UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI

"M.FANNO"

CORSO DI LAUREA MAGISTRALE IN

ENTREPRENEURSHIP AND INNOVATION

TESI DI LAUREA

"DEVELOPING A FINANCIAL RATIO-BASED MODEL FOR ESG SUSTAINABILITY PREDICTION IN PRIVATE FIRMS: A MACHINE LEARNING APPROACH "

RELATORE:

CH.MO PROF. ANTONIO PARBONETTI

LAUREANDA: OMNIYA SHWAKI ELKADY ELTAHER

MATRICOLA N. 2051959

ANNO ACCADEMICO 2023 – 2024

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.

Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

Omniya Eltaher

Acknowledgment

I would like to thank my supervisor, Professor Antonio Parbonetti for his continuous support and guidance,

and I would like to thank Dr. Francesco Ambrosini for all his support and effort.

To my mother, Dr. Hanan, and my father, Dr. Shawky. Thank you for supporting and inspiring me.

To my siblings: Thank you for being my backbone.

Table of Content

Introduction

Corporate sustainability has evolved into a fundamental aspect of modern business strategy, governance, and investment decision-making. The integration of Environmental, Social, and Governance (ESG) factors is no longer a peripheral concern but a central pillar of how companies are evaluated by investors, regulators, and the broader public. Companies are increasingly judged not only by their financial performance but also by how they manage their environmental impact, social responsibilities, and governance structures. As a result, ESG ratings have become a critical tool for assessing corporate sustainability, shaping investor decisions, and influencing corporate reputations.

However, traditional ESG ratings are not without their limitations. ESG scores, typically derived from a combination of qualitative and quantitative metrics, often suffer from a lack of standardization across rating agencies. Different ESG rating providers may apply varying methodologies, resulting in discrepancies between ratings for the same company. This inconsistency creates confusion among investors and stakeholders who rely on these ratings to make informed decisions. Furthermore, ESG ratings are often criticized for being overly focused on public companies, which are mandated to provide detailed sustainability disclosures, while private firms are largely excluded from such assessments due to the lack of formal reporting requirements.

Private companies, which comprise the vast majority of firms globally, play a critical role in the economy. In the European Economic Area (EEA), for instance, 99.87% of firms are private, accounting for 42.8% of corporate assets and employing 61.8% of the workforce (Beuselinck *et al.*, 2021)). Despite their economic importance, private firms typically do not face the same regulatory obligations to disclose ESG information as publicly listed companies. This creates a substantial data gap, leaving investors, regulators, and other stakeholders with limited visibility into the sustainability practices of a significant portion of the business landscape. The lack of transparency among private firms presents challenges for those engaged in sustainable investing, corporate governance, and risk management, as many private enterprises may not provide sufficient data to evaluate their long-term environmental and social risks or their adherence to ethical governance practices.

Additionally, even for public companies, ESG ratings are not without flaws. One major issue is the lack of transparency in the methodology used by different rating agencies. Some agencies may weigh certain ESG factors more heavily than others based on their proprietary models, leading to significant variations in scores across different raters. For instance, a company could receive a high ESG score from one provider due to its governance practices while receiving a lower score from

another because of weaker environmental performance. These discrepancies make it difficult for investors to form a consistent understanding of a company's overall ESG standing.

Another shortcoming of conventional ESG ratings is data reliability. ESG data is often self-reported by companies, leading to concerns about its accuracy and completeness. Companies may selectively disclose favorable information while omitting negative aspects of their environmental or social impact. This selective disclosure can lead to a skewed representation of a company's sustainability profile, which in turn undermines the credibility of the ESG scores. Furthermore, the infrequency of ESG reporting—often on an annual basis—means that the data quickly becomes outdated, failing to capture real-time changes in a company's sustainability performance or response to emerging risks.

A further criticism of standard ESG ratings is their over-reliance on subjective, qualitative assessments. Many ESG metrics, particularly in the social and governance pillars, are based on qualitative judgments rather than objective, quantifiable data. This subjectivity can introduce bias and reduce the comparability of ESG scores across companies and sectors. For instance, while environmental metrics like carbon emissions can be measured quantitatively, social factors such as employee welfare or community engagement are harder to assess uniformly across different firms, industries, and regions.

Moreover, ESG ratings have often been accused of failing to capture the full scope of sustainability issues. For example, some ratings may disproportionately focus on governance structures while underestimating a firm's social or environmental impact. In industries where environmental impact is a major concern, such as energy or manufacturing, this oversight can lead to incomplete assessments of the firm's overall sustainability performance. Additionally, many ESG frameworks fail to account for industry-specific risks or opportunities, which can lead to misjudgments about a company’s true sustainability efforts relative to its peers.

Given these limitations, particularly the lack of focus on private firms and the inconsistencies in ESG ratings for public companies, there is a pressing need to explore alternative methods for assessing sustainability. One such approach is leveraging financial data, which is more widely available and standardized across both public and private firms, to serve as a proxy for sustainability performance. Financial ratios, such as profitability, solvency, and liquidity indicators, can provide valuable insights into a company's operational efficiency and financial health, both of which are often linked to its capacity to engage in sustainable practices.

This study addresses the gap in ESG data for private firms by developing a predictive model that uses financial ratios to assess corporate sustainability. Financial data is readily available for most private firms, making it a useful tool for estimating sustainability performance in the absence of

formal ESG disclosures. The study employs a Random Forest classifier, a machine learning algorithm, to predict the sustainability status of firms based on their financial ratios. By training the model on public companies that provide comprehensive ESG disclosures and applying it to private firms, this research aims to offer a proxy measure for sustainability that is both scalable and reliable.

Chapter One: Corporate Sustainability, Sustainable Investing, and ESG: The Evolution

Environmental, Social, and Governance (ESG) criteria have evolved significantly over the past few decades, becoming a cornerstone of sustainable corporate practices and responsible investing. Initially rooted in the broader framework of Socially Responsible Investing (SRI) in the 1960s and 1970s, ESG has since expanded into a distinct and formalized set of principles that are now integral to corporate strategy, risk management, and investment decisions. The evolution of ESG reflects a growing awareness of the need to balance financial performance with environmental stewardship, social responsibility, and sound governance.

This chapter traces the historical development of ESG, from its early foundations in ethical investment practices to its current prominence in global financial markets. By examining key milestones, such as the creation of the Global Reporting Initiative (GRI), the introduction of the United Nations Principles for Responsible Investment (UNPRI), and the impact of international agreements like the Paris Climate Accord, this chapter provides a comprehensive overview of how both regulatory pressures and evolving stakeholder expectations have shaped ESG criteria. Moreover, the shift from exclusionary screening in SRI to the more integrated ESG frameworks seen today highlights the increasing recognition that sustainable business practices are not just ethically important but also critical for long-term value creation and risk mitigation.

Understanding the historical evolution of ESG makes it clear how the framework has become essential for corporations and investors alike, aligning financial goals with broader societal and environmental objectives. This chapter will lay the groundwork for further analysis of ESG measurement and prediction models in subsequent sections of the thesis.

1.1 Understanding ESG and corporate social responsibility

1.1.1 Historical Evolution of ESG

Environmental, Social, and Governance (ESG) factors represent a multidimensional framework for evaluating the sustainability and ethical impact of businesses. Originating from the broader concept of socially responsible investing (SRI), ESG has developed into a more structured and comprehensive approach, blending sustainability with financial metrics. ESG metrics provide investors and companies with tools to manage risks and opportunities associated with non-financial factors, which are increasingly seen as critical for long-term performance and corporate resilience.

The rise of ESG can be attributed to several key drivers. Climate change, social inequalities, and corporate governance scandals have highlighted the need for companies to integrate sustainability into their business models. The ESG framework evolved to address the environmental impacts of corporations, how they engage with society, and their governance structures. Each of these dimensions brings with it specific concerns and opportunities for businesses and investors.

Environmental factors focus on a company's interaction with the natural environment. This includes issues such as climate change, energy use, waste management, pollution, and biodiversity loss. With the growing global emphasis on sustainability—spurred by major events like the Kyoto Protocol in 1997 and the Paris Agreement in 2015—environmental sustainability has gained prominence in corporate strategies. The Task Force on Climate-related Financial Disclosures (TCFD), established in 2015, further stressed the importance of environmental transparency, encouraging companies to disclose climate-related risks and opportunities. Firms that fail to manage environmental risks, particularly those related to climate change, are increasingly viewed as facing financial instability, as highlighted by Amel-Zadeh and Serafeim (2018).

Social Factors examine how a company manages relationships with employees, suppliers, customers, and communities. This includes labor standards, human rights, diversity and inclusion, and the company's role in broader social issues. Historically, social concerns gained attention during the civil rights movements of the 1960s and 1970s, which laid the groundwork for the rise of socially responsible investing (Sparkes, 2001). Social factors further came to the forefront in the 2000s with movements such as #MeToo and Black Lives Matter, which pushed corporations to prioritize diversity, equity, and inclusion (DEI) within their strategies. As noted by Eccles and Stroehle (2018), social issues have become central to corporate risk management and investor decision-making, especially as companies face growing scrutiny on issues like supply chain ethics, human rights violations, and community engagement.

Governance factors encompass corporate leadership, executive compensation, board diversity, transparency, and shareholder rights. Governance failures, such as the Enron scandal in 2001 and the global financial crisis of 2008, highlighted the risks of poor corporate oversight and accountability. These events catalyzed regulatory reforms aimed at improving corporate governance practices globally. Good governance has become synonymous with ethical corporate behavior, promoting transparency and accountability, which are essential for building investor confidence (Bebchuk & Weisbach, 2010). Companies are increasingly linking strong governance practices to long-term financial performance as they strive to mitigate governance-related risks like fraud, corruption, and regulatory penalties.

The integration of ESG into corporate strategies and investment decisions gained significant momentum in the early 2000s. ESG evolved from a niche concern into a mainstream issue when investors began to recognize the financial materiality of ESG factors. In 2006, the United Nations launched the Principles for Responsible Investment (PRI), which provided a global framework for institutional investors to incorporate ESG into their decision-making processes (UNPRI, 2006). The PRI emphasized that ESG factors are not merely ethical considerations but are central to managing investment risks and improving long-term returns. Since its inception, PRI has grown rapidly, with over 3,000 signatories representing more than $100 trillion in assets under management by 2020 (UNPRI, 2020).

The academic community has also contributed to understanding the relationship between ESG and financial performance. Eccles et al. (2014) found that companies with high sustainability performance significantly outperformed their counterparts in both stock market and accounting performance. This study reinforced the idea that ESG factors, when properly managed, can drive financial success by mitigating risks, improving efficiency, and enhancing a company's reputation. Khan, Serafeim, and Yoon (2016) showed that there is a link between ESG factors, especially material ESG issues, and higher stock returns. This supports the idea that ESG can play a big part in creating financial value.

The environmental aspect, for instance, became increasingly important with the rise of climate-related regulations and the risks associated with failing to address sustainability. The financial impact of environmental mismanagement has become clear in cases like BP's Deepwater Horizon oil spill in 2010, which cost the company billions in fines, clean-up costs, and reputational damage. Conversely, companies that actively manage environmental risks, such as those investing in renewable energy and carbon reduction, have been able to reduce costs and capitalize on new market opportunities (Friede, Busch, & Bassen, 2015).

On the social front, the importance of managing human capital and community relations has also gained prominence. For example, companies with strong employee engagement, fair labor practices, and good relations with local communities are often seen as more resilient and better able to navigate social and operational challenges (Goss & Roberts, 2011). Firms that excel in social sustainability are often rewarded with greater customer loyalty and employee satisfaction, which ultimately translate into better financial performance.

The governance dimension remains a cornerstone of ESG, as strong governance practices are essential for maintaining investor trust and avoiding corporate scandals. Bebchuk, Cohen, and Ferrell (2009) showed that firms with strong governance frameworks, including board independence, shareholder rights, and transparent executive compensation, tend to perform better in the long run. Weak governance, by contrast, can lead to short-termism, conflicts of interest, and even financial collapse, as seen in high-profile cases like Lehman Brothers.

ESG has evolved from its early roots in socially responsible investing into a multifaceted framework that plays a critical role in both corporate strategy and investment decisions. Its integration into mainstream finance is driven by a recognition that addressing environmental, social, and governance issues is essential for managing risk, building resilience, and creating long-term value. As both regulators and investors continue to prioritize sustainability, ESG factors are likely to become even more central to corporate success in the future.

Tracing the roots of SRI:

SRI originated from faith-based values and civil rights movements, particularly in the 1960s and 1970s, with the Vietnam War acting as a significant catalyst. The early SRI focused on "avoidance screens," where investments were excluded from companies involved in industries or practices considered immoral, such as tobacco, alcohol, or weapons production.

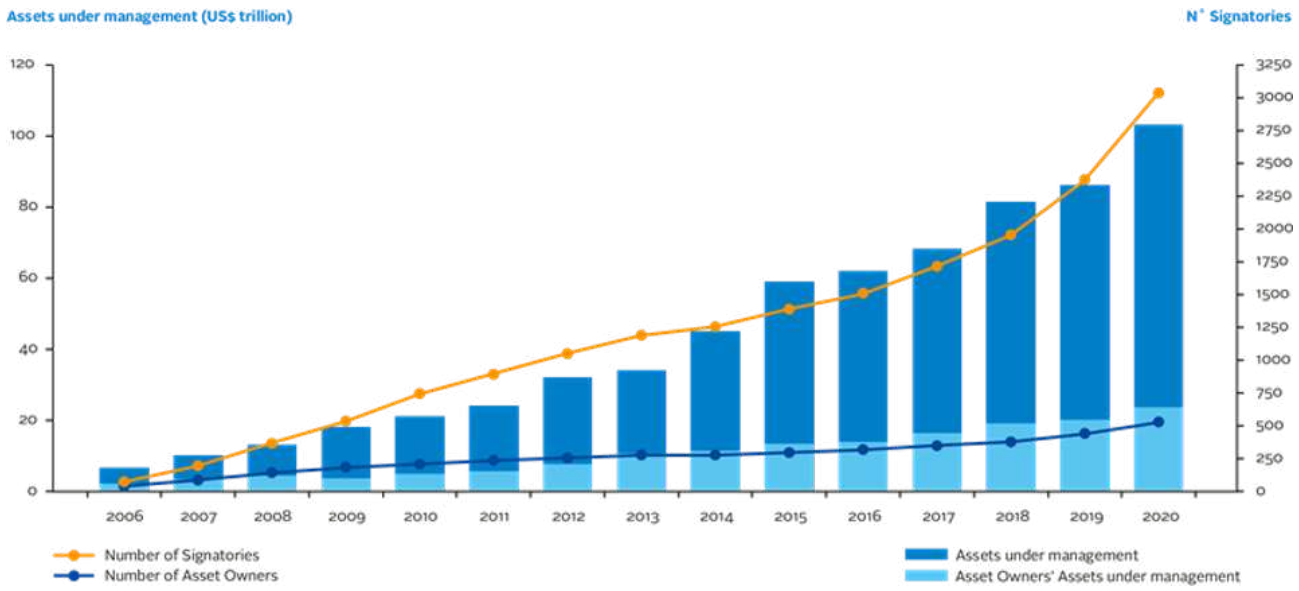

Fig. 1: Overview of the Number of Signatories

SOURCE: (PRINCIPLES FOR RESPONSIBLE INVESTMENT)

The rise of ESG and SRI came from the research-based link between financial performance and environmental, social, and governance risks. 'The investment risks posed by climate change and poor corporate governance provided a huge catalyst in the growth of ESG investing.' (Townsend, 2020). ESG investing brought a proactive approach to analyzing sustainability risks and corporate responsibility, focusing not only on avoiding harmful industries but also on identifying companies that lead in sustainability and corporate governance. Concerns over environmental sustainability, the use of fossil fuels, and global warming started to gain traction in the 1980s. The economic dangers associated with climate change were highlighted by incidents like the 1989 Exxon Valdez oil disaster and the establishment of groups devoted to the issue, such as the Intergovernmental Panel on Climate Change (IPCC). These events prompted pension funds and other institutional investors to include environmental risks in their portfolios. By the middle of the 2000s, studies such as the Freshfields Bruckhaus Deringer study (2005), which contended that ignoring climate change considerations could breach fiduciary obligation, had established a link between climate hazards and fiduciary duty. Moreover, the need for corporate environmental activities to be transparent was further heightened by climate change, which prompted the creation of disclosure frameworks like the Sustainable Accounting Standards Board (SASB) and the Global Reporting Initiative (GRI). Investors were able to evaluate companies' readiness for a low-carbon future more effectively thanks to these frameworks, which is important for their long-term financial performance. Consequently, ESG investing surfaced as a tactic for handling material environmental risk, including disruptions brought on by climate change as well as ethical issues. By the 2010s, global

investors' perspective on sustainability has changed as a result of the increasing importance of ESG factors—particularly climate risks—in mainstream investment decision-making.

However, Several significant challenges prevent socially responsible investing (SRI) from being widely used and having the desired impact. The absence of generally recognized standards for identifying what activities qualify as sustainable or socially responsible is one of the main obstacles. Due to the lack of standardization, it is challenging to compare ESG (environmental, social, and governance) practices throughout various industries and geographical areas, which confuses investors and businesses alike. Evaluations of corporate sustainability initiatives are further complicated by the uneven availability and quality of ESG data, as well as by the differences in reporting standards and criteria. Greenwashing is another problem that damages the reputation of SRI by firms misrepresenting their ESG performance. These issues are made worse by the fragmented regulatory frameworks between nations, as diverse behaviors result from regional laws and cultural norms.

Large institutional investors such as pension funds, sovereign wealth funds, and insurance companies have increasingly recognized the importance of ESG factors in mitigating long-term risks. Their demand for ESG information has pushed companies to improve their transparency and reporting practices. By incorporating ESG elements into their risk management and portfolio strategies, institutional investors have a significant impact on the advancement of SRI. Institutional investors can improve the stability and long-term performance of their investments by taking into account ESG-related risks, such as those resulting from inadequate corporate governance or environmental deterioration. Their impact also extends to promoting changes in corporate conduct, since pressure from large asset managers and pension funds tends to make firms more likely to adopt sustainable practices. Setting industry standards and advocating for increased accountability and transparency in ESG reporting are crucial tasks for institutional investors.

Through several regulatory actions, the European Union (EU) has assumed a leading role in addressing the challenges facing SRI. To provide clarity on what constitutes ecologically sustainable activities, the EU's Action Plan on Sustainable Finance is especially important in its efforts to provide a single taxonomy for sustainable investments. In addition, the plan mandates that asset managers and institutional investors disclose to investors how they include environmental, social, and governance considerations into their decision-making procedures. In order to promote

sustainable growth, the EU is also focusing on creating green bond criteria and benchmarks for low-carbon investments. The European Union (EU) hopes to improve the integration of ESG aspects in Europe and become a global leader in sustainable finance by encouraging regulatory harmonization and establishing clear rules.

The Evolution of Corporate Social Responsibility :

The evolution of Corporate Social Responsibility (CSR) reflects a dynamic interaction between businesses, society, and the evolving expectations of ethical corporate behavior. CSR began as a voluntary, often philanthropic effort and has since become an essential component of corporate strategy. By analyzing two significant works on CSR—Archie Carroll's 1999 article, Corporate Social Responsibility: Evolution of a Definitional Construct, and the 2008 chapter, A History of Corporate Social Responsibility: Concepts and Practices—it becomes evident that CSR's trajectory is closely tied to changing societal values and corporate adaptation. However, while CSR has grown in scope and complexity, the question remains whether it has genuinely transformed the core of corporate behavior or if it remains a strategic response to external pressures.

The evolution of CSR from a reactive, philanthropic effort to a formalized corporate strategy reflects both progress and limitations. While the growing emphasis on CSR has undoubtedly led to more responsible corporate behavior, much of its development appears to be driven by strategic motives rather than genuine ethical transformation. As CSR continues to evolve, its effectiveness will depend on whether businesses can move beyond seeing it as a necessity for survival and embrace it as an integral part of their ethical foundation. Until then, CSR remains a tool for navigating external pressures rather than a reflection of intrinsic corporate values.

CSR's modern origins in the 1950s were largely reactive, responding to public expectations rather than driven by proactive business ethics. Howard Bowen's 1953 book, Social Responsibilities of the Businessman, is widely regarded as the starting point of formal CSR discussions. Bowen emphasized that large corporations, wielding significant power, bore a responsibility to society beyond profit-making. This marked a shift from the earlier, more rudimentary philanthropic efforts of the early 20th century, where businesses primarily engaged in community welfare as an extension of their local relationships.

However, Bowen's framework was still limited by its focus on corporate responses to social pressure. It lacked a more profound, intrinsic sense of ethical obligation from corporations. Companies were still focused on their bottom line, and CSR remained largely voluntary and unsystematic, often a public relations tool rather than a deeply ingrained business philosophy.

By the 1960s and 1970s, CSR began to take on a more structured role within corporate strategies, driven in part by academic discourse. Scholars like Keith Davis expanded CSR's conceptualization by tying it to long-term business interests. Davis's "Iron Law of Responsibility" argued that businesses' social power must be matched by social responsibility, warning that failure to do so would erode their power over time. This approach revealed an inherent tension in CSR's evolution—businesses began to see social responsibility not as an ethical imperative but as a strategic necessity. The proliferation of CSR definitions during this period suggests that companies were increasingly aware of their social obligations, but many acted out of self-interest, viewing CSR as a means to sustain profitability and legitimacy.

Carroll's four-part CSR framework, introduced in 1979, crystallized this shift by categorizing responsibilities into economic, legal, ethical, and discretionary. While this framework acknowledged the multi-dimensional nature of CSR, it reinforced the idea that economic responsibility was foundational, suggesting that CSR was still primarily about aligning social responsibility with profit. The rise of corporate social performance (CSP) models during this period further emphasized measurable outcomes over genuine ethical commitment.

While CSR's formalization in the 1970s and 1980s marked progress in aligning corporate behavior with societal expectations, it also revealed a persistent gap between rhetoric and action. CSR became increasingly institutionalized, particularly with the introduction of standards, audits, and reporting mechanisms in the 1980s and 1990s. However, this institutionalization often seemed more about meeting regulatory requirements or improving public image than fostering genuine corporate ethics.

From this perspective, CSR's evolution appears less about an ethical transformation and more about strategic adaptation. Corporations have learned to navigate social expectations, incorporating CSR into their business models to maintain legitimacy. This raises questions about whether CSR's widespread adoption truly reflects a commitment to social responsibility or if it has been co-opted as a tool for risk management and profit maximization.

1.2 Corporate Sustainability: Definition and Evolution

1.2.1 What is corporate sustainability?

Sustainability can be considered a multi-dimensional concept because it demonstrates a company's ability to efficiently manage its environmental, social, and economic impacts in order to achieve long-term business success. Early definitions of sustainability focus on environmental preservation, while the contemporary corporate sustainability paradigm is a triple bottom line that comprises people, the planet, and profit (Elkington, 1998). Sustainability is increasingly recognized as part of a mainstream strategy rather than a peripheral concern, covering aspects of risk and opportunities that financial analysis misses (Montiel & Delgado-Ceballos, 2014).

Generally speaking, sustainability can be defined as the ability to meet present needs without sacrificing or compromising the ability of future generations to meet their needs. Brundtland Commission, 1987 In the corporate context, it would mean businesses need to balance short-term business objectives with longer-term social and environmental concerns. Major elements of corporate sustainability are responsible resource management, ethical treatment of labor, open governance, and community involvement.

Academic literature on corporate sustainability is more theoretical, philosophical, and focused on sophisticated multidimensional constructs, whereas Practitioner-oriented studies are more prescriptive, detailing clear steps managers can take to make sustainability a part of business strategy. They address how managers can comply with environmental legislation, control environmental risk, and innovate within the firm. Various definitions refer to the Brundtland Report of 1987, which emphasizes sustainable development without compromising the needs of future generations. Scholars allude to the triple-bottom-line approach in balancing environmental and social with economic outcomes (Montiel et al.). Thus, corporate sustainability is the natural extension of the sustainability concept and surrounds it with the same amount of complexity. Multidimensionality can be defined as a holistic approach to conducting business while achieving long-term environmental, social, and economic sustainability. Broader impacts of business operations on external factors are taken more into account as opposed to a solely profit-driven

strategy. Thus, it is built on 3 pillars. The first one is the environmental element, which focuses on the 'ecological footprint' of the company, such as energy consumption and carbon footprint, the second pillar is the social pillar, which is the main aspect of CSR but in a holistic view of corporate sustainability. It concerns two aspects, the internal one, such as diversity and inclusion and fair work practices for the employees, and an external one that focuses on developing in the local communities. The third and final pillar is governance. It aims at the ability of an enterprise to obtain profits and, in the long run, create economic value for all groups of stakeholders, such as shareholders, staff members, customers, suppliers, and the community at large, without infringing on the well-being of any groups. It concerns responsible financial management, proper investment decisions, and the pursuit of profitability with careful attention to the ethical means by which business is conducted.

Moreover, according to scholars, there are multiple approaches to measuring corporate sustainability that are diverse; accordingly, we can classify them into frameworks and theories. CS is measured through diverse frameworks like the Triple Bottom Line and Natural Resource-Based View (NRBV), with no consensus on standardized metrics.

Organizational theories, including stakeholder theory, institutional theory, and resource-based view, are frequently applied to analyze CS, but new frameworks like sustaincentrism have also emerged.

Fig. 2: Overview of Triple bottom line (Żak, 2015)

1.2.2 The evolution of Cs frameworks and theories

The theoretical foundations of corporate sustainability can be traced back to several key frameworks and theories that have shaped its development:

In figure 1 we can see the Triple Bottom Line (TBL) Framework, which was introduced by John Elkington in the 1990s. The Triple Bottom Line Framework emphasizes that businesses should focus not only on financial performance but also on social and environmental impacts. The TBL approach suggests that companies should measure their success based on three pillars: "people, planet, and profit." This framework has been instrumental in encouraging companies to adopt a more holistic view of their operations and to consider the broader implications of their business activities on society and the environment (Elkington, 1998).

Moreover, Building on the work of Freeman (1984), stakeholder theory posits that companies have a responsibility to a broad range of stakeholders, including employees, customers, suppliers, communities, and the environment, in addition to shareholders. This theory challenges the traditional shareholder-centric view of corporate governance and promotes a more inclusive

approach to value creation. In the context of corporate sustainability, stakeholder theory underscores the importance of engaging with various stakeholders to understand their expectations and to develop strategies that address their needs and concerns (Freeman, 1984). Furthermore, The Resource-Based View (RBV) posits that a firm's competitive advantage stems from its ability to acquire and manage valuable, rare, inimitable, and non-substitutable resources (Barney, 1991). In the context of corporate sustainability, this perspective suggests that companies that invest in sustainable practices, such as energy efficiency, waste reduction, and ethical labor practices, can develop unique capabilities that differentiate them from competitors and enhance their long-term profitability and resilience. Furthermore, institutional theory emphasizes the role of social, cultural, and regulatory norms in shaping organizational behavior (DiMaggio & Powell, 1983). This perspective is particularly relevant to corporate sustainability, as companies often face pressure from various institutions, including governments, regulatory bodies, industry associations, and non-governmental organizations, to adopt more sustainable practices. Institutional theory suggests that companies may engage in sustainability efforts to gain legitimacy, align with societal expectations, and mitigate regulatory risks.

Weak vs. Strong Corporate Sustainability :

Corporate sustainability may be differentiated into "weak" and "strong" forms. In the weak version, business interests are leading: environmental and social aspects are considered only insofar as they contribute to enhancing long-term financial results. Companies, within this framing, may thus act in an environmentally responsible way to the extent that it serves profitability, which very often results in superficial adherence to sustainability goals without major structural change. For instance, businesses involved in the oil and gas sectors may elucidate on the protection of the environment but put more effort into ascertaining the business is profitable rather than achieving something useful towards environmental improvement

On the other hand, strong sustainability seeks a shift in governance and legal mechanisms to acknowledge the ecological constraints of the planet. It emphasizes the call for corporations to work within the conditions set by the planetary boundary, respects ecosystems and foster human rights and social needs. Strong sustainability, taking SDGs as a reference for the world, calls for system-level changes in production processes and interactions between society and the natural environment. This form of sustainability calls for profound changes in business models and practices other than strategies oriented by profit alone. (Baumgartner & Rauter, 2017).

Moreover, scholars have defined another approach to sustainability, the system approach. The systems approach to sustainability views sustainable development as the intersection of three interlinked systems: the environmental, economic, and social systems. This approach emphasizes the need to balance the goals of these systems through trade-offs since focusing on one system in isolation may lead to negative impacts on the others. Sustainable development is depicted as the overlap between these systems, where each system's goals are met simultaneously, considering possible trade-offs and interdependencies between them.

1.2.3 Corporate Sustainability in Private Companies

While much of the academic and practical focus on corporate sustainability has traditionally been on publicly traded companies, the principles of sustainability are equally, if not more, critical for private companies. Private firms, which constitute a significant portion of the global economy, often operate under different governance structures and regulatory requirements than their publicly listed counterparts. This unique context presents both challenges and opportunities for integrating sustainability into business operations. Due to the lack of theories and literature on this point we in this paper tried to formulate what are the challenges that a private company could face, Unlike public companies, private firms are not subject to the same level of scrutiny and mandatory disclosure requirements related to ESG performance. This lack of transparency can hinder efforts to assess their sustainability practices accurately and consistently. Without standardized reporting, it becomes challenging for investors and other stakeholders to compare the sustainability performance of private companies, making it difficult to assess their long-term viability and ethical standing. Moreover, Private companies are subject to resource constraints , particularly small and medium-sized enterprises (SMEs), often have limited financial and human resources to dedicate to sustainability initiatives. The costs associated with implementing sustainable practices, such as upgrading to energy-efficient technologies or establishing robust governance frameworks, can be prohibitive for smaller firms. Additionally, private companies may lack the internal expertise needed to integrate sustainability effectively into their business strategies. Additionally The governance structures of private companies are often less formalized than those of public firms, which can impact their ability to implement and monitor sustainability initiatives. Private firms may not have dedicated sustainability committees or board members with ESG expertise, which can limit their capacity to develop and enforce comprehensive sustainability policies. However , according to some scholars there is an opportunity for private companies as well. Private companies, particularly SMEs, often have more flexibility and agility in their decision-making processes compared to larger public firms. This agility can enable them to respond more quickly to

sustainability challenges and opportunities, such as adopting new technologies, entering emerging markets with sustainable products, or aligning with changing consumer preferences toward more ethical and sustainable goods and services. Thus, Private firms are often characterized by a longer-term orientation in their strategic planning, as they are not subject to the same short-term pressures from shareholders as public companies. This focus on long-term value creation aligns well with sustainability goals, as sustainable practices often require significant upfront investment with longer-term payoffs. This alignment can provide private companies with a competitive advantage in markets where sustainability is increasingly becoming a critical differentiator. Additionally, they often have closer relationships with their stakeholders, including employees, customers, suppliers, and local communities. These relationships can provide valuable insights into stakeholder expectations and enable private firms to develop tailored sustainability strategies that address specific needs and concerns. By fostering strong stakeholder relationships, private companies can enhance their reputation, build trust, and differentiate themselves from competitors in the market.

1.3 Fundamental Data and ESG

1.3.1 Financial ratios as predictors of ESG rating

Financial ratios were also identified as some of the very insightful predictors of a firm's sustainability performance, especially with regard to ESG scores. Sustainability is growing in concern in terms of investment considerations, and the financial health of companies is increasingly being used to understand their capacity to handle ESG risks and opportunities. This is particularly beneficial for those companies that do not have direct sustainability reporting because financial ratios provide an idea of how effectively firms balance profitability and operational efficiency in relation to their efforts to manage risks.

Financial ratios depict the images of various features regarding the financial performance of a company. These include profitability, operational efficiency, liquidity, and leverage. These metrics are often related to how well the company under review can implement and keep sustainability practices. ESG scores have typically been associated with higher numbers by profitability indicators like return on assets (ROA) and return on equity (ROE). Companies that have a better ROE or ROA are better placed to spend resources on the implementation of projects regarding sustainability, such as the reduction of carbon emissions or betterment of labor conditions. In the "Heterogeneous Ensemble for ESG Ratings Prediction" research, it was proved beyond any doubt that fundamental

financial data could explain more than half of the variation in ESG ratings, including profitability ratios. This points to the relationship existing between a company's financial performance and its sustainability efforts.

Operational efficiency is another key indicator of sustainability performance. Ratios of efficiency, such as asset turnover and working capital turnover, would reveal a firm's capability in using its assets well.

Generally, firms with high-efficiency ratios tend to do well on ESG metrics, especially under the environmental criteria. Efficient utilization of assets means lower waste, less energy consumption, and, at the end of it, lower environmental footprints—all contributing to better ESG outcomes. The article "Forecasting the Environmental, Social, and Governance Rating of Firms by Using Corporate Financial Performance Variables: A Rough Set Approach" says that companies with higher operational efficiency also tend to have higher ESG scores. Evidently, firms with superior operation in resource management are likely to be sustainable, minimize their negative environmental impacts, and at the same time maximize the output. Sustainability also makes the role of leverage, reflected in ratios like debt to equity, multifaceted. High levels of debt can constrain companies from investment in long-term sustainability initiatives because servicing debt could possibly take priority over funding ESG projects. However, firms with moderate or low leverage are often better equipped to invest in sustainability without risking financial instability. Similarly, the ESG Score Predictor model under Moody's Analytics integrates the level of debt within the firm to be able to hold up the ESG investments over time. Companies with lower levels of debt would be aligned in the context of sustainability, as they would not want to cut back on ESG projects during times of financial crisis.

Financial ratios are a robust tool to predict a company's sustainability performance. They offer a quantitative foundation to assess how financial stability and efficiency really translate into ESG success. The reviewed studies revealed that the Moody's Analytics ESG Score Predictor and "Heterogeneous Ensemble for ESG Ratings Prediction" found that the financial metrics of profitability, operational efficiency, and leverage have quite a significant input into ESG scores. These ratios not only reflect the financial health of a firm but also the ability to reach the goal of sustainability and generate value in the long term.

With the conceptual backdrop in basic ideas of ESG score prediction, the most diverse methodologies—such as those reflected in the Rough Set Approach, Machine Learning, and Random Forest Classification—have been applied to the complicated problem of ESG score prediction. The models offer absolutely unique ways to manage the complex interrelations among

financial performance, operational data, and ESG outcomes. Adequate understanding of these techniques with details is of great importance for evaluation under a variety of contexts but especially for sustainability metric prediction.

The Rough Set Approach (RSA) is a well-suited mathematical method of working on incomplete or uncertain data. In this regard, the authors of "Forecasting the Environmental, Social, and Governance Rating of Firms by Using Corporate Financial Performance Variables: A Rough Set Approach" give RSA special importance in predicting ESG scores when financial data is available and no sustainability data is directly available. These models generate the decision rules from the data of financial performance like profitability, liquidity, and leverage; it is supportive to classify firms' ESG scores. RSA's strength lies in its ability to work with imprecise datasets. Most companies in the present day are not provided with extensive disclosures related to their sustainability issues, and this limits its use on vast data. Its other limitation includes scalability, and hence performance can be slowed on large and complicated data.

In contrast, ML models offer a much more dynamic approach to predicting the ESG score. For example, in the paper "Heterogeneous Ensemble for ESG Ratings Prediction," methods that involve feedforward neural networks and XGBoost or CatBoost in an ensemble model achieve better prediction capabilities. The methodology leverages large volumes of financial and operational data that can explain more than 54% of the variation in ESG ratings. The complex nature of the machine-learning models bestows them with the potential to deal with intricate non-linear relationships between variables, hence well-suited to predict ESG scores based on a set of multiple financial indicators. But it is exactly these properties that add so much complexity to the models and so often lead to a lack of transparency and decreased interpretability, specifically when there are requirements from stakeholders to provide insight into why specific predictions are made.

Another interesting machine learning approach that provides good performance in the field of ESG prediction is Random Forest Classification (RFC) models. The article "ESG Score Prediction Through Random Forest Algorithm" shows how it is possible to make high-accuracy predictions for ESG scores with the usage of RFC. RFC generates multiple decision trees using different random subsets of the given dataset, and the final output is a combination of all tree outputs. This is critically important and therefore makes this method especially effective in modeling nonlinear relationships between financial variables and sustainability outcomes, both in their accuracy and interpretability. RFC would give better predictive results with respect to traditional regression models for handling nonlinearity and variable interactions much more efficiently. Being another model based on machine learning, the RFC requires large data sets for optimal functioning and faces problems with overfitting if not controlled properly.

In the same scope, the paper "Fundamental Ratios as Predictors of ESG Scores: A Machine Learning Approach" makes use of the Random Forest to develop predictions of the ESG scores on the basis of financial ratios. The following study attempts to understand how structural data extracted from the balance sheets and income statements will be able to explain the ESG scores attributed by Bloomberg to companies in the STOXX Europe 600 index . They have found strong predictability of ESG scores from financial statement items such as profitability, liquidity, and solvency ratios. In particular, one study has highlighted the importance of factors like the Net Income-to-Sales ratio (NI_to_Sales) and Sales-to-Assets ratio. It is somewhat unexpected that a rise in NI_to_Sales does not lead to a higher ESG performance, as a higher profit from each sale might be considered indicative of a higher level of sustainability activities. This result is an indication of the strong link between financial performance and ESG performance, requiring inbuilt models that would support these nuances.

Machine learning techniques, specifically Random Forest, do well in recognizing non-linear patterns within the ESG data to be able to detect hidden relations that escape traditional regression models. With the ability to handle multicollinearity and noise in financial data, such flexibility makes these models better suited for predicting ESG scores rather than linear models. Besides, the flexibility of machine learning algorithms is very good, and they improve their predictive power over time as more data becomes available.

However, most machine learning models remain black boxes, which is very much a challenge in ESG contexts, where transparency and accountability are key. The Random Forest algorithm helps to overcome this drawback by including features like variable importance measures, which help identify the most important predictors of ESG scores. This feature makes the model interpretable, and it will therefore give stakeholders insight into which financial variables have the greatest influence on a company's sustainability performance.

Empirical tests further cement Random Forest Classification as a very strong predictor of ESG scores. In comparison to GLM, RFC constantly outperforms the traditional regression methods with much lower RMSE and MAPE scores. This shows how predictive accuracy can be increased using ensemble methods like Random Forest in highly difficult prediction problems such as forecasting ESG scores.

However, even with all these advantages, it is not without its challenges in terms of applying machine learning models to the prediction of ESG scores. Data availability is a major issue, especially for smaller companies or those based in regions where sustainability reporting is not so rigorous. Machine learning models work wonderfully with big data; on the other hand, they require

an enormous amount of data to function effectively. In this case, rough set techniques or simpler machine learning models are more appropriate where data are scarce or incomplete.

The Rough Set Approach and the Random Forest Classification are the two methods that offer corresponding valuable tools for the prediction of ESG scores, coming with their advantages and constraints. The Rough Set Approach has been found to be well adopted in all cases of incomplete or uncertain data, whereas the machine learning models depict higher accuracy and flexibility in handling complex data. In a word, Random Forest Classification is a method that strikes a trade-off between accuracy and interpretability, so it is one of the most effective in predicting ESG scores. These models are bound to play a vital role in the future as ESG reporting becomes quite sophisticated, offering stakeholders sound and transparent assessments of sustainability performance. Companies' integration of financial data, machine learning approaches, and advanced methods of classification will be crucial to evaluate ESG performance across company size and sectors as a basis to promote investments that are well-informed and responsible.

A majority of studies are conducted on public companies for which ESG data are available and more enriched due to necessary disclosures and regulations. This significantly lacks homogeneity in ESG data for private firms. In research, it often takes into account that required data are accessible in adequate amounts; in the case of private firms, data may be incomplete or non-existent. There is no methodology in addressing how to handle or fill in the missing data regarding private entities.

Much of the literature dealing with ESG, focuses on public firms and rarely if ever considers how the nature of private firms, under different regulatory environments and ownership structures, may vary. This then creates a dearth in understanding of the practical application of ESG frameworks in private firms, especially smaller enterprises.

Most of the ESG prediction models in the papers have been developed and aimed at public firms and make use of publicly disclosed data. This indicates the need for models that would be customized to the specificities of private firms, such as their operational scale, ownership structure, and limited disclosure practice (ERM).

Impact of Ownership Structures: Different ownership structures in private firms, such as family-owned businesses or venture-backed firms, greatly impact ESG priorities and reporting. This is another issue on which the current literature has not shed much light and, therefore, remains a gap in understanding heterogeneity within private firms (Bain).

More often than not, the ESG frameworks focus only on broad industries without getting into sector-specific issues for private firms. For example, peculiar challenges around ESG in agriculture, local manufacturing, or even tech startups are not covered in depth in the research.

Moreover, there are no standardized benchmarks for ESG for private firms; thus, it becomes a challenge to compare private firms among themselves or with the publicly listed ones. Such gaps still call for more research on the development of fully comprehensive methodologies to conduct ESG assessments in private firms.

1.3.2 AI intersection with corporate sustainability

Not only ML models can help firms be more accountable and therefore more sustainable, but AI innovations have been at the forefront of environmental, social, and governance concerns in major industries.

The materiality of ESG and environmental concerns and how companies have been trying to reduce waste and carbon footprint; therefore, a surge in the business-to-business collaboration has been growing since. Regarding water management, IBM 23 and VEOLIA 24 provide clients with a monitoring platform that guarantees the integration and optimization of data associated with water management. EMAGIN25 has created artificial intelligence capable of analyzing and learning from data gathered by sensors in water sources. This system forecasts future water usage and provides recommendations to improve efficiency, thereby minimizing water resource wastage.

AI technologies have already been implemented; for instance, ‘ENERGIENCY29 implemented an AI platform that analyzes energy performance. As a result, customers can save up to 20% on their energy bill. The company SENSING VISION30 created an enhanced AI model to identify, analyze, and correct energy anomalies in their buildings. It is based on a network of low-consumption sensors and AI algorithms that process collected data. The company guarantees that its clients will save between 20% and 40% on their energy bills.

Other papers have discussed the positive impact of AI on corporate sustainability performance. According to ” AI and Corporate Sustainability: Exploring the Environmental and Social Impacts of AI Integration," through deep analysis, they have enforced the concept of the positive impact of AI on corporate sustainability; a noteworthy finding was that “AI has the capability to dramatically reduce greenhouse gas emissions by gently adjusting energy consumption in production, logistics, and supply chain processes.” (Nosirov et al, 2021) Through the accurate prediction of market demand or operational demands, AI can reduce waste and energy consumption. Another positive impact is through increasing resource efficiency. By using smart grids and energy systems, firms can reduce their reliance on fossil fuels and enhance their use of renewable energy, therefore reducing their carbon footprint. A noteworthy case study on AI-integrated smart grids demonstrated

a 20% enhancement in energy distribution efficiency, resulting in diminished emissions and decreased operational expenses.

On the other hand, the presence of e-waste generation from the production of hardware as well as AI systems needs a lot of electricity to run, huge models and data centers, and the majority of that electricity is still produced using non-renewable resources. ”An AI model is an energy consumer through its lifecycle, yielding carbon emissions.” Furthermore, other concerns regarding the biases of ML models have been raised as ML models learn from data, and this data is still not regulated and can be biased or false. Moreover, there are more concerns about the privacy of users' data and transparency issues. Especially after the Cambridge Analytica scandal.

Furthermore, artificial intelligence enhances corporate social responsibility by improving transparency in ethical sourcing, particularly inside supply chains. Artificial intelligence and blockchain technology are employed to authenticate fair trade practices, enhancing consumer confidence and sales for ethically sourced items. As well as AI-driven market analysis instruments assist enterprises in penetrating emerging markets, fostering inclusivity and economic development. Nonetheless, optimism is counterbalanced by apprehensions of job displacement resulting from automation. A similar finding was reported in the paper “The Road to Corporate Sustainability: The Importance of Artificial Intelligence," where they have assured that artificial intelligence fosters sustainability by mitigating financial limitations, decreasing agency expenses, optimizing supply chain efficacy, and augmenting worker productivity and resource efficiency. These pathways enhance a company's operational efficiency while aligning it with environmental and social governance objectives.

According to EY “AI models consist of the application of computational tools to build models from examples, data, and experience, rather than following pre-programmed rules.” In the report, they have analyzed how AI can be integrated with ESG. The findings were interesting; for example, in the E component, we can see that AI can be used to manage energy consumption, hence reducing greenhouse emissions. Chen et al. (2021) developed a robust evaluation model utilizing AI approaches for forecasting energy efficiency and conservation. The proposed model demonstrates a notable energy efficiency rating of approximately 97.32%. And it can also be used to monitor climate change, assisting policymakers in creating effective strategies to mitigate the risks of climate change. Finally, it can be used in deforestation monitoring. This will enable conservation organizations to respond promptly.

1.3.3 Contribution to the Literature

This thesis fills a significant gap in the existing literature on ESG score prediction by focusing on private firms, a largely underexplored area. Most existing research on ESG ratings and corporate sustainability centers around publicly listed companies, where data is readily available. However, private firms represent a substantial portion of the global economy, and understanding their sustainability performance is crucial for both investors and stakeholders. By extending the scope of ESG prediction to private companies, this study contributes a novel perspective to the literature.

Additionally, this thesis employs financial variables over three years to predict ESG scores, a departure from the common approach of using publicly disclosed ESG data. While previous research has explored the relationship between corporate financial performance (CFP) and ESG ratings, there has been limited exploration of how financial indicators can be used as predictors for firms that do not publicly disclose their ESG performance. This adds a new dimension to the discussion of how private firms might be evaluated through the lens of corporate sustainability.

From a methodological standpoint, this study introduces machine learning techniques, specifically Random Forest classification, to predict ESG scores. Much of the existing literature relies on traditional linear regression models or the Rough Set Approach. By employing a non-linear, robust model like Random Forest, this thesis enhances the accuracy and reliability of ESG predictions, especially in cases where ESG ratings are unavailable, such as in private firms. The use of Random Forest classification allows for the capture of more complex patterns in the data, offering a more dynamic approach to ESG prediction.

Finally, the comparison between different predictive models—ranging from regression-based methods to machine learning techniques—provides insights into the most effective methodologies for ESG score prediction. This comparative analysis bridges a gap in the current literature by evaluating the efficacy of machine learning models in contrast to more conventional approaches. As a result, this study not only contributes a novel predictive model but also provides empirical evidence on the applicability of these models in different types of companies, particularly in the private sector.

Chapter Two: Institutional Setting

2.1 Overview of Sustainability Rating Agencies and ESG Reporting Frameworks

2.1.1 Sustainability Rating Agencies

ESG (Environmental, Social, and Governance) ratings have become a central tool for investors and stakeholders to assess a company's commitment to sustainability and ethical practices. Various agencies provide ESG ratings, each with its unique methodology, strengths, and challenges. While these ratings offer valuable insights, the inconsistency across providers can sometimes cause confusion. This overview discusses some of the most prominent ESG rating agencies and takes a closer look at Refinitiv, one of the leading providers in this space.

1. MSCI ESG Ratings

MSCI is one of the most widely recognized providers of ESG ratings. Their methodology focuses on evaluating companies based on 37 key ESG issues, grouped into three main pillars: environmental, social, and governance. These issues are further divided into ten themes, such as climate change, corporate behavior, and human capital management. MSCI assigns ratings on a scale from AAA (best) to CCC (worst), using peer comparisons within the same industry to determine a company’s risk exposure.

MSCI’s strength lies in its industry-specific approach, which accounts for how different sectors face unique challenges. However, one common criticism is that it often favors companies that disclose more information, regardless of the actual quality of their ESG practices. Companies that excel in transparency may receive higher ratings even if their sustainability impact is mediocre. This makes it important for investors to not solely rely on MSCI ratings without diving deeper into the company’s actual practices(ratings that don't rate).

2. Sustainalytics

Sustainalytics, a part of Morningstar, is another major player in ESG ratings. It scores over 7,000 companies on a scale of 0 to 100, focusing heavily on how well they manage their ESG risks. Sustainalytics divides its ratings into three categories: preparedness (how well a company is set up to handle future risks), disclosure (how transparent a company is about its ESG efforts), and performance (how well the company is actually doing on key ESG metrics).

One of Sustainalytics' strengths is its sectoral analysis, which ensures that companies are evaluated relative to others in their industry. This approach provides a more tailored assessment of how well

companies are dealing with ESG challenges specific to their operations. However, similar to MSCI, Sustainalytics has been criticized for rewarding companies that disclose more information, regardless of actual performance. This can lead to discrepancies in how companies with similar practices are rated (ratings that don't rate) .

3. RepRisk

RepRisk takes a different approach by focusing on ESG risks and controversies. Rather than relying on company disclosures, RepRisk gathers data from over 80,000 public sources, including news, regulatory filings, and social media. Its scoring system ranges from AAA to D, and it updates daily, providing real-time insights into a company's risk profile.

RepRisk's advantage is its focus on external data, which provides a more objective view of a company's actual ESG impact. It avoids the issue of inflated ratings due to strong disclosure practices by emphasizing real-world controversies and risks. This can be particularly valuable for investors who are concerned with identifying companies that might face reputational or regulatory risks in the future.

4. Refinitiv

Refinitiv, part of the London Stock Exchange Group, has become a leading provider of ESG ratings in recent years. Its methodology draws from over 400 data points across ten main categories, including emissions, diversity, and governance structure. Refinitiv uses publicly available data to generate scores for over 7,000 companies worldwide.

One of Refinitiv's strengths is its comprehensiveness. It covers a wide range of ESG metrics, ensuring that companies are assessed from multiple angles. Refinitiv also places a strong emphasis on governance, a crucial aspect of ESG that sometimes gets overshadowed by environmental and social issues. By incorporating both quantitative and qualitative data, Refinitiv provides a balanced view of a company's ESG performance.

However, like other agencies, Refinitiv faces challenges related to data availability and consistency. In cases where companies do not disclose enough information, Refinitiv may need to rely on estimates, which can introduce some uncertainty into the ratings. Despite this, its broad data coverage and transparent methodology make it a valuable tool for investors looking to integrate ESG considerations into their decision-making(What is ESG ratings) .

ESG ratings are an essential tool for evaluating corporate sustainability, but the lack of standardization across agencies like MSCI, Sustainalytics, RepRisk, and Refinitiv creates challenges for investors. Each agency has its unique strengths, whether it's MSCI's detailed

industry comparisons, Sustainalytics' focus on risk management, or RepRisk's emphasis on controversies. Refinitiv stands out for its comprehensive and transparent methodology, though it, too, is limited by the availability of data. To make informed decisions, investors should consider multiple ESG ratings and dig deeper into the data behind the scores.

ESG reporting frameworks play a crucial role in enhancing corporate transparency, accountability, and sustainability performance. These frameworks are designed to guide companies in their disclosure of environmental, social, and governance factors that investors and stakeholders increasingly prioritize. Expanding on the key frameworks, challenges, and emerging trends provides a more comprehensive understanding of the evolving ESG landscape.

2.1.2 Key ESG Reporting Frameworks and ESG Indices

Global Reporting Initiative (GRI): The Global Reporting Initiative (GRI) is one of the most widely recognized and comprehensive ESG reporting frameworks. It provides organizations with guidelines to report on their economic, environmental, and social impacts. The GRI framework emphasizes the importance of transparency in the disclosure process, encouraging companies to provide a holistic view of their sustainability performance. With GRI, organizations can use a set of standardized metrics to report on issues like emissions, resource usage, labor practices, and community impact. As sustainability reporting becomes a global expectation, GRI allows stakeholders to evaluate and compare corporate sustainability efforts across industries and regions.

Sustainability Accounting Standards Board (SASB): SASB focuses on developing industry-specific standards for reporting material sustainability information to investors. Its main goal is to help companies identify and report on sustainability factors that are likely to impact their financial performance. SASB's industry-specific approach ensures that companies disclose information relevant to their business activities. This framework is particularly favored by investors seeking to integrate sustainability factors into their investment decisions, as it allows for more precise and comparable ESG data based on sector-specific risks and opportunities.

Task Force on Climate-related Financial Disclosures (TCFD): TCFD has gained prominence as the primary framework for reporting climate-related financial risks and opportunities. It provides recommendations for disclosing information on the governance, strategy, risk management, and metrics and targets related to climate change. By aligning financial disclosures with climate risks, TCFD helps businesses communicate how they manage climate-related challenges and opportunities. This is particularly important in the context of global efforts to combat climate

change, as investors and regulators increasingly demand transparency on how companies are preparing for a low-carbon economy.

Integrated ESG Index: Emerging methodologies such as integrated ESG indices are being developed to address the challenges of inconsistent reporting. These indices aim to provide a more standardized approach to measuring and reporting ESG performance, allowing for better comparability across industries. For instance, the shipping sector is developing tailored methodologies to assess environmental impacts, given the sector's significant contributions to global emissions. These sector-specific indices are designed to enhance transparency while accounting for unique industry challenges.

As ESG reporting evolves, there is a growing recognition of the need for interdisciplinary approaches to enhance the robustness of disclosure mechanisms. This involves collaboration between sustainability experts, financial analysts, legal advisors, and data scientists to ensure that ESG reports are comprehensive and reliable. Interdisciplinary teams can help companies navigate complex ESG issues, such as regulatory compliance, risk management, and stakeholder engagement, ensuring that sustainability performance is integrated into core business strategies.

Moreover, Private firms, particularly those in private equity or family-owned businesses, tend to disclose significantly less information about their operations compared to public firms (Markarian et al., 2024). Since there is no legal mandate to report ESG data, many private firms do not provide detailed accounts of their sustainability initiatives. This limited transparency makes it difficult for stakeholders, such as investors or rating agencies, to conduct thorough ESG assessments. The lack of reliable ESG data hampers the ability of investors to evaluate the firm's long-term risks and opportunities related to sustainability. Consequently, private firms may be overlooked in ESG-conscious investment strategies, despite potential strong ESG performance.

Additionally, Private firms also face challenges in discerning the sincerity of ESG demands from stakeholders, especially investors (Zaccone & Pedrini, 2020). With the rise of ESG as a critical factor in investment decisions, firms must allocate resources to meet stakeholder expectations. However, not all stakeholders may be equally committed to sustainability; some may view ESG merely as a compliance or marketing tool. As a result, firms may struggle to identify which requests are genuine and which are driven by superficial trends or short-term interests. This uncertainty can lead to suboptimal resource allocation, where firms may invest in ESG areas that do not align with

their core values or long-term strategies, potentially undermining the overall impact of their sustainability efforts. Despite these challenges, the integration of ESG factors is becoming increasingly crucial for private firms seeking investment and long-term value creation. There is a growing recognition that sustainability practices enhance resilience, attract ESG-conscious investors, and improve brand reputation. As awareness of environmental and social impacts increases, private firms that proactively address ESG issues may gain a competitive advantage in the market. Investors and stakeholders are starting to demand more transparency and accountability, pushing private firms to embrace ESG frameworks despite the challenges associated with standardization, data accessibility, and authentic engagement with stakeholders.

Lastly, while private firms face unique obstacles in ESG assessment, addressing these challenges can offer significant rewards, fostering sustainable growth and long-term value in an evolving investment landscape.

3.2 CFP and ESG ratings:

3.2.1 Why corporations should focus on ESG efforts

1. Profitability and ESG ratings It is also well substantiated that companies with high ESG ratings frequently realize better profitability, including the metrics of return on assets and return on equity.

The meta-analysis in the "ESG and Financial Performance" paper finds that, overall, around 90% of studies report a positive or neutral relationship between ESG and financial performance. In lay terms, this means that companies pursuing business in a manner that is respectful of the environment and good for both people and planet actually end up with more money and, far from losing money, do not lose a thing financially.

Another paper, "Corporate Sustainability Performance and Firm Value," further specifies that a huge driver of profitability is the social aspect of ESG. Firms engaging with their communities, treating their employees well, and being transparent will, over time, show the fruits of these actions in the financial sector, as they build trust and loyalty from customers and investors alike.

2. Investment Efficiency and ESG Ratings Firms that have strong ESG practices are likely to be more efficient with their investments; in other words, they allocate resources much better. They make smarter decisions on where and how to invest their resources.

The Corporate Sustainability Performance and Firm Value paper illustrates this by finding that firms with high ESG ratings, particularly being rated good in the social dimension, lead to better investment decisions. Why? Because they have less information asymmetry (i.e., investors trust them more) and stronger relationships with stakeholders. This leads to better financial outcomes overall.

In addition to this, companies with good ESG make more capital. As demonstrated by Salfino and Kitzes in an article named "Financial Performance Shortfall and ESG Controversies," companies experiencing financial difficulty can obtain capital with a high success rate if the companies have an excellent ESG record because that proves to the investors that they will stand in good stead for the long term.

3. Risk Management and ESG Ratings High ESG-rated companies by and large manage their risks rather well. And the business performance of those firms—the volatility, generally—has been rather smoothed because it has taken all these proactive steps of managing environmental, social, and governance risks.

The "ESG and Financial Performance" study reads: "firms with higher ESG scores were less prone to financial shocks. Why? They're better prepared for things like regulatory changes, environmental disasters, or social backlash, so their earnings tend to be more stable."

Finally, "Beyond Dichotomy: The Curvilinear Relationship between Social Responsibility and Financial Performance" discusses that investment in ESG, but to a reasonable extent, will possibly lead to lower financial risks. Companies that strike the right balance in their ESG efforts are better placed to navigate uncertainty, and this ultimately helps protect their financial bottom line.

4. Cost of Capital and ESG Ratings A more concrete benefit from strong ESG performance is a reduced cost of capital. This includes reducing the associated costs of companies when borrowing money, either through loans or issuing shares.

In the paper "Corporate Sustainability Performance and Firm Value," it is shown that companies who practice good ESG, in particular governance, are in the lucky position of enjoying low borrowing costs. Investors and banks are more amenable to lending them at good rates because they trust these companies to be transparent and to manage their resources wisely.

The paper "ESG and Financial Performance" reinforces this view, arguing that firms with strong ESG performance could reduce their risk premium, meaning they might have cheaper access to capital, especially in emerging areas such as green bonds.

5. Firm Valuation and ESG Ratings Last but not least, companies with good ESG ratings usually realize a hike in their value. In other words, the stock market and investors show more interest in these companies, which elevates their market value.

The paper "Financial Performance Shortfall" presents that firms facing financial performance shortfall often resort to taking on ESG initiatives to maintain or increase their market valuation. ESG activities further strive toward enhancing a company's reputation so that it is able to recover from financial challenges and once again attract new investors.

It is in the "Strategic Management Journal" where it is mentioned that companies which moderately invest in ESG practices generally get the highest valuation gains. overdoing it all may not always be beneficial, but getting it right does lead to more investor confidence and, in due course, higher stock prices.

2.2.2 The investor revolution: why institutional investors care about sustainability.

The quantity of companies that quantify and disclose environmental (e.g., carbon emissions, water usage, waste production), social (e.g., employee, product, customer-related), and governance (e.g., political lobbying, anti-corruption, board diversity) data has surged dramatically in the last twenty-five years. This information is collectively termed ESG data. In the early 1990s, fewer than 20 organizations supplied ESG data; by 2016, almost 9,000 companies had published sustainability or integrated reports. As to EY, investors are progressively convinced that organizations excelling in ESG are less hazardous, more strategically positioned for the long term, and more equipped to handle uncertainty. (EY.2021)

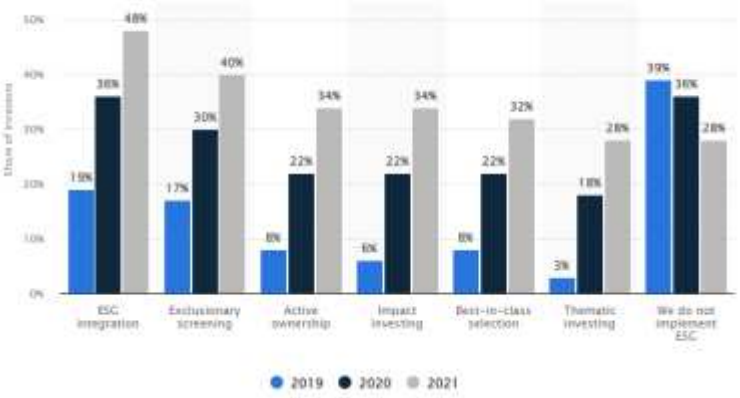

Fig. 3 Most common method for ESG adoption among institutional investors worldwide from 2019 to 2021 (Statista, 2021)

Recent research has demonstrated that ESG information correlates with several economically significant outcomes. ESG disclosures correlate with reduced capital limitations (Cheng et al. 2014), diminished cost of capital (Dhaliwal et al. 2011), decreased analyst forecast mistakes (Dhaliwal et al. 2012), and stock price fluctuations surrounding required ESG disclosure rules (Grewal, Riedl, and Serafeim 2017). Furthermore, industry-specific categories of materiality discern ESG information that is pertinent to value and indicative of organizations' future financial performance (Khan et al.). Although these studies illustrate substantial economic impacts, a comprehensive understanding of the reasons and methods by which investors utilize ESG information, as well as the associated challenges, remains elusive. To augment our understanding and supplement archive data, we conducted a survey in partnership with a multinational financial organization (i.e., BNY Mellon).

In the paper ‘Why and How Investors Use’ ESG information they sampled 413 responses from senior investment professionals with an average response rate across the questions in the survey of 9%. This is in line with other surveys that have collected responses from CFOs and obtained rates of 9, 8.4, and 5.4% (Graham and Harvey 2001; Graham, Harvey, and Rajgopal 2005; Dichev et al. 2013). They examined the factors that drive investors to utilize ESG data. A significant majority of respondents (82%) indicate that they utilize ESG information due to its financial relevance to investment performance. A greater percentage of US investors, compared to their European counterparts, believe that the information is immaterial for investment reasons (22% versus 4%) and that utilizing the information would breach their fiduciary duty (22% versus 8%). The latest finding is noteworthy given the US Department of Labor's guidelines, which underlines that the integration of ESG information in investing decisions aligns with fiduciary responsibilities. A considerable

proportion of the sample evaluates the material from an active ownership perspective (Dimson et al. 2015).

A considerable proportion of the sample evaluates the material from an active ownership perspective (Dimson et al. 2015). They contend that interaction with corporations can effectuate change in the corporate sector and tackle ESG issues; nonetheless, this conviction is predominantly prevalent among European investors. An identical proportion of the sample regards ESG information due to increasing client demand or official client directives.

This applies to larger asset managers, who tend to be responsive to the necessity of integrating ESG information. A diminished proportion of respondents regards such information as an ethical obligation, with European investors being more inclined to perceive it as such.

And to answer the question, how do institutional investors use the data? An equivalent proportion of investors utilize ESG information for engaging with firms or as a component in their value models (37% and 34%). A considerable proportion of the respondents utilize the information to delineate the investment universe via a screening that involves the exclusion of sin stocks and companies associated with ESG problems. Positive screening procedures, whether applied across industries or within a specific industry (i.e., best-in-class), remain infrequent at 13%. Approximately 13% of survey respondents utilize portfolio overlay, risk factor, and theme styles. Respondents indicate that negative screening is perceived as the most harmful to financial performance across several ESG methods. Complete integration into stock valuation, active ownership, and positive screening are deemed the most advantageous. (How a day investors use ESG), and for the third point, The future applications of ESG data for investors. Looking ahead, according to the authors the respondents anticipate that, among several ESG approaches, positive screening and active ownership will gain significance. Negative screening, theme investment, best-in-class evaluation, and comprehensive integration in stock valuation are anticipated to diminish in significance, on the other hand, the focus on the materiality of ESG is possibly on the rise.

The increasing importance of ESG matters to investors is shown in the recent 2020 EY Climate Change and Sustainability Services (CCaSS) Institutional Investor survey. The poll revealed that 98% of investors evaluate ESG, with 72% conducting a systematic analysis of ESG performance, in contrast to only 32% in the prior survey done two years ago. Furthermore, a significant portion of individuals employing an informal method intends to transition to a more stringent framework (39%). Moreover, in the article A Quest for Clarity by Mickesny & Company ESG is a significant consideration for almost 85% of the chief investment officers that respondents. Sixty percent of

participants evaluate their total portfolio for ESG factors, while over eighty percent analyze specific firm positions on the impact of ESG on projected cash flows. A notable majority are willing to pay a premium for companies that demonstrate a clear correlation between their ESG initiatives and financial performance.

Chapter Three: Empirical Analysis

Private companies are the backbone of the global economy and often remain unlisted throughout their existence. Their significance is especially pronounced in various regions, including Europe. According to data from Amadeus—a prominent research database on European public and private corporations maintained by Bureau van Dijk—99.87% of firms in the European Economic Area (EEA) are private. These companies account for 42.8% of total corporate assets and employ 61.8% of the European workforce (Beuselinck et al., 2021). Given their substantial impact, this thesis aims to explore how to determine the ESG ratings of private firms in Europe by developing a machine learning model.

Despite their economic importance, private companies often lack standardized ESG reporting. This absence poses a challenge for investors, regulators, and other stakeholders seeking to evaluate their sustainability performance. Unlike public firms, private companies are not legally obligated to disclose ESG information, resulting in a significant data gap. This study addresses this gap by investigating alternative methods for estimating ESG ratings, focusing on the use of financial ratios. The hypothesis is that financial ratios, commonly available in private companies' financial statements, can effectively predict their ESG ratings. By developing a machine learning model—specifically, a Random Forest classification model—trained on data from public firms with known ESG ratings, we aim to predict the sustainability performance of private companies based on their financial metrics.

The findings of this study will provide valuable insights into how financial data can be utilized to estimate the ESG performance of private companies. This can potentially guide investors and stakeholders in making informed decisions in the absence of direct ESG disclosures.

It is important to note that, while most private firms globally are not required to report their annual financial position, a considerable proportion of European companies are mandated to disclose their financial statements. The EU Accounting Directive for non-IFRS reporting for small to large companies ensures a certain level of consistency in financial reporting across Europe. This directive obliges all limited companies to submit their financial statements annually, including at minimum the balance sheet and profit and loss statements. As a result, most previous studies examining the relationship between ESG ratings and financial performance have focused on listed companies.

The existing literature on ESG and financial performance can be divided into two main areas. The first examines how ESG ratings influence financial performance, often finding a positive correlation between high ESG ratings and increased profitability. The second area focuses on predicting ESG ratings for public firms using both financial and non-financial variables. These financial variables are often related to the characteristics of listed companies, such as earnings per share or market capitalization. In terms of data, some researchers advocate using market financial data, while others emphasize the importance of fundamental financial ratios. This research builds on the latter, using fundamental ratios as the baseline for predicting the sustainability of private companies through a machine learning model trained on public company data.

Most previous research has concentrated on specific industries within limited geographic regions. In contrast, our study encompasses all industries across Europe. To enhance comparability with listed companies and reduce data noise, our research applies an even stricter threshold, focusing on companies with total assets of 500 million euros and annual revenue of 100 million euros. This threshold has been consistently applied to both private and public companies in our study over a three-year period.

3.1 Data collection and variables selection

3.1.1 Public Firm Data

ESG Data:

To indicate whether a company is sustainable or not and to get the most precise data, we have used environmental, social, and governance scores from LSEG data and analytics in the form of grade letters. LSEG Data & Analytics, formerly Refinitiv, is an American-British global provider of financial market data and infrastructure. LSEG's ESG scores seek to deliver transparent, precise,

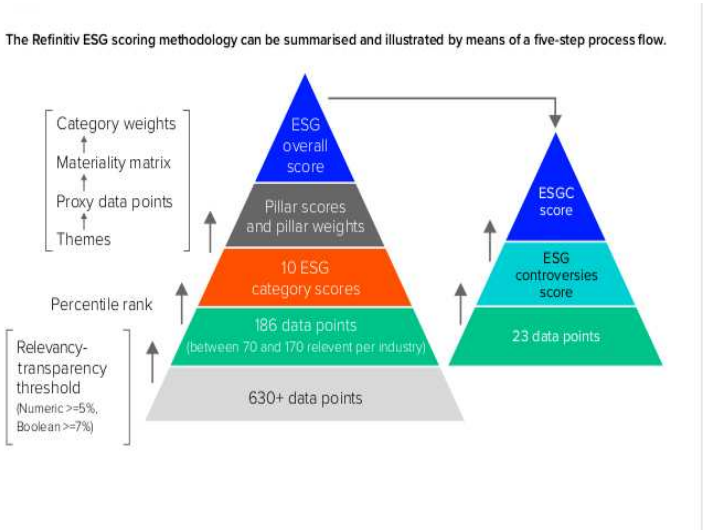

Fig. 4: Overview of Refintive Scoring Methodology

and comparable ESG data. The rankings derive from more than 630 distinct ESG measures encompassing over 12,500 public and private firms worldwide. The scores assess a company's relative ESG performance, dedication, and efficacy. Refinitiv's ESG methodology considers sector-specific materiality and firm size to minimize bias. Over 700 content research analysts worldwide compile ESG data. They collect data from several sources, including corporate reports, stock exchange filings, and news outlets. Data is regularly updated and recalculated weekly, emphasizing quality control through both algorithmic and manual verification.

The Figure by Refintive illustrates Refinitiv's ESG score methodology, commencing with the aggregation of over 630 distinct ESG data points from companies' public disclosures. From this

selection, only 186 are deemed relevant to any particular industry, guaranteeing that solely the most significant measures for that sector are utilized. The pertinent data points are categorized into 10 ESG categories, encompassing specific concerns like environmental management, human rights, emissions, and workforce diversity.

The category scores are aggregated into three primary pillars: Environmental, Social, and Governance. The scores are weighted based on their significance within each industry, indicating that the most pertinent issues for that sector exert a stronger impact on the overall ESG score. The overall score is derived from the aggregation of pillar scores and indicates the company's success in ESG based on publicly available data.

Refinitiv employs a percentile rank system to assess scores, thereby comparing a company's performance to that of its rivals. A transparency criterion exists whereby numeric data points are incorporated if over 5% of organizations report them, and boolean (yes/no) data points are included if more than 7% report them. This guarantees that the measures employed are broadly applicable and pertinent.

The ESG controversy score is presented on the right side. This score is derived from 23 contentious issues, including environmental disasters and human rights breaches, that may impact the company. This score is aggregated with the overall ESG score to yield the final ESGC score. This modification considers any substantial controversies, offering a more thorough assessment of the company's overall ESG performance.

The data we used had a threshold of 500 million euros in total assets or 100 million in total revenue; it covered several countries in Europe. The data presented 1,717 public firms We found out that according to the data of public companies, Finland, Germany, and Austria are the top 3 countries with the number of companies that have an above-average ESG rating in their public companies. Furthermore, the most frequent country in our dataset is Germany with 189 companies, followed by France with 159 companies. These findings need further exploration.

The rationale for using total assets as a proxy:

Total assets are a commonly employed criterion for evaluating firm size in empirical research, Augmented total assets may indicate enhanced resources, stronger financial stability, and an increased power to influence corporate practices such as ESG reporting and compliance. In the context of this thesis on corporate sustainability, large firms (with total assets above 500 million) are more predisposed to invest in ESG practices and provide transparent ESG disclosures. This may make them a more relevant focal group for such studies.

The 500 million criterion is not arbitrary; it serves to classify corporations as "large" organizations. The Dang et al. article underscores that size proxies represent various dimensions of corporate behavior. Large corporations face greater scrutiny, potentially increasing their propensity to adopt sustainable practices, which is linked to their corporate finance conduct and ESG reporting. Empirical studies frequently employ asset-based thresholds, such as 500 million, to diminish sample heterogeneity, facilitating a more concentrated examination of the financial strategy, corporate governance, and sustainability initiatives of major corporations. In another study, "How Firm Size Shapes the ESG and Financial Performance Nexus: Insights from the MENA Region," they related their findings that larger firms have a stronger positive relationship between ESG performance and financial performance with a resource-based view (RBV).The results indicate that employing business size as a proxy mitigates sample heterogeneity. The report differentiated organizations by size, offering insights into the influence of ESG performance on financial success across various firm sizes. The data indicates that larger organizations possess the means to participate in ESG activities and are also more inclined to reap financial advantages from such engagement because larger firms tend to have more resources, competitive advantages, and greater visibility. The regression performed by the authors indicates that the interaction between business size and ESG performance is significant, suggesting that firm size is pivotal in enhancing the financial gains derived from ESG activities. Specifically, larger firms in the sample demonstrated a more pronounced positive correlation between ESG performance and return on assets (ROA) than smaller firms.

As for the time period covered, our data covered 3 fiscal years. To make sure the data is cohesive, we have dropped any companies that failed to report their ESG scores or Refintive did not assign them a score for one fiscal year. Furthermore, we created an ordinal sequence following the same logic of A+=1, A=2, A-=3..etc) This ordinal sequence allowed us to create a binary variable called sustainable, allocating one to sustainable and zero to not sustainable to use it as two classes in order to train the random forest classifier. More will be discussed in the data processing part of the thesis.

Table 1: Financial Variables Description

Variable	Description	Justification
Sales_to_Assets	The firm's total revenue is divided by its average total assets.	Measures asset efficiency
Solvency Ratio	Directly obtained from ORBIS; Proportion of total debt to total assets.	Assesses the company's long-term financial risk.
Liquidity Ratio	Directly obtained from ORBIS; measures a company's ability to meet short-term obligations using current assets.	Indicates short-term financial health
NI_to_Sales	Net Income divided by total sales	Reflects the company's ability to convert revenues into profit
EBIT_to_sales	Earnings Before Interest and Taxes (EBIT) divided by total sales.	Indicates operational profitability in relation to sales
ROA	Directly obtained from ORBIS; measures profitability relative to total assets.	Higher ROA indicates more efficient resource use
Industry Classification	Categories based on Bureau van Dijk (BvD) classification.	Captures industry-specific ESG materiality
Sustainability Variable	Created based on the median ordinal scale of ESG letter grades (A+ to D-).	Provides a clear classification for companies based on their ESG performance

Fundamental Data:

In this study, given the restricted access to extensive ESG data from platforms like Refinitiv for private enterprises, we utilized the ORBIS database to acquire the requisite financial information. This study aims to predict ESG scores through quantitative financial indicators and one categorical variable, "industry," which has been one hot encoded, in contrast to ESG scores, which are generally derived from a blend of qualitative and quantitative measurements. The selected variables are essential financial ratios that are both commercially accessible and mandated by law for private

enterprises in the European Union, as they are obligated to report these measures annually. By establishing criteria for the firms in the dataset—500 million euros in total assets and 100 million euros in revenue—over a three-year fiscal period, we guarantee that our sample comprises firms of considerable magnitude, enhancing comparability and excluding smaller firms that may demonstrate inconsistent financial reporting or volatility (Peters & Romi, 2014) . For each prediction year, the ESG score typically corresponds to the financial performance and other relevant data from the previous fiscal year.

This analysis selects net income-to-sales ratio, sales-to-assets ratio, EBIT-to-sales ratio, solvency ratio, liquidity ratio, and ROA. In addition to these financial measurements, we included one qualitative variable: industry classification, categorized according to Bureau van Dijk (BvD). The incorporation of industry is essential, as companies across diverse sectors have distinct challenges and possibilities regarding sustainability (Khan, Serafeim, & Yoon, 2016). For example, companies in industrial and energy sectors may possess distinct environmental or social concerns relative to those in service-oriented industries. The use of a consistent classification system ensures sectoral analysis uniformity.

Four variables—liquidity ratio (current ratio), solvency ratio, ROA, and industry classification—were directly obtained from ORBIS. Nevertheless, for the remaining three financial ratios—Net Income-to-Sales Ratio, Sales-to-Assets Ratio, and EBIT-to-Sales Ratio—we computed them manually employing established financial formulas:

The sales-to-assets ratio was driven by dividing the firm's total revenue by the average of its total assets over time. This indicator assesses asset efficiency, indicating a corporation's effectiveness in leveraging its assets to create income.

The EBIT-to-sales (Operating Margin) ratio was determined by dividing the EBIT by sales. This ratio reflects the firm's operational profitability in relation to its total sales, offering insight into the efficiency of the company's operations in generating earnings.

The net income-to-sales (Net Profit Margin) ratio was calculated by dividing net profit by revenue for each fiscal year. This ratio indicates the company's capacity to transform revenues into profit, offering a transparent assessment of total profitability.

The inclusion of these ratios provides a comprehensive view of the firm's financial performance. Profitability measures, such as the net income-to-sales ratio and EBIT-to-sales ratio, reflect the firm's ability to generate earnings from operations as well as the company's overall profitability. These ratios are more informative to absolute values, contribute to improving the characterization of companies, and reflect a faithful and explicative understanding of the business activities and the financial performance of a company explaining the ESG scores ((D'Amato et al., 2021)) Conversely, solvency and liquidity ratios emphasize financial stability, offering insights into the firm's capacity to fulfill its long-term and short-term financial obligations, respectively. The amalgamation of these variables for a thorough evaluation of a firm's financial well-being, a crucial factor for long-term viability.

While ESG scores are predominantly derived from qualitative evaluations of a company's environmental, social, and governance procedures, financial performance significantly influences a company's capacity to attain sustainability objectives. A company exhibiting robust profitability and elevated solvency is more adept at investing in sustainable practices and mitigating environmental and social risks (Clark, Feiner, & Viehs, 2015). This study enhances the existing literature by correlating financial measures with ESG outcomes, aiming to quantify the relationship between financial performance and sustainability.

Table 2: Descriptive Statistics of Key Variables

Measures	***M***	***SD***	**Range**
Sales_to_Assets	0.73	0.51	4.055456021
Solvency Ratio	39.42	19.17	154.136
Liquidity Ratio	1.81	2.94	55.318
NI_to_Sales	21.67	287.55	7274.55468
EBIT_to_sales	0.15	0.19	1.79169
ROA	0.36	0.7	3.56

Table 2 presents descriptive statistics for key financial variables in the analysis. The average sales-to-assets ratio of 0.73, with a moderate standard deviation (0.51), suggests that most companies efficiently use their assets, although some variability exists. The solvency ratio shows an average of 39.42, indicating varying financial health across firms, as reflected in its higher standard

deviation (19.17) and broad range. The liquidity ratio has a mean of 1.81, implying adequate short-term asset coverage, but its large standard deviation (2.94) indicates differing liquidity levels among companies. Profitability measures like NI to Sales and EBIT to Sales show wide variability, as suggested by their high standard deviations, pointing to the diverse financial performance in the sample.

This research employs a quantitative methodology to forecast sustainability for private enterprises, utilizing financial data that is publicly disclosed by EU corporations and accessible through the ORBIS database. Employing conventional ratios guarantees uniformity and comparability, but incorporating industry as a qualitative element addresses sector-specific sustainability concerns. This work utilizes financial data as a surrogate for sustainability to provide a new approach for estimating ESG scores for companies without publicly accessible ESG ratings, thereby enhancing the existing literature on financial performance and corporate sustainability (Friede, Busch, & Bassen, 2015).

3.1.2 Variables Selection

This research employs a comprehensive methodology by analyzing the overall financial statements of organizations through fundamental financial parameters indicative of profitability, liquidity, and solvency. These financial ratios offer a more comprehensive and detailed perspective on a company's activities than absolute financial data. Ratios such as Sales to Assets, EBIT to Sales, and Net Income to Sales provide essential insights into a company's efficiency in utilizing its resources to generate revenue and profits. By concentrating on these ratios, we seek to delineate organizations in a manner that truly represents their financial stability and operational efficacy, therefore elucidating their ESG ratings more precisely.

Financial ratios are essential indicators in fundamental analysis, providing stakeholders insight into a company's business operations and financial performance. For decades, both institutional and private investors have utilized these ratios to evaluate a company's worth and predict its future potential. Profitability ratios, namely EBIT to Sales and Net Income to Sales, assess a company's capacity to transform sales into profits, a critical factor in evaluating a firm's sustainability and long-term survival. The sales-to-assets ratio offers insight into a company's efficiency in leveraging its assets to produce income. These ratios combined provide an essential component of the financial

analysis in this study, providing a more nuanced understanding of ESG scores in relation to operational efficiency and profitability.

Alongside profitability ratios, our research encompasses two critical balance sheet metrics: the Solvency Ratio and the Liquidity Ratio. These ratios offer an alternative viewpoint by emphasizing a company's capacity to fulfill its long-term and short-term financial commitments. The liquidity ratio, commonly known as the current ratio, assesses a company's capacity to meet its current liabilities using its current assets. This is a vital metric of a company's short-term financial viability, as entities with greater liquidity are more adept at managing unexpected problems and sustaining operational stability. The solvency ratio, defined as the proportion of total debt to total assets, indicates the company's capacity to fulfill its long-term financial commitments. A high solvency ratio signifies greater leverage, thus heightening financial risk, whereas a low solvency ratio implies a more conservative debt management strategy. When analyzed alongside profitability measures, these balance sheet ratios provide a comprehensive insight into a firm's financial status and performance.

Fundamental ratios, utilized in this study, have been employed for decades in stock market analysis to interpret a company's inherent value. Penman (2012) asserts that fundamental data denotes the "underlying factors that influence a company's actual operations and its future potential." Fundamental analysis posits that a company's stock price may not consistently represent its intrinsic worth in the near term; but, it will eventually realign with the company's financial fundamentals over time. This hypothesis posits that firms exhibiting robust financial performance will experience an appreciation in their stock prices over time as investors acknowledge their underlying value. Damodaran (2012) emphasizes that fundamental data—such as profitability, liquidity, and solvency ratios—provides a more reliable foundation for assessing a company's long-term potential than market indicators, which may be swayed by investor sentiment and external influences.

In this study, We utilized financial ratios, such as the solvency ratio, over a three-year period to predict the ESG performance of companies. Instead of treating each fiscal year as an isolated observation, I included the values of the financial ratios for three consecutive years as inputs into the machine learning model. This approach allowed the model to consider temporal variations in the financial data, capturing trends and changes over time that may have an impact on the company's ESG performance. By using this method, I aimed to identify patterns in the financial health and stability of firms that correlate with their sustainability status.

In this thesis, we apply this concept analogously to ESG ratings. Although ESG evaluations often encompass both qualitative and quantitative elements, we propose that a company's financial fundamentals should exhibit a robust correlation with its ESG performance over time. A company exhibiting good financial performance, marked by elevated profitability, liquidity, and solvency, is likely to possess the resources and stability required to invest in sustainable practices and uphold solid governance frameworks. Conversely, financially distressed organizations may lack the resources to prioritize sustainability, thereby impacting their ESG rankings. Consequently, we contend that financial data, obtained through standardized accounting methods and readily accessible for analysis, provides a dependable and comparable foundation for forecasting the sustainability of a firm,

The preliminary dataset for this research comprised 3,560 public enterprises across Europe, encompassing a diverse array of industries. Subsequently, after executing an inner join to align organizations with accessible ESG ratings, the final dataset was diminished to 975 firms. The diminished sample size was further processed to standardize the data, rectify missing values, and guarantee consistency. Numerical variables with absent data were imputed using the median, while categorical variables were imputed using the mode. This pre-processing phase was essential for preserving the dataset's integrity and ensuring that the model could generate accurate predictions without being distorted by absent or inconsistent data.

For our predictive model, we used a Random Forest classifier, which is an algorithm that is good at working with large datasets that have both correlated and uncorrelated features. We converted the ESG ratings into a binary variable by translating the ESG letter grades into numerical values, and then established a binary classification based on these scores. We were able to predict whether a company would fall into a higher or lower ESG category using this binary target variable. To elucidate temporal linkages and enhance the model's prediction efficacy, we utilized financial data spanning three years, considering these data as lagged variables. This methodology allowed the model to incorporate long-term financial patterns while remaining unaffected by short-term swings that may introduce data noise.

Financial data is known to be inherently noisy, and problems like multicollinearity—where independent variables exhibit strong correlation—can impact the outcomes of predictive models.

The relationships between financial statistics and corporate performance may be non-linear and contingent upon context, so complicating the study. D'Amato et al. (2023) assert that financial ratios, however beneficial, require cautious interpretation due to multicollinearity and other statistical complexities. Therefore, random forest is the optimal model for training as it's robust to multicollinearity.

3.1.3 Private Firm Data

When comparing the challenges of managing public firm data to those of managing private firm data, the former proved to be very simple. We initially collected data from the ORBIS database for 6,000 private enterprises. After cleaning the dataset, we discovered that 1,310 enterprises were missing critical sales data, which is necessary for computing vital financial ratios, including sales to assets and EBIT to sales. Due to the significance of sales data in evaluating operational efficiency and profitability, these organizations were omitted from the research. The final dataset included 2,035 private enterprises, after omitting the subsidiaries of big corporations.

Private enterprises in the European Union must adhere to particular financial reporting standards that guarantee openness and uniformity. The EU Accounting Directive (2013/34/EU) mandates that medium and large private enterprises compile and disclose detailed yearly financial statements, comprising a balance sheet, income statement, and accompanying notes. Companies, particularly those that satisfy specific size criteria, must adhere to International Financial Reporting Standards (IFRS) or national Generally Accepted Accounting Principles (GAAP), contingent upon their jurisdiction. The order requires big private companies to provide essential financial indicators, including total revenue, operational profit (EBIT), net income, and assets, facilitating the consistent calculation of key financial ratios across organizations.

The Sales to Assets ratio is determined by dividing total sales by the average total assets for the fiscal period, indicating asset efficiency. The EBIT to Sales ratio is calculated by dividing operating profit (EBIT) by total sales, offering insights into a company's capacity to create profit from its primary operations. Net Income to Sales is determined by dividing net income by total sales, reflecting overall profitability. These ratios, commonly employed in financial analysis and valuation, are crucial for evaluating the financial stability and sustainability of private enterprises.

In the EU, private corporations, especially large ones, are required to submit their financial accounts annually to national authorities, with the information frequently accessible via databases such as ORBIS. This legal structure guarantees a significant level of data consistency and comparability, crucial for studies pertaining to financial performance analysis. Significant private corporations are mandated to undergo financial statement audits, hence enhancing data reliability. The existence of strong financial reporting regulations guarantees that the essential ratios utilized in this research, such as profitability, liquidity, and solvency indicators, are derived from dependable and standardized financial disclosures (Penman, 2012).

In contrast to the public firm dataset, which necessitated the incorporation of ESG ratings and subsequent model adjustments, the private firm data just needed compliance with established financial criteria. This differentiation streamlined the investigation, enabling us to concentrate exclusively on the predicting efficacy of financial ratios. For private enterprises, ESG scores are either inaccessible or insufficient, rendering financial fundamentals the principal data source for research. This created an opportunity to evaluate the machine learning model, originally trained on public companies with ESG data, on private companies with solely financial principles.

The financial data utilized in this analysis originates from rigorously controlled reporting standards standardized across EU member states via the EU Accounting Directive. The directive mandates that private enterprises adhere to stringent regulations for revenue recognition, asset valuation, and liability treatment, thereby guaranteeing the accurate and comparable calculation of financial ratios (European Commission, 2013). Furthermore, the audit obligation for substantial enterprises ensures that the financial statements accurately represent the company's financial status.

The EU's regulatory framework renders private enterprise data optimal for financial analysis. Essential ratios—such as those indicating profitability, solvency, and liquidity—are required disclosures for big corporations and provide a dependable basis for analysis. This research benefits from the availability of high-quality financial data for private firms and strong regulatory control, ensuring that the fundamental ratios employed are both reliable and comparable across various companies and industries.

Initially, our dataset comprised 6,000 private enterprises; however, after filtering for missing data, and omitting subsidiaries, the final sample was diminished to 2,035 firms. These companies satisfied the financial criteria necessary for the analysis, and their financial statements complied with EU reporting regulations. The uniformity and adherence to regulations facilitated the application of the machine learning model to forecast outcomes based on financial fundamentals. The lack of ESG ratings in private companies may be perceived as a drawback; nevertheless, the presence of dependable financial data alleviated this concern and facilitated a concentrated assessment of financial performance as an indicator of sustainability.

The fundamental data for private enterprises in the EU is governed by the EU Accounting Directive (2013/34/EU) and IFRS, establishing a uniform framework for the computation of critical financial ratios. This legislative structure guarantees that the financial data of private enterprises is transparent, dependable, and comparable, rendering it appropriate for predictive modeling. Utilizing this high-quality dataset, we transferred the machine learning model developed for public businesses to the private firm dataset, evaluating the efficacy of financial fundamentals in predicting ESG-related outcomes.

Therefore, managing the private firm data was more straightforward owing to the uniform financial reporting standards in the EU and the accessibility of dependable data. Utilizing a final sample of 2.035 enterprises, we concentrated on the predictive capacity of financial fundamentals by employing essential profitability, liquidity, and solvency ratios to evaluate the financial health of the organizations. The uniformity in reporting standards guaranteed the dataset's appropriateness for the application of machine learning models, albeit the lack of ESG scores. This analysis illustrates the capacity of financial data to serve as a proxy for comprehending corporate sustainability in private enterprises.

The data reveals that our dataset encompasses companies from various European countries, spanning 28 distinct industries. The distribution of companies across these industries highlights the diverse economic sectors represented in the analysis.

3.2 Research Design

Several studies have endeavored to forecast Environmental, Social, and Governance (ESG) ratings with diverse machine learning methodologies, primarily concentrating on publicly traded corporations. The paper "Fundamental ratios as predictors of ESG scores: a machine learning approach" investigates how financial statement items affect ESG scores for companies listed in the STOXX Europe 600 Index. The authors use the Random Forest algorithm to analyze balance sheet and income statement items to explain the ESG Bloomberg ratings. The study finds that certain financial ratios, such as the Net Income to Sales ratio and Solvency Ratio, can significantly predict a company's ESG score. Meanwhile, Krappel (2019) employed fundamental financial data to forecast ESG scores; however, his methodology was restricted to public enterprises and may not be entirely applicable to private, unlisted, or limited liability entities, where such data may be inaccessible or less transparent. This poses a barrier for the application of ESG prediction models to private corporations, as these organizations frequently do not provide the same degree of information mandated for publicly traded firms. Gracia (2020) conducted a study aimed at forecasting the ESG performance of public companies across four specific sectors utilizing market financial data. This study offered significant insights into sector-specific ESG trends but primarily focused on public enterprises, without a defined methodology for private firms. Consequently, whereas prior studies have examined diverse models for predicting ESG scores of public companies, a significant vacuum persists in the literature about the applicability of these models to private companies. These studies, however instructive, do not provide a definitive strategy for tackling the issues of inadequate data transparency and financial disclosure in the private sector, where ESG monitoring is becoming progressively essential.

3.2.1 Data Preprocessing and Feature Engineering

This section delineates the precise processes undertaken to preprocess the data and execute feature engineering in preparation for machine learning analysis. The processes encompass addressing missing data, encoding categorical variables, and developing a sustainability variable essential for the classification task of forecasting ESG ratings.

Development of a Binary Sustainability Variable:

A crucial aspect of this investigation involved transforming the ESG letter grades into a format compatible with machine learning techniques. The dataset previously contained ESG ratings expressed as letter grades from A+ to D-. To streamline the analysis, I initially converted these letter grades to an ordinal scale, with values spanning from 1 (equivalent to D-) to 12 (equivalent to A+). This modification yielded a numerical representation of the ESG ratings, which is more suitable for computational models.

Table 3: ESG grade letter conversion

ESG Rating	Assigned Number
A+	1
A	2
A-	3
B+	4
B	5
B-	6
C+	7
C	8
C-	9
D+	10
D	11
D-	12

After converting the ESG scores to an ordinal scale, I computed the median ESG score for the dataset, which is 6. Companies with ESG scores over the median were categorized as "sustainable," while those at or below the median were designated as "non-sustainable." A binary variable named Sustainable was established, with "1" signifying a sustainable enterprise and "0" denoting a

non-sustainable firm. The binary variable was generated using Python, utilizing the Pandas and Numpy packages for data handling and transformation.

This stage guarantees that the machine learning model can accurately distinguish between companies based on their sustainability profiles, consistent with the objective of categorizing them as sustainable or unsustainable.

Missing data in continuous variables might bring bias into the model and diminish the accuracy of predictions. I utilized median imputation for all continuous variables to resolve this issue. The median is a resilient measure of central tendency that mitigates the influence of outliers, which is especially significant due to the potential skewness of financial variables in the dataset. This method guarantees that the imputed values accurately reflect the dataset's distribution while maintaining its structure.

The dataset's categorical variables, including industry, also exhibited missing values. To address them, we employed mode imputation, which populates missing values with the most prevalent category in the dataset. Utilizing the mode for imputation guarantees adequate representation of the predominant category, and this technique is proficient in preserving the integrity of categorical features. By imputing absent values with the mode, I ensured data preservation and the usability of categorical variables for subsequent analysis.

The industry variable, denoting the sector of each firm, is a categorical feature that is not immediately applicable in most machine learning techniques. We implemented one-hot encoding on the industry variable to address this issue. This procedure converted each industrial category into an individual binary variable. For instance, if the dataset comprises companies from sectors like manufacturing, technology, and finance, each sector would be denoted by a binary column (e.g., Industry_Manufacturing, Industry_Technology, Industry_Finance), where a "1" signifies the firm's affiliation with that sector, and a "0" denotes its absence from that sector.

Through the implementation of one-hot encoding, we guaranteed that the industry variable was accurately recorded and could be integrated into the machine learning models. This method inhibits the model from perceiving the categories as ordinal, so permitting each industry to be regarded independently.

3.2.2 Choosing the Model: Does It Make Sense?

The choice of using the Random Forest algorithm for the thesis is based on several important factors, both academically and practically, that align with the goals of the research on predicting ESG scores for private firms. Here are the reasons why we believe random forests are the most suitable approach for the study:

This thesis involves a complex set of financial and ESG-related variables, each potentially contributing to the prediction of corporate sustainability. Random Forests are particularly strong at handling high-dimensional datasets because they use multiple decision trees, each trained on random subsets of data and features. This ensures that the model can process and rank many variables effectively without overfitting (Breiman, 2001). Given that ESG score prediction involves many interconnected variables, Random Forests allow me to analyze them simultaneously and extract meaningful patterns (García et al., 2020).

One of the challenges in working with financial data is the risk of overfitting, where models perform well on training data but fail to generalize to new, unseen data. Random Forests use bootstrapping (random sampling with replacement) to reduce overfitting. By training each tree on different subsets of data, the model is more robust and less likely to memorize specific patterns that don't generalize well (Hull, 2020). Given the variability in the private firms I'm studying, this robustness ensures that the model remains stable across different samples, which is crucial for producing reliable results. A significant part of the thesis is to understand which financial and ESG-related variables are most important for predicting sustainability. Random forests offer a built-in mechanism for calculating feature importance, which ranks variables based on how much they reduce impurity (Breiman, 2001). This feature makes Random Forests particularly useful for the research because it provides clear insights into which financial metrics contribute the most to ESG scores, making the results both interpretable and actionable (Hull, 2020).

The relationship between financial performance and ESG scores is not always straightforward or linear. In fact, these relationships are often non-linear and involve complex interactions between multiple variables. Random Forests excel at capturing these non-linear patterns because each tree models different aspects of the data (D'Amato et al., 2023). This capability is essential for this thesis, where we expect the relationships between financial indicators and sustainability outcomes to be influenced by various interacting factors.

Since this thesis involves both public and private firms, we need a model that can generalize across different types of companies. Random Forests are known for their ability to handle heterogeneous

datasets because they aggregate predictions from multiple decision trees, each trained on different subsets of data. This generalization ability is particularly important for the research, as I am using public firm data to train the model and private firm data for testing. I'm confident that the model will adapt well to both datasets (Hull, 2020).

Finally, the use of random forests is widely accepted in both machine learning and finance for predictive tasks. Recent research has demonstrated that Random Forest models outperform traditional regression methods in predicting ESG scores due to their ability to capture complex, non-linear relationships (D'Amato et al., 2023). By incorporating this method into the thesis, I am aligning the work with current best practices in the field and contributing to the growing body of literature on ESG prediction and machine learning in finance (García et al., 2020).

In conclusion, Random Forests are the best choice for the thesis because they handle high-dimensional data, offer robustness against overfitting, provide insights into feature importance, and capture non-linear relationships—all of which are critical for accurately predicting ESG scores. Their ability to generalize across diverse datasets, as well as their wide acceptance in the academic and professional communities, further support their use in the research. I believe this approach will allow me to generate meaningful and reliable results, contributing to the field of ESG prediction and corporate sustainability.

3.3 The Machine Learning Model

3.3.1 Introduction to Random Forest Algorithm: Properties and Strengths

Random Forests (RF) represent a powerful ensemble learning method, particularly useful for both classification and regression tasks. The core concept behind random forests lies in aggregating predictions from multiple decision trees to enhance predictive performance. This method relies on the wisdom of crowds principle, where a large number of weakly correlated models (in this case, decision trees) work together to produce a more accurate and robust prediction than any single model could achieve alone. Hull (2020) emphasizes the strength of the Random Forest. algorithm is its ability to mitigate overfitting and improve prediction accuracy by leveraging ensemble learning techniques.

The Random Forest Algorithm

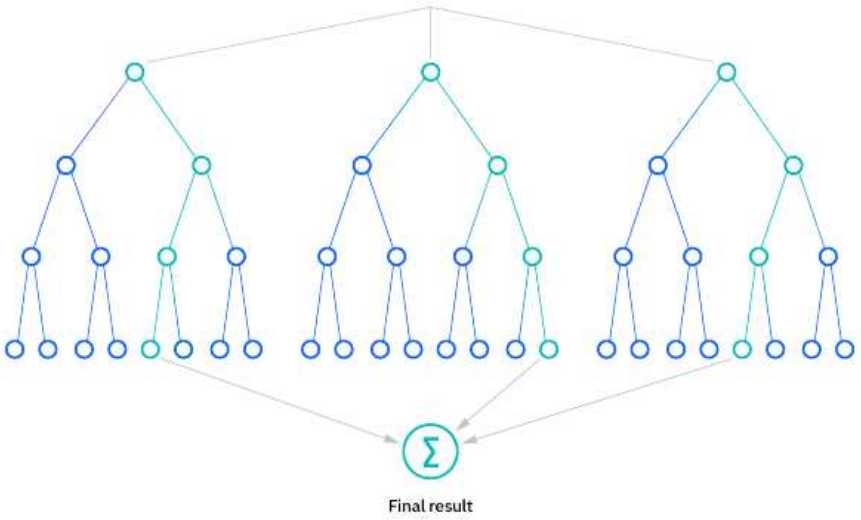

Fig. 5: Overview of Random Forest Classifier

Each decision tree in the forest is trained on a random subset of the training data, sampled with replacement (i.e., bootstrapping). Additionally, at each node of the tree, a random subset of the features is selected for splitting. This randomness is the foundation of the decorrelation between the trees in the forest, ensuring that individual trees do not consistently make the same errors. By combining these diverse, weak learners, the ensemble as a whole is able to outperform any single tree, effectively reducing both variance and bias (Hull, 2020).

Error reduction in random forests:

The efficacy of the Random Forest algorithm depends on two key factors:

1. Strength of individual trees: Each tree should perform better than random guessing, although it need not be perfect. This means that even weakly predictive trees can contribute positively to the overall accuracy of the forest.
2. Diversity of trees: The trees in the forest should be as uncorrelated as possible. The algorithm ensures this by introducing randomness in both data sampling and feature selection. When trees make uncorrelated errors, these errors are likely to cancel each other out, thereby reducing the overall model error (Breiman, 2001).

The bootstrap aggregation (or bagging) method used to train the individual trees introduces variability in the dataset, ensuring that each tree is exposed to a different subset of data. Simultaneously, the random selection of features at each split reduces the likelihood that trees will

consistently make the same decisions based on dominant features. This randomness is crucial for creating an ensemble that balances bias and variance, ensuring that the model generalizes well to unseen data (D'Amato et al., 2023).

3.3.2 Implementing Random Forest Algorithm

Public companies:

In the first phase of our model implementation, we focused on encoding independent categorical variables, particularly the "Industry" variable. Since machine learning models, especially tree-based models like Random Forest, require numerical inputs, it is essential to convert categorical data into a format that the algorithm can process effectively. To address this, we applied one-hot encoding, a well-established method for transforming categorical variables.

One-hot encoding converts each category within a categorical variable into a new binary variable. For example, if the "Industry" variable has 28 distinct categories, each category is transformed into a new binary column, where a value of 1 indicates the presence of that category for a given observation, while all other columns for that observation are set to 0. This approach ensures that the model does not infer any ordinal relationship between categories that do not naturally have one. As industries typically have no inherent hierarchy, this method prevents the model from mistakenly assigning a rank or order to the industries, which could skew the predictions (Friedman et al., 2001).

In the second phase of our implementation, we turned our attention to identifying the features and target variables for the model. We included a total of 45 features, which consisted of lagged independent variables for each year in our dataset. These features provided a comprehensive view of the companies' financial and operational data across multiple years, ensuring that the model had enough context to make accurate predictions.

The target variable in our study was labeled as "Sustainable." This binary classification variable indicated whether a company was considered sustainable based on its Environmental, Social, and Governance (ESG) grade obtained from Refinitiv LSEG. To avoid any data leakage, we excluded certain columns from the feature set, such as Company_Name, ESG_Grade_Numeric, and Sustainable. This ensured that the only variable being predicted was the sustainability status of the company, which was based on an ESG score transformation (Louppe, 2014).

In the third phase, we addressed the issue of missing values, which are common in large datasets. Missing data can introduce bias into machine learning models, For numerical data, we opted for median imputation. The median is a robust measure of central tendency that is less affected by outliers compared to the mean. By using the median, we ensured that extreme values did not skew the imputation process (Schafer, 1997). For categorical data, we used the mode (the most frequent value) as the imputation method. This choice ensured that the general distribution of the categorical variables was preserved, minimizing the impact of missing data on the overall structure of the dataset. The combined approach of median and mode imputation allowed us to maintain the integrity of the dataset while effectively dealing with missing values.

In the fourth phase, we prepared the data for model training by splitting it into training and testing subsets. Following machine learning best practices, we allocated 80% of the data to the training set and 20% to the testing set. To ensure that our results were reproducible, we set the random state to 42. Additionally, we employed stratified sampling based on the industry variable to maintain proportional representation of each industry in both the training and testing sets. This approach ensured that the model was exposed to a representative distribution of industries during training, preventing it from disproportionately favoring certain industries over others (Kohavi, 1995).

With the data preprocessed and split, we proceeded to train our Random Forest model. Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to make a final decision. It is particularly well-suited for datasets with a large number of features and is known for its robustness against both overfitting and noise (Breiman, 2001). After training the model on the public companies dataset, we evaluated its performance on the testing set. The model achieved an overall accuracy of 75%, meaning that it correctly classified sustainable and unsustainable companies 75% of the time. However, given the imbalanced nature of the dataset, where there were significantly more sustainable companies than unsustainable ones, accuracy alone was not a sufficient metric to evaluate the model's performance. To gain a deeper understanding of the model's effectiveness, we examined additional metrics such as precision, recall, and the F1-score.

3.3.3 Exploring threshold and cross-validation techniques

After training the Random Forest model, we initially evaluated its performance using the default classification threshold of 0.5. At this threshold, the model achieved a satisfactory accuracy of 75%. However, when we examined the precision and recall for sustainable (Class 1) and specificity of

unsustainable companies (Class 0), we noticed a significant imbalance. While the model performed well in identifying sustainable companies (with high recall and precision), its performance for unsustainable companies was considerably lower.

In a Random Forest classification model, the default threshold of 0.5 is commonly used to determine the classification decision. For probabilities greater than or equal to 0.5, the model predicts Class 1 (sustainable). For probabilities lower than 0.5, the model predicts Class 0 (unsustainable). However, when dealing with imbalanced datasets, where one class is more prevalent than the other, the default threshold may not provide optimal results for both classes (Louppe, 2014).

Recognizing this, we decided to experiment with different thresholds to better balance precision and recall for both classes. Adjusting the threshold allows us to control the trade-off between these two metrics. Lowering the threshold tends to increase recall but can reduce precision, while raising the threshold generally improves precision at the expense of recall (He & Ma, 2013).

Experimenting with threshold adjustments:

At the default threshold of 0.5, the model had a high recall for sustainable companies (87%) but struggled with unsustainable companies, which was only 42%. The specificity for unsustainable companies was even lower at 32%, indicating that the model was failing to identify a significant portion of unsustainable firms. This imbalance suggested that the model favored identifying sustainable companies but at the cost of neglecting unsustainable ones.

To address the model's underperformance for unsustainable companies, we first lowered the threshold to 0.4. This change slightly increased the specificity for unsustainable companies, allowing the model to recognize more of them. However, it also led to a decrease in precision for sustainable companies, as the model began predicting more companies as sustainable, including some false positives. Although the recall for unsustainable companies improved slightly, the overall balance between precision and recall for sustainable companies worsened, making this threshold less favorable (Friedman et al., 2001).

Next, we decided to raise the threshold to 0.6. This time, the results were much more promising. By increasing the threshold, the model became more cautious in predicting companies as sustainable, leading to an improvement in precision. At this threshold, the model's precision for sustainable companies rose to 81.6%, meaning that when the model predicted a company as sustainable, it was correct over 81% of the time. Interestingly, the recall for sustainable companies remained relatively high at 76.2%, indicating that the model still successfully identified a large portion of sustainable

firms, even with the more stringent threshold. This change struck a better balance between precision and recall, leading to a more accurate overall classification (Breiman, 2001).

While the recall for unsustainable companies remained lower than desired, the increase in precision for sustainable companies helped improve the model's overall reliability in real-world applications. We found that this threshold adjustment made the model less prone to false positives, making it a better fit for scenarios where correctly identifying sustainable companies was a higher priority.

We also experimented with thresholds higher than 0.6, such as 0.7 and 0.8. However, while precision for sustainable companies continued to increase at these thresholds, the balanced accuracy dropped significantly. The model became too conservative, often failing to identify sustainable companies and focusing primarily on avoiding false positives. This resulted in an overly restrictive classification system that was impractical for our use case, where identifying most sustainable companies was crucial (He & Ma, 2013). After evaluating the performance metrics at various thresholds, we settled on 0.6 as the optimal threshold for our model. This threshold offered the best trade-off between precision and recall, particularly for sustainable companies. It allowed us to maintain high precision while still achieving a respectable recall. Moreover, it minimized false positives for unsustainable companies, addressing some of the imbalance issues we had encountered with the default threshold (Schafer, 1997). Adjusting the classification threshold was a pivotal step in improving the model's performance. Through experimentation, we found that a threshold of 0.6 provided the best balance, enhancing the model's precision for sustainable companies while still capturing a significant portion of them resulting in a higher F-1 score and balanced accuracy. This adjustment made the Random Forest model a more reliable tool for predicting the sustainability status of companies, especially in an imbalanced dataset scenario.

As The cross-validation results it reveals a robust and consistent performance of the binary classification model across various data splits. The model excels in accurately identifying the majority of sustainable companies, striking a delicate balance between capturing true positives and minimizing false alarms. This balance, reflected in its overall performance metrics, underscores the model's reliability in forecasting corporate sustainability. Moreover, the uniformity of these results across all validation folds suggests a strong generalization capability, reinforcing the model's effectiveness as a predictive tool in this context.

In table 4, the hyperparameters used for the random forest classification have been analyzed, It is worth noting that only N-estimators and random states were obtained from previous literature and industry best practices, while the rest of the hyperparameters were a default in the Random

Hyperparameter	Value	Explanation	Reason for Use
n_estimators	100	The number of trees in the forest.	A larger number of trees generally leads to better performance but increases computation time. 100 is a common choice for a balance between accuracy and efficiency.
criterion	gini	The function used to measure the quality of a split. "Gini" calculates the impurity of the node.	Gini is used because it is computationally faster than other criteria (like entropy) and generally performs well in classification problems.
max_depth	None	The maximum depth of the tree. None means nodes are expanded until all leaves are pure or contain fewer than the minimum samples required to split.	Keeping it None allows the trees to grow fully, capturing more complex patterns in the data, which is useful when feature interactions are complex.
min_samples_split	2	The minimum number of samples required to split an internal node.	Using the smallest possible split value allows the model to consider more splits, enhancing its ability to learn complex patterns.
min_samples_leaf	1	The minimum number of samples required to be at a leaf node.	Setting this to 1 ensures that the model can capture small nuances in the data, avoiding overly generalized splits.
max_features	auto	The number of features to consider when looking for the best split. auto typically uses the square root of the number of features.	Using a subset of features helps in reducing overfitting and improves generalization by introducing diversity among the trees.
bootstrap	TRUE	Whether bootstrap samples are used when building trees.	Bootstrapping introduces randomness, helping to reduce overfitting by making each tree see a different subset of the data.
random_state	42	Controls the randomness of the model, ensuring reproducibility.	Setting a specific value ensures that the model produces the same results every time it runs, useful for debugging and comparison purposes.

Table 4: Overview of ML model parameters

Forest Classification. A random state hyperparameter of 42 was set to ensure that the model will obtain the same results each time it runs, resulting in consistency in the outputs of the model.

Second Part: Predicting Private Companies

The analysis of private firms' sustainability presents unique challenges, primarily due to the lack of direct ESG reporting, unlike their publicly traded counterparts. In this section, we delve deeper into the model's application to private firms, focusing on the predictive power of the Random Forest classifier in this context and the implications for understanding corporate sustainability across non-disclosing firms.

Private firms typically do not disclose ESG scores as they are not bound by the same reporting requirements as publicly listed companies. This creates a significant gap in data availability for assessing their sustainability practices. However, stakeholders, including investors, regulators, and consumers, increasingly demand transparency and accountability from all firms, not just public ones. Therefore, the ability to predict ESG performance for private firms based on financial metrics is a valuable tool for bridging this gap. For this analysis, we utilized financial variables as proxies for sustainability performance. The idea is rooted in the growing body of literature that suggests financial performance and sustainability are linked, especially in firms with long-term strategic outlooks. The model aimed to leverage these financial indicators to estimate sustainability for private companies.

Data Preprocessing and Feature Engineering for Private Firms

Before applying the Random Forest model to private firms, several key preprocessing steps were performed to ensure that the data was comparable to the public dataset and that the model could generate reliable predictions:

- Removal of Subsidiaries: Many private firms in the dataset were subsidiaries of publicly listed companies. Since subsidiaries often rely on their parent companies for sustainability reporting and do not have independent sustainability initiatives, they were removed from the dataset. This step ensured that the remaining firms were independent entities, thus making their financial data more reflective of their sustainability performance.
- Industry One-Hot Encoding: The industry of each firm was treated as a categorical variable. By one-hot encoding this variable, we transformed it into a format suitable for machine learning models. This allowed the model to consider the industry's impact on sustainability predictions, which is crucial since sustainability practices can vary significantly across industries (e.g., energy firms face different sustainability challenges than tech firms).
- Calculation of Financial Ratios: As stated in the public data section ROA, Solvency ratio and liquidity ratio were directly obtained from ORBIS while Net income to sales, Sales to assets, and Ebit to sales were calculated manually.

- Imputation of Missing Data: Missing numerical data were imputed using the median for each variable. This method was chosen to preserve the distribution of the data and avoid potential biases that could arise from removing companies with missing data or imputing with the mean, which could be skewed by outliers.

Manual Validation of the ML Model Prediction for Private Companies

After training the Random Forest classification model on public companies, the model was applied to predict the sustainability of private firms. Given the lack of direct ESG disclosures for private companies, we developed a comprehensive validation process using data from the LSEG Refinitiv Sustainability Dashboard, individual company websites, reports, and relevant controversies. This allowed us to effectively validate the machine learning model's predictions, ensuring accurate classification of private companies into sustainable and non-sustainable categories.

The LSEG Refinitiv Sustainability Dashboard provides a comprehensive approach to ranking public firms based on their ESG performance. It assigns scores across the Environmental, Social, and Governance pillars and ranks companies using percentile ranking. However, given the small sample size, we will follow a different approach. We will divide the companies using quartiles to group the 20 companies into four groups based on their scores.

- 1st Quartile (Q1): Bottom 25% (least sustainable).
- 2nd Quartile (Q2): Next 25%.
- 3rd Quartile (Q3): Next 25%.
- 4th Quartile (Q4): Top 25% (most sustainable).

We first gathered available information from company websites, reports, and the LSEG Refinitiv Dashboard (if applicable). Next, we used this data to construct an ESG score for each company, ranking them into quartiles based on this score. To simplify the process, a binary system was adopted where if a company has reports, news, or certifications in one of the ESG metrics, we will assign the value of 1 while if it doesn't have any proof of their work in this area, then they will be assigned a zero. Then we randomly selected 20 firms from the predicted private firms dataset, and to avoid a class imbalance, the selected firms constituted 10 companies that the model predicted to be sustainable and 10 that were not.

Based on the adopted binary system, firms can obtain a score ranging from 0 to 9 based on their performance in the 9 metrics we chose from Refintive LSEG metrics, evaluating three primary

pillars: environmental, social, and governance. For each pillar, specific metrics were used to quantify the company's performance.

For private firms, the data was sourced primarily from their websites, financial disclosures, and where available, sustainability reporting. This helped fill the gap left by the lack of formalized ESG reports that public companies typically publish. It is worth noting that In cases where companies were involved in significant controversies (e.g., environmental violations, governance scandals), the ESG score was adjusted downward. This provided a more balanced and accurate reflection of each firm's true sustainability profile, accounting for both positive contributions and negative impacts.

1. Environmental Pillar Metrics

The environmental pillar focuses on how well a company manages its environmental impact. This included evaluating:

- Emissions: Includes metrics like total emissions, reduction initiatives, and environmental policies.
- Innovation: Focuses on product innovation, environmental R&D, and sustainable products.
- Resource Use: Measures include water use, energy use, and waste reduction policies.

Each firm was assigned an environmental score based on these metrics. Higher scores were given to companies actively reducing emissions, using renewable energy, and minimizing waste. For example, firms that set ambitious targets for reducing their carbon footprint received a higher overall environmental score. Companies that lacked transparency on their environmental practices or were involved in environmental controversies were assigned lower scores.

2. Social Pillar Metrics

The Social pillar assesses a company's impact on society, including its relationships with employees, customers, and the communities it operates in. The key metrics evaluated in this pillar were:

- Labor practices and employee welfare: fair wages, diversity, inclusion policies, workplace safety, and employee development programs.
- Community Engagement: Philanthropy, volunteering, and investments in local communities.
- Diversity and inclusion: representation of minorities and women in leadership roles, as well as corporate policies promoting equality.

A company's social score reflected its commitment to fair labor practices, diversity, and community engagement. For instance, firms with strong employee benefits and diversity programs received

higher social scores. In contrast, companies with a history of poor labor conditions or social controversies were given zero in this category.

3. Governance Pillar Metrics

The governance pillar evaluates how a company is managed and the extent to which its leadership adheres to ethical practices. Metrics included:

- Board composition and independence: The diversity and independence of the board of directors.
- Transparency and reporting: The clarity and availability of financial and sustainability reporting.
- Anti-corruption policies: measures to prevent corruption, bribery, and unethical practices.

Companies with transparent reporting practices, diverse boards, and strong anti-corruption measures were given high governance scores. Conversely, firms with opaque reporting structures or governance scandals received lower governance scores.

Using these metrics will provide an approximation of the relevance of the Random Forest model to private companies. As in the results analysis section of this thesis, we will further analyze the results of the manual validation of private firms.

Chapter Four: Results, Discussion, and Conclusion

4.1 Key insights and results

Evaluation Metrics in Random Forest Classifier

Accuracy, recall, specificity, precision, and F1-score each offer unique insights into the model's performance, particularly when dealing with the prediction of sustainability in firms. Accuracy is a simple metric representing the proportion of correctly classified instances out of the total, but it can be misleading in imbalanced datasets (Hossin & Sulaiman, 2015). For instance, if non-sustainable companies far outnumber sustainable ones, a high accuracy might merely reflect the model's success in identifying the majority class rather than its ability to predict the minority class accurately. Recall (sensitivity), on the other hand, measures how well the model identifies actual positives, or sustainable companies, which is crucial in sustainability contexts to minimize false negatives. A high recall indicates that most sustainable companies are correctly identified. Specificity focuses on the model's ability to correctly classify negative instances, i.e., non-sustainable companies. It becomes critical when avoiding false positives is a priority since incorrectly classifying non-sustainable firms as sustainable can be misleading (Hossin & Sulaiman, 2015). Precision is equally important as it assesses the proportion of correctly predicted positives among all instances classified as positive, ensuring that firms labeled as sustainable are indeed sustainable, minimizing the cost of false positives. Finally, the F1-score serves as the harmonic mean of precision and recall, striking a balance between the two. It is particularly beneficial for imbalanced data scenarios, offering a single metric that reflects both the model's ability to find all positive cases and its accuracy in those predictions (Hossin & Sulaiman, 2015). Together, these metrics provide a comprehensive evaluation of the model’s performance in predicting sustainability, highlighting areas where it excels and where improvements are needed.

Table 5 : Random Forest Evaluation Metrics Summary Table

Metric	Formula	Purpose	Strengths	Weaknesses
Accuracy	TP+TN+FP+FNTP+TN	Measures the overall correctness of predictions.	Simple to understand and compute.	Misleading for imbalanced datasets.
Recall (Sensitivity)	TP+FNTP	Measures the ability to correctly identify positive cases.	Useful for minimizing false negatives.	May overlook false positives.
Specificity	TN+FPTN	Measures the ability to correctly identify negative cases.	Useful for avoiding false positives.	Can be less informative alone.
Precision	TP+FPTP	Measures the accuracy of positive predictions.	Useful when false positives are costly.	May overlook false negatives.
F1-Score	2×Precision+RecallPrecision×Recall	Balances precision and recall into a single metric.	Good for imbalanced data.	Less intuitive than other metrics.

4.1.1 Public Firm's Results

After training the model on the public companies dataset, we evaluated its performance on the testing set. The model achieved an overall accuracy of 75%, meaning that it correctly classified sustainable and unsustainable companies 75% of the time. However, given the imbalanced nature of the dataset, where there were significantly more sustainable companies than unsustainable ones, more than accuracy alone was needed to evaluate the model's performance. To gain a deeper understanding of the model's effectiveness, we examined additional metrics such as precision, recall, and the F1-score.

An Analysis of a Public Firm's Confusion Matrix

The confusion matrix reveals key insights into the predictive performance of the model. The high count of true positives (132) indicates that the model is proficient in identifying firms that align with sustainability criteria. This strong performance aligns with previous research suggesting that financial indicators, such as profitability and leverage, often correlate with higher ESG scores (García et al., 2022). The model's high recall (87.42%) suggests its robustness in detecting firms that genuinely implement sustainable practices, which is essential in ESG analysis for investors and stakeholders who prioritize sustainability in decision-making.

However, the matrix also exposes certain limitations. With only 14 true negatives, the model struggles to correctly identify non-sustainable firms, resulting in a low specificity of 31.82%. This issue is often observed in ESG prediction models due to the complex nature of sustainability, which is not always fully captured by financial metrics alone (Montiel & Delgado-Ceballos, 2014). The presence of 30 false positives further supports this observation, suggesting that while financial data can signal some aspects of sustainability, it may also lead to overestimations. This over-classification indicates the potential for a model bias toward labeling firms as sustainable, possibly due to financial characteristics commonly associated with ESG, such as high asset turnover or profitability, that may not comprehensively reflect the firm's true ESG performance (Moody's Analytics, 2021).

To improve the model's accuracy and balance between identifying both sustainable and non-sustainable firms, incorporating qualitative ESG metrics such as corporate governance practices, stakeholder engagement, or controversy data might be beneficial. This aligns with the findings of previous literature advocating for a multi-dimensional approach to ESG evaluation (UNPRI, accessed April 2024). Therefore, while the model shows significant promise in identifying sustainable firms, these results also highlight the importance of integrating broader ESG factors to capture a more nuanced and accurate prediction.

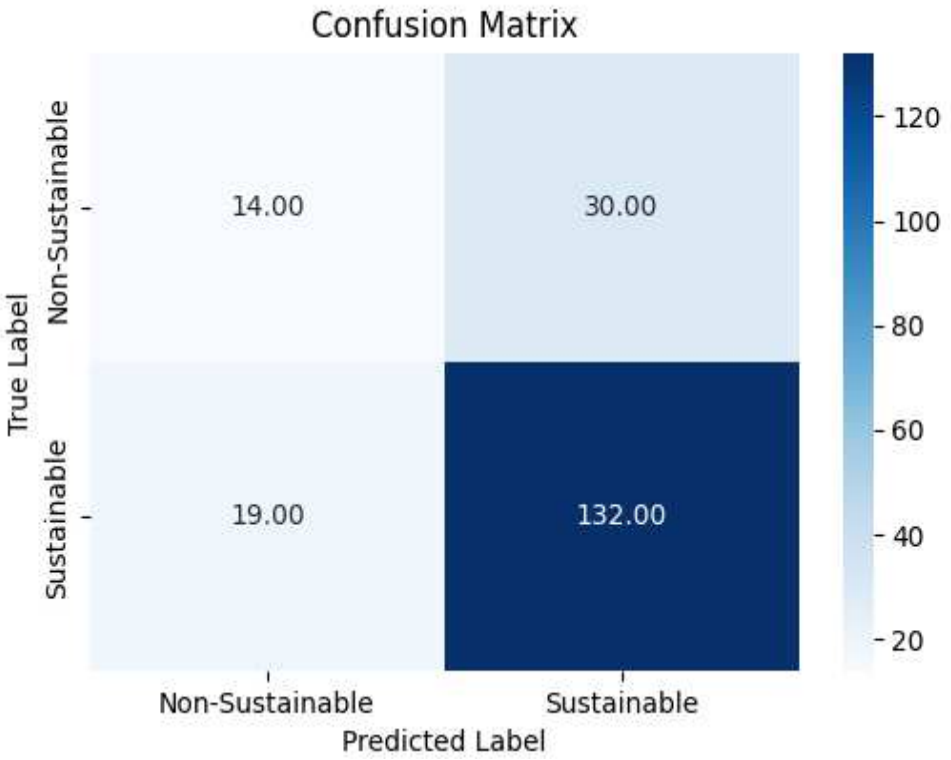

Fig. 7 : Confusion matrix "Top left: True Negatives; Top right: False Positives Bottom left: false negatives. Bottom right: True positives."

Analysis of Evaluation metrics

The model's performance in predicting the sustainability of public firms based on their fundamental data is quite promising, with several key strengths. An overall accuracy of 74.87% indicates a solid success rate in distinguishing between sustainable and non-sustainable firms. Notably, the F1 score of 84.35% shows a strong balance between precision and recall, suggesting that the model is effective in capturing firms that meet sustainability criteria. Additionally, the precision rate of 81.48% for the sustainable class implies a high level of confidence in the model's predictions; when the model identifies a firm as sustainable, it is likely accurate. The recall rate of 87.42% further reinforces the model's strength in identifying the most truly sustainable firms, supporting the use of fundamental financial data as meaningful predictors in ESG analysis. However, some weaknesses need to be addressed. The lower balanced accuracy of 59.62% hints at challenges in handling the imbalanced nature of the dataset, indicating that the model finds it difficult to differentiate between the two classes equally well. The specificity of just 31.82% points to a tendency to overestimate sustainability, revealing a struggle in correctly identifying non-sustainable firms. This limitation is

important to discuss, as it suggests that while financial data can be insightful, it may not fully capture the nuanced aspects of ESG criteria.

Table 6: Public Firm's Evaluation Metrics Analysis

Metric	Value	Significance
Accuracy	74.87%	Indicates the overall percentage of correctly classified instances, combining both sustainable and non-sustainable firms.
Balanced Accuracy	59.62%	Accounts for imbalanced classes, averaging the recall obtained for each class.
F1 Score	84.35%	Harmonic mean of precision and recall, providing a balance between the two. Useful for imbalanced datasets.
Precision (Class 1)	81.48%	Measures the proportion of true positives among all predicted positives (reliability of identifying sustainable firms).
Recall (Class 1)	87.42%	Indicates the model's ability to correctly identify actual sustainable firms.
Specificity (Class 0)	31.82%	Shows the model's ability to correctly identify non-sustainable firms, highlighting potential challenges in classifying them.

Analysis of Predictive Features of Corporate Sustainability

The analysis reveals that solvency ratios and liquidity ratios are the most important predictors of corporate sustainability. These financial ratios provide insight into a firm's ability to manage both long-term and short-term financial obligations, which significantly influences their sustainability performance.

1. Solvency Ratios: Solvency ratios measure a company's capacity to meet its long-term financial commitments, with higher ratios indicating stronger financial health. The model highlights that solvency is a crucial factor for sustainability. Companies with higher solvency ratios are better positioned to invest in sustainable practices because they have more financial flexibility and lower risks of financial distress (Bansal & DesJardine, 2014). This aligns with previous research that identifies financial stability as a critical enabler of corporate sustainability efforts (Elkington, 1997).
2. Liquidity Ratios: liquidity ratios, which indicate a firm's ability to meet short-term obligations, also play a key role in determining sustainability. Firms with higher liquidity

ratios are likely to manage their cash flow more effectively, which allows for greater investment in long-term sustainability initiatives (Dixon-Fowler et al., 2013). Companies that maintain healthy liquidity levels are more resilient to market fluctuations and can allocate resources towards environmental and social governance (ESG) activities.

The results are consistent and align with the Triple Bottom Line (TBL) framework, which emphasizes the importance of financial performance alongside social and environmental performance (Elkington, 1997). Firms that exhibit strong financial health through their solvency and liquidity ratios are more likely to adopt sustainable business models that balance profitability with social and environmental goals.

4.1.2 Private firm's results

Once the data was prepared, the Random Forest model, trained on the public dataset, was used to predict the sustainability of private firms. The model was well-suited for this task due to its ability to handle large datasets, its robustness against overfitting, and its strength in capturing non-linear relationships between financial metrics and sustainability outcomes.

Key Results:

After running the model on the private dataset, approximately 59.7% of private firms were predicted to be sustainable, while 40.3% were classified as non-sustainable. This distribution is reflective of the model's inherent bias toward predicting sustainability due to the imbalanced nature of the public dataset, where sustainable firms were overrepresented. Although the model was trained on public firm data, the prediction for private firms aligned with expectations based on industry insights. Private firms that exhibited strong financial performance, particularly in terms of profitability and leverage ratios, were more likely to be classified as sustainable. This is consistent with research suggesting that firms with healthier financial profiles are better positioned to invest in long-term sustainability strategies.

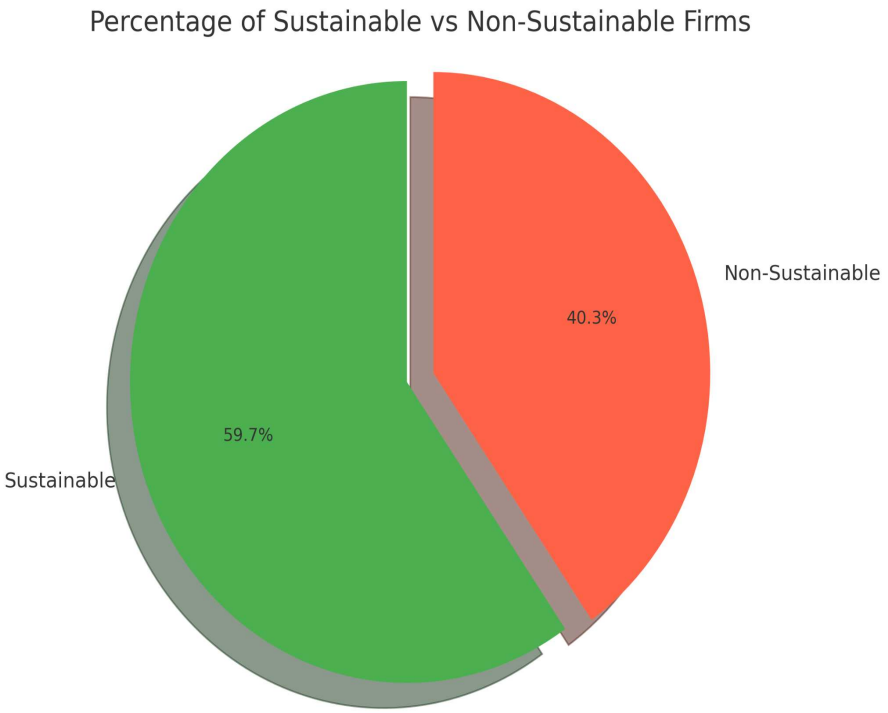

Fig. 8: A Pie chart of Private firm's results

The one-hot encoded industry variable allowed for a more granular analysis of sustainability predictions across sectors. For example: consumer goods companies were more frequently predicted as sustainable. This could be attributed to the low capital intensity and high profitability commonly associated with these industries, which affords them more resources to allocate toward sustainable practices. On the other hand, companies in the energy and manufacturing sectors showed a higher proportion of firms predicted as non-sustainable, likely due to the high capital requirements, regulatory pressures, and environmental challenges inherent to these industries. Such firms may struggle with adopting sustainability practices as quickly as their counterparts in less resource-intensive industries.

Real-world validation of Private Firm's Prediction

Using the composite ESG scores, the firms were ranked. Companies in the top quartiles generally had better ESG practices, as confirmed through data from the LSEG Refinitiv Sustainability Dashboard and company websites. specifically all companies in the 4th quartile had a strong sustainable practices, while the 3rd quartile included companies that were less transparent on their websites about their sustainability practices or did not have critical certificate such as ISO 14001 Certification for Environmental Management and ISO 14064 standards, which provides governments and businesses with an integrated set of tools for programs aimed at measuring,

quantifying and reducing greenhouse gas emissions. (*ISO 14064-1:2018*) but still have very promising plans or have other strong points in regards to their sustainability practices, such as diversity and inclusion.

Furthermore, the majority of the companies in the 1st and 2nd quartiles had no available data on sustainability on their websites or on the LSEG dashboard for private firms, or in some instances, such as TRIANEL GMBH, which had a publicly accessible annual report with no mention of their environmental or social activities and only a focus on the governance aspect of the firm.

To further validate the model's predictions, we have cross-referenced these rankings with real-world examples of sustainability practices. One such example is NISSIN Foods GMBH, which was consistently predicted by the model as a sustainable company. NISSIN's inclusion in the Dow Jones Sustainability Index for four consecutive years confirmed the accuracy of the model's prediction in predicting sustainable companies, as the Dow Jones Index is a globally recognized benchmark for sustainability. Another example is FrieslandCampina issuing 300 million euros ESG-linked Schuldschein.

4.2 Discussion of Model Performance for Private Firms

The strong performance of financial ratios such as solvency and liquidity in predicting sustainability highlights the significant role of financial health in driving corporate ESG outcomes. This supports the resource-based view (RBV) theory, which posits that firms with better access to resources, including financial stability, are more capable of adopting long-term, sustainable practices. Firms with higher solvency ratios are better positioned to meet long-term financial obligations and invest in sustainability initiatives, while liquidity ratios ensure firms can manage short-term risks and uncertainties, thus maintaining operational continuity even in times of economic fluctuation.

Moreover, these findings align with the Triple Bottom Line (TBL) framework, which suggests that financially stable companies are more likely to balance their financial objectives with environmental and social goals. Financially sound firms can allocate resources toward sustainable operations without sacrificing profitability. This indicates a reinforcing relationship between a company's financial health and its ability to engage in responsible corporate behavior, which is reflected in higher ESG scores.

The analysis of sustainability predictions across various industries reveals notable differences, particularly when considering European companies. The chemicals, petroleum, rubber, and plastics industry demonstrates one of the lowest average sustainability scores. This is consistent with broader trends observed in Europe, where these industries face significant challenges due to their

environmental impact, including high levels of greenhouse gas emissions and resource consumption. European regulatory frameworks, such as the EU Emissions Trading System (EU ETS), place these industries under stricter scrutiny, yet the transition to sustainable practices remains slow and difficult due to the nature of their operations.

Similarly, the Metals & Metal Products industry exhibits low sustainability scores, reflecting its resource-intensive production processes. Metal production, particularly steel, is a major source of carbon emissions, and despite efforts by European producers to adopt cleaner technologies, the industry's sustainability performance remains limited. The push towards a circular economy in Europe, encouraging metal recycling, has only marginally offset the environmental costs associated with primary production. In contrast, the wholesale industry, while not directly involved in manufacturing, shows a mid-range sustainability score. Wholesalers, as intermediaries in the supply chain, have a complex role in sustainability. Although they do not produce goods, they are part of broader supply chains that can have significant environmental impacts. European companies in this sector are increasingly pressured to ensure that the products they distribute meet environmental and sustainability standards, in line with the objectives of the EU Green Deal.

The Food & Tobacco Manufacturing industry, which scored higher in terms of sustainability, benefits from stringent European regulations on food production. The European focus on reducing carbon footprints, improving supply chain transparency, and adopting sustainable sourcing practices has positively influenced this sector's sustainability performance. However, the tobacco manufacturing segment within this industry continues to face environmental challenges, despite overall improvements in sustainability.

Lastly, the Retail industry shows relatively higher sustainability scores, reflecting ongoing efforts by European retailers to implement greener practices. These efforts include reducing plastic use, adopting sustainable packaging, and optimizing supply chains. Consumer demand for ethically produced goods has been a significant driver of these initiatives, alongside European regulatory pressures aimed at improving environmental standards in retail operations.

Overall, the data indicates that while high-impact industries such as chemicals and metals are slower to adopt sustainable practices, sectors like retail and food production are showing more substantial progress. This reflects broader European trends, where consumer demand, alongside stringent regulatory frameworks, is shaping the sustainability landscape across industries.

The absence of actual sustainability scores for private firms means that the predictions generated by the model cannot be directly validated against ground truth data. However, based on the model’s performance on public firms, it is reasonable to expect that the predictions for private firms reflect

meaningful patterns linked to financial health and industry characteristics. The model showed strong performances. It was able to generalize well from the public dataset to the private dataset due to the shared financial metrics across both types of firms. This generalization is critical because it demonstrates that financial performance indicators can be a reliable proxy for sustainability, even in firms that were not assigned an ESG score. it also showed a robust performance in regards to manual validation. The final validation revealed that 72% of the model's predictions aligned with the real-world ESG performance of the randomly selected companies, GRI reports, and other publicly available data. The private companies predicted as sustainable generally had robust reporting practices, and clear commitments to sustainability, both in their GRI reports and through independent sustainability credentials like the Dow Jones Sustainability Index. For firms predicted as non-sustainable, the lack of transparency, involvement in controversies, or absence of formal ESG reporting justified their classification. This combination of quantitative ESG scoring and qualitative real-world validation ensured that the model's predictions were both accurate and credible.

During the validation, we noticed that the majority of companies classified by the model as sustainable had Global Reporting Initiative (GRI) reports. These reports followed GRI reporting standards, which focus on transparency and provide detailed insights into the company's performance on all three ESG pillars. For example, firms like FrieslandCampina and Commerzbank, which were predicted as sustainable, adhered to GRI standards, and their reports provided extensive information on their sustainability efforts.

Moreover, these companies had engaging investor relations pages that offered in-depth information about their ESG initiatives. This included clear data on environmental impact, employee welfare, and governance practices, indicating a higher level of commitment to sustainability. The availability of such detailed information on their ESG performance reinforced the model's predictions.

Although geography was not a primary focus of the model, it is worth noting that firms operating in regions with stricter environmental regulations or higher levels of sustainability awareness were more likely to be predicted as sustainable. This observation aligns with the expectation that local regulatory environments can significantly influence corporate behavior regarding sustainability.

These results can be linked back to the thorough processing of the private dataset, including the imputation of missing values and hot encoding of industries, which contributed to the model's effectiveness in making predictions. By preserving the structure and integrity of the data, the model was able to identify meaningful relationships between financial metrics and sustainability.

The findings from this research have broad implications for both investors and policymakers. For investors, particularly those focused on responsible or sustainable investing, the ability to predict a company's ESG performance based solely on financial data is invaluable. It enables them to assess the sustainability of private firms, which often lack comprehensive ESG disclosures. This model provides an additional layer of analysis that can complement traditional financial assessments when making investment decisions.

For policymakers, the results highlight the importance of standardized ESG reporting frameworks for private firms. The predictive model's reliance on financial data underscores the need for more transparent and comprehensive ESG disclosures, particularly for firms that do not currently face mandatory reporting requirements. Policymakers should consider expanding ESG reporting guidelines to cover more firms, including private and smaller businesses, to ensure that all companies are held accountable for their sustainability practices.

The strong correlation between financial metrics and sustainability suggests that for non-reporting firms, investors and regulators can use financial data as a reliable proxy to assess ESG potential.

4.3 Model Limitations

The model, while demonstrating notable strengths in identifying sustainable firms, presents several limitations that need to be addressed for a more comprehensive evaluation of sustainability. A key limitation is the model's relatively low specificity, as evidenced by its struggle to correctly classify non-sustainable firms. This shortcoming suggests a tendency to overestimate sustainability, potentially due to the model's reliance on financial data alone, which may not capture the full scope of a firm's ESG performance. In real-world applications, this limitation can lead to misclassification, where firms that do not adhere to sustainable practices are incorrectly identified as sustainable, thereby undermining the integrity of ESG assessments.

Another limitation stems from the issue of class imbalance within the dataset. The model's high accuracy might reflect a bias toward the majority class (non-sustainable firms), rather than a balanced evaluation of both classes. Despite a high recall rate for sustainable firms, the model's lower balanced accuracy highlights the difficulty in equally recognizing non-sustainable firms, suggesting that the current dataset does not provide a sufficiently diverse representation of both classes. This imbalance, coupled with the observed low specificity, indicates that the model may not be robust enough for generalization in diverse contexts.

Moreover, the use of financial indicators as the sole predictors may inadequately address the qualitative aspects of sustainability, such as corporate governance, environmental policies, and stakeholder engagement. Previous research has emphasized the multi-dimensional nature of ESG assessment (Montiel & Delgado-Ceballos, 2014), implying that financial metrics alone cannot fully encapsulate a firm's sustainability profile. Therefore, while the model exhibits promising potential, these limitations underscore the necessity of integrating a broader range of data inputs, including qualitative ESG metrics, to enhance its predictive accuracy and reliability.

Furthermore, the use of financial ratios may only favor large corporations to classify them as sustainable and ignore mid to small-sized companies with more sustainable practices but lower financial performance.

In this research, we have tried to manually cross-reference the results of the private firm's model prediction through the real-world validation process indicated in Chapter 3. However, in the manual validation, we have referred from adjusting weights to different aspects of ESG with the goal of simplifying the process which omits a crucial part of ESG rating, which is sector materiality.

4.4 Conclusion

This research aimed to explore the potential of using financial ratios to predict corporate sustainability, specifically focusing on private firms that often lack formal Environmental, Social, and Governance (ESG) reporting. By utilizing a Random Forest classifier, the study highlighted the significance of solvency and liquidity ratios as primary indicators of sustainability, demonstrating a strong correlation between financial health and a firm's ability to implement sustainable practices. Firms with greater financial stability, as reflected in these ratios, are more likely to invest in long-term sustainability initiatives, which is consistent with existing theories on corporate finance and sustainability.

However, the analysis also revealed several limitations that need to be addressed in future research. The reliance on financial data alone oversimplifies the complexities of sustainability, particularly in terms of social and governance factors that are not easily captured by financial metrics. While financial ratios offer a useful proxy for sustainability assessment, they do not encompass the full spectrum of factors that contribute to a firm’s ESG performance. This limitation is particularly evident in private firms, where qualitative aspects such as governance practices, employee relations, and environmental impact are often not included in financial disclosures.

One of the primary challenges identified in this study is the imbalance in the dataset, which skewed the model’s ability to accurately classify non-sustainable firms. The training dataset, drawn from public companies with comprehensive ESG reporting, was heavily weighted toward firms with positive sustainability outcomes, leading to a bias in predictions when applied to private firms. This highlights the need for balanced datasets in machine learning models to ensure accurate predictions across both sustainable and non-sustainable firms. Techniques such as Synthetic Minority Over-sampling (SMOTE) or cost-sensitive learning could be explored in future research to mitigate this imbalance.

Geographical variations in sustainability outcomes also emerged as a critical factor, with firms based in regions with stronger regulatory frameworks and higher public awareness of sustainability showing better ESG performance. This underscores the importance of local regulatory environments and societal expectations in shaping corporate behavior. Firms operating in regions with stringent environmental regulations are more likely to adopt transparent and sustainable practices, highlighting the need for more region-specific variables in future models. Policymakers should consider strengthening ESG reporting requirements, particularly in regions where sustainability practices are lagging, to ensure a more uniform approach to corporate sustainability across different geographic locations.

In terms of validation, the model's predictions were supported by real-world examples, such as companies like NISSIN Foods GMBH and FrieslandCampina, which were correctly classified as sustainable based on their detailed ESG disclosures and inclusion in recognized sustainability indices like the Dow Jones Sustainability Index. However, the study also found that firms with fewer disclosures or those involved in controversies required manual adjustments to their sustainability scores. This points to the limitations of current predictive models in fully automating ESG assessments, particularly for firms that do not consistently disclose their sustainability practices. Future research could incorporate real-time data sources, such as news sentiment analysis or media monitoring, to better capture qualitative factors like governance scandals or environmental controversies that impact a firm's sustainability status.

The findings of this study carry important implications for both investors and policymakers. For investors, particularly those engaged in responsible or sustainable investing, the ability to predict a firm's ESG performance using financial ratios is a valuable tool. It allows for a more informed assessment of private firms, which often lack the same level of transparency as their publicly listed counterparts. This model provides an additional layer of analysis, complementing traditional financial metrics with sustainability insights, and offers a way for investors to evaluate long-term risk and value from an ESG perspective.

For policymakers, the results underscore the pressing need to establish standardized ESG reporting frameworks that include private firms. While the reliance on financial data has proven useful in this study, it also points to the necessity for more comprehensive and mandatory ESG disclosures. Expanding ESG reporting requirements to private firms would enhance transparency, reduce information asymmetry, and allow for more accurate assessments of corporate sustainability. This could further facilitate sustainable business practices and encourage firms to adopt responsible governance models that align with global sustainability goals.

Looking ahead, there are several directions in which future research can build upon the findings of this study. First, addressing the dataset imbalance and improving the model's ability to classify non-sustainable firms will be essential for developing more robust and reliable predictions. Second, incorporating qualitative data sources, such as stakeholder surveys, employee feedback, and independent governance ratings, could provide a more holistic view of a firm's sustainability profile. Lastly, future models could benefit from the integration of real-time data, allowing for dynamic assessments of firms' sustainability performance in response to changes in corporate behavior, market conditions, and regulatory environments.

This study has demonstrated the utility of financial ratios in predicting corporate sustainability, particularly for private firms without formal ESG reporting. While financial health is an important determinant of sustainability, it is clear that a more nuanced approach is needed—one that incorporates both quantitative and qualitative data to capture the full range of factors influencing ESG performance. As the demand for sustainable investing and corporate accountability continues to grow, the development of predictive models that combine financial and ESG factors will play a crucial role in shaping the future of sustainable finance and corporate governance.

Bibliography

Amel-Zadeh, A. and Serafeim, G., 2018. Why and how investors use ESG information: Evidence from a global survey. SSRN Electronic Journal, 74(3), pp.15-28.

Ambec, S. and Lanoie, P., 2008. Does it pay to be green? A systematic overview. Acadethe of Management Perspectives, 22(4), pp.45-62.

Bebchuk, L.A. and Weisbach, M.S., 2010. The state of corporate governance research. The Review of Financial Studies, 23(3), pp.939-961.

Buck, L., Brennan, J., Shandal, V., Brigl, M., Fischer, G., Stoffers, M. and Remillard, M., 2024. How private equity can converge on ESG data. BCG Global, pp.45-57.

Carroll, A.B., 1999. Corporate Social Responsibility: Evolution of a Definitional Construct. Business & Society, 38(3), pp.268-295.

Chen, S., Song, Y. and Gao, P., 2023. Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance. Journal of Environmental Management, 345, pp.45-60.

Clark, G.L., Feiner, A. and Viehs, M., 2014. From the stockholder to the stakeholder: How sustainability can drive financial outperformance. SSRN Electronic Journal, pp.25-38.

D'Amato, V., D'Ecclesia, R. and Levantesi, S., 2021. Fundamental ratios as predictors of ESG scores: A machine learning approach. Decisions in Economics and Finance, 44, pp.1087-1110.

Damodaran, A., 2012. Investment valuation: Tools and techniques for determining the value of any asset. John Wiley & Sons, pp.50-75.

Eccles, R.G., Ioannou, I. and Serafeim, G., 2014. The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), pp.2835-2857.

Elkington, J., 1998. Accounting for the triple bottom line. Measuring Business Excellence, 2(3), pp.18-22.

European Commission, 2013. Directive 2013/34/EU on the annual financial statements, consolidated financial statements, and related reports of certain types of undertakings. Official Journal of the European Union, pp.15-30.

Eurostat, 2008. NACE Rev. 2 Statistical classification of economic activities in the European Community, pp.10-45.

Forbes, K., 2024. What is corporate sustainability? Vanderbilt University, pp.34-50.

Friede, G., Busch, T. and Bassen, A., 2015. ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), pp.210-233.

Fu, T. and Li, J., 2023. An empirical analysis of the impact of ESG on financial performance: The moderating role of digital transformation. Frontiers in Environmental Science, 11, pp.112-125.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning. Springer Series in Statistics, pp.88-110.

Hull, J.C., 2020. Machine learning in business: An introduction to the world of data science. 2nd ed. University of Toronto, pp.25-50.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, pp.1137-1143.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R News, 2(3), pp.18-22.

Louppe, G., 2014. Understanding random forests: From theory to practice, pp.45-60.

McKinsey & Company, n.d. Investors want to hear from companies about the value of sustainability, pp.5-20.

Mintarya, L.N. et al., 2023. Machine learning approaches in stock market prediction: A systematic literature review. Procedia Computer Science, 216, pp.96-102.

Oundouri, P., Pittis, N. and Plataniotis, A., 2022. The impact of ESG performance on the financial performance of European area companies: An empirical examination. Environmental Sciences Proceedings, 15, pp.13-20.

Peters, G.F. and Romi, A.M., 2014. Does the voluntary adoption of corporate governance mechanisms improve environmental risk disclosures? Evidence from greenhouse gas emission accounting. Journal of Business Ethics, 125(4), pp.637-666.

Rakopoulou, V., 2011. A review of fundamental and technical stock analysis techniques. Journal of Stock Market Research, 12(2), pp.113-126.

Schafer, J.L., 2000. Analysis of incomplete multivariate data. Technometrics, 42(2), pp.213-225.

Serafeim, G., 2019. The investor revolution. Harvard Business Review, pp.30-45.

Sparkes, R., 2001. Ethical investment: Whose ethics, which investment? John Wiley & Sons, pp.60-75.

Van Dijk, L., Hijink, S. and Veld, L. in ’t, 2024. Corporate sustainability reporting. In EBI Studies in Banking and Capital Markets Law, pp.185-210.

WatchWire, n.d. The intersection of artificial intelligence and corporate sustainability, pp.50-65.

World Economic Forum, 2021. 3 paradigm shifts in corporate sustainability to ESG, pp.12-30.

García, F., González-Bueno, J., Guijarro, F. and Oliver, J., 2020. Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach. Sustainability, 12(8), pp.3324-3340.

Del Vitto, A., Marazzina, D. and Stocco, D., 2023. ESG ratings explainability through machine learning techniques. Annals of Operations Research, pp.75-90.

Krappel, T., Bogun, A. and Borth, D., 2021. Heterogeneous ensemble for ESG ratings prediction. arXiv, pp.15-25.

Hossin, M. & Sulaiman, M. N. (2015) 'A Review on Evaluation Metrics for Data Classification Evaluations', International Journal of Data Mining & Knowledge Management Process, 5(2), pp. 1-11.

Montiel, I. and Delgado-Ceballos, J., 2014. Defining and measuring corporate sustainability: Are we there yet? Organization & Environment, 27(2), pp.113-139.

Moody's Analytics, 2021. Using ESG score predictor: A methodological framework to estimate ESG scores, pp.20-45.

UNPRI, 2024. Principles for responsible investment, pp.5-25.

EY, n.d. Why ESG performance is growing in importance for investors, pp.50-65.

McKinsey, n.d. Investors want to hear from companies about the value of sustainability, pp.30-45.

WatchWire, n.d. The intersection of artificial intelligence and corporate sustainability, pp.15-30.

EY, n.d. Artificial intelligence & ESG stakes: Discussion paper, pp.25-40.

Harvard Law School Forum on Corporate Governance, 2023. ESG reporting for private companies, pp.45-60.

Liang, G., et al., 2024. Balancing sustainability and innovation: The role of artificial intelligence in shaping mining practices for sustainable mining development. Resources Policy, 90, pp.104-120.

Chen, P., Chu, Z. and Zhao, M., 2024. The road to corporate sustainability: The importance of artificial intelligence. Technology in Society, 76, pp.85-100.

Nosirov, I., Avulchayeva, F., Yormatov, I. and Yuldasheva, N., 2024. AI and corporate sustainability: Exploring the environmental and social impacts of AI integration. 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), pp.30-45.

Fakdawer, N.S., 2024. The role of accounting practices in advancing the agenda of green finance and impact investing. Advances in Applied Accounting Research, 2(2), pp.94-109.

Enyinful, K. and Aboagye, E., 2023. ESG and firm performance: Evidence from selected countries in Europe, University of Stavanger, pp.20-35.

Chen, S., Song, Y. and Gao, P., 2023. Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance. Journal of Environmental Management, 345, pp.45-60.

Alsayegh, M.F., Abdul Rahman, R. and Homayoun, S., 2023. Corporate sustainability performance and firm value through investment efficiency. Sustainability, 15, pp.305-320.

Ashrafi, M., Magnan, G.M., Adams, M. and Walker, T.R., 2020. Understanding the conceptual evolutionary path and theoretical underpinnings of corporate social responsibility and corporate sustainability. Sustainability, 12, pp.760-780.

Diez-Cañamero, B., Bishara, T., Otegi-Olaso, J.R., Minguez, R. and Fernández, J.M., 2020. Measurement of corporate social responsibility: A review of corporate sustainability indexes, rankings, and ratings. Sustainability, 12, pp.2153-2168.

Refinitiv, 2020. Worldscope Database - Data Definitions Guide (Issue 16), pp.30-50.

Melville, A., 2022. International financial reporting: A practical guide. 8th ed. Pearson Education Limited, pp.60-90.

Barnett, M.L. and Salomon, R.M., 2006. Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance. Strategic Management Journal, 27, pp.1101-1122.

KPMG International, 2019. Corporate tax: A critical part of ESG, pp.40-60.

Krüger, P., 2015. Climate change and firm valuation: Evidence from a quasi-natural experiment. Swiss Finance Institute Research Paper Series No. 15-40, pp.100-120.

Raza, H., Khan, M.A., Mazliham, M.S., Alam, M.M., Aman, N. and Abbas, K., 2022. Applying artificial intelligence techniques for predicting the environment, social, and governance (ESG) pillar score based on balance sheet and income statement data: A case of non-financial companies of USA, UK, and Germany. Frontiers in Environmental Science, 10, pp.150-175.

Drakopoulou, V., 2015. A review of fundamental and technical stock analysis techniques. Journal of Stock & Forex Trading, 5(1), pp.163-180.

GasTerra, 2023. Annual report 2023, pp.45-65.

Viterra, 2023. Full year report 2023, pp.50-80.

Zentiva, 2023. 2023 Non-financial disclosure sustainability report, pp.20-40.

Fu, T. and Li, J., 2023. An empirical analysis of the impact of ESG on financial performance: The moderating role of digital transformation. Frontiers in Environmental Science, 11, pp.125-140.

Koundouri, P., Pittis, N. and Plataniotis, A., 2022. The impact of ESG performance on the financial performance of European area companies: An empirical examination. Environmental Sciences Proceedings, 15, pp.13-25.

Sonko, K.N.M. and Sonko, M., 2023. Dethestifying environmental, social, and governance (ESG): Charting the ESG course in Africa. Springer Nature Switzerland AG, pp.60-80.

Del Vitto, A., Marazzina, D. and Stocco, D., 2023. ESG ratings explainability through machine learning techniques. Annals of Operations Research, pp.95-120.

Berg, F., Kölbel, J.F. and Rigobon, R., 2022. Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), pp.1315-1344.

Paul, Weiss, Rifkind, Wharton & Garrison LLP, 2021. ESG ratings and data: How to make sense of disagreement, pp.20-45.

Witzel, M. and Bhargava, N., 2023. AI-related risk: The merits of an ESG-based approach to oversight. Centre for International Governance Innovation, pp.75-90.

Friede, G., Busch, T. and Bassen, A., 2015. ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), pp.210-233.

Doyle, T.M., 2018. Ratings that don't rate: The subjective world of ESG ratings agencies. American Council for Capital Formation, pp.30-55.

D'Amato, V., D'Ecclesia, R. and Levantesi, S., 2021. Fundamental ratios as predictors of ESG scores: A machine learning approach. Decisions in Economics and Finance, 44, pp.1087-1110.

DasGupta, R., 2022. Financial performance shortfall, ESG controversies, and ESG performance: Evidence from firms around the world. Finance Research Letters, 46, pp.102-115.

Sheehy, B. and Farneti, F., 2021. Corporate social responsibility, sustainability, sustainable development, and corporate sustainability: What is the difference, and does it matter? Sustainability, 13(5965), pp.130-145.

Carroll, A.B., 2008. A history of corporate social responsibility: Concepts and practices. In Crane, A., McWilliams, A., Matten, D., Moon, J. and Siegel, D. (eds.) The Oxford Handbook of Corporate Social Responsibility. Oxford University Press, pp.60-75.

Carroll, A.B., 1999. Corporate social responsibility: Evolution of a definitional construct. Business & Society, 38(3), pp.268-295.

Townsend, B., 2020. From SRI to ESG: The origins of socially responsible and sustainable investing. The Journal of Impact and ESG Investing, 1(1), pp.19-35.

Derqui, B., 2020. Towards sustainable development: Evolution of corporate sustainability in multinational firms. Corporate Social Responsibility and Environmental Management, 27(6), pp.2712-2723.

Montiel, I. and Delgado-Ceballos, J., 2014. Defining and measuring corporate sustainability: Are we there yet? Organization & Environment, 27(2), pp.113-139.

Siew, R.Y.J., 2015. A review of corporate sustainability reporting tools (SRTs). Journal of Environmental Management, 164, pp.180-195.

Meuer, J., Koelbel, J. and Hoffmann, V.H., 2019. On the nature of corporate sustainability. Organization & Environment, 33(3), pp.319-341.

Renneboog, L., Ter Horst, J. and Zhang, C., 2008. Socially responsible investments: Institutional aspects, performance, and investor behavior. Journal of Banking & Finance, 32(9), pp.1723-1742.

Nofsinger, J.R., Sulaeman, J. and Varma, A., 2019. Institutional investors and corporate social responsibility. Journal of Corporate Finance, 58, pp.700-725.

Drempetic, S., Klein, C. and Zwergel, B., 2020. The influence of firm size on the ESG score: Corporate sustainability ratings under review. Journal of Business Ethics, 167(2), pp.333-360.

García, F., González-Bueno, J., Guijarro, F. and Oliver, J., 2020. Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach. Sustainability, 12(8), pp.3324-3340.

Krappel, T., Bogun, A. and Borth, D., 2021. Heterogeneous ensemble for ESG ratings prediction. In 4th KDD Workshop on Machine Learning in Finance (KDD-MLF '21), pp.25-40.

D'Amato, V., D'Ecclesia, R. and Levantesi, S., 2021. ESG score prediction through random forest algorithm. Computational Management Science, 19, pp.347-373.

Refinitiv, 2022. Environmental, social, and governance (ESG) scores from Refinitiv, pp.50-70.

Bureau van Dijk, 2011. Orbis user guide, pp.30-45.

Kayani, U.N., Gan, C., Rabbani, M.R. and Trichilli, Y., 2023. Is short-term firm performance an indicator of sustainable financial performance? Empirical evidence. Studies in Economics and Finance, pp.100-120.

Loiseau-Aslanidi, O., Piscaglia, S. and Gonzalez, B.S., 2022. Using ESG score predictor: A methodological framework to estimate ESG scores. Moody's Analytics, pp.75-90.

Martini, A., 2021. Socially responsible investing: From the ethical origins to the sustainable development framework of the European Union. Environment, Development and Sustainability, 23, pp.16874-16890.

Sustainable Development Solutions Network, 2022. Financing the joint implementation of Agenda 2030 and the European Green Deal, pp.45-65.

Morishita, M., Shimizu, N., Katori, T., Ikeda, E. and Chenet, H., 2020. Japan-EU comparative analysis on sustainable finance policy. Institute for Global Environmental Strategies, pp.40-55.

Uzsoki, D., 2020. Sustainable investing: Shaping the future of finance. International Institute for Sustainable Development (IISD), pp.30-50.

Moliterni, F., 2018. Sustainable investing and green finance: Boosting markets by solving ambiguities. Fondazione Eni Enrico Mattei (FEEM), pp.25-40.

Hull, J.C., 2020. Machine learning in business: An introduction to the world of data science. 2nd ed. John Hull University Press, pp.45-65.

Barbier, E.B. and Burgess, J.C., 2017. The sustainable development goals and the systems approach to sustainability. Economics: The Open-Access, Open-Assessment E-Journal, 11(28), pp.1-22.

Amel-Zadeh, A. and Serafeim, G., 2017. Why and how investors use ESG information: Evidence from a global survey. SSRN Electronic Journal, pp.15-35.

Sparkes, R., 2001. Ethical investment: Whose ethics, which investment? Business Ethics: A European Review, 10(3), pp.194-205.

Eccles, R.G. and Stroehle, J.C., 2018. Exploring social origins in the construction of ESG measures. SSRN Electronic Journal, pp.25-40.

Chen, C. et al., 2021. Artificial intelligence on economic evaluation of energy efficiency and renewable energy technologies. Sustainable Energy Technologies and Assessments, 47, pp.101-115.

Appendix

1. Importing Libraries

```
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

- **Explanation:** This section imports the necessary libraries:
  - pandas for data manipulation and analysis.
  - SimpleImputer for handling missing values by filling them with statistical measures (e.g., median).
  - OneHotEncoder to encode categorical data into a format suitable for machine learning models.
  - train_test_split to divide the dataset into training and testing subsets.
  - RandomForestClassifier for building a classification model based on the Random Forest algorithm.
  - classification_report and confusion_matrix from sklearn.metrics to evaluate the performance of the model.
  - numpy for numerical operations.
  - matplotlib and seaborn for visualizing data and model performance.

---

2. Loading the Data

```
def load_data(public_file, private_file):
    public_df = pd.read_excel(public_file)
```

```
    private_df = pd.read_excel(private_file)
    return public_df, private_df

public_companies_df, private_companies_df = load_data('Public12-encoded.xlsx', 'private_v2_with_ratios.xlsx')
```

- **Explanation:** Defines a function, load_data, to read the public and private companies' data from Excel files and store them in pandas DataFrames. The function accepts filenames as input and returns the loaded datasets. This structure simplifies the process of loading data and makes the code easier to adapt if the input files change.

---

3. Preprocessing the Dataset

```
def preprocess_data(df, categorical_cols, numerical_cols):
    non_numeric_strings = ['n.a.', 'n.s.', 'na', 'N/A']
    df.replace(non_numeric_strings, np.nan, inplace=True)

        df[numerical_cols] = df[numerical_cols].replace([np.inf, -np.inf], np.nan).astype(float)

    imputer = SimpleImputer(strategy='median')
    df[numerical_cols] = imputer.fit_transform(df[numerical_cols])

                encoder = OneHotEncoder(sparse_output=False, handle_unknown='ignore')
    encoded = encoder.fit_transform(df[categorical_cols])
                    encoded_df = pd.DataFrame(encoded, columns=encoder.get_feature_names_out(categorical_cols))

         final_df = pd.concat([pd.DataFrame(df[numerical_cols]), encoded_df], axis=1).reset_index(drop=True)
```

```
    return final_df, encoder, imputer
```

- **Explanation:** This function, preprocess_data, prepares the dataset for modeling:
    - **Replacing non-numeric values:** Replaces common placeholders for missing data (e.g., 'n.a.', 'N/A') with NaN to facilitate numerical operations.
    - **Handling numerical values:** Replaces infinite values with NaN and converts all numerical columns to float type to ensure consistency.
    - **Imputation:** Uses SimpleImputer to fill missing numerical values with the median, a robust method that reduces the effect of outliers.
    - **Encoding categorical variables:** Applies one-hot encoding to categorical columns, converting them into a binary format suitable for machine learning models.
    - **Combining columns:** Combines the imputed numerical data and encoded categorical data into a single DataFrame. This step ensures that both numerical and categorical information is included for modeling.
    - Returns the processed DataFrame, the encoder, and the imputer for use in other parts of the analysis.

---

4. Extracting Features and Target Variables

```
X_public = public_companies_df.drop(columns=['Company_Name', 'Sustainable'])
y_public = public_companies_df['Sustainable']
categorical_cols = X_public.select_dtypes(include=['object']).columns
numerical_cols = X_public.select_dtypes(include=[np.number]).columns
```

- **Explanation:** This section prepares the public dataset for analysis:
    - Separates the features (X_public) from the target variable (y_public), where 'Sustainable' is the target for classification.

- Identifies categorical and numerical columns in the dataset to handle them appropriately during preprocessing.

---

5. Preprocessing the Public Dataset

```
X_public_final, encoder, numerical_imputer = preprocess_data(X_public, categorical_cols, numerical_cols)
```

- **Explanation:** Calls the preprocess_data function to clean, impute, and encode the public dataset. It stores the resulting processed DataFrame in X_public_final, while also keeping the encoder and imputer objects for future use.

---

6. Splitting Data for Training and Testing

```
X_train, X_test, y_train, y_test = train_test_split(X_public_final, y_public, test_size=0.2, random_state=42)
```

- **Explanation:** Uses train_test_split to divide the processed public dataset into training and testing subsets. The test set is 20% of the total data, ensuring the model's performance can be evaluated on unseen data.

---

7. Training the Random Forest Model

```
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

- **Explanation:** Initializes a RandomForestClassifier with 100 decision trees (n_estimators=100) and trains the model on the training data. The random_state ensures reproducibility of results.

---

8. Evaluating Feature Importance

```
def plot_feature_importance(model, feature_names):
    importances = model.feature_importances_
    indices = np.argsort(importances)[::-1]

    plt.figure(figsize=(10, 6))
    plt.title("Feature Importance")
         plt.bar(range(len(feature_names)),  importances[indices],
align='center')
     plt.xticks(range(len(feature_names)), feature_names[indices],
rotation=90)
    plt.show()

plot_feature_importance(model, X_train.columns)
```

- **Explanation:** Defines plot_feature_importance to visualize the importance of features in the model. This plot helps interpret the model by showing which features most influence its predictions.

---

9. Model Evaluation

```
y_pred = model.predict(X_test)
print("Classification  Report:\n",  classification_report(y_test,
y_pred))
```

- **Explanation:** Uses the trained model to predict the target variable (Sustainable) for the test data. The classification_report provides performance metrics (precision, recall, F1-score) for evaluating the model's effectiveness.

---

10. Confusion Matrix Plotting

```
def plot_confusion_matrix(y_true, y_pred):
```

```
            conf_matrix    =    confusion_matrix(y_true,    y_pred,
normalize='all')
     sns.heatmap(conf_matrix, annot=True, fmt='.2f', cmap='Blues',
xticklabels=['Non-Sustainable',                    'Sustainable'],
yticklabels=['Non-Sustainable', 'Sustainable'])
    plt.xlabel('Predicted Label')
    plt.ylabel('True Label')
    plt.title('Confusion Matrix')
    plt.show()

plot_confusion_matrix(y_test, y_pred)
```

- **Explanation:** Defines plot_confusion_matrix to display the confusion matrix using a heatmap. This matrix shows how well the model distinguishes between sustainable and non-sustainable companies, providing insight into its accuracy and areas of improvement.

---

11. Preprocessing and Predicting for Private Companies

```
def  preprocess_private_data(df,  categorical_cols,  numerical_cols,
encoder, imputer, reference_cols):
          df.replace(['n.a.',   'n.s.',   'na',   'N/A'],   np.nan,
inplace=True)
        df[numerical_cols]  =  df[numerical_cols].replace([np.inf,
-np.inf], np.nan).astype(float)
    df[numerical_cols] = imputer.transform(df[numerical_cols])

    encoded = encoder.transform(df[categorical_cols])
                    encoded_df      =      pd.DataFrame(encoded,
columns=encoder.get_feature_names_out(categorical_cols))
```

```
        final_df = pd.concat([pd.DataFrame(df[numerical_cols]),
encoded_df], axis=1).reset_index(drop=True)

    for col in set(reference_cols) - set(final_df.columns):
        final_df[col] = 0

    return final_df[reference_cols]

X_private = private_companies_df.drop(columns=['Company_Name'])
X_private_final = preprocess_private_data(X_private,
categorical_cols, numerical_cols, encoder, numerical_imputer,
X_public_final.columns)

private_predictions = model.predict(X_private_final)
private_companies_df['Sustainability_Prediction'] =
private_predictions
```

- **Explanation:** The preprocess_private_data function applies similar preprocessing steps to the private dataset:
    - Cleans and encodes data using the previously trained encoder and imputer.
    - Ensures the private dataset's columns match those of the public dataset for compatibility with the model.
    - Uses the trained model to predict the sustainability status of private companies, adding these predictions to the DataFrame.

---

12. Saving the Predictions

```
private_companies_df[['Company_Name',
'Sustainability_Prediction']].to_csv('private_company_predictions.
csv', index=False)
print("Predictions saved to 'private_company_predictions.csv'")
```

- **Explanation:** Saves the predicted sustainability