NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

TRADITIONAL MACHINE LEARNING vs. GPT-4

A Comparative Study on ESG Score Prediction

Chiara Ciraso

Master Thesis

presented as partial requirement for obtaining the Master Degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

TRADITIONAL MACHINE LEARNING vs. GPT-4

A Comparative Study on ESG Score Prediction

by

Chiara Ciraso

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Data Analytics

Supervised by

Prof. Doctor José Américo Alves Sustelo Rio, NovaIMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15 July 2024

DEDICATION

If you are reading this, you are likely one of the reasons why I have reached this point and put the pieces back together, while acknowledging that complexity and contradictions are inherent parts of us all. Thank you for seeing beyond my sight when I could not, and for always having my back. You made this journey sweeter and my heart lighter. You are the best team I could wish for. Grazie.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Doctor José Américo Alves Sustelo Rio, for his unwavering support, guidance, and encouragement throughout the course of this master's thesis. His expertise, curious outlook and insightful feedback have been invaluable, and his dedication and enthusiasm have greatly contributed to the completion of this work. Thank you for your continuous mentorship.

ABSTRACT

This thesis examines the performance of traditional machine learning (ML) methods and large language models (LLMs), specifically GPT-4, within the context of predicting Environmental, Social and Governance (ESG) scores.

Leveraging on a dataset based on information by leading companies in the realm of environmental indicators and ESG rating, respectively Urgentem and Moody's, this study evaluates the models under three distinct scenarios: high-dimensional noisy data, lower-dimensional imbalanced data, and datasets with scarce information. Traditional ML methods, including Decision Trees (DT), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), are compared against GPT-4's capabilities in handling these conditions to explore strengths and weaknesses of both approaches across different aspects of the analytical chain, from data pre-processing to interpretation.

Results indicate that traditional ML models, particularly ensembles of DT, KNN, and SVM, excel in handling high-dimensional and noisy. In contrast, GPT-4 demonstrates superior performance with simpler, well-defined use-cases. Both approaches reveal specific advantages and limitations, highlighting the importance of model selection based on dataset characteristics and specific application needs.

The thesis underscores the importance of balancing automated approaches with expert control. The findings suggest a complementary application of traditional ML methods and LLMs, leveraging their respective strengths for more robust and comprehensive ESG scoring frameworks tailored to the practitioners' need.

KEYWORDS

Machine Learning; Large Language Models; Generative AI; GPT-4; ESG Scores; Sustainability

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CSRD	Corporate Sustainability Reporting Directive
CV	Cross-Validation
ELM	Extreme Learning Machines
ESG	Environmental, Social, and Governance
EU	European Union
GLM	Generalized Linear Models
GPT-4	Generative Pre-trained Transformer 4
KNN	K-Nearest Neighbors
LLM	Large Language Model
LR	Logistic Regression
MLE	Maximum Likelihood Estimation
NaN	Not a Number (missing values)
NFRD	Non-Financial Reporting Directive
NLP	Natural Language Processing
PCA	Principal Component Analysis
PLTR	Penalized Logistic Tree Regression
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
SFDR	Sustainable Finance Disclosure Regulation
SVM	Support Vector Machines
XGBoost	eXtreme Gradient Boosting

1 INTRODUCTION

The advent and widespread democratization of generative Artificial Intelligence (AI), particularly models like GPT-4, is transforming the way we operate, think, and tackle challenges. Its advanced text-generation capabilities, coupled with extensive topical breadth, present a significant opportunity for industries across various sectors. Of particular interest is the application of such technology to the field of sustainability, where the current environmental and social crises stand out as crucial issues requiring innovative and multifaceted solutions.

As awareness of environmental and social issues grows rapidly, so do the regulatory initiatives aimed at promoting long-term sustainable development by encouraging industries to adopt more responsible practices. These mandates compel companies to disclose more detailed non-financial information, supporting the new concept of sustainable finance and the incorporation of Environmental, Social, and Governance (ESG) metrics into scoring mechanisms.

This shift is driven by the growing pressure from climate change and the increased social awareness, leading to broader discussions on the impact of ESG practices on a firm's financial performance. ESG scores are used to assess a company's commitment to sustainable and ethical practices, which in turn influences their access to funding and investment opportunities. Environmental considerations include issues like climate change mitigation and adaptation, the safeguard of biodiversity, prevention of pollution also in view of a more circular economy. Social aspects refer to the spectrum of inequality, diversity, inclusiveness, working conditions, as well as human rights issues. The governance component examines aspects like management structure, executive remuneration, board diversity and structure, as well as political lobbying, corruption, and bribery, which all play a fundamental role in ensuring the inclusion of social and environmental considerations in the decision-making process.

This collective endeavor has taken further shape in the European Union (EU) policy context through a set of key initiatives, to embed ESG considerations into decision-making and sustainable finance, also in view of reaching the target set by the European Green Deal, aiming to make Europe the first climate-neutral continent by 2050. Among the most crucial programs mandating the disclosure of ESG information at the EU level, the following are the major ones. The EU Taxonomy Regulation provides clear criteria for identifying environmentally sustainable activities, enhancing transparency and consistency in ESG investments. The Sustainable Finance Disclosure Regulation (SFDR) mandates financial market participants to disclose how they integrate ESG risks into their processes, promoting greater transparency. The Corporate Sustainability Reporting Directive (CSRD), an update to the Non-Financial Reporting Directive (NFRD), requires companies to provide detailed ESG reports, extending reporting obligations to more firms with stricter standards.

However, integration of ESG metrics into financial decision-making does not come without any costs and is challenging due to several factors. Firstly, the recent surge in interest to disclose ESG information has resulted in a vast amount of data, where quality and consistency can vary significantly among industries and sectors. This issue is driven on the one hand by the relatively recent introduction of regulatory initiatives and on the other hand by the diverse nature of non-financial disclosures, which further complicate the comparability of ESG scores. Additionally, the high cost of collecting and accessing comprehensive ESG data constitutes a barrier for many practitioners and researchers.

In this scattered and complex scenario, a crucial role is played by institutions and rating agencies combining this information into transparent and comparable indicators and scores. ESG rating agencies act as independent third parties, evaluating the sustainability profile of companies and financial instruments through various methodologies. They measure exposure to sustainability risks and assess the impact on society and the environment. ESG scores are divided into three pillars - environmental, social, and governance - each weighted according to its significance for the specific industry group.

With the increasing volume of data and associated challenges, the evolving landscape of Machine Learning (ML) and AI offers potential solutions. Advanced techniques can address the bottlenecks in the analytical process, unlocking new avenues for methodology and topical research. While ML and LLMs have been used independently in the financial sector, there is limited research comparing their effectiveness in prediction tasks, especially in novel applications like ESG scoring. Specifically, by leveraging the capabilities of generative AI, the study seeks to understand how traditional machine learning methods measure up against LLMs, identifying strengths and weaknesses of each approach and providing insights into how these advanced technologies can support and enhance the assessment of ESG factors. This exploration is not only timely but also essential as companies and investors increasingly prioritize sustainability in their strategies. Furthermore, the findings from this study will contribute to the broader discourse on the role of generative AI in prediction tasks and analysis of quantitative data, providing insights for financial institutions looking to incorporate sustainability metrics into their assessments and AI practitioners deciding between machine learning approaches.

Aiming to explore whether the advent of AI can assist practitioners in navigating the challenges associated with ESG data and producing sound predictions of ESG scores, the analysis is framed around four main research questions (RQ):

RQ 1 – How do traditional methods and GPT-4 compare when predicting the environmental score in a noisy and high-dimensional space?

This question investigates the performance and robustness of traditional ML methods and GPT-4 in handling complex, high-dimensional, and incomplete datasets often encountered in ESG data.

RQ 2 – How do traditional methods and GPT-4 compare when predicting ESG score components in a lower-dimensional space also presenting class imbalance?

This question focuses on the effectiveness of both approaches in a simplified scenario where the data has been reduced in dimensionality and may exhibit class imbalance, a common challenge in ESG scoring.

RQ 3 – How do traditional methods and GPT-4 compare when predicting ESG scores with scarce information?

This question addresses the capability of traditional ML and GPT-4 in making accurate predictions when only limited ESG data is available, reflecting real-world situations where data may be incomplete as well as scarce, due to unavailability or high costs.

RQ 4 – How do they compare in terms of interpretability, accessibility, and other qualitative indicators?

Beyond predictive performance, this question evaluates the qualitative aspects of traditional ML methods and GPT-4, including how understandable and accessible their outputs and methods are to practitioners.

To address these questions, the thesis is structured as follows. The Introduction establishes the context, importance, and objectives of the research. The Literature Review provides an overview of existing studies on traditional ML methods applied to credit risk assessment and ESG scoring, and advancements in LLMs. The Methodology details the research design, data collection, pre-processing, implementation, and analysis methods used to compare traditional ML methods and GPT-4. The Results section presents the findings of the research, including comparative analyses of predictive performance and qualitative assessments. The Discussion interprets the results, highlighting the implications for ESG assessments and financial decision-making. Finally, the Conclusion summarizes the key findings, contributions to the field, and potential directions for future research.

2 LITERATURE REVIEW

The evolving landscape of ML and AI has significantly impacted various domains, gaining increasing attention in the financial landscape. In parallel, the growing interest towards sustainability and responsible lending has led to the integration of ESG metrics into companies' ratings and creditworthiness considerations.

As financial institutions increasingly leverage advanced technologies to enhance their decision-making processes, a diverse range of methodologies has emerged within the field of ML for corporate scoring. This literature review aims to provide a comprehensive overview of the current state of research, focusing on the application of traditional ML methods and LLMs for assessing companies' ratings.

A systematic approach has been employed to gather and analyze research spanning from 2015 to 2023, reflecting the recent interest and expansion of these topics and focusing on the most current developments in the field. The exploration encompassed the following details:

- **Electronic databases**: Scopus, Science Direct, ResearchGate and Google Scholar.
- **Search string**: "AI" OR "GPT" OR "Large Language Models" AND "Machine Learning" OR "Traditional Methods" AND "Credit Scoring" OR "Rating" OR "ESG".
- **Inclusion criteria**: studies that specifically addressed the application of ML and AI in credit scoring and those incorporating ESG metrics.
- **Exclusion criteria**: papers outside the period under analysis or with limited practical application.

This review reveals a rising interest in innovative approaches to assess a company's creditworthiness. Although traditional ML methods have undergone thorough analysis in the realm of credit scoring and, in a lower degree, of ESG rating, there exists a dearth of research on the application of LLMs to such use cases. Additionally, no studies have been identified as directly comparing the efficacy of these methodologies within this domain. While the lack of comprehensive comparative analyses highlights a research gap, the emerging relevance of sustainability metrics in credit scoring stresses the relevance of conducting new analysis in this field.

The outcome of the analysis delineates three major research trajectories. Firstly, a thorough investigation into the application of traditional ML in the broader field of credit scoring emerges as a central research pillar. Secondly, there is a discernible, yet minor, cluster of papers delving into the application of ML within the domain of ESG scoring, reflecting the recent but growing interest in embedding sustainability metrics into companies' assessments. Lastly, a smaller category involves the integration of Natural Language Processing (NLP) and LLM in the broader context of finance and analysis of ESG reports and news.

These three distinct groups not only serve as organizational pillars for the literature review but also underscore the multifaceted nature of contemporary research in the field of credit assessment and financial modelling.

2.1 TRADITIONAL MACHINE LEARNING IN CREDIT SCORING

The application of ML in credit scoring represents a substantial and cohesive area of study within the broader field of credit scoring. The latter involves the use of analytical methods to elaborate rating that signal creditworthiness, inform, and determine credit access and limits, as well as loan rates. Central to this body of work is the exploration, enhancement, and integration of various ML techniques aimed at improving the accuracy and efficacy of credit scoring models.

In this context, a vast variety of techniques has been extensively analyzed and compared. Simple linear classification models represent a widespread choice among credit institutions, largely due to their accuracy, straightforward implementation, and high interpretability (Lessmann et al., 2015).

Noteworthy is also the successful application of a broad range of ML techniques, with Decision Trees (DT), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) being the most prominent. In this scenario, DT models emerge as favorable performers in credit rating prediction, showcasing their superior accuracy compared to ANN and SVM (Golbayani et al., 2020).

Novel approaches in the context of supervised learning have also gained a growing interest, integrating elements of decision trees and logistic regression to create hybrid models designed to improve predictive performance and maintain interpretability. For example, the application of Penalized Logistic Tree Regression (PLTR) stands out for its ability to capture non-linear effects while maintaining transparency and interpretability (Dumitrescu et al., 2022).

Beyond traditional supervised learning techniques, unsupervised learning has introduced new dimensions to credit scoring methodologies. These applications demonstrate the potential of unsupervised techniques to uncover hidden patterns in credit data, allowing for a deeper understanding of the inherent relationships among data points, potentially revealing insights that might be challenging to identify through traditional supervised techniques alone (Bao et al., 2019). Next to that, ensemble models show significant performance in credit risk modelling thanks to their ability to produce more robust classifiers. These ensemble methods combine the predictions from all considered models, thereby enhancing overall predictive performance (Feng et al., 2018).

The evolving landscape of credit scoring continues to highlight the importance of balancing model accuracy and interpretability, a crucial factor in real-world financial decision-making (Florez-Lopez et al., 2015). Overall, the popularity of logistic regression is attributed to its

simplicity and transparency, while more complex ML models, including ensembles, offer enhanced performance but often at the cost of reduced interpretability. Notably, recent research underscores the potential of deep learning models in credit scoring, although their application remains relatively limited (Dastile et al., 2020). This signals a need for further exploration and application of advanced ML techniques, particularly within the realm of deep learning, to refine and enhance credit scoring models.

2.2 MACHINE LEARNING IN ESG CREDIT SCORING

The recent surge in incorporating ESG metrics into corporate risk assessment and scoring has sparked significant academic interest, with researchers increasingly exploring how ML techniques, traditionally used for financial credit rating, can be applied to non-financial factors. The industry interest progresses at the same pace as the availability of ESG scores from major rating agencies, including Bloomberg, Fitch, MSCI, S&P and Moody's. However, differences in methodology and proprietary models often make these scores opaque and difficult to interpret (Henriksson et al., 2019).

In this respect, many studies have tested the application of a broad range of statistical and ML methods to predict ESG ratings. While linear regression is an intuitive and easily interpretable method, research demonstrates that this approach might fall short in terms of performance due to the complex nature of ESG data, presenting significant gaps in terms of coverage and high noise especially in the case of smaller entities, less regulated industries, and emerging markets (Licari et al., 2023). On the other hand, further research highlights that simple models like Lasso and Ridge regularization can provide accuracy comparable to more complex models such as ANN and Random Forest (RF), despite the complexity of ESG datasets, which are often characterized by granular attributes, missing information, and variable redundancy (Del Vitto et al., 2023). Nevertheless, these linear methods are subject to overfitting in highly dimensional spaces, which may pose an issue when dealing with large datasets.

Advanced ML techniques have also been explored in this area, though to a lesser extent compared to the research conducted on corporate rating models. Techniques such as DT, RF, ANN, SVM, eXtreme Gradient Boosting (XGBoost), and Extreme Learning Machines (ELM) have demonstrated to uncover complex patterns and hidden relationships that may be difficult to identify using linear analysis methods.

Given the difficulty in accessing ESG data, an interesting research trajectory within this realm involves predicting ESG scores through ML techniques primarily by leveraging financial features. Among all, the RF algorithm registers a high prediction performance with respect to the classical regression approaches based on Generalized Linear Models (GLM), showcasing the ability to grasp the non-linear pattern of the predictors (D'Amato et al., 2021). Further results have been achieved in the application of various supervised learning techniques such as ELM, XGBoost, and SVM, which exhibit excellent performance and high accuracy when

prediction ESG ratings (Lin et al., 2023). Significant results have been also registered through the application of ensemble models constructed using Catboost, XGBoost and a feedforward neural network (Krappel et al., 2021), resulting in outperforming previous approaches.

In essence, these collective endeavors within ML in ESG scoring delineate a trajectory towards more sophisticated and comprehensive assessments integrating sustainability metrics. By harnessing advanced algorithms and adopting novel methodologies, these studies underscore the evolving trends within financial research, offering a glimpse into the potential of ML in the context of ESG scoring.

2.3 Natural Language Processing and Large Language Models in finance

Research on the application of NLP has primarily focused on extracting insights into ESG from textual information. However, to the best of our knowledge, no studies have utilized these methods to predict ESG scores employing non-textual ESG data.

One notable trend in this field involves using transformer-based language models to analyze ESG-related news data (Guo et al., 2020). Efforts towards this have seen the application of NLP, such as Bidirectional Encoder Representations from Transformers (BERT), to extract useful information from unstructured text data in social media and enhance ESG rating predictions. Despite these advancements, the precision and recall of such methods showed to be around 60%, indicating room for improvement in accurately capturing and interpreting ESG-related data from diverse text sources (Sokolov et al., 2021). Another significant study focuses on integrating NLP to address challenges such as discontinuity, imprecision, and time lag in original ratings. This research develops a multi-factor model to extract keywords from financial reports and counting their corresponding frequencies to construct percentage ESG ratings. The resulting model surpasses traditional ones in return rates, demonstrating the effectiveness of innovative methods in rating predictions and emphasizing the interpretability and predictive power of NLP models (Dong, 2023).

In the exploration of LLMs and AI within the broader financial domain, recent research has shed light on both their capabilities and limitations. One study engages with GPT-3 in a dialogue on climate finance, showcasing the model's eloquence but also highlighting its tendency for factual inaccuracies, a phenomenon known as hallucination, indicating instances where responses may lack grounding in facts (Leippold, 2023). This underscores the need for improved accuracy and reliability in LLMs. Additionally, research on financial reasoning in LLMs examined their proficiency in contributing to financial thinking and decision-making. With focus on various aspects, including task formulation, synthetic data generation, prompting methods, and evaluation capabilities, it emerged that coherent financial reasoning improves with better instruction or larger datasets (Son et al., 2023).

These studies underscore the promise of LLMs in contributing to financial discourse, whether in addressing climate finance matters or engaging in sophisticated financial analysis. However,

they also delineate the inherent limitations of these models, particularly the risk of hallucination and the need for continued refinement.

Overall, the dichotomy between traditional ML methods and LLMs unveils distinct strengths and weaknesses shaping their applicability in financial contexts. Traditional ML methods, including logistic regression and decision trees, boast simplicity, interpretability, and transparency as key strengths, providing a clear understanding of the factors influencing scoring decisions. However, simpler traditional methods exhibit limitations, as they may struggle to capture complex, non-linear relationships within data, hindering their ability to model more subtle patterns. In addition to that the risk of overfitting and the sensitivity to imbalanced datasets represent other weaknesses, as these methods might produce biased models in such scenarios.

On the other hand, LLMs unlock new opportunities, enabling the analysis of textual information in financial reports, news, and other complex narrative information, of relevance in the ESG field. This proficiency provides a more comprehensive understanding of creditworthiness. Yet, LLMs come with their set of disadvantages, as their black-box nature raises concerns about interpretability, making it challenging to understand the decision-making process within these models and, consequently, posing questions about accountability and transparency. Additionally, the risk of hallucination in LLMs is a notable concern, emphasizing the need for caution in relying solely on their outputs for critical financial decisions.

The current literature review highlights a notable interest in the intellectual community in measuring and comparing various algorithms aimed at enhancing decision-making processes in the field of ESG and credit scoring. Researchers demonstrated a keen focus on evaluating the efficacy and applicability of different algorithms to improve the accuracy and reliability of models used in financial assessments.

The present research aims to align with this trajectory of evaluating and comparing various algorithms for decision-making in ESG rating. However, it seeks to address a research gap that remains despite substantial progress in the exploration of traditional ML and LLM in this domain. While the existing literature highlighted the strengths and limitations of traditional ML and, to a smaller extent, of LLMs, it falls short in providing a comprehensive analysis of how they measure up against each other in terms of interpretability, effectiveness, and their prediction power.

Furthermore, no research has been found regarding the ability of LLMs to analyze quantitative data. Therefore, this thesis seeks to contribute to the current research landscape, by evaluating and comparing the performance of traditional ML and LLMs in the context of sustainability, thereby bridging the existing gap and offering practical insights for practitioner and researchers.

3 METHODOLOGY

This chapter delineates a comprehensive overview of the quantitative methodology employed to compare traditional ML methods with LLMs, namely GPT-4, in the context of ESG scoring. All RQs aim to compare the performance of traditional ML with the one of Generative AI. The sections that follow outline the RQs, present the data utilized, provide a detailed explanation of the study design, and cover the theoretical foundations supporting this study, as well as aspects of data processing and implementation.

3.1 THEORETICAL BACKGROUND – ALGORITHM AND LLM SELECTION

To determine which approach offers superior accuracy, reliability, and interpretability in evaluating ESG score, this study leverages on previous research demonstrating the improved performance of ensemble strategies over single ML credit scoring models. Therefore, this research evaluates the strengths, limitations, and performance benchmarks of combined ML methods vis-à-vis LLMs, rather than directly comparing LLM methods against individual traditional ML methods.

Ensemble learning, by combining multiple ML models, allows to produce enhanced predictive performance compared to a single model. This method does not only notably improve accuracy and robustness to avoid overfitting and mitigating bias and variance, but also takes advantage of complementary model features. The underlying idea is that different models may capture diverse patterns or biases in the data, and by aggregating their predictions, the ensemble model can compensate for individual model weaknesses and produce better overall performance. Furthermore, ensembles are proved to generalize better to unseen data.

Voting is the ensemble technique employed in the current study. As shown in Figure 1, this method aggregates the predictions from multiple models to reach a final estimate. Soft voting, also known as weighted voting, was deemed as suitable for the current study as it observes the probability scores of each base model for each class and combines the predictions by calculating the weighted average of these probabilities and then selecting the model with the highest weighted probability.

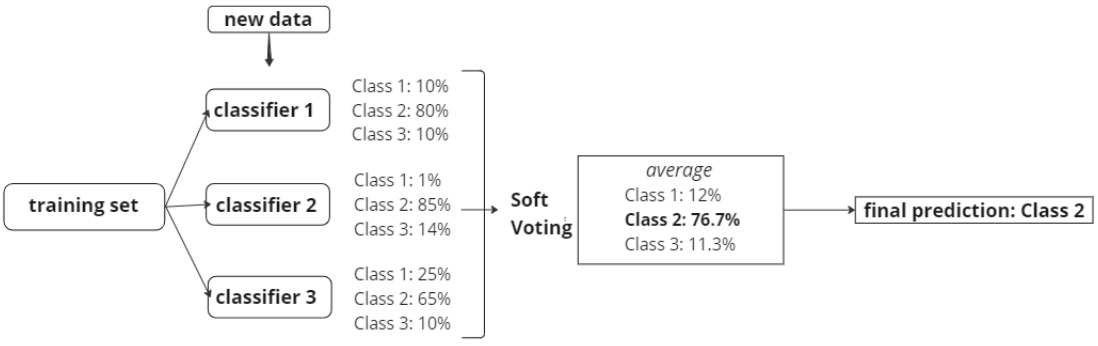

Figure 1 - Ensemble Soft Voting Prediction.

This technique has been selected as it offers several advantages, including increased robustness, improved generalization, as well as simple implementation and computation efficiency (Géron, 2019).

3.2 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm employed for both linear and non-linear classification tasks. Figure 2 illustrates the functioning of the SVM algorithm, which classifies data by delineating an optimal line or hyperplane that maximizes the distance between each class in a given space. The number of features used as input determines if the hyperplane is a line in a 2-D space or a plane in a n-dimensional space. The objective of the algorithm is to find a hyperplane that separates data points of one class from those of another class with the largest margin between the two classes, which represents the maximal width of the slab parallel to the hyperplane that has no interior data points. As SVM is a non-probabilistic model, it does not provide probabilities, whereas it outputs the distance from the hyperplane, which can be used as a measure of confidence in the prediction, and this can be transformed into a probability score using the logistic function.

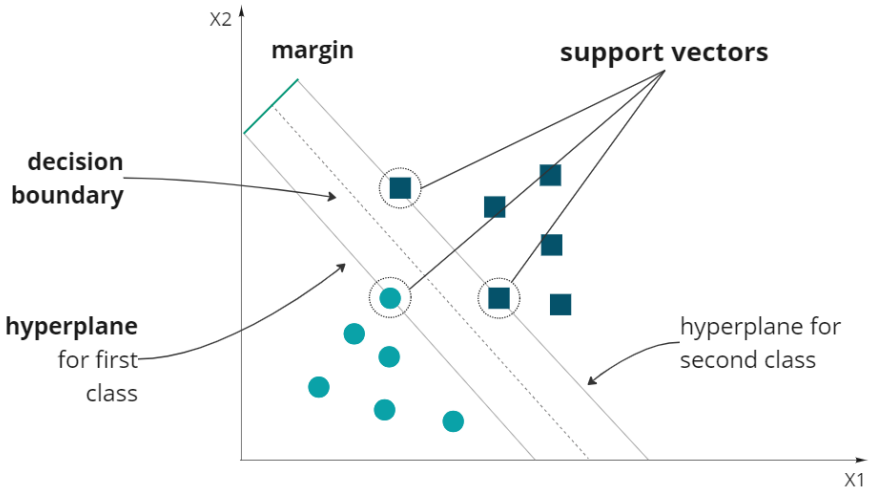

Figure 2 - Support Vector Machine Algorithm.

Overall, SVM have been selected for the study as they represent a powerful tool for predictions in contexts where the target variable and the input features unveil complex relationships (Abdullah et al., 2021).

3.3 Decision Trees

Decision Trees are a non-parametric supervised learning method used also for classification tasks. The algorithm is based on a recursive partitioning of the data into smaller subsets on the basis of the feature that is deemed as most informative. This process typically results in a tree-like structure, where each internal node of the tree represents a decision based on the value of a feature, and each leaf node represents a class label or a continuous value. In Figure 3, the process is depicted, showcasing DT hierarchical structure for making decisions. The

splitting is based on a set of splitting rules built on the classification feature and it usually works top-down, by choosing the variable that best splits the observations at each step. The partition criteria observe the feature that maximizes the information gain or minimizes the impurity, which is translated into a homogenous node and is calculated using measures such as entropy or Gini impurity.

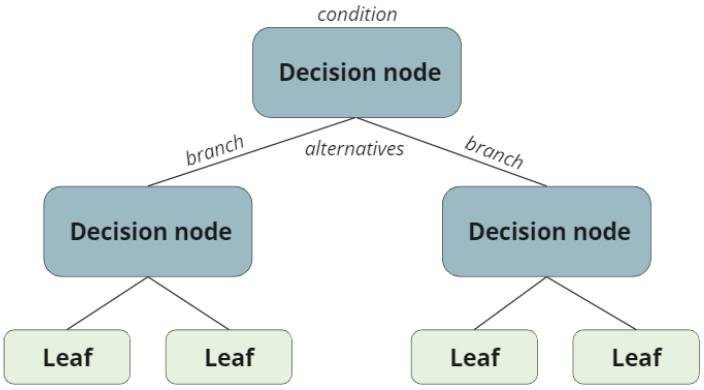

Figure 3 - Decision Tree Algorithm.

Decision trees have been selected due to their high interpretability, their capability of treating both numerical and categorical features in high dimensional spaces, their robustness against collinearity, their non-parametric nature, resulting in no need of assumptions about the shape of the underlying data distribution. However, it is important to note that they may be subject to overfitting, especially in cases in which the tree is too complex (Kotsiantis, 2013).

3.4 K-NEAREST NEIGHBORS

K-Nearest Neighbors is a non-parametric supervised learning algorithm also employed for classification problems. This method relies on a distance metric and is based on the primary assumption that similar points can be found close one another. As presented in Figure 4, it assigns a class label on the basis of the one that is most frequently represented around the k nearest neighbors of a data point. The k value is an hyperparameter and defines how many neighbors will be checked to determine the classification. This value is dependent on the nature of the variables, therefore input data with more outliers or noise is usually associated to higher values of k, which though can also lead to over-smoothing and poorer performance on test data. On the other hand, a smaller value of k can capture local structure better but may be sensitive to noise and outliers. Different distance metrics can be selected to run the algorithm, with the Euclidean distance being the most common method employed.

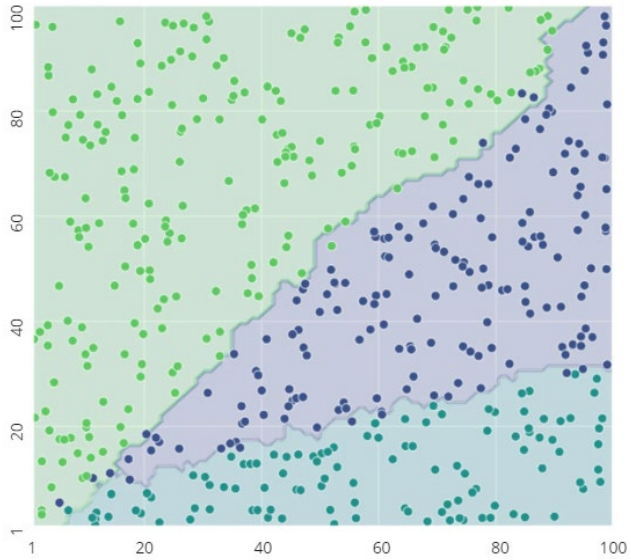

Figure 4 - K-Nearest Neighbors Algorithm. (Lanza, 2024)

As the other ML models outlined in the present chapter, the KNN has been selected for its simplicity and high interpretability. Additionally, being it a so-called lazy learning algorithm, it does not require an explicit training phase. Nevertheless, KNN may exhibit computational inefficiency with large datasets, as it needs to calculate the distance between a new data point and the entire training set, it can become slow as the size of the training data grows (Deng, 2016).

3.4.1 LARGE LANGUAGE MODEL VIA API (GPT-4)

LLMs are deep learning algorithms that use the transformer-based architecture, which converts text into numerical representations, also known as tokens, which are then converted into vectors. Through a multi-head attention mechanism, at each layer, tokens are contextualized within the scope of their context window, allowing the signal for key tokens to be amplified and less important tokens to be diminished. In other words, the transformer encompasses multiple layers of self-attention mechanisms, which enables the model to weigh the significance of each token in the input sequence concerning others, thereby capturing intricate linguistic patterns and semantic relationships. The model generates text by maximizing the likelihood of predicting the correct next token in each sequence (maximum likelihood estimation - MLE), following the probability distribution learned from the data it is trained on. This structure is exemplified in Figure 5 and lays the ground to the family of NLP tasks, which enables computer programs to understand and generate natural language autonomously (Haleem et al., 2022).

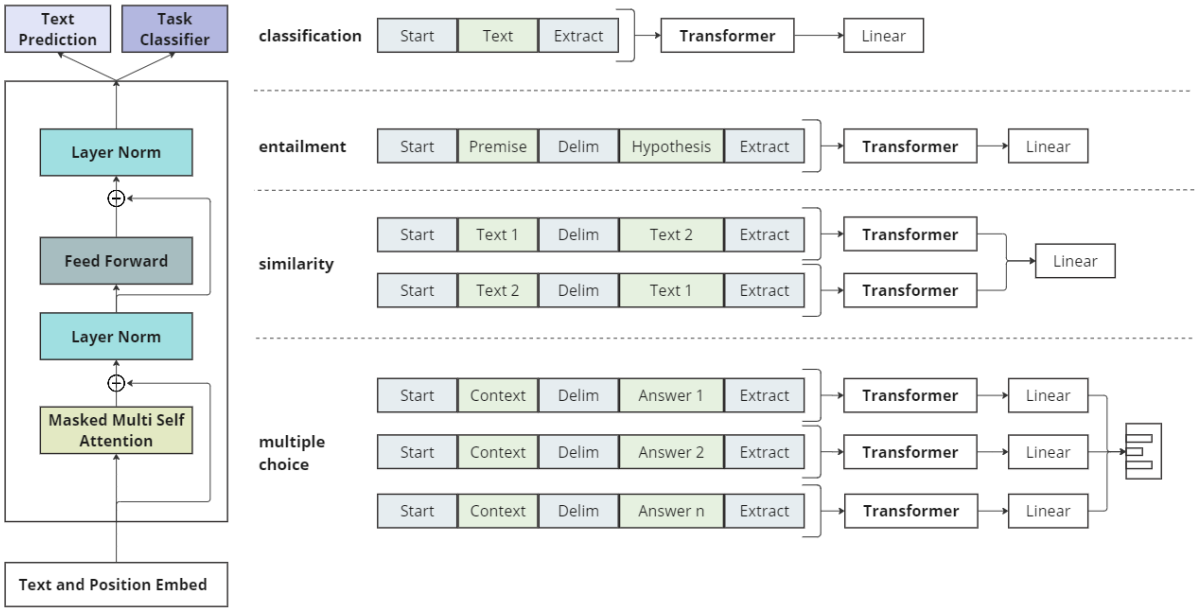

Figure 5 - GPT Model Architecture.

In this context, Generative Pre-trained Transformer 4 (GPT-4) is a large multimodal model created by OpenAI, accepting both text and image data as input. Before being released the model undergoes a pre-training phase, conducted by the provider, on massive sets of text, using both publicly available data and data licensed from third-party providers. After the deployment, the model is fine-tuned via reinforcement learning from human feedback and prompts. Specifically, GPTs can be refined on specific tasks, such as ESG rating prediction in this study, adjusting its parameters to optimize performance on the target task.

3.5 DATA

The dataset employed in this study is sourced from two primary providers: Urgentem, an independent entity specializing in carbon emission data and climate risk analytics for the finance sector, and Moody's, a leading global rating agency that provides in-depth insights into company-level ESG risks. Together, these sources contribute to a dataset that includes detailed information on more than 1,400 entities, covering over 200 specific ESG indicators. These indicators cover a wide array of critical areas, including corporate governance, diversity and inclusion, carbon emissions and intensity, impact metrics, and ethical standards.

The ESG ratings are taken from Moody’s, which supplies both overall scores aggregating all components and individual issuer profile scores for each ESG category. These scores are utilized as the output variables in the study’s model. The input features of the dataset are categorized into two types: granular and aggregated. The granular data, procured from the Urgentem database, encompass detailed measurements of carbon emissions, tracking of carbon intensity, and specific metrics evaluating environmental impacts. This level of detail allows for an exhaustive analysis of each entity's environmental performance and its overall

contribution to climate risk. Additionally, Moody's provides issuer category scores for the E, S, and G dimensions, which enrich the analysis by offering a detailed view of performance in these specific areas.

The structure and composition of the dataset are depicted in Figure 6, which illustrates the depth and breadth of the information gathered, outlining how both the granular and aggregated data come together to form a comprehensive dataset. This structured approach to data collection is essential for analyzing the complex interplay of ESG factors and for assessing how different data layers impact the performance of financial models in real-world scenarios. The full list of variables used in the study can be consulted in the Appendix A.

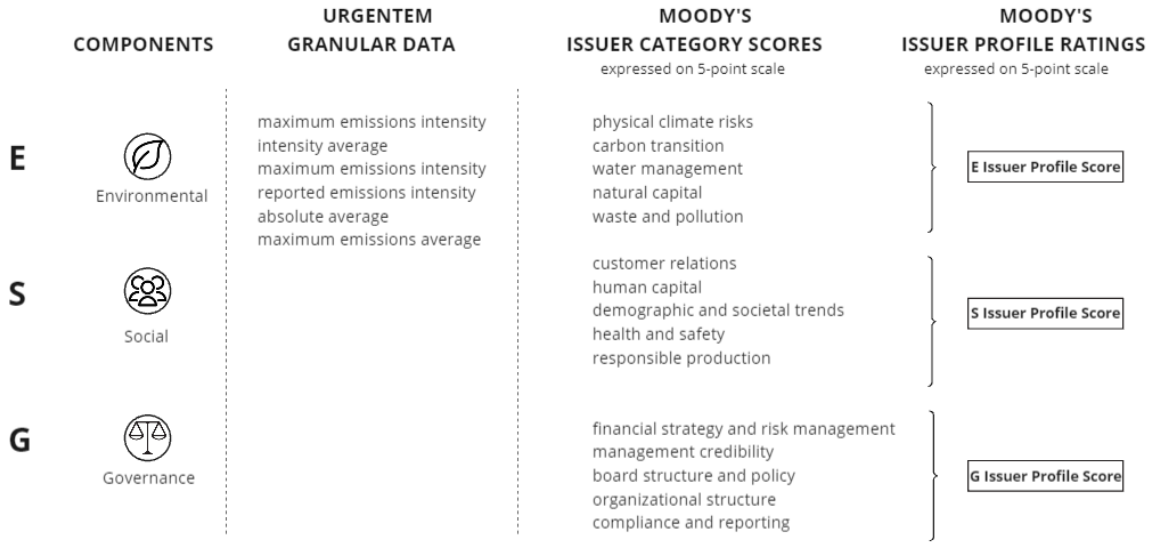

Figure 6 - Dataset composition.

3.6 RESEARCH QUESTIONS

The models presented in the previous section have been tested on the basis of three main research questions. Each RQ outlined in the following paragraph is designed to test how the models handle common challenges and complexities inherent to the data. Specifically, as shown in Figure 6, they delineate a number of classification problems in the realm of ESG rating under diverse conditions, such as varying levels of dimensionality, granularity, and noise, in order to investigate the behavior and performance of traditional ML approaches vis-á-vis LLMs under distinct conditions.

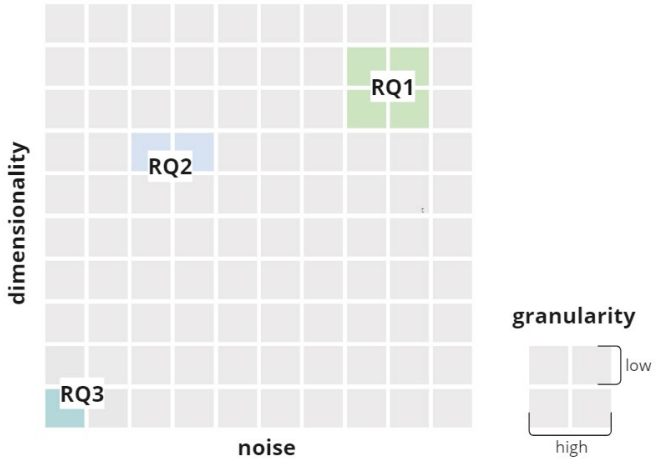

Figure 7 - Visualization of Study Design Research Questions.

3.6.1 RQ 1 – How do traditional methods and GPT-4 compare when predicting the environmental score in a noisy and high dimensional space?

- **Aim**: the objective of this RQ is the prediction of the environmental score, namely the environmental component of the ESG rating.
- **Data**: in this RQ, a highly complex dataset with substantial dimensionality, granularity, and noise is employed. The dataset's complexity is further emphasized by its diverse units of measurement, as well as the presence of numerous missing values, requiring advanced data pre-processing techniques.

  Key variables related to carbon emissions include:

  - **Intensity average**: the ratio of total greenhouse gas emissions (in CO2 equivalents) to economic output, standardizing emissions data across companies.
  - **Maximum emissions intensity**: the highest recorded emissions intensity over a specific period.
  - **Reported emissions intensity**: emissions intensity as reported by companies based on their activities.
  - **Absolute average**: the average of absolute emissions values reported over a period.
  - **Maximum emissions average**: the highest average of absolute emissions recorded over a specific timeframe.

  These variables provide a comprehensive view of environmental performance across different scopes, including Scope 1, 2, and 3 emissions[1].

[1] Scope 1 emissions refer to direct greenhouse gas (GHG) emissions from sources owned or controlled by the company, such as fuel combustion in company vehicles. Scope 2 emissions are indirect GHG emissions from the generation of purchased electricity, steam, heating, and cooling consumed by the company. Scope 3 emissions include all other indirect GHG emissions that occur in the value chain of the company, such as those from business travel, waste disposal, and product use.

3.6.2 RQ 2 – How do traditional methods and GPT-4 compare when predicting ESG score components in a lower dimensional space also presenting class imbalance?

- **Aim**: the objective of this RQ is to predict Environmental (E), Social (S), and Governance (G) scores separately, as well as the overall ESG rating.
- **Data**: this RQ tests the models using a dataset with lower dimensionality and aggregated data. The target variables here are predicted using intermediate indicators (issuer profile scores) scored by Moody's, whose values range from 1 to 5, as exemplified by Figure 6. In this scenario, the data reveals class imbalance, particularly in the governance score, which exhibit the highest level of disproportion.

  The indicators cover a wide range of dimensions, including:

  - **Environmental**: physical climate risks, carbon transition, water management, natural capital, waste, and pollution.
  - **Social**: customer relations, human capital, demographic and societal trends, health and safety, responsible production.
  - **Governance**: financial strategy and risk management, management credibility, board structure, organizational structure, compliance, and reporting.

3.6.3 RQ 3 – How do traditional methods and GPT-4 compare when predicting ESG score with scarce information?

- **Aim**: the objective of this task is to predict the overall ESG rating using partial E, S, G scores
- **Data**: in the final RQ, the target variable is estimated using intermediate E, S, and G scores, instead of direct indicators. Specifically, the overall ESG score is predicted using just three data points: the E, S, and G scores provided by Moody's.

Each RQ aims at providing insights into how each approach manages the complexities inherent in ESG data.

3.7 Study design

The analysis was structured in the following four main phases, represented in the diagram in Figure 8:

1. Phase I – Data collection and preparation
2. Phase II – Modelling
3. Phase III – Validation

4. Phase IV – Comparative Analysis and Evaluation

These phases were adapted from the Design Science Research (DSR) methodology, which typically outlines a structured approach to developing and evaluating innovative solutions to practical problems. In this adaptation, the steps were tailored to fit the specific objectives and context of the research, while DSR traditionally includes six steps.

Furthermore, the research methodology employed in this study followed a dual approach.

- **Manual application**: this refers to the execution of the analysis using traditional methods and tools, primarily using the Python programming language. In this approach, the entire analytical process, from data preparation to model development and evaluation, was performed manually.
- **Automated analysis via GPT**: parallel to the manual approach, the same analysis was carried out by prompting GPT-4. This involved interacting with the GPT-4 model, guiding it through the same stages of analysis as the manual application, and using its responses to perform the necessary tasks. Being the testing conducted from March to July 2024, GPT-4, launched by OpenAI on March 13, was used.

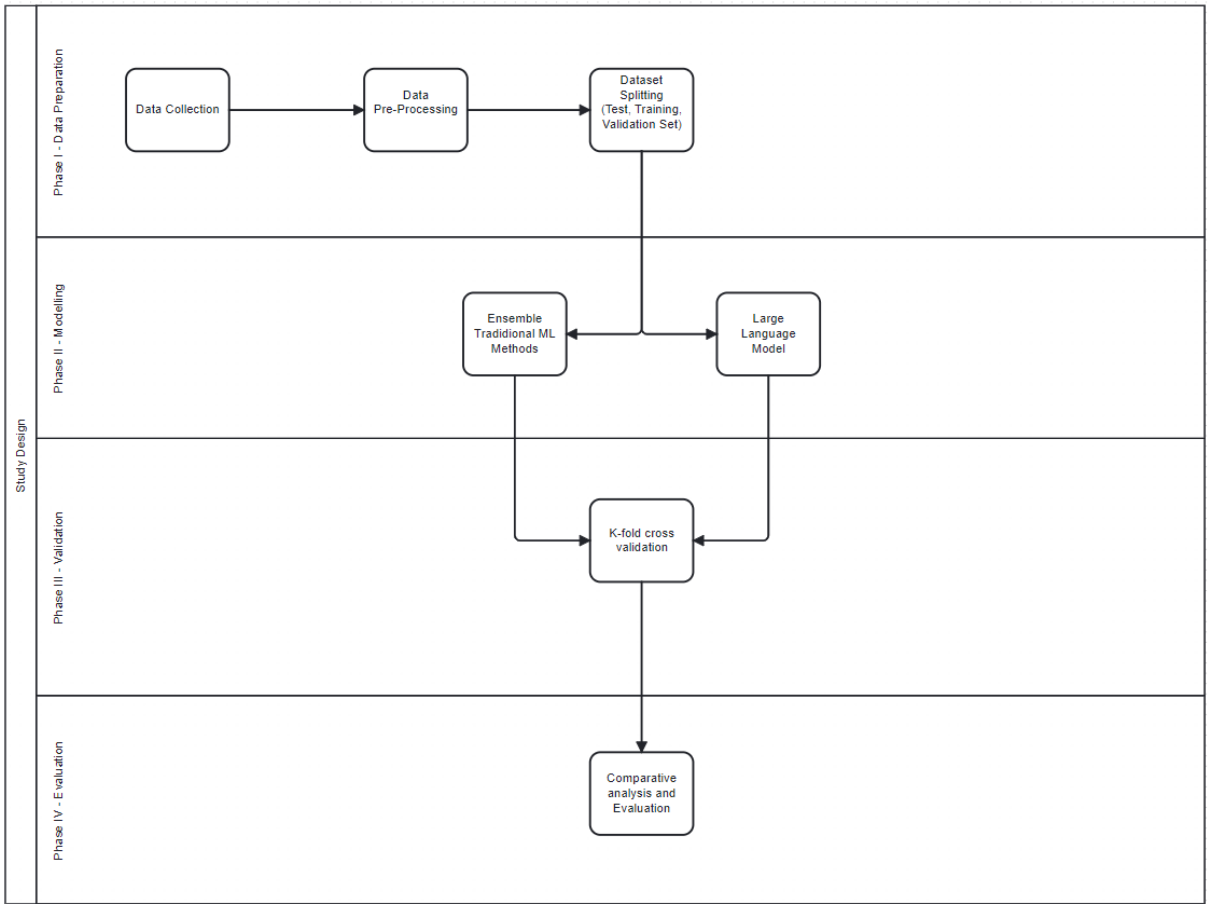

Figure 8 - Study design.

These phases are explained in the following sections. While the manual approach is described in detail, outlining the framework and specific steps taken, for the process involving GPT-4, only a brief overview is provided. Although the initial prompts adhere to the same methodological framework, the subsequent iterative steps - such as how GPT-4 handled pre-processing and modeling - are considered part of the implementation due to the flexibility given to the model.

3.7.1 PHASE I – DATA COLLECTION AND PREPARATION

Figure 9 - Phase I, Data Collection and Preparation.

The data collection step is described in detail in Paragraph 3.5.

3.7.1.1 MANUAL DATA PRE-PROCESSING

Given the complex and high-dimensional nature of the dataset presented in RQ 1, it required a meticulous approach to data preprocessing and dimensionality reduction.

A critical aspect of preprocessing involved addressing zero and missing values. Columns with a high proportion of zero values were evaluated; those exceeding a predefined threshold (i.e., 30%) were discarded as they likely lacked informative value. For other columns, zero values were replaced with missing values (NaNs) to prepare for imputation. This approach acknowledges that zeros in certain contexts may represent missing information rather than actual measurements.

Missing numerical values were imputed using the K-Nearest Neighbors Imputer. This method fills in missing data points by averaging the values of the nearest neighbors in the dataset. This technique preserves the inherent relationships and structure of the data, offering a more robust alternative to simpler imputation methods like mean or median imputation. Numerical features were standardized using the Standard Scaler, which transforms the data to have a mean of zero and a standard deviation of one. Standardization is essential to ensure that all features contribute equally to the analysis, especially in distance-based algorithms, which are sensitive to the scale of input features.

Categorical variables were transformed into binary vectors through one-hot encoding. This step converts categorical data into a numerical format that can be effectively utilized by most machine learning algorithms, ensuring that categorical differences are accurately represented in the analysis. Given the high dimensionality of the dataset, Principal Component Analysis (PCA) was employed to reduce the number of features while retaining 95% of the variance in the data. PCA transforms the data into a new set of orthogonal components, simplifying the

dataset without significantly compromising the informational content. This reduction mitigates the curse of dimensionality and enhances computational efficiency, making the subsequent analysis more manageable and interpretable. An overview of these key variables can be found in Appendix A, providing insight into the specific attributes pivotal for the ESG scoring model.

3.7.1.2 GPT-4 DATA PRE-PROCESSING

In this study, GPT-4's ability to handle data pre-processing tasks was evaluated with a high degree of flexibility. The model was provided with a raw dataset in CSV format from Urgentem, along with an additional dataset containing the full descriptive names of the columns. This supplementary dataset was necessary as the primary dataset used coded column names. Furthermore, a brief indication regarding the content of the dataset has been provided to guide the model in further understanding the data and generating relevant responses.

> *Task Description: the dataset I provided you as input is focused on ESG. Refer to the attached file "Urgentem Variable names and codes (Emissions dataset)" for a detailed description of all variables.*

Details on GPT-4 performance can be found under the section named GPT-4 Pre-processing.

3.7.2 PHASE II – MODELLING

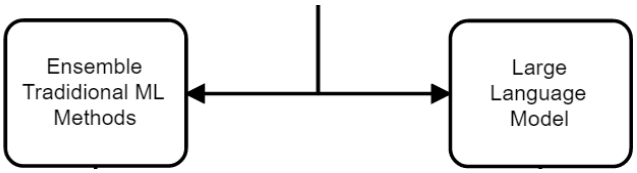

Figure 10 - Phase II, Modelling.

As depicted in Figure 10, Phase II involves the modelling techniques applied to the datasets Following the literature review results, two different combinations of traditional ML models have been compared to the application of a LLM via API.

3.7.2.1 ENSEMBLE MODELLING

1. Decision Trees and K-Nearest Neighbors (DT-KNN): the first ensemble model adopts Decision Trees and K-Nearest Neighbors as base learners. As presented above, DT and KNN were selected as base learners due to their simplicity, interpretability, and complementary characteristics. In this respect, the DT well capture non-linear relationships among features, while the KNN is effective in extrapolating local patterns and outliers. This combination aims to strike a balance between bias and variance, potentially leading to more stable predictions.
2. Decision Trees, K-Nearest Neighbors, and Support Vector Machines (DT-KNN-SVM): the second ensemble model includes SVM as an additional base learner alongside DT and

KNN, to incorporate a powerful model capable of capturing complex relationships in the data. SVMs excel in handling high-dimensional data and non-linear decision boundaries, complementing the capabilities of DT and KNN and potentially improving its overall generalization abilities. The training and prediction steps were similar to the DT-KNN approach, but this time, the ensemble included predictions from SVM as well. The Meta Model LR was used to combine the predictions of the base models, learning the optimal weights to assign to each base model's prediction to create a final ensemble prediction.

3.7.2.2 LARGE LANGUAGE MODEL VIA API (GPT-4)

For the LLM, a distinct approach was pursued. To further enhance GPT-4's capabilities for predicting ESG ratings, the model was asked to execute a task based on the dataset given as input.

The model was prompted to execute a task based on a provided dataset with the following instructions:

> *Task Description: the dataset I provided you as input is focused on ESG ratings, and I need your assistance in developing a predictive model to classify the environmental score[2]. Here is how you should proceed:*
>
> *First, divide the dataset into three subsets: training, validation, and testing set. Ensure that the data is randomized before splitting to prevent any potential bias. Select the variables you deem more appropriate for the task.*
>
> *Following this, develop a model, by choosing any algorithm(s) you find suitable and provide a justification for your choice. After developing the model, validate it using the validation set. This step includes tuning hyperparameters and evaluating the model's performance on the validation subset. Throughout the process, at every significant step, provide a Python (.py) file containing the detailed operations performed.*

Details on how GPT-4 handled the task can be found under the section named GPT-4 Model modelling.

3.7.3 PHASE III – VALIDATION

In Phase III, the predictive performance of the selected machine learning models was validated.

3.7.3.1 ENSEMBLE MODELLING VALIDATION

The ensemble models underwent validation using a robust and widely embraced technique known as K-fold cross-validation (CV). This method ensures a thorough evaluation of the

[2] This prompt has been used for the RQ 1. The same structure has been used for the other RQs, but it has been tailored based on the output variable.

models' performance while mitigating potential biases. K-Fold CV was chosen for its numerous advantages, including its robustness in providing a more accurate model performance estimate, its efficient use of data, its ability to avoid overfitting, and its facilitation of hyperparameter tuning.

To commence the validation process, the dataset underwent a shuffling procedure aimed at reducing any inherent biases. Subsequently, the dataset was partitioned into K subsets or "folds" with each one rotating as the validation set while the remaining ones served as the training data. This process iterated K times, ensuring each subset could serve as both training and validation data.

During each iteration of the K-fold CV process, a suite of performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were calculated to gain comprehensive insights into the models' predictive ability across various evaluation criteria. Upon completion of all K iterations, the performance metrics obtained from each fold were aggregated and averaged. This combination allowed for the estimation of the ensemble models' expected performance on unseen data, providing a robust indication of their predictive efficacy and generalization capability.

In view of optimizing the model performance, hyperparameter have been tuned as it follows. Hyperparameters are external to the model and cannot be learned from the training data. Instead, they must be set before training begins. Properly tuning these parameters can significantly improve model accuracy and generalization. The hyperparameters for DT, KNN, and SVM models were carefully chosen for this study to explore a wide range of combinations and ensure optimal performance.

Decision Tree

The Decision Tree algorithm has several hyperparameters that influence its performance. The following parameters shown in Table 1 were selected for tuning:

- **max_depth**: controls the maximum depth of the tree, effectively limiting how many splits can be made. This parameter helps in balancing model complexity and overfitting.
- **min_samples_split**: specifies the minimum number of samples required to split an internal node. This prevents the model from making splits based on a very small number of samples, thus reducing overfitting.

Table 1 - DT Hyperparameters Selection

Parameter	Values	Description
max_dept	None, 10, 20	None: no limit; 10, 20: restrict the tree depth.
min_samples_split	2, 5, 10	Minimum samples needed to split a node.

k-Nearest Neighbors

KNN's primary hyperparameter is the number of neighbors (n_neighbors). This determines how many nearest neighbors are considered when classifying a data point. The parameters selected for the hyperparameter tuning for the KNN algorithm are displayed in Table 2.

Table 2 - KNN Hyperparameters Selection

Parameter	Values	Description
n_neighbors	3, 5, 7	Number of neighbors to use for classification.

Support Vector Machine

SVM models have several important hyperparameters that affect their performance. The parameters chosen for tuning are shown in Table 3 and include:

- **C**: the regularization parameter. It controls the trade-off between achieving a low training error and a low testing error.
- **gamma**: defines the influence range of a single training example. It impacts the decision boundary shape in the model.

Table 3 - SVM Hyperparameter Selection

Parameter	Values	Description
C	0.1, 1, 10	Regularization parameter: lower values prioritize simpler models.
gamma	scale, auto	scale: Uses 1 / (number of features * variance of features);
		auto: Uses 1 / number of features.

The parameters were selected to provide a comprehensive exploration of each model's potential configurations, ensuring that both simple and complex models were considered:

- **Decision Tree**: the max_depth and min_samples_split parameters help control the tree's complexity. Limiting tree depth and requiring more samples for splits can prevent overfitting, thus improving generalization.
- **k-Nearest Neighbors**: the choice of neighbors affects the model's sensitivity to local variations. Fewer neighbors can capture detailed patterns but may overfit, while more neighbors can generalize better but might miss finer details.
- **Support Vector Machine**: the C parameter balances the trade-off between model complexity and error minimization. The gamma parameter influences how the model

interprets the data's complexity, with lower values allowing for a more global view and higher values focusing on local structures.

The selection balances the risk of overfitting with the need for accurate and generalized predictions. The tables above summarize the specific values explored during the tuning process.

3.7.3.2 LLM VALIDATION

Since it is a pre-trained LLM, GPT-4 underwent extensive validation during its training process by OpenAI. However, when the LLM is fine-tuned for a specific task, such as predicting ESG ratings, additional validation steps are necessary to ensure the effectiveness of the model employed in the context of the new task. Details on how GPT-4 handled this task can be found under the section called GPT-4 Model modelling.

3.7.4 PHASE IV – COMPARATIVE ANALYSIS AND EVALUATION

Once the model was validated, a comprehensive comparative analysis and evaluation of the models' performance were conducted, employing a range of standardized metrics and criteria.

3.7.4.1 QUANTITATIVE INDICATORS

Several key performance metrics were utilized to assess the models' predictive capabilities. These metrics offered insights into various aspects of models' performance.

Accuracy: it measures the proportion of correctly classified instances to all instances, providing an overall assessment of predictive correctness.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Instances} \quad (1)$$

Recall: also known as sensitivity, it evaluates the models' ability to correctly identify positive samples, crucial for tasks where identifying all positive instances is imperative.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

Precision: it quantifies the proportion of instances classified as positive that are truly positive, offering insights into the model's precision in positive predictions.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

F1-score: representing the harmonic mean of precision and recall, the F1-score captures the trade-off between these two metrics, offering a balanced assessment of models' performance.

$$\mathrm{F1} - Score = \ 2 * \frac{Precision \ * \ Recall}{Precision + Recall} \quad (4)$$

ROC-AUC: the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) assesses the models' ability to differentiate between classes by plotting the true positive rate against the false positive rate across different threshold values, with a higher AUC indicating superior performance.

$$ROC - AUC = Area\ Under\ ROC\ Curve \quad (5)$$

3.7.4.2 QUALITATIVE INDICATORS

Next to quantitative metrics, a list of qualitative indicators was utilized to assess other important features an optimal model should have.

- **Interpretability**: it refers to the ability of a model to explain its predictions or decisions in a human-understandable way. Interpretability can help in building trust with stakeholders, providing insights into how the model works, and identifying potential biases or errors.
- **Scalability**: it measures how the performance of a model changes as the size of the data or the complexity of the task increases. It is important for a model to be able to handle large datasets and complex tasks efficiently. This can involve considerations such as processing time, memory usage, and computational resources required.
- **Usability**: it focuses on how easy it is to use and deploy the model. For instance, a model that requires extensive pre-processing or complex feature engineering might be less usable than one that can handle raw data more effectively. Usability also involves considerations about how user-friendly the model is, both for developers implementing the model and for end-users who interact with it.
- **Robustness**: it refers to a model's ability to maintain its performance under different conditions, such as noisy or adversarial inputs, or changes in the underlying distribution of the data. It's important for a model to be robust to ensure reliable performance in real-world scenarios.

3.8 DATA PRE-PROCESSING

3.8.1 MANUAL PREPROCESSING

As described in the methodology chapter, in the manual analysis the data preprocessing involved the following key steps.

- **Treatment of missing values**: given the high noise within the dataset, NAs have been addressed by applying the KNN Imputer. This method averaged the values of the nearest neighbors to impute missing data points, preserving the inherent relationships in the dataset.
- **Standardization**: numerical features were standardized using the Standard Scaler, ensuring that all variables had a mean of zero and a standard deviation of one, which is crucial for the performance of distance-based algorithms.
- **One-Hot encoding**: categorical variables were transformed into binary vectors through one-hot encoding, making them suitable for use in machine learning models.
- **Dimensionality reduction**: PCA was employed to reduce the number of features while retaining 95% of the data's variance, mitigating the high dimensionality and enhancing computational efficiency.

3.8.2 GPT-4 PRE-PROCESSING

As GPT-4 had a high degree of autonomy in deciding the steps to take due to the flexible prompting approach, the preprocessing techniques it employed varied slightly between different runs. Nevertheless, the results presented in this section represent the most common approach selected for completing the task.

- **Treatment of missing values**: after loading the raw dataset and examining variable descriptions, GPT-4 initially applied mean imputation for numerical features and mode imputation for categorical features. While simple, this approach led to high approximations in columns with numerous missing values, introducing risks such as bias and distortion in the data due to artificial homogeneity, which not only increases the risk of overfitting, but also the risk of missing the true complexity and variance of the underlying data. To refine this approach, the model was prompted to reconsider its initial imputation strategy. When asked for confirmation, GPT-4 was guided to use K-Nearest Neighbors imputation, thereby reducing the likelihood of introducing theoretical issues and ensuring comparability of results across traditional methods, where KNN was adopted for handling missing values.
- **Feature selection**: in terms of feature selection, GPT-4 displayed varied behaviors across different iterations. Often, it chose to use the entire set of variables without performing any dimensionality reduction techniques. In fact, by applying the Random Forest algorithm, as explained in more details in the next paragraph, GPT-4 allowed the model to inherently rank features based on their importance by measuring the reduction in impurity each feature provided across the trees. The random feature selection process within each tree helped to mitigate the effects of high dimensionality

and reduced the likelihood of irrelevant features dominating the model. When guided to apply its domain knowledge, GPT-4 sometimes selected a subset of features based on their correlation matrix and their expected contribution to the prediction of the rating. When selecting the variables for the computation of E, S and G score respectively, under RQ 2, GPT-4 autonomously selected the variables for each prediction by initially assessing their correlations and identifying those most relevant to the component based on their names.

3.9 IMPLEMENTATION - MODEL SELECTION, TRAINING, AND EVALUATION

3.9.1 MANUAL MODELLING

In the manual approach, traditional machine learning models were trained and evaluated as follows.

- **Model selection**: based on the field research conducted beforehand, the study employed SVM, DT and KNN as the base models. Additionally, ensemble learning techniques were applied to combine these models using soft voting.
- **Hyperparameter tuning**: hyperparameters for each model were tuned using grid search methods to optimize performance. In view of optimizing the model performance, hyperparameter have been tuned following the configurations presented in Table 1, Table 2 and Table 3.
- **Validation**: K-fold CV was used to validate the models. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were computed for drawing the final insights on the performance.

3.9.2 GPT-4 MODELLING

As for the data processing phase, GPT-4 had a high degree of autonomy in deciding which model to adopt for the prediction in each RQ, the preprocessing techniques it employed varied slightly between different runs. Nevertheless, the results presented in this section represent the most common approach selected for completing the task.

- **Model selection**: GPT-4 proposed using a classification model, suggesting options like Random Forest, Gradient Boosting, and XGBoost. However, it frequently recommended the Random Forest model for this task. Consequently, this model was chosen to evaluate GPT-4's performance. As the models selected for the manual analysis, the RF is a well-known ensemble learning technique, that constructs multiple decision trees during training and outputs the average or majority vote of their predictions for classification. Random Forest uses bagging, where each tree in the forest is trained on a random subset of the training data. This means each tree sees a different portion of the data, which helps reduce overfitting. The reasoning provided by GPT-4 for this choice revolves around the fact that Random Forest is known for its robustness and versatility, performing well across a wide range of tasks. It effectively

handles both categorical and numerical features once they are encoded, making it adaptable to various types of data. Additionally, compared to more complex models like neural networks, RF is relatively easy to train and tune. It also provides strong performance with minimal initial tuning, which is beneficial for early evaluations. Given these strengths, Random Forest often serves as a reliable baseline and is frequently the final model selection for many predictive tasks.

- **Hyperparameter tuning**: GPT-4 initially defined a parameter grid for the RF model, which included the following hyperparameters and their respective values:

Hyperparameter	**Values**	**Description**
n_estimators	[100, 200, 300]	Number of trees in the forest.
max_depth	[10, 20, None]	Maximum depth of each tree. None means nodes are expanded until all leaves are pure.
min_samples_split	[2, 5, 10]	Minimum number of samples required to split an internal node.
min_samples_leaf	[1, 2, 4]	Minimum number of samples required to be at a leaf node.
bootstrap	[True, False]	Whether bootstrap samples are used when building trees.

- **Validation**: this grid covers a broad range of potential configurations, from less complex models (fewer trees, shallow depths) to more complex ones (more trees, deeper depths). A 3-fold cross-validation was adopted, splitting the dataset into three parts and the model was trained and validated three times, each time using a different part of the data for validation and the remaining parts for training.

4 RESULTS

This section provides a comprehensive overview of the main outcomes of applying two different approaches, traditional ML methods versus LLM (GPT-4), for the prediction of ESG scores under the conditions presented in the RQs, illustrating the results obtained for each RQ.

4.1 QUANTITATIVE PERFORMANCE COMPARISON

4.1.1 RQ 1 - PREDICTION OF ENVIRONMENTAL SCORE

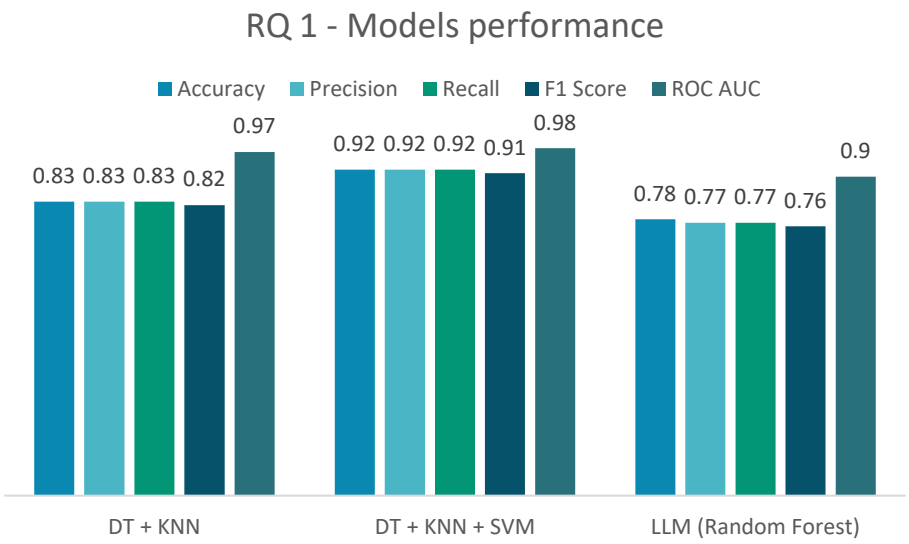

Figure 111 - Visualization of Models' Performance, RQ1.

The performance of different models for RQ1 is compared in Figure 11. When predicting the environmental score under the first RQ, the DT + KNN ensemble achieves consistent results across various metrics, each approximating 83%. This model demonstrates good generalization capabilities, although it encounters some difficulties in balancing true positives against false positives, as reflected by an F1 score of 82%. The integration of SVM substantially enhances performance, achieving 92% in accuracy, precision, and recall, alongside a 91% F1 score, indicative of effective noise management and robustness. Both ensemble models exhibit high ROC AUC values, exceeding 97%, which underscores their proficiency in handling the dataset. The LLM (RF), on the other hand, exhibits lower performance, with 78% accuracy and 77% in both precision and recall. The F1 score is similarly inferior at 76%. Nevertheless, a robust ROC AUC score of 90% indicates compelling capabilities in class distinction across various thresholds.

Table 4 - Models Results, Q1

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.83	0.83	0.83	0.82	0.97
DT + KNN + SVM	0.92	0.92	0.92	0.91	0.98
LLM (RF)	0.78	0.77	0.77	0.76	0.90

4.1.2 RQ 2 – Overview of predictions

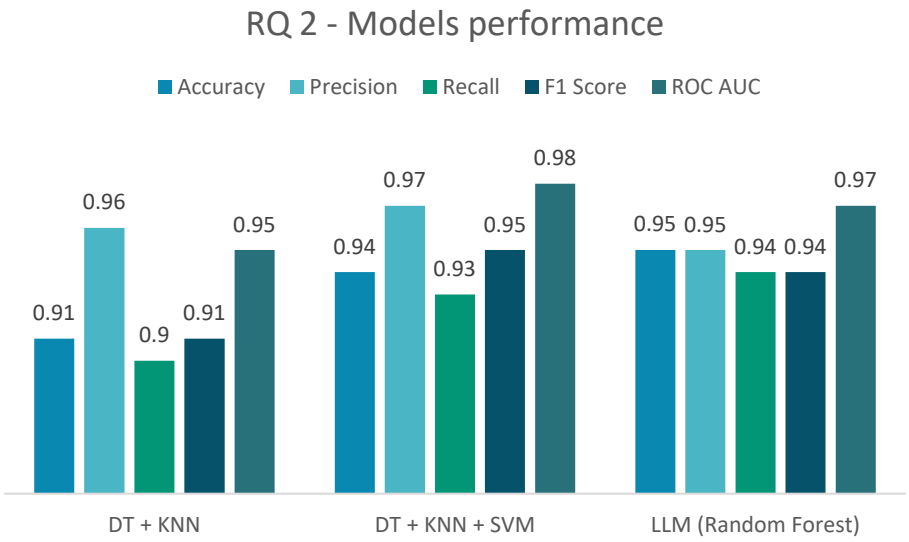

Figure 122 - Visualization of Average Models' Performance, RQ2.

Figure 12 illustrates the comparative performance of models for RQ2. Under the second RQ, DT + KNN model, while showing solid precision at 96%, has the lowest recall and accuracy among the three. The DT + KNN + SVM model presents higher values of the analyzed metrics, especially in precision and recall, respectively at 97% and 93% and a high value of ROC AUC at 98%, indicating enhanced class distinction capability. The LLM (RF) model demonstrates the highest overall performance with an average accuracy of 95% and robust balance across all metrics.

Table 5 – Average of Models Results, Q2

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.91	0.96	0.90	0.91	0.95
DT + KNN + SVM	0.94	0.97	0.93	0.95	0.98
LLM (RF)	0.95	0.95	0.94	0.94	0.97

4.1.2.1 RQ 2 - PREDICTION OF ENVIRONMENTAL SCORE

As it can be noted from Table 6, in the prediction of the environmental score, both the traditional machine learning approaches (DT + KNN and DT + KNN + SVM) and the LLM (RF) model demonstrate excellent performance. The DT + KNN model achieves high metrics across the board, achieving higher results when embedding the SVM. This increase suggests that adding SVM enhances the model's ability to generalize and effectively manage noise within the data. Meanwhile, the LLM (RF) model equals the performance of the DT + KNN + SVM model with identical scores.

Table 6 - Models Results, Q2 - environmental score

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.97	0.98	0.96	0.97	0.97
DT + KNN + SVM	0.99	0.99	0.99	0.99	0.99
LLM (RF)	0.99	0.99	0.99	0.99	0.99

4.1.2.2 RQ 2 - PREDICTION OF SOCIAL SCORE

Table 7 - Models Results, Q2 - social score

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.96	0.98	0.89	0.90	0.95
DT + KNN + SVM	0.98	0.99	0.96	0.97	0.99
LLM (RF)	0.98	0.98	0.98	0.98	0.99

As shown in Table 7, for predicting the social score, the performance of traditional machine learning models, especially the DT + KNN + SVM, and the LLM (Random Forest) is comparable. The DT + KNN model, while demonstrating high accuracy of 96% and precision of 98%, falls short in recall (89%) compared to its precision, indicating that it misses some true positive instances, which results in a lower F1 score of 90%. The addition of SVM to the DT + KNN significantly enhances the recall to 96%, better aligning it with its precision and yielding a more balanced and higher F1 score of 97%. The LLM (RF) model shows also high value across the metrics, with accuracy, precision, recall, and F1 score all at 98%, showcasing its robustness in efficiently predicting the social score.

4.1.2.3 RQ 2 - PREDICTION OF GOVERNANCE SCORE

Table 8 - Models Results, Q2 - governance score

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.91	0.99	0.90	0.92	0.97
DT + KNN + SVM	0.94	0.96	0.92	0.94	0.99
LLM (RF)	0.93	0.93	0.89	0.91	0.95

Table 8 shows that in predicting the governance score, the DT + KNN model, despite its high precision of 99%, displays a lower recall of 90%, which results in an F1 score of 92%. Adding SVM to the model improves the overall balance, with both recall and F1 score rising slightly to 92% and 94%, respectively, demonstrating a more consistent performance across metrics. The LLM (RF) model, while still performing well, has a precision of 93% and an F1 score of 91%, but a slightly lower recall of 89% compared to the DT + KNN + SVM approach. This suggests that for the governance score, the LLM model may slightly underperform compared to a traditional model enhanced with SVM.

4.1.2.4 RQ 2 - PREDICTION OF ESG SCORE

Table 9 - Models Results, Q2 - ESG score

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.80	0.90	0.83	0.86	0.92
DT + KNN + SVM	0.83	0.91	0.85	0.88	0.93
LLM (RF)	0.90	0.91	0.91	0.90	0.95

For predicting the overall ESG score, the LLM (RF) model demonstrates superior performance compared to traditional methods, as demonstrated by the results in Table 9. It achieves an

accuracy of 90%, outperforming both the DT + KNN model, which scores 80%, and the DT + KNN + SVM model at 83%. The Random Forest-based LLM maintains balanced and high precision and recall, both at 91%, leading to a strong F1 score of 90%. The model's ROC AUC of 95% also indicates its ability to effectively distinguish between different classes.

4.1.3 RQ 3 - PREDICTION OF ESG SCORE

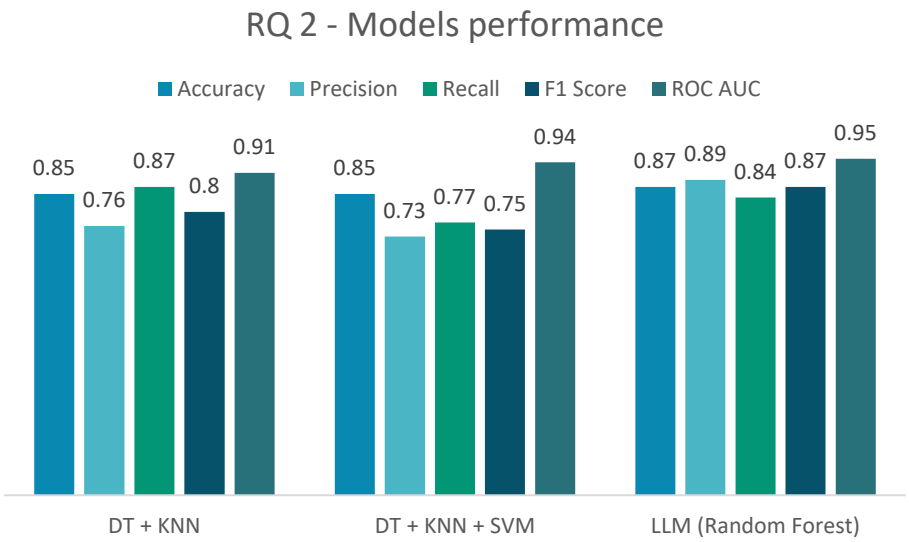

Figure 13 - Visualization of Models' Performance, RQ3.

The results of model performance for RQ3 are visualized in Figure 13. In predicting the ESG score using a simplified dataset, the LLM (Random Forest) model shows the highest performance among the compared models. It achieves an accuracy of 87%, slightly higher than both the DT + KNN and DT + KNN + SVM models, each of which have an accuracy of 85%. The LLM's precision of 89% is notably higher than that of the traditional ensemble models, indicating fewer false positives and greater confidence in its predictions. Nevertheless, the DT + KNN model presents the highest recall of 87%, followed by the LLM at 84% and the DT + KNN + SVM ensemble at This results in an F1 score of 77. The LLM's ROC AUC score of 98% is slightly higher than the ensemble models' ones.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DT + KNN	0.85	0.76	0.87	0.80	0.91
DT + KNN + SVM	0.85	0.73	0.77	0.75	0.94
LLM (RF)	0.87	0.89	0.84	0.87	0.95

5 DISCUSSION

The section that follows is structured along the entire analytical chain, from processing to evaluating, and aims at providing a detailed overview of the main quantitative and qualitative insights, shedding light on the challenges presented along the way, the main implication of the findings, together with the limitations encountered, laying the ground for final remarks and considerations.

5.1 DATA PREPROCESSING CHALLENGES AND INSIGHTS

Under the first RQ, the data preprocessing step represented one of complex task to be performed. In this context, the manual pre-processing resulted in an iterative task, which started from the inspection of the dataset along with the interpretation of the variables in more details. Given the high number of features, this activity was not only particularly time-consuming, but also required the application of more sophisticated methods to underpin the issues stemming from the data and address them. On the other hand, the use of GPT-4 significantly reduced the manual workload and time needed to screen the dataset. Although GPT-4 tended to prefer rather simplistic approaches, it quickly detected the main issues in the dataset and, when asked to apply more robust techniques, it proficiently handled the task as prompted, highlighting the benefits of the revised approach. Both the manual and the automated approach handled categorical and numerical values successfully, scaling them and encoding them where needed.

Features selection and dimensionality reduction through PCA, while not needed for the RQ 2 and 3, demonstrated to positively impact the predictions of the model under the first RQ. In this context, GPT-4 adopted a straightforward approach by analyzing the dataset's content and examining the correlation between the output variable and input features. However, a thorough analysis of the most important parameters selected by GPT-4 revealed that in the beginning also numerical IDs (e.g., Moody’s organization ID) were found to be correlated and were therefore used for the classification. This element demonstrates that despite generative models being very useful in reducing the time required for highly time-intensive tasks and claiming to work with highly correlated variables, they need to be thoroughly supervised to ensure the methodology remains sound and reasonable. These risks are mitigated by the fact GPT-4 consistently sought user confirmation at each step, allowing users to refine the approach as desired and experiment with more complex techniques if desired.

In cases where the dataset was less noisy, such as in RQ 2 and 3, had fewer dimensions, and included variables with clearly specified names and uniformly scaled values, pre-processing resulted in a simpler task. In these RQs, GPT-4 did not exhibit any major issues and successfully discriminated among the features, identifying the most suitable ones for predicting each output.

5.2 DATA MODELLING AND VALIDATION CHALLENGES AND INSIGHTS

The modelling process via traditional approaches resulted in an extensive and iterative processes, aimed at refining the models and enhance their performances. This task was particularly complex for RQ1 and RQ3, which involved scenarios with the most and the fewest data, respectively. On the other hand, GPT-4's automated approach to modelling and validation significantly reduced the time and effort typically required for these tasks. It showed proficiency in grasping the specificities of the problem, correctly identifying it as a classification issue. The manually selected models were based on extensive research and a deep understanding of both the algorithms and the topic at hand. In contrast, GPT-4, through its inspection of the dataset and comprehension of the study's objectives, consistently proposed the use of the Random Forest algorithm. This choice underscored the strengths of ensemble learning techniques. By prompting questions on the reasons why the model was selected, GPT-4 provided sound explanations grounded in theory and based on the nature of the data. These explanations highlighted the model's potential to effectively understand and process numerical data.

Throughout this phase, GPT-4 sought user confirmation at each step, allowing the exploration of alternative algorithms such as XGBoost. Although XGBoost was not the primary focus of the analysis, testing it highlighted the challenges GPT-4 faces with computationally intensive tasks, which can significantly impact its performance and occasionally cause session interruptions. Nevertheless, GPT-4 consistently provided the .py codes, allowing to run the tasks locally. This capability proved to be very useful for double-checking the execution of tasks and troubleshooting in instances where the sessions experienced resource errors.

GPT-4 employed structured approaches for tuning and validation, closely paralleling manual methods. It used GridSearchCV to systematically explore and evaluate different hyperparameter configurations, covering a broad range from simple to complex models and implemented 3-fold cross-validation, splitting the dataset into three parts and validating the model's performance across each fold.

Regarding the metrics used for evaluating the models, to ensure maximum comparability, these metrics were explicitly requested as input. Both approaches handled this final phase without encountering any major issues.

5.3 INTERPRETATION OF QUANTITATIVE FINDINGS

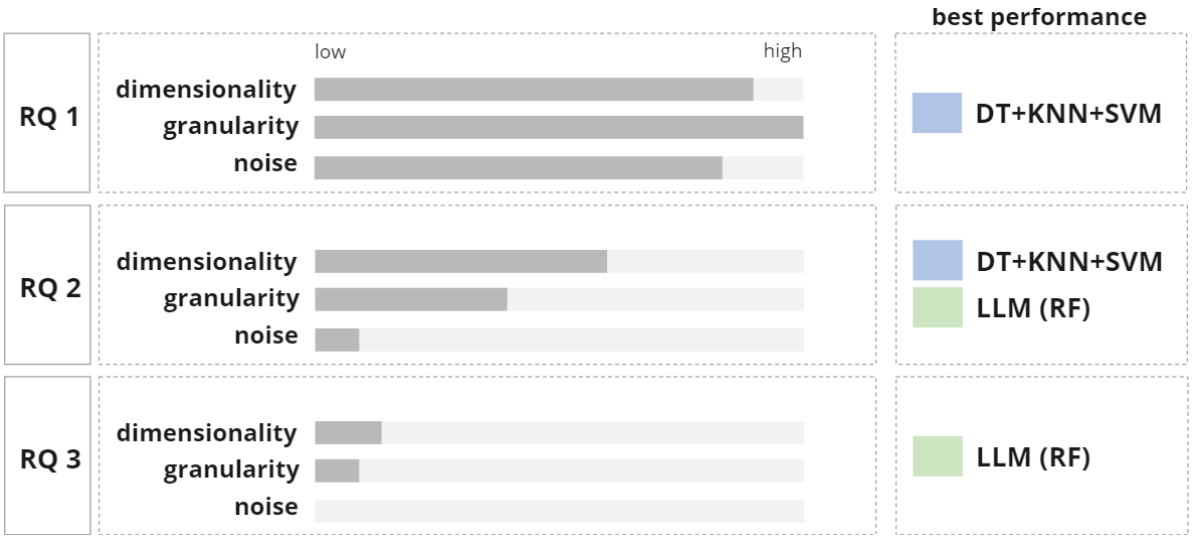

Figure 14 - Best Models Across RQs.

Figure 14 provides a comparative overview of the results across different research questions, which are further detailed in the following section.

RQ 1 - How do traditional methods and GPT-4 compare when predicting the environmental score in a noisy and high dimensional space?

RQ 1 explored predicting the environmental component of ESG scores using a highly complex and noisy dataset. In this context, the manual application of ensemble, in particular the combination of DT + KNN + SVM, demonstrated superior performance. These models benefited significantly from PCA for feature selection, which reduced the dataset's dimensionality and emphasized its most critical features. Each algorithm contributed to enhance and balance the prediction of the scores in different ways. Decision Trees leveraged on their ability to delineate decision paths efficiently also when datasets present several and noisy features, by breaking down the decision-making process into simpler steps and navigating the data's complexity effectively. K-Nearest Neighbors played a crucial role by averaging the outcomes from the nearest data points, providing a smoothing effect that mitigated the impact of noise and overfitting. This allowed the ensemble to generalize better across the noisy data. Support Vector Machine further enhanced the model's capability to handle complex and noisy datasets, maximizing the margin between classes, which is vital in high-dimensional spaces. This was reflected in the ensemble's strong performance, achieving a significantly high result for ROC AUC of 98%, showcasing its ability to capture non-linear relationships within the data.

In contrast, the RF model selected by GPT-4, though generally robust, underperformed in this RQ. Random Forests typically excel in high-dimensional settings through effective feature selection and averaging over many trees. However, in this case, the absence of dimensionality reduction techniques like PCA may have impacted the model due to the high noise, missing values, and unit diversity. This limitation likely constrained its performance, leading to lower scores (78% across most metrics and 90% in ROC AUC). This scenario suggests the importance of a careful and considerate pre-preprocessing phase, especially when managing highly complex and noisy datasets.

RQ 2 - How do traditional methods and GPT-4 compare when predicting ESG score components in a lower dimensional space also presenting class imbalance?

The results from RQ 2 underscore the generally enhanced performance of models when dealing with simpler datasets, characterized by low noise and highly relevant features. In this RQ, both the DT + KNN + SVM ensemble and the RF models demonstrated high performance. The DT + KNN + SVM model slightly outperformed in precision and ROC AUC, whereas the RF model reached higher results in terms of accuracy and recall. The F1 scores for both models were very close, reflecting a balanced approach between precision and recall. In this context, class imbalances, particularly in the governance score, significantly impact the model performance. In this case, the imbalance impacts recall, with the DT + KNN model achieving high precision but lower recall and F1 score. This suggests the model may miss minority class instances, however adding SVM improves the balance between precision and recall, demonstrating better overall performance in handling such issue. The LLM (RF) results indicate possible sensitivity to class imbalance. This indicates that imbalance impacts model performance, which is mitigated by integrating SVM and addressing class imbalances through preprocessing techniques.

Overall, looking at the average results achieved under this scenario, the differences between these models are minor, and selection can be based on the most relevant criteria for the research problem at hand. These results suggests that GPT-4's performance improves when the complexity inherent in the data is mitigated either by thorough pre-processing and advanced techniques to enhance the dataset's informative power (RQ 1) or by simpler and clearer input data.

RQ 3 - How do traditional methods and GPT-4 compare when predicting ESG score with scarce information?

RQ 3 underscores the strength of GPT-4 when working with datasets that are simplified yet contain informative features. The LLM’s superior performance in this RQ suggests that it benefits from clean, relevant input data where complex preprocessing is less critical. This finding highlights the potential of GPT-4 and similar models to excel in environments where data is limited but well-defined, leveraging its advanced capabilities to make accurate

predictions even with minimal input complexity. This can be particularly valuable in practical applications dealing with scarce input.

Overall, the manual approach provided a more controlled and consistent analysis, particularly beneficial for understanding and tuning complex models. GPT-4's automation, on the other hand, proved highly effective in rapidly analyzing large datasets, highlighting the main challenges, and executing standard procedures with minimal manual intervention. The combination of both approaches could offer a powerful framework for future research, leveraging GPT-4's efficiency for routine tasks while reserving manual methods and judgement for more complex data processing, model tuning, and validation.

In summary, the evaluation across these RQs highlights several key insights.

- Traditional models like DT + KNN and DT + KNN + SVM excel in handling highly complex and high-dimensional datasets when paired with thorough preprocessing techniques. This is evident in RQ 1, where these models effectively managed the high dimensionality and noise of the dataset under analysis.
- In scenarios where data complexity is reduced through aggregation or lower dimensionality (as in RQ 2 and RQ 3), LLMs demonstrate superior or equal performance compared to traditional methods.
- Class imbalance significantly impacts model performance and must be carefully considered during modelling and preprocessing stages. This is particularly evident in RQ 2, where incorporating SVM led to better performance. Notably, LLMs did not explicitly address class imbalance, which may have limited their performance in more imbalanced datasets.

5.4 INTERPRETATION OF QUALITATIVE FINDINGS

Beyond the evaluation of numerical metrics, this research explored the broader implications and practical aspects of using different modelling approaches. This comprehensive evaluation provided a richer perspective on the suitability and application of each method. The study highlights that each approach brings unique strengths and challenges suitable for different RQs. To facilitate a detailed comparison, the findings are organized into the following key areas for the (i) manual application of ensemble ML techniques and the (ii) usage of LLM via API like GPT-4.

1. Knowledge and skills requirements
2. Performance and adaptability
3. Transparency and interpretability
4. Application flexibility
5. Confidentiality and security
6. Model evolution and stability

5.4.1 KNOWLEDGE AND SKILLS REQUIREMENTS

(i) These require substantial expertise in programming and machine learning theory. Practitioners must have a deep understanding of the model's architecture and the ability to finely control and tune every aspect, from data pre-processing to feature engineering and model training. These models, under the expert's control, are usually based on robust statistical theories, providing a transparent view of their workings.

(ii) These models are designed for accessibility to users with varying levels of technical expertise. Their accessibility makes them particularly effective for less-technical users, despite a good knowledge of the theoretical aspects as well as topical knowledge is needed. While a solid grasp of theoretical principles and domain knowledge enhances their use, they generally do not require specialized coding skills, making them highly effective for non-technical users. This accessibility can reduce the barrier to entry for complex tasks.

Table 10 –Traditional ML vs GPT-4: Knowledge and Skills Requirements

Aspect	**Traditional ML**	**GPT-4**
Knowledge of coding	Requires proficiency in programming and software engineering	Accessible to non-technical users, can be used without any coding knowledge. However coding expertise may help assessing the quality of the responses
ML techniques and theory	Demands understanding of complex ML concepts and methods	Users need to master effective prompting, as well as have a goof level of understanding of the models and techniques used to critically assess the responses
Expert judgment	Essential for model selection, feature engineering, and tuning	Crucial for interpreting responses, critically assessing outputs and guiding conversation to achieve accurate results
Data pre-processing	Often needs extensive data cleaning and feature engineering	Data pre-processing is simpler, but expert judgment is crucial to ensure accurate data handling

5.4.2 PERFORMANCE AND ADAPTABILITY

(i) Known for their high customizability, these models allow precise tuning and optimization for specific problems. However, they require manual re-training and validation to adapt to new data, which can be labor-intensive. Despite this, they offer stable performance with detailed control over updates and changes, ensuring consistency in their outputs and reproducibility.

(ii) These models are highly adaptable, rapidly adjusting to new inputs and evolving contexts. They benefit from ongoing updates and fine-tuning, providing continuous performance enhancements without the need for manual adjustments. However, the non-deterministic nature of LLMs can lead to variability in outputs, meaning different sessions within GPT-4 can result in varied approaches to handling the same task, even with identical prompts. This variability highlights the model's ability to generate diverse solutions but also poses challenges for reproducibility. Despite this, repeated runs indicate that GPT-4 performs tasks with a high degree of consistency overall, suggesting reliable core functionality across iterations.

Table 11 - Traditional ML vs GPT-4: Performance and Adaptability

Aspect	**Traditional ML**	**GPT-4**
Time-Consuming	Building and training models can be very time-intensive	Produces fast results but may require careful prompting for accuracy
Performance Tuning	Allows for detailed performance optimization for specific use cases	Automatically improves with new updates and fine-tuning
Reproducibility	Models and results can be consistently reproduced under the same conditions	Less consistent due to evolving models. It may produce varied outputs in different iterations and sessions
Scalability	Can be optimized for large-scale data processing and analysis	Can handle multiple languages and extensive datasets with ease
Adaptability	Adapts slowly as changes require re-training and validation	Rapidly adapts to new inputs and contexts, making it highly responsive to evolving needs

5.4.3 TRANSPARENCY AND INTERPRETABILITY

(i) They are built following sound methodologies and ensure full transparency of their workings. However, interpretability can become challenging when complex pre-processing or ensemble techniques are employed.
(ii) Although potentially less transparent, their interpretability can be improved by prompting the model to explain its reasoning and choices.

Table 12 - Traditional ML vs GPT-4: Transparency and Interpretability

Aspect	Traditional ML	GPT-4
Transparency	Model can be thoroughly understood and visualized	Less transparent decision-making, though explanations can be requested
Interpretability	Can be very complex, especially with advanced models	Although complex, it offers explanations of outputs and reasoning if prompted

5.4.4 APPLICATION FLEXIBILITY

(i) These offer extensive flexibility and can finely tuned for specific tasks. Practitioners have control over applying various algorithms and techniques suited to the problem's needs, allowing for a tailored approach to model development and application.
(ii) Providing a broad range of capabilities, they accommodate various types of input files and can generate outputs that go beyond text and numerical results, including graphical representations and visualizations. This versatility allows for comprehensive analysis and presentation, catering to a wide array of data and output formats. Furthermore, their general-purpose nature makes them versatile for numerous applications, accommodating a wide array of problem types without requiring significant modifications.

Table 13 - Traditional ML vs GPT-4: Application Flexibility

Aspect	Traditional ML	GPT-4
Customizability	Can be tailored to specific problems with fine-tuned parameters	Provides a wide range of pre-trained models and different ML algorithms
Range of algorithms	Supports various algorithmic approaches like regression, clustering, etc.	Offers diverse algorithms and supports the usage of text

5.4.5 CONFIDENTIALITY AND SECURITY

(i) Typically trained and maintained locally, these models ensure secure handling of sensitive data, which is crucial to make sure that sensitive information remains protected and managed within secure environments.

(ii) Operating in environments where data is often processed externally, these models raise concerns about data security and privacy. The reliance on external processing can introduce risks that need to be carefully managed to protect data confidentiality.

Table 14 - Traditional ML vs GPT-4: Confidentiality and Security

Aspect	Traditional ML	GPT-4
Confidentiality	Models are trained locally, reducing issues with confidential data	Could be an issue when dealing with confidential data if not in a safe environment

5.4.6 MODEL EVOLUTION AND STABILITY

(i) These models maintain stable performance through detailed control over updates and changes. As they evolve through manual re-training and adjustment, they ensure consistent outputs, although can require substantial manual work to be adapted and modified. Nevertheless, this stability is beneficial for applications where reliability and predictability are paramount.

(ii) As they are continuously updated, these models often exhibit variability in outputs across different versions, with older versions being slowly deprecated when new versions are released. While this evolution keeps the models current, it can also lead

to inconsistencies and challenges for researchers, especially when the analysis is performed in periods of transitions from one model to the other.

Table 15 - Traditional ML vs GPT-4: Model Evolution and Stability

Aspect	Traditional ML	GPT-4
Model Evolution	Models are static once deployed unless manually updated	Models continuously evolve, which can affect consistency and stability
Stability	Generally stable and predictable in outputs once trained	May produce varied outputs in different iterations (different chats)

5.4.7 RELIANCE ON EXPERT OVERSIGHT

(i) Depend heavily on expert oversight for model tuning, feature selection, and validation. Experts ensure the model's alignment with theoretical principles and sound data handling practices. This oversight is essential for maintaining the integrity and effectiveness of the models, especially in complex or high-stakes applications.

(ii) Require careful supervision and prompting to avoid simplistic or erroneous outputs. These models can produce "hallucinations", meaning responses that are incorrect or biased. Expert guidance is pivotal to navigate complex tasks and ensure that the model's decisions are accurate and reliable. While the automation provided by GPT-4 streamlines many aspects of the coding process, human oversight remains essential to ensure that all task requirements are met and that the outputs are trustworthy.

Table 16 - Traditional ML vs GPT-4: Reliance on Expert Oversight

Aspect	Traditional ML	GPT-4
Expert judgement and oversight	Depend on expert oversight for model tuning, feature selection, and validation. Experts ensure alignment with theoretical principles and data handling practices	Require careful supervision and prompting to avoid simplistic or erroneous outputs; expert guidance is pivotal to ensure accuracy and reliability and to double check the soundness of the steps undertaken.

5.5 RECOMMENDATIONS FOR PRACTITIONERS

As outlined in previous sections, generative AI tools, such as GPT-4, offer significant capabilities for automating and enhancing the modelling process. To maximize their potential and ensure robust outcomes, practitioners should consider the following recommendations.

- Repeat the task in multiple sessions to progressively refine prompts and responses. This iterative process helps in understanding the behavior of the model and honing the instructions to optimize its performance for the given task and explore various approaches, ultimately leading to more robust and accurate solutions.
- Leverage the model's ability to learn from previous interactions. Each iteration provides an opportunity to incorporate learnings and adjust the approach, which is particularly useful for complex tasks.
- Break down larger tasks into smaller, manageable components. This stepwise approach simplifies the process, making it easier to monitor progress, troubleshoot issues, and refine each step. Furthermore, it also allows to execute each phase in a different session which might improve the model performance in case of particularly elaborated and complex tasks.
- Ensure the prompt contains a request for the codes (e.g., Python) that can be used to execute locally. This allows to validate the results at each step before proceeding and to thoroughly monitor the behavior of the model and, where needed, correct it.
- Allow the model some freedom to explore alternative methods and test multiple approaches quickly, potentially uncovering innovative solutions. At the same time, establish clear guidelines to ensure the model adheres to sound methodology. Challenge the model’s choices and validate them against established principles and best practices. This balance ensures that while the model explores, the solutions remain grounded in rigorous analytical frameworks.
- Ensure that all variables used in the model are well-defined and contextually explained. This clarity helps in understanding the meaning of each variable and prevents misinterpretation. In parallel, closely monitor the variables being used. Furthermore, identify and exclude any that are irrelevant, redundant, or potentially problematic (e.g., identifiers like IDs that may not be meaningful in the modelling context) or pre-emptively guide the model to disregard such variables to maintain the integrity and relevance of the analysis.

5.6 LIMITS OF THE STUDY

The empirical results reported herein should be considered in the light of some limitations.

- **Limited access to data**: due to the high costs related to retrieving ESG data, this study focused on a narrow portion of the available data at granular level. Specifically, in the predictions at the highest level of granularity, only variables related to the environmental component of the rating were considered. Consequently, the prediction of the full ESG score at the highest level of granularity was not possible due to the lack of data on the social and governance component. This limitation restricts the comprehensiveness of the predictions and prevents from understanding how these models performs in scenarios characterized by high data volume and variety of information on all components, which presents a recurrent use case for rating agencies and institutions having broad access to data.
- **Scope of the study**: the study's focus on specific research questions and datasets limits generalizability across different contexts. Therefore, replication of the tasks across multiple datasets is needed to test the significance of the results. Consequently, the findings may not be applicable to broader contexts or different industries, reducing the external validity of the study.
- **Lack of prior research studies on the topic**: there is a limited amount of prior research in this specific area, which may affect the robustness and comparability of the findings. This aspect represents a bottleneck, making it challenging to benchmark the results, potentially affecting the study's confidence in its conclusions.

6 CONCLUSIONS

6.1 SUMMARY OF KEY FINDINGS

The current research aimed to compare the performance of traditional ensemble methods, applied manually, with that of large language models when running a classification algorithm for predicting ESG scores. Given the inherent challenges often associated with this type of data, both methods have been tested under various scenarios to evaluate their performance.

The main findings are divided by RQ as it follows.

RQ 1 – How do traditional methods and GPT-4 compare when predicting the environmental score in a noisy and high dimensional space?

The analysis highlighted that in addressing the high-dimensional, noisy data of RQ 1, manual ensemble models combining DT, K-NN, and SVM outperformed GPT-4's Random Forest approach. The standard approach allowed to proficiently manage the high dimensionality and noise through more sophisticated pre-processing techniques like PCA and ensemble strategies. In contrast, GPT-4, while robust, struggled with the high noise and complex feature space due to its simplistic initial pre-processing approach. While capable of handling the data with minimal user intervention, its performance lagged the manual models, highlighting the need for more guided and nuanced data handling in such scenarios.

RQ 2 – How do traditional methods and GPT-4 compare when predicting ESG score components in a lower dimensional space also presenting class imbalance?

For RQ 2, which involved lower dimensionality, aggregated data and class imbalance, the manual ensembles performed strongly, particularly the DT + KNN + SVM combination. GPT-4 performance, on the other hand, closely matched or exceed the performance of the traditional ensemble models, with exclusion for the case characterized by high class imbalance. This underscores GPT-4's robustness when dealing with cleaner and aggregated data, although its effectiveness is reduced in more imbalanced scenarios.

RQ 3 – How do traditional methods and GPT-4 compare when predicting ESG score with scarce information?

In the context of simplified ESG score prediction for RQ 3, the traditional models maintained high performance but showed diminishing benefits compared to GPT-4. Despite the scarce information and the simplicity of the data, which required minimal processing, GPT-4 excelled in this scenario. The low complexity of the data allowed GPT-4 to leverage its pre-trained capabilities effectively, providing robust and consistent predictions with minimal input features.

RQ 4 – How do traditional methods and GPT-4 compare in terms of interpretability, scalability, usability, and robustness?

The comparison between traditional ML techniques and GPT-4 reveals that each approach has unique strengths and challenges. Traditional methods excel in terms of interpretability, control, and robustness, making them suitable for tasks requiring deep understanding and stable performance. However, they require significant expertise, resources, and time. GPT-4, on the other hand, offers scalability, usability, and adaptability, making accessible and efficient for a wide range of applications, though it requires careful management, especially in terms of data pre-processing, and supervision to handle its variability and ensure reliability. The choice between these approaches should be guided by the specific needs of the research question and the context of the application.

6.2 SIGNIFICANCE OF FINDINGS

The research advances ESG scoring models and machine learning by demonstrating that both traditional ML models and LLMs are viable for predicting ESG scores, each excelling in different scenarios, providing valuable insights for practitioners in selecting the appropriate model based on the data complexity and specific ESG components. GPT-4's adaptability and efficiency make it suitable for organizations with limited resources or minimal access to data, as it can make accurate predictions even with scarce information. This is particularly relevant also for those seeking to implement efficient, scalable solutions with limited computational resources or technical expertise. Furthermore, the research highlights that combining traditional ML models and LLMs could leverage their complementary strengths, leading to more robust solutions.

The findings in this thesis align with previous studies in several respects. Previous studies highlighted the high performance of DT, SVM, and KNN in credit scoring and ESG ratings. Furthermore, other studies stressed the robustness of ensemble methods combining these algorithms, similar to how the ensemble of DT, KNN, and SVM performed exceptionally well in high-dimensional and noisy data scenarios in this thesis. Additionally, the use of RF for predicting ESG scores aligns with findings by D'Amato et al. (2021) and Krappel et al. (2021), where RF showed high prediction performance due to its ability to handle non-linear patterns and its robustness against overfitting. In accordance with Leippold (2023), this thesis confirms that LLMs can hallucinate and that their responses may lack grounding in facts. Additionally, consistent with Son (2023), this thesis finds that better instruction tuning positively impacts results. However, it also highlights that larger datasets do not necessarily make the model perform better if these datasets are very noisy and complex. This contrasts with earlier studies that primarily focused on NLP tasks rather than quantitative predictions.

Overall, this thesis contributes to the ongoing discourse on the application of machine learning in ESG scoring and beyond, providing a framework for understanding the strengths and

limitations of different modeling approaches. It paves the way for future research and practical applications that integrate traditional ensemble techniques with cutting-edge AI to tackle diverse data challenges.

6.3 FUTURE RESEARCH RECOMMENDATIONS

Building on the findings and limitations presented, several avenues for future research are recommended. These suggestions focus on enhancing the study's methodology, as well as exploring new areas of investigation.

Future research should aim to utilize more comprehensive datasets including all ESG components at granular level to provide a more holistic analysis. Additionally, exploring further datasets, including the granular data directly utilized by rating agencies to score companies, could enhance the generalizability of the findings, and allow to highlight the strengths and weaknesses of these methods in various ESG data scenarios. Conducting cross-rating agency studies could also offer a more comprehensive understanding of ESG scoring models across different rating methodologies.

Exploring time-series models that not only predict current ESG scores but also forecast future trends based on historical data is another promising area of research. This approach could provide predictive insights into potential changes in a company's ESG performance.

Furthermore, given the rapid evolution of artificial intelligence, testing the scenarios using other LLMs, like future GPT versions, BERT or LLaMa could reveal strengths and limitations specific to each model. Fully exploring multimodal capabilities presents another research venue, leveraging LLMs to integrate both text and numerical data. This could involve using textual ESG reports and news about companies' practices and performance alongside structured numerical data, providing a richer context and improve model predictions.

These recommendations aim to provide a pathway for future research to build on the foundation established in this thesis, exploring new areas relevant to ESG scoring and beyond.

6.4 FINAL CONSIDERATIONS

In conclusion, this thesis demonstrates that GPT-4 results in a valid addition to the toolbox of practitioners interested in incorporating environmental, social and governance data in their predictions, and more generally to support different classification tasks. This method is, in most cases, comparable in accuracy to the classical ML models currently making up the bulk of the ESG prediction toolbox. The analysis ultimately underscores the importance of balancing automated approaches with expert control. Accordingly, this new methodology should be seen as a complement to, rather than a substitute for, existing techniques.

BIBLIOGRAPHICAL REFERENCES

Abdou, H., Pointon, J. (2011). CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3). https://doi.org/10.1002/isaf.325

Abdullah, D., Abdulazeez, A. (2021). Machine Learning Applications based on SVM Classification: A Review. *Qubahan Academic Journal*, 1(2). https://doi.org/10.48161/qaj.v1n2a50

Abhayawansa, S., Tyagi, S. (2021). Sustainable investing: The black box of environmental, social, and governance (ESG) ratings. Journal of Wealth Management, 24(1). https://doi.org/10.3905/JWM.2021.1.130

Abnoosian, K., Farnoosh, R., Behzadi, M. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics*, 24(1). https://doi.org/10.1186/s12859-023-05465-z

Anguita, D., Ghio, A., Greco, N., Oneto, L., Ridella, S. (2010). Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory. *In Proceedings of the International Joint Conference on Neural Networks.* https://doi.org/10.1109/IJCNN.2010.5596450

Ashofteh, A., Bravo, J. (2021). A conservative approach for online credit scoring. *Expert Systems with Applications*, 176. https://doi.org/10.1016/j.eswa.2021.114835

Aurélien Géron (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems, 2nd Edition. *O'Reilly Media* (ISBN: 9781492032649)

Bao, W., Lianju, N., Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128. https://doi.org/10.1016/j.eswa.2019.02.033

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

D'Amato, V., D'Ecclesia, R., Levantesi, S. (2021). Fundamental ratios as predictors of ESG scores: a machine learning approach. *Decisions in Economics and Finance*, 44(2). https://doi.org/10.1007/s10203-021-00364-5

D'Amato, V., D'Ecclesia, R., Levantesi, S. (2022). ESG score prediction through random forest algorithm. *Computational Management Science*, 19(2). https://doi.org/10.1007/s10287-021-00419-3

Dastile, X., Celik, T., Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, 91. https://doi.org/10.1016/j.asoc.2020.106263

De Diego, I., Redondo, A., Fernández, R., Navarro, J., Moguerza, J. (2022). General Performance Score for classification problems. Applied Intelligence, 52(10). https://doi.org/10.1007/s10489-021-03041-7

De Ranasinghe, I., Munasinghe, L. (2023). Comparison of performances of ML-Algorithms in the estimation of the execution time of non-parallel Java programs. *Journal of Science of the University of Kelaniya*, 16(1). https://doi.org/10.4038/josuk.v16i1.8074

Del Vitto, A., Marazzina, D., Stocco, D. (2023). ESG ratings explainability through machine learning techniques. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05514-z

Deng, Z., Zhu, X., Cheng, D., Zong, M., Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, 195. https://doi.org/10.1016/j.neucom.2015.08.112

Dong, Y. (2023). ESG Scoring System Construction: Portfolio Investment Based on Machine Learning. *Advances in Economics, Management and Political Sciences*, 3(1). https://doi.org/10.3905/jesg.2021.1.010

Dumitrescu, E., Hué, S., Hurlin, C., Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3). https://doi.org/10.1016/j.ejor.2021.06.053

Feng, X., Xiao, Z., Zhong, B., Qiu, J., Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. Applied Soft Computing Journal, 65. https://doi.org/10.1016/j.asoc.2018.01.021

Florez-Lopez, R., Ramon-Jeronimo, J. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. *A correlated-adjusted decision forest proposal. Expert Systems with Applications*, 42(13). https://doi.org/10.1016/j.eswa.2015.02.042

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3). https://doi.org/10.1080/15228053.2023.2233814

Golbayani, P., Florescu, I., Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and

decision trees. *North American Journal of Economics and Finance*, 54. https://doi.org/10.1016/j.najef.2020.101251

Guo, T. (2020). ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3593885

Haleem, A., Javaid, M., Singh, R. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(4). https://doi.org/10.1016/j.tbench.2023.100089.

Henriksson, R., Livnat, J., Pfeifer, P., Stumpp, M. (2019). Integrating ESG in portfolio construction. *Journal of Portfolio Management*, 45(4). https://doi.org/10.3905/jpm.2019.45.4.067

Khurana, D., Koli, A., Khatter, K., Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3). https://doi.org/10.1007/s11042-022-13428-4

Kim, J., Chua, M., Rickard, M., Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5). https://doi.org/10.1016/j.jpurol.2023.05.018

Kotsiantis, S. (2013). Decision trees: A recent overview. *Artificial Intelligence Review,* 39(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4

Lanza, S. (2024). Machine learning the breakdown of tame effective theories. *Cornell University Publications*. https://doi.org/10.48550/arXiv.2311.03437

Lee, O., Joo, H., Choi, H., Cheon, M. (2022). Proposing an Integrated Approach to Analyzing ESG Data via Machine Learning and Deep Learning Algorithms. *Sustainability (Switzerland)*, 14(14). https://doi.org/10.3390/su14148745

Leippold, M. (2023). Thus spoke GPT-3: Interviewing a large-language model on climate finance. *Finance Research Letters*, 53. https://doi.org/10.1016/j.frl.2022.103617

Lessmann, S., Baesens, B., Seow, H., Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1). https://doi.org/10.1016/j.ejor.2015.05.030

Lin, H., Hsu, B. (2023). Empirical Study of ESG Score Prediction through Machine Learning—A Case of Non-Financial Companies in Taiwan. *Sustainability (Switzerland)*, 15(19). https://doi.org/10.3390/su151914106

Liu, M. (2022). Quantitative ESG disclosure and divergence of ESG ratings. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.936798

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13(3). https://doi.org/10.1371/journal.pone.0194889

Melese, T., Berhane, T., Mohammed, A., Walelgn, A. (2023). Credit-Risk Prediction Model Using Hybrid Deep—Machine-Learning Based Algorithms. *Scientific Programming*, 2023. https://doi.org/10.1155/2023/6675425

Michalski, L., Low, R. (2021). Corporate credit rating feature importance: Does ESG matter?. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3788037

Mukid, M., Widiharih, T., Rusgiyono, A., Prahutama, A. (2018). Credit scoring analysis using weighted k nearest neighbor. *In Journal of Physics: Conference Series*, 1025(1). https://doi.org/10.1088/1742-6596/1025/1/012114

Nguyen, Q., Diaz-Rainey, I., Kuruppuarachchi, D. (2021). Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach*. Energy Economics*, 95. https://doi.org/10.1016/j.eneco.2021.105129

Pannakkong, W., Thiwa-Anont, K., Singthong, K., Parthanadee, P., Buddhakulsomsiri, J. (2022). Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM, and DBN. *Mathematical Problems in Engineering*, 2022. https://doi.org/10.1155/2022/8513719

Quinlan, J. (1986). Induction of Decision Trees*. Machine Learning,* 1(1). https://doi.org/ 10.1023/A:1022643204877

Rahaman, M., Ahsan, M., Anjum, N., Terano, H., Rahman, M. (2023). From ChatGPT-3 to GPT-4: A Significant Advancement in AI-Driven NLP Tools. *Journal of Engineering and Emerging Technologies,* 1(1). https://doi.org/10.52631/jeet.v1i1.188

Schonlau, M., Zou, R. (2020). The random forest algorithm for statistical learning. *Stata Journal*, 20(1). https://doi.org/10.1177/1536867X20909688

Seo, S. (2017). Beyond the Paris Agreement: Climate change policy negotiations and future directions. *Regional Science Policy and Practice*, 9(2). https://doi.org/10.1111/rsp3.12090

Shaik, A., Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model, 56. *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-13-2354-6_27

Sokolov, A., Mostovoy, J., Ding, J., Seco, L. (2021). Building Machine Learning Systems for Automated ESG Scoring. *The Journal of Impact and ESG Investing*, 1(3). https://doi.org/10.3905/jesg.2021.1.010

Son, G., Jung, H., Hahm, M., Jin, S., Na, K. (2023). Beyond Classification: Financial Reasoning in State-of-the-Art Language Models. *In FinNLP-Muffin 2023 - Joint Workshop of the 5th Financial Technology and Natural Language Processing and 2nd Multimodal AI For Financial Forecasting, in conjunction with IJCAI 2023 - Proceedings.* https://doi.org/10.48550/arXiv.2305.01505

Song, Y., Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2). https://doi.org/10.11919/j.issn.1002-0829.215044

Wang, H., Yang, Y., Wang, H., Chen, D. (2013). Soft-voting clustering ensemble. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7872. https://doi.org/10.1007/978-3-642-38067-9_27

Wang, J., Lu, S., Wang, S., Zhang, Y. (2022). A review on extreme learning machine. *Multimedia Tools and Applications*, 81(29). https://doi.org/10.1007/s11042-021-11007-7

Yang, Z., Zhang, L., Wang, X., Mai, Y. (2023). ESG Text Classification: An Application of the Prompt-Based Learning Approach. *Journal of Financial Data Science*, 5(1). https://doi.org/10.3905/jfds.2022.1.115

Zaremba, A., Demir, E. (2023). ChatGPT: Unlocking the future of NLP in finance. *Modern Finance*, 1(1). https://doi.org/10.61351/mf.v1i1.43

Zhang, D., Zhou, X., Leung, S., Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 37(12). https://doi.org/10.1016/j.eswa.2010.04.054

Zhang, N., Li, L., Chen, X., Deng, S., Bi, Z., Tan, C., Huang, F., Chen, H. (2022). DIFFERENTIABLE PROMPT MAKES PRE-TRAINED LANGUAGE MODELS BETTER FEW-SHOT LEARNERS. *In ICLR 2022 - 10th International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2108.13161

Zhang, P., Jia, Y., Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6). https://doi.org/10.1177/15501329221106935

Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, 8(3). https://doi.org/10.1145/2990508

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R. (2018). Efficient kNN classification with different numbers of nearest neighbors*. IEEE Transactions on Neural Networks and Learning Systems*, 29(5). https://doi.org/10.1109/TNNLS.2017.2673241

Zhang, Y., Wang, J. (2016). K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting*, 32(3). https://doi.org/10.1016/j.ijforecast.2015.11.006

APPENDIX A

Variable name	Source
Primary Issuer ISIN	Urgentem
LEI – Legal Entity Identifier	Urgentem
Company (Short) Name	Urgentem
ISIN – International Securities Identification Number	Urgentem
Disclosure Category	Urgentem
Primary Data Source	Urgentem
Number of Scope 3 Categories Disclosed	Urgentem
Emissions Year	Urgentem
Country	Urgentem
Region	Urgentem
NACE Level 1	Urgentem
NACE Level 2	Urgentem
NACE Level 3	Urgentem
NACE Level 4	Urgentem
Intensity Average Inference Scope 1, 2 & 3 Total (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 1 & 2 Total (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 1 & 2 Total (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Total (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Total (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 1 (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 1 (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 2 Location Based (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 2 Location Based (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 2 Market Based (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 2 Market Based (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Purchased Goods and Services (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Purchased Goods and Services (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Capital Goods (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Capital Goods (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Fuel and Energy Related Activities (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Fuel and Energy Related Activities (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Upstream Transport and Distribution (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Upstream Transport and Distribution (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Waste Generated (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Waste Generated (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Business Travel (tCO2e/$m Revenue)	Urgentem

Intensity Average Inference Scope 3 Business Travel (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Employee Commuting (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Employee Commuting (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Upstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Upstream Leased Assets (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Downstream Transport Distribution (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Downstream Transport Distribution (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Processing of Sold Products (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Processing of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Use of Sold Products (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Use of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 End of Life Treatment of Sold Products (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 End of Life Treatment of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Downstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Downstream Leased Assets (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Franchises (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Franchises (tCO2e/$m Revenue) - Source	Urgentem
Intensity Average Inference Scope 3 Investments (tCO2e/$m Revenue)	Urgentem
Intensity Average Inference Scope 3 Investments (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 1, 2 & 3 Total (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 1 & 2 Total (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 1 & 2 Total (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Total (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Total (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 1 (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 1 (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 2 Location Based (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 2 Location Based (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 2 Market Based (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 2 Market Based (tCO2e/$m Revenue) - Source	Urgentem

Maximum Emissions Intensity Scope 3 Purchased Goods and Services (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Purchased Goods and Services (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Capital Goods (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Capital Goods (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Fuel and Energy Related Activities (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Fuel and Energy Related Activities (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Upstream Transport and Distribution (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Upstream Transport and Distribution (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Waste Generated (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Waste Generated (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Business Travel (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Business Travel (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Employee Commuting (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Employee Commuting (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Upstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Upstream Leased Assets (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Downstream Transport Distribution (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Downstream Transport Distribution (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Processing of Sold Products (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Processing of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Use of Sold Products (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Use of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 End of Life Treatment of Sold Products (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 End of Life Treatment of Sold Products (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Downstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Downstream Leased Assets (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Franchises (tCO2e/$m Revenue)	Urgentem

Maximum Emissions Intensity Scope 3 Franchises (tCO2e/$m Revenue) - Source	Urgentem
Maximum Emissions Intensity Scope 3 Investments (tCO2e/$m Revenue)	Urgentem
Maximum Emissions Intensity Scope 3 Investments (tCO2e/$m Revenue) - Source	Urgentem
Reported Emissions Intensity Scope 1 & 2 (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 1 Gross (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 2 Location based (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 2 Market based (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 1. Purchased Goods and Services (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 2. Capital Goods (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 3. Fuel and Energy Related Activities (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 4. Upstream Transportation (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 5. Waste Generated in Operations (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 6. Business Travel (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 7. Employee Commuting (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 8. Upstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 9. Downstream Transportation (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 10. Processing of Sold Products (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 11. Use of Sold Products (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 12. End of Life Treatment of Sold Products (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 13. Downstream Leased Assets (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 14. Franchises (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 15. Investments (tCO2e/$m Revenue)	Urgentem
Reported Emissions Intensity Scope 3 Other Total (tCO2e/$m Revenue)	Urgentem
Reported Emissions Total Scope 1 & 2 (tCO2e)	Urgentem
Reported Emissions Total Scope 3 (tCO2e)	Urgentem
Reported Emissions Scope 1 Gross (tCO2e)	Urgentem
Reported Emissions Scope 2 Location based (tCO2e)	Urgentem
Reported Emissions Scope 2 Market based (tCO2e)	Urgentem
Reported Emissions Scope 3 1. Purchased Goods and Services (tCO2e)	Urgentem
Reported Emissions Scope 3 2. Capital Goods (tCO2e)	Urgentem
Reported Emissions Scope 3 3. Fuel and Energy Related Activities (tCO2e)	Urgentem
Reported Emissions Scope 3 4. Upstream Transportation (tCO2e)	Urgentem
Reported Emissions Scope 3 5. Waste Generated in Operations (tCO2e)	Urgentem
Reported Emissions Scope 3 6. Business Travel (tCO2e)	Urgentem
Reported Emissions Scope 3 7. Employee Commuting (tCO2e)	Urgentem

Reported Emissions Scope 3 8. Upstream Leased Assets (tCO2e)	Urgentem
Reported Emissions Scope 3 9. Downstream Transportation (tCO2e)	Urgentem
Reported Emissions Scope 3 10. Processing of Sold Products (tCO2e)	Urgentem
Reported Emissions Scope 3 11. Use of Sold Products (tCO2e)	Urgentem
Reported Emissions Scope 3 12. End of Life Treatment of Sold Products (tCO2e)	Urgentem
Reported Emissions Scope 3 13. Downstream Leased Assets (tCO2e)	Urgentem
Reported Emissions Scope 3 14. Franchises (tCO2e)	Urgentem
Reported Emissions Scope 3 15. Investments (tCO2e)	Urgentem
Reported Emissions Scope 3 Other Total (tCO2e)	Urgentem
Absolute - Average Inference Scope 1, 2 & 3 Total (tCO2e)	Urgentem
Absolute - Average Inference Scope 1 & 2 Total (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Total (tCO2e)	Urgentem
Absolute - Average Inference Scope 1 (tCO2e)	Urgentem
Absolute - Average Inference Scope 2 Location Based (tCO2e)	Urgentem
Absolute - Average Inference Scope 2 Market Based (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Purchased Goods and Services (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Capital Goods (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Fuel and Energy Related Activities (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Upstream Transport and Distribution (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Waste Generated (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Business Travel (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Employee Commuting (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Upstream Leased Assets (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Downstream Transport Distribution (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Processing of Sold Products (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Use of Sold Products (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 End of Life Treatment of Sold Products (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Downstream Leased Assets (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Franchises (tCO2e)	Urgentem
Absolute - Average Inference Scope 3 Investments (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 1, 2 & 3 (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 1 & 2 (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Total (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 1 (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 2 Location Based (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 2 Market Based (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Purchased Goods and Services (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Capital Goods (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Fuel and Energy Related Activities (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Upstream Transport and Distribution (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Waste Generated (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Business Travel (tCO2e)	Urgentem

Maximum Emissions Absolute - Scope 3 Employee Commuting (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Upstream Leased Assets (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Downstream Transport Distribution (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Processing of Sold Products (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Use of Sold Products (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 End of Life Treatment of Sold Products (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Downstream Leased Assets (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Franchises (tCO2e)	Urgentem
Maximum Emissions Absolute - Scope 3 Investments (tCO2e)	Urgentem
Rating Methodology	Moody's
Long Term Rating	Moody's
Long Term Rating Description	Moody's
Last Long Term Rating Action	Moody's
Long Term Rating Action Date	Moody's
Bank Long Term Rating	Moody's
Outlook	Moody's
CIS - Credit Impact Score	Moody's
IPS-E - Issuer Profile Score - Environmental	Moody's
Physical Climate Risks	Moody's
Carbon Transition	Moody's
Water Management	Moody's
Natural Capital	Moody's
Waste and Pollution	Moody's
IPS-S - Issuer Profile Score - Social	Moody's
Customer Relations	Moody's
Human Capital	Moody's
Demographic and Societal Trends	Moody's
Health and Safety	Moody's
Responsible Production	Moody's
Demographics	Moody's
Labor and Income	Moody's
Education	Moody's
Housing	Moody's
Health and Safety	Moody's
Access to Basic Services	Moody's
IPS-G - Issuer Profile Score - Governance	Moody's
Financial Strategy and Risk Management	Moody's
Management Credibility and Track Record	Moody's
Board Structure, Policies and Procedures	Moody's
Organizational Structure	Moody's
Compliance and Reporting	Moody's
Institutional Structure	Moody's
Policy Credibility & Effectiveness	Moody's
Budget Management	Moody's
Transparency and Disclosure	Moody's

CTI - Carbon Transition Indicator	Moody's
CTI Date	Moody's

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa