Understanding Corporate Sustainability using Machine Learning

EBB3 – Team 2 | Erasmus School of Economics

Vera Gak Anagrova

3 January, 2026

Contents

1 Introduction

This paper examines companies' different sustainability characteristics using hierarchical clustering and predicts ESG (Environmental, Social, and Governance) scores using three different models: logistic regression, decision trees, and random forests. In recent years, ESG scores have become influential in investing as these scores show a company's dedication to ethical and sustainable practices (Friede, Busch, and Bassen (2015)). For this reason, understanding the relationship between a company's environmental practices and its ESG score is important for investors and regulators, as responsible business is often considered smart business.

ESG scores exist to give a simple overview of a business's sustainability performance, but they may hide some variations in environmental performance (Berg, Kolbel, and Rigobon (2020)). Grouping firms based on environmental variables like energy, water, and carbon emissions allows for the identification of different sustainability profiles (Delmas, Etzion, and Nairn-Birch (2013)). This is achieved by applying multiple models and finding patterns in the data to interpret the distinct ESG scores.

The key research questions that this paper seeks to answer are whether environmental indicators can be used to classify companies into different sustainability profiles. Additionally, it investigates whether these profiles, along with other financial metrics in the dataset, can effectively predict whether a company has a high or low ESG score. By answering these questions, the study aims to provide insights into the relationship between environmental performance and ESG ratings, which will give regulators and investors a greater knowledge, helping them make better decisions.

To address these questions, the analysis combines both unsupervised and supervised machine learning approaches. First, firms with similar environmental characteristics are grouped using hierarchical clustering. Next, environmental and financial factors are used in classification models to predict ESG performance. Lastly, we interpret the final model to obtain a better perspective of what factors influence the different ESG predictions.

2 Data

The analysis uses a synthetic ESG and financial dataset that includes 1000 organizations across different industries and geographies from 2015 to 2025. Although the data is simulated, it was created to reflect realistic patterns commonly observed, making it relevant to this paper and very suitable for the analysis without compromising confidentiality.

Each firm is observed annually, keeping track of their ESG scores, environmental indicators (carbon emissions, water usage, energy consumption), and financial metrics (revenue, profit margin, growth rate, market capitalization). Overall ESG performance is captured by an aggregate ESG score ranging from 0 to 100, which is further divided into environmental, social, and governance individual scores.

The environmental indicators are important in this study because these variables are linked to businesses' operational activities rather than subjective analysis. Clustering algorithms are used to identify sustainability profiles based on these characteristics. Then, the supervised learning models combine financial and categorical firm variables, such as industry, region, and year, to predict different ESG performance levels.

3 Methodology

Before starting with the models, it is very important to prepare the data. The data was cleaned by removing the identifier variables (CompanyID and CompanyName) and converting categorical variables (industry, region, and year) into factors. Also, the variable GrowthRate had 1,000 missing values, which we filled in using the median value to keep the full sample size. Furthermore, data visualization showed that several numeric variables were highly skewed to the right. To address this, we log-transformed MarketCap, CarbonEmissions, WaterUsage, and EnergyConsumption to reduce scale dominance and improve numerical stability in the models. Revenue was kept in its original scale because it showed only moderate skewness. Additionally, we created a binary target variable called ESG_Class to classify companies as having either a "High" or "Low" ESG score based on whether the total ESG score was above or below the median. We scaled the environmental variables only for the cluster analysis to ensure equal contribution in the distance calculations. For the supervised learning models, variables were kept on their original or log-transformed scales, since logistic regression and tree-based models are not sensitive to variable

scaling. Finally, we split the dataset into training (70%) and test (30%) sets in order to estimate and assess the models.

To identify distinct sustainability profiles based on environmental performance, hierarchical clustering was performed using Ward's and Euclidean distance method on the scaled environmental indicators to produce compact and interpretable clusters. Euclidean distance was chosen as it works well with continuous variables, and Ward's method minimizes within-cluster variance, leading to more uniform groups. The dendrogram showed three different groups, which were labeled as "Sustainable," "Moderate impact," and "High impact" based on their average environmental profiles. The "Sustainable" cluster exhibited the lowest average carbon emissions, water usage, and energy consumption, while the "High impact" cluster showed the highest levels across these indicators.

Logistic regression was first used as a baseline classifier to model the probability that a company achieves a high ESG score. This offered a basic measure by which more complex machine learning models could be evaluated. A decision tree classifier was then estimated to capture nonlinear relationships between predictors. Decision trees offer a transparent structure, illustrating how different variables contribute to ESG classification through a sequence of decision rules (Breiman et al. (1984)). Tree complexity was controlled using cost complexity pruning, ensuring an interpretable decision tree and to prevent overfitting.

A random forest classifier was used to improve predictive performance even more. By aggregating predictions across many trees, the model reduces variance and improves generalization compared to a single decision tree (Breiman (2001)). This characteristic makes random forests appropriate for ESG classification, where nonlinear effects and interactions between financial and environmental variables are expected. To assess robustness, both an untuned and a lightly tuned version of the random forest were taken into consideration. Tuning focused on key hyperparameters that control model complexity and randomness, specially the number of predictors considered at each split. This parameter, which is essential to the bias–variance trade-off in random forests, allows the model to balance the strength of individual trees with the diversity of the ensemble. The number of trees was set sufficiently high to ensure stable predictions. Since random forests are typically robust and the goal was to evaluate model stability rather than enhance performance through aggressive tuning, extensive hyperparameter modification was avoided.

Model performance was evaluated using accuracy, sensitivity, specificity, balanced accuracy, and Cohen's kappa. Balanced accuracy was shown to account for potential class imbalance between high and low ESG firms. Confusion matrices were also used with heatmaps to examine classification errors in more detail.

Global model interpretation was conducted using variable importance measures from the random forest model. In addition, partial dependence plots were used to examine the effect of key continuous predictors on the probability of high ESG performance. For categorical predictors, group level predicted probabilities were compared to identify differences across regions. To complement the global analysis, local interpretation was performed using LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, and Guestrin (2016)). Two representative observations, one classified as high ESG and one as low ESG, were selected to illustrate how individual predictions are influenced by a small subset of characteristics. This provides insight into how global patterns can be translated into more specific company-level decisions.

4 Analysis & Results

4.1 Environmental Sustainability Clusters

The hierarchical clustering results indicate that there are three distinct sustainability profiles. The first cluster is formed by companies with lower carbon emissions, less water usage, and less energy consumption, being a more environmentally sustainable group. The second cluster shows moderate levels across all indicators, indicating a higher environmental impact than Cluster 1. Lastly, the third cluster displays substantially higher environmental impact, characterized by companies that have a negative impact on the environment.

Figure 1 displays the different clusters specified above, with the log-transformed water usage and carbon emissions. The graph has many points, each represents an individual company, and the different colors are associated with different clusters. One can almost see a line with all the firms close to each other, showing a positive relationship between the water usage and carbon emissions. Companies are separated into three distinct groups differentiated by different colors. These represent low, moderate, and high environmental impact, indicating that the clustering captures important variation in environmental performance rather than random noise.

Figure 2 illustrates the average environmental profiles of the three clusters. The sustainable cluster is associated with the lowest water usage and carbon emissions, while the high-impact cluster shows markedly higher values on both measures. The moderate-impact cluster is located in the middle of the graph. Overall, this graph provides insight into the average profile of each cluster, summarizing their environmental performance in a simple illustration.

Both figures, presented below, show that different companies can actually be grouped into different sustainability profiles based on different characteristics, providing a useful starting point for further ESG prediction research.

Figure 1: Environmental Sustainability Clusters

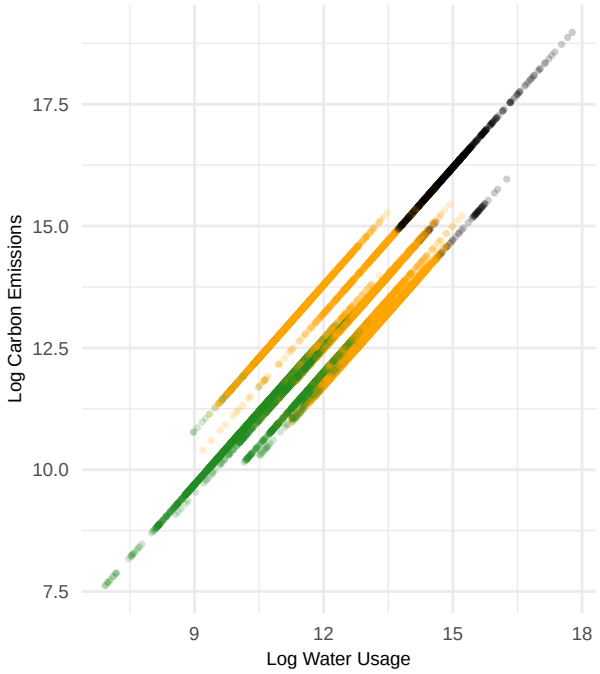


Figure 2: Average Environmental Clusters

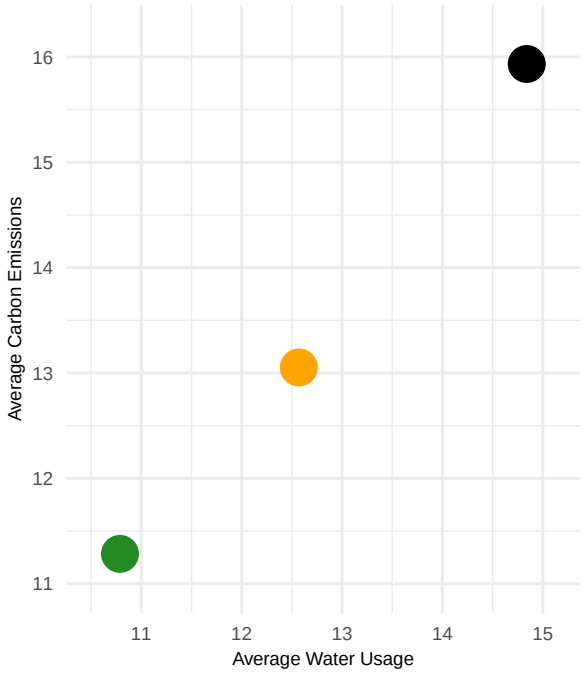


Table 1 complements Figures 1 and 2 by summarizing the average environmental variables and environmental ESG scores in numbers. Related to the patterns observed in Figure 2, Cluster 1 shows the lowest levels of carbon emissions (113,515), water usage (77,850), and energy consumption (284,848), and consequently has the highest environmental ESG score (81.6). Cluster 2 is formed by companies with a moderate environmental footprint, exhibiting higher resource use than Cluster 1 and as expected, a lower ESG score (45.0). By contrast, Cluster 3 displays extremely high environmental impact, with average carbon emissions over 11.6 million, water usage above 3.7 million, and energy consumption surpassing 118 million, getting the lowest environmental ESG score (34.1).

Table 1: Environmental Sustainability Cluster Profiles

Cluster	Carbon Emissions	Water Usage	Energy Consumption	Environmental ESG Score
1	113514.9	77849.7	284848.4	81.6
2	676703.5	446509.2	5279489.3	45.0
3	11644617.5	3761455.8	118022515.7	34.1

4.2 ESG classification models

The analysis started with a simpler model, such as the logistic regression, to use as a baseline. This way, we can better evaluate how well a simple linear decision boundary can predict ESG classifications. As shown in Table 2, the model performs very poorly, with both test accuracy and balanced accuracy around 0.27, which is below what would

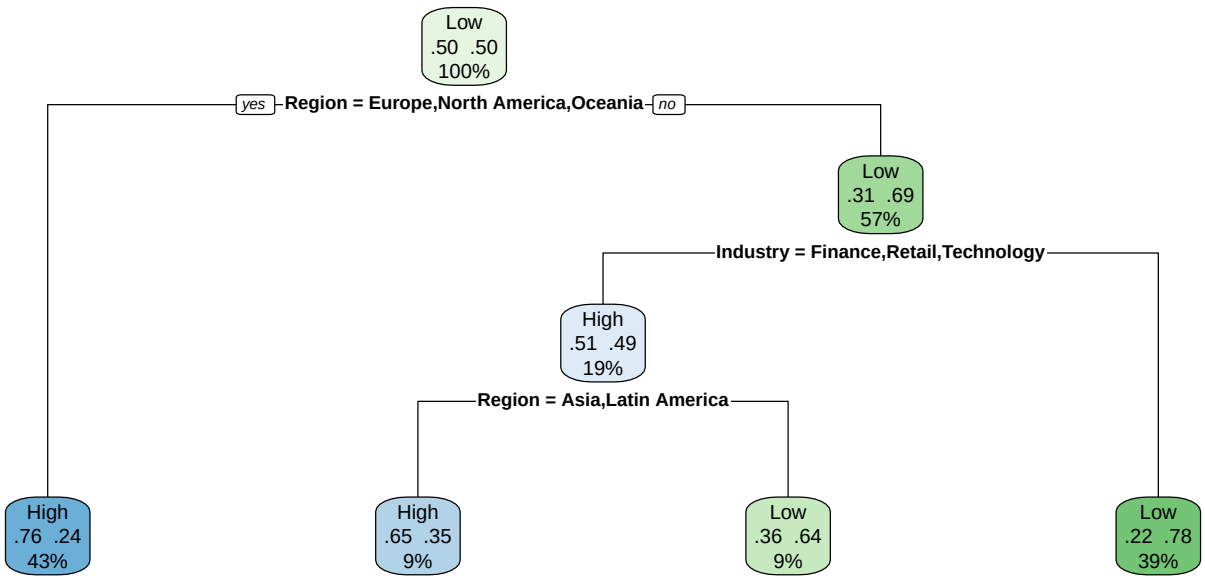
be expected from random guessing. The other numbers support that the model's predictions show little agreement with the observed ESG classes beyond random. In addition, the low sensitivity (0.258) and specificity (0.272) reveal that the model struggles to correctly identify any ESG score. This weak predictive ability can also be seen by its confusion matrix, which contains a large number of misclassified observations. Overall, these results emphasize the limitations of a linear model and motivate the use of more flexible, non-linear models.

Table 2: Performance Metrics for Logistic Regression

Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
0.25	0.265	-0.471	0.258	0.272	0.265

To move beyond the limitations of a linear model, a decision tree classifier was modeled to capture potential non-linear relationships and interaction effects between the variables. Figure 3 shows the pruned decision tree used to classify firms into high and low ESG categories. The first and most influential split is based on the region or geographic location. Firms that are located in Europe, North America, and Oceania have a more balanced mix of high and low ESG classifications (43% vs. 57%), while firms found in other regions are usually classified with a low ESG score. Also, in the other regions, industry association also separates companies into different ESG outcomes, pointing out the relation between regional and sector-specific characteristics. Several terminal nodes exhibit elevated class probabilities, thereby indicating the existence of clearly separated sustainability profiles. In sum, the decision tree provides an accessible perspective on the collective influence of firm characteristics on ESG classification, serving as a valuable intermediary between the simple logistic regression model and the more adaptable random forest methodology.

Figure 3: Decision Tree for ESG Classification



The decision tree classifier had better results compared to the logistic regression model, achieving a test accuracy of roughly 73% and a similar balanced accuracy, as detailed in Table 3. Sensitivity (0.762) and specificity (0.694) imply that the model has a good capacity to accurately classify both high and low ESG firms. This balanced performance profile suggests that the observed results are not influenced by class imbalance.

The confusion matrix supports this enhancement, revealing a greater percentage of accurately classified observations for each ESG category compared to the initial model. Although the decision tree does not achieve the same level of predictive accuracy as the random forest, it provides an intuitive structure that makes it easier to understand how different variables contribute to ESG classification.

Table 3: Performance Metrics for Descision Tree

Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
0.746	0.728	0.456	0.762	0.694	0.728

A random forest classifier was implemented to further improve predictive performance of the decision tree through variance reduction and the capture of more complex patterns. While the decision tree is easy to interpret, its performance can vary substantially depending on the specific sample used to build the tree. By averaging predictions across many trees trained on different bootstrap samples and subsets of predictors, the random forest produces more stable and generalizable results.

Table 4 demonstrates that the random forest model achieved the highest performance relative to all other models, with a test and a balanced accuracy of 84%. In contrast to the decision tree, the random forest significantly enhanced both overall accuracy and Cohen's Kappa, thereby indicating the model's capacity to outperform random classification. Despite attaining perfect accuracy on the training data, the test-set performance remained consistent, implying minimal overfitting.

Table 4: Performance Metrics for Random Forest

Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
1	0.84	0.679	0.824	0.855	0.84

A lightly tuned version of the random forest gave similar results (Table 5), with only minimal differences in test accuracy, balanced accuracy, and Cohen's Kappa. Figures 4 and 5 further confirm this similarity: the confusion matrices show almost the same distribution of correct classifications and errors across both classes, indicating that tuning does not meaningfully change the model's error profile. As the tuned version gave slightly better results, it was then retained as the final model for further interpretation. Overall, the random forest provides a robust and well-performing classifier, offering a strong balance between predictive accuracy and generalization among the considered approaches.

Table 5: Performance Metrics for Tuned Random Forest

Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
1	0.844	0.687	0.827	0.86	0.844

Figure 4: Confusion Matrix − Random Forest

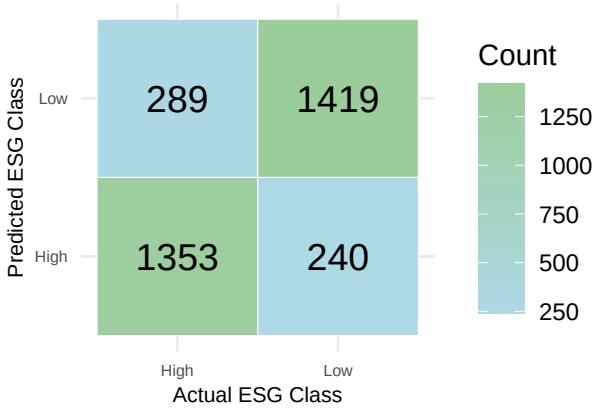
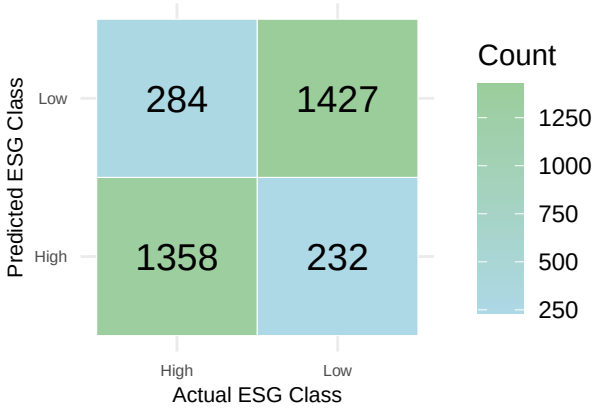


Figure 5: Confusion Matrix − Random Forest (Tuned)



4.3 Model performance comparison

Table 6 provides a nice overview of all supervised learning models considered in this study. To summarize, the logistic regression performs poorly across all metrics, indicating that linear decision boundaries are insufficient to capture the complexity of ESG classification. The decision tree represents a substantial improvement, offering both higher predictive accuracy and interpretable decision rules, particularly with respect to regional and industry-level characteristics.

The random forest achieves the strongest overall performance, consistently outperforming the decision tree in terms of accuracy, balanced accuracy, and agreement between predicted and observed ESG classes. While tuning the random forest leads to only marginal performance differences, the results remain stable across specifications. These findings point to a clear trade-off between interpretability and predictive accuracy, leading to the selection of the random forest as the primary model for subsequent global and local interpretation.

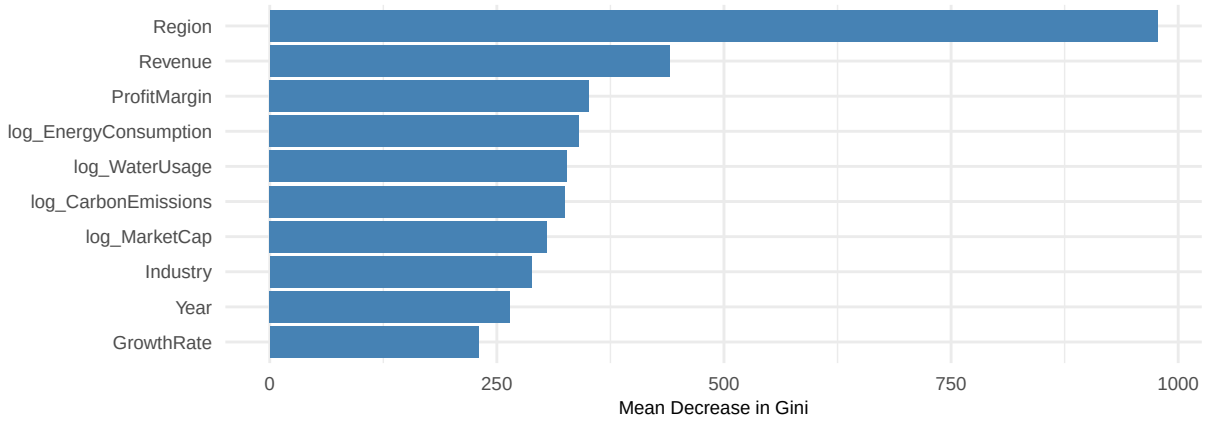
Table 6: Performance Metrics Comparison

	Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
Logistic_Regression	0.250	0.265	-0.471	0.258	0.272	0.265
Decision_Tree	0.746	0.728	0.456	0.762	0.694	0.728
Random_Forest	1.000	0.840	0.679	0.824	0.855	0.840
Random_Forest_Tuned	1.000	0.844	0.687	0.827	0.860	0.844

4.4 Model interpretation

Below, Figure 6 reports the top ten predictors in the random forest model. Region is the most important variable by a big margin, indicating that ESG classification is strongly influenced by the geographic location. Revenue and profitability are also very influential, followed by environmental measures such as energy, water, and carbon emissions. Overall, the variable importance results suggest that ESG outcomes reflect a combination of contextual factors and firm-level financial and environmental characteristics, rather than being influenced by a single variable.

Figure 6: Top 10 Variable Importance − Random Forest



Following the variable importance results, Figures 7 and 8 provide further insight into how the most influential predictors shape ESG classification. Figure 7 shows the partial dependence of revenue on the probability of high ESG performance. The relationship is positive but clearly nonlinear: increases in revenue are associated with substantial gains in predicted ESG probability at lower revenue levels, while the effect gradually levels off for larger firms. This pattern suggests diminishing marginal effects of firm size, indicating that beyond a certain scale, additional revenue contributes relatively little to improvements in ESG classification.

Figure 8 illustrates average predicted probabilities of high ESG classification across regions. Regional differences are noticable, with firms located in Europe, Oceania, and North America exhibiting substantially higher predicted ESG probabilities than in the other locations. Together, these figures show that ESG classification is influenced by both firm size and more general institutional and geographic factors, supporting the earlier clustering and decision tree

results. These differences continue even after adjusting for firm-level financial and environmental characteristics, illustrating the significance of regional context in ESG outcomes.

Figure 7: Partial Dependence of Revenue

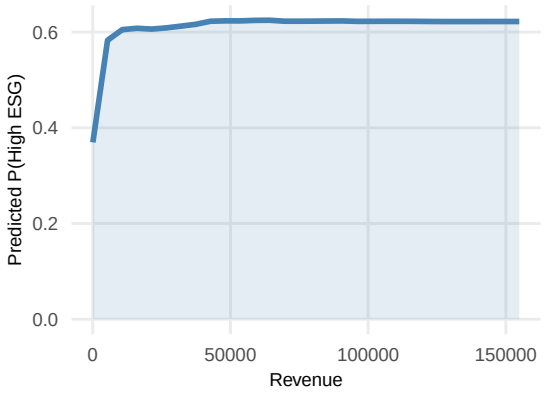
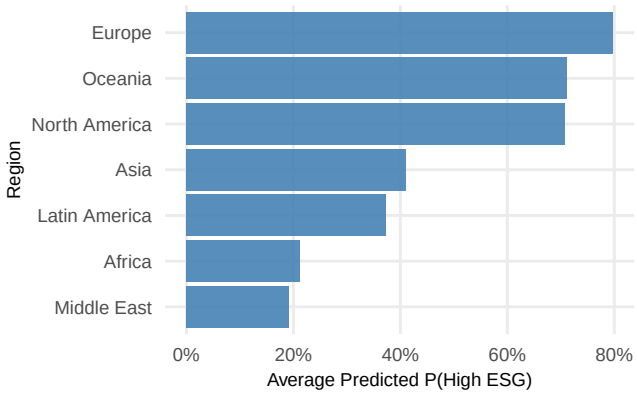


Figure 8: Average Predicted by Region



To complement the global interpretation, local explanations were generated using LIME for two representative test observations: one classified as High ESG with high confidence and one classified as Low ESG.

Figure 9 illustrates the local explanation for an observation predicted as High ESG with a probability of 0.99. The prediction is driven primarily by contextual and environmental factors. Location in North America is the strongest positive contributor, followed by relatively low carbon emissions, energy consumption, and water usage, all of which increase the predicted probability of high ESG performance. Industry affiliation in technology and a higher market capitalization further reinforce the classification. Importantly, all influential features act in the same direction, and no strong contradicting factors are present, indicating a coherent and internally consistent prediction.

Figure 9: Local explanation − High ESG example

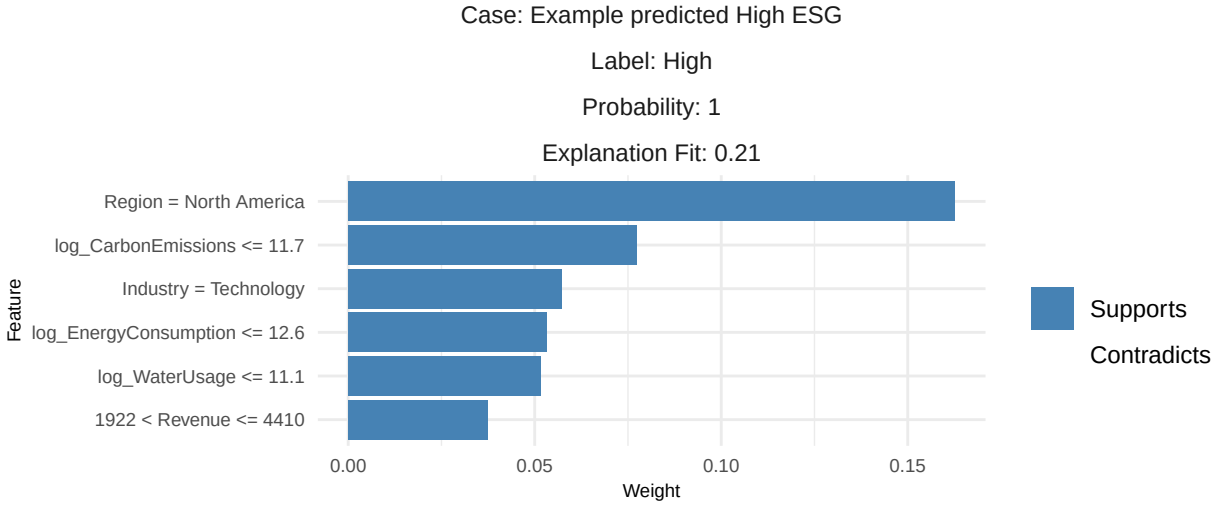
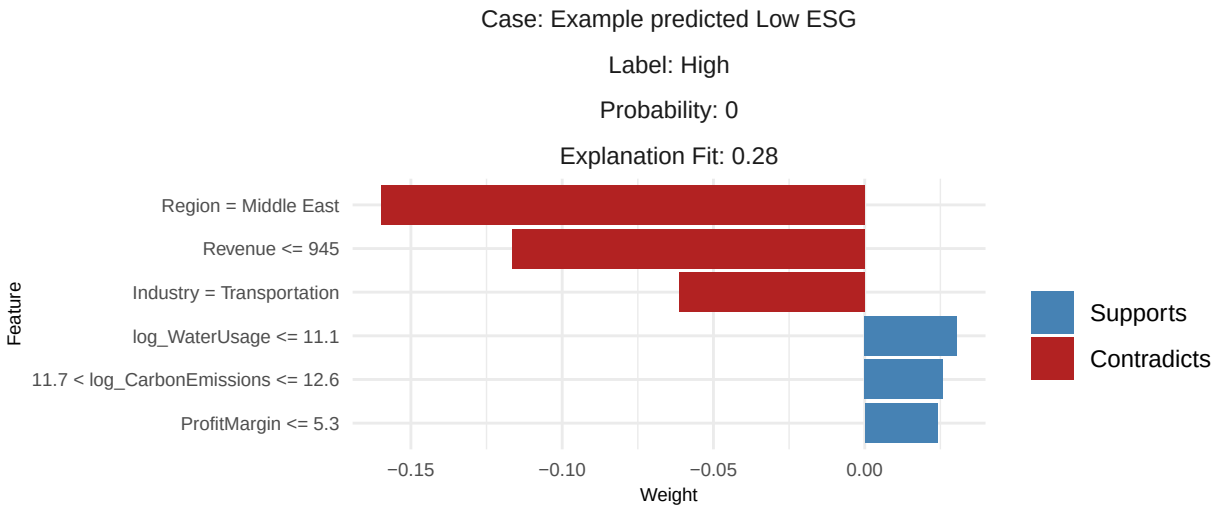


Figure 10 presents the local explanation for an observation predicted as Low ESG. In this case, the prediction is dominated by several strong negative contributors. Firms located in the Middle East and those with lower revenue show the strongest negative impact on the probability of being classified as high ESG, followed by industry affiliation in transportation. While some environmental indicators contribute weakly in favor of a higher ESG outcome, their effect is dominated by contextual and financial factors that lead to a Low ESG prediction. This comparison shows how the random forest combines environmental information with regional and industry characteristics for individual companies.

Figure 10: Local explanation − Low ESG example



Overall, the local explanations confirm that the model's predictions are not driven only by a single variable, but by the interaction of environmental performance, firm characteristics, and geographic context. These particular insights align closely with the global importance and partial dependence results, strengthening confidence in the interpretability and credibility of the model's decision process.

5 Conclusion

In conclusion, the objective of this paper was to determine whether environmental and financial variables and can be used to classify companies into different sustainability profiles. The analysis actually provides evidence after running several models (hierarchical clustering, logistic regression, decision trees, and random forests) that environmental and contextual factors play a central role in explaining ESG performance.

After running the hierarchical clustering, we observe that there are 3 different clusters characterized by low, moderate, and high environmental impact. Companies with lower emissions and lower resource usage consistently achieved higher environmental ESG scores. These findings suggest that ESG scores reflect underlying environmental behavior rather than random variation.

After estimating the supervised learning models it is visible that the logistic regression performs poorly, as it cannot capture the complex relationships in the data. Decision trees improve predictive performance while offering interpretable decision rules, whereas the random forest model achieves the highest predictive accuracy and balanced performance. At the end, global and local interpretation support that firm size and regional context systematically influence ESG predictions, and that these relationships can be translated to company-level classification decisions.

Also we have to take into consideration that there are some limitations. One of them is that the dataset used is simulated and not based on real companies information, which limits how applicable the results are to real world ESG evaluations. Another limitation is that the binary classification of ESG performance simplifies a very complex concept in which other factors influence this, hiding important variations within ESG categories. Finally, the analysis focuses on predictive associations and does not establish causal relationships between environmental performance and ESG outcomes, meaning that the analysis cannot determine whether environmental performance directly causes higher ESG scores or whether both are driven by other factors.

Some future research that could complement and expand this study to make it better could be to actually work on real ESG datasets that uses other different variables that could influence the different performance scores. Also, it would be interesting to explore causal inference approaches to see how firm behavior actually influences ESG ratings. However, this study is still relevant and the findings show us that machine learning method can be a very useful tool to identify sustainability profiles and finding different patterns to predict ESG performance.

References

Altmeyer, P., C. C. S. Liem, and A. van Deursen. 2023. "Explaining Black-Box Models Through Counterfactuals." In *The Proceedings of the JuliaCon Conferences (JCON)*. https://resolver.tudelft.nl/uuid:446dc879-2782-4f89-9e25-120e912448ae. https://doi.org/10.21105/jcon.00130.

Berg, Florian, Julian F. Kolbel, and Roberto Rigobon. 2020. "Aggregate Confusion: The Divergence of ESG Ratings." *Review of Finance* 26 (6): 1315–44. https://academic.oup.com/rof/article/26/6/1315/6590670.

Boehmke, Bradley, and Brandon M. Greenwell. 2025. "Chapter 11: Random Forests." https://bradleyboehmke.github.io/HOML/random-forest.html.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth International Group.

DataCamp. 2018. "Hierarchical Clustering in r: Dendrograms with `hclust`." DataCamp; https://www.datacamp.com/tutorial/hierarchical-clustering-R.

Delmas, Magali A., Dror Etzion, and Nicholas Nairn-Birch. 2013. "Triangulating Environmental Performance: What Do Corporate Social Responsibility Ratings Really Capture?" *Academy of Management Perspectives* 27 (3): 255–67. https://doi.org/10.5465/amp.2012.0123.

Ernst & Young. 2025. "Why ESG Performance Is Growing in Importance for Investors." EY Insights; https://www.ey.com/en_fi/insights/assurance/why-esg-performance-is-growing-in-importance-for-investors.

Friede, Gunnar, Timo Busch, and Alexander Bassen. 2015. "ESG and Financial Performance: Aggregated Evidence from More Than 2000 Empirical Studies." *Journal of Sustainable Finance & Investment* 5 (4): 210–33. https://www.tandfonline.com/doi/full/10.1080/20430795.2015.1118917.

GeeksforGeeks. 2025a. "Hierarchical Clustering in r Programming." GeeksforGeeks; https://www.geeksforgeeks.org/r-machine-learning/hierarchical-clustering-in-r-programming/.

———. 2025b. "Random Forest Approach in r Programming." https://www.geeksforgeeks.org/r-language/random-forest-approach-in-r-programming/.

Investopedia. 2025. "What Is a Black Box Model? Definition, Uses, and Examples." https://www.investopedia.com/terms/b/blackbox.asp.

Iris Ccarbon. 2024. "Why ESG Reporting Is More Important Than Ever: Key Benefits for Companies." Iris Carbon; https://iriscarbon.com/why-esg-reporting-is-more-important-than-ever-key-benefits-for-companies/.

Jagtap, Shriyash. 2023. "ESG and Financial Performance Dataset." https://www.kaggle.com/datasets/shriyashjagtap/esg-and-financial-performance-dataset; Kaggle.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. New York: Springer.

Keita, Zoumana. 2023. "Explainable AI: Understanding and Trusting Machine Learning Models." DataCamp; https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models.

Molnar, Christoph. 2025a. "LIME – Local Interpretable Model-Agnostic Explanations." https://christophm.github.io/interpretable-ml-book/lime.html. https://christophm.github.io/interpretable-ml-book/lime.html.

———. 2025b. "Methods Overview — Interpretable Machine Learning." https://christophm.github.io/interpretable-ml-book/overview.html. https://christophm.github.io/interpretable-ml-book/overview.html.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should i Trust You?: Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. https://doi.org/10.1145/2939672.2939778.