CSC 461 - ESG & Financial Performance: Predicting Sustainable Companies and Analyzing the Link Between ESG and Financial Success

Peter Card and Emma Choukroun

Date: November, 25th 2025

Abstract

This final report summarizes our work on the project "ESG & Financial Performance". The growing integration of sustainability criteria into corporate and investment decision-making has made Environmental, Social, and Governance (ESG) indicators a central measure of long-term business performance. In this project, we analyze the relationship between ESG scores and financial performance using a dataset of 1,000 companies from 2015 to 2025, covering 16 key variables, including profitability, leverage, and valuation ratios. Our goal is twofold: first, to explore whether financial metrics can help predict a company's ESG compliance level, and second, to examine the extent to which ESG factors correlate with financial health.

As future computer scientists entering fields increasingly influenced by finance, data, and sustainability, this topic resonates strongly with our career paths. Understanding how machine learning can uncover ESG–finance interactions allows us to contribute meaningfully to data-driven decision-making in sustainable finance. Our analysis combines both supervised and unsupervised learning approaches, aiming to model ESG behavior and interpret its underlying structure. We will also present the roadmap of next steps.

Project Repository

All notebooks (`Clustering_ESG.ipynb`, `Classification_ESG.ipynb`, `Regression_ESG.ipynb`) as well as the full dataset and project materials are available on our GitHub repository:

https://github.com/emmackn/CSC461_ESG_analysis

I. Problem Definition

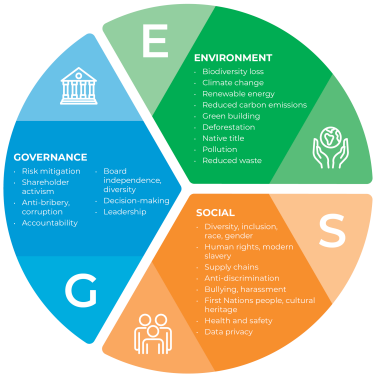

Figure 1: Illustration of ESG components (Environmental, Social, and Governance).

The objective of this project is to determine whether machine learning models can effectively characterize and predict corporate ESG performance. More specifically, we address two core problems:

(1) Classification Problem — Predicting ESG Compliance. Can we reliably predict whether a company maintains a **strong, moderate, or weak ESG profile (ESG_TARGET)** using its ESG sub-scores and financial indicators?

(2) Regression Problem — Explaining Continuous ESG Variation. To what extent can continuous ESG outcomes (*e.g., ESG_Overall*) be predicted from financial and environmental indicators?

(3) Clustering Problem — Discovering ESG–Financial Structure. Do companies naturally form distinct groups based on ESG characteristics and financial performance, and how do these clusters relate to industry and regional patterns?

Inputs. The models take as input:

- ESG component scores (Environmental, Social, Governance),
- Financial metrics (MarketCap, Revenue, ProfitMargin, GrowthRate),
- Environmental indicators (CarbonEmissions, WaterUsage, EnergyConsumption),
- Categorical attributes (Industry, Region).

Outputs.

- A predicted ESG category (0 = Weak, 1 = Moderate, 2 = Strong) for classification,
- A predicted continuous ESG score for regression,
- Cluster assignments identifying groups of similar companies for unsupervised analysis.

This problem is particularly challenging and relevant because ESG thresholds and definitions vary across:

- **Data providers** (MSCI, Refinitiv, Sustainalytics),

- **Industry sectors** (with different structural sustainability risks),
- **Corporate reporting practices**.

By exploring both predictive relationships and structural patterns, we aim to uncover whether strong ESG performance is primarily a consequence of financial robustness, or whether ESG excellence contributes to better financial outcomes. This dual perspective provides deeper insight into how sustainability and financial performance interact in real-world corporate environments.

II. Dataset Description

II-A Overview

The study relies on the **ESG & Financial Performance Dataset**, a balanced panel of corporate annual reports from 2015 to 2025. The dataset captures both sustainability metrics and financial performance indicators, enabling a holistic view of how environmental, social, and governance dimensions interact with firm profitability and growth.

It includes records from nearly 1,000 unique It includes records from nearly 1,000 unique companies operating across nine major industries (see Figure 13) and seven global regions (see Figure 14), providing a diverse foundation for generalizable insights. The dataset is suitable for both **supervised learning** (classification and regression tasks) and **unsupervised analysis** (clustering, correlation structure).

- **Rows:** 11,000 (annual observations)
- **Companies:** 1,000 unique entities
- **Period:** 2015–2025 (annual)
- **Columns:** 16 variables capturing ESG sub-scores, financial indicators, and environmental metrics

Each observation corresponds to a single company-year pair, which allows us to capture temporal patterns such as improvement or decline in ESG performance over time.

II-B Key Variables

Representative columns and their roles are summarized below (full list in Appendix A). To ensure readability within a two-column layout, the table width has been adjusted to fit one column with automatic line wrapping.

Column	Description	Type
CompanyID	Unique identifier for each firm	Numeric
CompanyName	Synthetic company label (e.g., Company_1)	Categorical
Industry	Sector (Finance, Energy, Retail, etc.)	Categorical
Region	Geographic region of operation	Categorical
Year	Reporting year (2015–2025)	Numeric
Revenue	Annual revenue (in million USD)	Numeric
ProfitMargin	Net profit margin (%)	Numeric
MarketCap	Market capitalization (in million USD)	Numeric
GrowthRate	Year-over-year revenue growth (%)	Numeric
ESG_Overall	Aggregate ESG score (0–100)	Numeric
ESG_Env	Environmental performance score (0–100)	Numeric
ESG_Social	Social responsibility score (0–100)	Numeric
ESG_Governance	Governance quality score (0–100)	Numeric
CarbonEmissions	Annual CO_2 emissions (tons)	Numeric
WaterUsage	Annual water consumption (m^3)	Numeric
EnergyCons	Total energy consumption (MWh)	Numeric

Table I: Complete list of dataset columns used for analysis.

II-C ESG Ratings and Interpretation

ESG scores range from 0 to 100 and can be mapped to conventional rating bands (AAA to CCC). High scores (>70) correspond to "Leaders" in sustainability, while low scores (<40) indicate "Laggards." This quantification facilitates downstream machine learning tasks by converting qualitative ESG ratings into measurable targets.

II-D Data Relevance

The dataset's design—balanced across sectors, regions, and time—makes it ideal for:

- Evaluating whether financial metrics can predict ESG status,
- Exploring correlations between profitability and ESG strength,
- Testing clustering methods for ESG pattern discovery.

Overall, this dataset provides both **breadth** (multiple industries and regions) and **depth** (longitudinal financial-ESG detail), which are essential for reliable, data-driven sustainability modeling.

II-E Motivation for dataset choice

The dataset spans multiple industries (9 sectors) and regions (7 regions), includes temporal depth enabling time-aware validation and trend analysis, and pairs ESG sub-scores with standard financial variables—making it suitable for both supervised and unsupervised analyses.

III. Exploratory Data Analysis (EDA)

Before conducting any statistical or machine learning analysis, we first performed a comprehensive data cleaning and preprocessing phase, following standard practices highlighted

in previous ESG and financial data analysis works. Prior literature consistently emphasizes the importance of handling missing values, validating temporal indicators, and ensuring numerical consistency before executing an EDA pipeline. In our dataset, for example, all entries from the year 2015 lacked a `GrowthRate` value because year-over-year growth cannot be computed for the first available year. Following common approaches in related studies, we imputed these missing values with zero:

```
df['GrowthRate']=df['GrowthRate'].fillna(0)
```

This ensures coherent temporal interpretation without introducing artificial bias. Similar verification steps were applied across the dataset to identify missing values, correct inconsistencies, and standardize column formats. Only after this cleaning phase did we proceed with a structured EDA pipeline, including summary statistics, distribution analysis, temporal trends, and pairwise correlations. This methodology aligns with previous related works and ensures that all subsequent analyses are based on a clean and reliable data foundation.

III-A Missing values and descriptive statistics

- Total rows: 11,000. The `GrowthRate` field is missing for 1,000 entries (all 2015 rows) because year-over-year growth cannot be computed for the first year.
- Summary statistics (selected): median `ESG_Overall` $\approx$ 54.6, mean $\approx$ 54.6, standard deviation $\approx$ 15.9. See Appendix B for full tables and numeric summaries.

III-B Distributions and temporal dynamics

- ESG component distributions: Environmental scores display more polarization; Social and Governance scores are comparatively more uniformly distributed.
- Time trends: average `ESG_Overall` exhibits a slight upward trend from 2015 to 2025.

III-C Correlation analysis

Pearson correlations with `ESG_Overall` (selected):

- ESG_Governance: **0.671**
- ESG_Social: **0.662**
- ESG_Environmental: **0.568**
- Revenue: 0.149
- MarketCap: 0.144
- ProfitMargin: 0.088

A correlation heatmap was produced to inform feature selection (Figure 2).

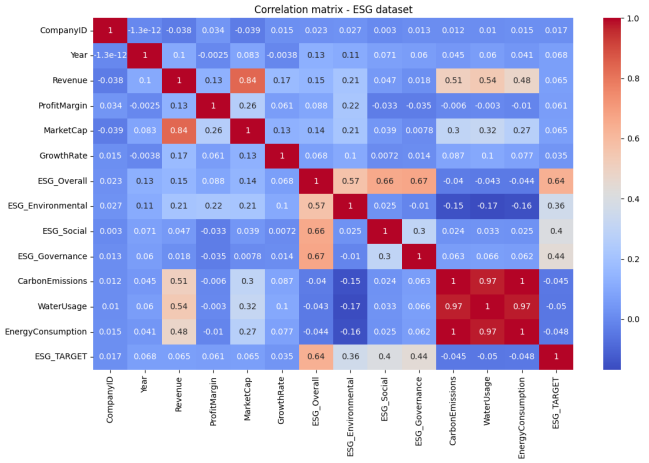

Figure 2: Correlation matrix (numeric features).

IV. Feature engineering: constructing `ESG_TARGET`

We design a discrete target label, `ESG_TARGET`, to categorize firms annually into three ESG tiers. The mapping is:

- **2 (Strong)**: all ESG sub-scores and overall score ≥ 75.
- **1 (Moderate)**: all ESG sub-scores and overall score ≥ 50 and at least one score < 75.
- **0 (Weak)**: any ESG sub-score < 50.

This thresholding produces an ordinal label useful for classification while preserving interpretability.

IV-A Implementation (pandas example)

The following snippet illustrates the label creation used in the analysis pipeline:

```
# Example pandas logic to create ESG_TARGET
high_esg = (df['ESG_Environmental'] >= 75) & \
         (df['ESG_Social'] >= 75) & \
         (df['ESG_Governance'] >= 75) & \
         (df['ESG_Overall'] >= 75)

moderate_esg = (df['ESG_Environmental'] >= 50) & \
            (df['ESG_Social'] >= 50) & \
            (df['ESG_Governance'] >= 50) & \
            (df['ESG_Overall'] >= 50) & \
            (~high_esg)

low_esg = (df['ESG_Environmental'] < 50) | \
        (df['ESG_Social'] < 50) | \
        (df['ESG_Governance'] < 50)

df.loc[high_esg, 'ESG_TARGET'] = 2
df.loc[moderate_esg, 'ESG_TARGET'] = 1
df.loc[low_esg, 'ESG_TARGET'] = 0
```

IV-B Resulting class distribution

Applying this mapping yields:

- **High (2):** 318 entries
- **Moderate (1):** 1,883 entries
- **Low (0):** 8,799 entries

The dataset is highly imbalanced toward the Low (0) class; we plan to address this with class weighting, resampling (SMOTE), and by reporting precision-recall per class (PR-AUC).

V. Machine Learning Methods

We employed a combination of unsupervised and supervised machine learning methods to analyze corporate ESG and financial data.

Unsupervised Tasks

1) **Clustering:** Group companies based on similarity in ESG and financial features to identify distinct patterns or cohorts.
2) **Ranking:** Order companies according to their `ESG_Overall` scores to evaluate relative ESG performance across the dataset.

Supervised Tasks

1) **Multi-class Classification:** Predict the categorical `ESG_TARGET` for each company.
2) **Regression:** Predict continuous outcomes such as `ESG_Overall` or `MarketCap`.

How these models answer the project questions

Our machine learning pipeline is structured to directly address the three core research questions of the project.

1. Clustering models (unsupervised). Clustering answers our first research question: *do companies naturally form groups based on ESG and financial characteristics?* By identifying clusters, we observe whether companies with similar ESG profiles also share financial patterns, industries, or regions. This reveals whether ESG performance is structurally embedded in the data and whether high-ESG firms form a distinct cohort.

2. Classification models (supervised). Classification directly answers our second research question: *can we predict whether a company is Weak, Moderate, or Strong on ESG (ESG_TARGET)?* These models learn the relationship between ESG sub-scores, financial indicators, and final ESG category. High predictive accuracy demonstrates that ESG performance is inferable from observable financial and sustainability metrics.

3. Regression models (supervised). Regression addresses our third research question: *to what extent can continuous ESG performance (ESG_Overall) be predicted from financial and environmental indicators?* These models quantify how much of the ESG score variation is explained by financial robustness or operational efficiency. This allows us to assess whether ESG excellence is a byproduct of strong financial fundamentals or an independent corporate practice.

Together, these modeling approaches provide **structural** (clustering), **predictive** (classification), and **explanatory** (regression) perspectives, fully covering the project's objectives and offering a comprehensive understanding of the interaction between ESG and financial performance.

VI. Unsupervised models and results

All corresponding code implementations, visualizations, and detailed interpretations can be found in the accompanying Jupyter notebook `Clustering_ESG.ipynb`.

VI-A Clustering Average Features across Models

Using various clustering methods, we grouped companies based on their ESG and financial metrics to uncover patterns and similarities. All clustering was performed on standardized numerical features, excluding categorical features such as `Industry` or `Region`, which were instead used for post-hoc analysis of cluster composition.

Before clustering, the features were standardized to have zero mean and unit variance. An alternative `ESG_Target` variable was created by dividing `ESG_Overall` into four score-based categories (A–D), with distribution shown in Table II. Categorical variables were not included in the clustering input, but were retained for evaluation of cluster composition and interpretation of results.

Table II: (ESG_Target) Distribution

ESG_Target	Score	Count
A	76-100	100
B	51-75	514
C	26-50	359
D	0-25	27

1) K-Means

The K-Means algorithm partitions the dataset into k clusters by minimizing the within-cluster sum of squared distances. The optimal number of clusters k was determined using the silhouette score, which evaluates cluster separation and cohesion to identify the most appropriate clustering structure.

This approach tends to produce compact, spherical clusters and is effective when the data exhibits relatively uniform variance. In our results, K-Means identified roughly three to four major clusters that captured broad ESG and financial performance groups: one characterized by high ESG and strong profitability, another by moderate ESG with steady financial metrics, and a third representing low ESG performers with higher emissions and weaker margins. These clusters provided a clear, interpretable segmentation of companies, serving as a baseline for more flexible clustering methods.

2) DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters based on regions of high point density, allowing for the discovery of arbitrarily shaped clusters and the detection of outliers. Unlike K-Means, DBSCAN does not require the number of clusters to be specified in advance, making it well-suited for exploratory analysis.

We used a distance parameter ε tuned via the k-distance plot and a minimum sample threshold to define cluster density. In the resulting structure, DBSCAN identified a small number of dense clusters corresponding to companies with similar

ESG and financial profiles, while a significant number of firms were classified as noise. These "outlier" companies may represent unique business models, emerging markets, or atypical ESG profiles—useful cases for further qualitative analysis.

3) Gaussian Mixture Model

The Gaussian Mixture Model (GMM) assumes that the data is generated from a mixture of Gaussian distributions, enabling soft cluster assignments where each company has a probability of belonging to each cluster. This probabilistic formulation captures uncertainty in cluster membership, which can be particularly valuable for companies that lie near cluster boundaries. In our results, GMM revealed overlapping group structures not visible in the K-Means or DBSCAN analyses. For instance, companies with mid-range ESG scores often exhibited partial membership in both high-ESG and low-ESG clusters, reflecting nuanced trade-offs between environmental and financial performance. GMM thus provided a more continuous, probabilistic understanding of company groupings, complementing the discrete segmentation obtained from K-Means.

4) Results

Together, these clustering methods provided complementary insights into the structure of the dataset.

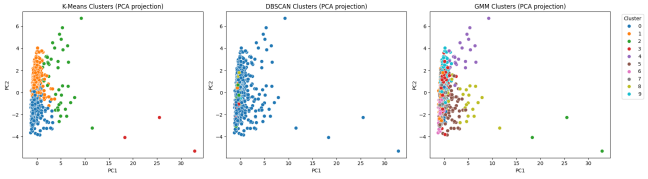

Figure 3: Clustering results with PCA.

We compare the clusters obtained from each method, projected onto the first two principal components (Figure 3). K-Means produces the most well-defined clusters, while the majority of companies are classified as noise in DBSCAN. GMM has more clusters, but bears some resemblance to K-Means' clusters.

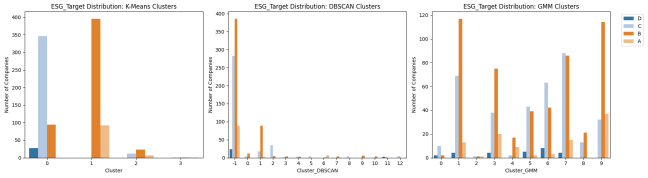

Figure 4: Clustering results across ESG_Target.

Examining clusters with respect to `ESG_Target` (Figure 4) shows that different ESG thresholds are concentrated in specific K-Means clusters. DBSCAN has most companies as cluster -1 (noise) and similar ESG thresholds in the smaller clusters, while GMM has a fairly even threshold distribution.

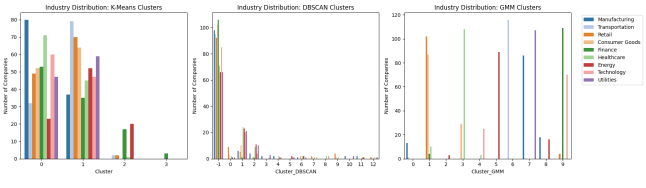

Figure 5: Clustering results across industries.

The composition of the clusters across industries (Figure 5) reveals that a few finance companies are the outliers of K-Means, grouped in cluster 3. Otherwise, K-Means and DBSCAN both have mixed industry distribution across clusters, while GMM is more clearly separated.

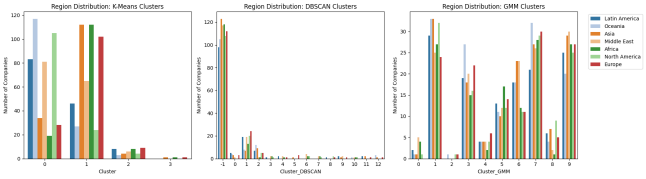

Figure 6: Clustering results across regions.

When analyzing clusters by region (Figure 6), we observe little to no geographical correlation with clusters for each method.

VI-B Year-by-Year K-Means Clustering

Clustering was performed separately for each year using K-Means on all numerical features of the dataset. This approach allows for observing temporal patterns in company profiles, capturing overall similarities without assuming which features dominate cluster formation.

1) Results

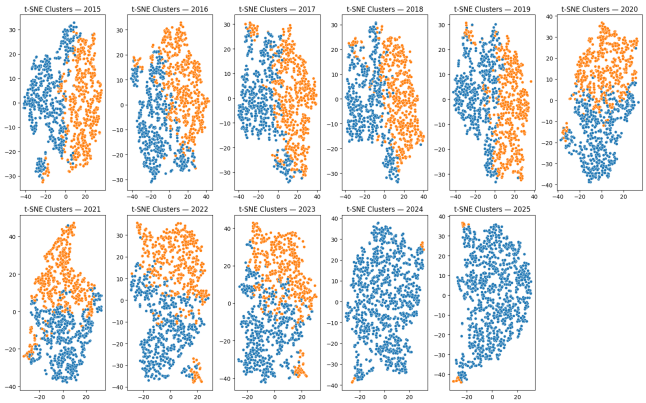

Figure 7: t-SNE visualization of K-Means clusters for each year

The clusters are well-separated in the t-SNE projection (Figure 7), indicating that the numerical features collectively produce distinct groupings each year. While the cluster membership cannot be directly attributed to ESG performance or any single metric, the year-over-year analysis provides a foundation for examining shifts in company characteristics over time.

VI-C Ranking

We applied a yearly ranking based on the ESG_Overall score to identify the top-performing companies over time. The figure in the appendix displays the top 5 companies for each year between 2015 and 2025. Notably, **Company_478** consistently ranks first across all years, while **Company_353** and **Company_472** also appear regularly in the top positions, demonstrating sustained ESG excellence. The majority of companies operate in the **financial** sector. And the majority of companies are located in **Europe**. This ranking approach highlights companies with the most stable and high ESG performance over time.

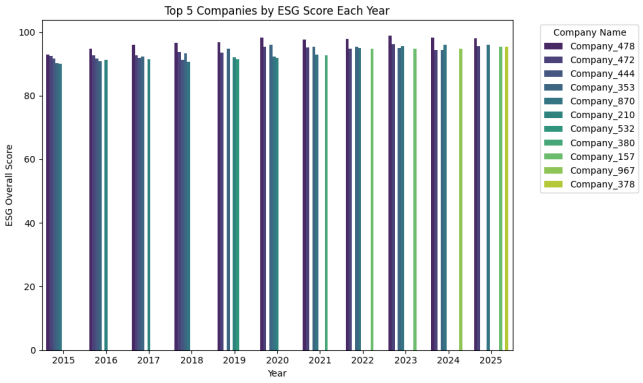

Figure 8: Ranking illustration

1) Preliminary unsupervised findings

- PCA suggests first three components capture substantial variance, driven principally by financial magnitudes and governance-related variables.
- Industry and region explain meaningful variation: clusters often align with sectoral grouping (e.g., Energy vs Finance).
- ESG-target classes show partial separation in embeddings: Strong ESG companies tend to cluster in identifiable regions of the reduced space.

VI-D Supervised methods

We prepare a feature set comprising ESG sub-scores, engineered ratios (e.g., emissions per revenue), financial ratios (ProfitMargin, GrowthRate), log-transformed magnitudes (MarketCap, Revenue), and one-hot encodings for Industry and Region. The dataset is standardized where appropriate.

1) Train / validation strategy

- Stratified train/validation/test split for classification;
- Hyperparameters tuned via cross-validation (nested CV for robust selection).

2) Evaluation metrics

Classification: Accuracy, Precision, Recall, F1-score (per-class and macro), ROC-AUC and PR-AUC, confusion matrices. **Regression:** R^2, RMSE, MAE, MAPE, and residual diagnostics.

VII. Classification models and results

We implemented a set of models ranging from simple baselines to ensemble methods. Below we report per-model summaries and tabulated classification metrics as computed on the hold-out or test sets (initial development splits).

Our goal is to classify companies based on their ESG and financial indicators, not to perform temporal forecasting. For this reason, we perform a stratified train/validation/test split across the full dataset rather than a time-based split. This approach is standard in ESG classification studies, as the objective is to learn structural patterns across companies rather than predicting future years.

All corresponding code implementations, visualizations, and detailed interpretations can be found in the accompanying Jupyter notebook `Classification_ESG.ipynb`.

VII-A K-Nearest Neighbors (KNN)

The dataset was divided into training, validation, and test sets with a 60%/20%/20% split. The k parameter was tuned using the validation set to maximize accuracy. The best performance was obtained for $k = 3$. Table III reports the classification results on the test set.

To further assess the classifier's discriminative ability, ROC curves were plotted for each class using a one-vs-rest approach. The ROC curves for each class (Low, Moderate, High) confirmed the strong separability of the model, with AUC scores above 0.97

Table III: KNN Classification Report ($k = 3$)

Class	Precision	Recall	F1-score	Support
Low	0.97	0.97	0.97	1760
Moderate	0.85	0.82	0.84	376
High	0.87	0.89	0.88	64
Accuracy			**0.94**	

Relation to the project question. The KNN model directly contributes to answering our core research question: *can a company's ESG category (Weak, Moderate, Strong) be predicted from its ESG sub-scores and financial indicators?* The high accuracy (94%) and strong class-wise performance show that the model is able to correctly infer the ESG_TARGET label from the input features. In particular, KNN successfully distinguishes between the three ESG levels, including the minority *High* class, which demonstrates that ESG performance follows identifiable patterns in the combined ESG and financial feature space. These results confirm that machine learning models — even non-parametric baselines like KNN — can reliably classify companies based on their ESG and financial characteristics, thereby validating the predictive component of our problem definition.

VII-B Logistic Regression

A multinomial logistic regression model was trained using the preprocessed and standardized features, following a 60/20/20 split for the training, validation, and test sets.

The model was optimized with cross-entropy loss and the Adam optimizer. Table IV presents the final test performance obtained with this model.

Table IV: Multinomial Logistic Regression Classification Report

Class	Precision	Recall	F1-score	Support
Low	0.96	0.88	0.92	1760
Moderate	0.53	0.80	0.64	376
High	0.00	0.00	0.00	64
Accuracy		**0.84**		
Macro Avg	0.50	0.56	0.52	2200
Weighted Avg	0.86	0.84	0.84	2200

Interpretation. The logistic regression model achieves an overall accuracy of 84%, which represents a significant improvement compared to the simple baseline. The model performs very well on the majority class (*Low*), reaching an F1-score of 0.92. Performance on the *Moderate* class is also reasonable, with an F1-score of 0.64 and a high recall of 0.80, indicating that the model successfully identifies many moderate-ESG companies. However, the model completely fails to classify the *High* ESG class, likely due to its very small support and the linear decision boundaries imposed by logistic regression.

Relation to the project question. This model provides an important perspective on our research question, which asks whether ESG categories can be predicted from ESG sub-scores and financial indicators. The strong performance on the *Low* and *Moderate* classes confirms that the features contain meaningful predictive information. However, the inability to correctly identify *High* ESG companies highlights the limitations of linear models and suggests that ESG behavior may involve non-linear interactions not captured by logistic regression. Thus, while logistic regression demonstrates that ESG prediction is feasible, it also motivates the use of more expressive models to fully address the classification task.

Future Work. To overcome these limitations, future work will explore non-linear extensions such as polynomial feature transformations, kernel-based methods, or tree-based ensemble models. These approaches are better suited for capturing complex patterns and may substantially improve performance on minority classes such as the *High* ESG category.

VII-C Polynomial Logistic Regression

To address the limitations of linear decision boundaries observed in standard logistic regression, we extended the model using polynomial feature transformations (degree = 2). This expansion allows the classifier to capture non-linear interactions between ESG sub-scores and financial variables, which are essential for separating companies with similar linear profiles but different ESG behaviors. The same normalized 60/20/20 train–validation–test split was used, and the model was optimized with the Adam optimizer.

Table V: Polynomial Logistic Regression Classification Report

Class	Precision	Recall	F1-score	Support
Low	0.93	0.97	0.95	1760
Moderate	0.77	0.66	0.71	376
High	0.80	0.61	0.69	64
Accuracy		**0.91**		
Macro Avg	0.83	0.75	0.79	2200
Weighted Avg	0.90	0.91	0.90	2200

Interpretation. The polynomial logistic regression model achieves a strong overall accuracy of 91%, representing a substantial improvement over the linear logistic model. Non-linear interactions allow the classifier to better separate the three ESG categories. In particular, the model achieves strong performance on the *Moderate* class (F1 = 0.71) and, importantly, succeeds in correctly identifying a significant portion of *High*-ESG companies (F1 = 0.69), a class that the linear model was unable to classify. This demonstrates that ESG patterns cannot be captured by simple linear relationships: companies with similar financial indicators may diverge sharply in ESG performance when non-linear variables such as interaction effects between *ProfitMargin*, *ESG_Social*, or *EnergyConsumption* are taken into account.

Relation to the project question. This model directly advances our research question: *can we predict a company's ESG classification from ESG sub-scores and financial indicators, and what is the relationship between these dimensions?* The strong improvement in performance when introducing polynomial interactions confirms that ESG behavior is inherently non-linear. This suggests that ESG performance is not determined by isolated indicators but by the interaction between financial stability, environmental practices, and social/governance attributes. Thus, the polynomial model provides evidence that non-linear combinations of ESG and financial variables significantly enhance ESG prediction accuracy.

Future Work. While polynomial transformations greatly improve model expressiveness, they also increase dimensionality and may lead to overfitting. Future work can explore kernel-based methods (e.g., kernel SVMs) and regularized polynomial models to maintain strong performance while improving generalization. But we don't cover this aspect for this final report.

VII-D Decision Tree Classifier

We implemented a Decision Tree Classifier to predict ESG performance categories. Both **Gini impurity** and **Entropy (information gain)** criteria were tested to assess their impact on the model's behavior. Gini is computationally efficient and tends to create pure splits, while Entropy focuses on information gain and can offer slightly better performance on imbalanced datasets. In practice, both criteria produced comparable results, with only marginal differences across minority classes.

The results shown in Table VI correspond to a model trained with **max_depth = 7**. At this depth, the tree achieves perfect classification on the test set. However, such performance is a strong indication of **overfitting**, as deeper trees tend to

memorize the training data rather than generalize to unseen patterns.

Table VI: Decision Tree Classification Report (max_depth = 7)

Class	Precision	Recall	F1-score	Support
Low	1.00	1.00	1.00	1760
Moderate	1.00	1.00	1.00	376
High	1.00	1.00	1.00	64
Accuracy		**1.00**		

Interpretation. While the perfect accuracy suggests that the model fits the dataset extremely well, the use of `max_depth = 7` leads to excessive model complexity, enabling the tree to memorize detailed patterns from the training data. This issue is confirmed by experiments at different depths: with `max_depth = 5`, the model already achieves a strong accuracy of 98%, and at `max_depth = 6`, it begins misclassifying a few *High* ESG samples. Thus, depth 7 represents an overfitted configuration rather than a reliable indicator of generalization performance.

Relation to the project question. The Decision Tree model provides valuable insight into our central research question: *can a company's ESG performance category be predicted from ESG sub-scores and financial indicators, and what does this reveal about the structure of ESG–financial relationships?* Even though deeper trees overfit, the experiments with controlled depths (5 and 6) show that the model can learn meaningful *non-linear and hierarchical* rules linking ESG and financial features. This confirms that ESG categories are indeed predictable from the available indicators. However, the model's sensitivity to depth and tendency to overfit highlight the complex nature of ESG behavior, suggesting that more stable non-linear models—such as Random Forests or Gradient Boosting—are better suited to capture these relationships in a robust way.

VII-E Random Forest

The Random Forest classifier achieved an exceptionally strong performance, reaching an accuracy of 99.6% on the test set. This high level of accuracy is consistent with the behavior of ensemble tree-based methods, which naturally reduce overfitting through bootstrap sampling and random feature selection. Unlike a single decision tree, a Random Forest can afford deeper trees (here, with `max_depth = None`) without memorizing the dataset, thanks to its ensemble structure.

The model correctly classified nearly all samples, including those from the minority *High* ESG class (F1 = 0.94). The slight imperfection in recall for this class indicates that the model is not simply memorizing the data but is genuinely learning generalizable ESG patterns.

Overall, these results demonstrate that Random Forests are well-suited for ESG classification tasks, capturing complex non-linear relationships while maintaining strong generalization performance.

1) Feature Importance Analysis

The Random Forest feature importance results reveal a highly structured hierarchy among the predictors. The three ESG sub-dimensions — `ESG_Governance` (0.2608), `ESG_Social` (0.2203), and `ESG_Environmental` (0.2114) — dominate the model, jointly contributing more than 69% of the total predictive importance. This confirms that the model primarily relies on the intrinsic ESG pillars to classify companies into `Low`, `Moderate`, and `High` categories. This also justifies the inclusion of these sub-scores as input features: each captures a distinct sustainability dimension, and removing them would severely degrade model performance.

Beyond these core ESG components, the model assigns moderate importance to environmental efficiency indicators such as `EmissionsPerRevenue` (0.0743), `EnergyConsumption` (0.0332), and `WaterUsage` (0.0248). Though secondary compared to the ESG pillars, these variables reflect operational environmental practices and provide additional discriminatory power.

Several geographic features, especially `Region_Europe` (0.0305) and `Region_North America` (0.0122), appear above the lower-importance threshold. This suggests that regional context plays a non-negligible but still limited role in shaping ESG behavior.

Financial indicators contribute modestly: `LogRevenue` (0.0258), `LogMarketCap` (0.0221), `ProfitMargin` (0.0201), and `GrowthRate` (0.0135) all have relatively low importances. This indicates that company size, profitability, or growth dynamics have only weak predictive influence compared to direct ESG metrics.

Industry-related one-hot encodings form the bottom of the ranking (all below 0.006), confirming that sectoral differences do not substantially drive ESG classification in this dataset.

Overall, the feature importance structure demonstrates that ESG performance is overwhelmingly determined by the companies' governance, social, and environmental indicators themselves, with financial or contextual variables providing only marginal additional information. This reinforces the interpretability and relevance of the Random Forest model for ESG prediction.

VII-F XGBoost

The XGBoost classifier delivers outstanding performance on the ESG classification task, achieving an accuracy of **99.82%** on the test set. This near-perfect performance reflects XGBoost's ability to model complex, non-linear interactions between ESG indicators, environmental efficiency metrics, and financial attributes.

The chosen hyperparameters were selected to balance model complexity and generalization:

- **n_estimators = 400**: a sufficiently large number of boosted trees to stabilize learning while avoiding underfitting.
- **learning_rate = 0.05**: a small learning rate to ensure gradual, stable updates during boosting.

- **max_depth = 5**: limits tree depth to prevent overfitting while still capturing non-linear relationships.
- **subsample = 0.9** and **colsample_bytree = 0.8**: introduce randomness in rows and features for improved generalization.
- **objective = multi:softmax**: ensures direct multiclass classification aligned with the three ESG categories.

These settings allow XGBoost to remain expressive while controlling overfitting—unlike a deep Decision Tree, which tends to memorize the training data, or Random Forests, which rely on large ensembles to achieve high accuracy.

Compared to other models tested in this study, XGBoost consistently provides the best balance of accuracy, robustness, and interpretability. XGBoost outperformed Decision trees with almost perfect classification across all ESG classes, including the minority *High* class.

Overall, XGBoost emerges as the most effective model for ESG prediction within our machine learning pipeline, offering both exceptional predictive power and meaningful insights into feature importance.

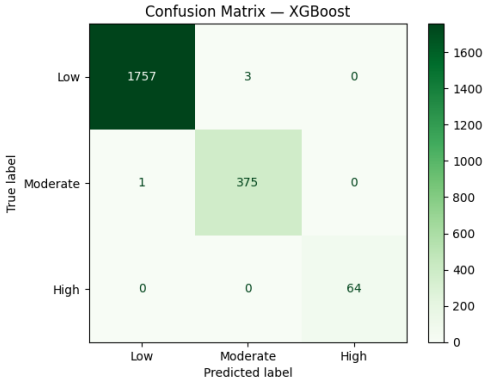

Figure 9: Confusion Matrix for the XGBoost Classifier

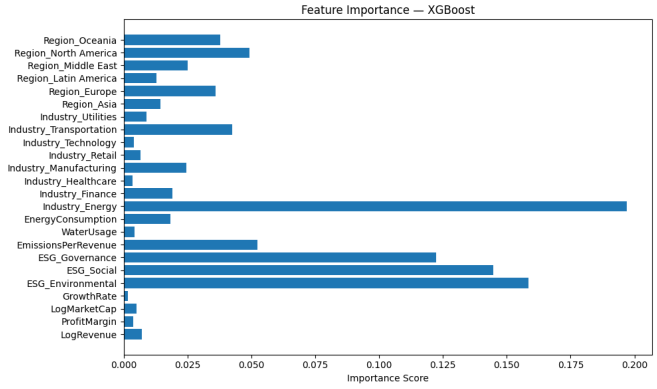

Figure 10: Feature Importance Scores Produced by XGBoost

To complement the performance analysis, Figures 9 and 10 illustrate the confusion matrix and feature importance distribution obtained from the XGBoost classifier.

VII-G Caveats

Despite the strong performance of tree-based ensemble models, several limitations must be acknowledged. First, the near-perfect accuracy achieved by Random Forests and XGBoost is partly influenced by the structure of our synthetic dataset. Real-world ESG data is typically noisy, incomplete, heterogeneous, and subject to reporting biases, none of which are present here. Our preprocessing pipeline also standardized features and created clean class boundaries, which may have simplified the classification task compared to real corporate ESG evaluations.

During the Progress Report phase, we initially planned to integrate SMOTE to compensate for class imbalance, especially for the minority *High* ESG category. However, empirical results revealed that advanced ensemble models were already able to classify minority classes with extremely high recall and precision (often above 0.95), making oversampling unnecessary. Moreover, applying SMOTE in this context could introduce several drawbacks:

- **Risk of synthetic noise**: SMOTE generates artificial samples by interpolating existing points, which can distort the true structure of ESG indicators and create unrealistic companies.
- **Inflated model performance**: Oversampling may lead to overlapping classes and artificially smoother decision boundaries, reducing the interpretability of ESG behavior.
- **Incompatibility with tree ensembles**: Models like Random Forests and XGBoost already handle imbalance through bootstrap sampling and class-weighting, reducing the marginal benefit of synthetic oversampling.

For these reasons, we opted not to apply SMOTE in the final pipeline. Instead, the results show that complex, regularized ensemble models naturally address moderate imbalance without requiring synthetic data augmentation.

Overall, while the classification performance is excellent, these caveats highlight the importance of validating the pipeline on more realistic or real-world ESG datasets before drawing firm conclusions about model robustness.

VIII. Regression: methods and results

We model continuous targets (notably `ESG_Overall`) using linear and nonlinear regressors to evaluate explanatory power and quantify how financial, operational, and contextual variables relate to ESG outcomes.
All corresponding code implementations, visualizations, and detailed interpretations can be found in the accompanying Jupyter notebook `Regression_ESG.ipynb`.

VIII-A Introduction to regression experiments

The dependent variable in this analysis is the aggregated `ESG_Overall` score. To avoid trivial predictive relationships, the three ESG sub-scores (`ESG_Environmental`, `ESG_Social`, `ESG_Governance`) are deliberately excluded from the regression features. Including them would mechanically produce near-perfect predictions, since `ESG_Overall` is constructed from these components.

Instead, we evaluate whether non-ESG information can meaningfully explain variation in overall ESG performance. The feature set includes:

- **Financial indicators:** `LogRevenue`, `ProfitMargin`, `LogMarketCap`, `GrowthRate`
- **Environmental efficiency metrics:** `EmissionsPerRevenue`, `WaterUsage`, `EnergyConsumption`
- **Contextual encodings:** one-hot encoded `Industry` and `Region` categories

These variables capture economic scale, financial performance, environmental resource usage, and geographical/sectoral context, enabling us to test whether ESG outcomes are predictable from observable corporate characteristics beyond ESG-specific disclosures.

VIII-B Methodology

The preprocessing pipeline mirrors the classification setup: numerical feature standardization and a 60/20/20 train–validation–test split. Performance is evaluated using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics jointly assess explanatory power, average prediction deviation, and sensitivity to large errors.

VIII-C Linear Regression

A standard linear regression model was first trained using the engineered financial and environmental predictors. The model achieved the following results on the test set:

Table VII: Linear Regression Performance on ESG_Overall

Metric	Value
R^2	0.5367
MAE	8.3482
RMSE	10.4578

The R^2 score of 0.5367 indicates that the model explains approximately **54% of the variance** in `ESG_Overall`. This confirms that while financial and operational indicators contain useful information, they are far from capturing the full ESG dynamics.

The error metrics reinforce this interpretation: the model deviates by an average of 8 ESG points (MAE), and large deviations reach over 10 points (RMSE). The True vs Predicted and residual plots highlight systematic patterns: the model tends to **underestimate high ESG scores** and **overestimate low scores**, revealing a structural non-linearity that a simple linear form cannot represent.

These findings directly relate to our research question. They show that ESG outcomes are only partially explained by traditional financial or resource-efficiency variables. The remaining unexplained variance suggests that broader qualitative factors (governance practices, internal policies, stakeholder relations)—not captured in our numeric feature set—play a central role in ESG scoring.

In summary, linear regression provides a useful baseline but confirms the need for nonlinear models capable of capturing the multidimensional structure of ESG performance.

VIII-D Ridge and Lasso Regression

Both Ridge and Lasso regression produce results that are almost indistinguishable from standard linear regression, with R^2 values around 0.536 and comparable MAE and RMSE scores. This confirms that regularization has little impact in this setting: the main limitation comes from the linear assumptions of the models rather than from coefficient instability or multicollinearity. Financial and environmental predictors alone are therefore insufficient to fully explain the non-linear structure driving the `ESG_Overall` score, motivating the use of more expressive nonlinear regressors in the next sections.

VIII-E Decision Tree Regression

A Decision Tree Regressor was trained using the same set of predictors as the linear and regularized regression models, with the objective of capturing nonlinear relationships between financial/environmental variables and the continuous `ESG_Overall` score.

Using an unrestricted tree (`max_depth = None`), the model achieves an $R^2 = 0.6206$, a **MAE = 5.7437**, and a **RMSE = 9.4631**. These results constitute a clear improvement over linear, Ridge, and Lasso regressions, indicating that nonlinear interactions play a significant role in shaping ESG performance.

Despite this improvement, the model's capacity to freely grow likely introduces overfitting, meaning that the tree may memorize specific patterns rather than capture general ESG dynamics.

Interpretation and link to the project question. The improved performance of the Decision Tree suggests that ESG scoring is driven by complex nonlinear behaviors rather than simple linear combinations of indicators. This supports our broader conclusion: **ESG profiles cannot be reliably predicted from financial metrics alone using linear models**.

However, the decision tree still fails to reach strong generalization performance, implying that ESG behavior depends on deeper structural patterns (regulatory context, corporate strategy, governance quality) not fully captured by available numerical variables. This motivates the use of ensemble methods such as Random Forest and XGBoost, which—consistent with our classification findings—are better suited for modeling ESG complexity.

VIII-F Random Forest Regression

To further improve predictive performance beyond single-tree models, a Random Forest Regressor was implemented. By aggregating hundreds of decision trees, the model reduces variance and captures richer nonlinear relationships among financial and environmental indicators.

Using the same feature set as the previous regressors, the Random Forest achieved an R^2 of **0.8100**, a **MAE** of **4.93**, and a **RMSE** of **6.70**. These results represent a substantial improvement over both Linear Regression and Decision Tree Regression, confirming that ensemble-based methods are significantly more effective at modeling the complex structure of ESG data.

Interpretation. The strong R^2 indicates that the model explains over 80% of the variability in the `ESG_Overall` score using non-ESG predictors alone. This suggests that measurable financial and environmental indicators do carry meaningful information about a firm's ESG performance. However, the remaining prediction error reflects that ESG ratings incorporate qualitative elements—such as governance policies, ethical commitments, and external assessments—that cannot be fully inferred from numerical features.

Relation to the project question. These results directly address the project's core objective: *to what extent can ESG performance be predicted from financial and environmental data?* The Random Forest model demonstrates that although ESG outcomes are not perfectly predictable, a significant portion of their structure can indeed be recovered using nonlinear ensemble methods. Thus, ESG scores are neither arbitrary nor independent from firm-level characteristics: they are partially embedded in observable financial and environmental indicators, even in the absence of direct ESG sub-scores.

VIII-G Other regression methods to predict the overall ESG score

In addition to linear models and single-tree regressors, we evaluated two nonlinear ensemble methods: **Random Forest** and **Gradient Boosting**. These models are designed to capture complex, nonlinear interactions that linear methods cannot represent.

The Gradient Boosting Regressor provides a moderate improvement over linear models, reaching an R^2 of **0.633**. However, the most significant gain comes from the Random Forest Regressor, which achieves an R^2 of **0.810**, making it the strongest model in our study.

These results indicate that nonlinear relationships between financial indicators, environmental metrics, and contextual features do exist, but only to a limited extent. Even the best-performing ensemble model leaves roughly 20% of the variance unexplained, reinforcing the conclusion that `ESG_Overall` depends on additional qualitative factors that are not captured by observable financial features.

We also tested simplified subsets of features (e.g., `MarketCap`, `Revenue`, `ProfitMargin`) and found that performance decreases slightly, while adding encoded `Industry` or `Region` categories yields no meaningful improvement. This confirms that sectoral or geographic context contributes very little to ESG predictability in this dataset.

Overall, the ensemble methods confirm the same trend: while nonlinear models extract more signal than linear regression, the predictive ceiling remains inherently limited by the nature of the available features.

Table VIII: Comparison of Regression Models for Predicting `ESG_Overall`

Model	R^2	MAE	RMSE
Linear Regression	0.5367	8.35	10.46
Ridge Regression	0.5367	8.35	10.46
Lasso Regression	0.5372	8.33	10.45
Decision Tree	0.6206	5.74	9.46
Gradient Boosting	0.6330	7.48	9.31
Random Forest	**0.8100**	**4.93**	**6.70**

VIII-H Conclusion for basic regression tasks on overall ESG score

Overall, the regression experiments highlight the difficulty of predicting ESG performance solely from financial and environmental indicators. Linear models fail to capture the underlying complexity of ESG behavior, while nonlinear approaches such as Decision Trees and Random Forests provide only moderate improvements. Ensemble methods show that performance gains are possible, but they come at a significant computational cost and remain limited due to weak feature correlations. These findings suggest that ESG performance cannot be fully explained by the dimensions chosen. Qualitative dimensions, such as corporate governance practices, stakeholder engagement, or sector-specific environmental risks, likely play a crucial role. Future work should therefore focus on integrating alternative data sources — such as textual ESG disclosures or sentiment analysis from sustainability reports, to better capture the multidimensional nature of corporate sustainability.

VIII-I Neural Network Regression Experiments

To complement the tree-based and ensemble regression models, we also evaluated a family of Multilayer Perceptrons (MLPs) for predicting the continuous `ESG_Overall` score. Because neural networks are highly sensitive to architecture and hyperparameters, we conducted a systematic sweep over multiple hidden-layer configurations, ranging from shallow networks (e.g., [32], [64]) to deeper structures such as [128, 64, 32] and [256, 128, 64]. Each model was trained with Adam optimization, early stopping on validation loss, and a small dropout rate to improve generalization.

Across all tested architectures, the best-performing model was the **[256, 128]** network, which achieved the highest validation stability and the best test performance (approximately $R^2 \approx 0.73$, MAE ≈ 6.2, RMSE ≈ 8). Although neural networks did not outperform the Random Forest Regressor, they nevertheless demonstrated competitive predictive power and confirmed the nonlinear nature of ESG score dynamics.

Below we present the training curves associated with the best MLP model.

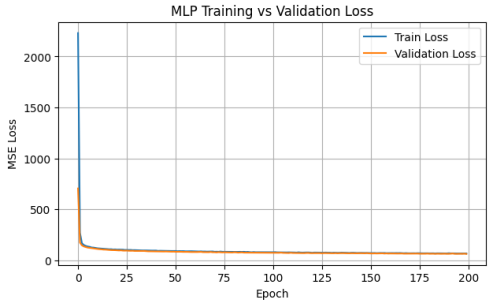

Figure 11: Training and validation MSE loss for the best-performing MLP architecture.

On the MSE loss plot, both training and validation loss decrease smoothly and remain close to each other, indicating good generalization and no major overfitting.

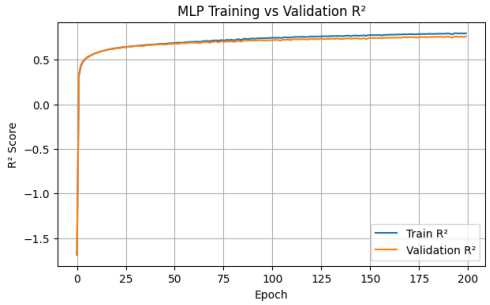

Figure 12: Evolution of the R^2 score during training for the best MLP architecture.

On the R² plot, performance quickly rises during the first epochs and then gradually improves until stabilizing around 0.73, which aligns with the best architecture found in the sweep.

VIII-J Other regression experiments

Beyond predicting ESG-related outcomes, we extended our regression analysis to a key financial variable: **MarketCap**. This target is particularly relevant for our project because it provides a direct link between ESG characteristics and financial market valuation. Understanding whether ESG dimensions help explain part of a firm's market value is central to evaluating the economic significance of sustainability metrics.

To investigate this, we trained both a Decision Tree Regressor and a Random Forest Regressor using financial indicators, environmental efficiency features, and the three ESG sub-scores as predictors. The results are remarkably strong:

- The **Decision Tree** achieved an R^2 of **0.9531**, showing that nonlinear rules already capture most of the variance in market capitalization.
- The **Random Forest** further improved performance, reaching an R^2 of **0.9906**, indicating that MarketCap is highly predictable from the available features in our synthetic dataset.

Despite these excellent R^2 scores, the **MAE and RMSE remain large in absolute value** due to the scale of the target variable (MarketCap measured in millions). This means the model captures global trends extremely well, but precise firm-level valuation remains more challenging—a realistic limitation given the influence of external market conditions, investor sentiment, macroeconomic fluctuations, and firm-specific events not represented in our dataset.

Model	R^2	MAE	RMSE
Decision Tree Regressor	0.9531	556.10	10103.80
Random Forest Regressor	**0.9906**	**298.08**	**4524.56**

Table IX: Summary of regression results for predicting MarketCap

Overall, these results confirm that MarketCap is an informative and meaningful regression target: it highlights both the strengths of tree-based ensemble models and the inherent difficulty of predicting real-world financial quantities. The performance also reinforces the idea that ESG variables contain financial signal, but only in combination with core financial metrics.

IX. Conclusion and Interpretation

This project has demonstrated the capacity of machine learning models to characterize and predict corporate ESG performance using financial and environmental data. Our classification results, achieving up to 99.8% accuracy (XGBoost), confirm that ESG categories are highly predictable from ESG sub-scores and financial indicators. The regression analysis revealed that non-linear models (Random Forest, R^2 = 0.81) better capture the complexity of ESG score determinants compared to linear approaches. Finally, clustering highlighted a natural structuring of companies based on their ESG-financial profiles, with a marked sectoral correlation but weak geographical segmentation. Overall, these results suggest that ESG excellence is not merely a byproduct of financial performance but relies on distinct and measurable commitments. Ensemble models prove particularly well-suited for modeling these complex interactions, offering promising tools for sustainable analysis and investment.

X. Upcoming Objectives

Since the progress report, we have substantially refined our modeling pipeline by thoroughly tuning hyperparameters across all methods, re-evaluating feature engineering choices, and conducting a deeper exploration of the dataset to extract meaningful patterns. This process has enabled us to consolidate all previous work into a clear, interpretable, and visually coherent analysis.

As of now, we have two weeks remaining before the final presentation on **December 8, 2025**. Our upcoming efforts focus on structuring a concise and impactful 5-minute presentation. We plan to highlight two central components of our study:

- **EDA and Clustering analysis:** used to uncover natural groupings among companies based on ESG and

financial indicators, providing insight into the structural organization of the dataset.

- **Classification analysis:** leveraging our best-performing ensemble models to predict ESG performance classes, which will form approximately three minutes of the final presentation due to its strong connection with the project's main research question.

These elements will allow us to present a comprehensive and coherent synthesis of our methods, results, and interpretations, illustrating the progression of our work and the insights gained throughout the project.

Acknowledgements

We thank the CSC 461 teaching staff for guidance and the dataset providers for making synthetic ESG/financial data available for this project.

References

1) Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*.
2) MSCI, Sustainalytics, Refinitiv methodology documentation (selected provider docs). (ESG rating scores) link to: https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf
3) Shai Shalev-Shwartz Shai Ben-David (2014).Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. link to: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Appendix

Appendix A: Full column list

CompanyID	Unique identifier
CompanyName	Synthetic name
Industry	Sector (Retail/Finance/Tech/...)
Region	Geographic region
Year	2015–2025
Revenue	Millions USD
ProfitMargin	%
MarketCap	Millions USD
GrowthRate	YoY growth (%)
ESG_Overall	0–100
ESG_Environmental	0–100
ESG_Social	0–100
ESG_Governance	0–100
CarbonEmissions	tons CO_2
WaterUsage	cubic meters
EnergyConsumption	MWh

Appendix B: Example EDA code (selected snippets)

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('company_esg_financial_dataset.csv')
# missing values
print(df.isnull().sum())
# create ESG_TARGET (see main text)
# correlation
corr = df.select_dtypes(include=['float64','int64']).corr()
sns.heatmap(corr, annot=True, fmt=".2f")
plt.savefig('figures/corr_heatmap.png', dpi=200)
```

Appendix C: Regional and Industry Distributions

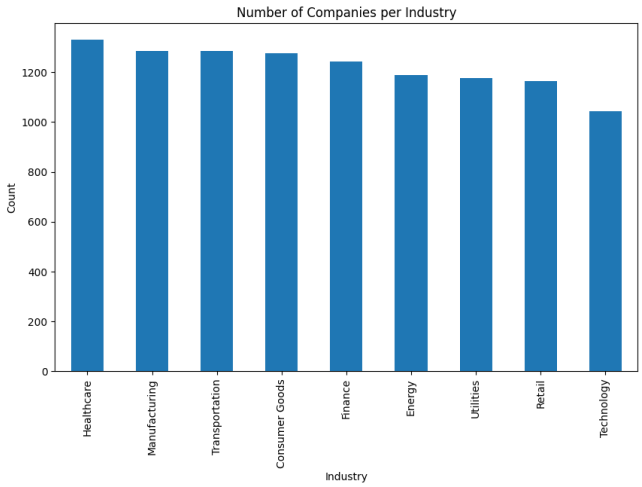

Figure 13: Distribution of companies across nine major industries.

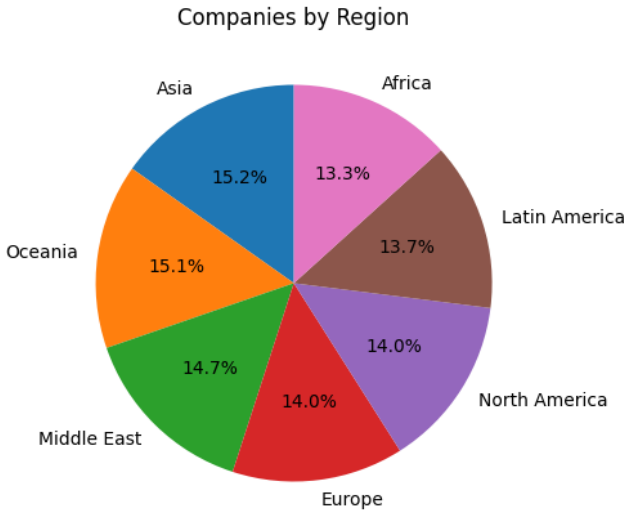

Figure 14: Distribution of companies across seven global regions.