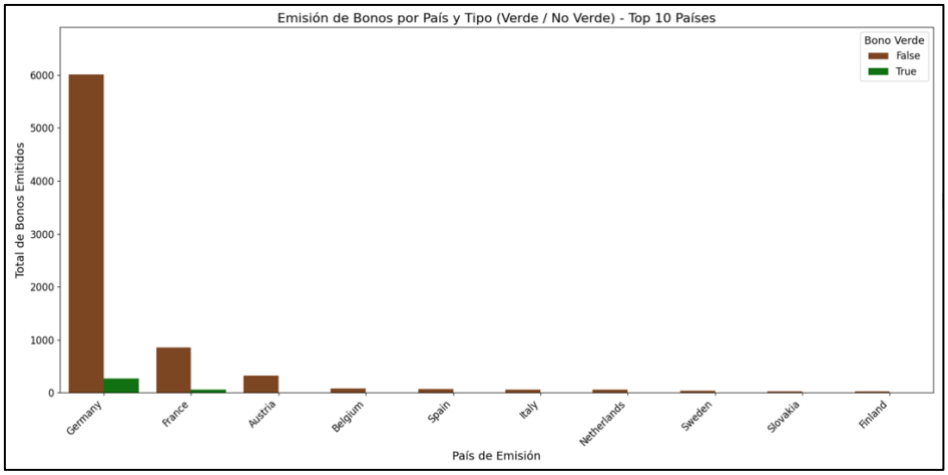
Gráfico 5: Total de emsiones de bonos verdes y no verdes por año



Fuente propia. Elaborado con la herramienta Python (resultados en script).

El gráfico nº6 muestra los 10 países de la UE con mayor volumen de emisión de bonos verdes y marrones (medido por el total de su valor nominal de emisión). Aunque en nuestro dataset hay más valores, son casi insignificantes y, casi no se ven en el gráfico. Por eso, se muestran aquellos con mayor volumen de emisión. Como vemos, Alemania es el país que domina la emisión de los bonos llegando a la cifra de emisión de 6000 bonos marrones y aproximadamente 200 bonos verdes.

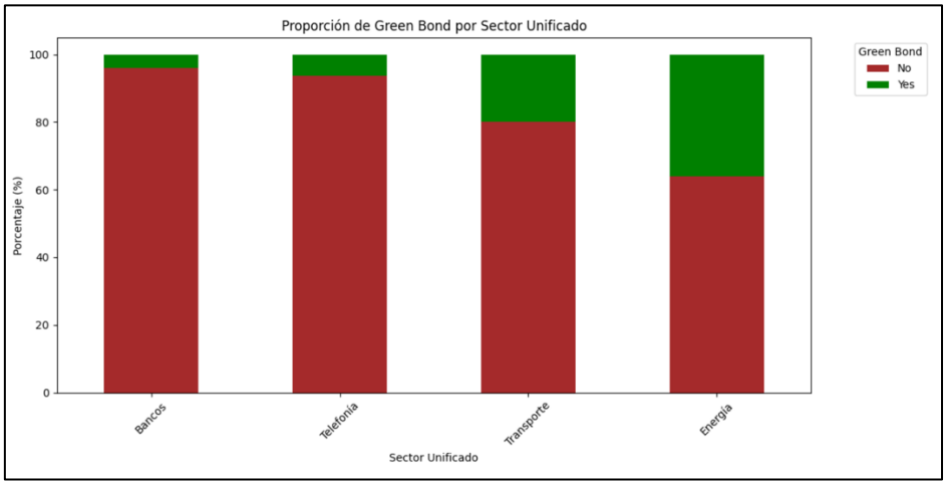
Gráfico 6: Top 10 de los países emisores por tipo de bono (nº de emisiones total respecto a nuestra muestra final)



Fuente propia. Elaborado con la herramienta Python (resultados en script).

El gráfico nº7 es interesante ya que nos muestra la proporción de bonos verdes por cada sector. Así, vemos que en el sector energía, en torno al 35% de sus bonos emitidos, son verdes mientras que en los bancos, los bonos verdes apenas representan el 2% del total de sus emisiones de bonos.

Gráfico 7: Proporción de emisión de tipo de bono por Sector Unificado (respecto a nuestra muestra final)

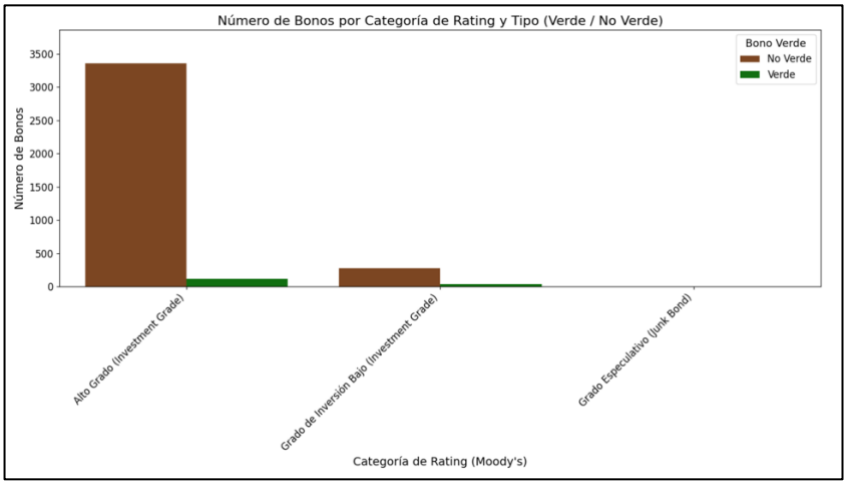


Fuente propia. Elaborado con la herramienta Python (resultados en script).

Para elaborar el gráfico nº8, hemos agrupado el rating crediticio de Moodys, la variable "Moody's Long Term Rating" en 3 categorías. De manera que los bonos con rating de tipo A se han clasificado como bonos de Alto grado. Los bonos con Rating igual a BA, cubre el grado de inversión baja. Los tipos de bono con un rating igual a B o CA, está destinados al grado especulativo. Por tanto, podemos asegurar que la mayoría de los

bonos del dataset, tanto de los verdes como de los marrones, forman parte del Alto Grado, es decir, con un rating más elevado.

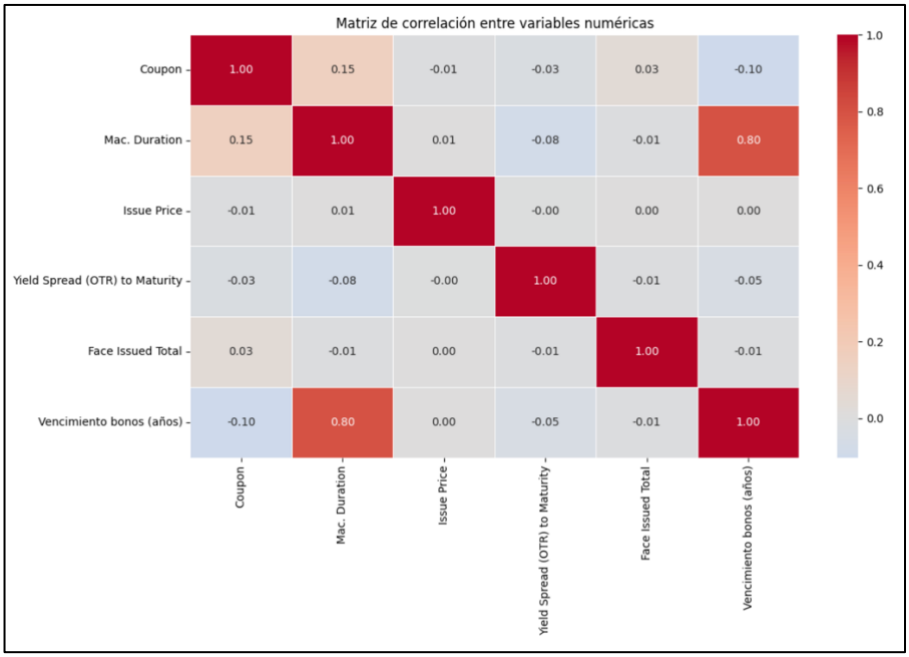
Gráfico 8: Distribución del rating de los bonos verdes y no verdes de nuestra muestra final



Fuente propia. Elaborado con la herramienta Python (resultados en script).

Por último, obtenemos la matriz de correlaciones de las variables numéricas. No existe una correlación significativa entre nuestras variables a excepción de la que sí existe entre Las variables Vencimiento del bono (en años) y la duración de Macaulay, con un valor de 0.80. Se trata de algo lógico y esperado, puesto que la variable vencimiento en años influye en el cálculo de la duración de Macaulay de forma directa y positiva. Por ello, no incluiremos la variable Vencimiento del bono (en años) como variable predictiva en nuestro modelo, sino la duración de Macaulay. En realidad, nosotros construimos al inicio la variable vencimiento del bono (en años) simplemente para filtrar aquellos bonos cuyo estudio nos interesaba (a medio y largo plazo), no con intención predictiva, pues ya disponíamos de la variable Duración de Macaulay y esperábamos que existiese una fuerte correlación entre ambas, como así ha ocurrido.

Gráfico 9: Matriz de correlaciones sobre las variables numéricas



Fuente propia. Elaborado con la herramienta Python (resultados en script).

2.2. METODOLOGÍA

Nuestra investigación quiere dar respuesta al debate sobre la existencia o no de una prima verde en los bonos. En este apartado se desarrolla el procedimiento que hemos seguido para la creación de un modelo de ML que prediga la greenium de los bonos. Esta predicción dará respuesta a nuestra investigación. El proceso metodológico resumido consta de los siguientes pasos: una vez obtenidos los datos, y realizado el preprocesamiento y análisis exploratorio minucioso de los mismos en el apartado anterior 3.1, a continuación, separamos la muestra total ya limpia y preprocesada en las dos submuestras de bonos verdes y bonos no verdes (marrones). Entrenamos tres modelos de ML (regresión lineal, random forest y gradient boosting) para predecir la prima de riesgo de los bonos marrones y evaluamos su desempeño, tanto su capacidad explicativa (con las métricas R^2 y R^2 ajustado), como su capacidad predictiva (a través del MAE y RMSE). Ello nos permitirá seleccionar el mejor modelo que mejor predice de los tres, con el cual ya realizaremos la predicción de la prima de riesgo de los bonos verdes. Se determina la prima verde como la resta de las medias de ambas rentabilidades predichas (verde menos marrón) y se testa la significatividad estadística de la greenium estimada mediante dos contrastes de diferencias de medias: uno sobre si las medias son iguales o no (testamos la

existencia o no de la prima verde) y otro sobre cuál de ambas medias de rentabilidad es mayor (testamos el signo positivo o negativo de la prima verde). Pasamos, pues, a detallarlo.

Como indicamos al definir las variables, no utilizamos todas las variables como predictoras (X), solo utilizamos ocho. Éstas son: 'Coupon', 'Mac. Duration', 'Issue Price', 'Face Issued Total', 'Sector Unificado', 'Country of Issue', "Moody's Long-term Issue Credit Rating" y 'Coupon Frequency'.

Por otro lado también debemos especificar nuestra variable objetivo (Y): “Yield Spread (OTR) to Maturity”. Esta es la variable que queremos predecir para los bonos verdes y los bonos marrones.

Comenzamos a crear nuestro modelo predictor con el nuevo conjunto de datos que hemos limpiado, tal y como indicamos en el paso anterior. Es importante aclarar que dividimos la muestra de los bonos. A partir de ahora los trataremos por separado, es decir, por un lado predecimos los bonos verdes y por otro lado los no verdes. Además, utilizaremos los resultados del modelo elegido en el caso de los bonos marrones para predecir el Yield Spread (OTR) to Maturity de los bonos verdes. Una vez aclarados estas dos observaciones, separamos los bonos marrones, obteniendo un total de **7.589 observaciones.**

```
X: (7589, 8)
y: (7589,)
df_model: (7589, 9)
```

Una vez que hemos definido nuestras variables predictoras y las variable objetivo, comenzamos a preparar la estructura del modelo predictivo. Comparamos tres modelos: Regresión lineal, Random Forest y Gradient Boosting. Para obtener los mejores resultados, utilizamos *RandomizedSearchCV* que, nos ayudará a encontrar de manera automatizada a generar combinaciones de hiperparámetros sacando los mejores resultados, así, no debemos cambiar las combinaciones manualmente. Por otro lado, hacemos Cross Validation (validación cruzada) para minimizar la varianza de error en nuestras predicciones (evitar el sobreajuste u overfitting). Además, utilizaremos un pipeline para asegurarnos que el proceso de entrenamiento y test se construye en base a las mismas estructuras.

Para realizar esta comparación hemos calculado las métricas apropiadas que nos muestran el desempeño de los modelos, tanto su bondad del ajuste (R^2 y R^2 ajustado) como su calidad predictiva (RMSE y MAE). Las métricas de bondad del ajuste se calculan para el conjunto de datos entrenamiento (train), mientras que las métricas de capacidad predictiva, las hemos calculado tanto para los subconjuntos de train (entrenamiento) y de test (prueba), como en la validación cruzada. En la Tabla nº1 del apartado 3.3. resultados se muestran los resultados obtenidos para todas estas métricas. Como queremos elegir el modelo que mejor prediga la rentabilidad de los bonos marrones, para aplicarlo después a la predicción de la rentabilidad de los bonos verdes, nos centraremos en las métricas RMSE y MAE, dos resultados que nos ayudan a elegir en base al error en la predicción. Por lo tanto, aquel modelo que minimice el error será el seleccionado.

Mediante la comparación del desempeño de los modelos en los datos de entrenamiento y en los datos de prueba, detectaremos la existencia o no de sobreajuste (u overfitting), que, como hemos dicho, hemos tratado mediante la estratificación del conjunto de datos en varios subconjuntos (5 folds en concreto), con lo que hemos obtenido una estimación más robusta del desempeño de los tres modelos. El overfitting es un problema que o significa que un modelo ha aprendido demasiadas particularidades en los datos de entrenamiento y no está prediciendo bien en los datos de prueba.

```
--- Random Forest ---
  RMSE Entrenamiento:      12.19
  RMSE Prueba:             28.49
  RMSE Ratio (Test/Train): 2.34
  MAE Entrenamiento:       6.84
  MAE Prueba:              17.37
  MAPE Entrenamiento:      0.12%
  MAPE Prueba:             0.31%
  MAPE Ratio (Test/Train): 2.58
   Posible sobreajuste detectado (error en test > 20% que en train).
```

Además, hemos querido dar un valor adicional a la predicción del modelo y es que, puede ser importante para futuros estudios saber qué variables son las que más influyen en la predicción, es decir, ¿qué variable tiene mayor peso? En el apartado siguiente mostraremos los resultados.

Tras el tratamiento del sobreajuste, estaremos en condiciones de establecer el mejor modelo seleccionado. En nuestro caso, el mejor modelo en base a las métricas de error es **Random Forest**, como mostraremos en el siguiente apartado de resultados.

Una vez seleccionado el modelo que mejor predice la rentabilidad de los bonos marrones, el siguiente paso será predecir los valores de nuestra variable target (rentabilidad)para los bonos verdes.

Una vez predichas las rentabilidades tanto de los bonos marrones como de los verdes, calcularemos las medias de las predicciones de ambos tipos de bonos. Y entonces estimaremos la greenium como la diferencia entre la rentabilidad media de los bonos verdes menos la rentabilidad media de los bonos marrones. Y miramos qué cuantía nos sale (o sea si existe greenium) y si es positiva o negativa. Así podremos, pues, dar respuesta a nuestras preguntas de investigación: ¿Existe una prima verde? ¿Es positiva o negativa? En los apartados siguientes, expondremos los resultados.

Y por último, para contrastar la significatividad estadística de esa greenium estimada, hacemos dos contrastes de hipótesis (ambos t-student con Spyci para muestras independientes, con un nivel de confianza del 95% y del 90%):

- Hacemos un contraste de igualdad de medias (t-test) del Yield spread de las muestras verde y marrón, para ver si existe diferencia significativa entre ambas medias. Es decir, testamos la significatividad estadística de nuestra greenium estimada. H0: igualdad de medias. H1: medias diferentes. Queremos rechazar H0 (rechazar que las medias son iguales, en cuyo caso no existe Greenium), y aceptar H1 (que las medias son distintas, o sea que sí existe greenium). Rechazaremos Ho si nuestro p-valor < 0,05 para el caso del 95% y si nuestro p-valor < 0,10 para el caso del 90%.

Con este 1er contraste testamos solo si existe greenium. Aún no testamos si esa grenium es positiva o negativa. Para eso hacemos el 2º contraste.

- Hacemos un contraste t-Student de la hipótesis H0: la media de rentabilidad de los bonos verdes es menor o igual que la de los marrones (la greenium existe y es negativa). H1: la media de rentabilidad de los bonos verdes es mayor que la de los marrones. Queremos rechazar la H0 de que la rentabilidad verde es menor o igual que la de los marrones y aceptar la H1(alternativa) de que la rentabilidad verde es mayor. Nuevamente, rechazaremos H0 si nuestro p-valor < 0,05 para el caso del 95% y si nuestro p-valor < 0,10 para el caso del 90%.

Estos dos contrastes son “complementarios”: el primero es sobre la existencia de prima verde, (medias diferentes) y el otro es sobre el signo de la prima (cuál de las medias es mayor)

2.3. RESULTADOS

Este apartado se dedica a la exposición de los resultados del proceso que hemos explicado anteriormente: la creación del modelo predictivo de ML sobre la Yield Spread (OTR) to Maturity. En el apartado siguiente, se discutirá la aportación de los resultados a nuestra investigación.

Como avanzábamos en el apartado anterior, hemos analizado la posibilidad de que existiera overfitting en nuestros modelos, y, efectivamente, lo hemos detectado en el caso del modelo random forest, tal y como mostramos en el siguiente output de nuestro script. Esta ilustración representa el output de la verificación de overfitting. Como podemos ver en el comentario. Para el modelo elegido de Random Forest, si existe overfitting así que, hemos de tratarlo para que nuestras predicciones sean los más reales y fuertes posibles.

```
--- Random Forest ---
  RMSE Entrenamiento:      12.19
  RMSE Prueba:             28.49
  RMSE Ratio (Test/Train): 2.34
  MAE Entrenamiento:       6.84
  MAE Prueba:              17.37
  MAPE Entrenamiento:      0.12%
  MAPE Prueba:             0.31%
  MAPE Ratio (Test/Train): 2.58
   Posible sobreajuste detectado (error en test > 20% que en train).
```

Pero lo hemos solucionado según lo descrito en el apartado anterior, obteniendo entonces las siguientes métricas que ya son ligeramente menores en el test que en el train.

```
Fitting 5 folds for each of 9 candidates, totalling 45 fits
Parámetros tras tratamiento: {'regressor__subsample': 0.6, 'regressor__max_features': 0.8}

Métricas tras tratamiento:
  RMSE train: 39.36
  RMSE test : 39.14
  MAE  train: 29.22
  MAE  test : 29.21

Brecha RMSE (test − train): -0.21
Ratio RMSE (test/train):    0.99
```

La siguiente tabla muestra los resultados que hemos obtenido al comparar los 3 modelos descritos. Para ello, se añaden las métricas de errores tanto para el conjunto de

entrenamiento como la prueba (o test). Además, también añadimos solo para el conjunto de entrenamiento el R2 y R2 Ajustado. En la última línea del resultado obtenemos que en base al RMSE en el conjunto de prueba, el mejor modelo, el que obtiene los mejores valores, es Random Forest. El **Random Forest**, es una herramienta de ML de aprendizaje automático predictiva de tipo ensemble. Esta técnica se caracteriza por el entrenamiento de muchos árboles de decisión con muestras aleatorias de los datos y con un conjunto de variables. De esta manera, a medida que su entrenamiento avanza se crea un modelo muy preciso. Este algoritmo usa elementos como out of bag que permite minimizar el error de generalización y, por tanto, al sobreajuste.

Tabla 1: Output de los resultados del entrenamiento y test para los modelos: Regresión Lineal, Random Forest y Gradient Boosting según las métricas señalizadas

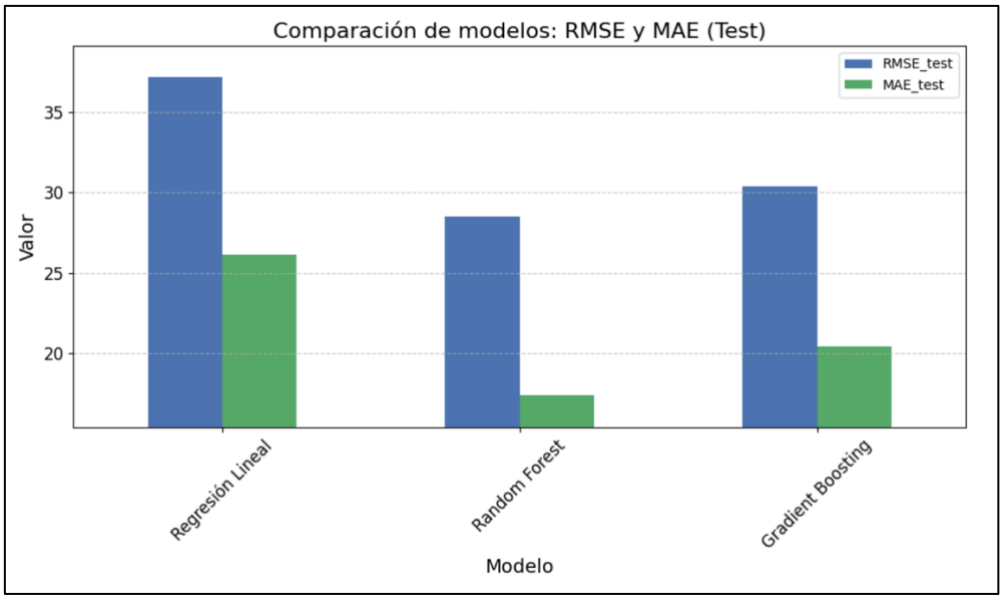
Modelo	R2_train	R2_Adj_train	RMSE_train	MAE_train	MAPE_train	RMSE_test	MAE_test	MAPE_test
Regresión Lineal	0.2889	0.2852	36.89	25.93	0.54	37.14	26.10	0.54
Random Forest	0.9223	0.9219	12.19	6.84	0.12	28.49	17.37	0.31
Gradient Boosting	0.5779	0.5757	28.42	19.09	0.41	30.39	20.44	0.44

Mejor modelo según RMSE en test: Random Forest

Elaborado con la herramienta Python (resultados en script).

El gráfico nº10 nos ayuda a visualizar de manera más sencilla los errores que obtenemos en la tabla anterior. Nos muestra como han desempeñado las métricas de los errores en cada uno de los modelos. Es claro que Random Forest tiene el valor más bajo de error.

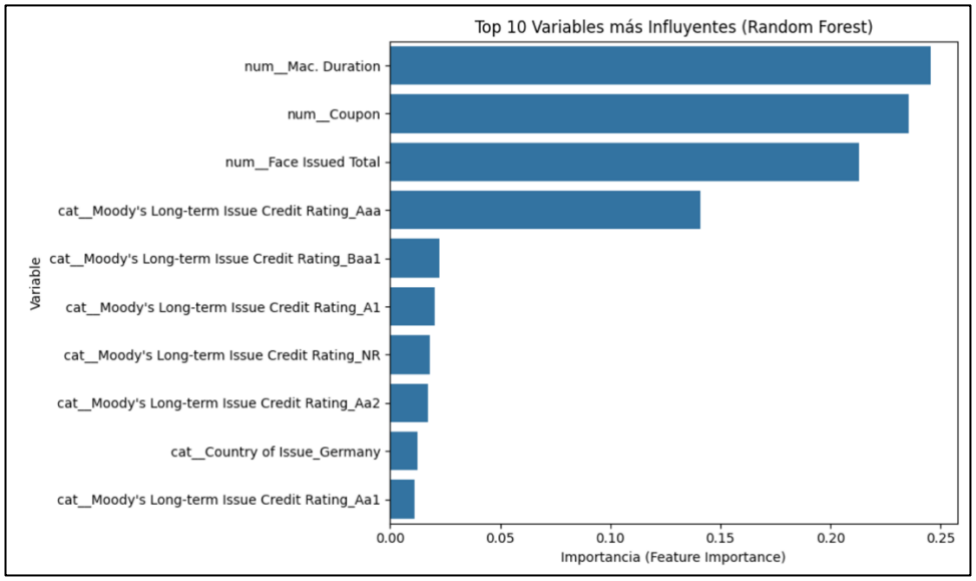
Gráfico 10: Comparación del RMSE test de los modelos entrenados



Elaborado con la herramienta Python (resultados en script).

El gráfico nº 11 muestra las variables más influyentes en la predicción de la rentabilidad (Yield Spread (OTR) to Maturity). Como era, previsible con base en la literatura previa existente, las 4 variables que nos salen más influyentes en la rentabilidad de un bono son: su riesgo de tipo de interés (medido en nuestro caso por la duración de Macaulay), su cupón (rentabilidad por cupón), el volumen de la emisión y que tenga bajo riesgo crediticio (AAA).

Gráfico 11: Variables más influyentes para la predicción del Yield Spread (OTR) to Maturity



Elaborado con la herramienta Python (resultados en script).

La siguiente tabla de resultados representa los valores predichos para los bonos marrones. En la ilustración vemos los datos originales del Yield Spread comparados con los valores predichos con el modelo RF.

El mejor modelo es: Random Forest
Pipeline reentrenado con todos los bonos marrones.

	Yield Spread (OTR) to Maturity	Yield Spread Predicho
4	200.795866	189.793345
7	53.613008	50.574181
34	50.768173	74.749533
36	176.001961	161.856715
38	175.573783	162.912965
41	59.550008	60.809244
42	76.715053	80.324451
43	162.906281	143.997028
45	78.287258	75.759097
46	135.949395	116.704936

Como hemos especificado anteriormente, el entrenamiento de los datos de los bonos marrones nos sirve para predecir los valores del Yield Spread de los bonos verdes. Así, en esta tabla visualizamos los valores predichos obtenidos para los bonos verdes.

El mejor modelo es: Random Forest
Pipeline reentrenado con todos los bonos verdes.

	Yield Spread (OTR) to Maturity	Yield Spread Predicho
135	54.608636	136.106163
22335	63.011979	71.002960
30466	83.721684	104.207059
30493	69.219463	74.454642
30570	92.260740	83.650147
38574	55.753194	75.426248
39180	43.530450	43.339025
39447	10.837647	35.280805
45489	122.625761	144.366808
45862	98.232853	114.115930

El paso siguiente a las predicciones es el cálculo de la media de los valores predichos. Es decir, la media de las predicciones del Yield Spread para los bonos marrones y, de la misma manera para los bonos verdes. De esta manera, encontramos una forma normalizada de comparación de las predicciones. El segundo resultado se refiere a la diferencia entre las medias de las predicciones. En otras palabras, el resultado a nuestra

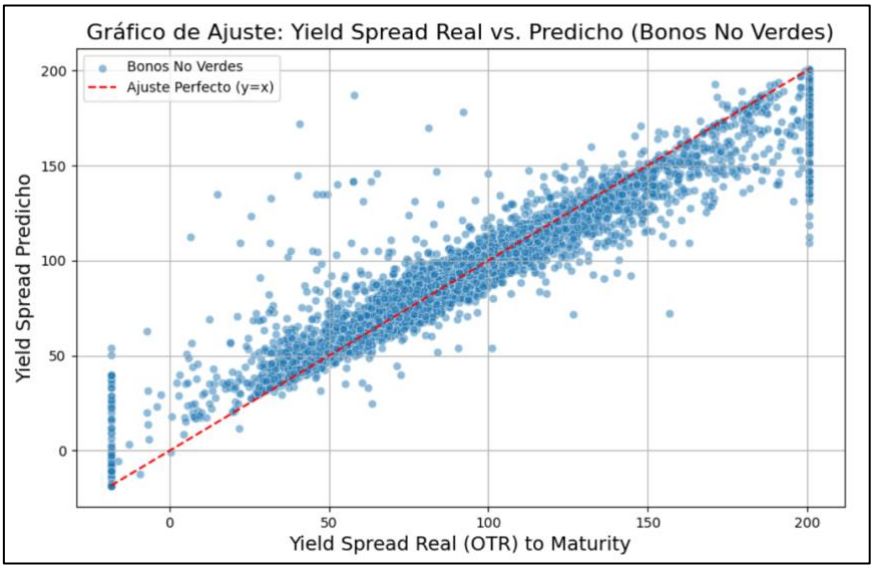
pregunta: ¿Existe una prima verde? Hay que notar que las variables de rentabilidades aparecen medidas en puntos básicos.

```
Media del Yield Spread Predicho para Bonos Verdes: 97.5096
Media del Yield Spread Predicho para Bonos No Verdes: 94.5169
```

```
Diferencia de medias predichas (Verdes - Marrones): 2.9927
```

Por último en este apartado de discusión sobre el modelo predictor de los Spreads, se han creado unos gráficos que visualizan el ajuste de las predicciones. El gráfico nº12 representa el ajuste de los spreads predichos de los bonos marrones respecto a los reales. Como vemos, podemos concluir que hay un buen ajuste de los datos repartiéndose a lo largo de la tendencia original. Aunque se puede deducir que el modelo tiene margen de mejora como se ve en los datos extremos, podríamos decir que el modelo se ajusta bien a las predicciones.

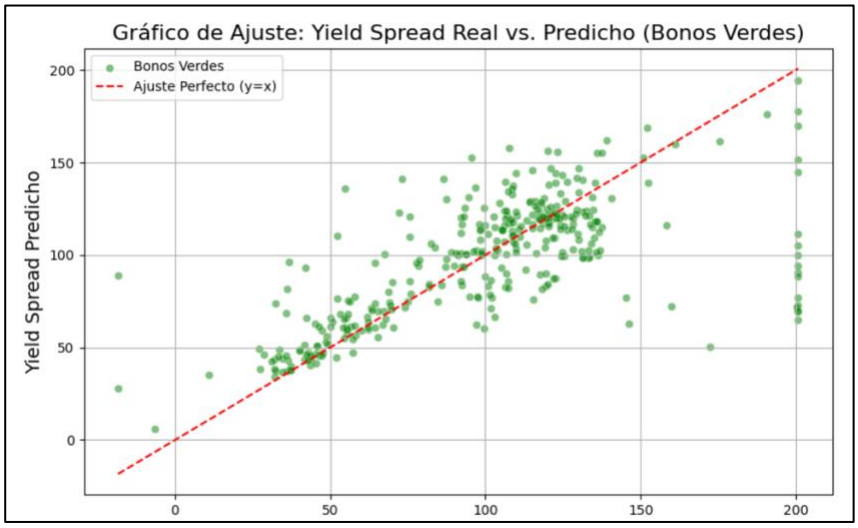
Gráfico 12: Ajuste de las predicciones de los bonos no verdes respecto a los datos reales utilizando el modelo Random Forest



Fuente propia. Elaborado con la herramienta Python (resultados en script).

En el mismo sentido, el gráfico nº13 dedicado a los spreads de los bonos verdes, se compara de manera visual el ajuste de los valores predichos y los reales. De nuevo, podemos afirmar que hay un buen ajuste en las predicciones, pero, al igual que ocurre con los bonos marrones, también los datos del extremo derecho se podrían mejorar.

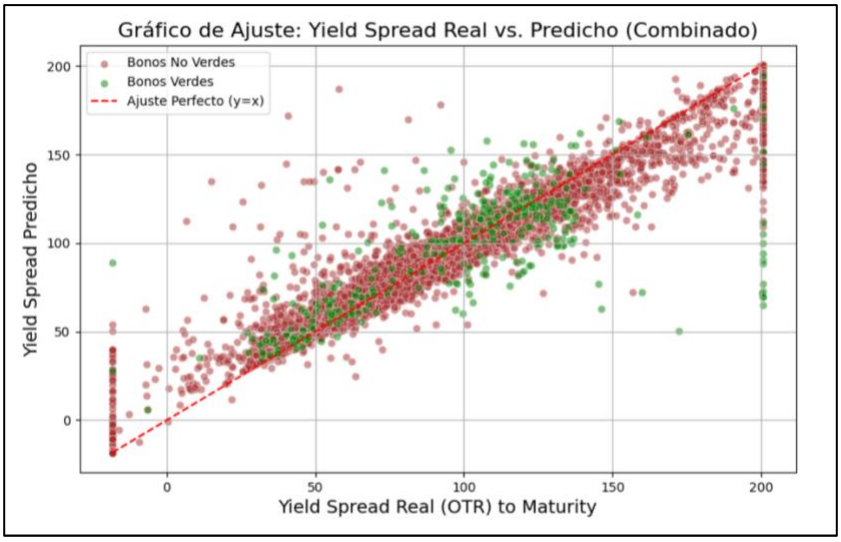
Gráfico 13: Ajuste de las predicciones de los bonos no verdes respecto a las observaciones reales utilizando el modelo Random Forest



Fuente propia. Elaborado con la herramienta Python (resultados en script).

Por último, el gráfico nº14, es una combinación de los gráficos anteriores. Es, por tanto, el gráfico de los valores predichos y valores reales para nuestros dos tipos de bonos: los bonos verdes y marrones.

Gráfico 14: Ajuste de las predicciones de los bonos verdes y no verdes respecto a las observaciones reales utilizando el modelo Random Forest



Fuente propia. Elaborado con la herramienta Python (resultados en script).

El último paso en nuestro modelo es la verificación de nuestras hipótesis. Por eso, hacemos dos contrastes de hipótesis (y ambos con niveles de confianza distintos).

En el primer contraste, sobre la existencia de greenium (test de igualdad de medias), obtenemos evidencia empírica estadísticamente significativa al 10% sobre la existencia de una ligerísima prima verde (positiva) en bonos corporativos de la UE en el periodo 2014-2025, si bien dicha existencia no es significativa al 5%.

En el 2º contraste, sobre el signo positivo de dicha ligera diferencia de rentabilidades o greenium (mayor la rentabilidad en los bonos verdes que en los marrones), obtenemos que dicha prima es positiva y significativa estadísticamente al 5%.

Así se obtiene en los siguientes outputs del script, el 1º de ellos con ambos contrastes para un nivel de confianza del 95% y el 2º con ambos contrastes para un nivel de confianza del 90%.

```
Test 1 (dos colas) — H0: medias iguales
  t = 1.6565, p-valor = 0.0984
  No rechazamos H0: no hay evidencia de diferencia de medias.

Test 2 (una cola) — H0: media_verdes ≤ media_marrones; H1: media_verdes > media_marrones
  t = 1.6565, p-valor (una cola) = 0.0492
  Rechazamos H0: la media de los bonos verdes es significativamente mayor.
```

```
 === Resumen de Greenium ===
 Media Spread verde : 97.50
 Media Spread marrón: 94.51
 Greenium (Δ = verdes − marrones): 2.99
 IC 90 % Greenium: [0.01, 5.97]

 Test 1 (dos colas) — H₀: μ_verdes = μ_marrones
   t = 1.6565, p-valor = 0.0984  (α = 0.1)
   Rechazamos H₀: existe diferencia significativa.

 Test 2 (una cola) — H₀: μ_verdes ≤ μ_marrones
                 H₁: μ_verdes > μ_marrones
   t = 1.6565, p-valor (una cola) = 0.0492  (α = 0.1)
   Rechazamos H₀: la media de los bonos verdes es significativamente mayor.
```

3. CONCLUSIONES

Existe en la literatura académica un debate aún inconcluso sobre si existe una prima verde (greenium) y, sobre el signo de esta. Se trata de un tema muy investigado pero para el cual la evidencia empírica no es unánime. Aunque la evidencia existente en la actualidad es mayoritaria para una greenium negativa (es decir, favorable para la emisión de bonos verdes por su menor coste de emisión para las instituciones emisoras y desfavorable para

el inversor), la diversidad de resultados sobre si la greenium es positiva o negativa depende en gran medida de las características de la muestra de bonos analizados en cuanto al emisor (soberanos, corporativos, que incluyan o no el sector financiero, etc.), geografía (EE.UU., China, Europa, etc.), periodo temporal, mercado (primario o secundario o ambos), etc. Por otro lado, la inmensa mayoría de los estudios publicados utilizan métodos econométricos tradicionales, siendo muy escasos los que utilizan modelos de ML. Por ello, nuestro estudio ha pretendido contribuir a dicho debate aportando nueva evidencia empírica sobre el tema, en concreto sobre los bonos corporativos de la UE (sector energía, transporte, banca y telefonía) durante el periodo 2014-2025, aplicando técnicas de ML.

Para identificar si existe greenium y analizar el signo de esta, sobre una muestra válida total de 7.589 bonos, de los cuales 351 son verdes (un 4,6% del total), hemos entrenado 3 modelos: regresión lineal, random forest y gradient boosting, para predecir la rentabilidad de los bonos marrones. El modelo con mejor capacidad predictiva ha sido random forest, el cual se ha utilizado para medir la rentabilidad de los bonos verdes. Se ha calculado la greenium como la diferencia de medias de ambas rentabilidades, de los bonos verdes menos la de los bonos marrones.

La conclusión que obtenemos al respecto es que sí existe una pequeña pero estadísticamente significativa prima verde de +2,9927 puntos básicos (estadísticamente significativa al 90%, pero no al 95%) y que ésta es positiva para el inversor, es decir, la rentabilidad a fecha de emisión de los bonos verdes es mayor que la rentabilidad de los bonos marrones (no obtenemos significatividad estadística para el signo positivo, al 95%).

Nuestro trabajo aporta nueva evidencia empírica, y es coherente con esa falta de unanimidad sobre la existencia o no (nuestra prima verde es positiva y solo estadísticamente significativa al 90%) y con esa disparidad en torno a su signo (en nuestro caso resulta positiva para el inversor y obtenemos evidencia estadísticamente significativa al 95%).

Así, pues, nuestros resultados están en línea con la corriente mayoritaria que confirma la existencia de greenium, y, por el contrario, se enmarca en la corriente minoritaria de estudios que obtienen una prima positiva (favorable para el inversor en términos de

rentabilidad y por tanto desfavorable para el emisor en términos de coste de financiación. Contribuimos, pues a reforzar la evidencia ya existente sobre la existencia de la greenium y a aumentar la minoritaria evidencia sobre el signo positivo de la misma.

Los resultados deben analizarse con precaución dadas las limitaciones de nuestro estudio. Por un lado hemos usado muestras independientes, no pareadas. Por otro lado, aunque nuestra muestra contiene un porcentaje de bonos verdes similar al que realmente existe en los mercados (nosotros un 4,4% frente a la realidad del mercado en torno al 3%), el volumen de los bonos verdes emitidos en el periodo de 2014 a 2025, apenas llegaban a 2.000.

Futuras líneas de trabajo

A partir de los resultados de nuestro trabajo y de sus limitaciones, proponemos dos líneas futuras de mejora e investigación:

- Obtener muestras pareadas de bonos verdes y marrones que sí sean directamente comparables en cuanto a las características de estos (vencimiento, emisor, etc.), en lugar de muestras independientes como ha sido nuestro caso.). Ya sea aplicando Propensity Matching Score (PMS) en la línea de Gianfrate & Peri (2019) o ya sea mediante otras técnicas de emparejamiento (matching) como la utilizada por Zerbib (2019).

- La aplicación de otras técnicas de ML que sigan estudiando el tema. Creemos útil utilizar Text mining o técnicas avanzadas de Natural Language Processing (NLP), como los Transformer-Large Language Models (LLMs) para extraer información textual relevante. Si bien ya existe algún estudio en este sentido como el de Fu et.al (2024), ellos lo aplican a China y utilizan análisis de sentimiento de medios de comunicación digitales e impresos. Nosotros plantearemos el análisis del contenido de los folletos de emisión de los bonos de manera que, se puedan incorporar variables predictoras nuevas al modelo de ML que hemos planteado en nuestra investigación de prima verde. Para incorporarlas deberíamos aplicar las transformaciones necesarias del text mining. De este modo, una vez que normalicemos las variables dándole un significado de interés, determinaremos si dichas variables textuales influyen o no en la prima verde, es decir, si son determinantes para nuestro estudio. Por otro lado, al comparar los modelos de ML predictivos solo con variables numéricas con los modelos que, además incorporan

las variables textuales, podremos ver si dichas variables textuales mejorarán el poder predictivo del modelo de ML propuesto para determinar de manera objetiva la prima verde.

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

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

Por la presente, yo, **MARÍA SÁNCHEZ GARCÍA**, estudiante de **BUSINESS ANALYTICS Y RELACIONES INTERNACIONALES** de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado **"¿Existe una prima verde (greenium) en el mercado de bonos corporativos de la U.E.? Evidencia empírica con modelos de Machine Learning",** declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Referencias: Usado juntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
4. Interpretador de código: Para realizar análisis de datos preliminares.
5. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
6. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
7. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
8. 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.

BIBLIOGRAFÍA

Ando, S., Fu, C., Roch, F., & Wiriadinata, U. (2024). How large is the sovereign greenium? *Oxford Bulletin of Economics and Statistics*, *86*(6), 1472-1483. https://doi.org/10.1111/obes.12619

Apergis, N., Chesini, G., & Poufina, T. (2024). The Yield of Green Bank Bonds. In: Lehner, O. M., Harrer, T., Silvola, H., & Weber, O. (Eds.). *The Routledge handbook of green finance,* 189-211. New York: Routledge.

Barclays (2022). Breaking down the greemium. Macro & Credit Research.

Barclays (2015). *The Cost of Being Green* [Credit research]. https://www.environmentalfinance.com/assets/files/US_Credit_Focus_The_Cost_of_Being_Green.pdf

Bhutta, U. S., Tariq, A., Farrukh, M., Raza, A., & Iqbal, M. K. (2022). Green bonds for sustainable development: Review of literature on development and impact of green bonds. *Technological Forecasting and Social Change*, *175*, 121378. https://doi.org/10.1016/j.techfore.2021.121378

Bonilla, J. R., & Vásquez, A. S. (2023). Greenium en Colombia: estudio de caso del mercado de bonos verdes a partir de un modelo estructural de dos factores. *Cuadernos de Economía (Santafé de Bogotá)*, *42*(90), 517-548.

Caramichael, J., & Rapp, A. C. (2024). The green corporate bond issuance premium. *Journal of Banking & Finance*, *162*, 107126. https://doi.org/10.1016/j.jbankfin.2024.107126

Chesini, G. (2024). Can Sovereign Green Bonds Accelerate the Transition to Net-Zero Greenhouse Gas Emissions? *International Advances in Economic Research*, *30*(2), 177-197. https://doi.org/10.1007/s11294-024-09900-6

Climate Bonds Initiative (CBI) (2024). *Climate Bonds Standard*. Globally recognised, Paris-aligned Certification of debt instruments, entities and assets using robust, science-based methodologies. Updated June 2024. Version 4.2. https://www.climatebonds.net/files/documents/CBI_Standard_V4-2.pdf

Climate Bonds Initiative (CBI) (November 2024). *Sustainable Debt Market Summary Q3 2024*. Retrieved from: https://www.climatebonds.net/files/reports/cbi_mr_q3_2024_01c.pdf

CNMV. (s.f.). Finanzas sostenibles. Recuperado de: https://www.cnmv.es/Portal/Finanzas-Sostenibles/Indice

Di Tommaso, C., Perdichizzi, S., Vigne, S., & Zaghini, A. (2024). *Is the government always greener?* (No. 718). CFS Working Paper Series. http://dx.doi.org/10.2139/ssrn.4746637

European Commission. (2020). REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020 on the establishment of a framework to facilitate sustainable investment and amending Regulation (EU) 2019/2088. Official Journal of European Union. Recuperado de: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020R0852&from=EN

Fatica, S., Panzica, R., & Rancan, M. (2021). The pricing of green bonds: are financial institutions special? *Journal of Financial Stability*, *54*, 100873. https://doi.org/10.1016/j.jfs.2021.100873

Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, *142*(3), 499-516. https://doi.org/10.1016/j.jfineco.2021.01.010

Fu, Y., He, L., Liu, R., Liu, X., & Chen, L. (2024). Does heterogeneous media sentiment matter the 'green premium'? Empirical evidence from the Chinese bond market. *International Review of Economics & Finance*, *92*, 1016-1027. https://doi.org/10.1016/j.iref.2024.02.076

Gianfrate, G., & Peri, M. (2019). The green advantage: Exploring the convenience of issuing green bonds. *Journal of Cleaner Production*, *219*, 127-135. https://doi.org/10.1016/j.jclepro.2019.02.022

Gibson Brandon, R., Glossner, S., Krueger, P., Matos, P. & Steffen, T. (2022). Do responsible investors invest responsibly? *Review of Finance, 26*(6), 1389–432. https://doi.org/10.1093/rof/rfac064

González Pacheco, V. (2019). *Una Breve Historia del Machine Learning*. Telefónica Tech. Recuperado de: https://telefonicatech.com/blog/una-breve-historia-del-machine-learning

Hinsche, I. C. (2021). A Greenium for the Next Generation EU Green Bonds Analysis of a Potential Green Bond Premium and its Drivers. *Center for Financial Studies Working Paper 663*. https://d-nb.info/124636414X/34

Hyun, S., Park, D., & Tian, S. (2020). The price of going green: the role of greenness in green bond markets. *Accounting & Finance*, *60*(1), 73-95. https://doi.org/10.1111/acfi.12515

ICMA (2024). *Green Bond Principles. Guidance Handbook*. November. https://www.icmagroup.org/assets/documents/Sustainable-finance/2024-updates/The-Principles-Guidance-Handbook-November-2024-041124.pdf

ICMA (2023). Climate Transition Finance Handbook. Guidance for Issuers. June. https://www.icmagroup.org/assets/documents/Sustainable-finance/2023-updates/Climate-Transition-Finance-Handbook-CTFH-June-2023-220623v2.pdf

ICMA (2021). *The Green Bond Principles: Voluntary Process Guidelines for Issuing Green Bonds*. https://www.icmagroup.org/assets/documents/Sustainable-finance/2022-updates/Green-Bond-Principles-June-2022-060623.pdf

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Karpf, A., & Mandel, A. (2018). The changing value of the 'green' label on the US municipal bond market. *Nature Climate Change*, *8*(2), 161-165. https://doi.org/10.1038/s41558-017-0062-0

Kocaarslan, B. (2024). The impact of liquidity conditions on the time-varying link between U.S. municipal green bonds and major risky markets during the COVID-19 crisis: A machine learning approach. *Energy Policy*, (184), 113911. https://doi.org/10.1016/j.enpol.2023.113911

Kocaarslan, B., & Soytas, U. (2023). The role of major markets in predicting the U.S. municipal green bond market performance: New evidence from machine learning models. *Technological Forecasting and Social Change, 196*, 122820. https://doi.org/10.1016/j.techfore.2023.122820

La Torre, M., & Leo, S. (2024). *Contemporary Issues in Sustainable Finance: Banks, Instruments, and the Role of Women*. Cham: Palgrave Macmillan.

Larcker, D. F., & Watts, E. M. (2020). Where's the greenium? *Journal of Accounting and Economics*, *69*(2-3), 101312. https://doi.org/10.1016/j.jacceco.2020.101312

Li, W., Hu, H., & Hong, Z. (2024). Green finance policy, ESG rating, and cost of debt. Evidence from China. *International Review of Financial Analysis*, *92*, 103051. https://doi.org/10.1016/j.irfa.2023.103051

Liang, H., & Renneboog, L. (2020). Corporate Social Responsibility and Sustainable Finance: A Review of the Literature. *Oxford Research Encyclopedia of Economics and Finance.* https://doi.org/10.1093/acrefore/9780190625979.013.592

Löffler, K. U., Petreski, A., & Stephan, A. (2021). Drivers of green bond issuance and new evidence on the "greenium". *Eurasian Economic Review*, *11*(1), 1-24. https://doi.org/10.1007/s40822-020-00165-y

MackAskill, S., Roca, E., Liu, B., Stewart, R.A., & Sahin, O. (2021). Is there a green premium in the green bond market? Systematic literature review revealing premium determinants. *Journal of Cleaner Production*, *280*, 124491. https://doi.org/10.1016/j.jclepro.2020.124491

Manzano Romero, D., et al (2024). Guía del Sistema Financiero Español. Madrid: AFI Global Education.

Meyer, S. & Henide, K (2020). Searching for 'Greenium'. Evidence of a green pricing premium in the secondary Euro-denominated investment grade corporate bond market. IHS Markit. https://www.icmagroup.org/assets/documents/Sustainable-finance/Public-research/Greenium-whitepaper-110521.pdf

Migliorelli, M. (2021). What do We Mean by Sustainable Finance? Assessing Existing Frameworks and Policy Risks. *Sustainability*, *13*(2), 975. https://doi.org/10.3390/su13020975

Naciones Unidas. (2018). *UNEP FI Annual Review*. Recuperado de: https://www.unepfi.org/wordpress/wp-content/uploads/2018/11/2017-UNEP-FI-Annual-Overview.pdf

Sergei, G., & Alesya, B. (2022). In search of greenium. Analysis of yields in the european green bond markets. *Procedia Computer Science*, *214*, 156-163. https://doi.org/10.1016/j.procs.2022.11.161

Tran, T. N. (2024). Determinants of the value of green bonds in the United Kingdom. *FinTech and AI in Finance (FinTAF)*, *2*(1). https://journals.shu.ac.uk/index.php/FinTAF/article/view/443

Zerbib, O. D. (2019). The effect of pro-environmental preferences on bond prices: Evidence from green bonds. *Journal of Banking & Finance*, *98*, 39-60. https://doi.org/10.1016/j.jbankfin.2018.10.012

Zervoudi, E., Moschos, N., & Christopoulos, A. (2025). From the Corporate Social Responsibility (CSR) and the Environmental, Social and Governance (ESG) Criteria to the Greenwashing Phenomenon: A Comprehensive Literature Review About the Causes, Consequences and Solutions of the Phenomenon with Specific Case Studies. *Sustainability*, *17*(5), 2222. https://doi.org/10.3390/su17052222

ANEXO: CÓDIGO

Enlace al código de Python: Enlace al script del código

Enlace a los datos: Excel bonos corporativos de la UE 2014-2025