POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Machine Learning for ESG Score Prediction: A GRI-Based Approach

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Federico Ballo, Thomas Mandato

Advisor: Prof. Daniele Marazzina

Academic year: 2024-2025

---

1. Introduction

In recent years, Environmental, Social, and Governance (ESG) factors have gained significant importance in evaluating corporate performance. Companies are increasingly required to integrate ESG criteria into their strategic decisions to ensure regulatory compliance, enhance reputation, and attract sustainability-focused investments. However, assessing ESG performance remains a complex challenge due to data heterogeneity and the absence of a standardized methodology for quantifying sustainability impacts.

This study explores the potential of Machine Learning techniques to predict ESG scores using Global Reporting Initiative (GRI) indicators. By leveraging Ridge Regression and Random Forest models, the research aims to determine the most influential GRI factors in ESG assessment and evaluate the predictive capability of these approaches. The findings contribute to bridging the gap between ESG reporting and data-driven sustainability evaluation, offering valuable insights for corporations, investors, and policymakers.

2. The GRI Index and Its Relationship with ESG

2.1. Brief History and Structure of GRI

The Global Reporting Initiative (GRI) is a key framework for sustainability reporting, providing structured guidelines for organizations to disclose their Environmental, Social, and Governance (ESG) performance.

Developed in the late 1990s, GRI was created to address the growing need for standardized corporate sustainability reporting. Over time, it evolved through multiple iterations, introducing the concept of materiality in G4, which emphasized reporting on the most relevant sustainability issues. Today, thousands of organizations worldwide use GRI standards to enhance corporate transparency and accountability [4].

The GRI Standards provide a structured framework for organizations to report on their environmental, social, and governance (ESG) impacts. They consist of Universal Standards, which apply to all organizations and ensure consistency in disclosures, Topic-Specific Standards, which focus on economic, environmental, and social aspects, and Sector Standards, which address industry-specific sustainability challenges. Each standard is identified by a unique numeric code, such as GRI 305 for greenhouse gas emissions and GRI 403 for occupational health and safety.

By adopting GRI Standards, companies enhance transparency and accountability, ensuring that sustainability reports meet the expectations of investors, regulators, and stakeholders. This harmonized approach not only supports regulatory compliance but also integrates sustainability into long-term corporate strategy, reinforcing a company's commitment to responsible business practices [10].

2.2. Relationship Between GRI and ESG

The GRI Index serves as a standardized framework for measuring corporate sustainability performance across the Environmental, Social, and Governance (ESG) dimensions. It enables organizations to report on key sustainability metrics, facilitating transparency and comparability.
From an environmental perspective, GRI reporting includes indicators such as greenhouse gas emissions, energy and water consumption, and biodiversity impact, helping companies assess and mitigate their ecological footprint [2]. The social dimension covers human rights, labor conditions, workplace diversity, and community engagement, providing insights into corporate ethical practices and employee well-being [8]. The governance aspect focuses on corporate integrity, transparency, and ethical business conduct, including board diversity, anti-corruption policies, and risk management [9].
GRI data is widely used to inform ESG ratings, influencing investor decisions and market perception. The structured nature of GRI reporting also facilitates advanced analytics, allowing organizations to link individual sustainability indicators to overall ESG scores. This enhances strategic decision-making by identifying key drivers of corporate sustainability performance [7].
However, integrating GRI with ESG frameworks presents challenges. While GRI emphasizes stakeholder concerns, ESG frameworks are often investor-driven, leading to differences in priorities [3]. Additionally, data collection complexity and the lack of global standardization make cross-industry comparisons difficult.
Despite these challenges, GRI reporting enhances transparency, accountability, and access to sustainable financing. It also enables companies to set measurable sustainability goals and leverage machine learning techniques to refine ESG strategies, driving long-term corporate sustainability improvements.

3. Data Collection and Exploratory Analysis

3.1. Collection Data Structure

This study investigates the relationship between ESG scores and corporate sustainability disclosures by leveraging the GRI Standards across three key categories: GRI 200 (Economic), GRI 300 (Environmental), and GRI 400 (Social). The dataset consists of 344 companies from the financial and manufacturing sectors in the United States and Europe, with ESG scores sourced from LSEG. Primary data sources include sustainability reports, annual reports, and proxy statements, focusing on companies that publish a GRI Content Index to ensure structured and comparable disclosures.
However, data collection posed significant challenges due to incompleteness and inconsistency in ESG reporting. Many companies provided qualitative rather than numerical data, making it difficult to quantify sustainability metrics. Additionally, reporting completeness varied across regions and industries, impacting data reliability and comparability.

3.2. Exploratory Analysis

A comparative analysis revealed that European companies exhibited higher reporting transparency, largely influenced by stricter regulations such as the Corporate Sustainability Reporting Directive (CSRD). In contrast, U.S. companies displayed more fragmented ESG disclosures, as reporting remains largely voluntary. Similarly, the manufacturing sector demonstrated more extensive sustainability reporting than the financial sector, primarily due to stricter environmental monitoring requirements related to carbon emissions, energy consumption, and resource management.
To enhance data usability, the dataset was segmented into six groups, categorized by ESG dimension (Environmental, Social, Governance), sector, and geography. Key frequently reported GRI indicators included energy consumption (GRI 302-1), CO_2 emissions (GRI 305), water usage (GRI 303), and workforce diversity (GRI

405). Conversely, significant reporting gaps were identified for external energy consumption (GRI 302-2), waste prevention (GRI 306-2), and workforce-related disclosures. These inconsistencies highlight the ongoing challenges in data standardization and emphasize the need for stronger regulatory alignment and transparency in ESG reporting.

This exploratory analysis lays the groundwork for correlation studies and machine learning applications, ensuring that the dataset is sufficiently structured to investigate the predictive power of GRI disclosures on ESG scores.

4. Dataset Completion and Correlation Analysis

4.1. Data Cleaning and Handling Missing Values

A structured and complete dataset is essential for reliable ESG analysis, particularly for correlation studies and machine learning applications. To enhance data quality, the dataset was segmented into Environmental, Social, and Governance (ESG) dimensions across financial and manufacturing sectors. Indicators with insufficient data availability below 15% for Environmental, 20% for Social, and 30% for Governance were excluded to ensure statistical robustness.

To address missing values, a penalty-based imputation method was applied, replacing gaps with the worst observed value for each indicator. This approach prevented incomplete disclosures from artificially improving ESG scores. Positive indicators, such as renewable energy use, were assigned the lowest observed value, while negative indicators, like emissions, were replaced with the highest observed value. This ensured data integrity and aligned with the assumption that non-disclosure may reflect weaker sustainability performance.

4.2. Correlation Analysis

A Pearson correlation analysis examined the relationship between GRI disclosures and ESG scores, using the formula:

$$\rho = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}, \quad (1)$$

where X_i and Y_i represent the individual values of the two variables being compared in this case, the GRI data and the ESG scores. The terms $\bar{X}$ and $\bar{Y}$ correspond to the mean values of each respective variable. The results showed strong correlations between higher disclosures in energy consumption reduction (GRI 302-4), CO_2 emissions management(GRI 305), and disposal and treatment of waste(GRI 306) with increased ESG scores. Social factors such as workforce diversity(GRI 405-1) and economic value distribution(GRI 201-1) correlated positively with ESG performance, while workplace injury rates(GRI 403-9) had a negative association. Governance indicators related to board composition and transparency(GRI 405-1) also showed expected positive correlations, with an interesting trend where firms with older board members scored higher in governance, possibly due to regulatory stability.

The correlation results validated the penalty-based imputation method, as expected trends remained consistent. The findings highlight the relevance of structured ESG disclosures and reinforce the role of GRI indicators in ESG score prediction. However, inconsistencies in voluntary reporting remain a challenge. This analysis provides a foundation for predictive modeling, demonstrating how structured sustainability data can enhance ESG assessment methodologies.

5. Machine Learning Methods for ESG Prediction

This study applies Machine Learning techniques to predict ESG scores, leveraging structured datasets that account for sectoral and regional variations. The objective is to construct a robust predictive framework that not only estimates ESG scores but also provides insights into the relative importance of sustainability indicators. To achieve this, two complementary Machine Learning methods are employed: Ridge Regression and Random Forest.

5.1. Ridge Regression

Ridge Regression is particularly suited for ESG data due to its ability to mitigate multicollinearity among sustainability indicators while preserving interpretability. It extends the Ordinary Least Squares (OLS) regression by incorporating an L2 penalty, which regularizes the regression coefficients to prevent overfitting [6]. The Ridge

estimator is defined as:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 + x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2}$$

where y_i represents the ESG score, x_i the vector of GRI indicators, β the regression coefficients, and λ a regularization parameter that controls the degree of shrinkage. The closed-form solution is given by:

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y, \tag{3}$$

where I is the p x p identity matrix, ensuring that $X^T X + \lambda I$ remains non-singular even in cases of highly correlated features. The optimal λ is determined using k-fold cross-validation, balancing bias and variance to enhance predictive accuracy [5].

5.2. Random Forest

To capture non-linear relationships within ESG data, Random Forest is employed as a flexible, ensemble-based approach that constructs multiple decision trees. By aggregating their outputs, it reduces variance and enhances generalization [1]. The final ESG score prediction is computed as the average output across all trees:

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} T_b(X), \tag{4}$$

where B represents the number of trees and $T_b(X)$ denotes the output of the bth tree. Unlike Ridge Regression, which assumes linearity, Random Forest can identify intricate interactions between sustainability indicators. Hyperparameter tuning is performed through Randomized Search Cross-Validation, optimizing key parameters such as tree depth, number of trees, and feature selection criteria. This allows the model to adapt to sector-specific and regional variations in ESG reporting.

6. Performance Evaluation of ESG Prediction Models

The effectiveness of Ridge Regression and Random Forest in predicting ESG scores was assessed using R-squared (R^2) and Mean Absolute Error (MAE) across six data blocks. While both models faced challenges, Random Forest generally outperformed Ridge Regression, particularly in predicting Social (S) and Environmental (E) scores within the financial sector.

6.1. Ridge Regression Results

Ridge Regression exhibited low predictive power, with R^2 values below 0.3 across all blocks. The highest performance was in the E-Manufacturing sector, where it achieved an R^2 of 0.2567 and an MAE of 13.4540. Governance blocks showed the weakest results.

Feature importance analysis highlights sector-specific drivers of ESG performance. For Environmental scores, Ridge Regression identified GRI 305-5a (reduction of greenhouse gas emissions) and 301-1 (material use) as key predictors in the financial sector, whereas in manufacturing, GRI 305-3a (other indirect GHG emissions) and 302-4a (energy consumption reduction) were most relevant. Social scores were primarily influenced by GRI 201-1 (economic value generation) across sectors. In the financial industry, GRI 418-1 (customer data losses) was a major factor, while in manufacturing, workplace safety (GRI 403-9a.iii) played a central role. Governance scores showed regional variations, with GRI 405-1br (diversity and age structure of the board at the end of the reporting period) being the most important factor in Europe, while in the US, legal actions related to anti-competitive behavior (GRI 206-1) had the strongest influence.

6.2. Random Forest Results

Random Forest demonstrated better predictive performance, particularly in Social-Financial (R^2 = 0.4489, MAE = 10.0159). The weakest performance was in Governance-Europe (R^2 = -0.3096, MAE = 18.8886), highlighting the difficulty of predicting governance scores.

For Random Forest, environmental performance in the financial sector was mainly driven by GRI 305-1a (direct GHG emissions), 305-3a (other indirect GHG emissions), and 302-1e (total energy consumption), while in manufacturing, GRI 305-3a (other indirect GHG emissions), 306-3a (waste generated), and 302-4a (energy consumption reduction) played a more significant role, emphasizing emissions control and resource management. In the social domain, fi-

nancial sector predictions were most influenced by GRI 201-1 (economic value generation), GRI 404-1a.ii.ii (management training hours), and GRI 405-1br.ix (employees aged 30-50), whereas in manufacturing, workplace safety (GRI 403-9a.iii) and employee age distribution (GRI 405-1br.x) were key factors. Governance in Europe was mainly shaped by board composition, with GRI 405-1ar.i (male board members) and 405-1ar.ii (female board members) at the beginning of the reporting period being critical, while in the US, compliance-related metrics such as GRI 206-1 (legal actions for anti-competitive behavior) and board demographics (GRI 405-1ar.vi and GRI 405-1ar.vii) were more significant, reflecting differences in governance priorities between regions.

6.3. Model Comparison

Comparing the two models, Random Forest outperformed Ridge Regression in most cases, particularly for Social and Environmental dimensions. Ridge Regression offered greater interpretability but struggled with complex interactions, whereas Random Forest captured non-linear patterns more effectively. Both models were affected by missing data and limited GRI indicators, emphasizing the need for improved data quality and feature selection.
The following table summarizes the R^2 and MAE values for both models across the six blocks.

	Ridge Regression		Random Forest	
	R^2	MAE	R^2	MAE
E_Financial	0.1722	17.2975	0.3546	13.6552
E_Manufacturing	0.2567	13.4540	0.1564	12.6451
S_Financial	0.2445	11.8830	0.4489	10.0159
S_Manufacturing	0.0969	14.9822	0.2905	11.4725
G_EU	0.0286	17.9949	-0.3096	18.8886
G_USA	0.1054	15.0550	0.1852	16.7240

Table 1: Comparison of Ridge Regression and Random Forest

7. Conclusion

This study explored the connection between GRI sustainability reporting and ESG scores, investigating whether machine learning models could effectively predict ESG ratings using GRI indicators. However, the results revealed significant challenges, including data limitations, inconsistencies between reporting frameworks, and methodological constraints. The relationship between GRI disclosures and ESG scores was not as strong as expected, particularly in governance, where key qualitative factors are not fully addressed by GRI standards.
A major hurdle was the quality and completeness of the dataset, as GRI reporting is voluntary and lacks full alignment with ESG scoring criteria. While some environmental indicators, such as CO_2 emissions and energy consumption, showed moderate correlations, overall predictive accuracy remained low. The effectiveness of predictions also varied by sector and region: manufacturing companies had more predictive environmental data than financial firms, while European businesses demonstrated greater consistency in ESG reporting compared to their U.S. counterparts, largely due to stricter regulations. To improve future research, better alignment between GRI and ESG frameworks, expanded datasets, and the use of advanced analytical tools like NLP and deep learning could enhance predictive capabilities. Although ESG score predictions based on GRI remain complex, the findings underscore the importance of standardized and transparent sustainability disclosures. As ESG considerations become more central to business strategy, data-driven approaches will play a crucial role in improving corporate sustainability assessments and fostering a more accountable and resilient economy.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Carbon Disclosure Project (CDP). Cdp climate change report, 2021. Retrieved from https://www.cdp.net.

[3] R. G. Eccles and M. P. Krzus. *The Integrated Reporting Movement: Meaning, Momentum, Motives, and Materiality.* John Wiley & Sons, 2018.

[4] Global Reporting Initiative (GRI). Gri standards, 2021. Retrieved from https://www.globalreporting.org.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Sta-*

tistical Learning: Data Mining, Inference, and Prediction. Springer Science Business Media, 2009.

[6] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[7] KPMG. The future of esg reporting: Global trends and best practices, 2020.

[8] Sustainability Accounting Standards Board (SASB). Sasb standards overview, 2021. Retrieved from https://www.sasb.org.

[9] Task Force on Climate-related Financial Disclosures (TCFD). Final report: Recommendations of the task force on climate-related financial disclosures, 2017. Retrieved from https://www.fsb-tcfd.org.

[10] World Business Council for Sustainable Development (WBCSD). Aligning esg and corporate reporting: A practical guide, 2021. Retrieved from https://www.wbcsd.org.