



www.turkishstudies.net/economy

Turkish Studies - Economics, Finance, Politics

eISSN: 2667-5625

Research Article / Araştırma Makalesi



Financial Stress and Market Volatility in Machine Learning-Based Greenwashing Detection: Evidence from S&P 500 Companies

Makine Öğrenmesi Tabanlı Yeşil Aklama Tespitinde Finansal Stres ve Piyasa Oynaklığı: S&P 500 Şirketlerinden Bulgular

Rüya Kaplan Yıldırım*

Abstract: As sustainability reporting has taken on an increasingly central role in corporate communications, the inconsistency between companies' stated environmental commitments and their actual practices has become a significant credibility issue for investors and regulatory bodies. The aim of this study is to assess the reliability of the environmental commitments made by companies in the S&P 500 index and, in this context, to develop a machine learning-based model for detecting greenwashing. A review of the existing literature on this topic reveals that the vast majority of studies focus solely on textual features and largely disregard financial and market-based indicators. Building on this gap, this study systematically examines whether integrating textual features with financial stress factors and market volatility into a single model improves classification performance. Accordingly, using data from 503 companies, natural language processing techniques (ClimateBERT, TF-IDF, sentiment analysis, readability metrics, and expressions of uncertainty), financial structure indicators (leverage ratio, return on assets, firm size, sector sensitivity), and market volatility indicators based on the GARCH (1,1) model—totaling 21 features—were created and fed as inputs to the XGBoost classifier. The dependent variable was obtained by converting the ISS governance risk score into a binary class using a dynamic thresholding method. According to the findings, the model demonstrated satisfactory classification performance with an ROC-AUC value of 0.704 and a Spearman rank correlation coefficient of 0.756. The results of the SHAP analysis revealed that the signals with the highest explanatory power were market volatility indicators, although textual features also made a significant contribution to model performance. Our findings also indicate that textual features play a complementary role by increasing the model's sensitivity during the decision-making process. In this context, the study demonstrates that, rather than relying on a single data source for greenwashing detection, multi-channel model architectures that simultaneously evaluate corporate discourse, financial structure, and market reaction produce more reliable and consistent results, and recommends the adoption of such integrated approaches for future research.

Keywords: Greenwashing, ESG, Financial Pressure, Machine Learning, Financial Structure Indicators.

* Asst. Prof. Dr., Adana Alparslan Türkeş Science and Technology University, Faculty of Economics, Administrative and Social Sciences, Department of International Trade and Finance

Dr. Öğr. Üyesi, Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, İktisadi, İdari ve Sosyal Bilimler Fakültesi, Uluslararası Ticaret ve Finansman Bölümü

ORCID <https://orcid.org/0000-0003-0455-568X>

rkaplan@atu.edu.tr

Cite as/ Atıf: Yıldırım, R. K. (2026). Financial stress and market volatility in machine learning-based greenwashing detection: Evidence from S&P 500 companies / Makine öğrenmesi tabanlı yeşil aklama tespitinde finansal stres ve piyasa oynaklığı: S&P 500 şirketlerinden bulgular. *Turkish Studies – Economics, Finance, Politics*, 21(Ö1), 1565-1629. [Special Issues Dedicated to Prof. Dr. Mahmut Özer], Ankara Bilim University Publications. <https://dx.doi.org/10.7827/TurkishStudies.95672>

Received/Geliş: 04 May/Mayıs 2026

Accepted/Kabul: 15 September/Eylül 2026

Published/Yayın: 20 September/Eylül 2026

Checked by plagiarism software



Öz: Sürdürülebilirlik raporlamasının kurumsal iletişimde giderek merkezi bir konuma gelmesiyle birlikte, şirketlerin beyan ettikleri çevresel taahhütler ile fiili uygulamaları arasındaki tutarsızlık, yatırımcılar ve düzenleyici kurumlar açısından önemli bir güvenilirlik sorunu haline gelmiştir. Bu çalışmanın amacı, S&P 500 endeksinde yer alan şirketlerin çevresel taahhütlerinin güvenilirliğini değerlendirmek ve bu bağlamda makine öğrenmesi tabanlı bir yeşil aklama (greenwashing) tespit modeli geliştirmektir. Konuyla ilgili mevcut literatür incelendiğinde, çalışmaların büyük çoğunluğunun yalnızca metinsel özelliklere odaklandığı ve finansal ile piyasa temelli göstergeleri büyük ölçüde göz ardı ettiği görülmektedir. Bu çalışma ise söz konusu boşluktan hareketle, metinsel özellikler ile finansal baskı unsurlarının ve piyasa oynaklığının tek bir modelde bütünleştirilmesinin sınıflandırma performansını artırıp artırmadığını sistematik biçimde incelemektedir. Bu doğrultuda, 503 şirkete ait veriler kullanılarak doğal dil işleme teknikleri (ClimateBERT, TF-IDF, duygu analizi, okunabilirlik ölçütleri ve belirsizlik ifadeleri), finansal yapı göstergeleri (kaldıraç oranı, aktif kârlılığı, firma büyüklüğü, sektörel duyarlılık) ve GARCH (1,1) modeline dayalı piyasa oynaklığı göstergeleri olmak üzere toplam 21 özellik oluşturulmuş ve bu özellikler XGBoost sınıflandırıcısına girdi olarak sunulmuştur. Bağımlı değişken, ISS yönetim risk skorunun dinamik eşikleme yöntemiyle ikili bir sınıfa dönüştürülmesiyle elde edilmiştir. Elde edilen bulgulara göre model, 0,704 ROC-AUC değeri ve 0,756 Spearman sıra korelasyon katsayısı ile tatmin edici bir sınıflandırma performansı sergilemiştir. SHAP analizi sonuçları, en yüksek açıklayıcılığa sahip sinyallerin piyasa oynaklığı göstergeleri olduğunu, ancak metinsel özelliklerin de model performansına önemli katkı sağladığını ortaya koymuştur. Bulgularımız ayrıca, metinsel özelliklerin karar aşamasında modelin duyarlılığını artırarak tamamlayıcı bir işlev üstlendiğine işaret etmektedir. Bu bağlamda çalışma, yeşil aklama tespitinde tek bir veri kaynağına dayanmak yerine, kurumsal söylemi, finansal yapıyı ve piyasa tepkisini eş zamanlı olarak değerlendiren çok kanallı model mimarilerinin daha güvenilir ve tutarlı sonuçlar ürettiğini göstermekte, gelecekteki çalışmalar için de bu tür bütünsel yaklaşımların benimsenmesini önermektedir.

Anahtar Kelimeler: Yeşil Aklama, ESG, Finansal Baskı, Makine Öğrenmesi, Finansal Yapı Göstergeleri.

Structured Abstract:

Most of the existing greenwashing detection literature relies almost exclusively on textual disclosure features, leaving the potential contribution of firm-level financial pressure and market-based signals largely unexamined. This study aims to build a machine-learning-based greenwashing detection model that evaluates the credibility of the environmental commitments made by S&P 500 companies by combining textual, financial, and market-volatility information. Specifically, we investigate whether incorporating GARCH-based conditional stock-return volatility as a proxy for investor skepticism toward corporate environmental claims, alongside natural language processing (NLP) and financial structure indicators, improves classification performance compared to text-only approaches. This multi-channel approach addresses a glaring gap in the literature: no previous study has combined time-series market-volatility dynamics with textual and financial indicators within a single ESG-credibility classifier, and so it is important to test whether “market skepticism” reflected in volatility has any additional and independent explanatory power beyond disclosure language and balance-sheet pressure.

The study is based on a quantitative and predictive machine-learning model on a sample of 503 S&P 500 companies, which are classified by GICS sector. We constructed three independent data channels. The NLP channel uses six features from company disclosure text such as TF-IDF (the weights of nine ESG-specific terms), a Loughran-McDonald ESG-extended sentiment score, Flesch and Gunning Fog readability indices, hedge word count and a ClimateBERT (DistilRoBERTa)-based greenwashing score with temperature scaling. Five features are incorporated in the financial channel (firm size, return on assets, leverage ratio and GICS sector sensitivity). The market channel uses ten cross-sectional features (company-specific GARCH (1,1) based on daily stock returns) based on the data from a firm, though we make use of a prediction before the forecast window as they are more likely to be leaked. The 21 resulting features were aggregated into a single matrix and fed into an XGBoost classifier, with hyperparameters tuned by RandomizedSearchCV and class imbalance dealt with through a dynamically adjusted `scale_pos_weight`. The binary dependent variable was built by applying a dynamic percentile threshold to the ISS Governance QualityScore. The dataset was split 80/20 (402 training, 101 test companies) using stratified sampling and SHAP (TreeExplainer) was used for explainability. Three robustness tests were designed: ablation test, where all GARCH features were removed, isolation test using only GARCH features, and lagged-GARCH test, where volatility parameters are re-estimated only on the first 70% of return history as an anti-data leakage check.

On the held-out test set of 101 previously unknown firms, the full model has an ROC-AUC of 0.704 and Spearman rank correlation of 0.756 which is an informative classification and a strong risk ranking consistency for a model built on a very small number of explanatory variables. SHAP analysis showed that market-volatility features are the focus of the explanation of the model outputs. Three of the top five and eight of the top thirteen predictors are from the GARCH channel, with `garch_vol_last` (SHAP = 0.668) being the strongest predictor before leverage and Gunning Fog readability. However, we found that volatility alone is not enough. When the GARCH features are used in isolation, the ROC-AUC is only slightly less (0.638, -2.3%), and Average Precision is only 20.7% worse (0.467), showing that the volatility is the most predictive feature in this case, even if there is no textual or financial context. When the GARCH features are removed completely, the ROC-AUC is still only 0.012 (1.8%) and the accuracy is 7.5%, indicating that NLP and financial features are very discriminating on their own and leverage (0.662) and Gunning Fog (0.618) are the most predictive features. The lagged-GARCH robustness test using volatility parameters estimated from only the first 70% of the return history outperformed the contemporary baseline model (ROC-AUC 0.704 vs. 0.653; Spearman ρ 0.756 vs. 0.739). This indicates that if the look-ahead data leaks, the market-volatility signal is not temporary but rather is a structural relationship.

The results suggest that greenwashing risk cannot be captured in only one information source: market volatility signals are the major drivers of the ranking performance but textual and financial features are also needed to enhance the decision-level precision and reduce the number of false positives. If we use multi-channel models which assess the corporate discourse, financial structure and market response jointly, we will find that greenwashing classification is more reliable than the single channel model. The study contributes to the literature in two ways. One is that the authors, for the first time, have introduced firm-level GARCH-derived volatility dynamics into an ESG-credibility classifier, which is key to the theory that heterogeneous investor beliefs about ESG authenticity are reflected in the variance in conditional returns. Second, the model's stability in ranking is also useful for portfolio ESG screening. The authors have performed three rigorous robustness

tests and demonstrated that volatility influences a good AUC while textual complexity and hedging language affects the classification accuracy. The study is predictive rather than causal and it relies on a single governance risk provider (ISS) and proxy disclosure reports instead of complete CSR/CDP/TCFD reports; this is a cross-sectional, single-period (2023-2024) study focused on S&P 500 firms and future studies using greenwashing cases, alternative ESG rating providers, panel data and instrumental-variable or Granger-causality designs should be undertaken to establish causality and generalizability.

Keywords: Greenwashing, ESG, Financial Pressure, Machine Learning, Financial Structure Indicators.

الملخص المنظم:

تعتمد معظم الأدبيات القائمة بشأن الكشف عن الغسل الأخضر اعتمادًا شبه حصري على خصائص الإفصاحات النصية، في حين تظل الإسهامات المحتملة للضغوط المالية على مستوى الشركات والإشارات المستمدة من السوق غير مدروسة إلى حد كبير. وتهدف هذه الدراسة S&P إلى بناء نموذج قائم على تعلم الآلة للكشف عن الغسل الأخضر، يُقِيم مصداقية الالتزامات البيئية التي تعلنها الشركات المدرجة في مؤشر من خلال دمج المعلومات النصية والمالية ومعلومات تقلبات السوق. وعلى وجه التحديد، تبحث الدراسة فيما إذا كان إدراج التقلب الشرطي 500 بوصفه مؤشرًا بديلًا لتشكُّك (GARCH) لعوائد الأسهم، المقدر باستخدام نماذج الانحدار الذاتي المعتمد للتباين الشرطي غير المتجانس والهيكل المالي، يُحسِّن أداء التصنيف مقارنةً (NLP) المستثمرين في الادعاءات البيئية للشركات، إلى جانب مؤشرات معالجة اللغة الطبيعية بالمقاربات المعتمدة على النصوص وحدها. وتعالج هذه المقاربة متعددة القنوات فجوة واضحة في الأدبيات؛ إذ لم تجمع أي دراسة سابقة بين ديناميات تقلبات السوق في السلاسل الزمنية والمؤشرات النصية والمالية ضمن مُصنَّف واحد لمصداقية الممارسات البيئية والاجتماعية ومن ثم، تبرز أهمية اختبار ما إذا كان «تشكُّك السوق» (المنعكس في التقلبات، يمتلك قدرة تفسيرية إضافية ومستقلة تتجاوز (ESG) والحوكمة لغة الإفصاح والضغوط المرتبطة بالميزانية العمومية

ومصنَّفة S&P 500 تستند الدراسة إلى نموذج كمي تنبؤي قائم على تعلم الآلة، طُبِّق على عينة من 503 شركات مدرجة في مؤشر وقد جرى إنشاء ثلاث قنوات مستقلة للبيانات. تستخدم قناة معالجة اللغة الطبيعية ست (GICS) قطاعيًا وفقًا للمعيار العالمي لتصنيف القطاعات سمات مستخرجة من نصوص إفصاحات الشركات، تشمل أوزان تسعة مصطلحات خاصة بالجوانب البيئية والاجتماعية والحوكمة، محسوبة ودرجة للمشاعر مستندة إلى معجم لوغرمان-ماكدونالد الموسع ليشمل (TF-IDF) وفق مقياس تكرار المصطلح-معكوس تكرار الوثيقة ومؤشري فليش وغاينينغ فوغ لقابلية القراءة، وعدد ألفاظ التحوُّط اللغوي، ودرجة للغسل الأخضر مستندة إلى نموذج ESG مصطلحات مع تطبيق معايير درجة الحرارة. وتضم القناة المالية خمس سمات، هي حجم الشركة، والعائد على (ClimateBERT (DistilRoBERTa) أما قناة السوق فتستخدم عشر سمات مقطعية مستمدة من بيانات كل GICS الأصول، ونسبة الرافعة المالية، وحساسية القطاع وفق تصنيف خاص بالشركة ومقدر باستخدام عوائد الأسهم اليومية؛ ومع ذلك، نستخدم تنبؤًا سابقًا لنافذة التنبؤ (1,1) GARCH شركة، بالاستناد إلى نموذج نظرًا إلى ارتفاع احتمال تسرُّب البيانات المرتبطة بها. وجمعت السمات الإحدى والعشرون الناتجة في مصفوفة واحدة، وأدخلت إلى مُصنَّف ومعالجة عدم توازن الفئات من خلال التعديل الديناميكي للمعلمة RandomizedSearchCV مع ضبط المعلمة الفائقة باستخدام XGBoost وISS وبنى المتغير التابع الثنائي بتطبيق عتبة ديناميكية على مؤشر جودة الحوكمة الصادر عن scale_pos_weight. وقُسمت مجموعة البيانات بنسبة 80/20، بواقع 402 شركة للتدريب و101 شركة للاختبار، باستخدام (ISS Governance QualityScore). وتفسير مخرجات النموذج. وصُنِّمت ثلاثة اختبارات للمتانة: اختبار TreeExplainer من خلال SHAP المعاينة الطيفية، واستُخدمت قيم المتأخر زمنيًا، الذي GARCH وحدها، واختبار GARCH واختبار عزل استُخدمت فيه سمات GARCH استبعاد أزيلت فيه جميع سمات. أُعيد فيه تقدير معلمة التقلب باستخدام أول 70% فقط من السجل التاريخي للعوائد، للتحقق من عدم تسرُّب البيانات

في مجموعة الاختبار المحجوزة، التي تضم 101 شركة لم يسبق للنموذج الاطلاع عليها، حقَّق النموذج الكامل مساحةً تحت منحنى بلغت 0.704، ومعامل ارتباط رتبتي لسبيرمان بلغ 0.756، بما يشير إلى قدرة تصنيفية ذات دلالة (ROC-AUC) خصائص تشغيل المستقبل أن SHAP معلوماتية واتساق قوي في ترتيب المخاطر، بالنسبة إلى نموذج بُني على عدد محدود جدًا من المتغيرات التفسيرية. وأظهر تحليل كما، GARCH سمات تقلبات السوق تستأثر بالدور الرئيس في تفسير مخرجات النموذج. فمن بين أهم خمسة متنبئات، تنتمي ثلاثة إلى قناة بلغت 0.668، متقدمًا على الرافعة SHAP أقوى المتنبئات، بقيمة garch_vol_last تنتمي إليها ثمانية من أهم ثلاثة عشر متنبئًا، وكان المتغير بمعزل عن غيرها GARCH المالية ومؤشر غانينغ فوغ لقابلية القراءة. ومع ذلك، تبين أن التقلب وحده غير كافٍ. فعند استخدام سمات (Average Precision) انخفاضًا طفيفًا فقط، لتبلغ 0.638، بانخفاض قدره 2.3%، بينما تراجع متوسط الإحكام ROC-AUC انخفضت قيمة بنسبة 20.7% فقط، ليبلغ 0.467؛ وهو ما يُظهر أن التقلب يمثل السمة الأعلى قدرةً على التنبؤ في هذه الحالة، حتى في غياب السياق النصي أو عند 0.012 فقط، أي 1.8%، وفي دقة التصنيف عند ROC-AUC بالكامل، يظل مقدار الانخفاض في GARCH المالي. وعند إزالة سمات بما يدل على أن سمات معالجة اللغة الطبيعية والسمات المالية تتمتع بقدرة تمييزية مرتفعة في حد ذاتها، وأن الرافعة المالية، بقيمة 7.5%، المتأخر زمنيًا GARCH ومؤشر غانينغ فوغ، بقيمة 0.618، هما أكثر السمات قدرةً على التنبؤ. وقد تفوَّق اختبار المتانة باستخدام 0.662، ROC- الذي استند إلى معلمة تقلب مقدرة من أول 70% فقط من السجل التاريخي للعوائد، على النموذج المرجعي المعاصر؛ إذ بلغت قيمة نحو 0.756 مقابل 0.739. ويشير ذلك إلى أن إشارة تقلبات السوق، في حال حدوث p نحو 0.704 مقابل 0.653، وبلغ معامل سبيرمان AUC تسرُّب للبيانات المستقبلية، لا تمثل ظاهرة مؤقتة، بل تعكس علاقة هيكلية

تشير النتائج إلى أن مخاطر الغسل الأخضر لا يمكن الإحاطة بها اعتمادًا على مصدر واحد للمعلومات؛ فإشارات تقلبات السوق هي المحركات الرئيسية لأداء ترتيب المخاطر، غير أن السمات النصية والمالية ضرورية أيضًا لتعزيز الإحكام على مستوى قرارات التصنيف وتقليل عدد الحالات الإيجابية الكاذبة. ويؤدي استخدام نماذج متعددة القنوات، يُقِيم خطاب الشركات وهيكلها المالي واستجابة السوق تقييماً مشتركاً، إلى

تصنيف للغسل الأخضر أكثر موثوقية من التصنيف القائم على نموذج أحادي القناة. وتسهم الدراسة في الأدبيات من وجهين. يتمثل الأول في وهو ما ESG على مستوى الشركات ضمن مُصنّف لمصادقية GARCH إدخال الباحثين، للمرة الأولى، ديناميات التقلب المستمدة من نماذج يُعد عنصرًا محوريًا في التصور النظري القائل بأن تباين معتقدات المستثمرين بشأن أصالة الممارسات البيئية والاجتماعية والحوكمة ينعكس في التباين الشرطي للعوائد. ويتمثل الثاني في أن استقرار النموذج في ترتيب المخاطر يفيد أيضًا في فحص المحافظ الاستثمارية وفق معايير وقد أجرى الباحثون ثلاثة اختبارات صارمة للمثانة، وأظهروا أن التقلب يسهم في تحقيق أداء جيد من حيث المساحة تحت المنحنى، في ESG، حين يؤثر التعقيد النصي ولغة التحوط في دقة التصنيف. وتتسم الدراسة بطابع تنبؤي لا سببي، وتعتمد على مزود واحد لتقييم مخاطر الحوكمة (CDP) ومشروع الإفصاح عن الكربون (CSR) وعلى تقارير إفصاح بديلة بدلاً من التقارير الكاملة للمسؤولية الاجتماعية للشركات، ISS هو كما أنها دراسة مقطعية تغطي فترة واحدة (2023-2024)، وتركز على (TCFD) وفرقة العمل المعنية بالإفصاحات المالية المتعلقة بالمناخ ومن ثم، ينبغي إجراء دراسات مستقبلية تستند إلى حالات الغسل الأخضر، ومزودين بديلين لتصنيفات S&P 500 الشركات المدرجة في مؤشر وبيانات لوحية، وتصاميم بحثية قائمة على المتغيرات الأدواتية أو سببية غرانجر، للتحقق من السببية وقابلية تعميم النتائج ESG،

الضغوط المالية، تعلم الآلة، مؤشرات الهيكل، (ESG) الكلمات المفتاحية: الغسل الأخضر، الجوانب البيئية والاجتماعية والحوكمة

المالي.

Résumé Structuré:

La majeure partie de la littérature consacrée à la détection de l'écoblanchiment repose presque exclusivement sur les caractéristiques textuelles des informations publiées par les entreprises, laissant largement inexplorée la contribution potentielle des pressions financières à l'échelle de l'entreprise et des signaux de marché. Cette étude vise à élaborer un modèle de détection de l'écoblanchiment fondé sur l'apprentissage automatique, permettant d'évaluer la crédibilité des engagements environnementaux des entreprises du S&P 500 en combinant des informations textuelles, financières et relatives à la volatilité du marché. Plus précisément, nous examinons si l'intégration de la volatilité conditionnelle des rendements boursiers, estimée au moyen de modèles autorégressifs à hétéroscédasticité conditionnelle généralisée (GARCH) et utilisée comme indicateur indirect du scepticisme des investisseurs à l'égard des allégations environnementales des entreprises, améliore les performances de classification lorsqu'elle est associée au traitement automatique du langage naturel (TALN) et aux indicateurs de structure financière, comparativement aux approches exclusivement textuelles. Cette approche multicanale répond à une lacune manifeste de la littérature : aucune étude antérieure n'a combiné les dynamiques de volatilité du marché issues de séries temporelles avec des indicateurs textuels et financiers au sein d'un même classificateur de crédibilité environnementale, sociale et de gouvernance (ESG). Il importe donc de déterminer si le « scepticisme du marché », reflété dans la volatilité, possède un pouvoir explicatif supplémentaire et indépendant, au-delà du langage employé dans les publications et des pressions liées au bilan.

L'étude repose sur un modèle quantitatif et prédictif d'apprentissage automatique appliqué à un échantillon de 503 entreprises du S&P 500, réparties par secteur selon la norme mondiale de classification des secteurs (GICS). Trois canaux indépendants de données ont été construits. Le canal TALN mobilise six caractéristiques extraites des textes publiés par les entreprises, comprenant les pondérations de neuf termes propres à l'ESG calculées selon la méthode de fréquence du terme-fréquence inverse du document (TF-IDF), un score de sentiment fondé sur le lexique de Loughran-McDonald étendu au domaine ESG, les indices de lisibilité de Flesch et de Gunning Fog, le nombre de marqueurs d'atténuation et un score d'écoblanchiment fondé sur ClimateBERT (DistilRoBERTa), avec calibration par mise à l'échelle de la température. Le canal financier intègre cinq caractéristiques : la taille de l'entreprise, la rentabilité des actifs, le ratio de levier financier et la sensibilité sectorielle selon la classification GICS. Le canal de marché utilise dix caractéristiques transversales tirées des données de chaque entreprise à partir d'un modèle GARCH (1,1) spécifique, estimé sur les rendements boursiers quotidiens ; toutefois, nous utilisons une prédiction antérieure à la fenêtre de prévision, compte tenu du risque accru de fuite de ces données. Les 21 caractéristiques obtenues ont été regroupées dans une matrice unique, puis fournies à un classificateur XGBoost. Les hyperparamètres ont été optimisés à l'aide de RandomizedSearchCV, tandis que le déséquilibre entre les classes a été traité par un ajustement dynamique du paramètre scale_pos_weight. La variable dépendante binaire a été construite en appliquant un seuil de percentile dynamique au score ISS Governance QualityScore. Le jeu de données a été divisé selon un ratio de 80/20, soit 402 entreprises pour l'entraînement et 101 pour le test, au moyen d'un échantillonnage stratifié. La méthode SHAP (TreeExplainer) a été utilisée pour assurer l'explicabilité du modèle. Trois tests de robustesse ont été conçus : un test d'ablation, supprimant l'ensemble des caractéristiques GARCH ; un test d'isolement, utilisant uniquement ces caractéristiques ; et un test GARCH décalé dans le

temps, dans lequel les paramètres de volatilité sont réestimés exclusivement sur les premiers 70 % de l'historique des rendements, afin de vérifier l'absence de fuite de données.

Sur le jeu de test réservé, composé de 101 entreprises non observées auparavant par le modèle, le modèle complet atteint une aire sous la courbe ROC (ROC-AUC) de 0,704 et une corrélation des rangs de Spearman de 0,756. Ces résultats témoignent d'une capacité de classification informative et d'une forte cohérence dans le classement des risques pour un modèle reposant sur un nombre très limité de variables explicatives. L'analyse SHAP montre que les caractéristiques de volatilité du marché occupent une place centrale dans l'explication des sorties du modèle. Trois des cinq principaux prédicteurs et huit des treize prédicteurs les plus importants proviennent du canal GARCH ; *garch_vol_last* constitue le prédicteur le plus influent (SHAP = 0,668), devant le levier financier et l'indice de lisibilité de Gunning Fog. Toutefois, nous constatons que la volatilité seule ne suffit pas. Lorsque les caractéristiques GARCH sont utilisées isolément, la ROC-AUC n'est que légèrement inférieure (0,638 ; -2,3 %), tandis que la précision moyenne (*Average Precision*) ne diminue que de 20,7 %, pour s'établir à 0,467. Ce résultat indique que la volatilité constitue, dans ce cas, la caractéristique la plus prédictive, même en l'absence de contexte textuel ou financier. Lorsque les caractéristiques GARCH sont entièrement supprimées, la diminution de la ROC-AUC demeure limitée à 0,012 (1,8 %) et celle de l'exactitude de classification à 7,5 %, ce qui indique que les caractéristiques TALN et financières possèdent, à elles seules, un fort pouvoir discriminant ; le levier financier (0,662) et l'indice de Gunning Fog (0,618) sont alors les caractéristiques les plus prédictives. Le test de robustesse GARCH décalé dans le temps, fondé sur des paramètres de volatilité estimés uniquement à partir des premiers 70 % de l'historique des rendements, surpasse le modèle de référence contemporain (ROC-AUC de 0,704 contre 0,653 ; ρ de Spearman de 0,756 contre 0,739). Cela indique que, si une fuite de données futures se produit, le signal de volatilité du marché n'est pas temporaire, mais reflète plutôt une relation structurelle.

Les résultats suggèrent que le risque d'écoblanchiment ne peut être appréhendé à partir d'une seule source d'information : les signaux de volatilité du marché constituent les principaux déterminants de la performance de classement, mais les caractéristiques textuelles et financières sont également nécessaires pour améliorer la précision des décisions de classification et réduire le nombre de faux positifs. L'utilisation de modèles multicanaux évaluant conjointement le discours des entreprises, leur structure financière et la réaction du marché permet d'obtenir une classification de l'écoblanchiment plus fiable que celle issue d'un modèle à canal unique. L'étude apporte deux contributions à la littérature. Premièrement, les auteurs introduisent, pour la première fois, des dynamiques de volatilité dérivées de modèles GARCH à l'échelle de l'entreprise dans un classificateur de crédibilité ESG. Cette contribution est centrale pour la proposition théorique selon laquelle l'hétérogénéité des croyances des investisseurs quant à l'authenticité des pratiques ESG se reflète dans la variance conditionnelle des rendements. Deuxièmement, la stabilité du modèle en matière de classement présente également un intérêt pour le filtrage des portefeuilles selon des critères ESG. Les auteurs ont réalisé trois tests de robustesse rigoureux et montré que la volatilité contribue à une bonne performance en matière d'AUC, tandis que la complexité textuelle et les marqueurs d'atténuation influencent l'exactitude de classification. L'étude est prédictive plutôt que causale et repose sur un seul fournisseur d'évaluations du risque de gouvernance (ISS), ainsi que sur des rapports de divulgation utilisés comme substituts, plutôt que sur des rapports complets de responsabilité sociale des entreprises (CSR), du Carbon Disclosure Project (CDP) ou de la Task Force on Climate-related Financial Disclosures (TCFD). Il s'agit d'une étude transversale couvrant une seule période (2023–2024) et centrée sur les entreprises du S&P 500. Des recherches futures mobilisant des cas d'écoblanchiment, d'autres fournisseurs de notations ESG, des données de panel et des dispositifs méthodologiques fondés sur les variables instrumentales ou la causalité au sens de Granger devraient donc être menées afin d'établir la causalité et la généralisabilité des résultats.

Mots-clés: Écoblanchiment, ESG, pression financière, apprentissage automatique, indicateurs de structure financière.

Resumen Estructurado:

La mayor parte de la literatura existente sobre la detección del lavado verde (*greenwashing*) se basa casi exclusivamente en las características textuales de la información divulgada, mientras que la posible contribución de las presiones financieras a nivel empresarial y de las señales de mercado permanece, en gran medida, inexplorada. Este estudio tiene como objetivo desarrollar un modelo de detección del lavado verde

basado en aprendizaje automático que evalúe la credibilidad de los compromisos ambientales de las empresas del S&P 500 mediante la integración de información textual, financiera y relativa a la volatilidad del mercado. Específicamente, se examina si la incorporación de la volatilidad condicional de los rendimientos bursátiles, estimada mediante modelos de heterocedasticidad condicional autorregresiva generalizada (GARCH), como variable sustitutiva del escepticismo de los inversores ante las declaraciones ambientales corporativas, junto con indicadores de procesamiento del lenguaje natural (PLN) y de estructura financiera, mejora el desempeño de clasificación frente a los enfoques exclusivamente textuales. Esta aproximación multicanal aborda una laguna manifiesta en la literatura: ningún estudio previo ha integrado las dinámicas de volatilidad del mercado observadas en series temporales con indicadores textuales y financieros en un único clasificador de credibilidad ambiental, social y de gobernanza (ESG). Por consiguiente, resulta pertinente comprobar si el «escepticismo del mercado» reflejado en la volatilidad posee una capacidad explicativa adicional e independiente respecto del lenguaje de divulgación y de las presiones del balance.

El estudio adopta un enfoque cuantitativo y predictivo basado en aprendizaje automático, aplicado a una muestra de 503 empresas del S&P 500, clasificadas por sector conforme al Estándar Global de Clasificación Industrial (GICS). Se construyeron tres canales independientes de datos. El canal de PLN utiliza seis características extraídas de los textos de divulgación corporativa, entre ellas las ponderaciones de nueve términos específicos de ESG calculadas mediante la frecuencia de término–frecuencia inversa de documento (TF-IDF), una puntuación de sentimiento basada en el léxico de Loughran–McDonald ampliado al ámbito ESG, los índices de legibilidad de Flesch y Gunning Fog, el recuento de expresiones de atenuación y una puntuación de lavado verde basada en ClimateBERT (DistilRoBERTa), calibrada mediante escalado de temperatura. El canal financiero incorpora cinco características: tamaño de la empresa, rentabilidad sobre los activos, ratio de apalancamiento y sensibilidad sectorial según el GICS. El canal de mercado utiliza diez características transversales obtenidas de los datos de cada empresa mediante un modelo GARCH (1,1) específico, estimado a partir de los rendimientos bursátiles diarios; no obstante, se utiliza una predicción anterior a la ventana de pronóstico, dada la mayor probabilidad de filtración de esos datos. Las 21 características resultantes se integraron en una única matriz y se introdujeron en un clasificador XGBoost, cuyos hiperparámetros se ajustaron mediante RandomizedSearchCV. El desequilibrio entre clases se abordó mediante un ajuste dinámico del parámetro `scale_pos_weight`. La variable dependiente binaria se construyó aplicando un umbral percentilar dinámico al ISS Governance QualityScore. El conjunto de datos se dividió en una proporción de 80/20, con 402 empresas para entrenamiento y 101 para prueba, mediante muestreo estratificado, y se utilizó SHAP (TreeExplainer) para interpretar las predicciones. Se diseñaron tres pruebas de robustez: una prueba de ablación, en la que se eliminaron todas las características GARCH; una prueba de aislamiento, que utilizó exclusivamente dichas características; y una prueba con GARCH rezagado, en la que los parámetros de volatilidad se reestimaron utilizando únicamente el primer 70 % del historial de rendimientos, como comprobación frente a la filtración de datos.

En el conjunto de prueba reservado, compuesto por 101 empresas no observadas previamente por el modelo, el modelo completo alcanza un área bajo la curva ROC (ROC-AUC) de 0,704 y una correlación de rangos de Spearman de 0,756. Estos resultados indican una capacidad de clasificación informativa y una elevada consistencia en la ordenación del riesgo para un modelo construido con un número muy reducido de variables explicativas. El análisis SHAP mostró que las características de volatilidad del mercado ocupan una posición central en la explicación de los resultados del modelo. Tres de los cinco principales predictores y ocho de los trece más relevantes proceden del canal GARCH; `garch_vol_last` constituye el predictor más influyente (SHAP = 0,668), por delante del apalancamiento y de la legibilidad medida mediante Gunning Fog. Sin embargo, los resultados muestran que la volatilidad por sí sola no es suficiente. Cuando las características GARCH se utilizan de forma aislada, el ROC-AUC es solo ligeramente inferior (0,638; –2,3 %), mientras que la precisión promedio (*Average Precision*) disminuye únicamente un 20,7 %, hasta situarse en 0,467. Esto indica que la volatilidad constituye la característica con mayor capacidad predictiva en este caso, incluso en ausencia de contexto textual o financiero. Cuando las características GARCH se eliminan por completo, la disminución del ROC-AUC sigue siendo de tan solo 0,012 (1,8 %) y la de la exactitud de clasificación es del 7,5 %, lo que indica que las características de PLN y las financieras poseen, por sí mismas, una elevada capacidad discriminativa; el apalancamiento (0,662) y Gunning Fog (0,618) son los predictores más relevantes. La prueba de robustez con GARCH rezagado, cuyos parámetros de volatilidad se estimaron utilizando únicamente el primer 70 % del historial de rendimientos, superó al modelo de referencia contemporáneo (ROC-AUC de 0,704 frente a 0,653; ρ de Spearman de 0,756 frente a 0,739). Ello indica que, si se produce una

filtración de información futura, la señal de volatilidad del mercado no es transitoria, sino que refleja una relación estructural.

Los resultados sugieren que el riesgo de lavado verde no puede captarse mediante una única fuente de información: las señales de volatilidad del mercado son los principales determinantes del desempeño en la ordenación del riesgo, pero las características textuales y financieras también son necesarias para mejorar la precisión de las decisiones de clasificación y reducir el número de falsos positivos. El uso de modelos multicanal que evalúan conjuntamente el discurso corporativo, la estructura financiera y la respuesta del mercado permite obtener una clasificación del lavado verde más fiable que la proporcionada por un modelo de un solo canal. El estudio contribuye a la literatura de dos maneras. En primer lugar, los autores incorporan, por primera vez, dinámicas de volatilidad derivadas de modelos GARCH a nivel empresarial en un clasificador de credibilidad ESG, un elemento central para el planteamiento teórico según el cual la heterogeneidad de las creencias de los inversores sobre la autenticidad de las prácticas ESG se refleja en la varianza condicional de los rendimientos. En segundo lugar, la estabilidad del modelo en la ordenación del riesgo también resulta útil para la selección de activos en carteras conforme a criterios ESG. Los autores realizaron tres pruebas rigurosas de robustez y demostraron que la volatilidad contribuye a un buen desempeño en términos de AUC, mientras que la complejidad textual y el lenguaje de atenuación influyen en la exactitud de clasificación. El estudio es predictivo, no causal, y se basa en un único proveedor de evaluaciones del riesgo de gobernanza (ISS) y en informes de divulgación utilizados como aproximaciones, en lugar de informes completos de responsabilidad social corporativa (CSR), del Proyecto de Divulgación de Carbono (CDP) o del Grupo de Trabajo sobre Divulgaciones Financieras Relacionadas con el Clima (TCFD). Asimismo, se trata de un estudio transversal correspondiente a un único período (2023–2024), centrado en empresas del S&P 500. Por tanto, se requieren investigaciones futuras que utilicen casos de lavado verde, proveedores alternativos de calificaciones ESG, datos de panel y diseños basados en variables instrumentales o en causalidad de Granger, con el fin de establecer la causalidad y la generalizabilidad de los resultados.

Palabras clave: Lavado verde, ESG, presión financiera, aprendizaje automático, indicadores de estructura financiera.

结构化摘要:

现有关于漂绿识别的研究几乎完全依赖企业披露文本的特征，而企业层面的财务压力及市场信号可能发挥的作用，在很大程度上尚未得到充分探讨。本研究旨在构建一个基于机器学习的漂绿识别模型，通过整合文本信息、财务信息与市场波动信息，评估标普500指数成分企业所作环境承诺的可信度。具体而言，本研究考察：在自然语言处理（NLP）特征和财务结构指标的基础上，引入基于广义自回归条件异方差（GARCH）模型估计的股票收益条件波动率，并将其作为投资者对企业环境声明持怀疑态度的代理变量，是否能够较仅使用文本信息的方法提升分类性能。这一多通道方法旨在填补现有文献中的显著空白：此前尚无研究在单一环境、社会与治理（ESG）可信度分类器中，将市场波动的时间序列动态与文本及财务指标相结合。因此，有必要检验波动率所反映的“市场怀疑”是否具有超越披露语言与资产负债表压力的额外且独立的解释力。

本研究采用定量、预测性的机器学习模型，以503家标普500指数成分企业为样本，并依据全球行业分类标准（GICS）对其进行行业分类。研究构建了三个相互独立的数据通道。NLP通道使用从企业披露文本中提取的六项特征，包括基于词频—逆文档频率（TF-IDF）计算的九个ESG专用术语权重、基于扩展至ESG领域的Loughran–McDonald词典计算的情感得分、Flesch与Gunning Fog可读性指数、模糊限制语词数，以及基于ClimateBERT（DistilRoBERTa）并采用温度缩放进行校准的漂绿得分。财务通道纳入五项特征，即企业规模、资产收益率、财务杠杆率及基于GICS分类的行业敏感性。市场通道利用各企业的数据提取十项横截面特征，这些特征基于企业特定的GARCH（1,1）模

型，并使用股票日收益率进行估计；不过，鉴于相关数据更容易发生泄漏，本研究使用预测窗口之前的预测结果。由此得到的21项特征被整合为一个统一矩阵，并输入XGBoost分类器；模型超参数通过RandomizedSearchCV进行调优，类别不平衡问题则通过动态调整scale_pos_weight参数加以处理。二元因变量通过对ISS Governance QualityScore应用动态百分位数阈值构建。研究采用分层抽样，将数据集按80/20的比例划分为训练集和测试集，分别包含402家和101家企业，并使用SHAP（TreeExplainer）解释模型输出。研究设计了三项稳健性检验：一是消融检验，移除全部GARCH特征；二是独立通道检验，仅使用GARCH特征；三是滞后GARCH检验，仅使用收益率历史序列前70%的数据重新估计波动参数，以检验是否存在数据泄漏。

在包含101家模型此前未见企业的留出测试集上，完整模型的受试者工作特征曲线下面积（ROC-AUC）为0.704，Spearman秩相关系数为0.756。对于一个仅由极少量解释变量构建的模型而言，这些结果表明其具有一定的分类信息价值，并在风险排序方面表现出较强的一致性。SHAP分析显示，市场波动特征在解释模型输出时占据核心地位。重要性排名前五的预测变量中，有三个来自GARCH通道；排名前三的预测变量中，有八个来自该通道。其中，garch_vol_last是最具预测力的变量（SHAP = 0.668），其重要性高于财务杠杆和Gunning Fog可读性指标。然而，研究发现，仅依靠波动率并不足够。当单独使用GARCH特征时，ROC-AUC仅略有下降，为0.638，降幅为2.3%；平均精确率（Average Precision）则仅下降20.7%，降至0.467。这表明，即使缺乏文本或财务背景信息，波动率在此情形下仍是最具预测力的特征。当完全移除GARCH特征时，ROC-AUC的降幅仍仅为0.012（1.8%），分类准确率的降幅为7.5%，说明NLP特征与财务特征本身也具有较强的区分能力，其中财务杠杆（0.662）和Gunning Fog指数（0.618）是最具预测力的特征。滞后GARCH稳健性检验仅使用收益率历史序列前70%的数据估计波动参数，其表现优于同期基准模型：ROC-AUC分别为0.704和0.653，Spearman秩相关系数 ρ 分别为0.756和0.739。这表明，如果存在未来信息泄漏，市场波动信号并非暂时性现象，而是反映了一种结构性关系。

研究结果表明，仅凭单一信息来源无法充分捕捉漂绿风险：市场波动信号是风险排序性能的主要驱动因素，但文本和财务特征同样不可或缺，有助于提高分类决策层面的精确率并减少假阳性数量。采用同时评估企业话语、财务结构与市场反应的多通道模型，能够获得比单通道模型更可靠的漂绿分类结果。本研究从两个方面拓展了现有文献。第一，作者首次将企业层面基于GARCH模型提取的波动动态引入ESG可信度分类器。这对于以下理论命题具有关键意义：投资者对ESG真实性的信念存在异质性，而这种异质性会反映在收益率的条件方差中。第二，模型在风险排序方面的稳定性，也有助于开展投资组合的ESG筛选。作者实施了三项严格的稳健性检验，并表明波动率有助于取得良好的AUC表现，而文本复杂性与模糊限制性语言则影响分类准确率。本研究属于预测性研究，而非因果研究；其依赖单一治理风险评估提供商ISS，并使用替代性披露报告，而非完整的企业社会责任（CSR）报告、碳信息披露项目（CDP）报告或气候相关财务信息披露工作组（TCFD）报告。此外，本研究是一项覆盖单一时期（2023—2024年）的横截面研究，研究对象仅限于标普500指数成分企业。未来应结合漂绿案例、其他ESG评级提供商的数据、面板数据，以及工具变量或格兰杰因果关系研究设计，进一步确立因果关系并检验研究结论的可推广性。

关键词：漂绿、环境社会与治理（ESG）、财务压力、机器学习、财务结构指标。

Структурированная Аннотация:

Большинство существующих исследований, посвящённых выявлению гринвошинга, опирается почти исключительно на текстовые характеристики раскрываемой корпоративной информации, тогда как потенциальный вклад финансового давления на уровне компаний и рыночных сигналов остаётся в значительной степени неизученным. Цель настоящего исследования — разработать модель выявления гринвошинга на основе машинного обучения, позволяющую оценивать достоверность экологических обязательств компаний, входящих в индекс S&P 500, посредством объединения текстовой, финансовой информации и данных о рыночной волатильности. В частности, исследуется, позволяет ли включение условной волатильности доходности акций, оцениваемой с помощью моделей обобщённой авторегрессионной условной гетероскедастичности (GARCH) и используемой в качестве прокси-показателя скептицизма инвесторов в отношении экологических заявлений компаний, наряду с признаками обработки естественного языка (NLP) и показателями финансовой структуры, повысить качество классификации по сравнению с подходами, основанными исключительно на текстовых данных. Этот многоканальный подход направлен на восполнение существенного пробела в литературе: ранее ни одно исследование не объединяло динамику рыночной волатильности, представленную временными рядами, с текстовыми и финансовыми показателями в рамках единого классификатора достоверности экологических, социальных и управленческих характеристик (ESG). Поэтому важно проверить, обладает ли «рыночный скептицизм», отражённый в волатильности, дополнительной и независимой объясняющей способностью сверх той, которую обеспечивают язык раскрытия информации и финансовое давление, отражаемое в бухгалтерском балансе.

Исследование основано на количественной прогнозной модели машинного обучения, применённой к выборке из 503 компаний индекса S&P 500, распределённых по секторам в соответствии с Глобальным стандартом отраслевой классификации (GICS). Были сформированы три независимых канала данных. Канал NLP использует шесть признаков, извлечённых из текстов корпоративных раскрытий, включая веса девяти специализированных терминов ESG, рассчитанные методом «частота термина — обратная частота документа» (TF-IDF), оценку тональности на основе словаря Лоурана — Макдональда, расширенного для охвата терминологии ESG, индексы удобочитаемости Флеша и Ганнинга — Фога, количество маркеров лингвистического хеджирования, а также оценку гринвошинга на основе ClimateBERT (DistilRoBERTa) с применением температурного масштабирования. Финансовый канал включает пять признаков: размер компании, рентабельность активов, коэффициент финансового рычага и отраслевую чувствительность согласно классификации GICS. Рыночный канал использует десять признаков поперечного среза, полученных из данных каждой компании на основе индивидуальной модели GARCH (1,1), оценённой по ежедневной доходности акций; при этом используется прогноз, сформированный до начала прогнозного окна, поскольку соответствующие данные в большей степени подвержены риску утечки. Полученный 21 признак был объединён в единую матрицу и передан классификатору XGBoost. Гиперпараметры настраивались с помощью RandomizedSearchCV, а дисбаланс классов учитывался посредством динамической корректировки параметра `scale_pos_weight`. Бинарная зависимая переменная была сформирована путём применения динамического процентильного порога к показателю ISS Governance QualityScore. Набор данных был разделён в соотношении 80/20 посредством стратифицированной выборки: 402 компании вошли в обучающую выборку, а 101 — в тестовую. Для интерпретации модели использовался метод SHAP (TreeExplainer). Были разработаны три проверки устойчивости: абляционный тест с исключением всех признаков GARCH; тест изолированного использования, включающий только признаки GARCH; и тест с лагированным GARCH, в котором параметры волатильности переоценивались исключительно на первых 70 % исторического ряда доходности для проверки отсутствия утечки данных.

На отложенной тестовой выборке из 101 компании, ранее не представленной модели, полная модель достигает значения площади под ROC-кривой (ROC-AUC), равного 0,704, и коэффициента ранговой корреляции Спирмена, равного 0,756. Эти результаты свидетельствуют об информативности классификации и высокой согласованности ранжирования рисков для модели, построенной на весьма ограниченном числе объясняющих переменных. Анализ SHAP показал, что признаки рыночной

волатильности занимают центральное место в объяснении результатов модели. Три из пяти и восемь из тринадцати наиболее значимых предикторов относятся к каналу GARCH; наиболее сильным предиктором является `garch_vol_last` (SHAP = 0,668), опережающий финансовый рычаг и индекс удобочитаемости Ганнинга — Фога. Вместе с тем установлено, что одной волатильности недостаточно. При изолированном использовании признаков GARCH значение ROC-AUC снижается лишь незначительно — до 0,638 (–2,3 %), тогда как средняя точность положительных предсказаний (Average Precision) уменьшается лишь на 20,7 %, до 0,467. Это показывает, что в данном случае волатильность обладает наибольшей прогностической способностью даже при отсутствии текстового или финансового контекста. При полном исключении признаков GARCH снижение ROC-AUC по-прежнему составляет всего 0,012 (1,8 %), а снижение доли правильных классификаций — 7,5 %. Это указывает на высокую самостоятельную различающую способность признаков NLP и финансовых показателей; наиболее сильными предикторами при этом выступают финансовый рычаг (0,662) и индекс Ганнинга — Фога (0,618). Проверка устойчивости с лагированным GARCH, использующая параметры волатильности, оценённые только по первым 70 % исторического ряда доходности, продемонстрировала превосходство над базовой моделью с текущими оценками (ROC-AUC: 0,704 против 0,653; коэффициент Спирмена ρ : 0,756 против 0,739). Это указывает на то, что в случае утечки будущих данных сигнал рыночной волатильности представляет собой не временное явление, а структурную взаимосвязь.

Результаты свидетельствуют о том, что риск гринвошинга невозможно в полной мере выявить на основе единственного источника информации: сигналы рыночной волатильности являются основными факторами, определяющими качество ранжирования, однако текстовые и финансовые признаки также необходимы для повышения точности положительных предсказаний на уровне классификационных решений и сокращения числа ложноположительных результатов. Применение многоканальных моделей, совместно оценивающих корпоративный дискурс, финансовую структуру и реакцию рынка, позволяет получить более надёжную классификацию гринвошинга по сравнению с одноканальной моделью. Исследование вносит вклад в литературу по двум направлениям. Во-первых, авторы впервые включают динамику волатильности на уровне компаний, полученную на основе моделей GARCH, в классификатор достоверности ESG. Это имеет ключевое значение для теоретического положения о том, что неоднородность убеждений инвесторов относительно подлинности практик ESG отражается в условной дисперсии доходности. Во-вторых, устойчивость модели в ранжировании рисков также полезна для отбора активов в инвестиционные портфели по критериям ESG. Авторы провели три строгие проверки устойчивости и показали, что волатильность способствует достижению высоких значений AUC, тогда как текстовая сложность и средства лингвистического хеджирования влияют на долю правильных классификаций. Исследование носит прогностический, а не причинно-объяснительный характер и опирается на единственного поставщика оценок рисков корпоративного управления — ISS, а также на отчёты о раскрытии информации, используемые в качестве заменителей, вместо полных отчётов по корпоративной социальной ответственности (CSR), Проекту раскрытия информации о выбросах углерода (CDP) и рекомендациям Рабочей группы по раскрытию финансовой информации, связанной с климатом (TCFD). Кроме того, это исследование поперечного среза, охватывающее один период (2023–2024 гг.) и ограниченное компаниями индекса S&P 500. Для установления причинно-следственных связей и обобщаемости результатов необходимы дальнейшие исследования с использованием случаев гринвошинга, данных альтернативных поставщиков рейтингов ESG, панельных данных, а также исследовательских дизайнов, основанных на инструментальных переменных или причинности по Грейнджеру.

Ключевые слова: гринвошинг, ESG, финансовое давление, машинное обучение, показатели финансовой структуры.

संरचित सार:

ग्रीनवॉशिंग की पहचान से संबंधित अधिकांश मौजूदा साहित्य लगभग पूर्णतः प्रकटीकरण की पाठगत विशेषताओं पर निर्भर करता है, जबकि कंपनी-स्तरीय वित्तीय दबाव और बाज़ार-आधारित संकेतों के संभावित योगदान का पर्याप्त परीक्षण नहीं किया गया है। इस अध्ययन का उद्देश्य मशीन लर्निंग पर आधारित एक ग्रीनवॉशिंग पहचान मॉडल विकसित करना है, जो पाठगत, वित्तीय और बाज़ार अस्थिरता संबंधी सूचनाओं को संयोजित करके S&P 500 कंपनियों द्वारा व्यक्त

पर्यावरणीय प्रतिबद्धताओं की विश्वसनीयता का मूल्यांकन करे। विशेष रूप से, हम यह जाँचते हैं कि प्राकृतिक भाषा प्रसंस्करण (NLP) तथा वित्तीय संरचना के संकेतकों के साथ, कॉर्पोरेट पर्यावरणीय दावों के प्रति निवेशकों के संशय के परोक्ष संकेतक के रूप में GARCH-आधारित सशर्त शेयर प्रतिफल अस्थिरता को सम्मिलित करने से, केवल पाठ पर आधारित दृष्टिकोणों की तुलना में वर्गीकरण निष्पादन में सुधार होता है या नहीं। यह बहु-चैनल दृष्टिकोण साहित्य में विद्यमान एक स्पष्ट शोध-अंतराल को संबोधित करता है: किसी पूर्व अध्ययन ने एक ही पर्यावरणीय, सामाजिक एवं शासन-संबंधी (ESG) विश्वसनीयता वर्गीकारक में काल-श्रृंखला आधारित बाज़ार अस्थिरता की गतिशीलता को पाठगत और वित्तीय संकेतकों के साथ संयोजित नहीं किया है। अतः यह परीक्षण करना महत्वपूर्ण है कि अस्थिरता में परिलक्षित “बाज़ार संशय” में प्रकटीकरण की भाषा और तुलन-पत्र संबंधी दबाव से परे कोई अतिरिक्त एवं स्वतंत्र व्याख्यात्मक क्षमता है या नहीं।

यह अध्ययन एक मात्रात्मक और पूर्वानुमानपरक मशीन लर्निंग मॉडल पर आधारित है, जिसे S&P 500 की 503 कंपनियों के नमूने पर लागू किया गया है। इन कंपनियों को वैश्विक उद्योग वर्गीकरण मानक (GICS) के अनुसार क्षेत्रवार वर्गीकृत किया गया है। हमने तीन स्वतंत्र डेटा चैनल निर्मित किए। NLP चैनल में कंपनियों के प्रकटीकरण पाठों से प्राप्त छह विशेषताओं का उपयोग किया गया है। इनमें पद आवृत्ति-प्रतिलोम दस्तावेज़ आवृत्ति (TF-IDF) के आधार पर निर्धारित नौ ESG-विशिष्ट पदों के भार, ESG शब्दावली तक विस्तारित Loughran–McDonald शब्दकोश पर आधारित भाव-विश्लेषण स्कोर, Flesch और Gunning Fog पठनीयता सूचकांक, भाषिक संकोच एवं सावधानी दर्शाने वाले शब्दों की संख्या, तथा तापमान स्केलिंग के माध्यम से अंशांकित ClimateBERT (DistilRoBERTa)-आधारित ग्रीनवॉशिंग स्कोर सम्मिलित हैं। वित्तीय चैनल में पाँच विशेषताएँ शामिल की गई हैं: कंपनी का आकार, परिसंपत्तियों पर प्रतिफल, वित्तीय उत्तोलन अनुपात और GICS के अनुसार क्षेत्रगत संवेदनशीलता। बाज़ार चैनल में प्रत्येक कंपनी के डेटा से प्राप्त दस अनुप्रस्थ विशेषताओं का उपयोग किया गया है, जो दैनिक शेयर प्रतिफलों के आधार पर अनुमानित कंपनी-विशिष्ट GARCH (1,1) मॉडल पर आधारित हैं; हालाँकि, संबंधित डेटा के लीक होने की अधिक संभावना को देखते हुए, हम पूर्वानुमान अवधि से पहले के पूर्वानुमान का उपयोग करते हैं। इस प्रकार प्राप्त 21 विशेषताओं को एकल आव्यूह में एकीकृत कर XGBoost वर्गीकारक में प्रविष्ट किया गया। हाइपरपैरामीटरों का समायोजन RandomizedSearchCV द्वारा किया गया, जबकि वर्ग-असंतुलन को scale_pos_weight के गतिशील समायोजन के माध्यम से संबोधित किया गया। द्विआधारी आश्रित चर का निर्माण ISS Governance QualityScore पर एक गतिशील प्रतिशतक सीमा लागू करके किया गया। स्तरीकृत प्रतिचयन द्वारा डेटासेट को 80/20 के अनुपात में विभाजित किया गया, जिसमें प्रशिक्षण के लिए 402 और परीक्षण के लिए 101 कंपनियाँ रखी गईं। मॉडल की व्याख्येयता के लिए SHAP (TreeExplainer) का उपयोग किया गया। तीन सुदृढ़ता परीक्षण तैयार किए गए: एक विलोपन परीक्षण, जिसमें सभी GARCH विशेषताओं को हटा दिया गया; एक पृथक परीक्षण, जिसमें केवल GARCH विशेषताओं का उपयोग किया गया; और एक समय-पश्चित GARCH परीक्षण, जिसमें डेटा रिसाव की जाँच के लिए प्रतिफलों के ऐतिहासिक अभिलेख के केवल प्रारंभिक 70% भाग पर अस्थिरता प्राचलों का पुनः आकलन किया गया।

मॉडल के लिए पूर्वतः अपरिचित 101 कंपनियों के पृथक रखे गए परीक्षण समुच्चय पर पूर्ण मॉडल ने 0.704 का ROC वक्र के अंतर्गत क्षेत्रफल (ROC-AUC) और 0.756 का स्पीयरमैन कोटि सहसंबंध प्राप्त किया। अत्यंत सीमित संख्या में व्याख्यात्मक चरों पर निर्मित मॉडल के संदर्भ में, ये परिणाम सूचनाप्रद वर्गीकरण क्षमता और जोखिम क्रम-निर्धारण में उच्च संगति दर्शाते हैं। SHAP विश्लेषण से स्पष्ट हुआ कि बाज़ार अस्थिरता संबंधी विशेषताएँ मॉडल के निर्गमों की व्याख्या में केंद्रीय भूमिका निभाती हैं। पाँच सर्वाधिक महत्वपूर्ण पूर्वानुमानकों में से तीन और तेरह सर्वाधिक महत्वपूर्ण पूर्वानुमानकों में से आठ GARCH चैनल से संबंधित हैं। इनमें garch_vol_last सबसे प्रभावशाली पूर्वानुमानक है (SHAP = 0.668), जिसका स्थान वित्तीय उत्तोलन और Gunning Fog पठनीयता से ऊपर है। तथापि, हमने पाया कि केवल अस्थिरता पर्याप्त नहीं है। जब GARCH विशेषताओं का पृथक रूप से उपयोग किया जाता है, तो ROC-AUC केवल थोड़ा कम होता है (0.638; -2.3%), जबकि औसत परिशुद्धता (Average Precision) केवल 20.7% घटकर 0.467 रह जाती है। इससे संकेत मिलता है कि पाठगत या वित्तीय संदर्भ के अभाव में भी, इस स्थिति में अस्थिरता सर्वाधिक पूर्वानुमानकारी विशेषता है। जब GARCH विशेषताओं को पूरी तरह हटा दिया जाता है, तब भी ROC-AUC में गिरावट केवल 0.012 (1.8%) और वर्गीकरण सटीकता में गिरावट 7.5% रहती है। यह दर्शाता है कि NLP और वित्तीय विशेषताएँ अपने आप में उच्च विभेदकारी क्षमता रखती हैं, तथा वित्तीय उत्तोलन (0.662) और Gunning Fog (0.618) सर्वाधिक पूर्वानुमानकारी विशेषताएँ हैं। प्रतिफलों के ऐतिहासिक अभिलेख के केवल प्रारंभिक 70% भाग से अनुमानित अस्थिरता प्राचलों का उपयोग करने वाले समय-पश्चित GARCH सुदृढ़ता परीक्षण ने समकालीन आधाररेखा मॉडल से बेहतर निष्पादन किया (ROC-AUC: 0.704 बनाम 0.653; स्पीयरमैन ρ : 0.756 बनाम 0.739)। इससे संकेत मिलता है कि यदि भविष्य की सूचना का रिसाव होता है, तो बाज़ार अस्थिरता का संकेत अस्थायी न होकर एक संरचनात्मक संबंध को प्रतिबिंबित करता है।

परिणाम दर्शाते हैं कि ग्रीनवॉशिंग जोखिम को किसी एक सूचना स्रोत के माध्यम से पूर्णतः नहीं समझा जा सकता। बाज़ार अस्थिरता के संकेत जोखिम क्रम-निर्धारण के निष्पादन के प्रमुख प्रेरक हैं, किंतु वर्गीकरण निर्णयों के स्तर पर परिशुद्धता बढ़ाने और मिथ्या सकारात्मक परिणामों की संख्या कम करने के लिए पाठगत तथा वित्तीय विशेषताएँ भी आवश्यक हैं। कॉर्पोरेट विमर्श, वित्तीय संरचना और बाज़ार प्रतिक्रिया का संयुक्त मूल्यांकन करने वाले बहु-चैनल मॉडलों का उपयोग, एकल-चैनल मॉडल की तुलना में अधिक विश्वसनीय ग्रीनवॉशिंग वर्गीकरण प्रदान करता है। यह अध्ययन साहित्य में दो प्रकार से योगदान देता है। पहला, लेखकों ने पहली बार कंपनी-स्तरीय GARCH-व्युत्पन्न अस्थिरता की गतिशीलता को ESG विश्वसनीयता वर्गीकरण में सम्मिलित किया है। यह उस सैद्धांतिक प्रतिपादन के लिए महत्वपूर्ण है जिसके अनुसार ESG प्रामाणिकता के संबंध में निवेशकों की धारणाओं की विषमता प्रतिफलों के सशर्त प्रसरण में परिलक्षित होती है। दूसरा, जोखिम क्रम-निर्धारण में मॉडल की स्थिरता निवेश पोर्टफोलियो की ESG-आधारित छँटाई के लिए भी उपयोगी है। लेखकों ने तीन कठोर सुदृढ़ता परीक्षण किए और प्रदर्शित किया कि अस्थिरता अच्छे AUC निष्पादन में योगदान देती है, जबकि पाठगत जटिलता तथा भाषिक संकोच एवं सावधानी दर्शाने वाली अभिव्यक्तियाँ वर्गीकरण सटीकता को प्रभावित करती हैं। यह अध्ययन कारणात्मक के बजाय पूर्वानुमानपरक है और शासन-संबंधी जोखिम के एकमात्र मूल्यांकन प्रदाता, ISS, पर निर्भर करता है। साथ ही, इसमें कॉर्पोरेट सामाजिक उत्तरदायित्व (CSR), कार्बन प्रकटीकरण परियोजना (CDP) और जलवायु-संबंधी वित्तीय प्रकटीकरण कार्यबल (TCFD) की संपूर्ण रिपोर्टों के स्थान पर वैकल्पिक प्रकटीकरण रिपोर्टों का उपयोग किया गया है। यह S&P 500 कंपनियों पर केंद्रित, एकल अवधि (2023–2024) को समेटने वाला अनुप्रस्थ अध्ययन है। अतः कारणात्मकता और निष्कर्षों की सामान्यीकरणीयता स्थापित करने के लिए भविष्य के अध्ययनों में ग्रीनवॉशिंग के मामलों, वैकल्पिक ESG रेटिंग प्रदाताओं, पैनेल डेटा तथा साधनात्मक चरों या ग्रेंजर कारणात्मकता पर आधारित शोध अभिकल्पों का उपयोग किया जाना चाहिए।

मुख्य शब्द: ग्रीनवॉशिंग, ESG, वित्तीय दबाव, मशीन लर्निंग, वित्तीय संरचना संकेतक।

Yapılandırılmış Özet:

Yeşil aklama tespitine ilişkin literatürün büyük bir kısmı neredeyse tamamen metinsel açıklama özelliklerine dayanmakta olup, şirket düzeyindeki finansal baskı ve piyasa temelli sinyallerin potansiyel katkısı büyük ölçüde incelenmemiştir. Bu çalışma, metinsel, finansal ve piyasa oynaklığı bilgilerini birleştirerek S&P 500 şirketleri tarafından verilen çevresel taahhütlerin güvenilirliğini değerlendiren, makine öğrenimi tabanlı bir yeşil yıkama tespit modeli oluşturmayı amaçlamaktadır. Özellikle, doğal dil işleme (NLP) ve finansal yapı göstergelerinin yanı sıra, şirketlerin çevresel iddialarına yönelik yatırımcı şüpheciliğinin bir göstergesi olarak GARCH tabanlı koşullu hisse senedi getiri oynaklığının dahil edilmesinin, yalnızca metne dayalı yaklaşımlara kıyasla sınıflandırma performansını iyileştirip iyileştirmedikini araştırıyoruz. Bu çok kanallı yaklaşım, literatürdeki bariz bir boşluğu doldurmaktadır: Daha önce hiçbir çalışma, zaman serisi piyasa oynaklığı dinamiklerini metinsel ve finansal göstergelerle tek bir ESG güvenilirlik sınıflandırıcısı içinde birleştirmedir; bu nedenle, oynaklıkta yansıyan “piyasa şüpheciliğinin”, açıklama dilinin ve bilanço baskısının ötesinde ek ve bağımsız bir açıklayıcı güce sahip olup olmadığını test etmek önemlidir.

Bu çalışma, GICS sektörlerine göre sınıflandırılmış 503 S&P 500 şirketinden oluşan bir örneklem üzerinde geliştirilen nicel ve öngörücü bir makine öğrenimi modeline dayanmaktadır. Üç bağımsız veri kanalı oluşturulmuştur. NLP kanalı, şirket açıklama metinlerinden TF-IDF (dokuz ESG'ye özgü terimin ağırlıkları, Loughran-McDonald ESG genişletilmiş duyarlılık puanı, Flesch ve Gunning Fog okunabilirlik endeksleri, hedge kelime sayısı ve sıcaklık ölçeklendirmesine sahip ClimateBERT (DistilRoBERTa) tabanlı greenwashing puanı) gibi altı özelliği kullanmaktadır. Finansal kanalda beş özellik yer almaktadır (şirket büyüklüğü, aktif kârlılığı, kaldıraç oranı ve GICS sektör duyarlılığı). Piyasa kanalı, bir şirketten elde edilen verilere dayalı on kesitsel özellik (günlük hisse senedi getirilerine dayalı şirkete özgü GARCH (1,1) kullanmaktadır; ancak bu verilerin sızma olasılığı daha yüksek olması nedeni ile tahmin penceresinden önceki bir tahmin kullanılmıştır. Elde edilen 21 özellik tek bir matris halinde birleştirilerek bir XGBoost sınıflandırıcısına beslenmiştir; hiperparametreler RandomizedSearchCV ile ayarlanmış ve sınıf dengesizliği dinamik olarak ayarlanan scale_pos_weight ile giderilmiştir. İkili bağımlı değişken, ISS Governance QualityScore'a dinamik bir persentil eşiği uygulanarak oluşturulmuştur. Veri kümesi, tabakalı örnekleme yöntemi kullanılarak 80/20 oranında (402 eğitim, 101 test şirketi) bölündü ve açıklanabilirlik için SHAP (TreeExplainer) kullanılmıştır. Üç adet sağlamlık testi tasarlanmıştır: tüm GARCH özelliklerinin kaldırıldığı ablasyon testi, yalnızca GARCH özelliklerinin kullanıldığı izolasyon testi ve veri sızıntısını önlemek amacıyla volatilité parametrelerinin getiri geçmişinin yalnızca ilk %70'i üzerinde yeniden tahmin edildiği gecikmeli GARCH testi.

Daha önce bilinmeyen 101 firmadan oluşan ayrılmış test kümesinde, tam model 0,704 ROC-AUC ve 0,756 Spearman sıra korelasyonu değerlerine sahiptir; bu, çok az sayıda açıklayıcı değişken üzerine inşa edilmiş bir model için bilgilendirici bir sınıflandırma ve güçlü bir risk sıralama tutarlılığı anlamına gelir. SHAP analizi, piyasa volatilitesi özelliklerinin model çıktılarının açıklanmasında odak noktası olduğunu göstermiştir. İlk beş tahmin değişkeninden üçü ve ilk on üç tahmin değişkeninden sekizi GARCH kanalından gelmektedir; garch_vol_last (SHAP = 0,668), kaldıraç ve Gunning Fog okunabilirliğinden önce en güçlü tahmin değişkenidir. Ancak, oynaklığın tek başına yeterli olmadığını tespit ettik. GARCH özellikleri tek başına kullanıldığında, ROC-AUC değeri sadece biraz daha düşük (0,638, -%2,3) ve Ortalama Hassasiyet sadece %20,7 daha kötü (0,467) çıkmıştır; bu da, metinsel veya finansal bağlam olmasa bile, bu durumda oynaklığın en iyi tahmin eden özellik olduğunu göstermektedir. GARCH özellikleri tamamen kaldırıldığında, ROC-AUC değeri yine de sadece 0,012 (%1,8) ve doğruluk oranı %7,5'tir; bu da NLP ve finansal özelliklerin tek başlarına oldukça ayırt edici olduğunu ve kaldıraç (0,662) ile Gunning Fog (0,618) değerlerinin en iyi tahmin eden özellikler olduğunu göstermektedir. Getiri geçmişinin yalnızca ilk %70'inden tahmin edilen oynaklık parametrelerini kullanan gecikmeli GARCH sağlamlık testi, çağdaş referans modelden daha iyi performans gösterdi (ROC-AUC 0,704'e karşı 0,653; Spearman ρ 0,756'ya karşı 0,739). Bu durum, ileriye dönük verilerin sızması halinde piyasa volatilitesi sinyalinin geçici değil, yapısal bir ilişki olduğunu göstermektedir.

Sonuçlar, yeşil aklama riskinin tek bir bilgi kaynağıyla tespit edilemeyeceğini göstermektedir: Piyasa oynaklığı sinyalleri, sıralama performansının başlıca belirleyicileridir; ancak karar düzeyindeki doğruluğu artırmak ve yanlış pozitif sonuçların sayısını azaltmak için metinsel ve finansal özelliklerin de dikkate alınması gerekmektedir. Kurumsal söylemi, finansal yapıyı ve piyasa tepkisini bir arada değerlendiren çok kanallı modeller kullanırsak, yeşil yıkama sınıflandırmasının tek kanallı modele kıyasla daha güvenilir olduğunu göreceğiz. Bu çalışma, literatüre iki yönden katkı sağlamaktadır. Birincisi, yazarların ilk kez şirket düzeyinde GARCH modelinden türetilen oynaklık dinamiklerini bir ESG güvenilirlik sınıflandırıcısına dahil etmeleridir; bu, ESG'nin gerçekliği hakkındaki heterojen yatırımcı inançlarının koşullu getirilerdeki varyansa yansıdığı teorisi açısından kilit öneme sahiptir. İkincisi, modelin sıralamadaki istikrarı, portföy ESG taraması için de yararlıdır. Yazarlar, üç titiz sağlamlık testi gerçekleştirmiş ve volatilitenin iyi bir AUC değerine katkıda bulunduğunu, metinsel karmaşıklık ile riskten korunma dilinin ise sınıflandırma doğruluğunu etkilediğini göstermiştir. Bu çalışma nedensel olmaktan ziyade öngörücü nitelikte olup, eksiksiz CSR/CDP/TCFD raporları yerine tek bir yönetim riski sağlayıcısına (ISS) ve vekaleten sunulan açıklama raporlarına dayanmaktadır. Çalışma, S&P 500 şirketlerine odaklanan kesitsel, tek dönemli (2023-2024) bir çalışmadır; nedensellik ve genelleştirilebilirliği tespit etmek için gelecekte yeşil aklama vakalarını, alternatif ESG derecelendirme sağlayıcılarını, panel verilerini ve araç değişken veya Granger nedensellik tasarımlarını kullanan çalışmalar yürütülmelidir.

Anahtar Kelimeler: Yeşil Aklama, ESG, Finansal Baskı, Makine Öğrenmesi, Finansal Yapı Göstergeleri

www.turkishstudies.net/economy**Turkish Studies - Economics, Finance, Politics**

eISSN: 2667-5625

**ANKARA SCIENCE
UNIVERSITY**

Financial Stress and Market Volatility in Machine Learning-Based Greenwashing Detection: Evidence from S&P 500 Companies

Introduction

As a result of global technological advancements in the manufacturing sector, global awareness of the damage caused to nature has grown over the years. In addition to preparing financial reports, companies have begun to prepare sustainability reports that include social and environmental disclosures. At this stage, stakeholders want to monitor not only companies' financial data but also their sustainability activities through these reports. Research shows that companies prioritizing sustainability are viewed as valuable by investors (Dağıstanlı & Çelik, 2023). In recent years, sustainability and sustainability reporting have taken center stage in corporate practices. At the same time, public awareness of environmental and sustainability issues has also increased. This situation has created increasing pressure on companies to integrate social and governance (ESG) factors into their business models. In addition, they are now required to disclose their performance in these areas to the public. Furthermore, pressure from stakeholders—including investors and governments—has led large companies to improve their corporate responsibility and sustainability reporting. Sustainability reports, which show the extent to which companies are progressing toward their sustainability goals and how these goals influence their strategies, are an important resource containing valuable information for all business stakeholders. The standards that have been developed not only ensure the consistency and comparability of sustainability disclosures but also enable information users to make more informed and strategic investment decisions. The disclosures included in these reports help reduce the information asymmetry that is likely to exist among investors. However, factors such as the largely voluntary nature of ESG disclosures, regulatory gaps, investor and customer expectations, and competitive pressures may lead some businesses to make misleading sustainability claims that go beyond their actual performance (Karagül & Selimoğlu, 2026; Suta et al., 2026; Liman, 2026).

Under current conditions, it is insufficient for businesses to present only their financial reports, as this fails to highlight their activities in environmental and social areas. The corporate activities through which businesses generate ecological impacts cannot be fully assessed, and the production of comparable and useful information is not ensured. Financial reporting has been criticized for failing to facilitate an understanding of environmental, social, and governance impacts. Consequently, the need for businesses to provide additional disclosures has emerged to ensure that these impacts are understood and balanced against their economic performance (Schaltegger & Burritt, 2010). In line with these developments, environmental, social and governance (ESG) reporting has emerged as a new reporting tool that allows companies to share their non-financial performance with stakeholders. However, the widespread adoption of ESG reporting has also led to new debates of the accuracy and reliability of the reported information.

Greenwashing has long been the subject of discussion (Çelik & Reis, 2026). Greenwashing is an interdisciplinary term, but there is no single definition, and in general is a misleading marketing strategy that portrays a product or service as environmentally friendly, green, or sustainable when in fact the product or service is not (Özkaya, 2026). Greenwashing can be defined as a form of deceptive marketing in which companies make false claims regarding the environmental performance of their products or services, thereby creating a gap between stated claims and actual practices that misleads stakeholders (Lyon & Maxwell, 2011; Shakourloo & Azimli, 2025). Greenwashing is considered unethical in the corporate world in which companies are seeking to raise capital, create a good impression of their company and regulatory stance by creating an environment-conscious image. The environmental and social aspects are not only associated with the risks but with the investor trust in financial markets. This is because the growing number of ESG-related investments in companies, while increasing the number of investors who are attracted to companies that have good environmental performance, also creates information asymmetry. Furthermore, the loss of trust in investors caused by greenwashing, coupled with the fact that companies demonstrate environmentally responsible behavior, exacerbates the problem. Indeed, the increasing prevalence of ESG-focused investment simultaneously encourages investors to favor companies with strong environmental performance and exacerbates the problem of information asymmetry (Shakourloo & Azimli, 2025).

Furthermore, greenwashing has effects such as eroding investor confidence and creating bias against companies that demonstrate environmentally responsible behavior (Tepeli & Büklü, 2025). Emphasizing that stakeholders play a vital role in whether greenwashing occurs, Seele and Gatti (2017) also noted that stakeholders' perceptions and reactions to greenwashing can significantly influence a company's strategies. However, the impact of these perceptions and reactions is not limited to the strategic level; concerns among stakeholders can also affect companies' market performance. Indeed, situations that cause concern among stakeholders can lead to sudden changes in stock prices (Chen and Dagestani, 2023). In particular, when financial pressures on firms increase, their rhetoric and risk management language may shift, and they may engage in behavior that contradicts reality by adopting more positive rhetoric.

Research has shown that greenwashing can take various forms, ranging from baseless claims to deliberate misrepresentations that undermine publicly traded companies' efforts to assess their actual environmental impacts (Atabay, 2026). In addition, companies that engage in greenwashing often use complex, vague language to try to conceal their true environmental performance. First introduced in 1986 by ecologist Jay Westerveld, the concept of greenwashing has been extensively studied and debated by scholars from various disciplines, including business, environmental management, and law.

One of the key studies in this area (Karagül & Selimoğlu, 2026) noted that greenwashing practices in businesses can lead to various financial problems, such as a loss of investor confidence, a decline in market value, damage to reputation, deliberate deceptive actions, fraud, and the resulting legal violations. In contrast, Li et al. (2023) have argued that greenwashing can be an effective tool for resolving the conflict between environmental pressures and stakeholder demands within organizations.

Research on greenwashing has employed a wide range of methodological approaches. A substantial body of this literature relies on case-based financial analysis (Laufer, 2003; Marquis et al., 2016; Tepeli & Büklü, 2025; Walker & Wan, 2012), while other studies adopt content analysis (Kuşku et al., 2017; Yu et al., 2020) or discourse analysis (Dorfleitner & Utz, 2024; Thewissen, 2026) to examine corporate disclosures. In addition, comparative analysis (Cocozza, 2026; Kathan et al., 2025) and text mining techniques (Kılınç et al., 2026) have been used to identify patterns across firms and industries. Other researchers have drawn on scores provided by ESG rating agencies

(Atabay, 2026; Kathan et al., 2025; Siddique et al., 2026) or conducted multiple case studies (Karagül & Selimoğlu, 2026) to assess the extent and consequences of greenwashing practices. However, the vast majority of existing studies rely solely on textual features, and research examining the combined effect of financial stress and market volatility on corporate discourse within a single model remains limited.

Machine learning is a subfield of computer science and was defined by Arthur Samuel in 1959 as “the ability to teach computers to learn without being explicitly programmed.” Machine learning in the field of artificial intelligence: It is based on prediction, learning from data and algorithms to make predictions, and learning from data. Machine learning technology can be used to extract meaningful data from the complex structures of data and can play a key role in multidimensional problem solving in the generation of artificial intelligence (Meral, 2025). Machine learning is very useful from a financial perspective in predicting market conditions and risks; the problem of risk prediction, the development of financial instruments demand and supply/demand balance (Özgür, 2021). In reality, large language models, natural language processing (NLP) and machine learning technology are not only automating the analysis of corporate reports. Text mining and NLP technology is transforming the analysis of corporate reports in the corporate report analysis process. These technologies are now used in a wide range of applications, such as the automated analysis of financial statements, sustainability disclosures, and risk reports; decision making processes; compliance audits and transparency (Başar & Doğruel, 2025). These techniques can analyse thousands of reports at once, reduce human bias and help in decision making.

This study aims at developing a machine learning system that assesses the reliability of the environmental commitments of companies in the S&P 500 index. The system can detect such risks not only on a textual basis but also on financial structure and market perception. The model uses not only corporate texts but also financial stress indicators and market volatility as explanations as well. Therefore, we find the influence of corporate discourse and financial conditions on greenwashing behavior in general.

Materials and Methods

Sample and Data Sources

The sample for this study is 503 companies, extracted from Wikipedia's "List of S&P 500 companies." The companies are classified according to the Global Industry Classification Standard (GICS) sector taxonomy, and the financial data was gathered through the Yahoo Finance (yfinance) API and is estimated to be from 4 January 2023 to 30 December 2024 (about 500 trading days). The data was chosen to consider the 2020-2022 period in a market free of pandemic-induced macroeconomic noise. Furthermore, the period is characterized by geopolitical shocks of energy markets, leading to market volatility and global ESG reporting standards (CSRD, SEC drafts) becoming legal. GARCH models can achieve good performance in moderate forecast windows to avoid structural market breaks and to maintain stability (Starica & Granger, 2005; Hwang & Valls Pereira, 2006).

The data we have collected are divided into three categories: Firstly, companies' business descriptions on the Yahoo Finance platform (longBusinessSummary) and sector-based sustainability proxy texts were used as input data for text analysis. Secondly, financial indicators like market capitalization (Market Cap), firm size (Ln_Firm_Size), return on assets (ROA), leverage ratio (Leverage), price-to-book ratio (Market-to-Book), dividend yield (Dividend Yield), etc. were taken into account. Third, the dependent variable is based on ISS (Institutional Shareholder Services) Governance QualityScore metrics obtained from Yahoo Finance. The ISS scoring system evaluates firms in the following order of the risk: 1 is the lowest risk (see ISS, 2023) and 10 is the highest risk (see ISS, 2023). The model's target variable (y) was obtained by converting these raw scores on the scale 1-10 into a binary class using dynamic quantile thresholding.

Table 1. Data sources

Source	Content	Coverage
Wikipedia S&P 500 table	Ticker list + sector	503 firms
yfinance API	Financial ratios, ESG risk scores, stock prices	503 firms
yfinance company descriptions	Proxy ESG disclosure text (NLP input)	503 firms
Daily log-returns	For GARCH modeling	249.610 rows, 501 firms

Missing Data Management

An imputation technique based on the GICS sector median was applied to fill in missing observations in the dataset. Under this approach, a firm's missing financial indicator was first imputed using the median value of firms in the same GICS sector; when the sector median was also missing, it was imputed using the overall sample median. This strategy is consistent with the assumption of sectoral homogeneity proposed by Little and Rubin (2019) in their theory of missing data. It limits the risk of individual outliers disrupting the imputation process.

Defining the Dependent Variable

The dependent variable was determined by converting the raw ISS (Institutional Shareholder Services) Governance QualityScore data (1-10 decile scale) from Yahoo Finance into a binary categorical variable through dynamic quantile thresholding. The 503 firms in the sample were ranked by their ISS governance risk score and the threshold value was the 75th percentile of the distribution ($\text{esg_risk_raw} = 8.0$). The firms with a risk score greater than or equal to this threshold were classified as "High Risk" ($y = 1$) while the lowest risk firms were classified as "Low Risk" ($y = 0$). This is done so that no data leaks are reported (Kaufman et al., 2012). In early trials, when the target variable was obtained from the model inputs, the artificially high accuracy rate (around 98%) was evaluated as a measure of this risk. As a result of this thresholding, 349 firms (69.4%) in the sample were marked as "Low Risk" ($y = 0$) and 154 firms (30.6%) as "High Risk" ($y = 1$). Due to a tie at a raw score of 8.0, the final class distribution was slightly off the theoretical 75%/25% ratio.

GICS Sector Sensitivity Map

Companies are assigned an environmental sensitivity weight ranging from 0 to 1 based on the GICS sector classification. These weights are determined based on sectors' carbon footprint, exposure to environmental regulations, and the historical frequency of ESG-related discussions.

Table 2. GICS Sector Sensitivity Map

Sector	Sensitivity Weight
Energy	0.95
Materials	0.80
Utilities	0.75
Industrials	0.65
Consumer Staples / Consumer Discretionary	0.50
Health Care	0.45
Financials	0.40
Real Estate	0.35
InformationTechnology	0.25
Communication Services	0.20

As shown in Table 2, although the Energy sector has the highest environmental sensitivity weight (0.95), it has a relatively small share of firms in the sample (4.4%). This means that the impact of the small number of firms in the sector on the model's output may be disproportionately large compared to their share in the sample. This is a methodological limitation that must be taken into account when interpreting the findings.

Three-Channel Architecture

Based on the assumption that a single data layer cannot fully represent greenwashing behavior, this study employs an architecture that evaluates three independent information channels together.

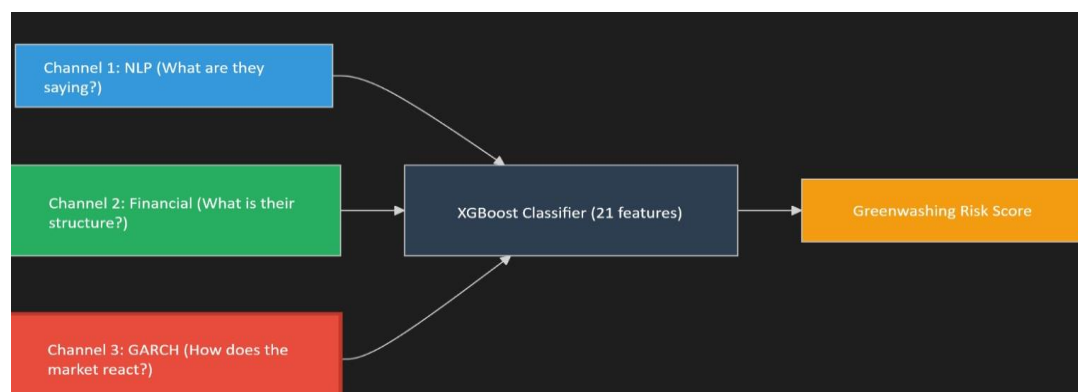


Figure 1. Three-pillar architecture

The first channel analyzes what companies say using natural language processing (NLP) techniques; the second channel explores whether the company's structure supports this narrative with financial indicators, and the third channel analyzes how the market reacts to these statements using the GARCH (1,1) model. The features learned from each channel are combined into one matrix of 21 variables and fed to the XGBoost classifier.

In the T2 (Volatility-Removed) model, while the Accuracy metric was increased by 7.5% from the Baseline model, the ROC-AUC level fell. This could be explained by the threshold-dependent nature of the Accuracy metric and the imbalanced class distribution in the target variable (69.4% vs. 30.6%). Rather than discriminating easy cases, the T2 model created an artificial increase in accuracy by focusing on the majority class (Low Risk), whereas the ROC-AUC value, which measures discriminatory power independent of the threshold, decreased.

On the other hand, the 20.7% drop in Average Precision across the T2 model, which is trained only on GARCH features, is very important. This is because market volatility is a very useful indicator for detecting potential greenwashing, but when it is used on its own it can be very noisy. Without the NLP channel, the model lost its precision by detecting all market shocks as greenwashing.

The function of each channel can be summarized by a question: the NLP channel seeks to answer "What is the company saying?" the financial channel asks "Does the company's structure support these claims?" and the GARCH channel asks "Does the market truly trust this company?" The data sources and sample features for each channel are summarized in Table 3.

Table 3. Three-Pillar architecture for greenwashing detection

Channel	Question	Data Source	Example Features
NLP	“What does the firm say?”	ClimateBERT, FinBERT, text analysis	GWL score, fog index, hedge word count
Financial	“Does the firm's structure support its claims?”	yfinance, S&P 500	Leverage, ROA, firm size
GARCH	“Does the market actually trust this firm?”	Daily stock returns → GARCH (1,1)	Conditional volatility, spike count, persistence

Channel 1: Natural Language Processing (NLP) Features

This channel is intended to investigate the text of what firms say in their environmental disclosures, at the text level. It is a well-known observation from the literature that companies which engage in greenwashing use complex and ambiguous language to hide the truth about their environmental performance. This is consistent with what is known already in financial reporting for the financial sector: Li (2008) found that annual reports of underperforming firms have a higher Fog index and are less understandable, while sustainable earnings reports are written in relatively clear English. In this paper, we believe that the same mechanism could also be true of environmental disclosures; we hypothesize that a high Fog index and low readability of environmental disclosure may be an indirect indication of a discourse strategy where information is presented formally but not properly communicated.

TF-IDF (Term Frequency–Inverse Document Frequency)

TF-IDF weights were calculated for nine key terms specific to the ESG field (sustainable, green, emissions, carbon, climate, environmental, renewable, decarbonization, net-zero). TF-IDF is a statistical method that measures the importance of a term in a specific document by weighting it with its rarity across the entire document collection (Salton & Buckley, 1988). In this way, the conceptual focus areas in companies’ sustainability discourse were quantified.

Loughran-McDonald ESG-Extended Sentiment Analysis

The sentiment score in the texts was calculated using a version of Loughran and McDonald’s (2011) finance-specific lexicon adapted to the ESG context. In addition to standard positive and negative terms, hedging words indicating uncertainty (might, could, may, perhaps, potentially) were also included and weighted by a factor of 0.7 before being added to the negative polarity.

The calculation proceeds as follows: first, the number of positive, negative, and uncertainty words in the text is counted; the number of uncertainty words is multiplied by 0.7 and added to the negative side to obtain the total negative value; finally, the formula.

$$\text{Emotion Score} = (\text{Positive} - \text{Total Negative}) \div (\text{Positive} + \text{Total Negative}) \quad (1)$$

is applied.

For example, if a text contains 10 positive, 4 negative, and 6 neutral words, the total negative score is calculated as $4 + (6 \times 0.7) = 8.2$, and the sentiment score is $(10 - 8.2) \div (10 + 8.2) \approx 0.10$, which indicates a slightly positive tone. The score ranges from -1 to $+1$; values approaching $+1$ are positive, values approaching -1 are negative, and values close to 0 indicate neutral or ambiguous discourse. This approach is consistent with findings suggesting that ambiguous language in corporate sustainability reports may be a potential indicator of greenwashing (Marquis et al., 2016).

Readability Metrics

Text complexity was assessed using two standard readability measures: the Flesch Reading Ease (Flesch, 1948), a score based on sentence length and syllable density, where lower values indicate more complex text; the Gunning Fog Index (Gunning, 1952), which is based on sentence length and the proportion of multisyllabic words, with higher values indicating text that is complex at a technical or academic level. It has been argued in the literature that high linguistic complexity in corporate sustainability reports may serve the purpose of intentionally obscuring information (obfuscation) and, therefore, could be considered a potential indicator of greenwashing (Li, 2008).

Counting Hedge Words (Expressions of Uncertainty)

To quantify the level of uncertainty in corporate reports, the number of hedge words was calculated using regular expressions (regex) (might, could, may, perhaps, possibly, probably, assume, suggest, potentially, believe). The density of hedge words per company was normalized by dividing it by the total text length.

Greenwashing Score with ClimateBERT

ClimateBERT (DistilRoBERTa) Model

ClimateBERT model developed by Webersinke et al. (2022) was used to obtain greenwashing risk scores. ClimateBERT is a transformer-based language model that is distilled to the DistilRoBERTa version of the RoBERTa architecture (Liu et al., 2019) and pre-trained on scientific publications, policy documents, and company reports on climate change. This domain-specific pre-training is helpful for the model to better describe sustainability discourse than the general language model.

For model parameters, we used a 6-layer DistilRoBERTa with around 82 million parameters; the pre-training corpus consisted of climate change research articles, IPCC reports, and corporate CDP disclosures. The total token length was limited to BERT's standard 512, and the batch size was 8 on the CPU and 16-32 on the GPU.

At this point, a clear conceptual distinction must be made: This study does not offer an econometric or causal impact analysis. The machine learning architecture developed is completely predictive. Market volatility or complex reporting languages are not the factors "causing" greenwashing; it is the algorithmic detection of the traces (symptoms) left by the company's lack of corporate transparency in the market and in the texts.

Calibration Using Temperature Scaling

We have applied the temperature scaling method to improve the reliability of the model outputs (Guo et al., 2017). This method aims at minimizing the sharpness of the model errors by dividing the raw logit values produced by the model by a temperature parameter T : $\text{adjusted score} = \text{raw score} \div T$; then the model estimates the probability with softmax. Low T values (e.g. $T = 0.3$) lead the model to produce sharper predictions closer to binary decisions while high T values (e.g. $T = 0.5$) lead to a softer probability distribution closer to intermediate values. Two different threshold values were used for the study: $T = 0.3$ was adopted for the Discourse Style Score (GWLS) to be able to distinguish the different greenwashing patterns in the discourse more clearly and $T = 0.5$ for the Symbolic Action Score (GWLE) to investigate more deeply the inconsistency in action-discourse on the action front.

Sentence-Level Inference

Each company text was split into sentences using the `'sent_tokenize'` function of the NLTK library and fed into the ClimateBERT model. The company GWLS and GWLE scores were

calculated by taking the arithmetic average of the sentence-level scores. This methodology can help to capture local greenwashing patterns which occur in long texts.

GWL Bileşik Skoru (Geometrik Ortalama)

The firm-level ultimate greenwashing (GWL) risk was calculated as the geometric mean of the discourse style (GWLS) and symbolic action (GWLE) scores:

$$GWL = \sqrt{GWLS \times GWLE} \quad (2)$$

For example, for a firm with $GWLS = 0.80$ and $GWLE = 0.60$, $GWL = \sqrt{(0.80 \times 0.60)} = \sqrt{0.48} \approx 0.693$.

The reason why we prefer the geometric mean to the arithmetic mean is that this method requires both components to be high simultaneously. For example, if one component is very high (0.90) and the other is very low (0.10), the geometric mean will only yield a low value like $\sqrt{(0.90 \times 0.10)} \approx 0.30$. This prevents companies that have issues only in the "rhetoric" or "action" dimension from being mistakenly called "high-risk." We assume that a genuine greenwashing risk requires both dimensions to be high simultaneously. The features of the NLP channel are summarized in Table 4.

Table 4. NLP features (Channel1)

#	Feature	Source	Description
1	NLP_GWL_Score	ClimateBERT	Greenwashing likelihood probability (0–1)
2	finbert_esg_sentiment	FinBERT	ESG sentiment polarity (–1 to +1)
3	flesch_reading_ease	Text analysis	Readability score (higher = simpler)
4	gunning_fog	Text analysis	Fog index (text complexity)
5	hedge_words_count	Lexicon-based	Count of hedging expressions
6	ESG_Gap	Derived	$0.3 \times \text{hedge} + 0.7 \times (1 - \text{readability}/100)$

Channel 2: Financial Characteristics

The second channel aims to evaluate the extent to which firms' financial structures support their environmental claims. So, log firm size (Ln_Firm_Size), ROA (return on assets) and leverage ratio (Leverage) were calculated from yfinance data; sector sensitivity (sector_sensitivity) was calculated, by engineering, as a weight between 0 and 1 of the GICS sector of those firms. These weights were taken into account the carbon footprint of entities, exposure to regulation and the frequency of ESG-related discussions for companies such as Energy companies (0.95 weight), and Information Technology companies (0.25 weight). The third channel is proxy_gwl_score, a text-based indicator. This is consistent with the corporate finance literature on the relationship between ESG and firm structure, in that companies with a high degree of leverage may be more likely to understate their ESG performance to lower their cost of capital. On the other hand large companies are more subject to public scrutiny (Gillan et al., 2021). In this context, the fact that sectors with high carbon intensity and regulatory exposure, such as the energy sector, are structurally riskier from the ESG perspective can also be viewed as an extension of the same logic.

Table 5. Financial features (*Channel2*)

#	Feature	Source	Description
7	Ln Firm Size	yfinance	Log market capitalization
8	ROA	yfinance	Return on assets
9	Leverage	yfinance	Debt-to-equity ratio
10	sector sensitivity	Engineered	GICS sector ESG exposure (0–1)
11	proxy_gwl_score	Engineered	Text-based composite GWL proxy

Channel 3: GARCH (1,1) Volatility Modeling

This channel aims to measure the market's reaction to corporate announcements through the conditional volatility of stock returns.

GARCH (1,1) Model

To capture market-based risk signals, a separate GARCH (1,1) model (Bollerslev, 1986) was estimated for each firm. The model operates on daily log returns (r_t , scaled by 100) with the following specification:

$$r_t = \mu + \varepsilon_t, \varepsilon_t = \sigma_t \cdot z_t, z_t \sim N(0,1) \quad (3)$$

$$\sigma_t^2 = \omega + \alpha \cdot \varepsilon_{t-1}^2 + \beta \cdot \sigma_{t-1}^2 \quad (4)$$

Here σ_t is conditional volatility (the market's perception of uncertainty on the relevant day). In the equation, the ω parameter is the firm's underlying volatility, the α parameter is the effect of the previous day's unexpected shock on today's volatility, and the β parameter is the extent to which yesterday's volatility is repeated today. The sum of α and β is generally taken as an indication of volatility persistence. When this sum is 1, it means that the effect of a shock lasts longer (Engle & Patton, 2001). So if this sum is 0.95 in a firm which is in a shock period of months, it might last for many months, and here it ends up fading after a few weeks.

The model was estimated using the ARCH Python library assuming a constant mean and normal distribution for firms with at least 100 days of observations; the estimation was completed for all 501 firms in the sample.

Derivation of Firm-Level Attributes from GARCH

The conditional volatility series for each firm was reduced to firm-level summary statistics and converted into a total of ten cross-sectional attributes, which are listed in Table 6.

The rationale for including these attributes is based on two distinct findings. First, it has been shown that investors exhibit strong and asymmetric price reactions to both positive and negative news regarding a firm's corporate social responsibility (CSR) practices; in particular, the reaction to negative events was found to be more pronounced than that to positive events (Krüger, 2015). In this study, it is hypothesized that a similar mechanism may extend beyond the price level to reflect in conditional volatility; that is, it is anticipated that investors' skepticism regarding a firm's ESG disclosures may manifest not only in returns but also in an increase in the dispersion of returns (conditional volatility). Second, it is noted that a firm's ESG/CSR profile is associated with its risk level, although findings regarding this relationship remain controversial in the literature (Gillan et al., 2021).

Table 6. Ten GARCH-derived features fed into XGBoost

#	Feature	Computation	Economic Meaning
1	garch_vol_mean	Mean of σ_t	Firm's average market uncertainty
2	garch_vol_std	Std. deviation of σ_t	How variable the uncertainty itself is
3	garch_vol_last	Final σ_T value	Market's most recent level of skepticism
4	garch_vol_max	Maximum σ_t	Most severe market reaction
5	garch_vol_q95	95th percentile of σ_t	Extreme risk measure
6	garch_spike_count	Days with z-score > 2	Count of abnormal volatility events
7	garch_spike_ratio	spike_count / total days	Normalized event frequency
8	garch_persistence	$\alpha + \beta$	How long shocks persist
9	garch_rolling_vol_60d	Mean of 60-day rolling σ	Volatility regime indicator
10	vol_return_corr	Corr (r_t , σ_t)	GARCH model fit quality

Data Leakage Prevention

Only historical returns before the forecast point (the period from January 4, 2023, to December 30, 2024) were used in deriving the GARCH parameters. As part of robustness tests, lagged volatility features were additionally generated using only the first 70% of each firm's return series, thereby confirming the absence of data leakage risk.

Feature Matrix and Model Input

A total of 21 features (six from the NLP channel, five from the financial channel and ten from the GARCH channel) were combined in a single matrix for the XGBoost classifier. The combined structure of the three channels is such that even though statistically the three channels are calculated independently, the final decision is in fact made from a combined evaluation.

Classification Model: XGBoost

XGBoost (Extreme Gradient Boosting) algorithm (Chen & Guestrin, 2016) was selected for greenwashing risk classification. XGBoost is a scalable and regularized version of gradient boosting to reduce the residual errors of each new decision tree compared to the previous one by building decision trees in a linear fashion (Friedman, 2001). The reason behind selecting this algorithm for this work can be summarized as follows: (i) it could be integrated into a model to model different scales and distributions of features such as NLP scores, financial ratios, and volatility metrics, (ii) it can provide an inherent protection against overfitting with L1/L2 regularization, (iii) it can model nonlinear interactions between features in a tree-like manner, and (iv) it can adapt to class imbalance in the form of a scale_pos_weight parameter (Lundberg & Lee, 2017).

Hyperparameter Optimization

The model hyperparameters were optimized using RandomizedSearchCV (Bergstra & Bengio, 2012); this method enables effective exploration in high-dimensional parameter spaces with lower computational cost compared to a full grid search. The search space covered the values {50, 100, 150, 200, 250, 300} for 'n_estimators', {0.01; 0.03; 0.05; 0.1; 0.15; 0.2}, and for 'max_depth': {2, 3, 4, 5, 6, 7}; the optimization was performed on 15 random combinations under 3-fold cross-validation, using the accuracy metric. At the end of this process, the values n_estimators=200, max_depth=4, and learning_rate=0.15 were selected as optimal. On the other hand, to preserve the

stochastic gradient boosting architecture and prevent the model from overfitting, the `subsample` and `colsample\_bytree` parameters were not included in the hyperparameter search space; instead, they were fixed at 0.8 in accordance with common best practices in the literature (Chen & Guestrin, 2016).

Class Imbalance Management

Percentile thresholding results in an imbalanced class distribution in the sample, with a ratio of approximately 3:1. To address this imbalance, XGBoost's `scale\_pos\_weight` parameter is dynamically adjusted based on the ratio of the number of firms in the safe class to the number of firms in the risky class. In this approach, when the model misclassifies a high-risk firm during training, the cost of this error is weighted at a level that is a multiple of the cost of misclassifying a low-risk firm; thus, the influence of the minority class (high-risk firms) on the learning process is increased. This strategy is consistent with the cost-sensitive learning paradigm (He & Garcia, 2009).

Training-Test Split

The dataset was split into 80% training and 20% test using stratified sampling (random_state=42); this split ensures that the class distribution in both subsets preserves the ratio found in the original sample. Consequently, the model was trained on 402 firms; performance evaluation was conducted on the remaining 101 firms, which the model had never seen during the training process. A summary of the model's training configuration is presented in Table 7.

Table 7. Model training configuration

Parameter	Value
Algorithm	XGBoost (gradient boosted trees)
Hyperparameter selection	RandomizedSearchCV (15 iterations, 3-fold CV)
n_estimators	200
max_depth	4
learning_rate	0.15
subsample	0.8
colsample_bytree	0.8
Train/test split	80% / 20% stratified (random_state=42)
Training set	402 firms
Test set	101 firms (never seen by the model)

Model Explainability: SHAP Analysis

To interpret the model's decision-making mechanism, SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) was applied, a game theory-based explainability method. The analysis was conducted on the XGBoost model using TreeExplainer, where the average contribution of each feature to the model is calculated based on the absolute SHAP values. This analysis helps to understand not only how well the model is performing but also to see which channel becomes dominant under which conditions.

Design of Robustness Tests

Three separate tests were designed to assess the contribution and reliability of GARCH integration. All tests were conducted under the same fixed hyperparameters (n_estimators=200, max_depth=4, learning_rate=0.15) and the same 80%/20% split (random_state=42); this ensures that differences between tests stem solely from the feature set.

Test 1 (without GARCH – ablation) involves removing all ten GARCH features and training the model using only the remaining eleven NLP and financial features; this test aims to determine the extent to which the GARCH channel contributes to the model’s overall performance.

Test 2 (GARCH only) is based on completely removing NLP and financial features and training the model using only the ten GARCH features; this design tests whether the market signal alone is sufficient.

Test 3 (lagged GARCH – anti-data-leakage test) involves re-estimating the GARCH(1,1) model for each firm using only the first 70% of the return history (the period from January 2023 to approximately September 2024); Data from the October–December 2024 period were not used in this forecast in any way. This design creates a precise temporal gap (lag) between the forecast point and the feature calculation point, thereby allowing us to test whether the GARCH signal is inflated by forward-looking information (contemporaneous information).

Summary of Data Leakage Precautions

Throughout the study, two key measures were taken to mitigate the risk of data leakage. The first is that the dependent variable is dynamically derived from the sample distribution rather than being calculated directly from the model inputs. The second is that GARCH parameters are calculated using only historical returns before the estimation point, and this design is further validated through a lagged re-estimation test.

Software and Hardware Environment

All analyses were conducted using the Python 3.10+ programming language. The core libraries used and their purposes are listed in Table 8.

Table 8. Libraries used in the analysis.

Library	Version	Purpose
transformers (HuggingFace)	≥ 4.30	ClimateBERT model inference
torch (PyTorch)	≥ 2.0	Deep learning framework
xgboost	≥ 1.7	Gradient boosting classifier
scikit-learn	≥ 1.2	Machine learning pipeline, evaluation metrics, cross-validation
arch	≥ 5.5	GARCH-based volatility modeling
shap	≥ 0.42	Model explainability (SHAP value) analysis
yfinance	≥ 0.2	Financial data acquisition
nltk	≥ 3.8	Text preprocessing and sentence tokenization
textstat	≥ 0.7	Computation of readability indices
pandas	≥ 2.0	Data manipulation and numerical processing
numpy	≥ 1.24	Data manipulation and numerical processing
matplotlib	≥ 3.7	Data visualization
seaborn	≥ 0.12	Data visualization

Findings

Descriptive Statistics

The distribution of the 503 firms comprising the sample across Global Industry Classification Standard (GICS) sectors is presented in Table 9. While the most heavily represented sectors in the sample were Manufacturing (15.7%) and Finance (15.1%), Energy (4.4%) and Communication Services (4.6%) stood out as the sectors with the lowest shares. Although the Energy sector has a low share within the sample, it carries a high sectoral sensitivity weight in the model due to its high carbon intensity. This result indicates that the impact of the small number of companies in this sector on the classification may be disproportionate.

Table 9. Sector distribution of the sample.

Sector	Firm Count	Share
Industrials	79	15.7%
Financials	76	15.1%
Information Technology	73	14.5%
Health Care	59	11.7%
Consumer Discretionary	48	9.5%
Consumer Staples	35	7.0%
Utilities	31	6.2%
Real Estate	31	6.2%
Materials	26	5.2%
Communication Services	23	4.6%
Energy	22	4.4%

The descriptive statistics for the ten features derived from the GARCH (1,1) model are presented in Table 10; this table shows that firm-level conditional volatility (`garch_vol_last`) averaged 1.70, but could reach a maximum of 7.61, indicating significant heterogeneity across firms.

Table 10. GARCH feature descriptive statistics (N = 501 firms).

Feature	Min	Mean	Max
<code>garch_vol_mean</code>	0.82	1.76	6.66
<code>garch_vol_std</code>	0.00	0.24	3.53
<code>garch_vol_last</code>	0.80	1.70	7.61
<code>garch_vol_max</code>	1.02	3.82	76.51
<code>garch_vol_q95</code>	0.94	2.14	11.77
<code>garch_spike_count</code>	0	18.6	40
<code>garch_spike_ratio</code>	0.00	0.04	0.08
<code>garch_persistence</code>	0.00	0.66	1.00
<code>garch_rolling_vol_60d</code>	0.81	1.75	6.54
<code>vol_return_corr</code>	-0.07	0.12	0.42

The analysis reveals that the `garch\_vol\_max` variable contains an outlier of 76.51, which is well above the mean (1.76). This value is believed to stem from the 53% single-day plunge in Globe Life Inc. (GL) stock on April 11, 2024, caused by a corporate scandal. It was also determined that the GARCH model accurately captured this sudden market shock. No data trimming (winsorization) was applied in this study. There are two reasons for this: first, since the XGBoost algorithm used is a tree-based model, it possesses natural resistance to outliers and does not distort the estimates as linear models do (Chen & Guestrin, 2016). Second, such extreme corporate confidence shocks are rare but valuable signals that reflect the very nature of greenwashing and ESG risk.

Model Performance: Held-Out Test Set Results

The model's performance on 101 firms it had never seen during the training process is summarized in Table 11. The sample of 503 firms was divided into an 80% training subset (402 firms) and a 20% test subset (101 firms) using random and stratified sampling; the model was developed solely on the 402 firms in the training set, and its performance was evaluated on the independent test set of 101 firms that were not included in the training process in any way. This distinction ensures that the reported results provide a performance estimate that is free from data leakage and reflects true out-of-sample validity.

Table 11. Held-out test set performance.

Metric	Value
ROC-AUC	0.704
Spearman ρ	0.756
F1 (macro)	0.657
Accuracy	0.703

The final model, which was developed using a lagged GARCH feature set, has an ROC-AUC of 0.704 and Spearman rank correlation of 0.756 on the held-out test set.

This is supported by the recent empirical studies on classification-based financial risk models, where it was demonstrated that AUC values around 0.70 are informative when working with a small number of explanatory variables, for example in classifying carbon emission risk using firm-level ESG ratings. In this case, the AUC values around 0.70 were obtained in the out-of-time robustness test and the result was interpreted as informative for practical screening (Kim & Na, 2026).

However, given the small sample size of this test (101 firms, 31 positive and 70 negative observations), a single misclassification in this sample size would change the AUC value by about 0.01 points. Indeed, the robustness tests presented in the next section result in AUC values of 0.64 to 0.70 for different feature types and hence point estimates should be interpreted with a certain degree of uncertainty.

The Spearman rank correlation of 0.756 indicates that the model ranks firms largely consistently according to their risk level. From the point of view of portfolio management applications, the stability of the relative risk ranking of firms is a more decisive criterion than the accuracy of binary classification and the performance of the model ranking is better than binary classification.

Robustness Test Findings

To evaluate the model's robustness under different feature configurations, three separate tests were conducted, and the results are presented in Table 12 in comparison with the baseline model.

Table 12. Robustness tests: side-by-side comparison.

Metric	Baseline	T1: No GARCH	T2: GARCH Only	T3: Lagged GARCH
ROC-AUC	0.653	0.641 (−1.8%)	0.638 (−2.3%)	0.704 (+7.8%)
Avg Precision	0.589	0.585 (−0.7%)	0.467 (−20.7%)	0.627 (+6.6%)
F1 (macro)	0.604	0.653 (+8.0%)	0.596 (−1.3%)	0.657 (+8.7%)
Precision (macro)	0.604	0.658 (+8.9%)	0.596 (−1.5%)	0.655 (+8.3%)
Recall (macro)	0.604	0.649 (+7.4%)	0.597 (−1.2%)	0.660 (+9.2%)
Accuracy	0.663	0.713 (+7.5%)	0.653 (−1.5%)	0.703 (+6.0%)
Spearman ρ	0.739	0.737 (−0.2%)	0.728 (−1.4%)	0.756 (+2.3%)

The baseline model—a configuration in which GARCH features are used contemporaneously—achieved ROC-AUC and Spearman ρ values of 0.653 and 0.739, respectively. Using these reference values as a baseline, the model's robustness was tested under three different feature configurations: an ablation test in which GARCH features were completely removed (T1), an isolation test in which only GARCH features were used (T2), and a timing test in which GARCH features were used with a one-period lag (T3). As seen in the summary table, these three configurations had differing effects on performance, both in terms of direction and magnitude.

Test 1 – When GARCH Features Are Removed

The purpose of this test is to assess the extent to which the model remains functional without its GARCH-based volatility features. Accordingly, all 10 GARCH features defined in the previous section were removed from the feature set, and the model was retrained using only the remaining 11 natural language processing (NLP) and financial features. This ablation test aims to determine whether the volatility signal makes a necessary contribution to the model's discriminative power or merely plays a supplementary role. The comparative results between the baseline model and the T1 configuration are presented in Table 13.

Table 13. TEST 1: Ablation (volatility features removed)

Metric	Baseline (21 feat)	T1: No Vol (11 feat)	Delta
ROC-AUC	0.653	0.641	−0.012
Spearman ρ	0.739	0.737	−0.002
F1 (macro)	0.604	0.653	+0.049

In this case, with GARCH features completely removed, the ROC-AUC remained only 0.012 points (1.8%) lower while Spearman's ρ remained virtually unchanged. The +0.049 points increase in the F1 metric (macro) as well as the increase in accuracy in Table 13 (7.5%) indicate that the model with fewer features is less likely to show a variance on the smaller test set. These results together suggest that NLP and financial features form a strong basis for discrimination on their own and that the model does not collapse in the absence of the GARCH-based volatility channel. This indicates that the GARCH signal is not a requirement of the model; volatility information is a feature of the model and not a prerequisite. On this basis it was shown that when the GARCH features were

removed, the variables "Leverage" (0.662) and "gunning fog" (0.618) become the most important features in the SHAP importance ranking. This redistribution indicates that in the absence of a market-driven volatility signal, the model has to compensate for this by the firm's financial structure (leverage ratio) and the linguistic complexity of the explanatory texts (Gunning Fog readability index).

Test 2 – Model Using Only GARCH Features

The purpose of this test is to assess the predictive power of the GARCH-based volatility signal on its own. To this end, all NLP and financial features were removed from the feature set, and the model was retrained using only the ten GARCH features. The comparative results between the baseline model and the T2 configuration are presented in Table 14.

Table 14. TEST 2: Volatility-only model

Metric	Baseline (21 feat)	T2: Vol Only (10 feat)	Delta
ROC-AUC	0.653	0.638	-0.015
Avg Precision	0.589	0.467	-0.122
Spearman ρ	0.739	0.728	-0.011

In the case where only GARCH features were used, the ROC-AUC declined slightly (0.638; -2.3%) and the Average Precision dropped 20.7% to 0.467. This indicates that the GARCH volatility signal can reasonably rank firms in terms of risk, but generates a significant number of false positives at the decision threshold. In other words, a lot of firms with high volatility tend to be flagged as high-risk even though they really are not.

This situation suggests that the use of high volatility alone as an environmental risk indicator is an issue since at the firm level one can expect volatility that is not directly related to the environment, such as sectoral growth dynamics, R&D intensity or market liquidity. In this context, NLP-based features such as the gunning fog variable, which measures text complexity, and the ESG gap variable, which captures uncertainty or cautious language, play an important role in distinguishing environmental-based suspicious volatility from normal volatility due to business operations. All in all, we have shown that only using market signals to detect volatility has a downside. Volatility is well served in terms of direction at the ranking level but at the decision level, the precision can't be good enough without NLP.

Test 3 – Lagged GARCH (Data Leakage Test)

This test aims to determine if features based on GARCH really have a structural and persistent signal or they are inflated by the contemporaneous information. For this purpose, we re-estimated the GARCH (1, 1) parameters of each firm by only the first 70% of the return history and thus we found there was a clear time gap between the estimation point and the feature calculation window. The results of our baseline model and T3 configuration are shown in Table 15.

Table 15. TEST 3: Lagged volatility (anti-leakage test)

Metric	Baseline (full data)	T3: Lagged (first 70%)	Delta
ROC-AUC	0.653	0.704	+0.051
Avg Precision	0.589	0.627	+0.038
F1 (macro)	0.604	0.657	+0.053
Accuracy	0.663	0.703	+0.040
Spearman ρ	0.739	0.756	+0.017

The ROC-AUC of 0.653 on the held-out test set is considered information-carrying for classification models with few explanatory variables as discussed in the previous section. It should be noted that the small test set ($N=101$) leads to higher metric variance. The model's true strength lies in its ranking quality (Spearman $\rho = 0.739$, $p < 10^{-18}$). This is the most actionable metric for portfolio screening. In this scenario, where GARCH parameters were re-estimated using only the first 70% of the return history, the model outperformed the baseline model trained on the full dataset across all metrics examined (ROC-AUC: 0.704 vs. 0.653; Spearman ρ : 0.756 vs. 0.739). This result, which may seem unexpected at first glance, can be interpreted as follows: when GARCH (1,1) parameters are estimated over a shorter and more stable window, out-of-sample generalization performance improves; conversely, the full sample window may also include short-term noise specific to the most recent period. The fact that performance improved rather than declined can also be interpreted as a strong indication that the GARCH signal is not inflated by look-ahead information. If there had been a timing inconsistency (data leakage) in the feature calculation, performance would have been expected to decline as the lagged model lost this advantage. However, this finding should be interpreted not as definitive proof on its own, but as a positive sign regarding the model's timing consistency. In this test the SHAP importance ranking has also changed, the `garch_rolling_vol_60d` variable is now in first place, well ahead of `garch_vol_last`. This effect can be explained by the fact that, in this case when most of the information on the most recent day's volatility is from months ago, the 60-day moving average is more representative of the firm's structural volatility regime.

SHAP Analysis Findings

SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) is an explainability method that measures the contribution of each feature to the model's output using a game theory-based approach. In this study, it was applied to the XGBoost model using the TreeExplainer algorithm. The SHAP feature importance rankings for the full model are presented in Table 16, a summary comparison by channel is shown in Table 17, and a comparison among the four different model variants is presented in Table 18. The visual distribution of the rankings is also shown in Figure 2.

Table 16. SHAP features importance ranking (full model).

Rank	Feature	SHAP(Mean)	Channel	Interpretation
1	<code>garch_vol_last</code>	0.668	GARCH	Market's current level of skepticism
2	<code>garch_vol_mean</code>	0.452	GARCH	Average uncertainty
3	<code>garch_vol_max</code>	0.448	GARCH	Most severe market reaction
4	Leverage	0.403	Financial	Debt-driven greenwashing incentive
5	<code>gunning_fog</code>	0.358	NLP	Text complexity / obfuscation
6	<code>garch_persistence</code>	0.340	GARCH	Shock persistence

7	vol_return_corr	0.308	GARCH	Model fit quality
8	garch_vol_q95	0.306	GARCH	Tail risk measure
9	ESG_Gap	0.300	NLP	Hedge word density
10	ROA	0.295	Financial	Profitability
11	garch_vol_std	0.293	GARCH	Uncertainty variability
12	sector_sensitivity	0.273	Financial	Sector ESG exposure
13	garch_spike_count	0.270	GARCH	Event frequency
14	NLP_GWL_Score	0.253	NLP	ClimateBERT score
15	proxy_gwl_score	0.251	NLP	GWL proxy
16	Ln_Firm_Size	0.229	Financial	Firm size
17	garch_rolling_vol_60d	0.216	GARCH	Regime indicator
18	flesch_reading_ease	0.200	NLP	Readability
19	finbert_esg_sentiment	0.135	NLP	ESG sentiment
20	garch_spike_ratio	0.090	GARCH	Normalized event rate
21	hedge_words_count	0.011	NLP	Raw hedge count

The fact that three of the top four features in the SHAP analysis originate from the GARCH channel empirically validates the “market skepticism” hypothesis at the core of this study. However, this situation also carries the risk of single-channel dependency in the model. It is particularly noteworthy that the Leverage and ROA variables—which represent the financial pressure hypothesis cited in the literature as the primary motivation for greenwashing—carry relatively limited weight in the SHAP ranking. This finding may support the thesis that large-scale, institutionalized firms, such as those in the S&P 500, use greenwashing not as a financial last resort but rather as a strategic positioning tool.

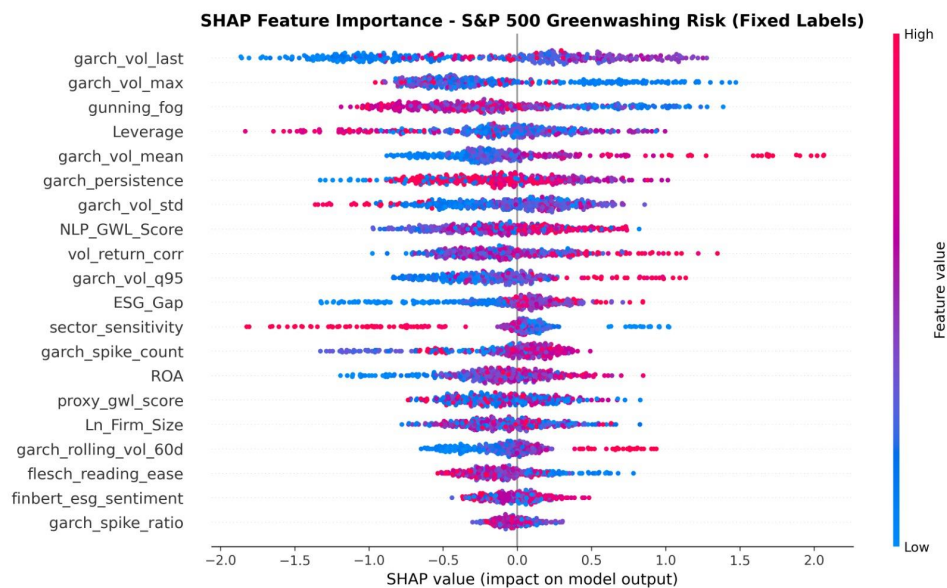


Figure 2. Visual distribution of the SHAP feature importance rankings**Table 17.** SHAP category summary

Channel	Features in Top 5	Features in Top 13	Overall SHAP Contribution
GARCH (Market)	3 / 5	8 / 13	Dominant
NLP (Text)	1 / 5	2 / 13	Complementary
Financial	1 / 5	3 / 13	Complementary

In the full model, three of the top five features (`garch_vol_last`, `garch_vol_mean`, `garch_vol_max`) come from the GARCH channel; the remaining two spots are shared by Leverage (financial channel) and `gunning_fog` (NLP channel). When expanded to the first thirteen features, this distribution becomes even more pronounced in favor of GARCH; eight come from the GARCH channel, three from the financial channel, and two from the NLP channel.

This pattern suggests that the market-based volatility signal is driving the overall explanation, while financial and textual channels play a complementary role with a limited but consistent contribution. This SHAP advantage may lead to the suspicion that volatility could be an endogenous determinant; however, we can alleviate this by comparing the results from the three robustness tests: (i) the 20.7% decline of the Average Precision for the model based on only GARCH features suggests that the volatility signal alone is not enough; (ii) the model remains functional even when GARCH features are completely removed, which is consistent with the idea that NLP and financial channels have independent explanatory power; (iii) the fact that the lagged (70%) volatility features outperform the full model suggests the relationship is not simply temporary and cyclical but rather very structural. Taken together, these three results show that the GARCH weight of SHAP is not just an artificial artifact of the model's data processing order or contemporaneity (contemporaneous relationship), but none of these three tests alone are evidence of causality. This is consistent with the framework of the theory of Krüger (2015) of “ESG concerns → volatility”; since the purpose of this study was not to prove a causal relationship, the overlaps should not be viewed as causality but should be seen as an indicator consistent with theoretical expectations.

Table 18. Cross-model SHAP comparison

Model	#1 Feature	GARCH in Top 5
Full Model	<code>garch_vol_last</code> (0.668)	3 / 5
T1: No GARCH	Leverage (0.662)	0 / 5
T2: GARCH Only	<code>garch_vol_last</code> (0.798)	5 / 5
T3: Lagged GARCH	<code>garch_rolling_vol_60d</code> (0.579)	4 / 5

We compare the four model variants and find that the model can be adapted to different channels without one channel. As far as GARCH features are removed (T1), we find that Leverage (0.662) is the most important feature, which means that even when no market signal is available, the model is able to adapt to the financial structure and text complexity. In T3 where we use GARCH features with a lag, we find that `garch_rolling_vol_60d` is the most important feature; this supports the view of the previous section that the 60-day regime average is a better signal than the instantaneous volatility measure. This flexibility can be understood as indirect proof that the proposed architecture does not depend on a single information channel.

Main Thesis: The Role of Market Skepticism in Detecting Greenwashing

The findings from the SHAP analysis reveal a pattern consistent with the study's underlying interpretive framework: the level of trust investors place in companies' ESG disclosures can translate into an indirectly observable signal via stock volatility. According to this framework, when a firm publicly discloses its environmental commitments but market participants do not find this disclosure convincing, this skepticism is expected to manifest in pricing behavior in three ways: an increase in the conditional variance of stock returns (σ^2_t), the persistence of this increase over time, and an increase in the frequency of periodic volatility shocks. This triplet of patterns—high volatility, high persistence, and frequent shocks—aligns with the variables `garch_vol_last`, `garch_persistence`, and `garch_spike_count`, which emerge as the strongest explanatory signals in the SHAP analysis.

This interpretive framework is not a causality test, but rather a proposal for a theoretical explanation consistent with the observed SHAP pattern; a direct causal test to determine whether the increase in volatility stems directly from investor skepticism falls outside the scope of this study. However, according to the results obtained in Test 2, the volatility signal alone is not sufficient; since high volatility can stem from many alternative causes (sectoral dynamics, firm size, liquidity conditions), it produces a significant number of false positives when used on its own. Therefore, NLP-based features such as text complexity (`gunning_fog`) and the density of cautious/ambiguous language (`ESG_Gap`) serve a complementary and necessary function in distinguishing volatility of environmental origin from normal volatility arising from business operations.

Discussion

The results of this study can be summarized in three sections. First, we have developed a model that classifies 503 S&P 500 firms according to greenwashing risk based on 21 features from NLP, financial and GARCH channels and has a Spearman rank correlation of 0.756 and ROC-AUC of 0.704 on an independent test set of 101 of them not picked up in training. A level of this performance is considered to be information-carrying for financial risk classification models that work with very few explanatory variables. In particular, the high-ranking consistency of the model indicates that the model can be used for portfolio screening.

Second, a SHAP analysis showed that the volatility characteristics in GARCH (i.e., `garch_vol_last` (SHAP = 0.668)) are the strongest explanatory signal from the model. This is consistent with the hypothesis that investors' lack of confidence in a firm's environmental disclosures may be reflected in the conditional variance of stock returns, even if it is not directly observable. However, robustness tests showed that this is not enough on its own: one model based on only GARCH (Test 2) saw a 20.7% decrease in Average Precision; that is, market volatility is a mix of factors that aren't related to the environmental data (e.g., sectors, liquidity, etc.) and when taken into account in isolation, leads to a large number of false positives. Thus, the NLP channel (`gunning_fog`, `ESG_Gap`) is a complementary channel to distinguish suspicious volatility from normal volatility due to business activity.

Third, the model's performance was only slightly lower (ROC-AUC: -0.012 points) in the ablation test (Test 1), where GARCH features were removed completely, showing that the remaining NLP and financial features have reasonable discriminatory power on their own. When we combine these three findings, the central thesis of the study is fulfilled. Greenwashing risk is not detected by a single approach based on the company's discourse (NLP), its financial structure (Finance) or its market response (GARCH) but the combination of these three information channels is necessary.

In order to be considered reliable from both academic and practical perspectives, the model should be evaluated in terms of data integrity, explainability and robustness, as well as its predictive performance. In this study we have tested the model in six dimensions.

The multi-source structure of our model was achieved by combining 21 features from the NLP, financial and GARCH channels into a single matrix. We have also computed SHAP values from TreeExplainer, and find that `garch_vol_last`, `garch_vol_mean`, and `garch_vol_max` are the top three features of the model with the most contribution, while `leverage` and `gunning_fog` are the financial and textual channels, respectively, among the top five features.

We have taken two measures to reduce the risk of data leakage: the dependent variable was generated based on a dynamic threshold based on the 75th percentile of the sample distribution, rather than on the model input and the GARCH parameters were derived using only the historical returns prior to the forecast point. We check this assumption further with our delayed re-estimation test (Test 3) where we only used the first 70% of the return history. The fact that the model's performance improved rather than declined in this test (ROC-AUC: 0.653 vs. 0.704) represents a positive signal that the signal was not inflated by forward-looking information.

We tested the model's robustness through three tests (ablation, isolation and delayed re-estimation) with the same fixed hyperparameters that confirmed the model's performance is not determined by one channel. The model's overfitting risk was mitigated by the model's performance in a 101-firm held-out test set that it had never seen before. Finally, the market sensitivity of our model is explained by the dominance of GARCH (1,1)-based features in the ranking of SHAP on top of the list (three of the top five features and eight of the top thirteen features).

Conclusion

We have developed an XGBoost model that classifies 503 companies within the S&P 500 based on their greenwashing risk with 21 features from three different information channels: natural language processing, financial indicators, and GARCH (1,1)-based market volatility. The model achieved an ROC-AUC of 0.704 and a Spearman rank correlation of 0.756 on an independent set of 101 previously unseen companies. This is supported by three independent robustness tests and anti-data-leakage validation with lagged GARCH features.

We find that the market-based volatility signal is the strongest explanatory component of the model, but this signal alone is not enough; NLP and financial channels are complementary to the model and increase sensitivity at the decision threshold. This is in line with the fact that instead of one data source in the company for greenwashing detection, multi-channel architectures that analyze a company's discourse, structure, and market response to the discourse can be more reliable. This study makes two significant contributions to the literature on greenwashing detection. First, to the best of our knowledge, it represents the first study to integrate time-series volatility dynamics into an ESG credibility classifier in the form of firm-level cross-sectional characteristics; the theoretical basis for this integration aligns with propositions suggesting that heterogeneous investor beliefs regarding ESG authenticity are reflected in the market in the form of conditional variance (Krüger, 2015; Pástor et al., 2021), and this integration has also been empirically validated through the ablation tests applied in the study. Second, these tests demonstrate that none of the three channels (NLP, financial, market) is sufficient on its own, but together they produce a synergistic result; in particular, the finding that volatility is decisive in terms of ranking performance (AUC), while NLP is decisive in terms of precision, offers a unique perspective on multi-source ESG analysis.

Beyond their theoretical contribution, these findings also offer concrete guidance for both policymakers and firms. For policymakers, the dominance of the market volatility signal—combined with its insufficiency on its own—suggests that greenwashing oversight should combine market-based risk indicators with disclosure-text review rather than relying on textual analysis alone; the documented divergence across rating providers (Berg et al., 2022) further supports mandating multi-provider ESG verification for firms flagged as high-risk. For firms, the results indicate that vague or overly positive disclosure language becomes especially visible—and costly to investor trust—during periods of elevated market volatility; firms are therefore advised to align sustainability claims closely

with verifiable financial and operational performance, particularly when under financial stress, to avoid being flagged by increasingly multi-channel greenwashing detection tools such as the one developed in this study.

Limitations and Future Research

While this study contributes to the empirical literature, it has some methodological limitations: First, the ISS Governance QualityScore data, which forms the basis of the dependent variable, should be evaluated in light of the phenomenon of ESG rating divergence documented by Berg, Kölbel, and Rigobon (2022). Different rating agencies (ISS, MSCI, Refinitiv) can systematically produce different risk scores for the same firm; because the study's dependent variable relies on a single data provider (ISS), it carries this risk of measurement error. In this context, the ISS Governance QualityScore is not a direct measure of greenwashing but rather an indirect indicator (proxy); since the concept of greenwashing, by definition, arises at the intersection of high corporate risk and low transparency (obfuscation) (Lyon & Maxwell, 2011), this study treats the governance risk score as only one dimension of this intersection and complements it with NLP-based textual transparency metrics (Fog Index, Hedge Words). Future studies using verified greenwashing case databases (SEC investigations, climate lawsuits) will strengthen this conceptual validity.

Second, the NLP channel uses company disclosure texts accessed via yfinance as a proxy instead of actual sustainability reports (CSR, CDP, GRI, TCFD). While this choice limits the information content of text-based signals, it does not affect the validity of the findings in terms of the methodological framework (GARCH integration, robustness verification); future studies conducted using actual CSR reports are expected to enhance the contribution of the NLP channel.

Third, the values ranging from 0 to 1 used to determine GICS sector sensitivity weights (sector_sensitivity), while based on concrete metrics such as carbon footprint and regulatory exposure, are partly heuristic in nature.

Fourth, the study employs a cross-sectional design and is based on a single time period (2023–2024). The use of a panel data structure and rolling GARCH windows will allow for an assessment of the model's performance under different market regimes. In this regard, the fact that the sample is limited to S&P 500 firms indicates that the findings are valid in the context of large-scale, institutionalized publicly traded firms; generalizing these findings to smaller-scale, non-publicly traded firms or those subject to different regulatory frameworks (e.g., emerging markets) should be approached with caution. Replication studies conducted on these firm groups will test the model's generalizability. Since market volatility and governance risk are measured in the same cross-section, endogeneity between variables (garch_vol_last) cannot be fully ruled out. Although the applied lagged variable tests prevent forward-looking data leakage, future studies should more precisely isolate this feedback loop using Granger causality or Instrumental Variable (IV) models.

Finally, this study is a predictive modeling study and does not claim to provide causal inference. The observed relationship between GARCH volatility and ESG risk should be interpreted as correlational; the causal nature of this relationship should be further tested in future studies based on instrumental variables or natural experiment designs. Furthermore, only the ISS Governance QualityScore was used as the target variable. In accordance with the “rating divergence” phenomenon recognized in the literature (Berg et al., 2022), it is recommended that future studies test the model's sensitivity through cross-provider validation using alternative providers such as Sustainalytics, MSCI, or Refinitiv. Another point to note is that, for the sake of scalability, company summary texts were used as proxies. In future studies, conducting a pilot application using full-text analysis of actual CSR, CDP, or TCFD reports will further clarify the model's mechanism.

www.turkishstudies.net/economy**Turkish Studies - Economics, Finance, Politics**

eISSN: 2667-5625

**ANKARA SCIENCE
UNIVERSITY**

Makine Öğrenmesi Tabanlı Yeşil Aklama Tespitinde Finansal Stres ve Piyasa Oynaklığı: S&P 500 Şirketlerinden Bulgular

Giriş

Dünyada yaşanan teknolojik gelişmelerin üretim sektöründeki yansımaları sonucunda doğaya verilen zarar kaynaklı küresel bilinç yıllar içinde artış göstermiştir. Firmalar, ekonomik rapor hazırlamak dışında, sosyal ve çevresel raporları içeren sürdürülebilirlik raporu hazırlamaya başlamıştır. Geline nokta paydaşlar firmaların sadece finansal verilerini değil sürdürülebilirlik faaliyetlerini de bu raporlar aracılığıyla takip etmek istemektedir. Yapılan araştırmalar, sürdürülebilirliğe önem veren firmaların yatırımcılar tarafından değerli görüldüğünü göstermektedir (Dağıstanlı & Dağıstanlı, 2023). Son yıllarda sürdürülebilirlik ve sürdürülebilirlik raporlaması kurumsal uygulamaların merkezine yerleşmiştir. Bununla birlikte çevresel ve sürdürülebilirlik konularına ilişkin toplumsal farkındalık da artmıştır. Bu durum işletmeler üzerinde çevresel, sosyal ve yönetim (Environmental, Social and Governance- ESG) unsurlarını iş modellerine entegre etme ihtiyacı ve bu alanlardaki performanslarını kamuya açıklama yönünde artan bir baskı oluşturmıştır. Bunlarla birlikte yatırımcılar ve hükümetler de dahil olmak üzere paydaşlardan gelen baskılar, büyük şirketlerin kurumsal sorumluluk ve sürdürülebilirlik raporlamasını geliştirmesine yol açtı. İşletmelerin sürdürülebilirlik hedeflerine ne ölçüde yaklaştıklarını ve bu hedeflerin stratejilerini nasıl etkilediğini gösteren sürdürülebilirlik raporları, tüm işletme paydaşları için faydalı bilgiler içeren önemli kaynaklardan biridir. Geliştirilen standartlar, sürdürülebilirlik açıklamalarının tutarlılığını ve karşılaştırılabilirliğini sağladığı gibi, bilgi kullanıcılarının yatırım kararlarını daha bilinçli ve stratejik olarak almalarına da olanak sağlamaktadır. Bu raporlarda yer alan açıklamalar, yatırımcılar arasında var olması muhtemel bilgi asimetrisinin azaltılmasına yardımcı olmaktadır. Ancak ESG açıklamalarının büyük ölçüde gönüllülük esasına dayanması, düzenleyici boşluklar, yatırımcı ve müşteri beklentileri ile rekabet baskısı gibi faktörler, bazı işletmelerin fiili performanslarının ötesinde yanıltıcı sürdürülebilirlik beyanlarında bulunmasına yol açabilmektedir (Karagül & Selimoğlu, 2026; Suta et al., 2026; Liman, 2026).

Günümüz koşullarında işletmelerin yalnızca finansal raporlarını sunmaları, çevresel ve sosyal alanlarda gerçekleştirdikleri faaliyetlerini ortaya koyamamasından dolayı yetersiz kalmaktadır. İşletmelerin ekolojik etki yarattığı kurumsal faaliyetler tam olarak değerlendirilememekte, karşılaştırılabilir ve kullanışlı bilgi üretimi de sağlanamamaktadır. Finansal raporlama, çevresel, sosyal ve kurumsal etkilerin anlaşılmasını mümkün kılmadığı gerekçesiyle eleştirilere maruz kalmıştır. Bu nedenle, söz konusu etkilerin anlaşılmasını ve bu etkilerin işletmelerin ekonomik performansıyla dengelenmesini sağlamak amacıyla, işletmelerin ilave açıklamalar yapması gerektiği gündeme gelmiştir (Schaltegger & Burritt, 2010). Bu gelişmeler doğrultusunda çevresel, sosyal ve yönetim (ÇSY) raporlaması, işletmelerin finansal olmayan performanslarını paydaşlara şeffaf biçimde aktarmalarını amaçlayan tamamlayıcı bir raporlama aracı hâline gelmiştir. Ancak ÇSY raporlamasının yaygınlaşması, raporlanan bilgilerin doğruluğu ve

güvenilirliği konusunda yeni tartışmaları da beraberinde getirmiştir. Bu tartışmaların merkezinde yeşil aklama (greenwashing) sorunu bulunmaktadır (Çelik & Reis, 2026). Disiplinlerarası niteliğinden dolayı, evrensel düzeyde kabul görmüş ortak bir tanımı bulunmamasına rağmen yeşil aklama genel anlamda; bir ürün veya hizmeti gerçekte çevre dostu, yeşil ya da sürdürülebilir olmadığı halde çevre dostu, yeşil ya da sürdürülebilir olarak gösteren yanıltıcı bir pazarlama stratejisi olarak tanımlanmaktadır (Özkaya, 2026). Başka bir tanımla yeşil aklama, işletmelerin çevresel duyarlılığa ilişkin beyanları ile bu beyanları destekleyen gerçek uygulamalar arasındaki uyumsuzluk durumu ve bir firmanın çevresel performansı hakkında paydaşlarını yanıltma pratiğidir (Shakourloo & Azimli, 2025; Lyon & Maxwell, 2011).

Kurumsal bağlamda yeşil aklama ise, işletmelerin çevre dostu bir imaj inşa ederek sermaye temin etme, tüketici nezdinde olumlu algı oluşturma ve düzenleyici baskıları sınırlama amacıyla başvurdukları etik dışı bir strateji olarak değerlendirilmektedir. Bu tür uygulamalar yalnızca çevresel ve toplumsal etkiler bakımından değil, aynı zamanda finansal piyasalarda güvenin tesis edilmesi açısından da önemli riskler barındırmaktadır. Zira ESG odaklı yatırımların yaygınlaşması, yatırımcıların çevresel performansı yüksek işletmelere yönelmesini teşvik ederken, eş zamanlı olarak bilgi asimetrisi sorununu da derinleştirmektedir (Shakourloo & Azimli, 2025). Ayrıca yeşil aklamanın yatırımcılar üzerinde oluşturduğu güven kaybı ile çevreci davranışlar gösteren firmalara önyargıda bulunma gibi etkileri bulunmaktadır (Tepeli & Büklü, 2025).

Yeşil aklamanın (greenwashing) gerçekleşmesinde veya gerçekleşmemesinde paydaşların hayati bir öneme sahip olduğunu vurgulayan Seele and Gatti (2017), paydaşların yeşil aklamaya ilişkin algıları ve tepkilerinin, bir şirketin stratejilerini önemli ölçüde etkileyebileceğini de dile getirmiştir. Bununla birlikte, söz konusu algı ve tepkilerin etkisi stratejik düzeyde sınırlandırılmayıp; paydaşlarda oluşan endişeler şirketlerin piyasa performansına da yansımabilmektedir. Nitekim paydaşlar açısından endişe yaratan durumlar, pay fiyatlarında ani değişimlere neden olabilmektedir (Chen and Dagestani, 2023). Özellikle firmalar üzerindeki finansal baskılar arttığında firmaların söylemleri ile risk yönetim dili değişebilmekte ve daha olumlu söylemlerde bulunarak gerçeğe aykırı davranışlar ortaya çıkabilmektedir. Ortaya koyulan çalışmalar, yeşil aklamanın asılsız iddialardan, halka açık firmaların gerçek çevresel etkilerini değerlendirme çabalarını sekteye uğratan bilinçli yapılan yanlış beyanlar ile farklı türden uygulamalarla meydana gelebileceğini göstermiştir (Atabay, 2026). Bunların yanı sıra yeşil aklama yapan firmalar genellikle karmaşık, belirsiz ifadeler kullanarak gerçek çevresel performanslarını gizlemeye çalışır.

İlk kez 1986 yılında ekolojist Jay Westerveld tarafından ortaya atılan yeşil aklama kavramı işletme, çevre yönetimi ve hukuk dahil olmak üzere çeşitli disiplinlerden akademisyenler tarafından kapsamlı bir şekilde incelenmiş ve tartışılmıştır. Bu kapsamdaki önemli çalışmalardan birisini gerçekleştiren (Karagül & Selimoğlu, 2026), yeşil aklama uygulamalarının işletmelerde yatırımcı güveni, piyasa değer kaybı, itibar kaybı, bilinçli yapılan aldatıcı eylemler, dolandırıcılıklar ile bunlar sonucunda meydana gelen yasal ihlaller gibi çeşitli finansal problemlere neden olabileceğini belirtmiştir. Buna karşın (Li et al., 2023) işletmelerde çevresel baskı ve ortakların talepleri arasındaki uyumsuzluğun ortadan kaldırılabilmesinde yeşil aklamanın etkin bir araç olabileceğini öne sürmüştür. Greenwashing üzerine yapılan çalışmaların önemli bir kısmı vaka temelli finansal analiz yöntemi (Laufer, 2003; Walker & Wan, 2012; Marquis et al., 2016; Tepeli & Büklü, 2025), içerik analizi (Özdemir et al., 2017; Yu et al., 2020;), söylem analizi (Dorfleitner & Utz, 2024; Thewissen, 2026), karşılaştırmalı analiz (Kathan et al., 2025; Cocozza, 2026) metin madenciliği (Kılınç et al., 2025;), ESG derecelendirme kuruluşlarının skorları (Atabay, 2026; Kathan et al., 2025, Siddique et al., 2026), ve çoklu örnek olay incelemesine (Karagül & Selimoğlu, 2026) dayanmaktadır. Son yıllarda doğal dil işleme ve makine öğrenmesi kullanılmaya başlanmıştır. Ancak mevcut çalışmaların büyük çoğunluğu yalnızca metinsel özellikleri kullanmaktadır. Finansal baskı ve piyasa volatilitesinin kurumsal söylem üzerindeki etkisini aynı model içerisinde değerlendiren çalışmalar oldukça sınırlıdır.

Makine Öğrenmesi bilgisayar bilimlerinin bir alt dalı olup 1959 yılında Arthur Samuel tarafından “bilgisayarlara açıkça programlanmadan öğrenme alışkanlığı kazandırma” şeklinde tanımlanmaktadır. Yapay zekâ çerçevesinde hesaplamalı öğrenme teorisi ve desen tanıma çalışmalarıyla evrimleşen makine öğrenmesi, tahmin yapabilme, verilerle öğrenme, algoritmaların geliştirilmesi ve incelenmesini esas alır. Bu öğrenme teknikleri, karmaşık verilerden anlamlı bilgi çıkarılmasına imkân tanıyarak yapay zekâ sistemlerinin çok boyutlu problem çözme süreçlerinde önemli bir rol oynamaktadır (Meral, 2025).

Finansal açıdan ele alındığında makine öğrenmesi, değişen piyasa koşulları ve bunların ölçülmesinde, risklerin tahmin edilmesinde, riske sebebiyet verebilecek durumların saptanmasında, finansal araçların arzında talebin ne ölçüde olacağını ve arz talep dengesinin sağlanmasında kolaylıklar sağlamaktadır (Özgür, 2021). Nitekim büyük dil modelleri, doğal dil işleme ve makine öğrenmesi teknikleri kurumsal raporların analizini de otomatik hale getirmektedir. Bu kapsamda metin madenciliği ve doğal dil işleme (NLP) teknolojileri, büyük hacimli ve yapılandırılmamış metinsel verilerden anlamlı içgörüler elde edilmesini sağlayarak, kuruluşların kurumsal rapor analiz süreçlerini kökten değiştirmektedir. Söz konusu teknolojiler, finansal tabloların, sürdürülebilirlik bildirimlerinin ve risk raporlarının otomatik analizinde, karar alma mekanizmalarının desteklenmesinde, uyumluluk denetimlerinde ve şeffaflığın artırılmasında günden güne daha yaygın bir alanda kullanılmaktadır (Başar & Doğruel, 2025). Bu yöntemler binlerce raporu aynı anda analiz edebilmekte, insan önyargısını azaltmakta ve karar destek sistemi olarak kullanılabilirliği artırmaktadır.

Bu çalışma, S&P 500 endeksine dahil firmaların çevresel taahhütlerinin güvenilirliğini değerlendiren bir makine öğrenimi sistemi geliştirmek amacıyla tasarlanmıştır. Sistem, bu riski yalnızca metinsel analize dayanarak değil, finansal yapı ve piyasa algısını da dikkate alarak tespit etmektedir. Mevcut çalışmalardan farklı olarak model yalnızca kurumsal metinleri değil, aynı zamanda finansal baskı göstergelerini ve piyasa volatilitelerini de açıklayıcı değişkenler olarak kullanmaktadır. Böylece kurumsal söylem ile finansal koşullar arasındaki ilişkinin yeşil aklama davranışı üzerindeki etkisi birlikte değerlendirilmektedir. Çalışmanın hem literatüre hem de yatırımcılar, düzenleyici kurumlar ve denetim otoriteleri açısından uygulamaya katkı sağlaması beklenmektedir.

Materyal ve Yöntem

Örneklem ve Veri Kaynakları

Araştırmanın örneklemi, Wikipedia'nın “List of S&P 500 companies” sayfasından otomatik olarak çekilen 503 firma oluşturmaktadır. Firmalar, Global Industry Classification Standard (GICS) sektör taksonomisine göre sınıflandırılmış; finansal veriler ise Yahoo Finance (yfinance) API'si aracılığıyla toplanmış olup 4 Ocak 2023 ile 30 Aralık 2024 tarihleri arasındaki dönemi (yaklaşık 500 işlem günü) kapsamaktadır. Bu dönemin seçilme amacı 2020-2022 dönemini pandemi kaynaklı makro-ekonomik gürültüden arındırılmış bir piyasa rejiminde çalışmaktır. Ayrıca, enerji piyasalarındaki jeopolitik şokların piyasa volatilitelerini domine ettiği ve global ESG raporlama standartlarının (ör. CSRD, SEC taslakları) yasal bir yaptırım gücüne kavuştuğu bu dönem ve 'yeşil aklama' pratiğinin salt bir PR sorunundan çıkıp piyasa volatilitesiyle (GARCH) doğrudan fiyatlanan finansal bir riske dönüştüğü zaman aralığıdır. GARCH modellerinin yapısal piyasa kırılmalarından kaçınmak ve stabilize sağlamak için ılımlı tahmin pencerelerinde optimal performans gösterdiği tespit edilmiştir (Starica & Granger, 2005; Hwang & Valls Pereira, 2006).

Toplanan veriler üç kategoriye ayrılmaktadır. İlk olarak, firmaların Yahoo Finance platformundaki iş açıklamaları (longBusinessSummary alanı) ile sektör bazlı sürdürülebilirlik proxy metinleri, metinsel analiz için girdi kümesini oluşturmuştur. İkinci olarak, piyasa değeri (Market Cap), firma büyüklüğü (Ln_Firm_Size), aktif kârlılığı (ROA), kaldıraç oranı (Leverage), piyasa-defter değeri oranı (Market-to-Book), temettü verimi (Dividend Yield) gibi finansal göstergeler derlenmiştir. Üçüncü olarak, bağımlı değişken, Yahoo Finance aracılığıyla elde edilen ISS

(Institutional Shareholder Services) Governance QualityScore metrikleri kullanılarak oluşturulmuştur. ISS skorlaması, firmaları 1 (en düşük risk) ile 10 (en yüksek risk) arasında göreceli bir desil (decile) ölçeğinde değerlendirmektedir (ISS, 2023). Modelin hedef değişkeni (y), bu 1-10 ölçeğindeki ham skorların dinamik kuantil eşiklemesi ile ikili sınıfa (binary) dönüştürülmesiyle elde edilmiştir.

Tablo 1. Veri Kaynakları

Kaynak	İçerik	Kapsam
Wikipedia S&P 500 tablosu	Ticker listesi + sektör	503 firma
yfinance API	Finansal oranlar, ESG risk skorları, hisse fiyatları	503 firma
yfinance şirket açıklamaları	Proxy ESG bildirim metni (NLP girdi)	503 firma
Günlük log-getiriler	GARCH modellemesi için	249.610 satır, 501 firma

Eksik Veri Yönetimi

Veri setindeki eksik gözlemlerin tamamlanmasında GICS sektör medyanına dayalı imputation tekniği uygulanmıştır. Bu yaklaşımda bir firmanın eksik finansal göstergesi, öncelikle aynı GICS sektöründeki firmaların medyan değeriyle; sektör medyanının da eksik olduğu durumlarda genel örneklem medyanıyla doldurulmuştur. Bu strateji, Little ve Rubin'in (2019) eksik veri teorisinde önerdiği sektörel homojenlik varsayımıyla uyumlu olup, tekil aykırı gözlemlerin doldurma sürecini bozma riskini sınırlamaktadır.

Bağımlı Değişkenin Tanımlanması

Bağımlı değişken, Yahoo Finance aracılığıyla elde edilen ISS (Institutional Shareholder Services) Governance QualityScore ham verisinin (1–10 desil ölçeği) dinamik kuantil eşiklemesi yoluyla ikili bir sınıf değişkenine dönüştürülmesiyle elde edilmiştir. Örneklemdeki 503 firma, ISS yönetim risk puanlarına göre küçükten büyüğe sıralanmış ve dağılımın 75. yüzdeline karşılık gelen eşik değeri hesaplanmıştır ($esg_risk_raw = 8.0$). Risk puanı bu eşik değerine eşit veya bu değerin üzerinde olan firmalar "Yüksek Risk" ($y = 1$), eşik altında kalan firmalar ise "Düşük Risk" ($y = 0$) olarak sınıflandırılmıştır. Bu tercihin gerekçesi, Kaufman ve arkadaşlarının (2012) tanımladığı veri sızıntısı (data leakage) riskinden kaçınmaktır. Erken aşama denemelerde, hedef değişkenin doğrudan model girdilerinden türetilmesi durumunda gözlenen yapay olarak yüksek doğruluk oranı (yaklaşık %98), bu riskin somut bir göstergesi olarak değerlendirilmiştir. Uygulanan bu eşikleme sonucunda örneklemdeki 349 firma (%69.4) "Düşük Risk" ($y = 0$), 154 firma (%30.6) ise "Yüksek Risk" ($y = 1$) sınıfında yer almıştır. Ham skorda 8.0 değerinde gözlenen gözlem yığılması (tie) nedeniyle, nihai sınıf dağılımı teorik %75/%25 oranından bir miktar sapmıştır.

GICS Sektör Hassasiyet Haritası

Firmalar, GICS sektör sınıflandırmasına göre 0 ile 1 arasında değişen bir çevresel hassasiyet ağırlığına sahiptir. Bu ağırlıklar, sektörlerin karbon ayak izi, çevresel düzenleyici maruziyeti ve tarihsel ESG tartışma sıklığı temel alınarak belirlenmiştir.

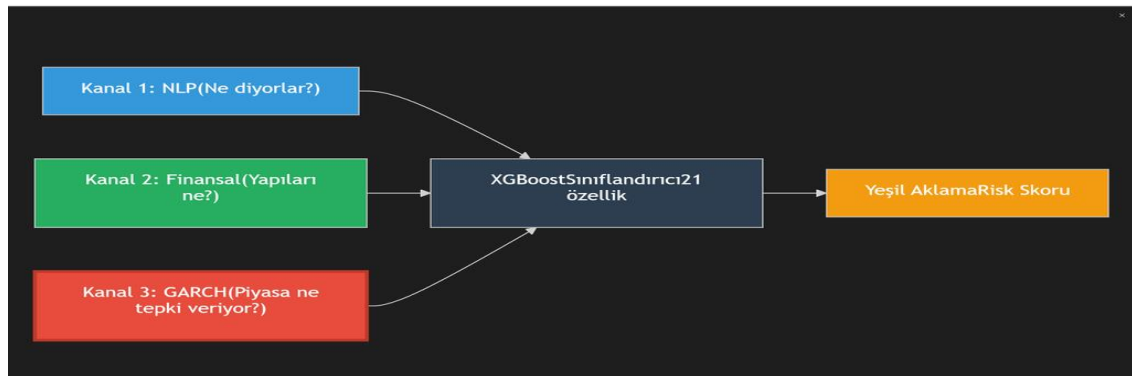
Tablo 2. GICS Sektör Hassasiyet Haritası

Sektör	Hassasiyet Ağırlığı
Enerji	0.95
Malzemeler	0.80
Kamu Hizmetleri	0.75
Sanayi	0.65
Temel Tüketim / Zorunlu Tüketim Malları	0.50
Sağlık Hizmetleri	0.45
Finansal Hizmetler	0.40
Gayrimenkul	0.35
Bilgi Teknolojileri	0.25
İletişim Hizmetleri	0.20

Tablo 2'de görüldüğü üzere, Energy sektörü en yüksek çevresel hassasiyet ağırlığına (0.95) sahip olmakla birlikte, örnekleme nispeten düşük bir firma payına (%4.4) karşılık gelmektedir. Bu durum, ilgili sektördeki az sayıdaki firmanın model çıktısı üzerindeki etkisinin, örneklem içindeki payıyla orantısız biçimde büyük olabileceğine işaret etmekte olup, bulguların yorumlanmasında dikkate alınması gereken metodolojik bir sınırlılık oluşturmaktadır.

Üç Kanallı Mimari

Yeşil aklama davranışının tek bir veri katmanıyla eksiksiz biçimde temsil edilemeyeceği öngörüsünden hareketle, çalışmada üç bağımsız bilgi kanalını birlikte değerlendiren bir mimari kullanılmıştır.

**Şekil 1.** Üç Saç Ayağı Mimarisini

Birinci kanal, doğal dil işleme (NLP) teknikleriyle firmaların ne söylediğini; ikinci kanal, finansal göstergeler aracılığıyla firma yapısının bu söylemi ne ölçüde desteklediğini; üçüncü kanal

ise GARCH (1,1) modellemesiyle piyasanın söz konusu açıklamalara nasıl tepki verdiğini analiz etmektedir. Her kanaldan elde edilen özellikler, toplam 21 değişkenden oluşan tek bir matriste bir araya getirilerek XGBoost sınıflandırıcısına girdi olarak sunulmuştur.

T1 (Volatilitenin Çıkarıldığı) modelinde Accuracy (Doğruluk) değerinin Baseline modele göre %7.5 artarken, ROC-AUC değerinin düşmesi ilk bakışta çelişkili görünebilir. Ancak bu durum, Accuracy metriğinin eşik-bağımlı (threshold-dependent) doğası ve hedef değişkendeki dengesiz sınıf dağılımı (%69.4'e %30.6) ile açıklanmaktadır. T1 modeli, zor vakaları ayırt etmek yerine çoğunluk sınıfına (Düşük Risk) ağırlık vererek yapay bir doğruluk artışı sağlamıştır; oysa eşikten bağımsız ayrılcılık gücünü ölçen ROC-AUC düşmüştür.

Öte yandan, sadece GARCH özellikleriyle eğitilen T2 modelinde Ortalama Hassasiyetin (Average Precision) %20.7 gibi dramatik bir düşüş yaşaması kritik bir bulgudur. Bu düşüş, piyasa volatilitésinin yeşil aklama şüphesini yakalamada çok güçlü ancak tek başına bırakıldığında 'gürültülü (noisy)' bir sinyal olduğunu kanıtlar. NLP kanalının yokluğunda model, her piyasa şokunu yeşil aklama sanarak hassasiyetini (precision) kaybetmiştir.

Her kanalın işlevi, kendisine karşılık gelen bir soruyla özetlenebilir: NLP kanalı “firma ne söylüyor?”, finansal kanal “firmanın yapısı bu iddiaları destekliyor mu?” ve GARCH kanalı “piyasa bu firmaya gerçekten güveniyor mu?” sorularına yanıt aramaktadır. Kanallara ait veri kaynakları ve örnek özellikler Tablo 3’de özetlenmiştir.

Table 3. Yeşil aklamamanın tespitine yönelik üç sütunlu mimari

Kanal	Soru	Veri Kaynağı	Örnek Özellikler
NLP	“Firma ne söylüyor?”	ClimateBERT, FinBERT, metin analizi	GWL skoru, fog indeksi, hedge kelime sayısı
Finansal	“Firmanın yapısı iddiaları destekliyor mu? ”	yfinance, S&P 500	Kaldıraç, ROA, firma büyüklüğü
GARCH	“Firmanın yapısı iddiaları destekliyor mu? ”	Günlük hisse getirileri → GARCH(1,1)	Koşullu volatilité, spike sayısı, persistence

Kanal 1: Doğal Dil İşleme (NLP) Özellikleri

Bu kanal, firmaların çevresel açıklamalarında ne söylediğini metin düzeyinde analiz etmeyi amaçlamaktadır. Yeşil aklama yapan firmaların, gerçek çevresel performanslarını gizlemek amacıyla karmaşık ve belirsiz bir dil kullandığı literatürde sıklıkla dile getirilen bir gözlemdir. Bu görüş, finansal açıklamalar alanındaki bulgularla da örtüşmektedir: Li (2008), zayıf performans gösteren firmaların yıllık raporlarının daha yüksek Fog indeksine sahip ve daha az okunabilir olduğunu, buna karşılık kazançları kalıcı olan firmaların raporlarının nispeten sade bir dille yazıldığını göstermiştir. Bu çalışmada da benzer bir mekanizmanın çevresel açıklamalar için geçerli olabileceği varsayılmış; yüksek Fog indeksi ve düşük okunabilirliğin, bilginin biçimsel olarak sunulduğu ancak etkin biçimde iletilmediği bir söylem stratejisinin dolaylı göstergesi olabileceği öngörülmüştür.

SHAP analizinde ilk dört sıradaki özellikten üçünün GARCH kanalından gelmesi, çalışmanın merkezinde yer alan 'piyasa şüphesi' tezini ampirik olarak doğrulamaktadır. Ancak bu durum, modelin tek-kanal bağımlılığı (single-channel dependency) riskini de beraberinde getirmektedir. Bilhassa literatürde yeşil aklamamanın temel motivasyonu olarak gösterilen finansal baskı hipotezinin temsilcileri olan Leverage (Kaldıraç) ve ROA (Kârlılık) değişkenlerinin SHAP sıralamasında nispeten sınırlı bir ağırlığa sahip olması dikkat çekicidir. Bu bulgu, S&P 500 gibi büyük ölçekli ve kurumsallaşmış firmaların, yeşil aklamayı finansal bir çaresizlikten ziyade, stratejik bir konumlandırma aracı olarak kullandıkları tezini destekleyebilir.

TF-IDF (Terim Sıklığı–Ters Belge Sıklığı)

ESG alanına özgü dokuz anahtar terim (sustainable, green, emissions, carbon, climate, environmental, renewable, decarbonization, net-zero) için TF-IDF ağırlıkları hesaplanmıştır. TF-IDF, bir terimin belirli bir belgedeki önemini, tüm belge koleksiyonundaki seyrekliğiyle ağırlıklandırarak ölçen istatistiksel bir yöntemdir (Salton & Buckley, 1988). Bu yolla, firmaların sürdürülebilirlik söylemlerindeki kavramsal odak alanları nicelleştirilmiştir.

Loughran-McDonald ESG-Genişletilmiş Duygu Analizi

Metinlerdeki duygu skoru, Loughran ve McDonald'ın (2011) finans metinlerine özgü sözlüğünün ESG bağlamına uyarlanmış bir versiyonuyla hesaplanmıştır. Standart pozitif ve negatif terimlere ek olarak, belirsizlik bildiren hedge kelimeler (might, could, may, perhaps, potentially) de dahil edilmiş ve bu kelimeler 0,7 katsayısıyla ağırlıklandırılarak negatif polariteye eklenmiştir.

Hesaplama şu şekilde işlemektedir: önce metindeki pozitif, negatif ve belirsizlik kelimeleri sayılır; belirsizlik kelime sayısı 0.7 ile çarpılarak negatif tarafa eklenir ve toplam negatif elde edilir; son olarak

$$\text{Emotion Score} = (\text{Positive} - \text{Total Negative}) \div (\text{Positive} + \text{Total Negative}) \quad (1)$$

formülü uygulanır.

Örneğin bir metinde 10 pozitif, 4 negatif ve 6 belirsizlik kelimesi varsa, toplam negatif $4 + (6 \times 0.7) = 8.2$ olarak hesaplanır ve duygu skoru $(10 - 8.2) \div (10 + 8.2) \approx 0.10$ çıkar; bu da hafif pozitif bir tona işaret eder. Skor -1 ile +1 arasında değişmekte, +1'e yaklaşan değerler olumlu, -1'e yaklaşan değerler olumsuz, 0'a yakın değerler ise nötr veya belirsiz söylemi ifade etmektedir. Bu yaklaşım, kurumsal sürdürülebilirlik raporlarındaki belirsiz dilin potansiyel bir yeşil aklama göstergesi olabileceğine dair bulgularla tutarlıdır (Marquis vd., 2016).

Okunabilirlik Metrikleri

Metin karmaşıklığı iki standart okunabilirlik ölçütüyle değerlendirilmiştir: Flesch Reading Ease (Flesch, 1948), cümle uzunluğu ve hece yoğunluğuna dayalı bir skor olup düşük değerler daha karmaşık metne işaret etmektedir; Gunning Fog Index (Gunning, 1952) ise cümle uzunluğu ve çok heceli kelime oranına dayanmakta, yüksek değerler teknik veya akademik düzeyde karmaşık bir metni göstermektedir. Kurumsal sürdürülebilirlik raporlarındaki yüksek dilsel karmaşıklığın, bilginin kasıtlı olarak muğlaklaştırılması (obfuscation) amacı taşıyabileceği ve bu nedenle potansiyel bir yeşil aklama göstergesi olarak değerlendirilebileceği literatürde ileri sürülmüştür (Li, 2008).

Hedge Words (Belirsizlik İfadeleri) Sayımı

Kurumsal raporlardaki belirsizlik düzeyini nicelleştirmek amacıyla, düzenli ifadeler (regex) kullanılarak hedge kelime sayıları hesaplanmıştır (might, could, may, perhaps, possibly, probably, assume, suggest, potentially, believe). Firma başına hedge kelime yoğunluğu, toplam metin uzunluğuna oranlanarak normalleştirilmiştir.

ClimateBERT ile Yeşil Aklama Skoru

ClimateBERT (DistilRoBERTa) Modeli

Yeşil aklama risk skorlarının üretiminde, Webersinke ve arkadaşları (2022) tarafından geliştirilen ClimateBERT modeli kullanılmıştır. ClimateBERT, RoBERTa mimarisinin (Liu vd., 2019) DistilRoBERTa boyutuna damıtılmış ve iklim değişikliğiyle ilgili bilimsel yayınlar, politika belgeleri ve kurumsal raporlar üzerinde alana özgü ön eğitimden geçirilmiş bir transformer tabanlı dil modelidir. Bu alana özgü ön eğitim, modelin sürdürülebilirlik söylemindeki nüansları genel amaçlı dil modellerine kıyasla daha etkin yakalamasını sağlamaktadır.

Model parametreleri: mimari olarak 6 katmanlı, yaklaşık 82 milyon parametrelili DistilRoBERTa kullanılmış; ön eğitim korpusu iklim değişikliği araştırma makaleleri, IPCC raporları ve kurumsal CDP açıklamalarından oluşmuş; maksimum token uzunluğu BERT'in standart limiti olan 512 ile sınırlandırılmış; batch boyutu CPU'da 8, GPU'da 16-32 olarak ayarlanmıştır.

Bu noktada kavramsal bir sınırın net çizilmesi gerekmektedir: Bu çalışma ekonometrik veya nedensel (causal) bir etki analizi iddia etmemektedir. Geliştirilen makine öğrenmesi mimarisi bütünüyle öngörüsül (predictive) bir yapıdadır. Piyasa volatilitesi veya karmaşık raporlama dilleri, yeşil aklamaya 'sebep' olan faktörler değil; aksine, firmanın kurumsal şeffaflık eksikliğinin piyasada ve metinlerde bıraktığı izlerin (septomların) algoritmik olarak tespit edilmesidir.

Temperature Scaling ile Kalibrasyon

Model çıktılarının güvenilirliğini artırmak amacıyla temperature scaling yöntemi uygulanmıştır (Guo vd., 2017). Bu teknik, modelin ürettiği ham logit değerlerini bir sıcaklık parametresine (T) bölerek olasılık tahminlerinin keskinliğini kontrol etmektedir: ayarlanmış puan = ham puan ÷ T, ardından bu değerler softmax fonksiyonuyla olasılıklara dönüştürülmektedir. Düşük T değerleri (örn. T = 0.3) modelin daha keskin, ikili kararlara yakın tahminler üretmesini sağlarken; yüksek T değerleri (örn. T = 0.5) daha yumuşak, orta değerlere yakın bir olasılık dağılımı vermektedir.

Çalışmada iki farklı sıcaklık değeri kullanılmıştır: Söylem Stili Skoru (GWLS) için T = 0.3 seçilmiş, böylece söylemdeki belirgin yeşil aklama kalıpları daha keskin biçimde ayırt edilmesi hedeflenmiştir; Sembolik Eylem Skoru (GWLE) için ise T = 0.5 kullanılarak eylem-söylem tutarsızlığının daha nüanslı değerlendirilmesi amaçlanmıştır.

Cümle Bazlı Çıkarım

Her firma metni NLTK kütüphanesinin sent_tokenize fonksiyonuyla cümlelere ayrılmış, her cümle ayrı ayrı ClimateBERT modeline verilmiştir. Firma düzeyindeki GWLS ve GWLE skorları, cümle bazlı skorların aritmetik ortalaması alınarak hesaplanmıştır. Bu yaklaşım, uzun metinlerdeki yerel yeşil aklama kalıplarının yakalanma kapasitesini artırmaktadır.

GWL Bileşik Skoru (Geometrik Ortalama)

Firma düzeyindeki nihai yeşil aklama (GWL) riski, söylem stili (GWLS) ve sembolik eylem (GWLE) skorlarının geometrik ortalaması olarak hesaplanmıştır:

$$GWL = \sqrt{GWLS \times GWLE} \quad (2)$$

Örneğin GWLS = 0.80 ve GWLE = 0.60 olan bir firma için $GWL = \sqrt{(0.80 \times 0.60)} = \sqrt{0.48} \approx 0.693$ elde edilir.

Aritmetik ortalama yerine geometrik ortalamanın tercih edilmesinin nedeni, bu yöntemin iki bileşenin de eş zamanlı olarak yüksek olmasını gerektirmesidir. Örneğin bileşenlerden biri çok yüksek (0.90) diğeri çok düşükse (0.10), geometrik ortalama yalnızca $\sqrt{(0.90 \times 0.10)} \approx 0.30$ gibi düşük bir değer üretmektedir. Böylece yalnızca söylem veya yalnızca eylem boyutunda sorun olan firmaların yanlışlıkla "yüksek riskli" sınıfına dahil edilmesi önlenmekte; gerçek yeşil aklama riskinin, her iki boyutun da eş zamanlı olarak yüksek olmasını gerektirdiği varsayılmaktadır. NLP kanalına ait öznetelikler tablo 4'de özetlenmiştir.

Tablo 4. NLP Özellikleri (Kanal 1)

#	Özellik	Kaynak	Açıklama
1	NLP GWL Score	ClimateBERT	Yeşil aklama olasılık skoru (0–1)
2	finbert esg sentiment	FinBERT	ESG duygu polaritesi (–1 to +1)
3	flesch reading ease	Metin analizi	Okunabilirlik skoru (yüksek = basit)
4	gunning fog	Metin analizi	Fog indeksi (metin karmaşıklığı)
5	hedge words count	Sözlük tabanlı	Belirsizlik ifadesi sayısı
6	ESG_Gap	Türetilmiş	$0.3 \times \text{hedge} + 0.7 \times (1 - \text{okunabilirlik} / 100)$

Kanal 2: Finansal Özellikler

İkinci kanal, firmaların finansal yapısının çevresel iddiaları ne ölçüde desteklediğini değerlendirmeye yöneliktir. Bu doğrultuda log-firma büyüklüğü (Ln_Firm_Size), aktif kârlılığı (ROA) ve kaldıraç oranı (Leverage) yfinance verisinden elde edilmiş; sektör hassasiyeti (sector_sensitivity), her firmanın GICS sektörüne atanan 0-1 arası bir ağırlık olarak mühendislik yoluyla türetilmiştir. Bu ağırlıklar, sektörlerin karbon ayak izi, düzenleyici maruziyeti ve tarihsel ESG tartışma sıklığı dikkate alınarak belirlenmiştir; örneğin Enerji sektörüne 0.95, Bilgi Teknolojileri sektörüne ise 0.25 gibi görece düşük bir ağırlık atanmıştır. Kanala ayrıca metin tabanlı bileşik bir gösterge olan proxy_gwl_score dahil edilmiştir. Yüksek kaldıraçlı firmaların sermaye maliyetini düşürmek amacıyla ESG performanslarını olduğundan iyi gösterme eğiliminde olabileceği, büyük firmaların ise kamuoyu denetimine daha fazla maruz kaldığı, kurumsal finans literatüründeki ESG-firma yapısı ilişkisine dair tartışmalarla tutarlıdır (Gillan, vd., 2021). Bu çerçevede enerji gibi karbon yoğunluğu ve düzenleyici maruziyeti yüksek sektörlerin, ESG açısından yapısal olarak daha riskli kabul edilmesi de aynı mantığın bir uzantısı olarak değerlendirilebilir.

Tablo 5. Finansal Özellikler (Kanal 2)

#	Özellik	Kaynak	Açıklama
7	Ln Firm Size	yfinance	Log piyasa değeri
8	ROA	yfinance	Aktif karlılığı
9	Leverage	yfinance	Borç/özkaynak oranı
10	sector sensitivity	Mühendislik	GICS sektörü ESG hassasiyeti (0–1)
11	proxy_gwl_score	Mühendislik	Metin tabanlı bileşik GWL proxy

Kanal 3: GARCH (1,1) Volatilite Modellemesi

Bu kanal, piyasanın firma açıklamalarına verdiği tepkiyi, hisse senedi getirilerindeki koşullu volatilité üzerinden ölçmeyi amaçlamaktadır.

GARCH (1,1) Modeli

Piyasa tabanlı risk sinyallerini yakalamak amacıyla, her firma için ayrı ayrı GARCH (1,1) modeli (Bollerslev, 1986) tahmin edilmiştir. Model, günlük log-getiriler (r_t , 100 ile ölçeklenmiş) üzerinde aşağıdaki spesifikasyonla çalışmaktadır:

$$r_t = \mu + \varepsilon_t, \varepsilon_t = \sigma_t \cdot z_t, z_t \sim N(0,1) \quad (3)$$

$$\sigma_t^2 = \omega + \alpha \cdot \varepsilon_{t-1}^2 + \beta \cdot \sigma_{t-1}^2 \quad (4)$$

Burada σ_t , koşullu volatilitiyi (piyasanın ilgili gündeki belirsizlik algısını) temsil etmektedir. Denklemdaki ω parametresi firmanın taşıdığı temel dalgalanma düzeyini, α parametresi bir önceki günkü beklenmedik şokun bugünkü volatiliteye etkisini, β parametresi ise dünkü volatilitenin bugüne ne ölçüde taşındığını ifade etmektedir. α ile β 'nin toplamı, yaygın olarak volatilitenin kalıcılığının bir göstergesi olarak yorumlanmaktadır; bu toplamın 1'e yaklaşması, bir şokun etkisinin daha uzun sürdüğü anlamına gelmektedir (Engle & Patton, 2001). Örneğin bu toplamın 0.95 olduğu bir firmada piyasa şokunun etkisi aylarca sürebilirken, 0.70 olduğu bir firmada birkaç hafta içinde sönümlenmektedir.

Model, arch Python kütüphanesiyle sabit ortalama (constant mean) ve normal dağılım varsayımı altında, en az 100 günlük gözlemi bulunan firmalar için tahmin edilmiş; örneklemdeki 501 firmanın tamamında tahmin başarıyla tamamlanmıştır.

GARCH'tan Firma Bazlı Öznitelik Türetimi

Her firmaya ait koşullu volatilité serisi, firma düzeyinde özet istatistiklere indirgenerek toplam on adet kesitsel (cross-sectional) öznitelige dönüştürülmüş ve Tablo 6'da listelenmiştir.

Bu özniteliklerin dahil edilme gerekçesi iki ayrı bulguya dayanmaktadır. İlk olarak, yatırımcıların bir firmanın kurumsal sosyal sorumluluk (CSR) uygulamalarına ilişkin olumlu ve olumsuz haberlere güçlü ve asimetrik fiyat tepkileri verdiği gösterilmiştir; özellikle olumsuz olaylara verilen tepkinin olumlu olaylara kıyasla daha belirgin olduğu bulunmuştur (Krüger, 2015). Bu çalışmada, benzer bir mekanizmanın fiyat düzeyinin ötesine geçerek koşullu volatiliteye de yansiyebileceği varsayılmış; yani yatırımcıların bir firmanın ESG beyanlarına duyduğu şüphenin, yalnızca getirilerde değil, getirilerin dağılımındaki artışta (koşullu volatilité) da kendini gösterebileceği öngörülmüştür. İkinci olarak, ESG/CSR profilinin firmaların risk düzeyiyle ilişkili olduğu, ancak bu ilişkiye dair bulguların literatürde hâlâ tartışmalı olduğu belirtilmektedir (Gillan, vd., 2021).

Tablo 6. GARCH Tabanlı Özelliklerin XGBoost Modelinde Kullanımı

#	Özellik	Hesaplama	Ekonomik Anlamı
1	garch_vol_mean	σ_t 'nin ortalaması	Firmanın ortalama piyasa belirsizliği
2	garch_vol_std	σ_t 'nin standart sapması	Belirsizliğin ne kadar değişken olduğu
3	garch_vol_last	Son σ_T değeri	Piyanın en güncel şüphe seviyesi
4	garch_vol_max	Maksimum σ_t	En şiddetli piyasa tepkisi
5	garch_vol_q95	σ_t 'nin 95. yüzdeliği	Uç değer risk ölçüsü
6	garch_spike_count	z-skoru > 2 olan gün sayısı	Kaç kez anormal volatilité yaşandı
7	garch_spike_ratio	spike count / toplam gün	Olay sıklığının oranı
8	garch_persistence	$\alpha + \beta$	Şokların ne kadar süre devam ettiği
9	garch_rolling_vol_60d	60 günlük hareketli σ ortalaması	Volatilité rejimi göstergesi
10	vol_return_corr	Kor ($ r_t $, σ_t)	GARCH model uyum kalitesi

Veri Sızıntısı Önlemi

GARCH özelliklerinin türetilmesinde yalnızca tahmin noktasından önceki tarihsel getiriler (2023-01-04 ile 2024-12-30 arası dönem) kullanılmıştır. Sağlamlık testleri kapsamında, her firmanın

getiri serisinin yalnızca ilk %70'i kullanılarak gecikmeli (lagged) volatilité özellikleri ayrıca üretilmiş ve bu yolla veri sızıntısı riskinin bulunmadığı doğrulanmıştır.

Özellik Matrisi ve Model Girdisi

NLP kanalından altı, finansal kanaldan beş ve GARCH kanalından on olmak üzere toplam 21 öznitelik, tek bir matriste birleştirilerek XGBoost sınıflandırıcısına girdi olarak sunulmuştur. Bu bütünleşik yapı, üç kanalın istatistiksel olarak bağımsız biçimde hesaplanmasına karşın, nihai kararın izole değil birleşik bir değerlendirme üzerinden verilmesini sağlamaktadır.

Sınıflandırma Modeli: XGBoost

Yeşil aklama riskinin sınıflandırılmasında XGBoost (Extreme Gradient Boosting) (Chen & Guestrin, 2016) algoritması tercih edilmiştir. XGBoost, gradyan artırma (gradient boosting) çerçevesinin ölçeklenebilir ve düzenlileştirilmiş bir uygulaması olup, karar ağaçlarını ardışık biçimde inşa ederek her yeni ağacın bir öncekinin artık hatalarını (residual errors) azaltmasını hedeflemektedir (Friedman, 2001). Bu çalışma için tercih edilme gerekçeleri şu şekilde özetlenebilir: (i) NLP skorları, finansal oranlar ve volatilité metrikleri gibi farklı ölçek ve dağılıma sahip öznitelikler tek bir modelde etkin biçimde bütünleştirebilmesi; (ii) L1/L2 düzenlileştirme yoluyla aşırı öğrenmeye karşı yerleşik koruma sunması; (iii) ağaç tabanlı yapısı sayesinde öznitelikler arası doğrusal olmayan etkileşimleri örtük biçimde modelleyebilmesi; (iv) scale_pos_weight parametresiyle sınıf dengesizliğine uyum sağlayabilmesi; ve (v) SHAP (Lundberg & Lee, 2017) ile doğrudan uyumlu, yorumlanabilir bir yapı sunmasıdır.

Hyperparameter Optimizasyonu

Model hiperparametreleri, RandomizedSearchCV (Bergstra & Bengio, 2012) ile optimize edilmiştir; bu yöntem, tam grid aramasına kıyasla daha düşük hesaplama maliyetiyle yüksek boyutlu parametre uzaylarında etkin bir keşif imkânı sunmaktadır. Arama uzayı, n_estimators için {50, 100, 150, 200, 250, 300}, learning_rate için {0.01; 0.03; 0.05; 0.1; 0.15; 0.2} ve max_depth için {2, 3, 4, 5, 6, 7} değerlerini kapsamış; optimizasyon 3 katlı çapraz doğrulama altında 15 rastgele kombinasyon üzerinden, doğruluk (accuracy) metriği esas alınarak yürütülmüştür. Bu süreç sonunda n_estimators=200, max_depth=4 ve learning_rate=0.15 değerleri optimum olarak seçilmiştir. Öte yandan, stokastik gradyan artırma (stochastic gradient boosting) yapısını korumak ve modelin aşırı öğrenmesini (overfitting) engellemek amacıyla subsample ve colsample_bytree parametreleri hiperparametre arama uzayına dahil edilmemiş; literatürdeki yaygın best-practice uygulamaları (Chen & Guestrin, 2016) doğrultusunda 0.8 olarak sabitlenmiştir.

Sınıf Dengesizliği Yönetimi

Yüzdelik dilim eşiklemesi, örnekleme yaklaşık 3:1 oranında dengesiz bir sınıf dağılımı ortaya çıkarmaktadır. Bu dengesizliği gidermek amacıyla XGBoost'un scale_pos_weight parametresi, güvenli sınıftaki firma sayısının riskli sınıftaki firma sayısına oranı olarak dinamik biçimde ayarlanmıştır. Bu yaklaşımda, model eğitim sırasında bir riskli firmayı yanlış sınıflandırdığında bu hatanın maliyeti, güvenli bir firmayı yanlış sınıflandırma maliyetinin katları düzeyinde ağırlıklandırılmakta; böylece azınlık sınıfının (riskli firmalar) öğrenme sürecindeki etkisi artırılmaktadır. Bu strateji, maliyet duyarlı öğrenme (cost-sensitive learning) paradigmasıyla uyumludur (He & Garcia, 2009).

Eğitim-Test Ayrımı

Veri seti, tabakalı örnekleme (stratified sampling) ile %80 eğitim ve %20 test olarak ayrılmıştır (random_state=42); bu bölünme, her iki alt kümedeki sınıf dağılımının orijinal örneklemedeki oranı korumasını sağlamaktadır. Sonuç olarak model 402 firma üzerinde eğitilmiş; performans değerlendirmesi, modelin eğitim sürecinde hiç görmediği kalan 101 firma üzerinde gerçekleştirilmiştir. Model eğitim yapılandırmasının özeti Tablo 7'de sunulmuştur.

Tablo 7. Model eğitim yapılandırması

Parametre	Değer
Algoritma	XGBoost (gradient boosted trees)
Hiperparametre seçimi	RandomizedSearchCV (15 iterations, 3-fold CV)
n_estimators	200
max_depth	4
learning_rate	0.15
subsample	0.8
colsample_bytree	0.8
Eğitim/test ayrımı	80% / 20% stratified (random_state=42)
Eğitim seti	402 firma
Test seti	101 firma (model tarafından hiç görülmemiş)

Model Açıklanabilirliği: SHAP Analizi

Modelin karar mekanizmasını yorumlamak amacıyla, oyun teorisi temelli bir açıklanabilirlik yöntemi olan SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) kullanılmıştır. Analiz, XGBoost modeline TreeExplainer aracılığıyla uygulanmış; her özneteliğin model çıktısına yaptığı ortalama katkı, mutlak SHAP değerleri üzerinden hesaplanmıştır. Bu analiz, yalnızca modelin genel performansını değil, hangi kanalın hangi koşullarda baskın hale geldiğini de ortaya koymayı amaçlamaktadır.

Dayanıklılık(Robustness) Testlerinin Tasarımı

GARCH entegrasyonunun katkısını ve güvenilirliğini sınamak amacıyla üç ayrı test tasarlanmıştır. Testlerin tümü, aynı sabit hiperparametreler ($n_estimators=200$, $max_depth=4$, $learning_rate=0.15$) ve aynı tabakalı %80/%20 ayrımı ($random_state=42$) altında yürütülmüştür; bu sayede testler arası farkların yalnızca öznetelik kümesinden kaynaklandığı garanti edilmiştir.

Test 1 (GARCH olmadan – ablasyon), on GARCH özneteliğinin tamamının çıkarılarak modelin yalnızca kalan on bir NLP ve finansal öznetelikle eğitilmesini içermektedir; bu test, GARCH kanalının modelin genel başarısına ne ölçüde katkı sağladığını ortaya koymayı amaçlamaktadır.

Test 2 (yalnızca GARCH), NLP ve finansal özneteliklerin tamamen çıkarılıp modelin yalnızca on GARCH özneteliğiyle eğitilmesine dayanmaktadır; bu tasarım, piyasa sinyalinin tek başına yeterli olup olmadığını sınamaktadır.

Test 3 (gecikmeli GARCH – veri sızıntısı karşıtı test), her firma için GARCH(1,1) modelinin, getiri geçmişinin yalnızca ilk %70'i (Ocak 2023 – yaklaşık Eylül 2024 dönemi) kullanılarak yeniden tahmin edilmesini içermektedir; Ekim–Aralık 2024 dönemine ait veriler bu tahminde hiçbir şekilde kullanılmamıştır. Bu tasarım, tahmin noktası ile öznetelik hesaplama noktası arasında kesin bir zamansal boşluk (lag) yaratarak, GARCH sinyalinin ileriye dönük bilgiyle (contemporaneous information) şişirilip şişirilmediğinin sınanmasına imkân tanımaktadır.

Veri Sızıntısı Önlemlerinin Özeti

Çalışma boyunca veri sızıntısı riskine karşı iki temel önlem alınmıştır. Birincisi, bağımlı değişkenin örneklem dağılımından dinamik olarak türetilmesi ve doğrudan model girdilerinden hesaplanmamasıdır. İkincisi, GARCH özneteliklerinin yalnızca tahmin noktasından önceki tarihsel

getiriler kullanılarak hesaplanması ve bu tasarımın gecikmeli bir yeniden tahmin testiyle ayrıca doğrulanmasıdır.

Yazılım ve Donanım Ortamı

Tüm analizler Python 3.10+ programlama dilinde gerçekleştirilmiştir. Kullanılan temel kütüphaneler ve bunların kullanım amaçları Tablo 8’de listelenmiştir.

Tablo 8. Analizde kullanılan kütüphaneler

Kütüphane	Versiyon	Kullanım Amacı
transformers (HuggingFace)	≥ 4.30	ClimateBERT model çıkarımı
torch (PyTorch)	≥ 2.0	Deep learning altyapısı
xgboost	≥ 1.7	Gradyan artırma sınıflandırıcısı
scikit-learn	≥ 1.2	ML pipeline, metrikler, çapraz doğrulama
arch	≥ 5.5	GARCH volatilité modelleme
shap	≥ 0.42	Model açıklanabilirlik analizi
yfinance	≥ 0.2	Finansal veri çekimi
nltk	≥ 3.8	Metin ön işleme, cümle tokenizasyonu
textstat	≥ 0.7	Okunabilirlik indeksleri
pandas	≥ 2.0	Veri manipülasyonu
numpy	≥ 1.24	Görselleştirme
matplotlib	≥ 3.7	ClimateBERT model çıkarımı
seaborn	≥ 0.12	Deep learning altyapısı

Bulgular

Tanımlayıcı İstatistikler

Örnekleme oluşturan 503 firmanın Global Industry Classification Standard (GICS) sektörlerine göre dağılımı Tablo 9’da sunulmuştur. Örnekleimde en fazla temsil edilen sektörler Sanayi (%15.7) ve Finans (%15.1) olurken, Enerji (%4.4) ve İletişim Hizmetleri (%4.6) en düşük paya sahip sektörler olarak dikkat çekmektedir. Enerji sektörü örneklem içerisinde düşük paya sahip olmasına rağmen yüksek karbon yoğunluğu nedeniyle modelde yüksek bir sektörel hassasiyet ağırlığına sahiptir. Bu sonuç sektördeki az sayıdaki firmanın sınıflandırma üzerindeki etkisinin orantısız olabileceğine işaret etmektedir.

Tablo 9. Örneklemin sektör dağılımı

Sektör	Firma sayısı	Oran
Endüstriyel Ürünler	79	15.7%
Finans	76	15.1%
Bilgi Teknolojileri	73	14.5%
Sağlık Hizmetleri	59	11.7%
İsteğe Bağlı Tüketim	48	9.5%
Temel Tüketim Ürünleri	35	7.0%

Kamu Hizmetleri	31	6.2%
Gayrimenkul	31	6.2%
Ham Madde / Malzeme	26	5.2%
İletişim Hizmetleri	23	4.6%
Enerji	22	4.4%

GARCH (1,1) modelinden türetilen on özelliğe ait tanımlayıcı istatistikler Tablo 10'da yer almaktadır; bu tablo, firma bazında koşullu volatilitenin (garch_vol_last) ortalama 1.70 düzeyinde seyrettiğini, ancak maksimum değerin 7.61'e kadar çıkabildiğini, dolayısıyla firmalar arası belirgin bir heterojenliğin bulunduğunu göstermektedir.

Tablo 10. GARCH özelliklerine ilişkin tanımlayıcı istatistikler (N = 501 firma)

Özellik	Min	Ortalama	Max
garch_vol_mean	0.82	1.76	6.66
garch_vol_std	0.00	0.24	3.53
garch_vol_last	0.80	1.70	7.61
garch_vol_max	1.02	3.82	76.51
garch_vol_q95	0.94	2.14	11.77
garch_spike_count	0	18.6	40
garch_spike_ratio	0.00	0.04	0.08
garch_persistence	0.00	0.66	1.00
garch_rolling_vol_60d	0.81	1.75	6.54
vol_return_corr	-0.07	0.12	0.42

Analizler neticesinde garch_vol_max değişkeninde 76.51 gibi ortalamanın (1.76) çok üzerinde bir uç değer (outlier) bulunduğu görülmektedir. Bu değer Globe Life Inc. (GL) firmasının 11 Nisan 2024 tarihinde yaşadığı %53'lük tek günlük kurumsal skandal çöküşünden kaynaklandığı tahmin edilmektedir. GARCH modeli bu ani piyasa şokunu doğru bir şekilde yakaladığı da tespit edilmiştir. Çalışmada herhangi bir veri kırpma (winsorization) işlemi uygulanmamıştır. Bunun iki nedeni vardır; birincisi, Kullanılan XGBoost algoritması ağaç tabanlı (tree-based) bir model olduğu için uç değerlere karşı (outlier-robust) doğal bir dirence sahiptir ve lineer modellerdeki gibi tahmini bozmaz (Chen & Guestrin, 2016). İkincisi ise, bu tür ekstrem kurumsal güven şokları, yeşil aklama ve ESG riskinin bizzat doğasını yansıtan nadir ama değerli sinyallerdir.

Model Performansı: Held-Out Test Seti Sonuçları

Modelin, eğitim sürecinde hiç görmediği 101 firma üzerindeki performansı Tablo 11'de özetlenmiştir. Örnekleme, 503 firma üzerinden rastgele ve katmanlı (stratified) örnekleme yöntemiyle %80 eğitim (402 firma) ve %20 test (101 firma) alt kümelerine ayrılmış; model yalnızca eğitim setindeki 402 firma üzerinde geliştirilmiş, performansı ise eğitim sürecine hiçbir şekilde dahil edilmemiş 101 firmalık bağımsız test setinde değerlendirilmiştir. Bu ayırım, raporlanan sonuçların veri sızıntısından (data leakage) arındırılmış, gerçek dışsal geçerliliği (out-of-sample validity) yansıtan bir performans tahmini sunmasını sağlamaktadır.

Tablo 11. Ayrılmış test seti performansı

Metric	Değer
ROC-AUC	0.704
Spearman ρ	0.756
F1 (macro)	0.657
Accuracy	0.703

Nihai model, gecikmeli (lagged) GARCH özellik setiyle geliştirilen konfigürasyon, held-out test setinde 0.704 ROC-AUC ve 0.756 Spearman sıralama korelasyonuna ulaşmıştır.

Sınıflandırma tabanlı finansal risk modellerinde, kısıtlı sayıda açıklayıcı değişkenle çalışıldığında 0.70 civarındaki AUC değerlerinin bilgi taşıyıcı (informative) kabul edildiği yakın tarihli ampirik çalışmalarda da desteklenmektedir. Örneğin, firma düzeyinde ESG derecelendirmelerini kullanarak karbon emisyon riskini sınıflandıran bir çalışmada, zamansal (out-of-time) dayanıklılık testinde 0.70 civarında AUC değerleri elde edilmiş ve bu sonuç pratik tarama (screening) amaçları için bilgi taşıyıcı olarak değerlendirilmiştir (Kim & Na, 2026)

Bununla birlikte, test setinin sınırlı büyüklüğü (101 firma; 31 pozitif, 70 negatif gözlem) dikkate alındığında, bu örneklem hacminde tek bir yanlış sınıflandırmanın AUC değerini yaklaşık 0.01 puan oynatabileceği göz önünde bulundurulmalıdır. Nitekim izleyen bölümde sunulan dayanıklılık testleri, farklı özellik konfigürasyonları arasında 0.64-0.70 aralığında değişen AUC değerleri üretmekte olup, bu durum nokta tahminlerin belirli bir belirsizlik payıyla yorumlanması gerektiğini doğrulamaktadır.

Spearman sıralama korelasyonunun 0.756 düzeyinde gerçekleşmesi, modelin firmaları risk düzeyine göre büyük ölçüde tutarlı biçimde sıraladığını göstermektedir. Portföy yönetimi uygulamaları açısından, ikili sınıflandırma doğruluğundan ziyade firmaların göreceli risk sıralamasının istikrarlı olması daha belirleyici bir kriterdir; bu açıdan modelin sıralama performansı, ikili sınıflandırma metriklerine kıyasla daha güçlü bir sonuç ortaya koymaktadır.

Dayanıklılık (Robustness) Testi Bulguları

Modelin farklı özellik konfigürasyonları altındaki dayanıklılığını değerlendirmek amacıyla üç ayrı test uygulanmış ve sonuçlar, baseline (referans) modelle karşılaştırmalı olarak Tablo 12'de sunulmuştur.

Tablo 12. Dayanıklılık testleri: Yan yana karşılaştırma

Metrik	Temel Model	T1: GARCH Yok	T2: Yalnızca GARCH	T3: Gecikmeli GARCH
ROC-AUC	0.653	0.641 (-1.8%)	0.638 (-2.3%)	0.704 (+7.8%)
Avg Precision	0.589	0.585 (-0.7%)	0.467 (-20.7%)	0.627 (+6.6%)
F1 (macro)	0.604	0.653 (+8.0%)	0.596 (-1.3%)	0.657 (+8.7%)
Precision (macro)	0.604	0.658 (+8.9%)	0.596 (-1.5%)	0.655 (+8.3%)
Recall (macro)	0.604	0.649 (+7.4%)	0.597 (-1.2%)	0.660 (+9.2%)
Accuracy	0.663	0.713 (+7.5%)	0.653 (-1.5%)	0.703 (+6.0%)
Spearman ρ	0.739	0.737 (-0.2%)	0.728(-1.4%)	0.756 (+2.3%)

Baseline model — GARCH özelliklerinin eşzamanlı (contemporaneous) biçimde kullanıldığı konfigürasyon — 0.653 ROC-AUC ve 0.739 Spearman ρ değerlerine ulaşmıştır. Bu referans değerler temel alınarak, modelin dayanıklılığı üç farklı özellik konfigürasyonu altında sınanmıştır: GARCH özelliklerinin tamamen çıkarıldığı bir ablasyon testi (T1), yalnızca GARCH özelliklerinin kullanıldığı bir izolasyon testi (T2) ve GARCH özelliklerinin bir dönem geciktirilerek kullanıldığı bir zamanlama testi (T3). Genel tabloda görüldüğü üzere, bu üç konfigürasyon performans üzerinde farklı yönde ve farklı büyüklükte etkiler yaratmıştır.

Test 1 –GARCH Özellikleri Çıkarıldığında

Bu testin amacı, modelin GARCH tabanlı volatilité özellikleri olmaksızın ne düzeyde işlevsel kaldığını sınamaktır. Bu doğrultuda, önceki bölümde tanımlanan 10 GARCH özelliğinin tamamı özellik setinden çıkarılmış ve model yalnızca kalan 11 doğal dil işleme (NLP) ve finansal özellikle yeniden eğitilmiştir. Bu ablasyon testi, volatilité sinyalinin modelin ayrıştırma gücüne ne ölçüde zorunlu bir katkı sağladığını, yoksa yalnızca tamamlayıcı bir rol mü üstlendiğini ortaya koymayı amaçlamaktadır. Baseline model ile T1 konfigürasyonu arasındaki karşılaştırmalı sonuçlar Tablo 13’de sunulmuştur.

Tablo 13. TEST 1: Ablasyon (volatilité özellikleri Çıkarılması)

Metrik	Temel Model (21 özellik)	T1: Volatilité yok (11 özellik)	Fark
ROC-AUC	0.653	0.641	-0.012
Spearman ρ	0.739	0.737	-0.002
F1 (macro)	0.604	0.653	+0.049

GARCH özelliklerinin tamamen çıkarıldığı bu senaryoda ROC-AUC yalnızca 0.012 puan (%1,8) gerilemiş, Spearman ρ ise neredeyse değişmeden kalmıştır. F1 (macro) metriğinde gözlenen +0.049 puanlık artış, Tablo 13’de raporlanan Accuracy artışıyla (+%7,5) birlikte değerlendirildiğinde, daha az özellikle çalışan modelin küçük test setinde daha düşük varyans sergilediğine işaret etmektedir.

Bu bulgular bir bütün olarak değerlendirildiğinde, NLP ve finansal özelliklerin tek başına dahi güçlü bir ayrıştırma temeli oluşturduğu ve modelin GARCH tabanlı volatilité kanalının yokluğunda çökmediği görülmektedir. Bu durum, GARCH sinyalinin modelin performansı için zorunlu bir bileşen değil, tamamlayıcı (complementary) bir sinyal olarak işlev gördüğü şeklinde yorumlanabilir; volatilité bilgisi, model için bir bağımlılık değil bir katkı kaynağıdır.

Bu yorumu destekler nitelikte, GARCH özellikleri çıkarıldığında SHAP önem sıralamasında Leverage (0.662) ve gunning_fog (0.618) değişkenlerinin en belirleyici özellikler haline geldiği gözlenmiştir. Bu yeniden dağılım, modelin piyasa kaynaklı volatilité sinyali mevcut olmadığında, firmanın finansal yapısına (kaldıraç oranı) ve açıklama metinlerinin dilsel karmaşıklığına (gunning_fog okunabilirlik endeksi) dayanarak telafi mekanizması geliştirdiğini göstermektedir.

Test 2 – Yalnızca GARCH Özellikleri ile Model

Bu testin amacı, GARCH tabanlı volatilité sinyalinin tek başına ne düzeyde tahmin gücüne sahip olduğunu sınamaktır. Bu doğrultuda, tüm NLP ve finansal özellikler özellik setinden çıkarılmış ve model yalnızca on GARCH özelliğiyle yeniden eğitilmiştir. Baseline model ile T2 konfigürasyonu arasındaki karşılaştırmalı sonuçlar Tablo 14’de sunulmuştur.

Table 14. TEST 2: Yalnızca Volatilite Modeli

Metrik	Temel Model (21 özel.)	T2: Yalnızca Volatilite (10 özel.)	Fark
ROC-AUC	0.653	0.638	-0.015
Avg Precision	0.589	0.467	-0.122
Spearman ρ	0.739	0.728	-0.011

Yalnızca GARCH özelliklerinin kullanıldığı bu senaryoda ROC-AUC sınırlı bir düşüş göstermiş (0.638; -%2.3), buna karşın Average Precision metriği %20.7 oranında gerileyerek 0.467'ye inmiştir. Bu örüntü, GARCH volatilite sinyalinin firmaları risk sıralamasına makul ölçüde yerleştirebildiğini, ancak karar eşiğinde (threshold) belirgin sayıda yanlış pozitif ürettiğini göstermektedir; başka bir ifadeyle, yüksek volatiliteye sahip firmaların önemli bir kısmı, gerçekte öyle olmadıkları hâlde yüksek riskli olarak işaretlenme eğilimindedir.

Bu durum, yüksek volatilitenin tek başına çevresel risk göstergesi olarak yorumlanmasının metodolojik bir sınırlılık taşıdığına işaret etmektedir; zira firma düzeyinde volatilite, sektörel büyüme dinamikleri, araştırma-geliştirme yoğunluğu veya piyasa likiditesi gibi çevresel içerikten bağımsız çok sayıda faktörden kaynaklanabilir. Bu bağlamda, NLP tabanlı özelliklerin — özellikle metin karmaşıklığını ölçen *gunning_fog* ve muğlak/temkinli dil kullanımını yakalayan *ESG_Gap* değişkenlerinin — çevresel kökenli şüpheli volatiliteyi işletme faaliyetinden kaynaklanan olağan volatiliteden ayırt etmek bakımından belirleyici bir rol üstlendiği değerlendirilmektedir. Bulgular bir bütün olarak, yalnızca piyasa sinyaline dayanan bir tespit stratejisinin risk taşıdığını; volatilitenin sıralama düzeyinde doğru yönü işaret etmesine karşın, NLP desteği olmaksızın karar eşiğinde hassasiyetin (precision) belirgin biçimde zayıfladığını ortaya koymaktadır.

Test 3 – Gecikmeli GARCH (Veri Sızıntısı Testi)

Bu test, GARCH tabanlı özelliklerin gerçekten yapısal ve kalıcı bir sinyal mi taşıdığını, yoksa çağdaş (contemporaneous) bilgiyle şişirilip şişirilmediğini sınamayı amaçlamaktadır. Bu doğrultuda, her firma için GARCH (1,1) parametreleri yalnızca getiri geçmişinin ilk %70'lik dilimi kullanılarak yeniden tahmin edilmiş; böylece tahmin noktası ile özellik hesaplama penceresi arasında açık bir zamansal boşluk (temporal gap) yaratılmıştır. Baseline model ile T3 konfigürasyonu arasındaki karşılaştırmalı sonuçlar Tablo 15'de sunulmuştur.

Tablo 15. TEST 3: Gecikmeli volatilite (Veri sızıntısını önleme testi)

Metrik	Temel Model (tam veri)	T3: Gecikmeli (ilk 70%)	Fark
ROC-AUC	0.653	0.704	+0.051
Avg Precision	0.589	0.627	+0.038
F1 (macro)	0.604	0.657	+0.053
Accuracy	0.663	0.703	+0.040
Spearman ρ	0.739	0.756	+0.017

Held-out test setindeki 0.653 ROC-AUC, bir önceki bölümde tartışıldığı üzere, kısıtlı sayıda açıklayıcı değişkenle çalışan sınıflandırma modelleri için bilgi taşıyıcı kabul edilen bir düzeyi yansıtmaktadır. Küçük test setinin (N=101) metrik varyansını artırdığı göz önünde

bulundurulmalıdır. Modelin asıl gücü sıralama kalitesindedir (Spearman $\rho = 0.739$, $p < 10^{-18}$) ki bu portföy tarama uygulamaları için en aksiyonel metriktir. GARCH parametrelerinin yalnızca getiri geçmişinin ilk %70'i kullanılarak yeniden tahmin edildiği bu senaryoda, model incelenen tüm metriklerde tam veri setiyle eğitilen baseline modeli geride bırakmıştır (ROC-AUC: 0.704 karşısında 0.653; Spearman ρ : 0.756 karşısında 0.739). İlk bakışta beklenmedik görünen bu sonuç, GARCH(1,1) parametrelerinin daha kısa ve istikrarlı bir pencere üzerinde tahmin edildiğinde örneklem dışı (out-of-sample) genelleme performansının arttığı; buna karşılık tam örneklem penceresinin, son döneme özgü kısa vadeli gürültüyü de içerebileceği biçiminde yorumlanabilir. Performansın azalmak yerine artmış olması, aynı zamanda GARCH sinyalinin ileriye dönük (look-ahead) bilgiyle şişirilmediğine dair güçlü bir gösterge olarak değerlendirilebilir: eğer özellik hesaplamasında bir zamanlama tutarsızlığı (data leakage) söz konusu olsaydı, gecikmeli modelin bu avantajı kaybetmesiyle performansın azalması beklenirdi. Bununla birlikte, bu bulgu tek başına kesin bir kanıt olarak değil, modelin zamanlama tutarlılığı açısından olumlu bir işaret olarak yorumlanmalıdır. Bu testte SHAP önem sıralaması da değişmiş; garch_rolling_vol_60d değişkeni garch_vol_last'ı geride bırakarak birinci sıraya yükselmiştir. Bu değişim, son gün volatilitesi bilgisinin aylar öncesine ait bir gözlem hâline geldiği bu senaryoda, 60 günlük hareketli ortalamanın firmanın yapısal volatilité rejimini daha güvenilir biçimde temsil etmesiyle açıklanabilir.

SHAP Analizi Bulguları

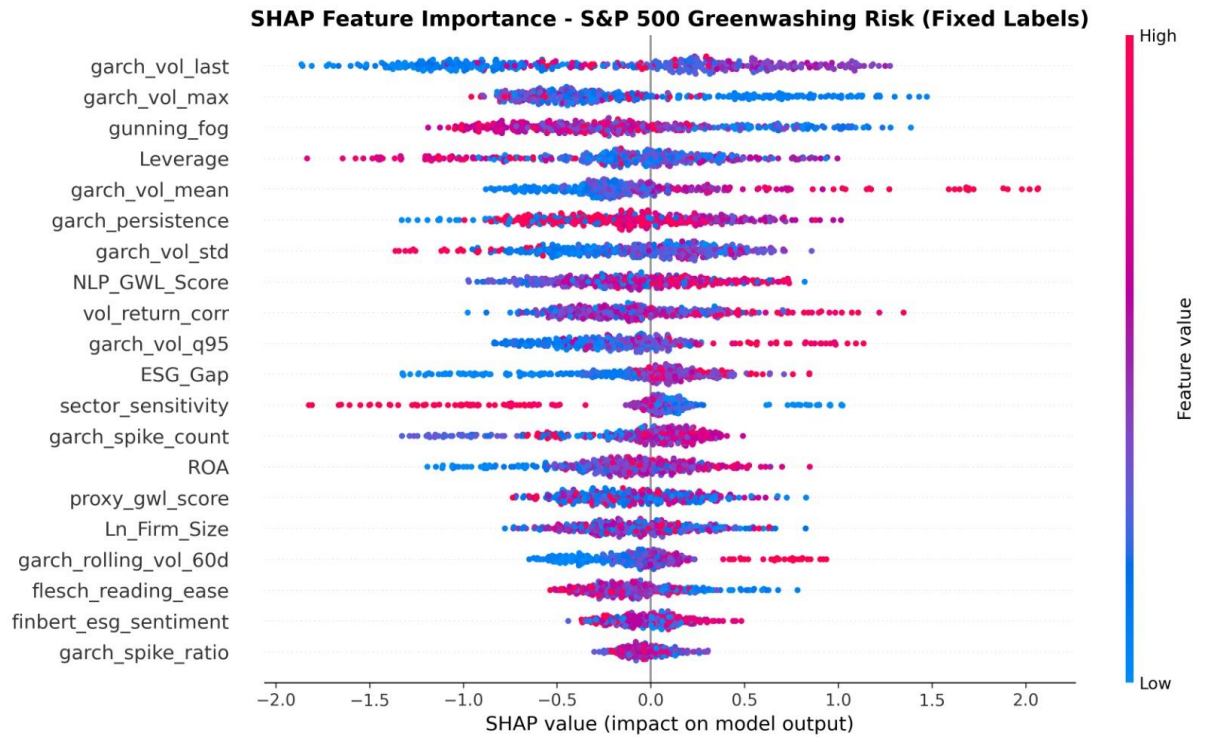
SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), her özelliğin model çıktısına yaptığı katkıyı oyun teorisi temelli bir yaklaşımla ölçen bir açıklanabilirlik yöntemidir. Bu çalışmada, TreeExplainer algoritması kullanılarak XGBoost modeline uygulanmıştır. Tam modele ait SHAP özellik önem sıralaması Tablo 16'da, kanal bazında özet karşılaştırma Tablo 17'de ve dört farklı model varyantı arasındaki karşılaştırma Tablo 18'de sunulmaktadır. Sıralamanın görsel dağılımı ayrıca Şekil 2'de yer almaktadır.

Tablo 16. SHAP özellik önem sıralaması (tam model)

Sıralama	Özellik	SHAP ortalama	Kanal	Yorum
1	garch_vol_last	0.668	GARCH	Piyananın güncel şüphe seviyesi
2	garch_vol_mean	0.452	GARCH	Ortalama belirsizlik
3	garch_vol_max	0.448	GARCH	En şiddetli piyasa tepkisi
4	Leverage	0.403	Finansal	Borçlanma baskısı
5	gunning_fog	0.358	NLP	Metin karmaşıklığı
6	garch_persistence	0.340	GARCH	Şokların kalıcılığı
7	vol_return_corr	0.308	GARCH	Uç risk ölçüsü
8	garch_vol_q95	0.306	GARCH	Hedge kelime yoğunluğu
9	ESG_Gap	0.300	NLP	Karlılık
10	ROA	0.295	Finansal	Belirsizlik dalgalanması
11	garch_vol_std	0.293	GARCH	Sektör hassasiyeti
12	sector_sensitivity	0.273	Finansal	Olay sıklığı
13	garch_spike_count	0.270	GARCH	ClimateBERT skoru
14	NLP_GWL_Score	0.253	NLP	GWL proxy
15	proxy_gwl_score	0.251	NLP	Firma büyüklüğü

16	Ln Firm Size	0.229	Finansal	Rejim göstergesi
17	garch_rolling_vol_60d	0.216	GARCH	Okunabilirlik
18	flesch_reading_ease	0.200	NLP	ESG duygu analizi
19	finbert_esg_sentiment	0.135	NLP	Normalleştirilmiş olay oranı
20	garch_spike_ratio	0.090	GARCH	Model uyum kalitesi
21	hedge_words_count	0.011	NLP	Ham hedge sayısı

SHAP analizinde en yüksek öneme sahip dört değişkenden üçünün GARCH kanalından türemiş olması, bu çalışmanın temelini oluşturan "piyasa şüpheliği" hipotezini ampirik olarak doğrulamaktadır. Bununla birlikte, bu durum modelde tek kanala dayalı bir bağımlılık riskini de beraberinde getirmektedir. Özellikle, literatürde yeşil aklamanın başlıca motivasyonu olarak gösterilen finansal baskı hipotezini temsil eden Kaldıraç (Leverage) ve Aktif Kârlılığı (ROA) değişkenlerinin SHAP sıralamasında görece sınırlı bir ağırlık taşıması dikkat çekicidir. Bu bulgu, S&P 500 gibi büyük ölçekli ve kurumsallaşmış firmaların greenwashing'i finansal bir son çare olarak değil, stratejik bir konumlandırma aracı olarak kullandığı tezini destekler niteliktedir.



Şekil 2. SHAP özellik önem sıralamasının görsel dağılımı

Table 17. SHAP kategori özeti

Kanal	İlk 5'teki Özellik	İlk 13'teki Özellik	Toplam SHAP Katkısı
GARCH (Piyasa)	3 / 5	8 / 13	Baskın
NLP (Metin)	1 / 5	2 / 13	Tamamlayıcı
Finansal	1 / 5	3 / 13	Tamamlayıcı

Tam modelde ilk beş özelliğin üçü (garch_vol_last, garch_vol_mean, garch_vol_max) GARCH kanalından gelmekte; geriye kalan iki sırayı ise Leverage (finansal kanal) ve gunning_fog (NLP kanalı) paylaşmaktadır. İlk on üç özelliğe genişletildiğinde bu dağılım GARCH lehine daha da belirginleşmekte; sekizi GARCH, üçü finansal ve ikisi NLP kanalından oluşmaktadır. Bu örüntü, piyasa tabanlı volatilité sinyalinin genel açıklayıcılıkta öne çıktığını, finansal ve metinsel kanalların ise sınırlı ancak düzenli bir katkıyla tamamlayıcı rol üstlendiğini göstermektedir. Söz konusu SHAP üstünlüğü, volatilitenin endojen bir belirleyici olabileceği şüphesini doğurabilir; bu endişe, üç robustness testinden elde edilen bulgularla birlikte değerlendirildiğinde kısmen giderilmektedir: (i) yalnızca GARCH özellikleriyle çalışan modelin Average Precision metriğinde %20,7 oranında gerilemesi, salt volatilité sinyalinin tek başına yeterli olmadığına işaret etmektedir; (ii) GARCH özellikleri tamamen çıkarıldığında modelin hâlâ işlevsel kalması, NLP ve finansal kanalların bağımsız bir açıklayıcı güce sahip olduğuyla tutarlıdır; (iii) gecikmeli (%70) volatilité özelliklerinin tam modelin performansını geçmesi, ilişkinin geçici/döngüsel değil, nispeten yapısal bir nitelik taşıdığı yorumunu desteklemektedir.

Bu üç bulgu bir arada, SHAP'te gözlenen GARCH ağırlığının yalnızca modelin veri işleme sırasına veya eşzamanlılığa (contemporaneous ilişkiye) bağlı yapay bir artefakt olmadığını göstermekte, ancak bu üç testin hiçbirisi tek başına kesin bir nedensellik kanıtı sunmamaktadır. Bu bulgular, Krüger (2015)'in önerdiği "ESG endişesi → volatilité" yönündeki teorik mekanizmayla tutarlı bir örüntü sergilemekte; ancak bu çalışmanın tasarımı nedensel bir ilişkiyi test etmeye yönelik olmadığından, gözlenen örtüşme doğrudan bir nedensellik kanıtı olarak değil, teorik beklentiyle uyumlu bir gösterge olarak değerlendirilmelidir.

Table 18. Cross-model SHAP comparison

Model	#1 Özellik	İlk 5' te GARCH
Tam Model	garch_vol_last (0.668)	3 / 5
T1: GARCH Yok	Leverage (0.662)	0 / 5
T2: Sadece GARCH	garch_vol_last (0.798)	5 / 5
T3: Gecikmeli GARCH	garch_rolling_vol_60d (0.579)	4 / 5

Dört model varyantı arasındaki karşılaştırma, modelin herhangi bir kanalın yokluğunda diğer kanallara uyum sağlayabildiğini göstermektedir. GARCH özellikleri tamamen çıkarıldığında (T1), en belirleyici özellik konumuna Leverage (0.662) geçmekte; bu durum, piyasa sinyali mevcut olmadığında modelin finansal yapı ve metin karmaşıklığına dayanarak telafi ettiğini doğrulamaktadır. GARCH özelliklerinin gecikmeli (lagged) kullanıldığı T3 konfigürasyonunda ise garch_rolling_vol_60d birinci sıraya yükselmekte; bu bulgu, önceki bölümde tartışılan 60 günlük rejim ortalamasının anlık volatilité ölçümüne kıyasla daha güvenilir bir sinyal taşıdığı yorumunu desteklemektedir. Bu esneklik, önerilen mimarinin tek bir bilgi kanalına aşırı bağımlı olmadığını dolaylı bir kanıtı olarak değerlendirilebilir.

Temel Tez: Piyasa Şüphesinin Yeşil Aklama Tespitindeki Rolü

SHAP analizinden elde edilen bulgular, çalışmanın temel yorumlama çerçevesiyle tutarlı bir örüntü ortaya koymaktadır: yatırımcıların firmaların ESG beyanlarına duyduğu güven düzeyi, hisse senedi volatilitesi aracılığıyla dolaylı olarak gözlemlenebilir bir sinyale dönüşebilir. Bu çerçeveye göre, bir firma çevresel taahhütlerini kamuya açıkladığında ancak piyasa katılımcıları bu beyanı ikna edici bulmadığında, söz konusu şüphenin fiyatlama davranışına üç şekilde yansması beklenebilir: hisse getirilerinin koşullu varyansında (σ^2_t) bir artış, bu artışın zaman içinde kalıcılık göstermesi ve

periyodik volatilité şoklarının sıklaşması. Bu üçlü örüntü, yüksek volatilité, yüksek kalıcılık ve sık şoklar, SHAP analizinde en güçlü açıklayıcı sinyal olarak öne çıkan `garch_vol_last`, `garch_persistence` ve `garch_spike_count` değişkenleriyle örtüşmektedir.

Bu yorumlama çerçevesinin bir nedensellik testi değil, gözlemlenen SHAP örüntüsüyle tutarlı bir teorik açıklama önerisi olduğu vurgulanmalıdır; volatilité artışının doğrudan yatırımcı şüphesinden kaynaklandığına dair doğrudan bir nedensel test bu çalışmanın kapsamı dışındadır. Bununla birlikte, Test 2'de elde edilen sonuca göre, volatilité sinyali tek başına yeterli değildir; yüksek volatilité birçok alternatif nedenden (sektörel dinamikler, firma büyüklüğü, likidite koşulları) kaynaklanabileceğinden, tek başına kullanıldığında belirgin sayıda yanlış pozitif üretmektedir. Bu nedenle, metin karmaşıklığı (`gunning_fog`) ve temkinli/muğlak dil yoğunluğu (`ESG_Gap`) gibi NLP tabanlı özellikler, çevresel kökenli şüpheli volatilitéyi işletme faaliyetinden kaynaklanan olağan volatiliteden ayırt etmek bakımından tamamlayıcı ve gerekli bir işlev görmektedir.

Tartışma

Bu çalışmanın bulguları, üç ana başlık altında özetlenebilir. İlk olarak, geliştirilen model, 503 S&P 500 firmasını NLP, finansal ve GARCH kanallarından türetilen 21 özellik üzerinden yeşil aklama riskine göre sınıflandırmakta ve eğitim sürecinde hiç görülmemiş 101 firmalık bağımsız test setinde 0,704 ROC-AUC ile 0,756 Spearman sıralama korelasyonuna ulaşmaktadır. Bu düzeydeki bir performans, kısıtlı sayıda açıklayıcı değişkenle çalışan finansal risk sınıflandırma modelleri için bilgi taşıyıcı kabul edilen bir aralıkta yer almakta ve özellikle sıralama tutarlılığının yüksek olması, modelin portföy tarama (screening) amaçlı kullanımlar için pratik değer taşıdığına işaret etmektedir.

İkinci olarak, SHAP analizi, GARCH tabanlı volatilité özelliklerinin — özellikle `garch_vol_last` değişkeninin ($SHAP = 0.668$) — modelin en güçlü açıklayıcı sinyali olduğunu ortaya koymuştur. Bu bulgu, yatırımcıların bir firmanın çevresel beyanlarına duyduğu güvensizliğin, doğrudan gözlemlenebilir olmasa da, hisse senedi getirilerinin koşullu varyansına yansiyebileceği yönündeki teorik önermeyle uyumludur. Ancak bu üstünlüğün tek başına yeterli olmadığı, robustness testleriyle ayrıca sınanmıştır: yalnızca GARCH özellikleriyle çalışan model (Test 2), Average Precision metriğinde %20.7 oranında bir düşüş sergilemiş; bu da piyasa volatilitésinin, çevresel içerikten bağımsız pek çok nedenden (sektörel dinamikler, likidite koşulları vb.) kaynaklanabileceğini ve tek başına kullanıldığında belirgin sayıda yanlış pozitif ürettiğini göstermiştir. Bu noktada NLP kanalının, metin karmaşıklığını ölçen `gunning_fog` ve temkinli dil kullanımını yakalayan `ESG_Gap` değişkenlerinin, çevresel kökenli şüpheli volatilitéyi işletme faaliyetinden kaynaklanan olağan volatiliteden ayırt etmede tamamlayıcı bir işlev üstlendiği görülmektedir.

Üçüncü olarak, GARCH özelliklerinin tamamen çıkarıldığı ablasyon testinde (Test 1) modelin performansının yalnızca sınırlı ölçüde gerilemesi (ROC-AUC: -0.012 puan), kalan NLP ve finansal özelliklerin tek başına da makul bir ayırıştırma gücüne sahip olduğunu göstermektedir. Bu üç bulgu bir arada değerlendirildiğinde, çalışmanın temel tezi doğrulanmaktadır: yeşil aklama riskinin güvenilir biçimde tespit edilebilmesi, ne yalnızca firmanın söylemine (NLP), ne yalnızca finansal yapısına, ne de yalnızca piyasanın tepkisine (GARCH) dayanan tekil bir yaklaşımla değil, bu üç bilgi kanalının birbirini tamamlayıcı biçimde kullanılmasıyla mümkün olmaktadır.

Modelin akademik ve pratik açıdan güvenilir kabul edilebilmesi için, tahmin performansının ötesinde veri bütünlüğü, açıklanabilirlik ve dayanıklılık boyutlarının da değerlendirilmesi gerekmektedir. Bu çalışmada geliştirilen model altı temel açıdan sınanmıştır.

Modelin çok kaynaklı yapısı, NLP, finansal ve GARCH kanallarından elde edilen 21 özelliğin tek bir matriste birleştirilmesiyle sağlanmıştır. Açıklanabilirliği, TreeExplainer aracılığıyla hesaplanan SHAP değerleriyle desteklenmiş; analiz, `garch_vol_last`, `garch_vol_mean` ve `garch_vol_max` değişkenlerinin en yüksek katkıya sahip ilk üç özellik olduğunu, Leverage ve

gunning fog değişkenlerinin ise finansal ve metinsel kanalları ilk beşlikte temsil ettiğini ortaya koymuştur.

Veri sızıntısı riskine karşı iki önlem alınmıştır: bağımlı değişken, model girdilerinden doğrudan türetilmek yerine örneklem dağılımının 75. yüzdelik dilimine dayalı dinamik bir eşiklemeyle oluşturulmuş; GARCH özellikleri ise yalnızca tahmin noktasından önceki tarihsel getirilerle hesaplanmıştır. Bu varsayım, getiri geçmişinin yalnızca ilk %70'i kullanılarak yapılan gecikmeli bir yeniden tahmin testiyle (Test 3) ayrıca sınanmış; modelin performansının azalmak yerine artması (ROC-AUC: 0.704 karşısında 0.653), sinyalin ileriye dönük bilgiyle şişirilmediğine dair olumlu bir gösterge olarak değerlendirilmiştir.

Modelin dayanıklılığı, aynı sabit hiperparametreler altında yürütülen üç ayrı testle (ablasyon, izolasyon, gecikmeli yeniden tahmin) sınanmış; bu testler, hiçbir tekil kanalın performansı tek başına belirlemediğini doğrulamıştır. Aşırı öğrenme riskinin bertaraf edildiği, modelin hiç görmediği 101 firmalık held-out test setindeki performansı ile ortaya konmuştur. Son olarak, modelin piyasa duyarlılığı, GARCH (1,1) tabanlı özelliklerin SHAP sıralamasında baskın konumda yer almasıyla (ilk beş özelliğin üçü, ilk on üç özelliğin sekizi) desteklenmektedir.

Sonuç

Bu çalışma, S&P 500 kapsamındaki 503 firmayı, doğal dil işleme, finansal göstergeler ve GARCH (1,1) tabanlı piyasa volatilitesi olmak üzere üç bağımsız bilgi kanalından türetilen 21 özellik aracılığıyla yeşil aklama riskine göre sınıflandıran bir XGBoost modeli geliştirmiştir. Model, hiç görülmemiş 101 firmalık bağımsız test setinde 0.704 ROC-AUC ve 0.756 Spearman sıralama korelasyonuna ulaşmış; bu performans, üç ayrı robustness testiyle ve gecikmeli GARCH özellikleriyle yürütülen veri sızıntısı karşıtı doğrulamayla desteklenmiştir.

Bulgular, piyasa tabanlı volatilité sinyalinin modelin en güçlü açıklayıcı bileşeni olduğunu, ancak bu sinyalin tek başına yeterli olmadığını göstermektedir; NLP ve finansal kanallar, özellikle karar eşliğindeki hassasiyeti artırarak tamamlayıcı bir rol üstlenmektedir. Bu bulgu, yeşil aklama tespitinde tekil bir veri kaynağına dayanan yaklaşımların yerine, firmanın söylemi, yapısı ve piyasanın bu söyleme verdiği tepkiyi bütünlükte değerlendiren çok kanallı mimarilerin daha güvenilir sonuçlar üretebileceğine işaret etmektedir.

Bu çalışma, yeşil aklama tespiti literatürüne iki temel katkı sunmaktadır. İlk olarak, zaman serisi volatilité dinamiklerinin, firma düzeyinde kesitsel özellikler biçiminde bir ESG güvenilirlik sınıflandırıcısına entegre edilmesi bakımından, bildiğimiz kadarıyla bu yönde bir ilk teşkil etmektedir; bu entegrasyonun teorik temeli, ESG otantikliğine ilişkin heterojen yatırımcı inançlarının koşullu varyans biçiminde piyasaya yansıdığı yönündeki önermelerle (Krüger, 2015; Pastor vd., 2021) örtüşmekte ve bu entegrasyon, çalışmada uygulanan ablasyon testleriyle ampirik olarak da doğrulanmıştır. İkinci olarak, bu testler aracılığıyla üç kanalın (NLP, finansal, piyasa) hiçbirinin tek başına yeterli olmadığı, ancak birlikte sinerjik bir sonuç ürettiği gösterilmiştir; özellikle volatilitenin sıralama performansı (AUC) açısından, NLP'nin ise hassasiyet (precision) açısından belirleyici olduğu bulgusu, çok kaynaklı ESG analizine yönelik özgün bir perspektif sunmaktadır.

Sınırlılıklar ve Gelecek Çalışmalar

Bu çalışma ampirik literatüre katkı sağlamakla birlikte, bazı metodolojik sınırlılıklar barındırmaktadır: İlk olarak, hedef değişkenin temelini oluşturan ISS Governance QualityScore verisi, Berg, Kölbel ve Rigobon'un (2022) belgelediği ESG derecelendirme ayrışması (rating divergence) olgusu ışığında değerlendirilmelidir. Farklı derecelendirme kuruluşlarının (ISS, MSCI, Refinitiv) aynı firma için sistematik olarak farklı risk puanları üretebildiği bilinmektedir; çalışmanın bağımlı değişkeni tek bir veri sağlayıcısına (ISS) dayanması, bu ölçüm hatası riskini taşımaktadır. Bununla bağlantılı olarak, ISS Governance QualityScore, yeşil aklamanın doğrudan bir ölçüsü değil, dolaylı bir göstergesidir (proxy); greenwashing kavramı tanımı gereği yüksek kurumsal risk ile düşük

şeffaflığın (obfuscation) kesişiminde ortaya çıktığından (Lyon & Maxwell, 2011), bu çalışma yönetim risk skorunu bu kesişimin yalnızca bir boyutu olarak ele almakta ve bunu NLP kaynaklı metinsel şeffaflık metrikleriyle (Fog Index, Hedge Words) tamamlamaktadır. Doğrulanmış greenwashing vaka veritabanlarıyla (SEC incelemeleri, iklim davaları) yapılacak gelecek çalışmalar bu kavramsal geçerliliği güçlendirecektir.

İkinci olarak, NLP kanalı, gerçek sürdürülebilirlik raporları (CSR, CDP, GRI, TCFD) yerine yfinance üzerinden erişilen şirket açıklama metinlerini proxy olarak kullanmaktadır. Bu tercih, metin tabanlı sinyallerin bilgi içeriğini sınırlandırmakla birlikte, bulguların metodolojik çerçeve (GARCH entegrasyonu, robustness doğrulaması) açısından geçerliliğini etkilememektedir; gerçek CSR raporlarıyla yürütülecek gelecek çalışmaların NLP kanalının katkısını artırması beklenmektedir.

Üçüncü olarak, GICS sektör hassasiyet ağırlıklarının (sector_sensitivity) belirlenmesinde kullanılan 0-1 arası değerler, karbon ayak izi ve düzenleyici maruziyet gibi somut ölçütlere dayanmakla birlikte, kısmen yargısal (heuristic) bir nitelik taşımaktadır.

Dördüncü olarak, çalışma kesitsel (cross-sectional) bir tasarıma sahiptir ve tek bir zaman dilimine (2023-2024) dayanmaktadır. Panel veri yapısının ve dönemsel olarak güncellenen (rolling) GARCH pencerelerinin kullanılması, modelin farklı piyasa rejimlerindeki performansının değerlendirilmesine imkân tanıyacaktır. Bununla bağlantılı olarak, örneklemin yalnızca S&P 500 firmalarıyla sınırlı olması, bulguların büyük ölçekli ve kurumsallaşmış halka açık firmalar bağlamında geçerli olduğunu; daha küçük ölçekli, halka açık olmayan veya farklı düzenleyici çerçevelere tabi (örn. gelişmekte olan piyasa) firmalara genellenmesinin ihtiyatla değerlendirilmesi gerektiğini göstermektedir. Bu firma gruplarında yapılacak replikasyon çalışmaları, modelin genellenebilirliğini test edecektir.

Piyasa volatilitesi ile yönetim riski eşzamanlı kesitte ölçüldüğü için, değişkenler arası endojenite (garch_vol_last) tam anlamıyla dışlanamaz. Uygulanan gecikmeli (lagged) özellik testleri öngörüsül veri sızıntısını engellese de, gelecekteki çalışmaların Granger-Nedensellik (Granger-causality) veya Araç Değişken (Instrumental Variable - IV) modelleri kullanarak bu geri-besleme döngüsünü daha kesin izole etmesi gerekmektedir.

Son olarak, bu çalışma öngörücü (predictive) bir modelleme çalışması olup nedensel çıkarım (causal inference) iddiası taşımamaktadır. GARCH volatilitesi ile ESG riski arasında gözlenen ilişki korelasyonel olarak değerlendirilmeli; bu ilişkinin nedensel niteliği, gelecekte araç değişken (instrumental variable) veya doğal deney tasarımlarına dayalı çalışmalarla ayrıca sınanmalıdır. Ayrıca Hedef değişken olarak yalnızca ISS Governance QualityScore kullanılmıştır. Literatürde kabul gören "derecelendirme ayrışması" olgusu gereği (Berg vd., 2022), gelecek araştırmaların model duyarlılığını Sustainability, MSCI veya Refinitiv gibi alternatif sağlayıcılarla çapraz doğrulama (cross-provider validation) testlerine tabi tutması önerilmektedir. Bir diğer husus ise ölçeklenebilirlik adına şirket özet metinleri proxy olarak kullanılmıştır. İleriki çalışmalarda gerçek CSR, CDP veya TCFD raporlarının tam metin analizi ile bir pilot uygulama yapılması, modelin mekanizmasını daha da somutlaştıracaktır.

References

- Atabay, E. (2026). Yeşil söylem mi, yeşil gerçeklik mi? BIST sürdürülebilirlik endeksi'nde listelenen şirketler arasında yeşil aklama (greenwashing) olasılıklarının ampirik bir değerlendirmesi. *Mali Çözüm*, 36(194), 489–520.
- Başar, Ü., & Doğruel, M. (2025). Metinsel veri madenciliği ile Türkiye'deki üniversitelerin sürdürülebilirlik raporlarının değerlendirilmesi. *Beykent Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, 18(2), 75–95. <https://doi.org/10.20854/bujse.1842519>
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315–1344. <https://doi.org/10.1093/rof/rfac033>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Çelik, N., & Reis, Ş. G. (2026). Sürdürülebilirlik raporlaması, yeşil yıkama ve bağımsız denetim: Bibliyometrik bir inceleme. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi*, 25(2), 560–573. <https://doi.org/10.21547/jss.1882859>
- Chen, P., & Dagestani, A. A. (2023). Greenwashing behavior and firm value: From the perspective of board characteristics. *Corporate Social Responsibility and Environmental Management*, 30(5), 2330–2343. <https://doi.org/10.1002/csr.2488>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Cocozza, R., Curcio, D., Gallo, S., & Vioto, D. (2026). Unveiling the dark side of sustainability: Are banks' ESG misrepresentations truly worthwhile? *Journal of Financial Stability*, 85, Article 101554. <https://doi.org/10.1016/j.jfs.2026.101554>
- Dağıstanlı, H. A., & Çelik, İ. (2023). Sürdürülebilirlik raporlaması ve firma performansı: BIST sürdürülebilirlik endeksi üzerine bir uygulama. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 76, 1–16. <https://doi.org/10.51290/dpusbe.1153330>
- Dorfleitner, G., & Utz, S. (2024). Green, green, it's green they say: A conceptual framework for measuring greenwashing on firm level. *Review of Managerial Science*, 18, 3463–3486. <https://doi.org/10.1007/s11846-023-00718-w>
- Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, 1(2), 237–245. <https://doi.org/10.1088/1469-7688/1/2/305>
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233. <https://doi.org/10.1037/h0057532>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Gillan, S. L., Koch, A., & Starks, L. T. (2021). Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal of Corporate Finance*, 66, Article 101889. <https://doi.org/10.1016/j.jcorpfin.2021.101889>
- Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). <https://doi.org/10.48550/arXiv.1706.04599>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of GARCH estimates and variance recovery. *Applied Financial Economics*, 16(2), 149–158. <https://doi.org/10.1080/13518470500039436>
- Institutional Shareholder Services. (2023). *Governance QualityScore: Methodology guide*. ISS ESG.
- Karagül, A. A., & Selimoğlu, S. K. (2026). Sürdürülebilirlik raporlarında yeşil aklama uygulamaları ve finansal yansımaları: Uluslararası örnek olay analizi. *Muhasebe ve Finansman Dergisi*, 111, 59–90.
- Kathan, M. C., Utz, S., Dorfleitner, G., Eckberg, J., & Chmel, L. (2025). What you see is not what you get: ESG scores and greenwashing risk. *Finance Research Letters*, 74, Article 106710. <https://doi.org/10.1016/j.frl.2024.106710>
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15, 1–21. <https://doi.org/10.1145/2382577.2382579>
- Kılınç, Y., İnce, M. R., & Badem, A. C. (2026). Sürdürülebilirlik raporlamasında yeşil yıkamayı tespit etmek için çok boyutlu metinsel bir çerçeve. *Discover Sustainability*, 7, Article 482. <https://doi.org/10.1007/s43621-026-02890-x>
- Kim, C. G., & Na, H. J. (2026). Exploring the impact of ESG ratings on corporate carbon emissions in Korean firms: Evidence from machine learning and deep learning models. *Sustainability*, 18(9), Article 4553. <https://doi.org/10.3390/su18094553>
- Krüger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, 115(2), 304–329. <https://doi.org/10.1016/j.jfineco.2014.09.008>
- Kuşku, Ö. E., Topsümer, F., & Göztaş, A. (2017). Yeşil aklama kampanyaları ve medya gündemine yansımaları: Volkswagen emisyon krizinin Amerikan basınında sunumu. *Uluslararası Bilimsel Araştırmalar Dergisi (IBAD)*, 2(2), 357–370. <https://doi.org/10.21733/ibad.2141>
- Laufer, W. S. (2003). Social accountability and corporate greenwashing. *Journal of Business Ethics*, 43(3), 253–261. <https://doi.org/10.1023/A:1022962719299>
- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3), 221–247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Li, R., Zhong, M., & Shahab, Y. (2023). Can penalties for environmental violations deter firms from engaging in greenwashing? *Business Ethics, the Environment & Responsibility*, 34, 189–214. <https://doi.org/10.1111/beer.12623>
- Liman, A. (2026). İklimle ilgili açıklamaların Türkiye sürdürülebilirlik raporlama standartları çerçevesinde analizi: BIST sürdürülebilirlik 25 endeksi işletmeleri örneği. *Muhasebe ve Vergi Çalışmaları Dergisi*, 19, 29–54. <https://doi.org/10.29067/muvu.1650671>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. <https://doi.org/10.1002/9781119482260>

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. <https://doi.org/10.48550/arXiv.1907.11692>
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30, 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
- Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. *Journal of Economics & Management Strategy*, 20(1), 3–41. <https://doi.org/10.1111/j.1530-9134.2010.00282.x>
- Marquis, C., Toffel, M. W., & Zhou, Y. (2016). Scrutiny, norms, and selective disclosure: A global study of greenwashing. *Organization Science*, 27(2), 483–504. <https://doi.org/10.1287/orsc.2015.1039>
- Meral, R. (2025). Gıda biliminde yapay zekâ uygulamaları: Makine öğrenmesi tabanlı yaklaşımlar ve güncel eğilimler. *Bartın University International Journal of Natural and Applied Sciences*, 8(1), 196–216.
- Özgür, S. B. (2021). Algoritmalar, yapay zekâ, makine öğrenmesi, derin öğrenme ve uygulamaları: Beşerî fayda üretiminin yazılımlar tarafından karşılanması. *Ekonomi ve Yönetim Araştırmaları Dergisi*, 10(1), 1–28.
- Özkaya, B. (2026). İklim değişikliği temalı reklamların yeşil aklama bağlamında incelenmesi. *Abant Sosyal Bilimler Dergisi*, 26(1), 543–568. <https://doi.org/10.11616/asbi.1824181>
- Pástor, L., Stambaugh, R. F., & Taylor, L. A. (2021). Sustainable investing in equilibrium. *Journal of Financial Economics*, 142(2), 550–571. <https://doi.org/10.1016/j.jfineco.2020.12.011>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Schaltegger, S., & Burritt, R. L. (2010). Sustainability accounting for companies: Catchphrase or decision support for business leaders? *Journal of World Business*, 45(4), 375–384. <https://doi.org/10.1016/j.jwb.2009.08.002>
- Seele, P., & Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies. *Business Strategy and the Environment*, 26, 239–252. <https://doi.org/10.1002/bse.1912>
- Shakourloo, A., & Azimli, A. (2025). The role of greenwashing scandals on investors' herding behavior. *Journal of Economics and Finance*, 49, 733–767. <https://doi.org/10.1007/s12197-025-09725-x>
- Siddique, M. A., Karim, S., Haque, M. R., & Mia, P. (2026). The truth behind sustainability claims: Examining carbon risk, ESG disclosures and greenwashing. *International Review of Financial Analysis*, 109, Article 104735. <https://doi.org/10.1016/j.irfa.2025.104735>
- Stărică, C., & Granger, C. W. J. (2005). Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87(3), 503–522. <https://doi.org/10.1162/0034653054638305>

-
- Suta, A., Limbach, Z., & Tóth, Á. (2026). Varieties of capitalism and sustainability reporting: Text analysis-based evidence on European companies. *Sustainable Futures*, 11, Article 101865. <https://doi.org/10.1016/j.sfr.2026.101865>
- Tepeli, Y., & Büklü, E. (2025). Çevre dostu açıklamalar, yeşil pazarlama ve piyasa gerçekliği: Yeşil aklamanın finansal etkileri. *Uluslararası Yönetim Akademisi Dergisi*, 8(4), 1102–1114. <https://doi.org/10.33712/mana.1831443>
- Thewissen, J., Shrestha, P., Arslan-Ayaydin, Ö., & Beibei, Y. (2026). Topics that matter: ESG reports and information asymmetry. *Journal of Financial Research*, 1-54. <https://doi.org/10.1111/jfir.70053>
- Walker, K., & Wan, F. (2012). The harm of symbolic actions and green-washing: Corporate actions and communications on environmental performance and their financial implications. *Journal of Business Ethics*, 109(2), 227–242. <https://doi.org/10.1007/s10551-011-1122-4>
- Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). ClimateBERT: A pretrained language model for climate-related text. In *Proceedings of the AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges* (Arlington, VA). <https://doi.org/10.48550/arXiv.2110.12010>
- Yu, E. P., Luu, B. V., & Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance*, 52, Article 101192. <https://doi.org/10.1016/j.ribaf.2020.101192>

ETHICAL DECLARATIONS

Research and Publication Ethics

The author(s) declare that this manuscript is an original work prepared in accordance with the principles of scientific research and publication ethics. The manuscript has not been previously published and is not currently under consideration for publication elsewhere.

All sources used in the preparation of this study have been properly cited and referenced. The author(s) confirm that the work contains no plagiarism, fabrication, falsification, or manipulation of data. The research findings are presented honestly and transparently in accordance with the principles of academic integrity.

Author Contributions

All authors have significantly contributed to the conception, design, analysis, and preparation of the manuscript. All authors have read and approved the final version of the manuscript and agree to its submission and publication.

Conflict of Interest

The author(s) declare that there is no conflict of interest regarding the publication of this manuscript.

Funding

This research did not receive any specific grant, financial support, or funding from any public, commercial, or non-profit funding agencies.

Artificial Intelligence (AI) Usage Disclosure

During the preparation of this manuscript, artificial intelligence tools may have been used solely for limited purposes such as language editing, translation assistance, and stylistic improvement. Any AI-assisted translations or language corrections were carefully reviewed and verified by the author(s). Artificial intelligence tools were not used to generate the core scientific content, research design, analysis, or scholarly conclusions. All academic responsibility for the content of the manuscript rests entirely with the author(s).

Copyright and Permissions

The author(s) confirm that all necessary permissions for copyrighted materials have been obtained where required and that the manuscript does not violate any intellectual property rights.

Author Approval

All authors confirm that they have read and approved the final version of the manuscript and consent to its submission and publication.

ETİK BEYANLAR

Araştırma ve Yayın Etiği

Yazar(lar), bu makalenin özgün bir çalışma olduğunu ve bilimsel araştırma ve yayın etiği ilkelerine uygun olarak hazırlandığını beyan eder. Makale daha önce herhangi bir yerde yayımlanmamış olup başka bir dergide değerlendirme sürecinde değildir.

Çalışmada kullanılan tüm kaynaklar akademik kurallara uygun biçimde atıflanmıştır. Yazar(lar), çalışmanın intihal, veri uydurma, veri tahrifi veya manipülasyon içermediğini; araştırma bulgularının bilimsel dürüstlük ilkelerine uygun şekilde sunulduğunu beyan eder.

Yazar Katkısı

Tüm yazarlar çalışmanın tasarlanması, yürütülmesi, analizi ve makalenin hazırlanmasına önemli ölçüde katkı sağlamıştır. Tüm yazarlar makalenin son hâlini okuyup onaylamış ve yayımlanmasını kabul etmiştir.

Çıkar Çatışması

Yazar(lar), bu makalenin yayımlanmasına ilişkin herhangi bir çıkar çatışması bulunmadığını beyan eder.

Finansal Destek

Bu araştırma herhangi bir kamu kurumu, ticari kuruluş veya kâr amacı gütmeyen kuruluş tarafından finansal olarak desteklenmemiştir.

Yapay Zekâ Kullanım Beyanı

Makalenin hazırlanması sürecinde yapay zekâ araçlarından yalnızca dil düzenleme, çeviri desteği ve anlatım iyileştirmesi amacıyla sınırlı ölçüde yararlanılmış olabilir. Yapay zekâ araçları tarafından yapılan çeviri ve dil düzenlemeleri yazar(lar) tarafından dikkatle kontrol edilmiştir. Yapay zekâ araçları çalışmanın bilimsel içeriğinin, araştırma tasarımının, analizlerinin veya akademik sonuçlarının oluşturulmasında kullanılmamıştır. Makalenin içeriğine ilişkin tüm akademik sorumluluk yazar(lar)a aittir.

Telif Hakları ve İzinler

Yazar(lar), telif hakkı gerektiren materyaller kullanılmışsa gerekli izinlerin alındığını ve çalışmanın herhangi bir fikrî mülkiyet hakkını ihlal etmediğini beyan eder.

Yazar Onayı

Tüm yazarlar makalenin son hâlini okuduklarını, onayladıklarını ve yayımlanmasını kabul ettiklerini beyan eder.