THE INTEGRATION OF MACHINE LEARNING AND DECISION SUPPORT SYSTEM IN SUSTAINABILITY PERFORMANCE MANAGEMENT

NG JIA YING

A project report submitted in partial fulfilment of the requirements for the award of the degree of Bachelor of Civil Engineering (Environmental) with Honours

Faculty of Engineering and Green Technology
Universiti Tunku Abdul Rahman

September 2024

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature :

Name : Ng Jia Ying

ID No. : 20AGB06527

Date : 20th September 2024

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"THE INTEGRATION OF MACHINE LEARNING AND DECISION SUPPORT SYSTEM IN SUSTAINABILITY PERFORMANCE MANAGEMENT"** was prepared by **NG JIA YING** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Civil Engineering (Environmental) with Honours at Universiti Tunku Abdul Rahman.

Approved by,

Signature : _________________________

Supervisor : Assoc. Prof. ChM. Ts. Dr. Tan Kok Weng

Date : 21/9/2024

ACKNOWLEDGEMENTS

I would like to thank everyone who contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Assoc. Prof. ChM. Ts. Dr. Tan Kok Weng for his invaluable advice, guidance and his enormous patience throughout the development of the research.

THE INTEGRATION OF MACHINE LEARNING AND DECISION SUPPORT SYSTEM IN SUSTAINABILITY PERFORMANCE MANAGEMENT

ABSTRACT

This study developed a framework of a machine learning embedded decision support system that supports company sustainability performance management activities through the assessment of sustainability reports and generation of sustainability scores. The sustainability report assessment function hopes to assist companies in compliance with sustainability reporting standards to improve stakeholder engagement and enhance financing prospects. A rule-based system complemented by Natural Language Processing (NLP) technology is adopted for the system. The role of sustainability scores is to provide a direct indicator of the company sustainability performance. The machine learning model, Random Forest Regressor, is deployed to evaluate the performance of the machine learning model in generating sustainability scores under a supervised learning style. The data used in the development of the machine learning model is extracted from company sustainability reports available online. The results of model testing deliver promising results with the performance of the model improving with sample size. However, the model failed to deliver consistently accurate predictions, mainly due to the small data size and the imbalance distribution of data in the database. Lastly, recommendations for the challenges of machine learning integration with sustainability performance management are suggested for the improvement in the data collection and processing during database preparation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

n	the total number of datasets in the sample
P_i	predicted value of the dependent variable
R^2	Coefficient of Determination
var	variance
Y_i	actual value of the dependent variable
$\bar{Y}$	means of the actual dependent variables

AHP	Analytical Hierarchy Process
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANOVA	Analysis of Variance
BERT	Bidirectional Encoder Representations from Transformers
BIST	Borsa Istanbul
CART	Classification and Regression Tree
CLARA	Clustering Large Applications
CLARANS	Clustering Large Applications based on RANdomized Search
CSR	Corporate Social Responsibility
CSV	Comma Separated Values
DBMS	Database Management System
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DENCLUE	Density-based Clustering
DSS	Decision Support System
ELM	Extreme Learning Machine
ESG	Environmental, Social and Governance
EVS	Explained Variance Score
GDBSCAN	Generalized Density-based Spatial Clustering of Applications with Noise
GDPR	General Data Protection Regulation
GRI	Global Reporting Initiative

IBM	International Business Machines Corporation
IEEE	The Institute of Electrical and Electronics Engineers
JPEG	Joint Photographic Experts Group
KPI	Key Performance Indicators
LLP	Limited Liability Partnership
LSEG	London Stock Exchange Group
MAE	Mean Absolute Error
MCDA	Multi-Criteria Decision Analysis
MLR	Multiple Linear Regression
MITI	Ministry Investment, Trade and Industry
MSCI	Morgan Stanley Capital International
MSE	Mean Squared Error
NIMP	National Industrial Master Plan
NIP	National Investment Plan
NLP	Natural Language Processing
PDCA	Plan, Do, Check, Act
PDF	Portable Document Format
PLC	Publicly Listed Company
PNG	Portable Network Graphics
R&D	Research and Development
RF	Random Forest
RMSE	Root Mean Squared Error
SDG	Sustainable Development Goals
SME	Small and Medium-sized Enterprises
SPE	Sustainability Performance Evaluation
SQL	Structured Query Language
SVM	Support Vector Machines
T5 Transformer	Text-to-Text Transfer Transformer
TCFD	Task Force on Climate-related Financial Disclosures
TEJ	Taiwan Economic Journal
UN	The United Nations
URL	Uniform Resource Locators
US	The United States
USD	United States Dollar

XBRL	eXtensible Business Reporting Language
XGB	eXtreme Gradient Boosting

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

In 2015, the United Nations (UN) launched the 2030 Agenda for Sustainable Development with its 17 Sustainability Development Goals (SDGs) as a response to growing environmental and humanitarian pressure (United Nations, 2015). The agenda and the SDGs were developed to secure a peaceful and prosperous future for present and future youths. There are 17 SDGs tackling economic, humanitarian, environmental, and governance issues. Countries around the world have adopted the SDGs and integrated them into country policies and development plans. Malaysia has incorporated sustainability into numerous national policies, such as the 12th Malaysia Plan, the National Investment Plan (NIP), and the National Industrial Master Plan (NIMP) for 2030 (Ministry of Investment, Trade and Industry, 2023). The emphasis on sustainability has trickled down to the corporate world.

Sustainability performance management has gained a newfound importance for businesses around the world, including Malaysian businesses. Meeting the sustainability agenda is influential to long-term business survival as it affects financial prospects, company image, and consumer trust in companies. At the beginning of 2020, global investment in sustainable projects had risen to 35.3 trillion US Dollars (Global Sustainable Investment Alliance, 2021) while the sustainable transition of the economy in Southeast Asia is projected to generate 1 trillion US Dollars' worth of economic opportunity (Hardcastle and Mattios, 2020). Business operations have been negatively impacted by the current progression of anthropogenic climate change through financial losses, supply chain disruption, and physical threats (Kalogiannidis *et al.*, 2024). Practising sustainability performance management is beneficial to the survival hood of a business in a climate-change future.

In practice, sustainability performance management is carried out in feedback loops consisting of several processes, namely sustainability goal setting, action plan development, progress measurement, reporting of progress and continuous, strategic adjustments to the action plans to yield better results. Sustainability reporting is the practice of publicizing company strategies and progress on sustainability matters. It has now become an important part of company management due to its importance in attaining investment, promoting transparency, and fostering consumer trust.

Meanwhile, the development of machine learning technology has led to significant improvements in the efficiency of data processing, posing solutions to many of society's problems, one of which is sustainability issues. Machine learning models have been widely used in environmental, economic, and social topics (Nishant, Kennedy and Corbett, 2020). The pairing of machine learning and decision support systems (DSS) can potentially propel the sustainability transition in industries through the provision of information insights for sustainability decision-making.

1.2 Problem Statements

Despite the benefits of sustainability performance management and the potential repercussions of not managing sustainability issues, businesses in Malaysia remain reluctant to adopt sustainability principles. The Malaysian Business Sustainability Pulse Report 2022 published by the UN Global Compact Network Malaysia and Brunei reported that 47% of the Malaysian private sector is not committed to any SDGs with 34 % of the surveyed companies perceiving SDGs as not relevant to their business operations. The findings of the report revealed the lack of progress in sustainability practice adoption in Malaysian companies (UN Global Compact Network Malaysia &

Brunei, 2022). The Edge had cited companies having a “Business-As-Usual” mindset and the lack of sustainability expertise among others as the challenges of Malaysian private sectors (Nadar, 2023).

Although sustainability has gained significant importance in the corporate climate in the international community, the progress of sustainability transition in the Malaysian industry remains stagnant. A report from the Ministry of Investment, Trade and Industry (MITI) cited the lack of sustainability expertise and resources in private companies as one of the reasons hampering the progress of sustainability transition in Malaysia as the companies would not understand their current sustainability standings. Inaccessibility to the assessment system further complicates the adoption of sustainability practices into company operations (Ministry of Investment, Trade and Industry, 2023). The launch of the i-ESG framework by MITI serves as a starting point for private companies to incorporate sustainability principles and practices into their organizations. However, the compliance of the produced report with the established standards and the interpretation of the results remains a problem. Hence, there is a need for a system that can assess company sustainability performance and the generated sustainability reports and provide informational insights from the results displayed in the report to optimize sustainability efforts of an organization.

1.3 Aims and Objectives

The objectives of the thesis are shown as the following:

i) To develop a machine learning integrated decision support system (DSS) framework for the evaluation and management of sustainability performance through literature review means.

ii) To assess the feasibility of machine learning approach in the sustainability decision support system (DSS).

iii) To suggest solutions to the challenges of machine learning applications in sustainability performance management.

CHAPTER 2

LITERATURE REVIEW

2.1 Corporate Sustainability

Corporate sustainability is used ambiguously in literature, it can refer to the company's contribution to sustainability development or performance on corporate social responsibility or it can simply be the long-term survival of the company (Cantele, Landi and Vernizzi, 2024). Despite its indistinct usage in past literature, corporate sustainability communicates the idea of a company’s responsibilities extending beyond its financial performance and including social and environmental obligations.

Although corporate sustainability and corporate social responsibility are used interchangeably in literature, the two are rooted in different aspects. Corporate social responsibility is of an ethical origin that mainly focuses on social issues while corporate sustainability was mainly concerned with environmental matters at the beginning of its conception (Cantele, Landi and Vernizzi, 2024). However, in subsequent development, the two terms have evolved into umbrella terms for policies and practices taken by the company for non-financially oriented causes like the environment, reducing inequalities, etc. Kantabutra and Ketprapakorn (2020) defined corporate sustainability as a leadership and management approach for a company to achieve financial profitability while fulfilling its economic, social, and environmental obligations.

Corporate sustainability has roots in various management theories such as the triple bottom line theory and the stakeholder theory.

The triple bottom line theory was proposed by John Elkington, a British Management Consultant, in 1994 (Alhaddi, 2015). The theory suggests that other than financial output, businesses should also pay attention to environmental, and societal factors when assessing company performance and success to balance sustainability with financial growth (Alhaddi, 2015). Therefore, the triple bottom line theory assigns the same importance to the three lines, environment, society, and economy when discussing company performance. The economic line focuses on the interaction of the business with the economy to ensure company survival and longevity through the evaluation of business output and profitability. The social line concerns with how the company treats people within and outside of the organization. Companies should be contributing to the betterment of the community and employees through the provision of fair wages and healthcare. The environmental line looks at how a company manages its natural resources and their environmental impacts. Ideally, a company should be working on reducing its environmental footprint and responsible allocation of natural resources.

Stakeholder theory is a business ethics and organizational management theory that hopes to benefit all stakeholders which are people and parties who are influential or under the influence of the company (Mahajan *et al.*, 2023). According to the theory, the company should consider the needs of both the shareholders and the stakeholders in operations. Stakeholder theory affects corporate sustainability and corporate social responsibility by incorporating the well-being of external communities into the company decision-making process. Banerjee, Iyer and Kashyap (2003) found that pressure from the public has a significant impact on the practice of corporate sustainability in high environmental impact industry. Such a conclusion is intuitive because maintaining a good reputation is financially beneficial to companies as reputation plays an important role in attracting and sustaining customers. The findings of Okafor, Adusei and Adeleye (2021) revealed that the higher spending on CSR causes is related to higher revenue in tech companies in the States.

These theories shaped corporate sustainability into a cross-dimensional topic that discusses the company's impact on all stakeholders on economic, social, and environmental matters.

2.1.1 Environmental, Social and Governance (ESG)

The environmental, social, and governance (ESG) was first used by the United Nations Global Compact in 2004 to discuss the role of ESG in addressing world issues such as environmental degradation, corruption, and human rights crisis (Jacobs, 2024). Pérez et al. (2022) explained the concept of ESG with examples of issues in each dimension (Table 2.1). Although the environmental and social dimensions had been defined previously, the governance dimension has yet to be discussed. The governance aspect is more related to the initiatives taken by the organization when handling sustainability affairs. A company with a proactive approach to sustainability management is beneficial to the organization's prospects in a climate change future.

Table 2.1: Description and Examples of Environmental, Social and Governance (ESG) Dimensions (Pérez *et al.*, 2022)

Pillars	**Description**	**Examples**
Environmental	Addresses impact on the physical environment and the risk of a company and its suppliers/partners from climate events	• Climate change and Greenhouse-gas emissions (GHG) • Air pollution (non-GHG) • Water and wastewater management • Waste and hazardous-materials management • Circularity

		• Biodiversity and ecosystems rehabilitation
Social	Addresses social impact and associated risk from societal actions, employees, customers, and the communities where it operates	• Labour practices • Health and safety • Community engagement • Diversity and inclusion • Community relations, local economic contribution • Product and service attributes
Governance	Assesses timing and quality of decision making, governance structure, and the distribution of rights and responsibilities across different stakeholder groups, in service of positive societal impact and risk mitigation	• Business ethics, data security • Capital allocations, supply chain management • Governance structure • Engagement and incentives • Policies • External disclosures • Position and advocacy

Since then, ESG has been included by the finance industry as a principle for sustainable investing which aims to generate financial returns and promote social and environmental well-being through capital injections into sustainable companies or projects (Li *et al.*, 2021). This idea assumes that funding will drive companies towards adopting sustainability principles into their operation which would translate into sustainability performance and enhance the environmental and social conditions of its surroundings.

Existing literature provides inconclusive evidence for the impact of ESG factors on the financial performance of companies. Li et al. (2021) concluded that the ESG factors have a complex relationship with company financial performance. Some studies showed that the management of ESG elements positively influenced company values. Matsumura, Prakash and Vera-Muñoz (2014) examined the effects of carbon emission disclosure on company value. Their findings revealed that although carbon emissions have an inverse relationship with company value, those companies who did not publish their emission data were valued less with a median of 2.3 billion USD. Disclosures of CSR can lead to improvement in industrial effluent quality and reduction in air pollutant emissions, but it is at the expense of the company's profitability (Chen, Hung and Wang, 2018). Segura et al. (2024) observed a relationship between company market value and social and environmental factors, but such phenomena were not seen in the governance elements.

Moreover, ESG is effective at company risk mitigations. Flammer (2013) argued that the adoption of environmental practices might not necessarily have a significant positive impact on stock price, but it could prevent drastic depreciation of stock price due to environmental misconduct. The insurance effect of the social dimensions in ESG was more prominent in companies with high growth opportunities, improving their financial stability (Kim, Lee and Kang, 2021).

2.2 Sustainability Performance Management

Sustainability performance management is performed in loops involving establishing sustainability goals, setting strategy, and the subsequent measuring, and assessing the performance of the company on sustainability issues after implementation of the action plan. Before the commencement of a new loop, improvement on the strategy plan should be discussed and incorporated by the decision-maker into the next plan. An important aspect of sustainability performance management is the measurement of sustainability progress. Quoting Peter Drucker, “What gets measured gets managed.”, understanding the current standing on sustainability issues allows companies to take the appropriate measures to reduce their environmental footprints and improve societal impacts (UN Global Compact Network Malaysia & Brunei, 2022). This also ensures good governance practices on the company’s side which is integral to long-term sustainable changes in companies.

Kantabutra (2024) proposed a framework for sustainability performance management after reviewing various sources of literature on the topic shown in Figure 2.1. The process consists of 6 interconnected concepts. Sustainability vision, values and assumptions in the company make up the sustainability culture of the company which plays a defining role in the successful execution and management of sustainability strategies. Company strategy should align with company culture to prevent internal rejection that leads to failed execution (Akpamah, Ivan-Sarfo and Matkó, 2021). Hence, cultivating a sustainable company culture is imperative to the sustainable development of companies. Sustainability strategies involve the process of formulating sustainability goals and plans for goal achievement. The process is followed by the execution of the strategy. The outcomes are measured and evaluated to conclude the performance of the company on its sustainability goals.

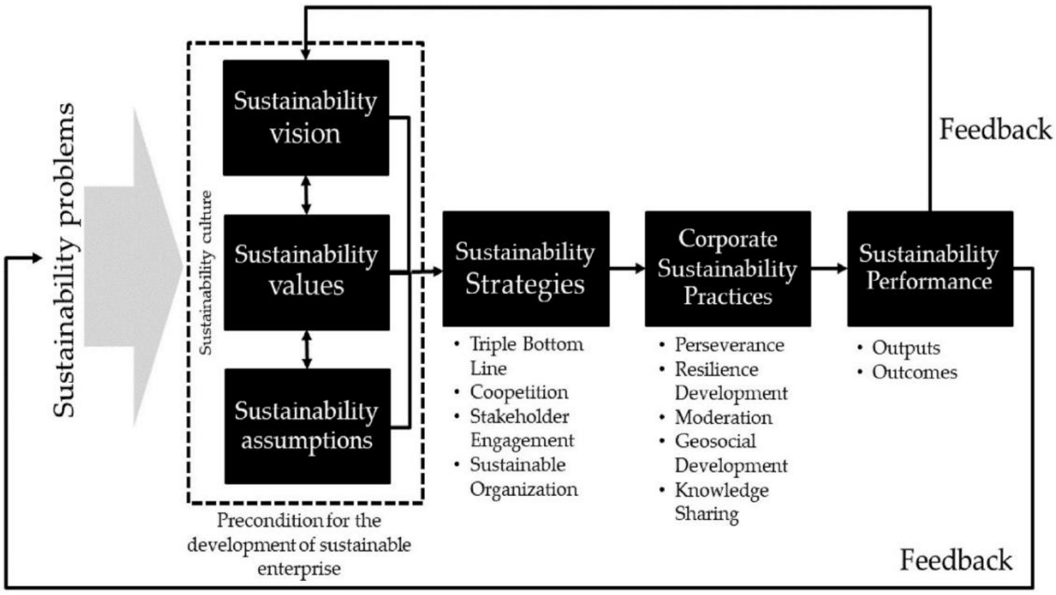


Figure 2.1: Sustainability Performance Management Framework by Kantabutra (2024)

2.2.1 Sustainability Strategy Development and Execution

The development and implementation of sustainability strategies are crucial to the growth and competitiveness of the company through the optimization of resources for maximum positive impact on financial and non-financial sectors (Rodrigues and Franco, 2019). The goal is to create financial value and optimize the societal and environmental impact of the company through the strategic allocation of company resources.

Engert and Baumgartner (2016) studied the determining factors of successful sustainability strategy implementation through case studies on global car producers in Europe. Elements of governance such as organizational structure, culture, leadership, management control, employee motivation qualifications, and communication are found to be crucial in the success of sustainability strategy execution with a prerequisite that sustainability was embedded in the core strategy of the company. This shows the importance of governance in sustainability performance management. Smith and Sharicz (2011) found that having a governance structure that cares about sustainability drives the organization towards sustainability-oriented decision-making in company activities and leads to long-term changes in companies. Moreover, the company must conduct a thorough examination of its resources and competencies before establishing the plan (Epstein and Roy, 2001). The external and internal sustainability drivers related to the operation of the company need to be incorporated for a holistic approach to company sustainability performance management (Epstein and Roy, 2001).

León - Soriano, Jesús Muñoz - Torres and Chalmeta - Rosaleñ (2010) provided the process for the formulation of a sustainability strategy. They highlighted the importance of having a consensus on sustainability goals and the preparation of company strategy by linking the goals with company activities through the cause-and-effect of the actions. Although each company has different sets of sustainability goals, the authors emphasised the sequence of the goal establishment should start from the company sustainability mission statement to the goals on social, environmental, and economic matters and detail the issues under each sector to ensure that all goals are in alignment with each other.

2.2.2 Sustainability Performance Assessment

After the implementation of the sustainability strategy, the ESG impact of the company needs to be evaluated to determine the effectiveness and efficiency of the strategy for the continuous improvement of sustainability performance.

Büyüközkan and Karabulut (2018) defined sustainability performance evaluation (SPE) as the quantification of an organization’s performance based on performance indicators to evaluate the economic, social, and environmental impact of the organization’s policies, decisions, and actions. SPE includes the process of sustainability accounting and assessment to conclude the performance of the company after the implementation of the sustainability strategy.

Sustainability accounting refers to the identification and collection of sustainability data within the organizations and its influences. The sustainability data are collected according to the key performance indicators (KPI) for the sustainability goals of the company. The KPIs in sustainability performance management are the sustainability indicators which are metrics designed to reflect the company's performance on social, environmental, governance, and economic dimensions (Contini and Peruzzini, 2022). Companies can refer to the sustainability indicator sets published by organizations such as the Global Reporting Initiative (GRI) during the performance measurement process. The GRI does provide some indicator sets that are industry-specific but only for a handful of industries like the agriculture and the oil and gas industry. Therefore, companies should conduct the material assessment prior to data collection to determine the indicators that are relevant to their company activities. The data of the sustainability indicators can be quantitative or qualitative, most social indicators have qualitative values. Moldan, Janoušková and Hák (2012) suggest the establishment of baselines and targets against the sustainability indicators as reference points for easy interpretation of sustainability progress by a distance-to-target method.

The sustainability performance assessment process analyses the data collected and transforms them into meaningful results that reflect the sustainability impact of the company (Büyüközkan and Karabulut, 2018). This is achieved through the employment of analytical integration techniques such as the Analytical Hierarchy Process (AHP) and Multiple-Criteria Decision Making. The analytical SPE approach transforms the collected data into meaningful information through mathematical models such as Fuzzy Logic, etc. Pislaru, Herghiligiu and Robu (2019) used fuzzy logic to reduce the dimensionality of corporate sustainability performance assessment, enhancing the interpretability and transparency of the assessment model.

Furthermore, conceptual models such as composite sustainability index and sustainability ratings are widely adopted as they are designed to be user-friendly and to help stakeholders better understand, categorize, and account for the company's sustainability impact. However, composite sustainability indexes cannot be representative of the sustainability impact due to information loss in the aggregation process, trade-offs, and compromises of indicators of different factors. Dočekalová and Kocmanová (2016) addressed some weaknesses of the composite sustainability index by proposing a set of aggregate indicators with several composite sub-indicators reflective of company performance in different sustainability aspects. The addition of composite sub-indicators compensates for the information loss in the aggregation process. Similar logic can be applied to sustainability ratings that are used to demonstrate the sustainability impact of companies. The sustainability ratings were launched by financial institutes like Morgan Stanley Capital Investment (MSCI) and the London Stock Exchange Group (LSEG) to showcase the sustainability performance of companies for investment purposes (Diez-Cañamero *et al.*, 2020). It can also be used as a straight-forward communication tool for management who are not familiar with sustainability to see the performance of the company. The disclosure of sustainability ratings has beneficial effects on the company financial performance and confidence in long-term investors (Diez-Cañamero *et al.*, 2020). However, sustainability ratings are criticized for the reasons of bias, trade-offs, and the lack of standardization, transparency, and credibility of source information, solutions suggested by Windolph (2011) are shown in Table 2.2. The bias caused by financial and economic prioritization of the financial institutes can also be another concern for the application of sustainability ratings in company sustainability performance management (Diez-Cañamero *et al.*, 2020).

Table 2.2: Challenges, Causes and Possible Improvements of Sustainability Ratings (Windolph, 2011)

Rating Challenge	Cause	Possible Improvements
Lack of standardization	Complexity of corporate sustainability	• Find a common corporate sustainability

		including several perspectives, coordinate research
Lack of credibility of information	Lack of data availability	• Include non-government organizations and third parties for external verification
Bias	Financial background of ratings' users	• Sensitize ratings' users for the integrative character of corporate sustainability, open ratings for a wider audience
Trade-offs	Demand of ratings' users	
Lack of transparency	Commercial use of ratings	• Disclose methodology
Lack of independence	Intermingled business of raters	• Avoid business relations to companies • Include independent third parties

SPE provides insightful information on the strengths and weaknesses of the organization's current sustainability performance management practice and illuminates future areas of improvement. Such features of SPE greatly aid the decision-making process in sustainability performance management and optimize sustainability efforts.

2.2.3 Sustainability Reporting

Sustainability reporting, also known as corporate sustainability reporting or non-financial reporting, refers to the disclosure of a company's environmental, social, and governance (ESG) performance and impacts (Oprean-Stan *et al.*, 2020). The report includes the company's practices, goals, achievements, and challenges on the topic of ESG. Sustainability reporting aims to comprehensively communicate the sustainability performance of a company to the stakeholders including investors, customers, employees, regulators, and communities to increase the transparency of company operations. Overall, sustainability reporting concludes the findings of the sustainability performance assessment of the company in a report, usually published annually to inform the stakeholders on the sustainability impact and progress of the company and to disclose their future plans on sustainability matters.

The practice of sustainability reporting has a well-established mechanism, collecting data, data analysis, and reporting of analysis results. The reporting is done following sustainability reporting standards and frameworks published by organizations like the Global Reporting Initiative (GRI) and the Task Force on Climate-related Financial Disclosures (TCFD) for companies engaging in voluntary reporting. Stock markets around the world have included sustainability reporting as a criterion for market listing, some, like Malaysia's Bursa Malaysia, would launch their reporting standards for the publicly listed companies (PLCs) to follow. However, due to the vast number of standards available for the company, the format of sustainability reporting lacks standardization which reduces the comparability of the reports (Stolowy and Paugam, 2023; Wagenhofer, 2024). Moreover, due to sustainability reporting being a newly emerging practice, it has yet to be widely adopted by companies while many of the sustainability reports are found to be incompliant with the reporting standards (Lozano, Nummert and Ceulemans, 2016).

A survey conducted by Lozano, Nummert and Ceulemans (2016) revealed that the main reasons for the company disclosure of sustainability reports are company transparency, sustainability assessment, publicity of sustainability efforts, stakeholder engagement, enhancement of company reputation, and instigation changes. Practising sustainability reporting is also linked to better financial performance in publicly listed companies (Oncioiu *et al.*, 2020; Thayaraj and Karunarathne, 2021). Experts have suggested that sustainability reporting can profit small and medium-sized enterprises (SME) through company image enhancement and differentiation from similar companies on the market, giving the company a more competitive edge (Castilla-Polo and Guerrero-Baena, 2023). Furthermore, sustainability reporting can serve as a starting point for companies to integrate sustainability principles into operations, improve their sustainability performance, and start fundamental behavioural changes in the daily operations of the company (Lozano, Nummert and Ceulemans, 2016).

2.3 Machine Learning

Machine learning is a subset of artificial intelligence that enables computers to better their performance in certain tasks through past experiences. A machine learning model with predefined sets of parameters learns to carry out certain tasks from input training data without explicit programming (Oladipupo Ayodele, 2010). Machine learning enables systems to identify patterns, make predictions, and derive insights from complex datasets, thereby facilitating decision-making and problem-solving in diverse domains.

The process for machine learning model development is shown in Figure 2.2 below. The objective of machine learning is to develop a model that can autonomously perform tasks without human supervision. The first stage of machine learning model development is to determine the task for the model. The performance of machine learning models depends significantly on the quantity and quality of the training data (Sarker, 2021), Hence, data collection and data preparation process are crucial to the development of a machine learning model. Algorithm selection depends on the nature of the task and the characteristics of the data, different types of machine learning algorithms have their own strengths and weaknesses in handling data and task performance. Hypothesis class is the pool of models that the learning algorithm will choose from to produce our final model while the loss function quantifies the prediction error in the output of the model. The training data is fed to the learning algorithm during the model training process to produce the machine learning model. The model is assessed with novel data to examine its performance. If the model produces satisfactory performance, the model can be deployed, otherwise, it goes through model tuning to yield a better model. Model interpretability and explainability are becoming relevant due to the issue of ethics, hence, an emphasis is being put on the ease of the interpretation of model output and the decision-making process behind the output.

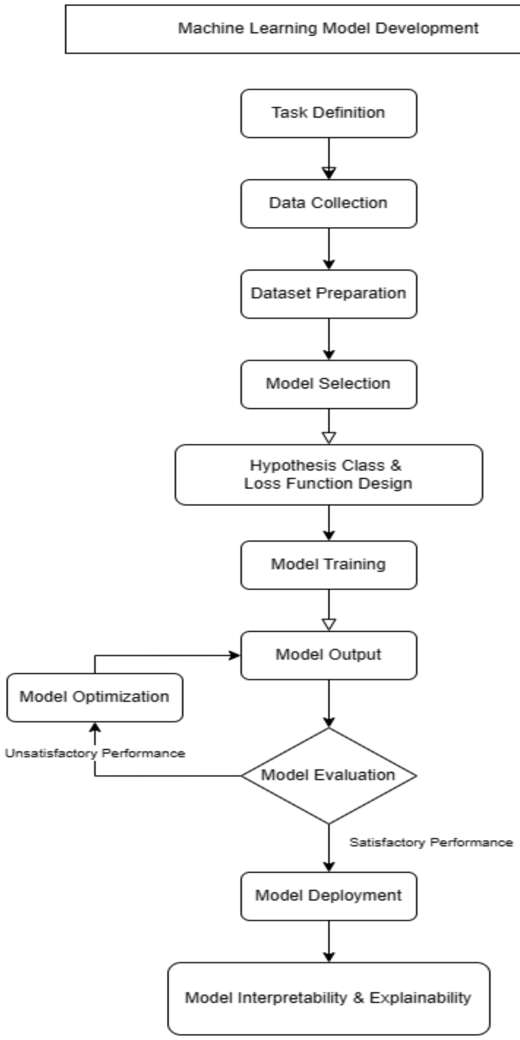


Figure 2.2: Machine Learning Model Development Process

2.3.1 Reinforcement Learning

Reinforcement learning enables systems to learn through interactions with the environment. The objective of reinforcement learning is to receive positive feedback (reward) while avoiding negative feedback (penalty) from the environment, thus, the model adjusts its output after receiving feedback from the environment. Optimization of the machine learning model happens autonomously in the process of interactions. It is widely applied in the systems of robotics, autonomous driving tasks, and recommendation algorithms to effectively increase their efficiency as these systems need to react to a highly dynamic environment (Sarker, 2021).

2.3.2 Unsupervised Learning

Unsupervised learning trained on unlabelled data to extract information from large amounts of data without human involvement, hence, it would not be affected by human bias. The output of unsupervised learning is the grouping of similar data and the detection of outliers in the dataset. This allows unsupervised learning to be used in tasks such as clustering, density estimation, dimensionality reduction, and anomaly detection (Sarker, 2021). Unsupervised learning can also discover hidden relationships in unlabelled data. Therefore, unsupervised learning is used in trends and association rule identification.

Clustering algorithms are used to agglomerate similar data points into clusters in an unlabelled dataset. This machine learning technique is useful in identifying patterns, detecting anomalies, and extracting features in data. The relationships between data points are assessed using similarity measures. Centroid-based clustering algorithms mainly use Euclidian Distance, Manhattan Distance, and Minkowski Distance as their similarity measures. The datasets are partitioned into several predetermined clusters through their value of similarity measures. Table 2.3 highlighted some examples of centroid-based clustering algorithms and their characteristics (Kumar Uppada, 2014).

Table 2.3: Centroid Based Clustering Algorithms and Characteristics adapted from Kumar Uppada (2014)

Algorithm	Sensitive to noise	Outlier	Structure-centric	Minimize intra-cluster variance
k-means	Very high	Very sensitive	Yes	No
k-medoids	Optimum	Sensitive	Yes	No
CLARA	Optimum	Kick-off to study	No	Yes

CLARANS	Very low	Deals with outliers	No	Yes
k-Harmonic means	High	Sensitive	Yes	No
Fuzzy c-means	Optimum	Kick-off to study	Yes	No

Density-based clustering algorithms automatically group data points in clusters based on density. Clusters are separated by regions of low-density data which are the outliers and noise in the datasets (Campello *et al.*, 2020). This type of algorithm outperforms centroid-based clustering algorithms in grouping datasets into irregular clusters or non-exclusive clusters. DBSCAN, GDBSCAN, and DENCLUE are some of the classic density-based clustering algorithms. Connectivity-based clustering algorithms conduct the clustering process through several iterations of combining clusters of similar features. Hierarchical clustering uses two different approaches: divisive and agglomerative clustering. Divisive clustering uses top-down approaches that separate different clusters from a big group that contains all the data points while agglomerative clustering uses a bottom-up approach that assimilates data points that were regarded as individual clusters in the beginning (Kameshwaran and Malarvizhi, 2014). Distribution-based clustering algorithms organize data points using probability distributions (Uniform, Gaussian, Inverse Gaussian), meaning the data points are clustered according to their probability of belonging in a cluster. Each cluster has a central point which is used to determine the probability of a data point falling into the cluster. A data point with a longer distance from the central point has a lower chance of being in the cluster.

Association analysis is conducted to identify hidden patterns in datasets containing large numbers of data items and their interdependency. These patterns are called association rules which explain the relationship between two data items. The association rules are identified when the features appear together in high frequency. The association patterns found in the datasets could be positive and negative. Positive association rules refer to the relationship between data items present in the datasets while negative association rules are the relationship between the present data items and the absent attributes.

2.3.3 Supervised Learning

Supervised learning machine learning builds the algorithms on the functions that represent the relationship between input-output datasets. The task of the final machine learning model is to be able to predict the output from the input data. For any given problem, the training of the supervised machine learning model is done using labelled data. Figure 2.3 illustrates the training process of supervised machine learning. Supervised learning machine learning algorithms excel in tasks such as classification and regression analysis. Although both tasks require the prediction of outcomes from input data, classification produces qualitative outcomes such as labels while regression yields quantitative results like numbers (Sarker, 2021).

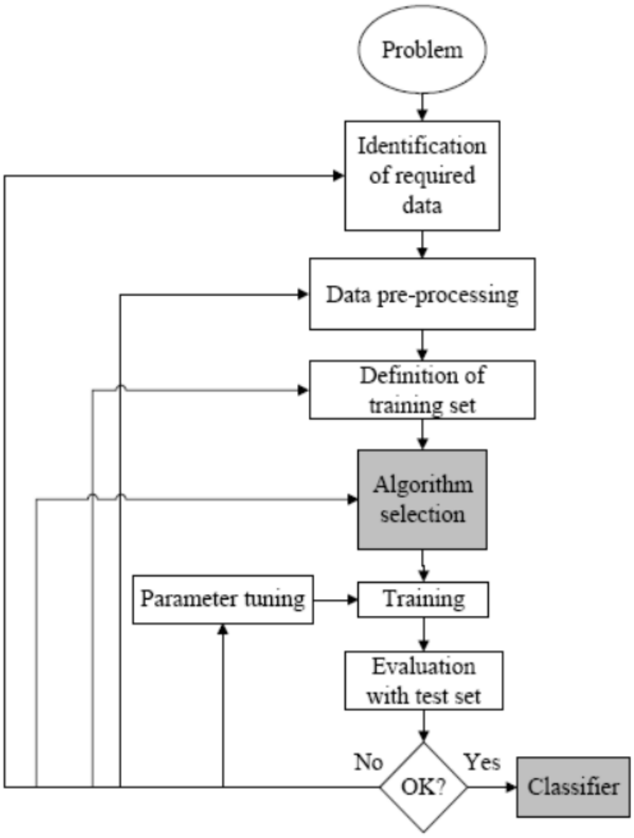


Figure 2.3: Training Process of Supervised Machine Learning Model (Sarker, 2021)

Regression models are trained with labelled datasets; hence, they fall under supervised learning. A trained regression model aims to determine the relationship between input and output of a continuous nature. Depending on the relationship between input data and output value, regression techniques are categorized into linear regression, non-linear regression, ridge regression, lasso regression, and elastic net regression. Linear regression uses linear functions to predict the targeted output while non-linear regression uses non-linear functions like polynomial and logistic functions. Ridge, lasso and elastic net regression are regularized regressions that differ by the method of regularization to reduce overfitting. Ridge regression uses L1 regularization and lasso regularization uses L2 regularization while elastic net regression uses both L1 and L2 regularization. Other than the prediction of continuous outputs, regression is used for feature selection and hyperparameter tuning to improve machine learning models.

Classification models predict a class label for input data. It aims to develop a function that can map the relationship of input and output data that would enable it to classify similar data that was not included in the training dataset. Based on the number of labels the dataset will have in the results, classification problems are categorized into three types, binary classification, multiclass classification, and multi-label classification. Binary classification problems are classification tasks that have only two class labels, often the results are “normal” or “abnormal”. An example of the binary classification problem is the classification of spam email which is tasked to determine whether the e-mails go into the spam folders. Multiclass classification problems are those with more than two class labels. The results of prediction are different categories for the data that are mutually exclusive in each category. Multi-label classification problems also have multiple class labels; however, the data can simultaneously belong to more than one category.

2.3.4 Application of Machine Learning Model in Sustainability Performance Management

There is a wide usage of machine learning models in research on sustainability performance management. A variety of white-box machine learning models and black-box machine learning models to improve the explainability of ESG ratings (Del Vitto, Marazzina and Stocco, 2023). The White-box model refers to the supervised machine learning models that use algorithms with transparent decision-making processes, hence, high explainability while black-box models are the models created with low explainability algorithms. The list of machine learning algorithms used in the study is listed below in Table 2.4.

Table 2.4: Machine Learning Algorithms Used By Del Vitto, Marazzina and Stocco (2023)

White-box model	Black-box model
Linear regression	Random forest regressor
Ridge regression	Ada boost regressor
Lasso regression	Artificial neural network
K-neighbor regressor	
Decision tree regressor	

Natural language processing is used by Fischbach et al. (2023) and Kang and Kim (2022) in text mining tasks on data from social media sourced and corporate sustainability reports. Modapothala, Issac and Jayamani (2010) conducted the analysis on the reporting variables (organizational, environmental, social, and economic performance) selected by different industries using One Way ANOVA and Multivariate Discriminant Analysis. Ni et al. (2023) utilized the Large Language Model to extract Key Indicators from text-based media to achieve the automated analysis of the Corporate Sustainability Report. Fuzzy C-means, Classification and Regression Tree (CART), and Adaptive Neuro-Fuzzy Inference System (ANFIS) were applied to assess country sustainability through a large number of indicators set (Nilashi *et al.*, 2019). Shahi, Issac and Modapothala (2012) used Naïve Bayes, Neural

Network, C4.5, and Decision Table to conduct the automated scoring of Corporate Sustainability Reports following GRI standards. Vivas et al. (2019) utilized Multiple Linear Regression (MLR), a generalized linear regression algorithm, in the development of a hybrid multi-criteria decision analysis (MCDA) model to carry out prediction tasks for the assessment of sustainability performance. Laskar (2018) used Logistic Regression to determine the effects of Corporate Social Reporting on Asian firm Value. However, the researchers highlighted the overconfidence of the logistic regression model as one of the limitations of the study. Least-squares regression, panel data regression, and logistic regression were used by Wang (2017) to study the relationship between sustainability disclosure and firm characteristics of the Taiwan 50 Index-listed companies. Logistic regression was used again by Akbulut and Kaya (2019) to the relationship between sustainability reporting, firm value, and financial leverage in the automobile industry. Chang et al. (2019) used multiple regression models to investigate the factors affecting sustainability reporting quality in the financial sector and the impact of the equator principle on moderation. Sariyer and Taşkın (2022) used Kmeans++ clustering algorithm to conduct cluster analysis on the ESG score of companies included on the Borsa Istanbul (BIST) Sustainability Index for the identification of the relationship between ESG score and firm financial and ESG performance. Kanmani et al. (2020) constructed a framework for the assessment of the environmental sustainability of countries utilizing Self-Organized Maps, a clustering technique, to form clusters of countries with similar environmental performance and to compare the countries in each cluster in different timeframes. Galindo, Vaz and de Noronha (2015) applied hierarchical clustering Ward's methods to form clusters of companies of alike sustainability profiles to understand their contribution to their country's sustainability performance. Kmeans algorithm was used by Li and Rockinger (2024) to investigate the changes in bank sustainability reporting focus over the years.

2.4 Decision Support System (DSS)

A Decision Support System (DSS) is an interactive computer-based system tasked with enhancing decision-making in companies and organizations (Jain and Raju, 2016).

A DSS enhances the decision-making process in the form of providing data-driven informational insights to decision-makers in the organizations. However, the DSS is designed to only support managerial decisions not to replace the role of management in company decision-making (Deogun, 1988). The application of DSS has been proven to produce an improved outcome. The implementation of DSS also leads to beneficial outcomes including cost reduction, improvement in efficiency and productivity within organizations, and better organizational control (Di Matteo *et al.*, 2021). For example, Bright et al. (2012) studied the use of DSS in the clinical field and concluded that the use of DSS has led to general improvement of healthcare service processes such as reduction in morbidity, more appropriate order of treatment method, etc. However, the impact of DSS implementation on clinical workload, efficiency, and economic outcomes remains undefined due to the lack of evidence.

2.4.1 Types of Decision Support System (DSS)

There are five (5) types of decision support system (DSS), namely (1) Data-driven DSS, (2) Model-driven DSS, (3) Document-driven DSS, (4) Communication-driven DSS, (5) Knowledge-driven DSS (Hasan *et al.*, 2017). Table 2.5 includes the description for each type of DSS. The DSS is categorized according to their source of information. Data-driven DSS supports decision-making processes through data processing techniques. This type of DSS consumes vast amounts of data stored usually in a data warehouse system. Model-driven DSS employs mathematical and analytical models to assist the decision-makers with the analysis of a situation (Power, 2002). Model-driven DSS is suitable for budgeting, forecasting, and planning activities. Documentation is a process that document-driven DSS can aid (Fernando and Baldelovar, 2022). Communication-driven DSS supports company operations by improving the connection between management and employees to enhance efficiency in business conduct (Fernando and Baldelovar, 2022). Knowledge-based DSS is constructed with the vast bodies of knowledge of business management and experts related to the problem in question to recommend measures to decision-makers (Power, 2002).

Table 2.5: Type of Decision Support System (DSS)

Type	**Description**
Data-driven DSS	• Conduct data analysis on large database • Maintain ease of access to data
Model-driven DSS	• Perform various modelling tasks that are problem specific • No need for large database as the required data & parameters are given by users
Document-driven DSS	• Provide relevant documents and websites relevant to the problem
Communication-driven DSS	• Improve communication within the organization with the use of instant email, message, and video chat
Knowledge-driven DSS	• Suggest solutions to problems

2.4.2 Components of Decision Support System (DSS)

A Decision Support System (DSS) consists of four (4) components, which are (1) data management subsystem, (2) model base management subsystem, (3) user interface subsystem and (4) users.

The data management subsystem contains the database necessary for the decision-making process. The database contains collections of data relevant to the decision-making process in a structured form that is usable by the computers. The data management subsystem uses a database management system (DBMS) that allows users to make modifications to the database and access the data within. The database management system is often computer software such as Oracle DBMS, Access and SQL Server from Microsoft, DB2 from IBM, and the Open-source DBMS MySQL (Jain and Raju, 2016).

Model base management subsystem contains the computational models required for decision-making analysis using data from the data management subsystem. Depending on the type of DSS, this component contains different applications to fulfil the needs of the users. Commonly applied models are the forecasting models, optimization models, and simulation models. The forecasting models use variables to predict future situations to help companies make the best choices with their company strategy. The availability, and accuracy of data are crucial to the performance of the forecasting models (Power, 2002). Optimization models are often used for activities like project planning and resource allocation to reduce operational costs and increase the profitability of companies (Power, 2002). The simulation models are utilized to analyse the outcome of different situations and their benefits and risks (Power, 2002).

The user interface subsystem is the platform of the DSS that communicates and interacts with the users. The user interface includes an input and an output system. The input system enables the users to access and modify the database and the model while the output system displays the result of the modelling and data for the users' reference (Jain and Raju, 2016). The user interface subsystem also utilizes graphics and tables to display the results for easy digestion of information on the users' side (Power, 2002).

The last component of DSS is the users who interact with all the other components of the DSS through the user interface. The targeted users of the DSS framework in this study are the management of companies who wish to continuously improve the long-term sustainability performance of their company and sustainability consultants as an assessment and progress tracking tool. The DSS provides support in tracking their progress in the sustainability journey and realignment of sustainability strategies with company goals when companies fail to make advancements on sustainability issues.

2.4.3 Application of Decision Support System (DSS)

Decision support systems (DSS) have been applied in the sustainability field to optimize problems. Zarte, Pechmann and Nunes (2019) conducted a systematic literature review on the utilization of DSS in sustainability manufacturing. Their findings concluded that the existing studies focused on the integration of all dimensions of sustainability (environment, society, and governance) and company strategy planning, but it did not impact the operation activities as they were mostly related to single sustainability dimensions. The authors pointed out that the selection of sustainability indicators for the DSS was mainly done subjectively. Such practice could lead to negligence in certain aspects of sustainability in the decision-making process. The situation was reflected by the lack of focus on the social dimension in the existing literature (Zarte, Pechmann and Nunes, 2019). Moreover, their findings revealed that within the environmental dimension, the emphasis was put on emission reduction. They also found that small-medium enterprises are not given as much attention in existing studies.

Shin et al. (2017) developed a DSS for the optimization of manufacturing processes to improve firm sustainability performance. They developed a DSS prototype that is reusable and adaptable in different situations in the manufacturing process. They applied their prototype in two scenarios, resource allocation and parameter selection. Their case studies have shown that their DSS prototype is effective in reducing energy consumption in the manufacturing process.

Chalmeta and Ferrer Estevez (2023) used a balanced scorecard to assist the implementation of sustainability practices in university settings. Their methodology for the sustainability balanced scorecard covered all aspects of sustainability when developing the sustainability strategy of the universities in studies. The authors found that the application of the proposed sustainability scorecard has led to positive changes across ESG. Some example changes were the improved transparency and accountability seen in the governing bodies, mitigation of environmental impacts, and the allocation of institutional resources to create value for the common good. This study showed that traditionally business-oriented practices could be applied to sustainability issues with some modifications integrating sustainability principles and stakeholder expectations.

Mattiussi, Rosano and Simeoni (2014) combined life cycle inventory, impact assessment, multi-objectives, and multi-attribute decision-making modelling in a sustainability DSS designed for the development of energy plants. The main factors in the decision-making process were economic (Net Present Value) and social (Human Health Impact Reduction) oriented, but the social criteria are affected by environmental factors which are the air pollutants emissions from the plants. Mathematical models were used in this decision support system to develop the optimal solution to the problem, in this study, it was the design of an energy plant. The decision support system achieved its objective for their case study in the Kwinana Industrial Area, but the process was labour-intensive as the calculations were performed manually using Microsoft Excel.

Juan, Gao and Wang (2010) developed a DSS that improves the energy sustainability of office buildings. The DSS utilized an A* graph search algorithm coupled with general algorithms in the development of optimal strategies. The coupling of the two algorithms enhanced the calculation efficiency of the system as they compensated for each other's weaknesses when searching for a solution. The criteria of the decision-making process were adopted by the US Green Building Council to ensure that the optimal solution was developed with the principle of sustainability. The authors implemented the DSS in a renovation project in Taiwan. Three scenarios, energy demand of building without renovation (Scenario I), with renovation that partially adopted the recommended solutions (Scenario II), and with renovation that fully implemented the suggested solutions (Scenario III), were compared to assess the effectiveness of the DSS, the data is shown in Figure 2.4. The result showed that the implementation of the solutions suggested by the DSS, even with partial adoption, significantly reduced the energy consumption of the office building.

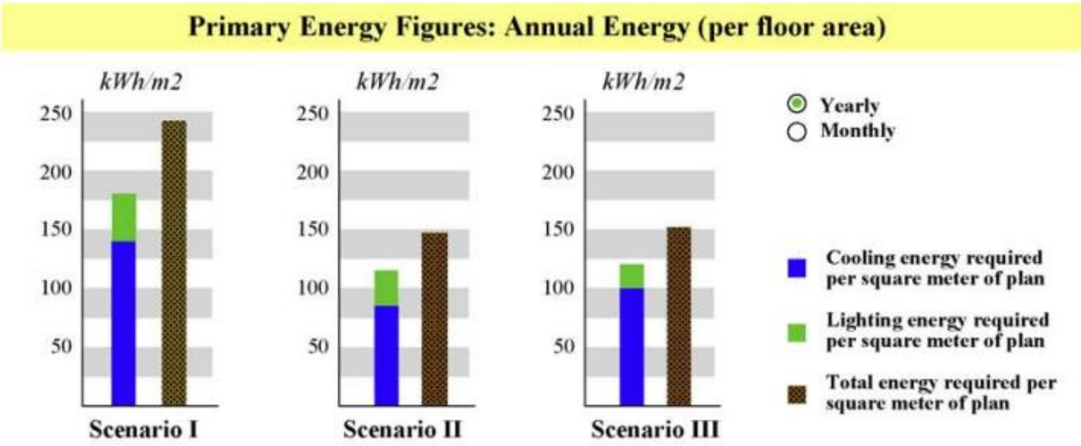


Figure 2.4: Energy Consumption of the Office Building in Different Scenarios (Juan, Gao and Wang, 2010)

CHAPTER 3

METHODOLOGY

3.1 Flow of Study

The overall flow of this study was summarized in Figure 3.1. Details of each process are elaborated in subsequent sections.

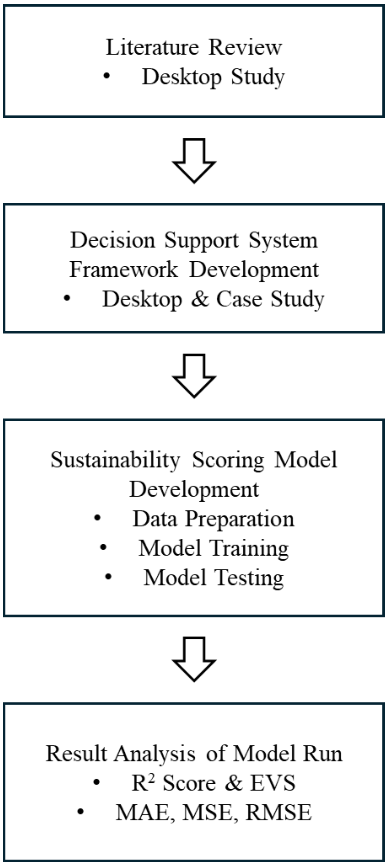


Figure 3.1: Flow of Study

3.2 Desktop Study for Subject Overview

Desktop studies were conducted on the topic of Sustainability Performance Management, Machine Learning, and Decision Support System (DSS) to gain a comprehensive understanding of these topics before the system development process.

This research used secondary literature data in the process of conceptual framework development for the DSS. These secondary literature data include peer-reviewed journal papers and review articles on the subject of the DSS, machine learning and its application, and corporate sustainability performance management. They were collected from academic research databases such as Elsevier Scopus and IEEE Explore through academic search engines Google Scholar using keywords like “Sustainability Performance Management”, “Machine Learning” and “Decision Support System”. The review of these secondary literature data enabled a comprehensive understanding of subjects related to this research project in a relatively short amount of time. The journal papers and research articles provided an overview of the methods, principles, and current practices of corporate sustainability performance management. The subject of machine learning was understood through studying articles and research papers that provided information on the machine learning types, abilities, strengths, and weaknesses of machine learning models and applications. Studies on DSS illuminated the type and components of DSS, their development, and usages in real-life cases. These literature reviews laid the foundation for this research project and the development of the conceptual framework.

3.3 Development of Decision Support System (DSS) Framework

The framework for the decision support system (DSS) was developed through case studies of several research papers on DSS development and applications of DSS in different industries. The process for the DSS development was adapted from Di Matteo et al. (2021) and shown below in Figure 3.2.

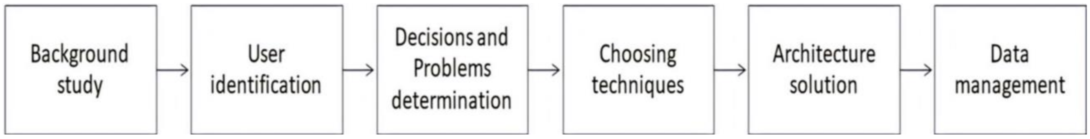


Figure 3.2: Development of Decision Support System (DSS)

The flow of the study followed the figure, except the scope of this study was limited to Architecture Solutions only. Background study and decision problem determination were done previously through literature reviews. The users were identified to be the decision makers for sustainability strategies in companies.

The case studies gave a basic structure of DSS. There are mainly three components in a DSS, they are the data management system for the storage and handling of data necessary for the decision-making process, the model management system which holds the computation model that enables the functioning of the system, knowledge management system which comprise of information collected from all sources including input from experts and excerpts of sustainability strategy from company sustainability reports with excellent sustainability score and the user interface that allows interactions between users and the systems.

The research of Di Matteo et al. (2021) applied their DSS prototype for sustainable cultural asset management in a museum setting. This research provides a general structure for the development of DSSs through comprehensive explanations of the selection of framework components. Baffo et al. (2023) developed a DSS that assesses the sustainability of investment projects using Sustainable Development Goal (SDG) related criteria. Although the DSS in the study is not equipped with machine learning elements, it demonstrated the evaluation of sustainability performance using indicators in a DSS. Markopoulos, Al Katheeri and Al Qayed (2023) integrated the Refinitiv ESG Score database in the DSS framework they developed for the calculation of the ESG score of SMEs. Angelakoglou and Gaidajis (2020) designed a composite indicator to assess the environmental sustainability of a mining industrial facility. They constructed an assessment framework for environmental sustainability shown in

Figure 3.3. In this study, stage 1 of the assessment framework proposed by Angelakoglou and Gaidajis (2020) is used in the construction of the indicator sets while stage 2 of the assessment framework will be carried out using a machine learning model for the computation of sustainability scores.

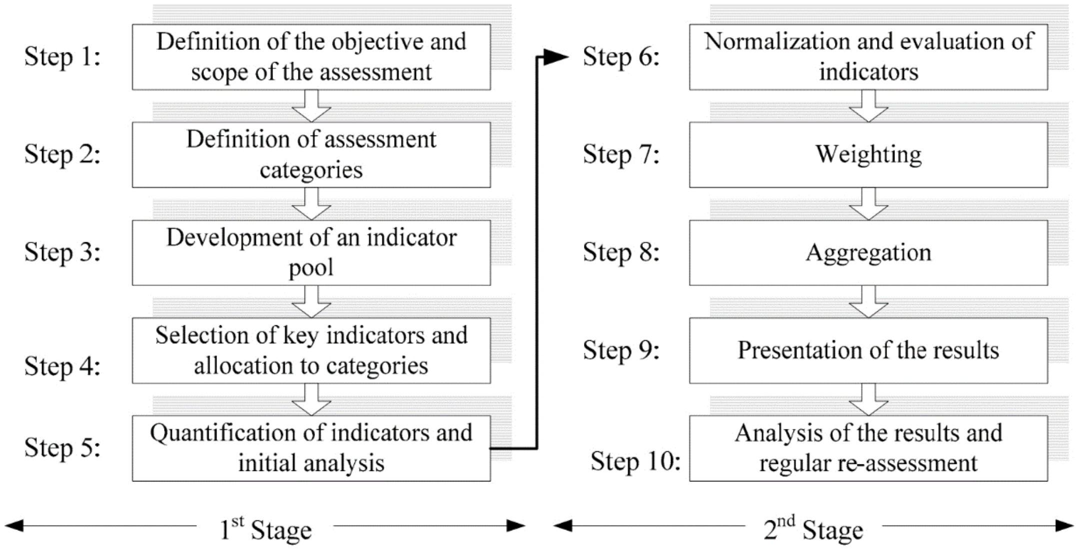


Figure 3.3: Environmental Sustainability Assessment Framework by Angelakoglou and Gaidajis (2020)

Dočekalová and Kocmanová (2016) demonstrated the mechanism behind the aggregation of the composite sustainability index. The information provided by these case studies is integral in the process of conceptual framework in this study.

Further desktop studies were conducted to have an in-depth understanding of each component. The data management system handles the data collection and management in the DSS. The data management system is crucial to the DSS as machine learning models rely heavily on large amounts of quality data to produce excellent results. The flow of data in the system is mapped out to determine the necessary tools needed for the system. Techniques and tools to collect large amounts of data for the system were researched to achieve automation of sustainability data collection from the internet, including web scraping and PDF parsing techniques. The model management system is the components that perform the main tasks of the system. Literature on machine learning applications was read to determine the models used for each function in the DSS. The model selection was done based on the performance of the models from numerous research projects published in journal papers. The knowledge management system keeps a database of information that can assist decision-making to solve the problems. The source of the information needs to be identified. The DSS also needs to include tools to retrieve and store the information in the database. The user interface is a platform that allows interaction between users and the system. The user interface needs to be able to connect with the data management and the model management system so that the user can input reports into the system. Visualization using graphs and charts is important for users to easily understand the data and results of model calculations which is integral to the decision-making process.

3.4 Development of Sustainability Scoring Model

3.4.1 Data Preparation

ESG scores for 15 electrical and electronic companies were extracted from the LSEG ESG Score website (formally Refinitiv) and recorded in an Excel database. The LSEG ESG Score Database was used in this study as it was an easily accessible option with scoring categories that can be traced back to the sustainability indicator set. The sampled companies were selected randomly and consisted of Malaysian and international companies of different sizes. The information extracted from the website was the ESG scores and the year that the data was collected. The scoring categories in the LSEG Scoring System are shown in Figure 3.4. The company sustainability reports of those companies were collected from their company website according to the year of data collection. The environmental data were extracted from the sustainability reports according to the indicators listed in the GRI Standards. The indicators in consideration were GRI 302, GRI 303, GRI 305, and GRI 306. Three environmental indicators were excluded, namely GRI 301 Material, 304 Biodiversity, and 308 Supplier Environmental Assessment, as these data points were absent in the sustainability reports of many companies. The main reason for the exclusion was to reduce the presence of missing data because missing data points could significantly affect the performance of the prediction model negatively. The extracted data consisted of data with different units, hence, unit conversion was performed to ensure that data were all in unison and avoid disrupting the result of prediction.

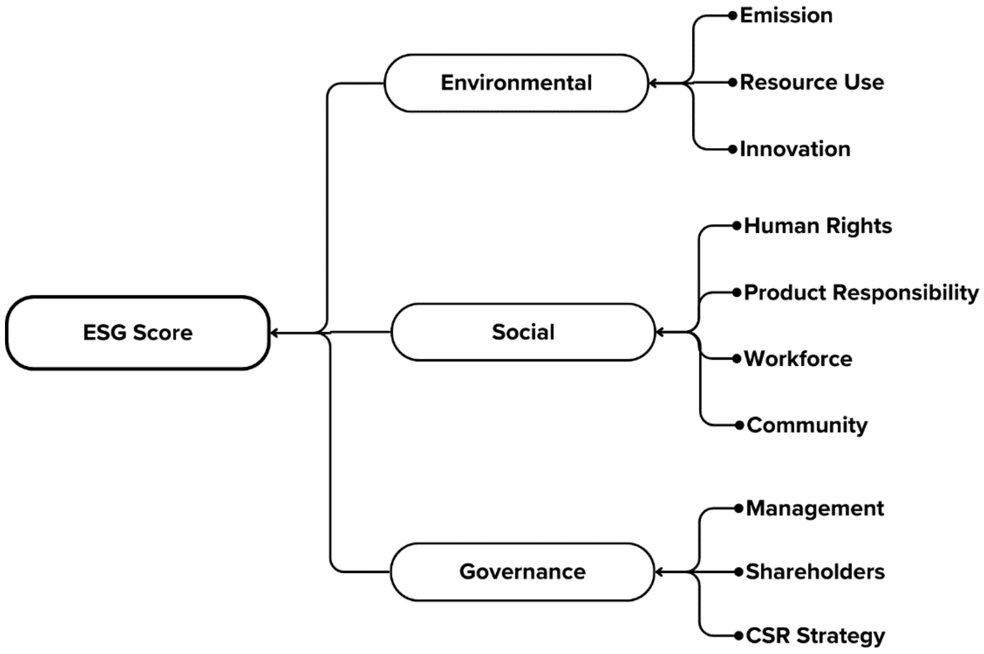


Figure 3.4: Structure of LSEG ESG Score

The datasets were separated into three samples with the data sizes of 5, 10, and 15. As the size of the datasets of this experiment was minute, the regression model was not expected to produce mature ESG score predictions. Hence, applying different sizes of data samples allowed observation of how data sizes would impact the performance of the random forest regressor model on ESG score prediction from the data extracted from the sustainability reports. The data were split into 60% training data and 40% testing data with the training_testing_split function from the Scikit Learn library.

3.4.2 Model Training

The training of the machine learning was conducted on Jupyter Notebook (https://jupyter.org/) using the Python programming language. The Python libraries such as Pandas, Numpy, Matlibplot, and Sickit-learn were used for different purposes. The Pandas library provided data structures and functions to support data manipulation and analysis operations including data cleaning and exploration. Numpy is a Python package for scientific computing that provides arrays, matrices, and mathematical functions that can handle multi-dimensional arrays like the datasets used in machine learning. Matplolib handles visualization of the data and prediction for an intuitive understanding of data distribution and model performance. Matplotlib can generate various types of charts such as line and scatter plots and bar charts for the display of data and results. All machine learning related tools and functions were called from Sickit-learn, a machine learning library in Python. Machine learning models were called from the Sickit-learn library during training. The library also supplied data splitting functions that separate the datasets into training and testing data. Model performance assessment and validation was conducted using tools and functions from Sickit-learn as well.

Random Forest Regressor was used for the scoring model. It is a machine learning of supervised learning style that can be used for regression tasks. It is an ensemble learning model that produces prediction by combining the output of multiple decision trees in the regressor. The aggregation of multiple outputs from different decision trees allows the random forest regressor to have the benefit of accurate and stable prediction. Random forest regressor also can quantify feature importance which is useful in determining indicators influential to sustainability performance.

The training of a random forest regressor has three steps: bootstrap sampling, decision tree building and prediction generation. Bootstrap sampling is the random selection of data points to create multiple subsets of the training data. A sample data can coexist in multiple sample subsets as it is randomly selected. This method is called bootstrapping. A decision tree is constructed for each sample subset, randomly selected some features to create split nodes in the tree as this can reduce the correlation in the decision trees. Therefore, each tree uses different sets of features to form the split node. The prediction made by each decision tree is aggregated to produce the final prediction of the random forest regressor.

The N-estimators, and the random state value are some of the parameters that can affect the performance of the regression model. The N-estimators refers to the number of trees in a random forest regressor while the random state value deals with the randomness of the data sampling, setting it to a fixed integer ensures reproducibility of the model training results which is important for comparable and consistent experiment. A higher number of trees in the random forest regressor can improve the performance of the model.

The N-estimators of the random forest regressor used in the experiment were set to 30. The random state value was assigned to the integer of 30 to ensure the reproducibility of the experiment.

3.4.3 Model Performance Assessment

The task of the regression model was to successfully map out a formula that could explain the relationship between the independent variables and dependent variables or the input and output data. The performance of the regression model was evaluated with various parameters. This research used the Coefficient of Determination (R^2 Score), the Explained Variance Score (EVS), the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE). These parameters are explained below in detail.

The Coefficient of Determination (R^2 Score), often referred to as R^2, indicates how well the regression model fits the actual datasets. It explained the variance of the relationship between the independent and dependent variables. The value of the R^2

Score starts from 0 to 1. The value of 0 means that the regression model cannot explain the variance and produce inaccurate prediction while the value of 1 shows that the model can explain the variance in the input and output data completely and predict the actual output data. The R^2 Score indicates the ability of the model to correctly predict the output data. The R^2 Score is given by

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(Y_i - P_i)^2}{\sum_{i=0}^{n}(Y_i - \bar{Y})^2} \qquad (3.1)$$

The Y_i is the actual value of the dependent variable while the P_i is the prediction made by the model. $\bar{Y}$ is taken by the means of the actual dependent variables.

The Explained Variance Score (EVS) is similar to the R^2 score, but it measures how well the model explains the variance in the dependent variables from the independent variables in the dataset used in the model development. The range of EVS also falls between 0 and 1. A low EVS represents that the model failed to explain the variability of the target variable while an EVS approaching 1 says the opposite. The formula of EVS is

$$EVS = 1 - \frac{var(y_i - p_i)}{var(y_i)} \qquad (3.2)$$

According to Scikit Learn, the main difference between R^2 Score and EVS is that EVS does not take the systematic offset of the prediction into account which could lead to the wrong conclusion on the performance of the model.

The Mean Absolute Error (MAE) represents the average magnitude of the prediction error in absolute value. Hence, it does not consider the sign of the errors between prediction and actual data, the error can be neutralized by differences with opposite signs which can lead to misjudgements of the model’s performance. The MAE is calculated with the formula below.

$$MAE = \frac{\sum_{i=0}^{n} |Y_i - P_i|}{n} \tag{3.3}$$

"n" is the total number of datasets in the sample. The aim is to have a low MAE value for the regression model as an indicator of its accurate prediction.

Similar to MAE, the Mean Squared Error (MSE) measures the prediction accuracy of the regression model. It is given by the average of the squared difference between the actual and the predicted value, so the MSE is not affected by the sign of the difference The value of the MSE is indicative of how close the predictions are to the actual data, larger value of MSE shows that the model produces predictions that are further away from the actual data. However, the value of MSE is easily affected by large errors in the predictions, even if there is only very few of the large errors. MSE is given by the formula written below.

$$MSE = \frac{\sum_{i=0}^{n} (Y_i - P_i)^2}{n} \tag{3.4}$$

The Root Mean Squared Error (RMSE) is taken by the root square of MSE. RMSE is not affected by the sign of the prediction error and still considers the existence of a large error in the predictions without giving the larger error too much weight. It represents an estimation of the prediction error standard deviation. The value of RMSE tells on average how much the prediction will differ from the actual value. The formula of RMSE is given below.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n} (Y_i - P_i)^2}{n}} \tag{3.5}$$

3.4.4 Model Testing

The model testing was conducted through the prediction of sustainability scores in the testing datasets of the samples. The testing data was selected randomly by the training_testing_split function. The comparison of predicted and actual values of the ESG scores and the analysis of the evaluation metrics were performed to observe the performance of the regression model.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 The development of the Decision Support System (DSS) Framework

The framework of the decision support system (DSS) is displayed in Figure 4.1. The details of each component are discussed in subsequent sections.

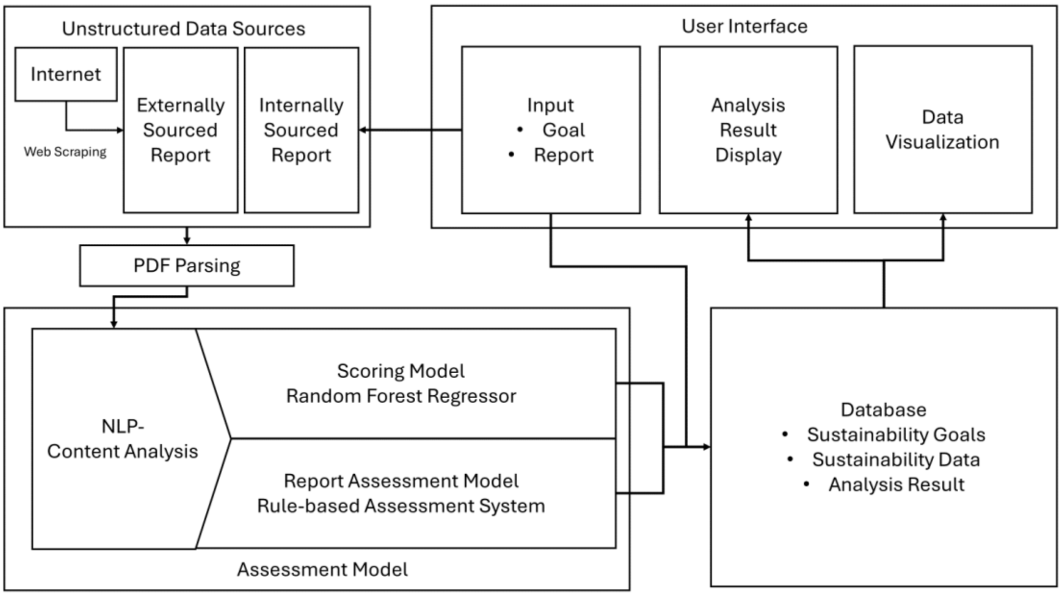


Figure 4.1: The developed Decision Support System (DSS) Framework

4.1.1 Data Management System

The first component of the decision support system (DSS) is a data management system that consists of three subsystems: data collection, extraction, storage, and processing systems.

The data collection system utilizes search query commands and web scraping tools to enable the automation of company sustainability report collection. The search query command searches the internet using keywords such as “sustainability report”, “Electrical and Electronics Company”, “ESG Data”, etc. to collect the URL of target data into a list. The web scraping tool is used to extract reports in Portable Document Format (PDF) and ESG data from the URL.

The PDF parsing is used to extract information from the company sustainability reports. Currently, most sustainability reports are only available in the form of PDF documents which are not readable to computers. PDF parsing allows computers to access the information in sustainability reports through the extraction of text, tables, and images from the PDF files and storing them in a structured database. PDF parsing techniques such as image extraction detect image objects like charts in the PDF file and save them in standard image formats like the Joint Photographic Experts Group (JPEG) and the Portable Network Graphics (PNG). Tables that are used extensively in sustainability reports can also be extracted using PDF parsing techniques. The PDF parser can identify the table structure in the PDF file and convert it into Comma Separated Values (CSV) files and Excel formats that are accessible to computers. There are several Python libraries that can perform PDF parsing, including PyPDF2, textract, etc.

Data processing is conducted using Natural Language Processing (NLP) to understand the content of the extracted data and categorize the data under the three pillars of ESG. NLP will also identify the reporting standards from the content of the report. Charts are widely used in the reports to display company ESG data, hence, computer vision is used to interpret and extract data points from the images. The system can be adapted to any reporting standards used by the companies. For the assessment of the sustainability report following the GRI standards, the data extracted will be categorized according to the indicators in the universal and topic standards. The sustainability data extracted are stored in the database.

The structure of the database is demonstrated in Figure 4.2. Each box represents a table in the database. The data included in the database are information on the company, its sustainability goals, information on the sustainability reports, the ESG data extracted from the sustainability reports, and recommendations on sustainability strategy improvement.

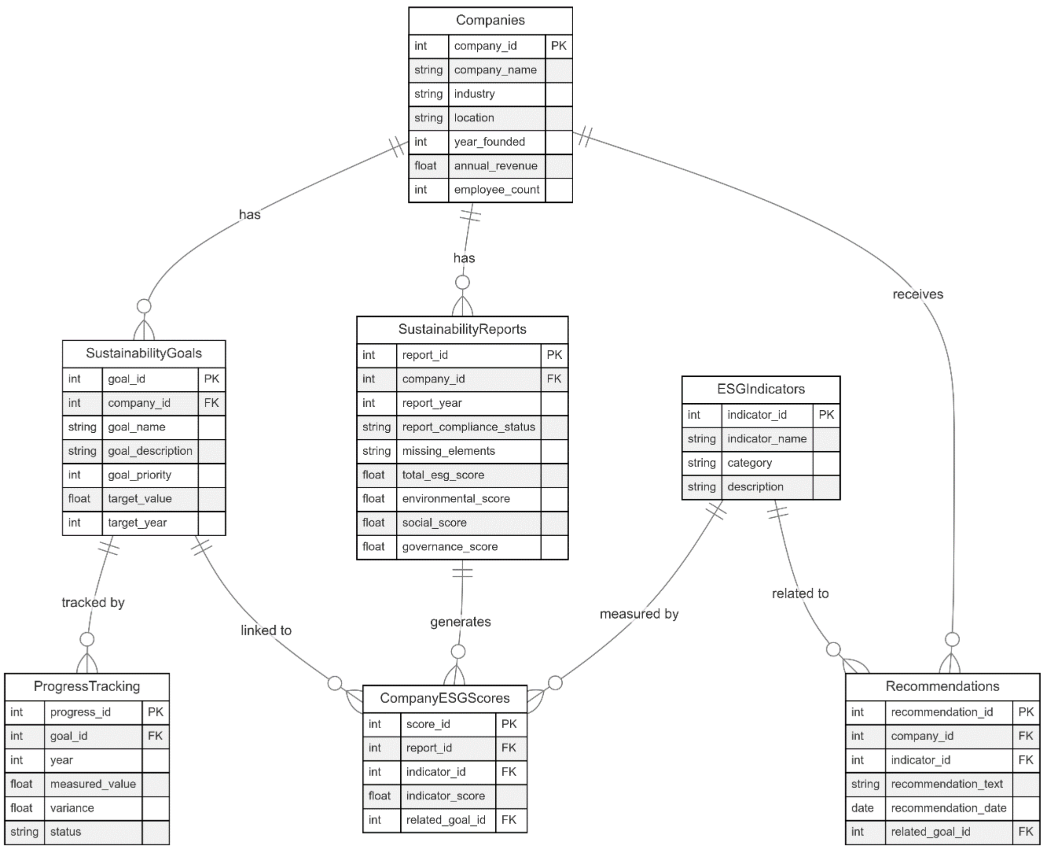


Figure 4.2: The Structure of ESG Database

The information of the company and sustainability goals are input by the users through the user interface. The data for the sustainability report and the ESG indicators are extracted through the data collection system. The data for progress tracking, company ESG score, and recommendations are generated by the DSS.

4.1.2 Model Management System

There are three computational models in the DSS framework, each for a different purpose in sustainability performance assessment and management. The decision problems of this DSS are 1) to conduct a compliance assessment of the sustainability report to its reporting standards, 2) to perform analysis on corporate sustainability performance and 3) to recommend sustainability strategy improvement.

Compliance with sustainability reporting standards is rewarding on company financial performance through the increase of firm value (Moses, 2022; Sreepriya, Suprabha and Prasad, 2023). However, assessing the level of standard compliance requires some levels of expertise and familiarity with the sustainability reporting standards which is not easily accessible to most companies in Malaysia as sustainability reporting is still in its infancy in the country. The proposed solution to the first problem of standard compliance assessment is a system that utilizes both machine learning and rules to identify the required reporting matters in the report. This combination was proposed by Hamdani et al. (2021) to automate the assessment of General Data Protection Regulation (GDPR) compliance of company data privacy policies in Europe. Although the classification task was not targeted toward sustainability reporting standards, they share some common attributes in terms of the requirement of mandatory information disclosure and having text-based data sources that are not directly machine-readable. They used NLP models as the text classifier that assigns categories to the text segments to assist the rule-based approach to the checking of compliance requirements of GDPR. Natural language processing models like transformer-based language models pre-trained on databases of the niche languages used in the reporting of ESG matters can be applied to classify the text sections into their respective categories. Transformers differentiate themselves from other language models through their ability to understand words in the context of their usage (Brugger *et al.*, 2023). Text-to-Text Transfer Transformer (T5 Transformer) demonstrated high performance on text classification tasks (Hamdani *et al.*, 2021). Webersinke et al. (2021) had pretrained the climateBERT on a large database of climate-related excerpts which improved its performance on tasks such as text classification and sentiment analysis. ClimateBERT is a transformer-based language model that is capable of conducting text classification tasks for climate-related texts. Research conducted by Brugger et al. (2023) focused on the classification of text from the social pillar in the reports. Their sentence transformer text classifier demonstrated promising result in text classification of text related to human rights in the constraint of limited database. However, they concluded that due to the limitation of text parsing technology and the difficulties of extracting information out of non-textual content of the sustainability report i.e. tables and images which are used frequently when presenting ESG data.

After the text fragments have been categorized, they can enter the process of automatic compliance checking with the requirement rules of the standards. This study uses the reporting requirements of the GRI standards for sustainability reporting as an example. The rules of reporting according to GRI standards are listed here:

i) Disclose all disclosures in GRI 2: General Disclosure 2012
ii) Disclose Materiality Assessment Process Using GRI 3: Material Topics 2021
iii) Disclose Material Topics
iv) Disclose Non-Reporting Disclosures Under Material Topics and Reasons for Omissions of Disclosures Items
v) Publish GRI Index
vi) Produce Statement of Use

GRI 2: General Disclosure 2012 falls under universal standards that are applicable to all industries. The first two sections of the company require the company to provide a general overview of company structure, operation, and details concerning company sustainability reporting practice. The third section focuses on the governance body and policy of the company. The last two sections report on the company's sustainability strategy development process and stakeholder engagement adopted by the company. The company is allowed to exclude disclosure items from GRI 2 with permitted reasons justifying the exclusion except for Disclosure 2-1: Organizational Details, Disclosure 2-2: Entities included in the organization's sustainability reporting, Disclosure 2-3: Reporting Period, frequency and contact point, Disclosure 2-4: Restatements of information, Disclosure 2-5: External assurance. These five disclosures are mandatory reporting items that should be included in the report. The GRI allows the four reasons in Table 4.1 with explanations for the exclusion of disclosures.

Table 4.1: Permitted Reasons for Disclosure Omissions (Global Reporting Initiative, 2023)

Reasons for omission	**Required explanation**
Not applicable	Explain why the disclosure or the requirement is considered not applicable.
Legal prohibitions	Describe the specific legal prohibitions.
Confidentiality constraints	Describe the specific confidentiality constraints.
Information unavailable/ incomplete	Specify which information is unavailable or incomplete, specify which part is missing (e.g., specify the entities for which the information is missing). Explain why the required information is unavailable or incomplete. Describe the steps being taken and the expected time frame to obtain the information.

The company shall include the material topics and their materiality assessment process in their report using GRI 3: Material Topics 2021. There are three sections in GRI 3. The first two sections must be reported while the third section on material topics management can be omitted with reasons and explanation included in Table 4.1. The GRI publishes sectoral standards for several industries (Oil and Gas, Coal, Agriculture, Aquaculture, and Fishing, Mining). These sectoral standards provide a list of potential topics for companies in their respective industries. Companies in these industries are required to report on the relevant disclosures in the sectoral standard of their own industry. The company adopting the sectoral standards shall include explanations of the reasons for “not applicable” on the omitted topics. Companies in other industries will have to conduct materiality assessments on the topic standards published and report on the topic material to their operation and impacts. The company can cite one of the four reasons in Table 4.1 for the exclusion of disclosures under the material topic standards and support with explanations.

The GRI requires all companies adopting GRI standards to include a GRI index in their report. The GRI index contains the statement of use, all the topic standards and disclosures reported by the company, the reasons for omissions, and supporting explanations for topic standards and disclosures. The company using sectoral standards needs to include the GRI Sector Standard reference numbers. The location of the reported disclosure in the report shall also be included in the index. In the event that the GRI index is published separately from the sustainability report, a link shall be provided in the report for the location of the index.

The rules are incorporated into the report standard compliance assessment system. The decision-making process for the system is shown in Figure 4.3 to Figure 4.6 for GRI standard compliance assessment. The categorized text segments go through the compliance verification process and come to the conclusion of whether the report complies with the requirements of the reporting standards. The system checks the inclusion of reporting on GRI 2, materiality assessment, reporting of material topics, and the GRI index. At the end of the checking process, if the report complies with all requirements, the result of “In Compliance with GRI Standards” will be concluded, recorded, and sent to the user interface for display. If the report is not in compliance with the standards, the list of the missing elements and the conclusion will be recorded and sent to the user interface.

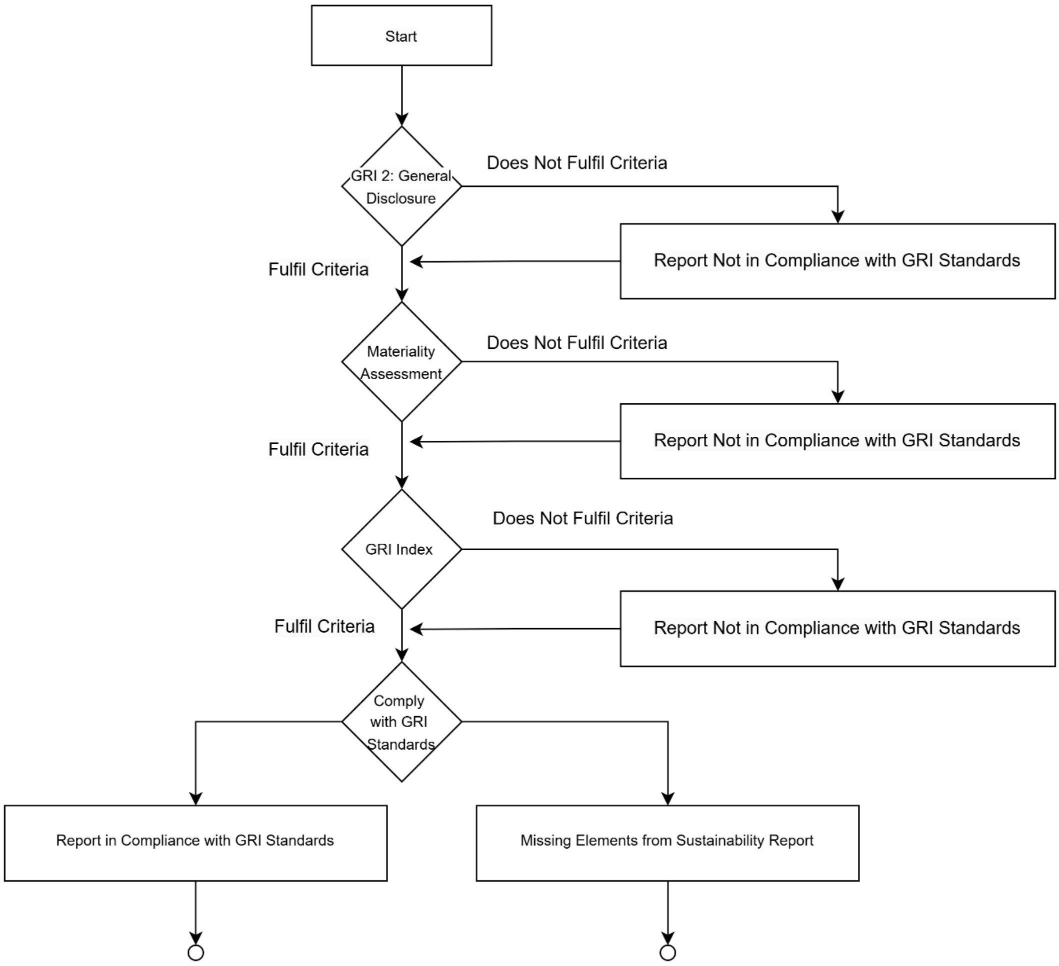


Figure 4.3: Overall Decision-making Process for Rule-based Report Standard Compliance Assessment System

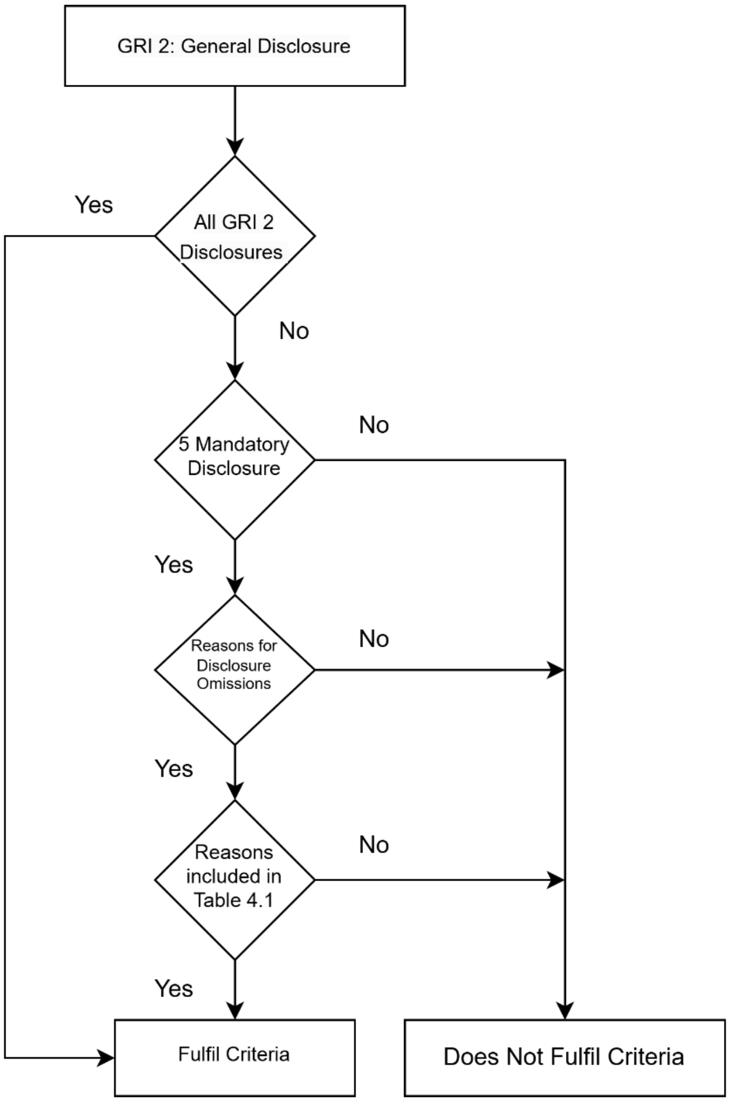


Figure 4.4: Decision-making Process for Segments of GRI 2 General Disclosure

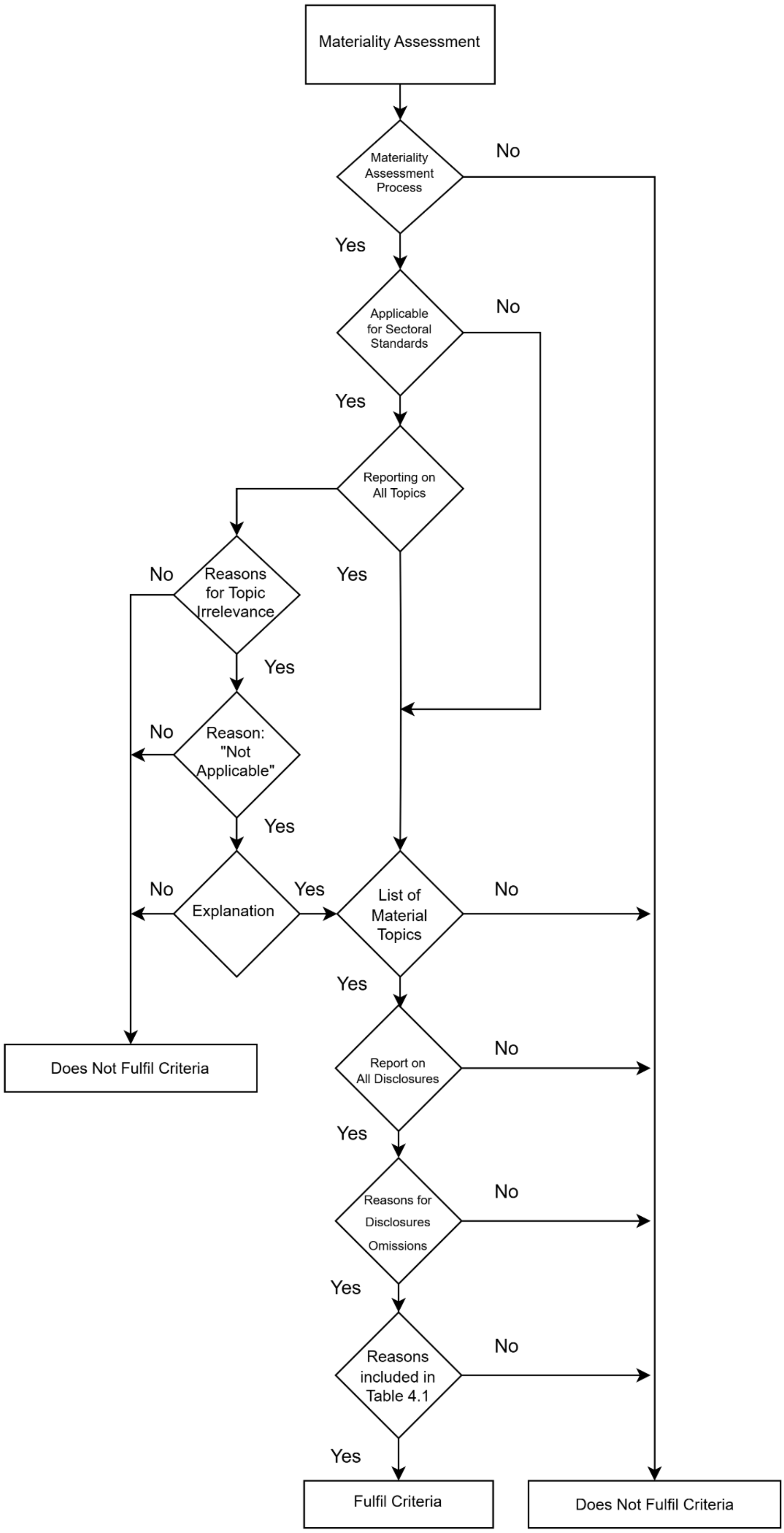


Figure 4.5: Decision-making Process for Segments of Materiality Assessment

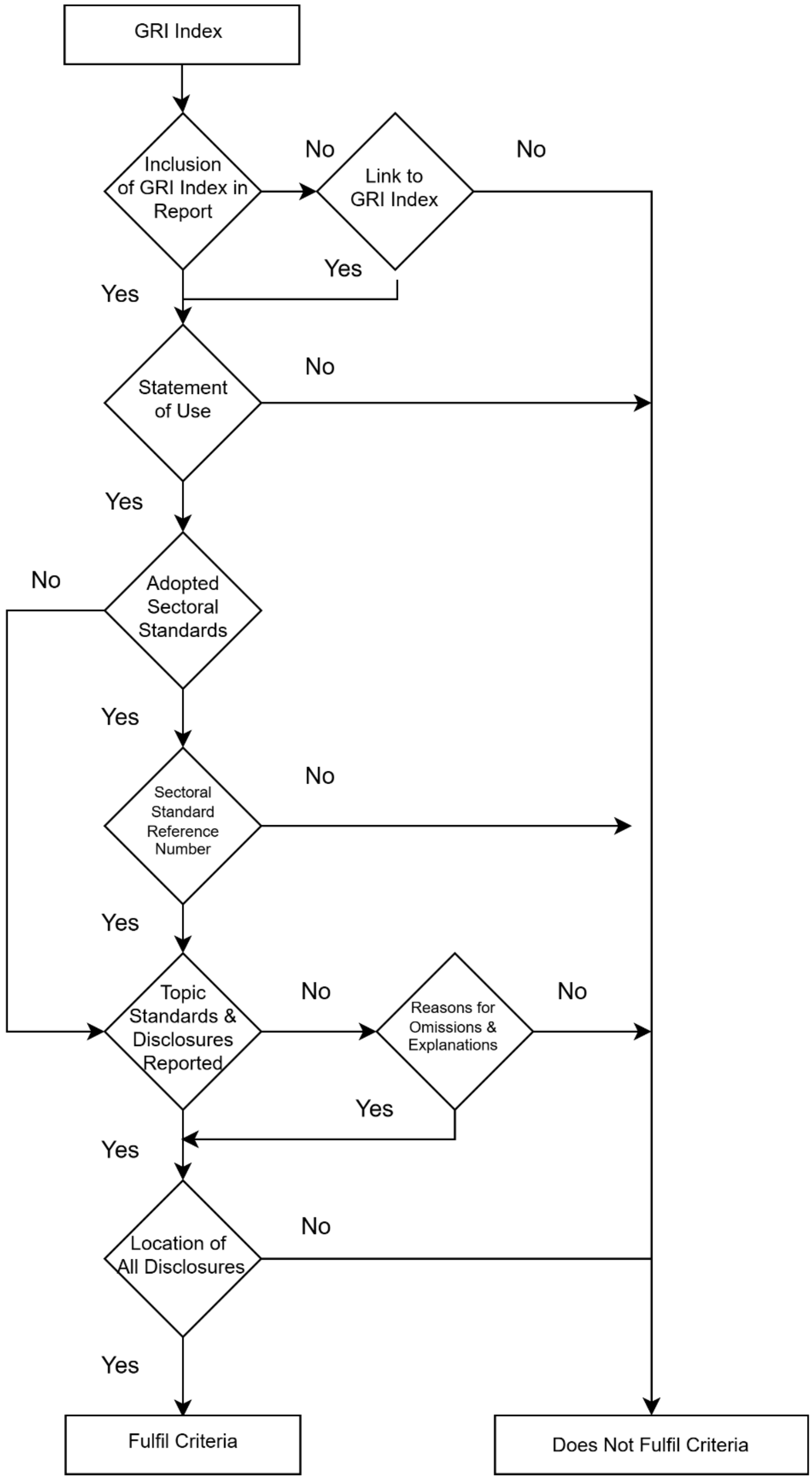


Figure 4.6: Decision-making Process for Segments of GRI Index

Sustainability scoring is used next to assess the sustainability performance of the company. The machine learning regression model is applied in the system to predict the sustainability score of the company based on ESG data extracted from the sustainability report. Currently, the ability to generate sustainability scores remains in the hands of a few numbers of organizations like S&P Global, LSEG and MSCI. This reduces the transparency of sustainability performance evaluation and prevents SMEs companies from getting an ESG score along with its benefits on finance and sustainability matters. Larger companies are likely to be awarded with ESG scores than smaller scaled companies, lowering their chances in obtaining green investment (Zumente and Lāce, 2021). This study hopes to improve the accessibility of sustainability scores to companies through the utilization of machine learning regression model that was trained to predict ESG scores from ESG data.

Many researchers had leveraged various machine learning models in regression tasks to predict ESG ratings with financial and non-financial data disclosed by the company. Del Vitto, Marazzina and Stocco (2023) used a variety of white box and black box regression models to reproduce Refinitiv ESG scores for companies from different backgrounds. The employment of both white box and black box machine learning models enabled understanding of the assessment scoring mechanisms used by scoring organizations, improving rating transparency which, in turn, enhances public trust in the rating system. The models used in the research successfully predicted the Environmental and Social scores from different regions with high accuracy while the prediction of Governance scores fluctuated with regions. This fluctuation was explained by the limitations of data available from the regions in which the models underperformed and the presence of noise in the data. However, the researchers believed that the prediction could be improved by incorporating more data into the training process. The findings of the research revealed that simple models can perform as well as the more complex models. These findings provided a possibility for the developers to select a model that is less demanding of computation capacity when designing the system to optimize the available resources. Furthermore, the researchers utilized the regression models to perform feature selection to reveal the influence of different indicators in the ESG score. Clarifying the importance of indicators in ESG ratings allowed companies with limited resources to focus on the indicators important to their operations, optimizing their ESG efforts.

Lin and Hsu (2023) had deployed a series of machine learning models on ESG scoring calculations for Taiwanese companies using indicators extracted from Taiwan Economic Journal (TEJ). The models they used include Extreme Learning Machine (ELM), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGBoost) and Random Forest (RF), a series of regression models that were tasked with uncovering the relationship between the indicators and the final ESG score. All the models were able to understand the relationships between the indicators and the ESG scores and demonstrated the ability to generate close predictions of the actual ESG scores. The conclusions of this research pointed out that supervised machine learning is faster at solving complex prediction problems when compared with mathematical models. The researchers recommended the consideration of sustainability policies in ESG evaluation. However, the result of the experiment led to the conclusion that Random Forest was not performing as well as other models. The potential reason for this phenomenon could be related to the calibration of the Random Forest models. The researchers had only included 20 trees in the model which could lead to the low effectiveness of the model as the number of trees is an important factor in improving the model's performance. The increase in the number of trees can result in improvement of model performance as proven by previous studies. Contreras et al. (2021) studied the optimization of Random Forest Regressor application on the modelling of rainfall runoff and forecasting in the Andean Mountains. They concluded that the increase of the n_estimator value, which is the parameter that decides the number of trees in a random forest model, has a significant positive effect on model performance in the range of 0 to 100. This finding is consistent with the conclusion of Nadi and Moradi (2019) which stated that the model with large numbers of smaller trees demonstrated better performance.

The literature introduced above demonstrates the predictive ability of various machine learning regression models to generate ESG scores for companies using sustainability data that can be found in their sustainability reports.

The ESG score has multiple functions in sustainability performance management. The ESG score is indicative of the sustainability performance of the company. The sustainability scoring model uses data of the sustainability indicators to generate sustainability scores as a way to quantify the sustainability performance of the company without human intervention. The sustainability data are extracted from sustainability reports from various companies according to sustainability indicators and used for sustainability scoring. The companies are categorized into different categories reflective of the sustainability performance of the companies. This study uses the LSEG scoring system as an example for explanation. The LSEG scoring system separates companies into four categories seen in Table 4.2 according to their ESG scores.

Table 4.2: Categories in LSEG ESG Scoring System (London Stock Exchange Group, 2023)

Range of ESG Score	Category	Descriptions
0 to 25	First Quartile	• Poor Relative ESG Performance • Insufficient Degree of Transparency in Reporting Material ESG Data Publicly.
>25 to 50	Second Quartile	• Satisfactory Relative ESG Performance • Moderate Degree of Transparency in Reporting Material ESG Data Publicly.
>50 to 75	Third Quartile	• Good Relative ESG Performance • Above Average Degree of Transparency in Reporting Material ESG Data Publicly.
>75 to 100	Fourth Quartile	• Excellent Relative ESG Performance • High Degree of Transparency in Reporting Material ESG Data Publicly

From the categories awarded to the companies, they can understand their standings in terms of sustainability performance. The predicted ESG Score can point out the weaknesses in the company sustainability strategy as the ESG score reflects the sustainability performance of the company. The score of each subcategory under ESG serves as measurements for the sustainability performance of the company. The scores can be traced back to the data of the related indicator set which is the benchmarks that allow the system to track the progress of the company on the sustainability goal achievement. The scores can also be related to a set of indicators that are indicative of the area of improvement. Using the environment score in the LSEG scoring system as an example, the score of each subcategory reveals information on the company's environmental performance on emissions, resource use, and innovation. The sustainability indicators related to each category are displayed in Table 4.3.

Table 4.3: Sustainability Indicators Related to the Categories in LSEG ESG Scoring System (Twinamatsiko and Kumar, 2022)

Pillar	**Category**	**Theme**
Environmental	Emission	Emission Waste Biodiversity Environmental Management System
	Innovation	Product Innovation Green Revenues R&D and Capital Expenditure
	Resource Use	Water Energy Sustainable Packaging Environmental Supply Chain
Social	Community	Equally Important to All Industries, hence, a median weight of five is assigned to all
	Human Rights	Human Rights
	Product Responsibility	Responsible Marketing Product Quality Data Privacy
	Workforce	Diversity and Inclusion Carrer development and Training

		Working Conditions Health and Safety
Governance	CSR Strategy	CSR Strategy ESG Reporting and Transparency
	Management	Structure (Independence, Diversity, Committees) Compensation
	Shareholders	Shareholder Rights Takeover Defences

The result of the assessment and analysis conducted by the machine learning model and its generated identification number are sent to the database to be recorded. The result of the assessment and analysis becomes benchmarks that will be compared with historical data to measure the progress of the company on the achievement of the sustainability goals of the company. The system will suggest improvements to the sustainability indicators related to the sustainability goals company. Priority will be placed on the indicators relating to the sustainability goals that are lacking progress.

4.1.3 User Interface

The user interface is the platform that allows users to interact with the decision support system (DSS). The user interface serves several functions including sustainability goal establishment, result display, and data visualization. The users will input the report draft into the DSS through the user interface. After the system has completed the analysis, the user interface will display the result of the standard compliance analysis, the generated scores, and the analysis of the strengths and weakness of the sustainability management strategy based on the ESG sectoral scores. The user interface will make use of visualization tools to display the data and results through charts and tables for ease of user comprehension.

4.2 Role in Sustainability Performance Management

The decision support system (DSS) can support many activities in sustainability performance management, standard compliance assessment of sustainability reports, generation of sustainability scores, analysis of sustainability data, sustainability progress measurement, and recommendations for company sustainability strategy improvement. These functions allow the DSS to contribute to company sustainability performance management through the enhancement of reporting standard compliance, measurement of sustainability progress, achievement of company sustainability goal, and support for stakeholder communication.

The utilization of machine learning technique assisted rule-based model in assessing sustainability reports automates and democratizes the checking for reporting standard compliance. The model checks the criteria for report compliance with the standards. In the situation where the report does not adhere to the standards, the model points out the missing elements of the report to help companies improve their reporting. This function assists companies in ensuring their report adheres to the standards which enable them to receive the benefits of sustainability reporting like increased favours in green financing opportunities and strengthening of consumer trust from company transparency (Deloitte & Touche LLP, 2022). Compliance with sustainability reporting standards prevents financial penalties in situations of compulsory reporting for the Publicly Listed Companies (PLC) listed on the markets that require sustainability reporting as a listing prerequisite and companies operating in countries with mandatory reporting requirements.

The DSS optimizes company sustainability efforts through strategic alignment with their sustainability goals. As all companies have their own sustainability priorities and financial limitations, a standardized sustainability management strategy might not be applicable. Hence, companies have the freedom to introduce their own set of sustainability goals that are in line with the company values and priorities to receive assistance from the DSS for goal achievement. As the recommendations suggested are goal-oriented, adoption of these measures ensures the company resources are used in areas with the most significant impact. The annual ESG scoring and benchmarking with the data of sustainability indicators quantify the impact of the current sustainability management strategy, allowing the company to monitor the progress and make timely amendments if necessary.

The ESG scoring and the sustainability data visualization and tracking features are great tools in the improvement of stakeholder engagement. ESG scores can communicate clearly to the consumers, employees, investors, and shareholders on the sustainability performance of the company. Sustainability performance has become an important factor for consumers when choosing a product (Boufounou *et al.*, 2023). Transparent communication is an integral part of maintaining consumer trust and the disclosure of ESG score is an easy way of communicating a company's sustainability impact. The ESG scores can also provide justifications for the implementation of sustainability measures to the shareholders when substantial capital investments are necessary. The visualization tool of sustainability data is crucial for straightforward communication of company performance. Study conducted by (Kim, Setlur and Agrawala, 2021) concluded that visual charts can more effectively communicate with viewers than when two were presented together. Therefore, the application of visual charts for data demonstrations and progress tracking can enhance stakeholder communication in conveying the company sustainability performance.

The DSS helps the company to incorporate sustainability into company operations through participation in the Plan-Do-Check-Act cycle. The various functions of the DSS support the planning of sustainability performance management strategies and action plans for the achievement of sustainability goals. The recommendations provided by the DSS assist the "Do" part of the cycle, enabling concrete sustainability actions that effectively contribute to the enhancement of company sustainability. The assessment of sustainability reports and analysis of indicator data enable a comprehensive understanding of the impact of the implemented measures and how they are contributing towards the sustainability goals. In turn, allows companies to act on their insufficiency and make effective and impactful adjustments for the next year. The utilization of the DSS in the PDCA cycle ensures that sustainability is not just an afterthought but an integral part of company operation.

4.3 Elements of Machine Learning

The sustainability performance management decision support system (DSS) incorporates many elements of machine learning such as data-driven decision-making processes, automated improvement, and optimization.

The DSS adopts data-driven decision-making principles through analysis of a significant amount of company ESG data and scores from different sources. The ESG data are extracted from company sustainability reports that are collected from the Internet through the web scraping method. The ESG data are the numerical and categorical values of the sustainability indicators from different sustainability reporting standards. The same data collection method is used for ESG score collection. The data is fed into the machine learning model that produces predictions of the ESG scores that correlate to the ESG data input. The machine learning model in use here adopts supervised learning that requires the training data to be labelled. The role of the model is to discover and develop the relationship between the independent and dependent variables. The optimum performance of the model is to consistently predict accurate results (ESG scores) from the input data (ESG data). However, the performance of a machine learning model is highly reliant on the quality of the data. In particular, the completeness, consistency, and the presence of noise in the data are influential to the performance of a supervised machine learning model. A complete dataset without missing values prevents bias and inaccuracy in the prediction result. Having consistent data in time series data allows the model to better understand the underlying relationship in the input and output data. The presence of non-relevant features in the data leads to less accurate predictions. Therefore, the data collection and processing steps are critical to produce a well-performed supervised machine learning model. The three factors can pose problems to the ESG scoring model in the beginning when the database includes a relatively small number of companies as the current practice of some SMEs during reporting of ESG data is incomplete, and inconsistent in the reported indicators over the years. However, as the database grows with the inclusion of more and more companies and collects more ESG data that improves the completeness and consistency of the data, over time, the performance of the model will improve. With the addition of new data, the model takes the new information into consideration when predicting the ESG scores and amends the decision-making process of score prediction. As mentioned in previous sections, the current practice of ESG scoring still is highly reliant on the judgement of human experts, therefore, the results could be inconsistent and unreliable due to human biases. The employment of a data-driven supervised machine learning model in ESG scoring can improve the reliability and consistency of ESG scores generated. The adoption of data-driven decision-making principles in the DSS ensures that the insights provided to the decision-makers are objective, and evidence-based to help them implement effective and efficient strategies to achieve their sustainability goals.

The data-driven decision-making approach of the system enables adaptation to the dynamic sustainability standards that shift focus as new scientific discoveries and development sustainability. The machine learning integrated system is different from the conventional approach in the necessity of explicit coding. The conventional approach to decision-making depends on the predefined rules and algorithms built into the system by the developers. These systems rely heavily on the input of experts in the development phase. A machine learning-integrated DSS gains insights from historical data through the utilization of various machine learning models to solve the decision problem without explicit programming. The integration of machine learning with DSS offers a competitive edge over the conventional approach in terms of dynamic analysis which allows the system to adjust and improve its decision-making process over time to adapt to changes in sustainability standards. As sustainability is an all-encompassing concept that includes many topics under ESG, the material topics of each industry can have drastic differences depending on the nature of the industry. The electrical and electronics industry has more focus on topics like energy usage and efficiency, and materials while financial institutes are more affected by governance topics such as business ethics and data and technology (Mohr, Riquelme and Quick, 2022). The materiality of sustainability issues also evolves over time due to changes in the development of social and environmental well-being. Conventional DSS would require two different system development approaches for these two distinct industries and the logic behind the decision-making process can be rendered invalid due to the dynamic nature of materiality. The machine learning integrated DSS can be adapted for different industries using different sets of training data containing ESG data and scores of companies from the target industry. As the supervised machine learning model makes adjustments to its prediction strategies with the introduction of new data points, the system can remain in service even when there is a drastic change in scoring methods or in sustainability topic priority.

4.4 The Performance of Scoring Model

The ESG scoring model is executed using the machine learning model, random forest regressor, with a supervised learning style. The task of the model is to predict the sustainability scores of companies from sustainability data extracted from company sustainability reports according to environmental indicators listed in the GRI standards. The number of trees in a random forest model is set to be 30 and remains for the entire experiment. The experiments are run three times respectively with three samples consisting of sustainability data from 5 companies (Sample 1), 10 companies (Sample 2), and 15 companies (Sample 3). The dataset was split into 60% training and 40% testing data. The results are demonstrated in Table 4.4, Figure 4.7, Figure 4.8, Figure 4.9, Figure 4.10 and Figure 4.11.

Table 4.4: Values of the Performance Assessment Metrics for 3 Samples

N_estimator	30		30		30	
Sample	1		2		3	
Data Size	5		10		15	
Dataset	Training	Testing	Training	Testing	Training	Testing
R^2 Score	0.64	-1379.64	0.81	-23.19	0.86	0.25
EVS	0.66	-83.64	0.84	0.18	0.86	0.58
MAE	12.87	18	5.79	36.56	7.31	12.93
MSE	189.08	345.16	80.21	1383.51	69.68	213.73
RMSE	13.75	18.58	8.96	37.2	8.35	14.62

The performance of the scoring model is assessed through five parameters, the Coefficient of Determination (R^2 score), Explained Variance Score (EVS), Mean Average Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Overall, the model demonstrates improvement in the prediction of emission score with the increase in sample size.

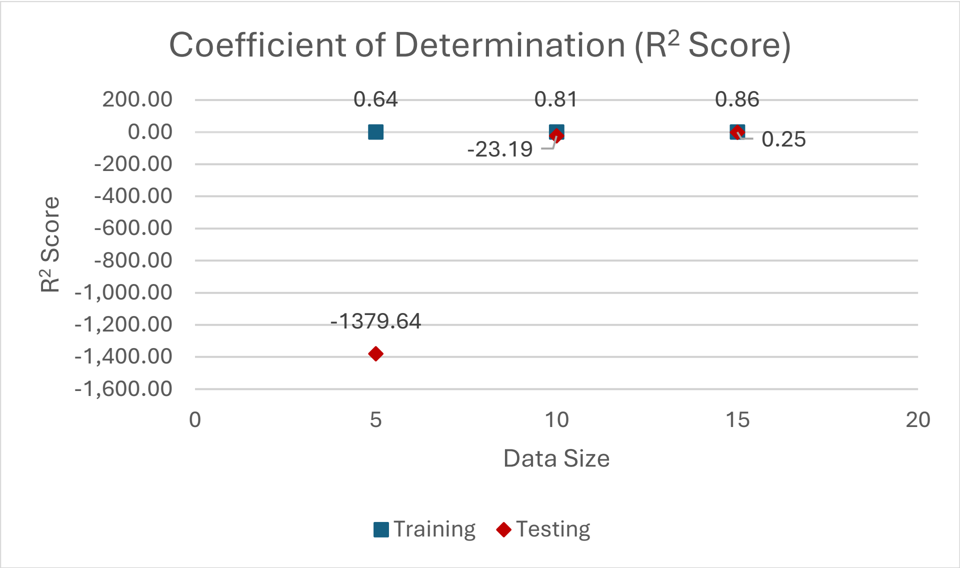


Figure 4.7: Value of Coefficient of Determination (R^2 Score) for Training and Testing Dataset

The value of the Coefficient of Determination or the R^2 score for the training dataset consistently increases from 0.64 to 0.86 when the data size grows from 5 to 15 sets of data. The R^2 score for the testing data set starts from a negative score of -1379.64 for sample 1 climbs to -23.19 for sample 2 and reaches 0.25 for sample 3 which contains 15 sets of data. The negative R^2 score for the sample 1 and 2 means that the model does not fit the data well. The model cannot capture the variability in the relationship between the ESG data and the emission score with the data provided. Although the R^2 score for sample 3 is quite low but the positive sign says that the model is improving. The trend of the R^2 score for both datasets suggests that the introduction of new data can produce a better performing model.

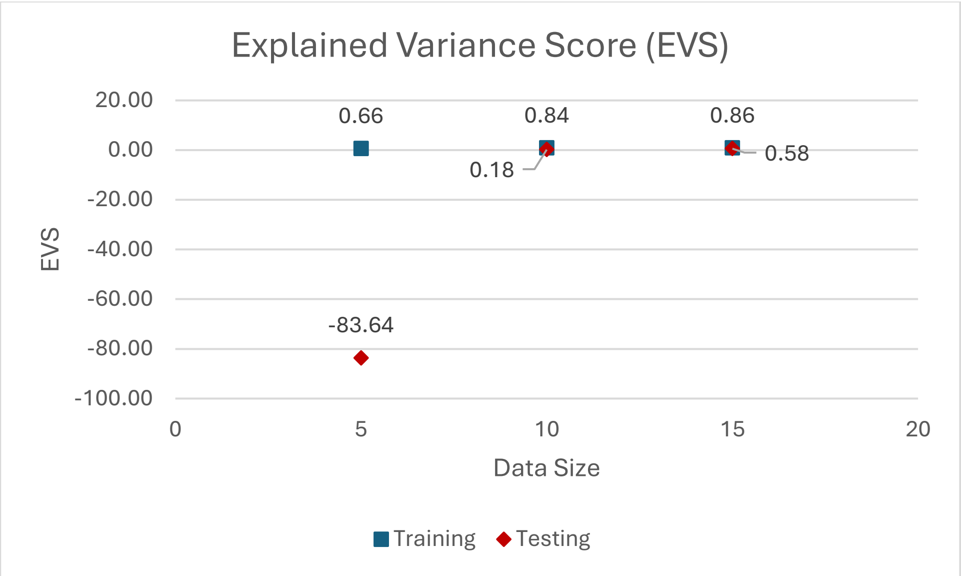


Figure 4.8: Value of Explained Variance Score (EVS) for Training and Testing Dataset

The results of the explained variance score (EVS) are consistent with the result of the R^2 score. The EVS for both training and testing datasets increases with the data size. For training data, the EVS for sample 1 is 0.66. After the addition of 5 sets of data, the EVS rises to 0.81. The EVS for sample 3 is 0.86. Similar to the R^2 score, the testing EVS for sample 1 is of negative value as well, standing at -83.64. The additional 5 data sets lead to substantial improvement in EVS, causing it to leap to 0.18. The improvement sustains, the EVS for sample 3 is 0.58. From the result of the EVS, the performance of the regression model appears to benefit from the increase in data.

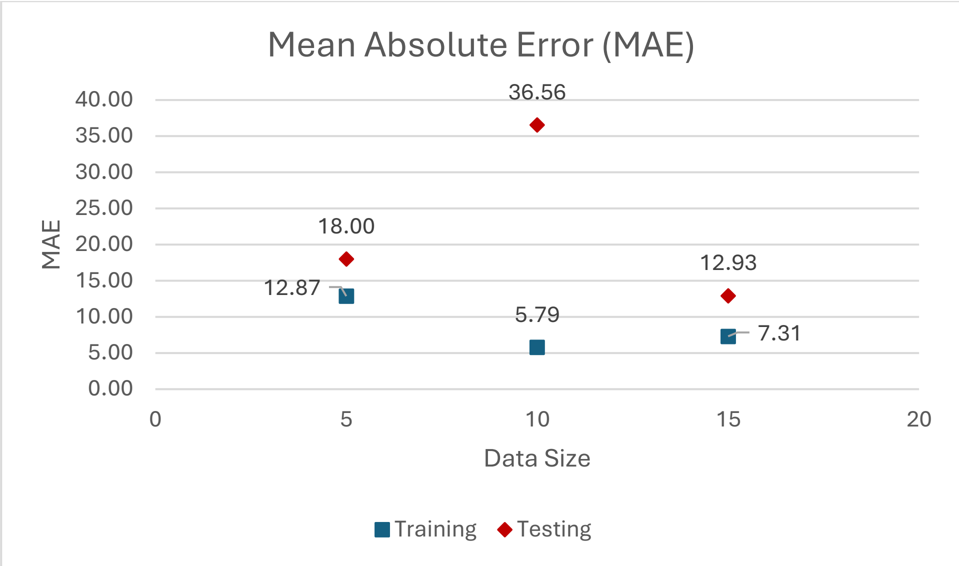


Figure 4.9: Value of Mean Absolute Error (MAE) for Training and Testing Dataset

The mean absolute error measures the overall difference between the predicted and the actual values. Training MAE for sample 1 with 5 sets of data is 12.87. It drops to 5.79 for sample 2 but increases slightly to 7.31 for sample 3. The MAE for the testing data is 18.00 for sample 1, doubles to 36.56 for sample 2, and falls back to 12.93 for sample 3. The changes in data size seem to have an inconsistent impact on the MAE of training and testing data.

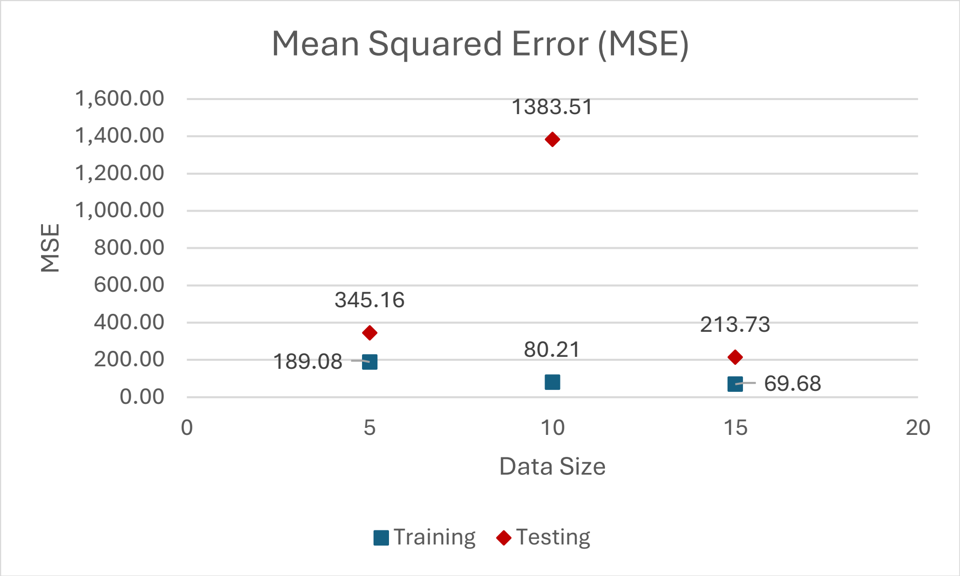


Figure 4.10: Value of Mean Squared Error (MSE) for Training and Testing Dataset

The metric Mean Squared Error accentuates the big error in the prediction of the model. The training MSE demonstrates a consistent fall with the increase in training data size, the training MSE is 189.08, 80.21, and 69.68 for the sample 1, 2, and 3 respectively. The same pattern of changes from MAE is observed in the testing MSE. The MSE of the testing dataset for sample 1 is 345.16 which spikes to 1383.51 for sample 2 with 10 data sets and experiences a decline to 213.73 for the largest data size of 15.

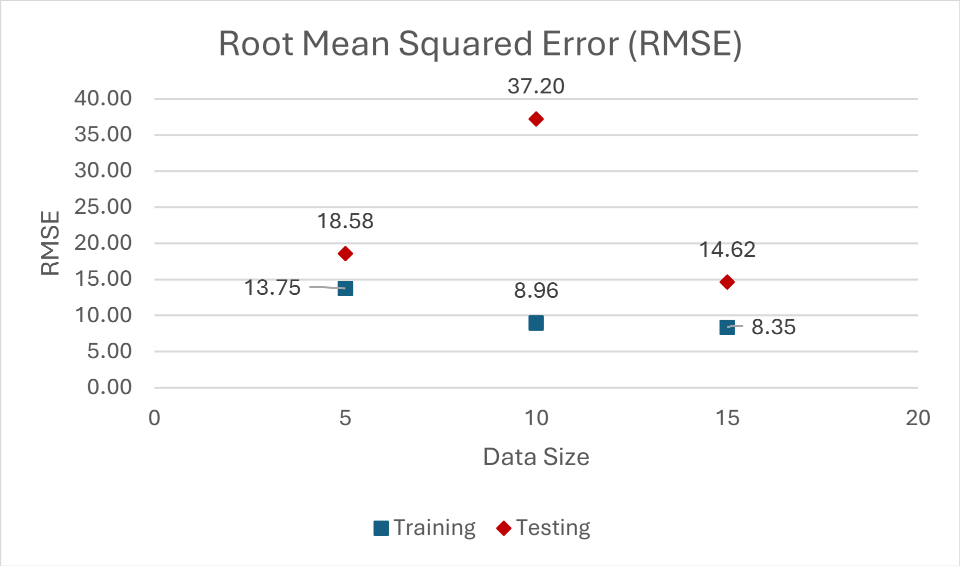


Figure 4.11: Value of Root Mean Squared Error (RMSE) for Training and Testing Dataset

The root means squared error is taken by square rooting the MSE. This metric reveals information on the error margin of the prediction made by the model. The training RMSE stands at 13.75 for the smallest data size. The increase in data brings the RMSE down to 8.96 (Sample 2) and 8.35 (Sample 3). The RMSE for the testing data starts at 18.58 and grows to 37.2 after doubling the data size and decreases to 14.62 for the data size of 15 companies.

Based on the performance metrics, the performance of the regression model is improving with the growth of data size. The R^2 score and EVS demonstrate consistent improvement with the introduction of additional data. Their behaviours communicate that, with a large amount of training data, the regression model is capable of capturing the variance in the input and output data. The access to new data causes irregular changes in the error evaluation metrics. For the training data, new data can generally effectively bring down the value of the metrics but the metrics for the testing data show a rise-and-fall pattern when new data are introduced. To understand the reasons for such behaviour of the error metrics, attentions are shifted to the composition of the dataset used in this experiment. Table 4.5 reveals the statistical information of the three samples used. All three samples have a mean of over 70 with a standard deviation of around 20. The 50 percentiles of all three samples had reached the value of 80 while the minimum value in all three samples is only 44.

Table 4.5: Statistical Information of Three Samples

Sample	1	2	3
Data Size	5	10	15
Mean	75.40	72.8	75.60
Standard Deviation	23.15	21.57	20.89
Minimum	44.00	44.00	44.00
25 Percentile	58.00	52.00	55.50
50 Percentile	88.00	81.00	85.00
75 Percentile	89.00	88.75	93.00
Maximum	98.00	98.00	99.00

It is seen that the value of the dependent variable in the database is skewed towards a higher end. During the database construction phase, the companies selected to be included in the database were done randomly without consideration of the balance of Emission score in the database. This leads to an imbalanced database consisting of mostly companies with high emission scores. Due to the small sample size and the randomness in the split of training and testing datasets, the distribution of the data points in the training and testing dataset becomes asymmetric. The data points in the training and testing dataset in all samples are shown in Table 4.6. The training datasets for the sample 1 and 2 are mostly populated by low data points while having the majority of high data points in the testing dataset. The training and testing datasets of Sample 3 have a better mix of low and high data points. However, it must be noted that the database of this experiment does lack diversity. The minimum emission score in the database consisting of 15 sets of data is 44 while the median stands at 85, a result of the 8 emission scores having values higher than 80 out of the 15 scores included.

Table 4.6: Data Points in Three Samples

Sample	1		2		3	
Data Size	5		10		15	
No.	Training	Testing	Training	Testing	Training	Testing
1	98	89	58	77	89	66

2	58	88	88	98	98	88
3	44		44	85	58	95
4			50	89	50	44
5			95		44	77
6			44		91	85
7					99	
8					97	
9					53	
Median	58.00	88.50	54.00	87.00	89.00	81.00
Mean	66.67	88.50	63.17	87.25	75.44	75.83
Std Dev	22.88	0.50	20.68	7.56	22.11	16.89

A comparison of the predicted and the actual value of the testing datasets for all samples is given by Table 4.7, Figure 4.12, Figure 4.13 and Figure 4.14.

Table 4.7: Value of the Predicted and Actual Emission Score in the Testing Database

Sample	1		2		3	
Data Size	5		10		15	
No.	Actual	Predicted	Actual	Predicted	Actual	Predicted
1	89	75.6	77	50.70	66	53.30
2	88	65.4	98	52.43	88	86.83
3			85	48.47	95	78.63
4			89	51.17	44	53.7
5					77	63.17
6					85	61.20

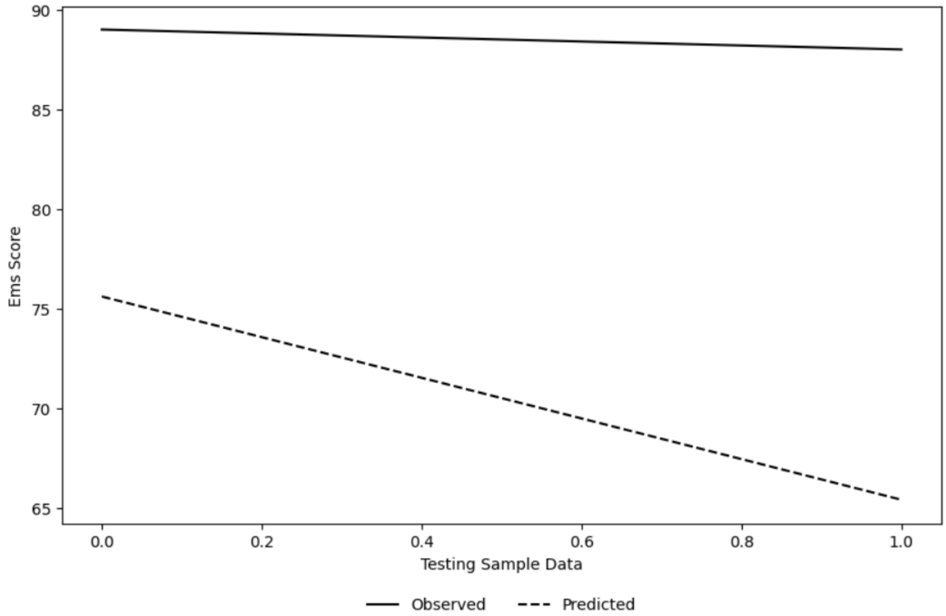


Figure 4.12: Predicted and Actual Value of Testing Dataset for Sample 1

The comparison of the prediction and the actual values for the testing data of sample 1 shows that the model has yet to understand the relationship between the data of the sustainability indicators and the emission scores with the training data size of 3. The X-axis of the figure is the index of the sample data starting from 0.00 to 1.00 for two samples while the Y-axis represents the emission score. The two axes represent the same things for all the three figures. The solid line maps the actual emission score in the testing dataset while the dashed line maps the predictions. There is a wide gap between the solid and dashed lines with the solid line floating near 90 with a small degree downward slope and the dashed line sitting a little above 75 and falling to 65 with a steep slope. The two lines show very little degree of correlation, indicating the model does not fit the data well.

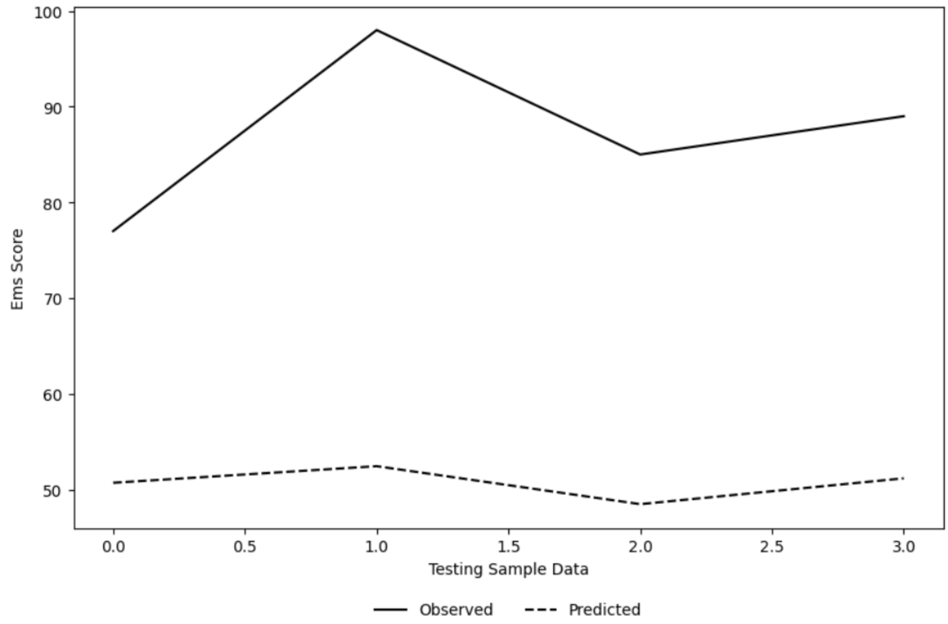


Figure 4.13: Predicted and Actual Value of Testing Dataset for Sample 2

The model was trained with 9 sets of data from sample 2 which have a comparatively balanced population of data points. The testing dataset for sample 2 has four data points. Looking at Figure 4.13, the gap between the solid and the dashed lines is still significant. The model begins to show signs of understanding the pattern of scoring, seen through the behaviour of the dashed line mirroring the solid line with less intense variability. All the actual emission scores are still much higher than the predictions of the model.

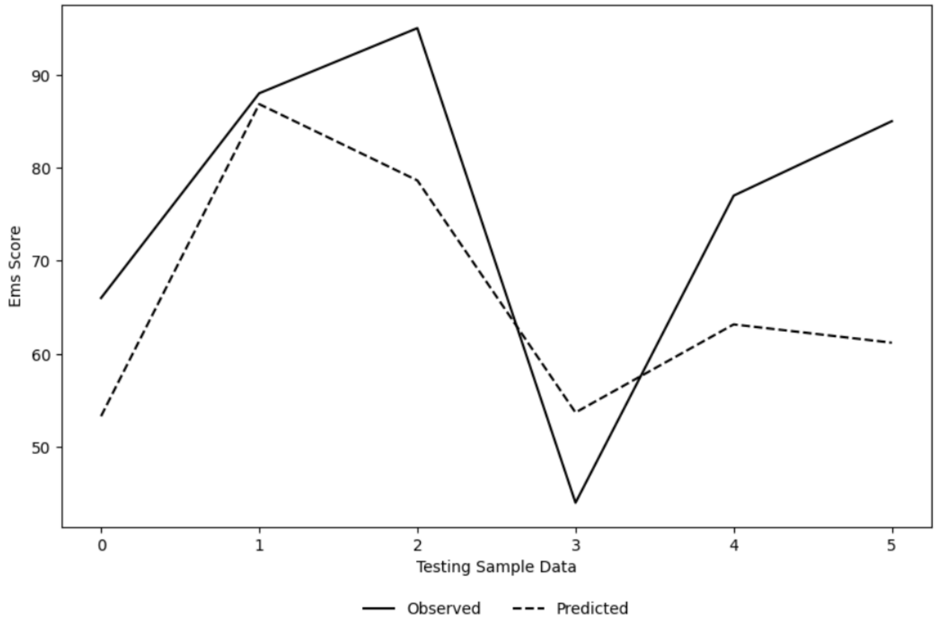


Figure 4.14: Predicted and Actual Value of Testing Dataset for Sample 3

Sample 3 has 6 data points in the testing dataset. From Figure 4.13, the gap between the solid and dashed lines has reduced significantly, meaning that the predictions are getting closer to the actual emission scores. Although the general trend of the dashed line is starting to align with the solid line, there is still some discrepancy between the actual and predicted values of several data points. This suggests that the model still requires further training to produce accurate predictions consistently.

A close examination of the predicted emission score explains the behaviours of the performance metrics. The predicted emission scores for the testing dataset of the sample 1 and 2 are lower than the actual emission scores by at least 20 points. The large difference in the actual and the predicted values leads to the surge of all the error metrics for sample 2. The model also cannot capture the variance of the relationship of the sample 1 and 2 datasets which explains the negative R^2 Score for both sample and the EVS of sample 1.

The random forest regressor has shown potential in the task of ESG score prediction in this experiment even within the limitations of small sample size and asymmetric data composition. The results of the experiment suggest that data size has a positive relationship with the performance of the random forest regressor model on emission score prediction. This finding is supported by previous studies. Cui and Gong (2018) experimented with 25 data sizes ranging from 20 to 700 in prediction tasks using machine learning regression models and concluded that the performance of the regression models stabilizes and improves with the increase in sample size. Bouasria et al. (2023) found the sample size has a positive effect on increasing the R^2 Score within 300 samples, after which the effect of further sample expansion became insignificant. Their conclusion aligned with the results of Bailly et al. (2022), saying that the increase of data from 1000 to 100000 has little effect on regression model performance. The effects of data size on model performance diminish after a certain threshold. The sample size used in these studies is far from sufficient for the random forest regressor to generate accurate predictions of emission stably, but the predictions of the model trained with sample 3 provide promising aspects of ESG scoring with a supervised machine learning model.

It has been iterated before that the database in use is an imbalanced database due to negligence during database construction. The database is not representative of the distribution of the real ESG scores, which causes imbalance bias in the emission score prediction model (Gu and Oelke, 2019). The model is prone to predict lower emission scores if they are the main population in the training dataset seen in the results of model prediction for sample 1 and 2. The difference in the distribution of data in the training and testing data sets negatively affects the model learning (Ben-David *et al.*, 2010). This results in the spike of MAE, MSE, and RMSE seen for the case of sample 2. The predictions for the testing dataset of sample 3, the sample that has a relatively balanced distribution in both datasets, show signs that the random forest regressor is learning the scoring patterns from the data. This conclusion aligns with the results of the performance metrics which are the best out of all three model runs.

4.5 The Challenges of Machine Learning Integration with Sustainability Performance Management

The lack of an accessible sustainability database for the training of machine learning model. The existing sustainability databases are mostly available for paid users which lowers the accessibility of the data. A majority of databases on the market focus on western countries like the members of the European Union and the United States. The development of the sustainability database is further complicated by the format of sustainability reports as the current default format is the Portable Document Format which is not readily machine readable. Therefore, the construction of a sustainability database requires manual extraction of sustainability data from the reports, which is time consuming and prone to human error.

The change of default report format into the eXtensible Business Reporting Language format (XBRL) can potentially provide solutions to the first two problems mentioned. The XBRL files are document files that are digitally tagged for machine readability. The XBRL reporting is used extensively in financial reporting for the benefits of data accessibility and comparability. The XBRL reporting is also adopted in Malaysia for financial reporting as mandated by the Suruhanjaya Syarikat Malaysia, Securities Commission Malaysia and Inland Revenue Board of Malaysia (Ilias, Ghani and Azhar, 2019). Tawiah and Borgi (2022) concluded that XBRL reporting helps improve information efficiency and enhance data processing led to an increase in the quality of financial reports. The adoption of the XBRL format in financial reporting also data quality in terms of accessibility, accuracy, and consistency in format (Wang and Gao, 2012). Furthermore, research shows that the XBRL financial reporting mandate resulted in enhanced structural comparability of financial statements (Yang, Liu and Zhu, 2018). The XBRL reporting format has yet to be used in sustainability reporting practice. However, the European Union had incorporated XBRL reporting in the draft of the Corporate Sustainability Reporting Directives as the electronic reporting format in the hope to improve governance efficiency (European Parliament, 2022). Malaysia can follow in their footsteps and introduce digital reporting into our sustainability reporting framework.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study has developed a decision support system (DSS) framework integrated with machine learning techniques for company sustainability performance management. The DSS analyses the sustainability reports on their compliance with the reporting standards and extracts data from them to assess the sustainability performance of companies. A rule-based approach complemented with Natural Language Processing (NLP) Technology is applied for the assessment of the standard compliance evaluation process. The machine learning regression model is used for the task of sustainability scoring as the assessment of a company's performance on sustainability matters. The integration of machine learning and DSS enables a data-driven decision-making process in companies and allows the DSS to evolve with the data it consumes.

The experiment is conducted in this study to assess the performance of the machine learning model on sustainability scoring. The machine learning regression model, Random Forest Regressor is applied in the experiment to predict the sustainability scores from data of the GRI sustainability indicators. The performance of the regression model is examined through performance metrics including Coefficient of Determination (R^2 Score), Explained Variance Score (EVS), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The value for the R^2 Score and the EVS increases with the amount of data in the sample for both training and testing data, the increase in data size has a larger impact on the result for the testing data. Bigger data size has an inverse relation with

MAE, MSE, and RMSE, the value of these metrics declines as the data size expands, and the expansion of data size has a similar impact on training and testing datasets. Although the random forest regressor in the experiment is not mature enough to produce consistently accurate predictions, the result shows that the increase in data size can significantly improve the model performance in sustainability scoring tasks. However, the importance of curating a database with a balanced distribution is emphasised as an imbalanced database can result in systematic bias in the model. The problem is further complicated by the small data size as observed in the comparisons between the predicted and actual sustainability scores of three samples. The result reveals that the data distribution in the training dataset directly affects the regression scoring strategy. Hence, it is crucial to ensure that the training datasets are representative of real-life data distribution. As observed in the result of sample 3 where the training data are more balanced, the predictions made by the regression model start to align with the actual value in the database, indicating that the model is capturing the scoring pattern in the datasets.

The main challenge is identified in this study which related to the difficulties in sustainability database construction. Due to the format of sustainability reports being machine inaccessible Portable Document Format (PDF), the development of the database is time-consuming and prone to errors. This study suggests that the adoption of the eXtensible Business Reporting Language as the default report format for its data processing benefits.

5.2 Recommendations

This study provided a conceptual framework for a sustainability performance management DSS embedded with machine learning technology. This study only covers the development of the framework. Further studies can consider the continuation on the construction and implementation of the DSS. Moreover, the experiment with sustainability scoring model can be expanded with larger and more diverse datasets as the database used in this experiment is small in size and has an uneven distribution which can affect the scoring strategy of the scoring model. This study focuses on the prediction of Emission score which is only one subcategory under the LSEG ESG Scores. Therefore, future studies can consider expanding the scope of the scoring targets to include the other categories in the LSEG ESG Scoring system, as well as incorporating other scoring systems.

REFERENCES

Akbulut, D.H. and Kaya, I. (2019) 'Sustainability reporting and firm performance', *Pressacademia*, 9(9), pp. 81–84. Available at: https://doi.org/10.17261/pressacademia.2019.1071.

Akpamah, P., Ivan-Sarfo, E. and Matkó, A. (2021) 'Organizational Culture As A Strategy', *Cross-Cultural Management Journal*, XXIII, pp. 15–26.

Alhaddi, H. (2015) 'Triple Bottom Line and Sustainability: A Literature Review', *Business and Management Studies*, 1(2), pp. 6–10.

Angelakoglou, K. and Gaidajis, G. (2020) 'A conceptual framework to evaluate the environmental sustainability performance of mining industrial facilities', *Sustainability (Switzerland)*, 12(5). Available at: https://doi.org/10.3390/su12052135.

Baffo, I. *et al.* (2023) 'A decision support system for measuring and evaluating solutions for sustainable development', *Sustainable Futures*, 5. Available at: https://doi.org/10.1016/j.sftr.2023.100109.

Bailly, A. *et al.* (2022) 'Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models', *Computer Methods and Programs in Biomedicine*, 213. Available at: https://doi.org/10.1016/j.cmpb.2021.106504.

Banerjee, S.B., Iyer, E.S. and Kashyap, R.K. (2003) 'Corporate Environmentalism: Antecedents and Influence of Industry Type', *Journal of Marketing*, 67, pp. 106–122.

Ben-David, S. *et al.* (2010) *Impossibility Theorems for Domain Adaptation*.

Bouasria, A. *et al.* (2023) 'Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents', *Ecological Informatics*, 78. Available at: https://doi.org/10.1016/j.ecoinf.2023.102294.

Boufounou, P. *et al.* (2023) 'ESGs and Customer Choice: Some Empirical Evidence', *Circular Economy and Sustainability*, 3(4), pp. 1841–1874. Available at: https://doi.org/10.1007/s43615-023-00251-8.

Brugger, F. *et al.* (2023) *Analysing Sustainability Reports Using Machine Learning*. Available at: https://nadel.ethz.ch/.

Büyüközkan, G. and Karabulut, Y. (2018) 'Sustainability performance evaluation: Literature review and future directions', *Journal of Environmental Management*, 217, pp. 253–267. Available at: https://doi.org/10.1016/j.jenvman.2018.03.064.

Campello, R.J.G.B. *et al.* (2020) 'Density-based clustering', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 10(2). Available at: https://doi.org/10.1002/widm.1343.

Cantele, S., Landi, S. and Vernizzi, S. (2024) 'Measuring corporate sustainability in its multidimensionality: A formative approach to integrate ESG and triple bottom line approaches', *Business Strategy and the Environment* [Preprint]. Available at: https://doi.org/10.1002/bse.3872.

Castilla-Polo, F. and Guerrero-Baena, M.D. (2023) 'The business case for sustainability reporting in SMEs: consultants' and academics' perceptions', *Sustainable Development,* 31(5), pp. 3224–3238. Available at: https://doi.org/10.1002/sd.2576.

Chalmeta, R. and Ferrer Estevez, M. (2023) 'Developing a business intelligence tool for sustainability management', *Business Process Management Journal*, 29(8), pp. 188–209. Available at: https://doi.org/10.1108/BPMJ-03-2023-0232.

Chang, W.F. *et al.* (2019) 'Drivers of sustainability reporting quality: financial institution perspective', *International Journal of Ethics and Systems*, 35(4), pp. 632–650. Available at: https://doi.org/10.1108/IJOES-01-2019-0006.

Chen, Y.-C., Hung, M. and Wang, Y. (2018) 'The effect of mandatory CSR disclosure on firm profitability and social externalities: Evidence from China', *Journal of Accounting and Economics*, 65(1), pp. 169–190. Available at: https://doi.org/10.1016/j.jacceco.2017.11.009.

Contini, G. and Peruzzini, M. (2022) 'Sustainability and Industry 4.0: Definition of a Set of Key Performance Indicators for Manufacturing Companies', *Sustainability*, 14(17), p. 11004. Available at: https://doi.org/10.3390/su141711004.

Contreras, P. *et al.* (2021) 'Influence of random forest hyperparameterization on short-term runoff forecasting in an andean mountain catchment', *Atmosphere*, 12(2). Available at: https://doi.org/10.3390/atmos12020238.

Cui, Z. and Gong, G. (2018) *The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features*.

Deloitte & Touche LLP (2022) *Purpose-driven ESG in the consumer industry: Sustainability as a value-creator*. Available at: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-purpose-driven-esg-in-the-consumer-industry-infograhic.pdf (Accessed: 15 August 2024).

Deogun, J.S. (1988) 'A Conceptual Approach to Decision Support System Models ', *Information Processing & Management*, 24(4), pp. 429–448.

Diez-Cañamero, B. *et al.* (2020) 'Measurement of Corporate Social Responsibility: A Review of Corporate Sustainability Indexes, Rankings and Ratings', *Sustainability*, 12(5), p. 2153. Available at: https://doi.org/10.3390/su12052153.

Dočekalová, M.P. and Kocmanová, A. (2016) 'Composite indicator for measuring corporate sustainability', *Ecological Indicators*, 61, pp. 612–623. Available at: https://doi.org/10.1016/j.ecolind.2015.10.012.

Engert, S. and Baumgartner, R.J. (2016) 'Corporate sustainability strategy – bridging the gap between formulation and implementation', *Journal of Cleaner Production*, 113, pp. 822–834. Available at: https://doi.org/10.1016/j.jclepro.2015.11.094.

Epstein, M.J. and Roy, M.-J. (2001) 'Sustainability in Action: Identifying and Measuring the Key Performance Drivers', *Long Range Planning*, 34(5), pp. 585–604. Available at: https://doi.org/10.1016/S0024-6301(01)00084-X.

European Parliament, C. of the E.U. (2022) *Corporate Sustainability Reporting Directives, European Parlianment, Council of the European Union*. European Union: European Parliament.

Fernando, J.G. and Baldelovar, M. (2022) 'Decision Support System: Overview, Different Types and Elements', *Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery*, 2(2).

Fischbach, J. *et al.* (2023) 'Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool', in *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023*. Institute of Electrical and Electronics Engineers Inc., pp. 2823–2830. Available at: https://doi.org/10.1109/BigData59044.2023.10386488.

Flammer, C. (2013) 'Corporate Social Responsibility and Shareholder Reaction: The Environmental Awareness of Investors', *Academy of Management Journal*, 56(3), pp. 758–781. Available at: https://doi.org/10.5465/amj.2011.0744.

Galindo, P.V., Vaz, E. and de Noronha, T. (2015) 'How corporations deal with reporting sustainability: Assessment using the multicriteria logistic biplot approach', *Systems*, 3(1), pp. 6–26. Available at: https://doi.org/10.3390/systems3010006.

Global Reporting Initiative (2023) 'GRI 1: Foundation 2021', *GRI Standard*. Global Reporting Initiative, p. 14.

Global Sustainable Investment Alliance (2021) *Global Sustainable Investment Review 2020*. Available at: www.robeco.com.

Gu, J. and Oelke, D. (2019) 'Understanding Bias in Machine Learning'. Available at: http://arxiv.org/abs/1909.01866.

Hamdani, R. El *et al.* (2021) 'A combined rule-based and machine learning approach for automated GDPR compliance checking', in *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*. Association for Computing Machinery, Inc, pp. 40–49. Available at: https://doi.org/10.1145/3462757.3466081.

Hardcastle, D. and Mattios, G. (2020) *Meeting the region's sustainability challenges offers potential for a trillion dollar annual economic opportunity by 2030*.

Hasan, M.S. *et al.* (2017) *Decision Support System Classification And Its Application In Manufacturing Sector: A Review*. Available at: www.jurnalteknologi.utm.my.

Ilias, A., Ghani, E.K. and Azhar, Z. (2019) 'XBRL in the Malaysian Financial Reporting Landscape: The Regulators' Perspective', *Fact Periodicals*, 2 December. Available at: https://accountancy.uitm.edu.my/index.php/en/research-ican/fact-periodicals/61-2020-february/137-xbrl-in-the-malaysian-financial-reporting-landscape-the-regulators-perspective (Accessed: 30 August 2024).

Jacobs, B.L. (2024) 'From CSR and TBL to ESG and the SDGs: Roots from Resistance to Regularization', *Louisiana Law Review*, 84(4), pp. 1251–1262.

Jain, R. and Raju, S.S. (2016) *Decision support system in agriculture using quantitative analysis*. Agrotech Publishing Academy.

Juan, Y.-K., Gao, P. and Wang, J. (2010) 'A hybrid decision support system for sustainable office building renovation and energy performance improvement', *Energy and Buildings*, 42(3), pp. 290–297. Available at: https://doi.org/10.1016/j.enbuild.2009.09.006.

Kalogiannidis, S. *et al.* (2024) 'Relationship between Climate Change and Business Risk: Strategies for Adaptation and Mitigation: Evidence from a Mediterranean Country', *WSEAS Transactions on Environment and Development*, 20, pp. 276–294. Available at: https://doi.org/10.37394/232015.2024.20.28.

Kameshwaran, K. and Malarvizhi, K. (2014) *Survey on Clustering Techniques in Data Mining*. Available at: www.ijcsit.com.

Kang, H. and Kim, J. (2022) 'Analyzing and Visualizing Text Information in Corporate Sustainability Reports Using Natural Language Processing Methods', *Applied Sciences (Switzerland)*, 12(11). Available at: https://doi.org/10.3390/app12115614.

Kanmani, A.P. *et al.* (2020) 'Assessing global environmental sustainability via an unsupervised clustering framework', *Sustainability (Switzerland)*, 12(2). Available at: https://doi.org/10.3390/su12020563.

Kantabutra, S. (2024) 'Toward a sustainability performance management framework', *Heliyon*. Elsevier Ltd. Available at: https://doi.org/10.1016/j.heliyon.2024.e33729.

Kantabutra, S. and Ketprapakorn, N. (2020) 'Toward a theory of corporate sustainability: A theoretical integration and exploration', *Journal of Cleaner Production*, 270. Available at: https://doi.org/10.1016/j.jclepro.2020.122292.

Kim, D.H., Setlur, V. and Agrawala, M. (2021) 'Towards understanding how readers integrate charts and captions: A case study with line charts', in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. Available at: https://doi.org/10.1145/3411764.3445443.

Kim, S., Lee, G. and Kang, H. (2021) ‘Risk management and corporate social responsibility’, *Strategic Management Journal*, 42(1), pp. 202–230. Available at: https://doi.org/10.1002/smj.3224.

Kumar Uppada, S. (2014) *Centroid Based Clustering Algorithms-A Clarion Study*. Available at: www.ijcsit.com.

Laskar, N. (2018) ‘Impact of corporate sustainability reporting on firm performance: an empirical examination in Asia’, *Journal of Asia Business Studies*, 12(4), pp. 571–593. Available at: https://doi.org/10.1108/JABS-11-2016-0157.

León-Soriano, R., Jesús Muñoz-Torres, M. and Chalmeta-Rosaleñ, R. (2010) ‘Methodology for sustainability strategic planning and management’, *Industrial Management & Data Systems*, 110(2), pp. 249–268. Available at: https://doi.org/10.1108/02635571011020331.

Li, T.T. *et al.* (2021) ‘Esg: Research progress and future prospects’, *Sustainability (Switzerland)*. MDPI. Available at: https://doi.org/10.3390/su132111663.

Li, Y. and Rockinger, M. (2024) ‘Unfolding the Transitions in Sustainability Reporting’, *Sustainability (Switzerland)*, 16(2). Available at: https://doi.org/10.3390/su16020809.

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

London Stock Exchange Group (2023) *LSEG ESG Scores*. Available at: https://www.lseg.com/en/data-analytics/sustainable-finance/esg-scores#t-score-range (Accessed: 30 July 2024).

Lozano, R., Nummert, B. and Ceulemans, K. (2016) ‘Elucidating the relationship between Sustainability Reporting and Organisational Change Management for Sustainability’, *Journal of Cleaner Production*, 125, pp. 168–188. Available at: https://doi.org/10.1016/j.jclepro.2016.03.021.

Mahajan, R. *et al.* (2023) ‘Stakeholder theory ’, *Journal of Business Research*, 166.

Markopoulos, E., Al Katheeri, H. and Al Qayed, H. (2023) ‘A decision support system architecture for the development and implementation of ESG strategies at SMEs’, in *Proceedings of the 6th International Conference on Intelligent Human Systems Integration (IHSI 2023) Integrating People and Intelligent Systems, February 22–24, 2023, Venice, Italy*. AHFE International. Available at: https://doi.org/10.54941/ahfe1002916.

Matsumura, E.M., Prakash, R. and Vera-Muñoz, S.C. (2014) ‘Firm-Value Effects of Carbon Emissions and Carbon Disclosures’, *The Accounting Review*, 89(2), pp. 695–724. Available at: https://doi.org/10.2308/accr-50629.

Di Matteo, E. *et al.* (2021) ‘Development of a decision support system framework for cultural heritage management’, *Sustainability (Switzerland)*, 13(13). Available at: https://doi.org/10.3390/su13137070.

Mattiussi, A., Rosano, M. and Simeoni, P. (2014) ‘A decision support system for sustainable energy supply combining multi-objective and multi-attribute analysis: An Australian case study’, *Decision Support Systems*, 57, pp. 150–159. Available at: https://doi.org/10.1016/j.dss.2013.08.013.

Ministry of Investment, Trade and Industry. (2023) *i-ESG National Industry Environmental, Social & Governance Framework Phase 1.0:"Just Transition"*.

Modapothala, J.R., Issac, B. and Jayamani, E. (2010) ‘Appraising the corporate sustainability reports - Text mining and multi-discriminatory analysis’, in *Innovations in Computing Sciences and Software Engineering*. Kluwer Academic Publishers, pp. 489–494. Available at: https://doi.org/10.1007/978-90-481-9112-3_83.

Mohr, K., Riquelme, M. and Quick, P.M. (2022) *Double Materiality for Financial Institutions Survey Findings and Recommendations*.

Moldan, B., Janoušková, S. and Hák, T. (2012) ‘How to understand and measure environmental sustainability: Indicators and targets’, *Ecological Indicators*, 17, pp. 4–13. Available at: https://doi.org/10.1016/j.ecolind.2011.04.033.

Moses, L.T. (2022) *Sustainability Reporting Compliance and Financial Performance of Companies Listed On the Nigeria Stock Exchange*, *European Journal of Accounting, Auditing and Finance Research*. Available at: https://www.eajournals.org/.

Nadar, F. (2023) ‘Views: Critical gaps remain in adoption of sustainability in Malaysia’, *The Edge*, 21 June.

Nadi, A. and Moradi, H. (2019) ‘Increasing the views and reducing the depth in random forest’, *Expert Systems with Applications*, 138. Available at: https://doi.org/10.1016/j.eswa.2019.07.018.

Ni, J. *et al.* (2023) ‘Paradigm Shift in Sustainability Disclosure Analysis: Empowering Stakeholders with CHATREPORT, a Language Model-Based Tool’. Available at: http://arxiv.org/abs/2306.15518.

Nilashi, M. *et al.* (2019) ‘Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach’, *Journal of Cleaner Production*, 240. Available at: https://doi.org/10.1016/j.jclepro.2019.118162.

Nishant, R., Kennedy, M. and Corbett, J. (2020) ‘Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda’, *International Journal of Information Management*, 53, p. 102104. Available at: https://doi.org/10.1016/j.ijinfomgt.2020.102104.

Okafor, A., Adusei, M. and Adeleye, B.N. (2021) ‘Corporate social responsibility and financial performance: Evidence from U.S tech firms’, *Journal of Cleaner Production*. Elsevier Ltd. Available at: https://doi.org/10.1016/j.jclepro.2021.126078.

Oladipupo Ayodele, T. (2010) *Introduction to Machine Learning 1*. Available at: www.intechopen.com.

Oncioiu, I. *et al.* (2020) 'Corporate Sustainability Reporting and Financial Performance', *Sustainability*, 12(10), p. 4297. Available at: https://doi.org/10.3390/su12104297.

Oprean-Stan, C. *et al.* (2020) 'Impact of sustainability reporting and inadequate management of esg factors on corporate performance and sustainable growth', *Sustainability (Switzerland)*, 12(20), pp. 1–31. Available at: https://doi.org/10.3390/su12208536.

Pérez, L. *et al.* (2022) *Does ESG really matter - and why?*, *McKinsey & Company*. Available at: https://www.mckinsey.com/capabilities/sustainability/our-insights/does-esg-really-matter-and-why#/ (Accessed: 30 August 2024).

Pislaru, M., Herghiligiu, I.V. and Robu, I.-B. (2019) 'Corporate sustainable performance assessment based on fuzzy logic', *Journal of Cleaner Production*, 223, pp. 998–1013. Available at: https://doi.org/10.1016/j.jclepro.2019.03.130.

Power, D.J. (2002) *Decision Support Systems: Concepts and Resources for Managers*. UNI Scholar Works, University of Northern Iowa.

Rodrigues, M. and Franco, M. (2019) 'The Corporate Sustainability Strategy in Organisations: A Systematic Review and Future Directions', *Sustainability*, 11(22), p. 6214. Available at: https://doi.org/10.3390/su11226214.

Sariyer, G. and Taşkın, D. (2022) 'Clustering of firms based on environmental, social, and governance ratings: Evidence from BIST sustainability index', *Borsa Istanbul Review*. Borsa Istanbul Anonim Sirketi, pp. S180–S188. Available at: https://doi.org/10.1016/j.bir.2022.10.009.

Sarker, I.H. (2021) 'Machine Learning: Algorithms, Real-World Applications and Research Directions', *SN Computer Science*. Springer. Available at: https://doi.org/10.1007/s42979-021-00592-x.

Segura, L.C. *et al.* (2024) 'ESG Dimensions and Corporate Value: Insights for Sustainable Investments', *Sustainability*, 16(17), p. 7376. Available at: https://doi.org/10.3390/su16177376.

Shahi, A.M., Issac, B. and Modapothala, J.R. (2012) *Intelligent Corporate Sustainability Report Scoring Solution Using Machine Learning Approach to Text Categorization*. IEEE.

Shin, S.J. *et al.* (2017) 'Developing a decision support system for improving sustainability performance of manufacturing processes', *Journal of Intelligent Manufacturing*, 28(6), pp. 1421–1440. Available at: https://doi.org/10.1007/s10845-015-1059-z.

Smith, P.A.C. and Sharicz, C. (2011) 'The shift needed for sustainability', *The Learning Organization*, 18(1), pp. 73–86. Available at: https://doi.org/10.1108/09696471111096019.

Sreepriya, J., Suprabha, K.R. and Prasad, K. (2023) ‘Does GRI compliance moderate the impact of sustainability disclosure on firm value?’, *Society and Business Review*, 18(1), pp. 152–174. Available at: https://doi.org/10.1108/SBR-06-2022-0172.

Stolowy, H. and Paugam, L. (2023) ‘Sustainability Reporting: Is Convergence Possible?’, *Accounting in Europe*, 20(2), pp. 139–165. Available at: https://doi.org/10.1080/17449480.2023.2189016.

Tawiah, V. and Borgi, H. (2022) ‘Impact of XBRL adoption on financial reporting quality: a global evidence’, *Accounting Research Journal*, 35(6), pp. 815–833. Available at: https://doi.org/10.1108/ARJ-01-2022-0002.

Thayaraj, M.S. and Karunarathne, W.V.A.D. (2021) ‘The Impact of Sustainability Reporting on Firms’ Financial Performance’, *Journal of Business and Technology*, 5(2), pp. 51–73. Available at: https://doi.org/10.4038/jbt.v5i2.33.

Twinamatsiko, E. and Kumar, D. (2022) ‘Incorporating ESG in Decision Making for Responsible and Sustainable Investments using Machine Learning’, in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, pp. 1328–1334. Available at: https://doi.org/10.1109/ICEARS53579.2022.9752343.

UN Global Compact Network Malaysia & Brunei (2022) *Malaysia Businesses Sustainability Pulse Report 2022*.

United Nations (2015) *The 17 Goals, Department of Economic and Social Affairs, United Nations*.

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

Vivas, R. *et al.* (2019) ‘Measuring sustainability performance with multi criteria model: A case study’, *Sustainability (Switzerland)*, 11(21). Available at: https://doi.org/10.3390/su11216113.

Wagenhofer, A. (2024) ‘Sustainability Reporting: A Financial Reporting Perspective’, *Accounting in Europe*, 21(1), pp. 1–13. Available at: https://doi.org/10.1080/17449480.2023.2218398.

Wang, M.C. (2017) ‘The relationship between firm characteristics and the disclosure of sustainability reporting’, *Sustainability (Switzerland)*, 9(4). Available at: https://doi.org/10.3390/su9040624.

Wang, Z. and Gao, S.S. (2012) ‘Are XBRL-based Financial Reports Better than Non-XBRL Reports? A Quality Assessment’, *International Journal of Economics and Management Engineering*, 6(4). Available at: http://216.241.101.197/viewer.

Webersinke, N. *et al.* (2021) ‘ClimateBert: A Pretrained Language Model for Climate-Related Text’. Available at: http://arxiv.org/abs/2110.12010.

Windolph, S.E. (2011) ‘Assessing Corporate Sustainability Through Ratings: Challenges and Their Causes’, *Journal of Environmental Sustainability*, 1(1), pp. 1–22. Available at: https://doi.org/10.14448/jes.01.0005.

Yang, S., Liu, F.-C. and Zhu, X. (2018) ‘The Impact of XBRL on Financial Statement Structural Comparability’, in *Network, Smart and Open*, pp. 193–206. Available at: https://doi.org/10.1007/978-3-319-62636-9_13.

Zarte, M., Pechmann, A. and Nunes, I.L. (2019) ‘Decision support systems for sustainable manufacturing surrounding the product and production life cycle – A literature review’, *Journal of Cleaner Production*, 219, pp. 336–349. Available at: https://doi.org/10.1016/j.jclepro.2019.02.092.

Zumente, I. and Lāce, N. (2021) ‘Esg rating—necessity for the investor or the company?’, *Sustainability (Switzerland)*, 13(16). Available at: https://doi.org/10.3390/su13168940.

APPENDICES

APPENDIX A: Sample 1 of the Sustainability Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	GEI	DGE	APE	WstG	DWst	WstD	DwWR	Ems	RsU	Invt	Env
2	414000	237600	6.2733	183	16	167	0.0018	8145	116536	50198	0.7498	5680	1	840	703	137	0.8369	89	98	34	70
3	579209	24927	20.04	431	0	431	0.0143	11732	37555	49287	1.6349	-1263	1	11292	11101	191	0.9831	98	95	50	80
4	9E+06	2E+07	18.859	19130	16980	2150	0.0469	523000	194000	717000	1.7563	200	1	678300	584400	93900	0.8616	58	99	86	81
5	4E+07	226703	39.884	91471	87269	4202	0.0934	222295	6E+06	6E+06	5.8783	463162	0	742488	730624	11864	0.984	88	67	28	63
6	102207	2753	60.505	457.41	126.04	331.36	1.8404	14.624	19703	17428	96.766	2290.2	0	131.31	121.68	9.63	0.9267	44	28	0	22

APPENDIX B: Sample 2 of the Sustainability Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	GEI	DGE	APE	WstG	DWst	WstD	DvWR	Ems	RsU	Invt	Env
2	414000	237600	6.273347	183	16	167	0.001818	8145	116536	50198	0.74983	5680	1	840	703	137	0.836905	89	98	34	70
3	579209	24927	20.03977	431	0	431	0.014297	11732	37555	49287	1.634897	-1263	1	11292	11101	191	0.983085	98	95	50	80
4	9068400	18637200	18.85896	19130	16980	2150	0.046856	523000	194000	717000	1.756281	200	1	678300	584400	93900	0.861566	58	99	86	81
5	38838433	226703.1	39.88393	91471	87269	4202	0.093388	222295	5535324	5757620	5.878298	463162	0	742488.3	730624.1	11864.17	0.984021	88	67	28	63
6	102206.5	2752.999	60.50548	457.406	126.043	331.363	1.840384	14.624	19703.15	17427.54	96.76591	2290.231	0	131.305	121.675	9.63	0.926659	44	28	0	22
7	44103.1	0	80.74385	22.448	0	22.448	0.041098	45.69	8383.7	8426.39	15.42702	-5303.39	0	231.783	167.0222	64.76079	0.720597	50	29	17	28
8	439375	0	324.5015	713.687	0	713.687	0.460443	314	80286	80600	0.052	-1272	1	965.3	670	295.3	0.694085	85	64	0	46
9	3030.556	2500	0.016139	45726	35775	9971	0.029096	1190885	347558	1538443	4.489279	30000	1	282134.5	265206.5	16928.07	0.94	95	100	78	90
10	2988000	1440	90.29833	3980000	3208	772	0.43313	65500	230500	296000	1.756281	12300	1	13200	7200	6000	0.545455	77	97	43	64
11	23271.26	0	51.37598	30293	0	30293	50.43899	0	4313.92	4313.92	9.523843	-580.35	1	145.7586	0	145.7586	0	44	35	0	19
12																					

APPENDIX C: Sample 3 of the Sustainability Data

NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	GEI	DGE	APE	WstG	DWst	WstD	DvWR	Ems	RsU	Invt	Env
414000	237600	6.273347	183	16	167	0.001818	8145	116536	50198	0.74983	5680	1	840	703	137	0.836905	89	98	34	70
579209	24927	20.03977	431	0	431	0.014297	11732	37555	49287	1.634897	-1263	1	11292	11101	191	0.983085	98	95	50	80
9068400	18637200	18.85896	19130	16980	2150	0.046856	523000	194000	717000	1.756281	200	1	678300	584400	93900	0.861566	58	99	86	81
38838433	226703.1	39.88393	91471	87269	4202	0.093388	222295	5535324	5757620	5.878298	463162	0	742488.3	730624.1	11864.17	0.984021	88	67	28	63
102206.5	2752.999	60.50548	457.406	126.043	331.363	1.840384	14.624	19703.15	17427.54	96.76591	2290.231	0	131.305	121.675	9.63	0.926659	44	28	0	22
44103.1	0	80.74385	22.448	0	22.448	0.041098	45.69	8383.7	8426.39	15.42702	-5303.39	0	231.783	167.0222	64.76079	0.720597	50	29	17	28
439375	0	324.5015	713.687	0	713.687	0.460443	314	80286	80600	0.052	-1272	1	965.3	670	295.3	0.694085	85	64	0	46
3030.556	2500	0.016139	45726	35775	9971	0.029096	1190885	347558	1538443	4.489279	30000	1	282134.5	265206.5	16928.07	0.94	95	100	78	90
2988000	1440	90.29833	3980000	3208	772	0.43313	65500	230500	296000	1.756281	12300	1	13200	7200	6000	0.545455	77	97	43	64
23271.26	0	51.37598	30293	0	30293	50.43899	0	4313.92	4313.92	9.523843	-580.35	1	145.7586	0	145.7586	0	44	35	0	19
6306523	107208	127.1036	15806	12689	3117	0.476474	353771	719066	1072837	21.26086	66155	1	18812.3	17257.38	1554.915	0.917346	91	74	89	85
72828000	7884000	241.2055	175520	70920	104600	0.312594	2018789	9539765	10982818	32.82184	-1536322	1	744019	684497.5	59521.52	0.92	99	98	34	73
13488318	1894759	174.5062	24228	18824	5403	0.061292	1110000	1060000	2170000	24.61656	-90000	1	50673	46073	5317	0.909222	97	97	95	97
775121	0	435.012	3674000	2881000	793000	0.000446	1800	144800	146600	0.08236	-9700	1	1954	1044	910	0.534289	53	69	0	37
21778.2	3008.16	3.927615	39.5	0	39.5	0.051573	12.3	3848.6	3860.9	5.040998	534	1	134.7024	59.63504	75.0674	0.442717	66	48	34	48

APPENDIX D: Development of the Random Forest Regressor with Sample 1

In [24]:
```
import pandas as pd
data=pd.read_csv(r"C:\Users\User\Desktop\ESG 1.csv")
print(data)
```

```
            NRE           RE         EI        WW        WD       WC  \
0  4.140000e+05  2.376000e+05   6.273347    183.000     16.000   167.000
1  5.792090e+05  2.492700e+04  20.039772    431.000      0.000   431.000
2  9.068400e+06  1.863720e+07  18.858957  19130.000  16980.000  2150.000
3  3.883843e+07  2.267031e+05  39.883929  91471.000  87269.000  4202.000
4  1.022065e+05  2.752999e+03  60.505477    457.406    126.043   331.363

         WI         GE1          GE2          TGE  ...         DGE  APE  \
0  0.001818    8145.000   116536.000     50198.000  ...    5680.000    1
1  0.014297   11732.000    37555.000     49287.000  ...   -1263.000    1
2  0.046856  523000.000   194000.000    717000.000  ...     200.000    1
3  0.093388  222295.000  5535324.000   5757620.000  ...  463162.000    0
4  1.840384      14.624    19703.148     17427.541  ...    2290.231    0

         WstG         DWst         WstD      DvWR  Ems  RsU  Invt  Env
0     840.0000     703.0000    137.00000  0.836905   89   98    34   70
1   11292.0000   11101.0000    191.00000  0.983085   98   95    50   80
2  678300.0000  584400.0000  93900.00000  0.861566   58   99    86   81
3  742488.2848  730624.1194  11864.16543  0.984021   88   67    28   63
4     131.3050     121.6750      9.63000  0.926659   44   28     0   22

[5 rows x 21 columns]
```

In [25]:
```
data.describe()
```

Out[25]:

	NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	...	DGE	APE	WstG	DWst	WstD	DvWR	Ems	RsU	Invt	Env
count	5.000000e+00	5.000000e+00	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.000000e+00	5.000000e+00	...	5.00000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000
mean	9.800450e+06	3.825837e+06	29.112296	22334.481200	20878.208600	1456.272600	0.399349	153037.32480	1.180624e+06	1.318307e+06	...	94013.84620	0.600000	286610.317960	265389.958880	21220.359086	0.918447	75.400000	77.400000	39.600000	63.20000
std	1.666533e+07	8.280528e+06	21.275685	39494.252616	37830.929582	1731.986288	0.806341	226961.41422	2.435337e+06	2.498993e+06	...	206376.57606	0.547723	387550.191383	361697.233087	40946.628571	0.067879	23.147354	30.615356	31.603797	24.20124
min	1.022065e+05	2.752999e+03	6.273347	183.000000	0.000000	167.000000	0.001818	14.62400	1.970315e+04	1.742754e+04	...	-1263.00000	0.000000	131.305000	121.675000	9.630000	0.836905	44.000000	28.000000	0.000000	22.00000
25%	4.140000e+05	2.492700e+04	18.858957	431.000000	16.000000	331.363000	0.014297	8145.00000	3.755500e+04	4.928700e+04	...	200.00000	0.000000	840.000000	703.000000	137.000000	0.861566	58.000000	67.000000	28.000000	63.00000
50%	5.792090e+05	2.267031e+05	20.039772	457.406000	126.043000	431.000000	0.046856	11732.00000	1.165360e+05	5.019800e+04	...	2290.23100	1.000000	11292.000000	11101.000000	191.000000	0.926659	88.000000	95.000000	34.000000	70.00000
75%	9.068400e+06	2.376000e+05	39.883929	19130.000000	16980.000000	2150.000000	0.093388	222295.00000	1.940000e+05	7.170000e+05	...	5680.00000	1.000000	678300.000000	584400.000000	11864.165430	0.983085	89.000000	98.000000	50.000000	80.00000
max	3.883843e+07	1.863720e+07	60.505477	91471.000000	87269.000000	4202.000000	1.840384	523000.00000	5.535324e+06	5.757620e+06	...	463162.00000	1.000000	742488.284800	730624.119400	93900.000000	0.984021	98.000000	99.000000	86.000000	81.00000

8 rows × 21 columns

In [26]:
```
X=data.drop(['Ems', 'RsU', 'Invt', 'Env'], axis = 1)
```

In [27]:
```
X.shape
```

Out[27]:
```
(5, 17)
```

```
In [28]: print(data[:6])
```

```
            NRE           RE         EI        WW        WD       WC  \
0  4.140000e+05  2.376000e+05   6.273347    183.000    16.000   167.000
1  5.792090e+05  2.492700e+04  20.039772    431.000     0.000   431.000
2  9.068400e+06  1.863720e+07  18.858957  19130.000 16980.000  2150.000
3  3.883843e+07  2.267031e+05  39.883929  91471.000 87269.000  4202.000
4  1.022065e+05  2.752999e+03  60.505477    457.406   126.043   331.363

         WI        GE1         GE2         TGE  ...        DGE  APE  \
0  0.001818   8145.000   116536.000    50198.000  ...    5680.000    1
1  0.014297  11732.000    37555.000    49287.000  ...   -1263.000    1
2  0.046856 523000.000   194000.000   717000.000  ...     200.000    1
3  0.093388 222295.000  5535324.000  5757620.000  ...  463162.000    0
4  1.840384     14.624    19703.148    17427.541  ...    2290.231    0

          WstG          DWst         WstD      DVWR  Ems  RsU  Invt  Env
0     840.0000      703.0000    137.00000  0.836905   89   98    34   70
1   11292.0000    11101.0000    191.00000  0.983085   98   95    50   80
2  678300.0000   584400.0000  93900.00000  0.861566   58   99    86   81
3  742488.2848   730624.1194  11864.16543  0.984021   88   67    28   63
4     131.3050      121.6750      9.63000  0.926659   44   28     0   22

[5 rows x 21 columns]
```

```
In [29]: print(X[:6])
```

```
            NRE           RE         EI        WW        WD       WC  \
0  4.140000e+05  2.376000e+05   6.273347    183.000    16.000   167.000
1  5.792090e+05  2.492700e+04  20.039772    431.000     0.000   431.000
2  9.068400e+06  1.863720e+07  18.858957  19130.000 16980.000  2150.000
3  3.883843e+07  2.267031e+05  39.883929  91471.000 87269.000  4202.000
4  1.022065e+05  2.752999e+03  60.505477    457.406   126.043   331.363

         WI        GE1         GE2         TGE        GEI         DGE  APE  \
0  0.001818   8145.000   116536.000    50198.000   0.749830    5680.000    1
1  0.014297  11732.000    37555.000    49287.000   1.634897   -1263.000    1
2  0.046856 523000.000   194000.000   717000.000   1.756281     200.000    1
3  0.093388 222295.000  5535324.000  5757620.000   5.878298  463162.000    0
4  1.840384     14.624    19703.148    17427.541  96.765913    2290.231    0

          WstG          DWst         WstD      DVWR
0     840.0000      703.0000    137.00000  0.836905
1   11292.0000    11101.0000    191.00000  0.983085
2  678300.0000   584400.0000  93900.00000  0.861566
3  742488.2848   730624.1194  11864.16543  0.984021
4     131.3050      121.6750      9.63000  0.926659
```

```
In [30]: y = data['Ems']
         from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.40)
```

```
In [31]: print (X_train.size, X_test.size, y_train.size, y_test.size)
```

```
51 34 3 2
```

```
In [32]: print (X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(3, 17) (2, 17) (3,) (2,)
```

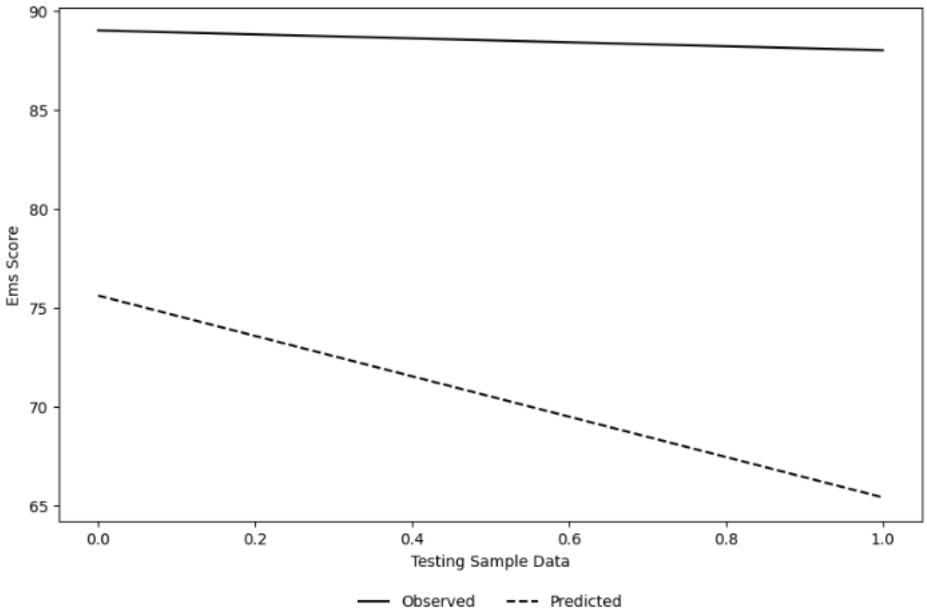
```
In [33]: from sklearn.ensemble import RandomForestRegressor

In [34]: model = RandomForestRegressor(n_estimators = 30, random_state = 30)

In [35]: Ems1_rf = model.fit(X_train, y_train)

In [36]: X.dtypes

Out[36]: NRE     float64
         RE      float64
         EI      float64
         WW      float64
         WD      float64
         WC      float64
         WI      float64
         GE1     float64
         GE2     float64
         TGE     float64
         GEI     float64
         DGE     float64
         APE       int64
         WstG    float64
         DWst    float64
         WstD    float64
         DvWR    float64
         dtype: object

In [37]: from sklearn.metrics import mean_absolute_error, mean_squared_error, explained_variance_score, r2_score

In [39]: print('The training r_sq is: %.2f'% Ems1_rf.score(X_train, y_train))

The training r_sq is: 0.64

In [40]: # r_sq - how well the prediction align the datapoints

In [43]: ytrain_pred = Ems1_rf.predict(X_train)

In [44]: print('The MAE is: %.2f'% mean_absolute_error(y_train, ytrain_pred))

The MAE is: 12.87

In [45]: # MAE for Model Performance

In [46]: print ('The MSE is :%.2f'% mean_squared_error(y_train, ytrain_pred))

The MSE is :189.08

In [47]: import numpy as np
         print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_train, ytrain_pred)))

The RMSE is:13.75

In [48]:  print('The EVS is :%.2f'% explained_variance_score(y_train, ytrain_pred))

The EVS is :0.66

In [49]: ytest_pred = Ems1_rf.predict(X_test)

In [50]: print(ytest_pred[:6])

[75.6 65.4]

In [51]: print(ytest_pred[:10])

[75.6 65.4]
```

In [53]:
```
print (y_test[:6])
```
```
0    89
3    88
Name: Ems, dtype: int64
```

In [55]:
```
print('The testing r_sq is: %.2f'% r2_score(y_test, ytest_pred))
```
```
The testing r_sq is: -1379.64
```

In [54]:
```
print('The MAE is: %.2f'% mean_absolute_error(y_test, ytest_pred))
print ('The MSE is:%.2f'% mean_squared_error(y_test, ytest_pred))
print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_test, ytest_pred)))
print('The EVS is :%.2f'% explained_variance_score(y_test, ytest_pred))
```
```
The MAE is: 18.00
The MSE is:345.16
The RMSE is:18.58
The EVS is :-83.64
```

In [56]:
```
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (10,6)
x_ax = range(len(X_test))
plt.plot(x_ax, y_test, label = 'Observed', color = 'k', linestyle = '-')
plt.plot(x_ax, ytest_pred, label = 'Predicted', color = 'k', linestyle = '--')
plt.ylabel('Ems Score')
plt.xlabel('Testing Sample Data')
plt.legend(bbox_to_anchor = (0.5, -0.2), loc = 'lower center', ncol = 2, frameon = False)
plt.show()
```



APPENDIX E: Development of the Random Forest Regressor with Sample 2

In [1]:
```
import pandas as pd
data=pd.read_csv(r"C:\Users\User\Desktop\ESG 2.csv")
print(data)
```

```
           NRE           RE          EI          WW         WD         WC  \
0  4.140000e+05  2.376000e+05    6.273347      183.000     16.000    167.000
1  5.792090e+05  2.492700e+04   20.039772      431.000      0.000    431.000
2  9.068400e+06  1.863720e+07   18.858957    19130.000  16980.000   2150.000
3  3.883843e+07  2.267031e+05   39.883929    91471.000  87269.000   4202.000
4  1.022065e+05  2.752999e+03   60.505477      457.406    126.043    331.363
5  4.410310e+04  0.000000e+00   80.743853       22.448      0.000     22.448
6  4.393750e+05  0.000000e+00  324.501477      713.687      0.000    713.687
7  3.030556e+03  2.500000e+03    0.016139    45726.000  35775.000   9971.000
8  2.988000e+06  1.440000e+03   90.298330  3980000.000   3208.000    772.000
9  2.327126e+04  0.000000e+00   51.375980    30293.000      0.000  30293.000

          WI          GE1          GE2          TGE  ...         DGE  APE  \
0   0.001818     8145.000   116536.000    50198.000  ...    5680.000    1
1   0.014297    11732.000    37555.000    49287.000  ...   -1263.000    1
2   0.046856   523000.000   194000.000   717000.000  ...     200.000    1
3   0.093388   222295.000  5535324.000  5757620.000  ...  463162.000    0
4   1.840384       14.624    19703.148    17427.541  ...    2290.231    0
5   0.041098       45.690     8383.700     8426.390  ...   -5303.390    0
6   0.460443      314.000    80286.000    80600.000  ...   -1272.000    1
7   0.029096  1190885.000   347558.000  1538443.000  ...   30000.000    1
8   0.433130    65500.000   230500.000   296000.000  ...   12300.000    1
9  50.438988        0.000     4313.920     4313.920  ...    -580.350    1

            WstG          DWst          WstD      DvWR  Ems  RsU  Invt  Env
0     840.000000     703.00000    137.000000  0.836905   89   98    34   70
1   11292.000000   11101.00000    191.000000  0.983085   98   95    50   80
2  678300.000000  584400.00000  93900.000000  0.861566   58   99    86   81
3  742488.284800  730624.11940  11864.165430  0.984021   88   67    28   63
4     131.305000     121.67500      9.630000  0.926659   44   28     0   22
5     231.782999     167.02221     64.760790  0.720597   50   29    17   28
6     965.300000     670.00000    295.300000  0.694085   85   64     0   46
7  282134.535000  265206.46290  16928.072100  0.940000   95  100    78   90
8   13200.000000    7200.00000   6000.000000  0.545455   77   97    43   64
9     145.758625       0.00000    145.758625  0.000000   44   35     0   19

[10 rows x 21 columns]
```

In [2]:
```
data.describe()
```

Out[2]:

	NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	...
count	1.000000e+01	1.000000e+01	10.000000	1.000000e+01	10.000000	10.000000	10.000000	1.000000e+01	1.000000e+01	1.000000e+01	...
mean	5.250003e+06	1.913312e+06	69.249726	4.168428e+05	14337.404300	4905.349800	5.339950	2.021931e+05	6.574160e+05	8.519316e+05	...
std	1.213203e+07	5.876942e+06	94.696627	1.252312e+06	28163.698408	9429.096315	15.856093	3.854038e+05	1.717599e+06	1.791110e+06	...
min	3.030556e+03	0.000000e+00	0.016139	2.244800e+01	0.000000	22.448000	0.001818	0.000000e+00	4.313920e+03	4.313920e+03	...
25%	5.862896e+04	3.600000e+02	19.154161	4.376015e+02	0.000000	356.272250	0.032096	1.127675e+02	2.416611e+04	2.539241e+04	...
50%	4.266875e+05	2.626500e+03	45.629955	9.921844e+03	71.021500	742.843500	0.070122	9.938500e+03	9.841100e+04	6.539900e+04	...
75%	2.385802e+06	1.762591e+05	75.684259	4.186775e+04	13537.000000	3689.000000	0.453615	1.830962e+05	2.213750e+05	6.117500e+05	...
max	3.883843e+07	1.863720e+07	324.501477	3.980000e+06	87269.000000	30293.000000	50.438988	1.190885e+06	5.535324e+06	5.757620e+06	...

	DGE	APE	WstG	DWst	WstD	DvWR	Ems	RsU	Invt	Env
count	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	50521.349100	0.700000	172972.896642	160019.327951	12953.568694	0.749237	72.800000	71.200000	33.600000	56.300000
std	145340.781368	0.483046	296664.211236	276844.046882	29064.576782	0.299022	21.565404	30.875377	31.213245	25.978838
min	-5303.390000	0.000000	131.305000	0.000000	9.630000	0.000000	44.000000	28.000000	0.000000	19.000000
25%	-1092.337500	0.250000	383.837249	292.766657	139.189656	0.700713	52.000000	42.250000	4.250000	32.500000
50%	1245.115500	1.000000	6128.650000	3951.500000	243.150000	0.849235	81.000000	81.000000	31.000000	63.500000
75%	10645.000000	1.000000	214900.901250	201680.097175	10398.124072	0.936665	88.750000	97.750000	48.250000	77.500000
max	463162.000000	1.000000	742488.284800	730624.119400	93900.000000	0.984021	98.000000	100.000000	86.000000	90.000000

8 rows × 21 columns

```
In [3]: X=data.drop(['Ems', 'RsU', 'Invt', 'Env'], axis = 1)
```

```
In [4]: X.shape
```

```
Out[4]: (10, 17)
```

```
In [43]: print(data[:10])
```

```
            NRE           RE          EI          WW         WD        WC  \
0  4.140000e+05  2.376000e+05    6.273347     183.000     16.000   167.000
1  5.792090e+05  2.492700e+04   20.039772     431.000      0.000   431.000
2  9.068400e+06  1.863720e+07   18.858957   19130.000  16980.000  2150.000
3  3.883843e+07  2.267031e+05   39.883929   91471.000  87269.000  4202.000
4  1.022065e+05  2.752999e+03   60.505477     457.406    126.043   331.363
5  4.410310e+04  0.000000e+00   80.743853      22.448      0.000    22.448
6  4.393750e+05  0.000000e+00  324.501477     713.687      0.000   713.687
7  3.030556e+03  2.500000e+03    0.016139   45726.000  35775.000  9971.000
8  2.988000e+06  1.440000e+03   90.298330 3980000.000   3208.000   772.000
9  2.327126e+04  0.000000e+00   51.375980   30293.000      0.000 30293.000

          WI          GE1          GE2          TGE  ...        DGE  APE  \
0   0.001818     8145.000   116536.000     50198.000  ...    5680.000    1
1   0.014297    11732.000    37555.000     49287.000  ...   -1263.000    1
2   0.046856   523000.000   194000.000    717000.000  ...     200.000    1
3   0.093388   222295.000  5535324.000   5757620.000  ...  463162.000    0
4   1.840384       14.624    19703.148     17427.541  ...    2290.231    0
5   0.041098       45.690     8383.700      8426.390  ...   -5303.390    0
6   0.460443      314.000    80286.000     80600.000  ...   -1272.000    1
7   0.029096  1190885.000   347558.000   1538443.000  ...   30000.000    1
8   0.433130    65500.000   230500.000    296000.000  ...   12300.000    1
9  50.438988        0.000     4313.920      4313.920  ...    -580.350    1

            WstG          DWst          WstD      DvWR  Ems  RsU  Invt  Env
0     840.000000     703.00000    137.000000  0.836905   89   98    34   70
1   11292.000000   11101.00000    191.000000  0.983085   98   95    50   80
2  678300.000000  584400.00000  93900.000000  0.861566   58   99    86   81
3  742488.284800  730624.11940  11864.165430  0.984021   88   67    28   63
4     131.305000     121.67500      9.630000  0.926659   44   28     0   22
5     231.782999     167.02221     64.760790  0.720597   50   29    17   28
6     965.300000     670.00000    295.300000  0.694085   85   64     0   46
7  282134.535000  265206.46290  16928.072100  0.940000   95  100    78   90
8   13200.000000    7200.00000   6000.000000  0.545455   77   97    43   64
9     145.758625       0.00000    145.758625  0.000000   44   35     0   19

[10 rows x 21 columns]
```

```
In [44]: print(X[:10])
```

```
            NRE           RE          EI          WW         WD         WC  \
0  4.140000e+05  2.376000e+05    6.273347     183.000     16.000    167.000
1  5.792090e+05  2.492700e+04   20.039772     431.000      0.000    431.000
2  9.068400e+06  1.863720e+07   18.858957   19130.000  16980.000   2150.000
3  3.883843e+07  2.267031e+05   39.883929   91471.000  87269.000   4202.000
4  1.022065e+05  2.752999e+03   60.505477     457.406    126.043    331.363
5  4.410310e+04  0.000000e+00   80.743853      22.448      0.000     22.448
6  4.393750e+05  0.000000e+00  324.501477     713.687      0.000    713.687
7  3.030556e+03  2.500000e+03    0.016139   45726.000  35775.000   9971.000
8  2.988000e+06  1.440000e+03   90.298330 3980000.000   3208.000    772.000
9  2.327126e+04  0.000000e+00   51.375980   30293.000      0.000  30293.000

          WI         GE1          GE2          TGE        GEI         DGE  \
0   0.001818     8145.000   116536.000    50198.000   0.749830    5680.000
1   0.014297    11732.000    37555.000    49287.000   1.634897   -1263.000
2   0.046856   523000.000   194000.000   717000.000   1.756281     200.000
3   0.093388   222295.000  5535324.000  5757620.000   5.878298  463162.000
4   1.840384       14.624    19703.148    17427.541  96.765913    2290.231
5   0.041098       45.690     8383.700     8426.390  15.427015   -5303.390
6   0.460443      314.000    80286.000    80600.000   0.052000   -1272.000
7   0.029096  1190885.000   347558.000  1538443.000   4.489279   30000.000
8   0.433130    65500.000   230500.000   296000.000   1.756281   12300.000
9  50.438988        0.000     4313.920     4313.920   9.523843    -580.350

   APE          WstG         DWst          WstD      DvWR
0    1     840.000000    703.00000    137.000000  0.836905
1    1   11292.000000  11101.00000    191.000000  0.983085
2    1  678300.000000  584400.00000  93900.000000  0.861566
3    0  742488.284800  730624.11940  11864.165430  0.984021
4    0     131.305000    121.67500      9.630000  0.926659
5    0     231.782999    167.02221     64.760790  0.720597
6    1     965.300000    670.00000    295.300000  0.694085
7    1  282134.535000  265206.46290  16928.072100  0.940000
8    1   13200.000000   7200.00000   6000.000000  0.545455
9    1     145.758625      0.00000    145.758625  0.000000
```

```
In [7]: y = data['Ems']
        from sklearn.model_selection import train_test_split
        X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.40)
```

```
In [8]: print (X_train.size, X_test.size, y_train.size, y_test.size)
```

```
102 68 6 4
```

```
In [9]: print (X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(6, 17) (4, 17) (6,) (4,)
```

```
In [10]: from sklearn.ensemble import RandomForestRegressor
```

```
In [11]: model = RandomForestRegressor(n_estimators = 30, random_state = 30)
```

```
In [12]: Ems2_rf = model.fit(X_train, y_train)
```

```
In [13]: X.dtypes

Out[13]: NRE     float64
         RE      float64
         EI      float64
         WW      float64
         WD      float64
         WC      float64
         WI      float64
         GE1     float64
         GE2     float64
         TGE     float64
         GEI     float64
         DGE     float64
         APE       int64
         WstG    float64
         DWst    float64
         WstD    float64
         DVWR    float64
         dtype: object

In [14]: from sklearn.metrics import mean_absolute_error, mean_squared_error, explained_variance_score, r2_score

In [15]: print('The training r_sq is: %.2f'% Ems2_rf.score(X_train, y_train))

The training r_sq is: 0.81

In [16]: # r_sq - how well the prediction align the datapoints

In [17]: ytrain_pred = Ems2_rf.predict(X_train)

In [18]: print('The MAE is: %.2f'% mean_absolute_error(y_train, ytrain_pred))

The MAE is: 5.79

In [19]: # MAE for Model Performance

In [20]: print ('The MSE is :%.2f'% mean_squared_error(y_train, ytrain_pred))

The MSE is :80.21

In [21]: import numpy as np
         print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_train, ytrain_pred)))

The RMSE is:8.96

In [22]:  print('The EVS is :%.2f'% explained_variance_score(y_train, ytrain_pred))

The EVS is :0.84

In [23]: ytest_pred = Ems2_rf.predict(X_test)

In [24]: print(ytest_pred[:6])

[50.7        52.43333333 48.46666667 51.16666667]

In [25]: print(ytest_pred[:10])

[50.7        52.43333333 48.46666667 51.16666667]

In [26]: print (y_test[:6])

8    77
1    98
6    85
0    89
Name: Ems, dtype: int64
```

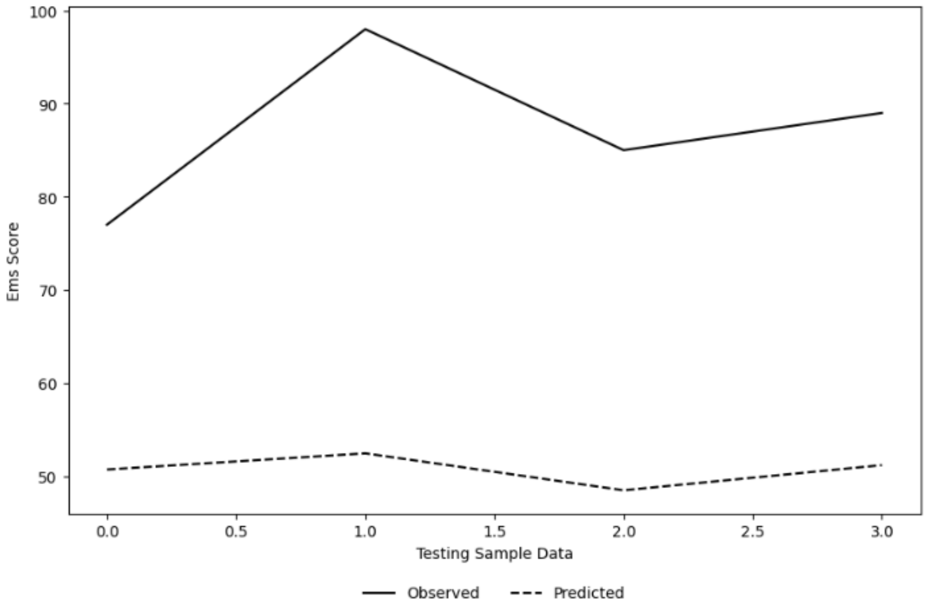
In [27]:
```
print('The testing r_sq is: %.2f'% r2_score(y_test, ytest_pred))
```

```
The testing r_sq is: -23.19
```

In [28]:
```
print('The MAE is: %.2f'% mean_absolute_error(y_test, ytest_pred))
print ('The MSE is:%.2f'% mean_squared_error(y_test, ytest_pred))
print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_test, ytest_pred)))
print('The EVS is :%.2f'% explained_variance_score(y_test, ytest_pred))
```

```
The MAE is: 36.56
The MSE is:1383.51
The RMSE is:37.20
The EVS is :0.18
```

In [29]:
```
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (10,6)
x_ax = range(len(X_test))
plt.plot(x_ax, y_test, label = 'Observed', color = 'k', linestyle = '-')
plt.plot(x_ax, ytest_pred, label = 'Predicted', color = 'k', linestyle = '--')
plt.ylabel('Ems Score')
plt.xlabel('Testing Sample Data')
plt.legend(bbox_to_anchor = (0.5, -0.2), loc = 'lower center', ncol = 2, frameon = False)
plt.show()
```



APPENDIX F: Development of the Random Forest Regressor with Sample 3

In [1]:
```
import pandas as pd
data=pd.read_csv(r"C:\Users\User\Desktop\ESG 3.csv")
print(data)
```

```
             NRE           RE          EI          WW          WD  \
0   4.140000e+05  2.376000e+05    6.273347     183.000      16.000
1   5.792090e+05  2.492700e+04   20.039772     431.000       0.000
2   9.068400e+06  1.863720e+07   18.858957   19130.000   16980.000
3   3.883843e+07  2.267031e+05   39.883929   91471.000   87269.000
4   1.022065e+05  2.752999e+03   60.505477     457.406     126.043
5   4.410310e+04  0.000000e+00   80.743853      22.448       0.000
6   4.393750e+05  0.000000e+00  324.501477     713.687       0.000
7   3.030556e+03  2.500000e+03    0.016139   45726.000   35775.000
8   2.988000e+06  1.440000e+03   90.298330 3980000.000    3208.000
9   2.327126e+04  0.000000e+00   51.375980   30293.000       0.000
10  6.306523e+06  1.072080e+05  127.103578   15806.000   12689.000
11  7.282800e+07  7.884000e+06  241.205526  175520.000   70920.000
12  1.348832e+07  1.894759e+06  174.506216   24228.000   18824.000
13  7.751210e+05  0.000000e+00  435.012049 3674000.000 2881000.000
14  2.177820e+04  3.008160e+03    3.927615      39.500       0.000

            WC         WI          GE1          GE2           TGE  ...  \
0      167.000   0.001818     8145.000   116536.000  5.019800e+04  ...
1      431.000   0.014297    11732.000    37555.000  4.928700e+04  ...
2     2150.000   0.046856   523000.000   194000.000  7.170000e+05  ...
3     4202.000   0.093388   222295.000  5535324.000  5.757620e+06  ...
4      331.363   1.840384       14.624    19703.148  1.742754e+04  ...
5       22.448   0.041098       45.690     8383.700  8.426390e+03  ...
6      713.687   0.460443      314.000    80286.000  8.060000e+04  ...
7     9971.000   0.029096  1190885.000   347558.000  1.538443e+06  ...
8      772.000   0.433130    65500.000   230500.000  2.960000e+05  ...
9    30293.000  50.438988        0.000     4313.920  4.313920e+03  ...
10    3117.000   0.476474   353771.000   719066.000  1.072837e+06  ...
11  104600.000   0.312594  2018789.000  9539765.000  1.098282e+07  ...
12    5403.000   0.061292  1110000.000  1060000.000  2.170000e+06  ...
13  793000.000   0.000446     1800.000   144800.000  1.466000e+05  ...
14      39.500   0.051573       12.300     3848.600  3.860900e+03  ...

            DGE  APE           WstG           DWst          WstD     DvWR  \
0      5680.000    1     840.000000     703.000000    137.000000  0.836905
1     -1263.000    1   11292.000000   11101.000000    191.000000  0.983085
2       200.000    1  678300.000000  584400.000000  93900.000000  0.861566
3    463162.000    0  742488.284800  730624.119400  11864.165430  0.984021
4      2290.231    0     131.305000     121.675000      9.630000  0.926659
5     -5303.390    0     231.782999     167.022210     64.760790  0.720597
6     -1272.000    1     965.300000     670.000000    295.300000  0.694085
7     30000.000    1  282134.535000  265206.462900  16928.072100  0.940000
8     12300.000    1   13200.000000    7200.000000   6000.000000  0.545455
9      -580.350    1     145.758625       0.000000    145.758625  0.000000
10    66155.000    1   18812.295350   17257.380260   1554.915090  0.917346
11 -1536322.000    1  744019.000000  684497.480000  59521.520000  0.920000
12   -90000.000    1   50673.000000   46073.000000   5317.000000  0.909222
13    -9700.000    1    1954.000000    1044.000000    910.000000  0.534289
14      534.000    1     134.702442      59.635042     75.067400  0.442717
```

```
    Ems  RsU  Invt  Env
0    89   98    34   70
1    98   95    50   80
2    58   99    86   81
3    88   67    28   63
4    44   28     0   22
5    50   29    17   28
6    85   64     0   46
7    95  100    78   90
8    77   97    43   64
9    44   35     0   19
10   91   74    89   85
11   99   98    34   73
12   97   97    95   97
13   53   69     0   37
14   66   48    34   48

[15 rows x 21 columns]
```

In [2]:
```
data.describe()
```

Out[2]:

	NRE	RE	EI	WW	WD	WC	WI	GE1	GE2	TGE	...	DGE	APE	WstG	DWst	WstD	DvWR	Ems	RsU	Invt	
count	1.500000e+01	1.500000e+01	15.000000	1.500000e+01	1.500000e+01	15.000000	15.000000	1.500000e+01	1.500000e+01	1.500000e+01	...	1.500000e+01	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.00
mean	9.727985e+06	1.934807e+06	111.616816	5.372014e+05	2.084538e+05	63680.866533	3.620125	3.670869e+05	1.202776e+06	1.526362e+06	...	-7.094130e+04	0.800000	169688.130948	156608.318321	13127.612629	0.747730	75.600000	73.200000	39.200000	60.20
std	2.019021e+07	5.050181e+06	129.787286	1.337729e+06	7.398410e+05	203547.304699	12.960499	6.069314e+05	2.697115e+06	3.020022e+06	...	4.236749e+05	0.414039	294698.629962	273719.282874	27084.545091	0.270959	20.890189	27.326596	34.086445	25.41
min	3.030556e+03	0.000000e+00	0.016139	2.244800e+01	0.000000e+00	22.448000	0.000446	0.000000e+00	3.848600e+03	3.860900e+03	...	-1.536322e+06	0.000000	131.305000	0.000000	9.630000	0.000000	44.000000	28.000000	0.000000	19.00
25%	7.315482e+04	7.200000e+02	19.449365	4.442030e+02	0.000000e+00	381.181500	0.035097	1.798450e+02	2.862907e+04	3.335727e+04	...	-3.287695e+03	1.000000	535.891500	418.511105	141.379312	0.619770	55.500000	56.000000	8.500000	41.50
50%	5.792090e+05	3.008160e+03	60.505477	1.913000e+04	3.208000e+03	2150.000000	0.061292	1.173200e+04	1.448000e+05	1.466000e+05	...	2.000000e+02	1.000000	11292.000000	7200.000000	910.000000	0.861566	85.000000	74.000000	34.000000	64.00
75%	7.687462e+06	2.321515e+05	150.804897	6.859850e+04	2.729950e+04	7687.000000	0.446786	4.383855e+05	5.333120e+05	1.305640e+06	...	8.990000e+03	1.000000	166403.767500	155639.731450	8932.082715	0.923330	93.000000	97.500000	64.000000	80.50
max	7.282800e+07	1.863720e+07	435.012049	3.980000e+06	2.881000e+06	793000.000000	50.438988	2.018789e+06	9.539765e+06	1.098282e+07	...	4.631620e+05	1.000000	744019.000000	730624.119400	93900.000000	0.984021	99.000000	100.000000	95.000000	97.00

; rows × 21 columns

In [3]:
```
X=data.drop(['Ems', 'RsU', 'Invt', 'Env'], axis = 1)
```

In [4]:
```
X.shape
```

Out[4]:
```
(15, 17)
```

```
In [18]: print(data[:15])
```

```
             NRE           RE          EI          WW           WD  \
0   4.140000e+05  2.376000e+05    6.273347     183.000       16.000
1   5.792090e+05  2.492700e+04   20.039772     431.000        0.000
2   9.068400e+06  1.863720e+07   18.858957   19130.000    16980.000
3   3.883843e+07  2.267031e+05   39.883929   91471.000    87269.000
4   1.022065e+05  2.752999e+03   60.505477     457.406      126.043
5   4.410310e+04  0.000000e+00   80.743853      22.448        0.000
6   4.393750e+05  0.000000e+00  324.501477     713.687        0.000
7   3.030556e+03  2.500000e+03    0.016139   45726.000    35775.000
8   2.988000e+06  1.440000e+03   90.298330 3980000.000     3208.000
9   2.327126e+04  0.000000e+00   51.375980   30293.000        0.000
10  6.306523e+06  1.072080e+05  127.103578   15806.000    12689.000
11  7.282800e+07  7.884000e+06  241.205526  175520.000    70920.000
12  1.348832e+07  1.894759e+06  174.506216   24228.000    18824.000
13  7.751210e+05  0.000000e+00  435.012049 3674000.000  2881000.000
14  2.177820e+04  3.008160e+03    3.927615      39.500        0.000

            WC         WI          GE1          GE2           TGE  ...  \
0      167.000   0.001818     8145.000   116536.000  5.019800e+04  ...
1      431.000   0.014297    11732.000    37555.000  4.928700e+04  ...
2     2150.000   0.046856   523000.000   194000.000  7.170000e+05  ...
3     4202.000   0.093388   222295.000  5535324.000  5.757620e+06  ...
4      331.363   1.840384       14.624    19703.148  1.742754e+04  ...
5       22.448   0.041098       45.690     8383.700  8.426390e+03  ...
6      713.687   0.460443      314.000    80286.000  8.060000e+04  ...
7     9971.000   0.029096  1190885.000   347558.000  1.538443e+06  ...
8      772.000   0.433130    65500.000   230500.000  2.960000e+05  ...
9    30293.000  50.438988        0.000     4313.920  4.313920e+03  ...
10    3117.000   0.476474   353771.000   719066.000  1.072837e+06  ...
11  104600.000   0.312594  2018789.000  9539765.000  1.098282e+07  ...
12    5403.000   0.061292  1110000.000  1060000.000  2.170000e+06  ...
13  793000.000   0.000446     1800.000   144800.000  1.466000e+05  ...
14      39.500   0.051573       12.300     3848.600  3.860900e+03  ...

            DGE  APE           WstG           DWst          WstD      DvWR  \
0      5680.000    1     840.000000     703.000000    137.000000  0.836905
1     -1263.000    1   11292.000000   11101.000000    191.000000  0.983085
2       200.000    1  678300.000000  584400.000000  93900.000000  0.861566
3    463162.000    0  742488.284800  730624.119400  11864.165430  0.984021
4      2290.231    0     131.305000     121.675000      9.630000  0.926659
5     -5303.390    0     231.782999     167.022210     64.760790  0.720597
6     -1272.000    1     965.300000     670.000000    295.300000  0.694085
7     30000.000    1  282134.535000  265206.462900  16928.072100  0.940000
8     12300.000    1   13200.000000    7200.000000   6000.000000  0.545455
9      -580.350    1     145.758625       0.000000    145.758625  0.000000
10    66155.000    1   18812.295350   17257.380260   1554.915090  0.917346
11 -1536322.000    1  744019.000000  684497.480000  59521.520000  0.920000
12   -90000.000    1   50673.000000   46073.000000   5317.000000  0.909222
13    -9700.000    1    1954.000000    1044.000000    910.000000  0.534289
14      534.000    1     134.702442      59.635042     75.067400  0.442717

    Ems  RsU  Invt  Env
0    89   98    34   70
1    98   95    50   80
2    58   99    86   81
3    88   67    28   63
4    44   28     0   22
5    50   29    17   28
6    85   64     0   46
7    95  100    78   90
8    77   97    43   64
9    44   35     0   19
10   91   74    89   85
11   99   98    34   73
12   97   97    95   97
13   53   69     0   37
```

In [19]:
```
print(X[:15])
```

```
             NRE           RE          EI          WW          WD  \
0   4.140000e+05  2.376000e+05    6.273347     183.000      16.000
1   5.792090e+05  2.492700e+04   20.039772     431.000       0.000
2   9.068400e+06  1.863720e+07   18.858957   19130.000   16980.000
3   3.883843e+07  2.267031e+05   39.883929   91471.000   87269.000
4   1.022065e+05  2.752999e+03   60.505477     457.406     126.043
5   4.410310e+04  0.000000e+00   80.743853      22.448       0.000
6   4.393750e+05  0.000000e+00  324.501477     713.687       0.000
7   3.030556e+03  2.500000e+03    0.016139   45726.000   35775.000
8   2.988000e+06  1.440000e+03   90.298330 3980000.000    3208.000
9   2.327126e+04  0.000000e+00   51.375980   30293.000       0.000
10  6.306523e+06  1.072080e+05  127.103578   15806.000   12689.000
11  7.282800e+07  7.884000e+06  241.205526  175520.000   70920.000
12  1.348832e+07  1.894759e+06  174.506216   24228.000   18824.000
13  7.751210e+05  0.000000e+00  435.012049 3674000.000 2881000.000
14  2.177820e+04  3.008160e+03    3.927615      39.500       0.000

            WC         WI          GE1          GE2           TGE        GEI  \
0      167.000   0.001818     8145.000   116536.000  5.019800e+04   0.749830
1      431.000   0.014297    11732.000    37555.000  4.928700e+04   1.634897
2     2150.000   0.046856   523000.000   194000.000  7.170000e+05   1.756281
3     4202.000   0.093388   222295.000  5535324.000  5.757620e+06   5.878298
4      331.363   1.840384       14.624    19703.148  1.742754e+04  96.765913
5       22.448   0.041098       45.690     8383.700  8.426390e+03  15.427015
6      713.687   0.460443      314.000    80286.000  8.060000e+04   0.052000
7     9971.000   0.029096  1190885.000   347558.000  1.538443e+06   4.489279
8      772.000   0.433130    65500.000   230500.000  2.960000e+05   1.756281
9    30293.000  50.438988        0.000     4313.920  4.313920e+03   9.523843
10    3117.000   0.476474   353771.000   719066.000  1.072837e+06  21.260857
11  104600.000   0.312594  2018789.000  9539765.000  1.098282e+07  32.821840
12    5403.000   0.061292  1110000.000  1060000.000  2.170000e+06  24.616563
13  793000.000   0.000446     1800.000   144800.000  1.466000e+05   0.082360
14      39.500   0.051573       12.300     3848.600  3.860900e+03   5.040998

            DGE  APE           WstG           DWst          WstD      DvWR
0      5680.000    1     840.000000     703.000000    137.000000  0.836905
1     -1263.000    1   11292.000000   11101.000000    191.000000  0.983085
2       200.000    1  678300.000000  584400.000000  93900.000000  0.861566
3    463162.000    0  742488.284800  730624.119400  11864.165430  0.984021
4      2290.231    0     131.305000     121.675000      9.630000  0.926659
5     -5303.390    0     231.782999     167.022210     64.760790  0.720597
6     -1272.000    1     965.300000     670.000000    295.300000  0.694085
7     30000.000    1  282134.535000  265206.462900  16928.072100  0.940000
8     12300.000    1   13200.000000    7200.000000   6000.000000  0.545455
9      -580.350    1     145.758625       0.000000    145.758625  0.000000
10    66155.000    1   18812.295350   17257.380260   1554.915090  0.917346
11 -1536322.000    1  744019.000000  684497.480000  59521.520000  0.920000
12   -90000.000    1   50673.000000   46073.000000   5317.000000  0.909222
13    -9700.000    1    1954.000000    1044.000000    910.000000  0.534289
14      534.000    1     134.702442      59.635042     75.067400  0.442717
```

In [7]:
```
y = data['Ems']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.40)
```

In [8]:
```
print (X_train.size, X_test.size, y_train.size, y_test.size)
```

```
153 102 9 6
```

In [9]:
```
print (X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(9, 17) (6, 17) (9,) (6,)
```

```
In [10]: from sklearn.ensemble import RandomForestRegressor

In [11]: model = RandomForestRegressor(n_estimators = 30, random_state = 30)

In [12]: Ems3_rf = model.fit(X_train, y_train)

In [13]: X.dtypes

Out[13]: NRE     float64
         RE      float64
         EI      float64
         WW      float64
         WD      float64
         WC      float64
         WI      float64
         GE1     float64
         GE2     float64
         TGE     float64
         GEI     float64
         DGE     float64
         APE       int64
         WstG    float64
         DWst    float64
         WstD    float64
         DvWR    float64
         dtype: object

In [14]: from sklearn.metrics import mean_absolute_error, mean_squared_error, explained_variance_score, r2_score

In [15]: print('The training r_sq is: %.2f'% Ems3_rf.score(X_train, y_train))

        The training r_sq is: 0.86

In [16]: # r_sq - how well the prediction align the datapoints

In [20]: ytrain_pred = Ems3_rf.predict(X_train)

In [21]: print('The MAE is: %.2f'% mean_absolute_error(y_train, ytrain_pred))

        The MAE is: 7.31

In [22]: # MAE for Model Performance

In [23]: print ('The MSE is :%.2f'% mean_squared_error(y_train, ytrain_pred))

        The MSE is :69.68

In [24]: import numpy as np
         print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_train, ytrain_pred)))

        The RMSE is:8.35

In [25]:  print('The EVS is :%.2f'% explained_variance_score(y_train, ytrain_pred))

        The EVS is :0.86
```

```
In [26]: ytest_pred = Ems3_rf.predict(X_test)
```

```
In [27]: print(ytest_pred[:6])
```

```
[53.3        86.83333333 78.63333333 53.7        63.16666667 61.2       ]
```

```
In [28]: print(ytest_pred[:10])
```

```
[53.3        86.83333333 78.63333333 53.7        63.16666667 61.2       ]
```

```
In [29]: print (y_test[:6])
```

```
14    66
3     88
7     95
9     44
8     77
6     85
Name: Ems, dtype: int64
```

```
In [30]: print('The testing r_sq is: %.2f'% r2_score(y_test, ytest_pred))
```

```
The testing r_sq is: 0.25
```

```
In [31]: print('The MAE is: %.2f'% mean_absolute_error(y_test, ytest_pred))
         print ('The MSE is:%.2f'% mean_squared_error(y_test, ytest_pred))
         print('The RMSE is:%.2f'% np.sqrt(mean_squared_error(y_test, ytest_pred)))
         print('The EVS is :%.2f'% explained_variance_score(y_test, ytest_pred))
```

```
The MAE is: 12.93
The MSE is:213.73
The RMSE is:14.62
The EVS is :0.58
```

```
In [32]: import matplotlib.pyplot as plt
         plt.rcParams['figure.figsize'] = (10,6)
         x_ax = range(len(X_test))
         plt.plot(x_ax, y_test, label = 'Observed', color = 'k', linestyle = '-')
         plt.plot(x_ax, ytest_pred, label = 'Predicted', color = 'k', linestyle = '--')
         plt.ylabel('Ems Score')
         plt.xlabel('Testing Sample Data')
         plt.legend(bbox_to_anchor = (0.5, -0.2), loc = 'lower center', ncol = 2, frameon = False)
         plt.show()
```