Impacts of Economic Complexity and HDI on Marine and Terrestrial Ecosystem: A Machine Learning Approach

How do economic complexity and human development impact environmental sustainability, specifically within the contexts of SDGs 14 and 15?

by

Muhammad Nabeel Numan and Indronil Kundu

Lund University - Data Analytics and Business Economics (DABE)
DABN01
Master's Thesis (15 credits ECTS)
May 2024
Supervisor: Behnaz Pirzamanbein

Declaration

I declare that this thesis is my own work and has not been submitted elsewhere in any form for another degree or diploma. Where other sources of information have been used, they have been acknowledged. This thesis employed OpenAI's GPT to enhance writing clarity, grammar, maintain coherence across contributions from multiple authors, and assist in debugging code. If any information has been inadvertently not referenced, it is purely an error.

Abstract

This thesis delves into the intricate interplay between economic complexity (ECI), human development (HDI), and environmental sustainability, with a special focus on Sustainable Development Goals (SDGs) 14 (Life Below Water) and 15 (Life on Land). By leveraging advanced machine learning models—including Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN)—we predict environmental outcomes and illuminate the dynamics driving these relationships. Our study seamlessly integrates theoretical frameworks, empirical findings, and SHAP (SHapley Additive exPlanations) analysis to offer a nuanced understanding of how ECI and HDI impact environmental indicators such as forest cover, forest biomass, mountain key biodiversity areas (KBAs), beach litter, chlorophyll, and sustainable fisheries.

Our findings underscore the exceptional performance of Random Forest models, particularly in predicting SDG 15 indicators. The SHAP analysis reveals a significant positive impact of ECI on forest cover and biomass, lending credence to the Environmental Kuznets Curve (EKC) hypothesis. In the realm of SDG 14, the influence of ECI and HDI varies across different indicators, with especially notable effects observed in chlorophyll and sustainable fisheries.

The implications of our research are profound, highlighting the necessity for integrated policies that foster economic complexity and human development while safeguarding the environment. This study offers invaluable insights for policymakers striving to achieve sustainable development by balancing economic growth, human development, and environmental sustainability. Through the application of sophisticated machine learning models, our research presents a comprehensive approach to understanding and addressing the multifaceted challenges of sustainable development.

Contents

List of Figures

List of Tables

1 Introduction

Over the past few decades, the world has undergone rapid transformations, which are still ongoing without a doubt. As the world keeps evolving, the rate of change has become more accelerated. It took several decades for humankind to progress from the first Industrial Revolution, characterized by steam power and mechanization, towards the second Industrial Revolution, marked by electricity and mass production (Horvath, B., 2018). However, it took less time to transition from the second into the third Industrial Revolution, driven by electronics and information technology in the late 20th century. And then comes the most recent and swift leap forward: the fourth Industrial Revolution, often termed Industry 4.0, characterized by robotics, artificial intelligence, and beyond (Horvath, B., 2018). Evidently, the pace of change has quickened with each successive leap forward, pushing the boundaries of possibilities, yet simultaneously introducing many new challenges.

While social, economic, political, and technological development redefine and evolve the way humankind interacts with itself and with the world around it, the aftermath of development is not always auspicious. In 1962, Rachel Carson wrote *Silent Spring*, which sheds light on the interaction between human activity and nature. The book highlights that the relationship between humans and the environment can sometimes resemble a parasitic relationship where human actions harm or exploit the environment. A few years later, in 1987, the United Nations (UN), in its report *Our Common Future*, introduced the concept of sustainable development and defined it as a development that "meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland, G., 1987).

1.1 Evolution of Sustainable Development

The concept of sustainable development led to a fundamental shift in how humans view growth and progress. Barbier, E.B. (1987) illustrated the model of sustainable development using three dimensions: Economy, Society, and Environment (shown in Figure 1). Undeniably, the relationship between these three dimensions is complex and multifaceted. Arguably, humankind has had a shallow definition of growth and development for a long time, one that neglected the well-being of human (societal dimension) (Ul Haq, M., 1995) and well-being of envrionment (Horvath, B., 2018).

Over time, as the understanding of sustainable development has evolved, economists, researchers, institutions, and nations have acknowledged that traditional growth measures need to be improved to reflect true progress that is more inclusive of the societal and environmental dimensions. Mahbub Ul Haq (1995), in *Reflections on Human Development*, emphasizes a pivotal perspective that true development is ob-

tained through investing in people and their well-being rather than solely chasing monetary economic growth. Mahbub Ul Haq laid the foundations for human development theory and reshaped how we measure socio-economic prosperity through the Human Development Index (HDI) (Stanton, E.A., 2007). Haq elucidated that in many societies, gross national income (GNP) can increase while human lives deteriorate, highlighting the futility of traditional economic growth indicators that fail to account for social and human welfare (Stanton, E.A., 2007).

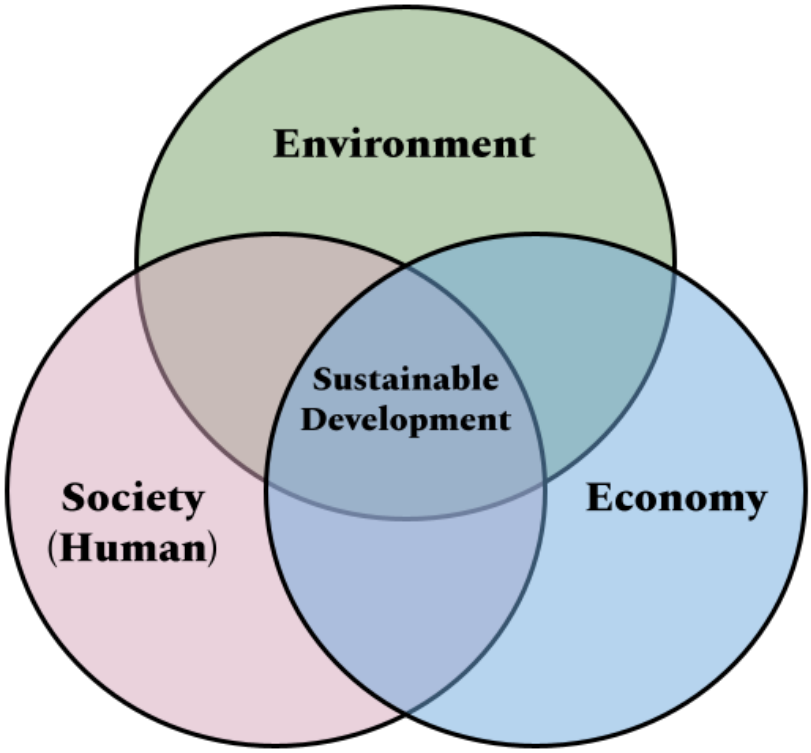

Figure 1: Sustainable development model by Barbier, E.B. (1987) (Source: (Numan, N. M., 2022)

Recognition of these limitations, has lead to newer indicators been developed to provide a more comprehensive view of socio-economic progress. In addition to the Human Development Index (HDI) and various other measures, the Economic Complexity Index (ECI), developed by Hidalgo C.A. and Hausmann, R. in 2009, stands out. They aimed to create a measure that evaluates the diversification of an economy, reflecting its underlying production capabilities and knowledge base (Hidalgo, C.A., and Hausmann, R., 2009). Similarly, advancements have been made in evaluating environmental sustainability, with the ecological footprint (EF) emerging as one of the most significant indicator. EF measures "how fast we consume resources and generate waste compared to how fast nature can absorb our waste and generate resources" (Global Footprint Network, 2024).

Moreover, in recent decades, significant global efforts have been made to promote sustainable development, locally and globally, attempting to rectify the mistakes made by human activities in the past and the present. The UN launched the Millennium Development Goals (MDGs), set from 2000 to 2015, to engage in a global war against "hunger, disease, unmet schooling, gender inequality, and environmental degradation" (Sachs, J.D., 2012). In 2015, the effort and scope were expanded by the UN to achieve sustainable development as it adopted Agenda 2030, which sets forth the Sustainable Development Goals (SDGs) (Katila, P. et al., 2019). SDGs consist of 17 goals and 169 sub-goals (targets) which aim to provide a pathway for sustainable development inclusively by addressing human development, social development, economic development , and environmental challenges (Katila, P. et al., 2019).

1.2 Research Objective and Significance

However, achieving the SDGs and their targets is a complex task. This is because the SDGs, much like the three dimensions of sustainable development—economic, social, and environmental—are deeply interconnected, and this interconnectedness presents both trade-offs and synergies (Numan, M. N., 2022). As stated previously, that Barbier E.B. (1987) depicted sustainable development using three-circles to illustrate the interplay between economic, social, and environmental dimensions. While effective in showing their interconnectedness, it misses the dynamic complexities of their relationships. Numan's (2022) conceptualized "Relational sustainable development model", which is a conceptual model addresses these limitations by positioning these dimensions at the vertices of a triangle with bidirectional arrows, symbolizing their continuous and reciprocal interactions, shown in in 2. In this refined model, the environment serves as the sustaining dimension while the social and economic dimension regarded as developing. This foundational role emphasizes the necessity of environmental preservation for long-term sustainability and highlights that true development must align with environmental preservation (Numan, M. N., 2022).

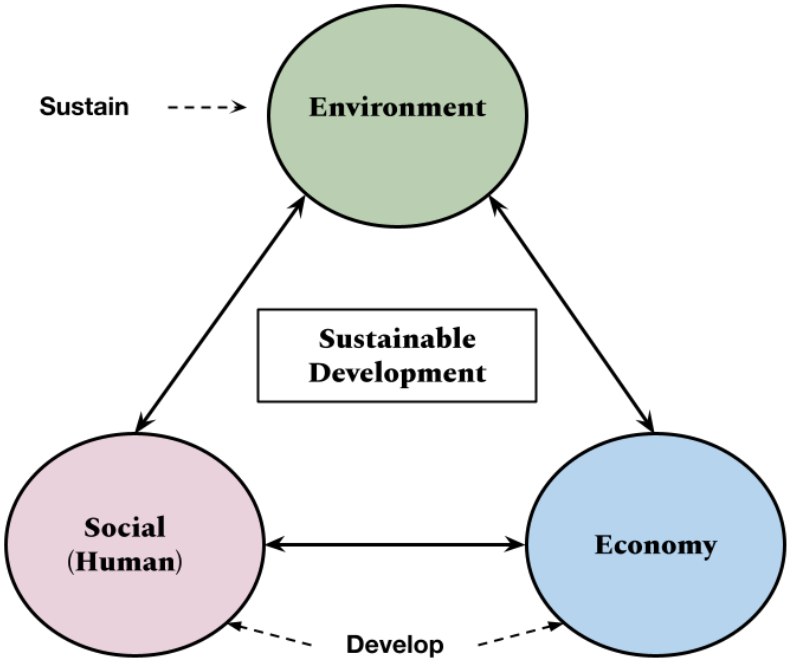

Figure 2: Relational Sustainable Development Model (Numan, N. M., 2022)

This conceptual framework illustrates that while balancing all dimensions is ideal, inherent synergies, trade-offs, and feedback loops make this challenging. Therefore, it emphasizes a strategic approach where sustainable development aims to advance economic and social dimensions in harmony with environmental preservation, which is of critical importance in achieving long-term sustainability. The SDGs share this preservation focused perspective, advocating for progress that ensures environmental sustainability is not compromised as encapsulated by the inclusion of SDG 13 (Climate Action), SDG 14 (Life Below Water) and SDG 15 (Life on Land), shown by 3. Among these goals, SDG 14 (Life Below Water) and SDG 15 (Life on Land) are critical for conserving the natural environment. SDG 14 focuses on conserving and responsibly managing marine ecosystems and their resources, while SDG 15 focuses on protecting terrestrial ecosystems and preventing their degradation and biodiversity loss (Katila, P. et al., 2019).

Figure 3: Agenda 2030: Sustainable Development Goals (The Global Goals., 2024)

Katila P. et al. (2019) describes that SDGs do not explicitly address the interlinkages, synergies, and trade-offs among targets. however, the pursuit of one goal or target can directly or indirectly influence other goals, either positively or negatively (Katila, P. et al., 2019). To effectively monitor and achieve these goals, it is essential to understand the intricate relationships between economic activities, human development, and environmental health. There is a growing need to delve into the interactions among these dimensions, particularly between the sustaining dimension (environment) and the developing dimensions (economic and social). By doing so, we can better understand how to balance these aspects to achieve sustainable outcomes.

In this context, the Economic Complexity Index (ECI) and the Human Development Index (HDI) provide valuable insights into these relationships.The ECI measures the complexity and knowledge intensity of an economy (Hausmann, R. et al., 2014), while the HDI assesses human development through "the inclusion of proxies for three important ends of development: access to health, education, and goods." (Stanton, E.A., 2007, p.3). By examining how these indices interact with environmental sustainability indicators, we aim to understand the impact of economic and social development on marine and terrestrial ecosystems.

To date, much of the literature has focused on analyzing relationships through various statistical techniques, often looking at isolated impacts of economic or social variables on environmental outcomes. While these methods have provided significant

insights, they sometimes fall short in capturing the dynamic and complex nature of these interactions. Predictive modeling offers a different approach by focusing on forecasting outcomes based on current and historical data, thereby identifying patterns and trends that traditional statistical methods may overlook.

1.2.1 Research Question

The primary objective of this research is to explore how economic complexity and human development influences the achievement of SDG 14 (Life Below Water) and SDG 15 (Life on Land). This involves investigating the relationships between economic and social progress and environmental health. Advanced analytical techniques are used to develop predictive models that identify key patterns and relationships within the data. These models help us understand which factors are most strongly associated with environmental outcomes.

The focus on predictive modeling, through the use of advanced machine learning (ML) methods, provides robust estimates and predictions that enhance our understanding of these interactions. This can reveal underlying patterns and provide actionable insights that may not be evident through traditional analysis. The objective of this approach is to leverage data-driven insights to uncover complex relationships, enabling a comprehensive examination of how economic complexity and human development can be harmonized with environmental sustainability, offering valuable perspectives for policymakers.

This lead us to our research question:

> ***How do economic complexity and human development influence environmental sustainability, particularly in the context of achieving SDG 14 (Life Below Water) and SDG 15 (Life on Land), and what key factors drive these relationships?***

Conclusively, by addressing this research question, the study aims to provide a comprehensive understanding of the interplay between economic, social, and environmental dimensions of sustainable development within the framework of the SDGs. The insights gained from this research will not only contribute to academic knowledge but also offer valuable guidance for policymakers in designing strategies that harmonize economic growth and human development with environmental preservation.

2 Literature Review

The intersection of economic complexity, human development, and environmental sustainability is a critical area of study, particularly in the context of Sustainable Development Goals (SDGs) 14 and 15, which focus on life below water and life on land, respectively. This literature review synthesizes research on the impacts of economic complexity and human development on environmental sustainability, exploring how these factors interplay to influence the achievement of these SDGs. Additionally, it highlights the potential role of machine learning in enhancing sustainability analysis, an area that has been underutilized in existing research.

2.1 Economic Complexity and Environmental Sustainability

Economic complexity, defined as the diversity and sophistication of a country's productive capabilities, has significant implications for environmental sustainability. High economic complexity often leads to more diversified and technologically advanced industries, which can both positively and negatively impact the environment.

Nguyen Van Tran et al. (2021) conducted a comprehensive study examining the relationship between economic complexity and environmental sustainability. They found that countries with higher economic complexity tend to have more efficient production processes, potentially reducing environmental degradation. However, they also noted that this increased efficiency could lead to greater resource extraction and higher emissions due to more intensive industrial activities (Nguyen Van Tran et al., 2021).

Rafique et al. (2021) conducted an in-depth analysis using a panel data approach to investigate how economic complexity affects environmental sustainability across different countries. Their findings suggest that complex economies are better equipped to implement sustainable practices and technologies, which can mitigate the negative environmental impacts of industrial activities. The study highlighted the role of innovation in promoting sustainability, as more complex economies tend to invest more in research and development, leading to more sustainable production methods (Rafique et al., 2021). This research underscores the importance of fostering economic complexity to achieve environmental sustainability, particularly through policy measures that promote innovation and technological advancement.

Mehrjo and Yuzbashkand (2021) explored the relationship between economic complexity, information and communication technology (ICT), biomass energy consumption, and environmental degradation in Iran. Their study used a time-series analysis to assess the impacts of these factors on environmental sustainability. They

found that while higher economic complexity and ICT usage can lead to improved environmental outcomes through more efficient resource use and lower emissions, the positive effects are often offset by increased biomass energy consumption, which can contribute to environmental degradation. This highlights the need for a balanced approach that considers the environmental impacts of different energy sources and promotes cleaner alternatives (Mehrjo and Yuzbashkand, 2021).

A broader context is provided by Rafique et al. (2021), who conducted a world sample study to examine the general trends in the relationship between economic complexity and environmental performance. Their findings indicate that while economic complexity generally promotes better environmental performance, the benefits are not uniformly distributed. Countries with high economic complexity often export their environmental burdens to less complex economies through global supply chains. This phenomenon, known as environmental load displacement, underscores the need for more integrated and equitable global environmental policies that address the cross-border impacts of economic activities (Rafique et al., 2021).

Economic complexity captures a wide variety of components within production, including skill, knowledge, and technological innovation, making it a strong indicator for the Sustainable Development Goals (SDGs). Complex economies have moved from agriculture-based systems to complex industrial frameworks, resulting in a notable increase in energy consumption and greenhouse gas emissions (Rafique et al., 2022). This shift highlights the demand for knowledge-intensive technology and sustainable energy solutions. By adopting energy-efficient products and renewable energy sources, advanced economies, with their varied and complex product bases, use economic complexity to support sustainable growth and environmental protection. As a result, economic complexity promotes information sharing and the adoption of cleaner technologies, which both advance economic development and environmental sustainability (Rafique et al., 2022).

Hausmann and Hidalgo (2011) highlight that economic complexity is determined by the diversity and ubiquity of products in a country's export basket. Research indicates that economies with higher economic complexity tend to adopt knowledge-intensive technologies and produce energy-efficient goods, which can potentially reduce pollution emissions (Hausmann et al., 2014). Studies by Swart and Brinkmann (2020) further support the notion that economic complexity affects ecological footprint through the productive structure, technological adoption, and renewable energy generation. This body of work underscores the multifaceted nature of economic complexity and its significant implications for environmental sustainability (Hausmann and Hidalgo, 2011; Hausmann et al., 2014; Swart and Brinkmann, 2020).

2.2 Human Development and Environmental Sustainability

The Human Development Index (HDI), which measures life expectancy, education, and per capita income, has traditionally been used to assess a country's development. However, its impact on environmental sustainability is complex and multifaceted. Higher HDI levels are generally associated with better environmental awareness and the capacity to implement sustainable practices.

Pelinescu (2020) conducted a study examining the impact of human capital on economic growth and environmental sustainability. The study found that improvements in education and health, key components of HDI, lead to more environmentally conscious behaviors and better implementation of sustainable practices. However, the study also noted that these benefits are more pronounced in developing countries, where improvements in human capital can lead to significant reductions in environmental degradation (Pelinescu, 2020).

Hickel (2020) argues that the traditional HDI model does not account for ecological impacts and often promotes high levels of income that correlate with increased environmental degradation. To address this, Hickel proposed the Sustainable Development Index (SDI), which adjusts HDI by incorporating ecological efficiency, measured by CO2 emissions and material footprint. This approach highlights the importance of integrating ecological considerations into human development metrics to ensure sustainability. Hickel's study underscores the need for a paradigm shift in how development is measured, advocating for an approach that balances human development with ecological sustainability (Hickel, 2020).

Tran et al. (2021) investigated the trade-offs between environment, energy consumption, and human development across different levels of economic development. Their study found that in developing countries, improvements in HDI components such as education and health lead to reduced carbon emissions, as educated populations are more likely to adopt sustainable practices. Conversely, in developed countries, the correlation is less significant due to already established high consumption patterns. This research highlights the differing impacts of human development on environmental sustainability across different stages of economic development, emphasizing the need for tailored policy approaches that consider these differences (Tran et al., 2021).

Charfeddine (2021) examined the ecological footprint of MENA countries, focusing on the influence of socio-political factors alongside economic development. The study found that higher levels of human development can lead to better environmental policies and practices, provided that socio-political structures support sustainability initiatives. This research highlights the importance of good governance and political stability in achieving environmental sustainability, suggesting

that socio-political factors can significantly influence the effectiveness of environmental policies (Charfeddine, 2021).

The literature on development economics has been enhanced by the concept of human development and SD, especially in the last three decades (Costantini and Monni, 2008). The UN recognizes a hybrid of both concepts, i.e., sustainable human development, meaning human development should occur within the SD path. Integrating these two concepts has encouraged researchers to look at their relationship. As EKC is the most commonly used theory when studying SD, Costantini and Monni (2008) studied the EKC for general environmental stress as a function of HDI (excluding income). Their results show that at 0.6 HDI (excluding income), that is, medium development, environmental degradation will reverse. The main explanation is that policy will shift towards environmental friendliness as the environment is no longer a luxury good since the economy is no longer underdeveloped (Costantini and Monni, 2008). The second is that high development leads to technological innovation resulting in a changed productive structure that is more service-oriented, thus producing low emissions (Costantini and Monni, 2008).

2.3 Interconnection between Economic Complexity and Human Development

The relationship between economic complexity and human development is complex and interdependent. Higher economic complexity can lead to improved human development outcomes by creating better employment opportunities and higher income levels.

Pelinescu (2020) highlighted that economic complexity and human development together can drive significant improvements in environmental performance, especially in OECD countries where advanced economic structures and high human development coexist. The study emphasizes the need for policies that simultaneously promote economic complexity and human development to achieve sustainable outcomes (Pelinescu, 2020).

Arica and Kurt (2021) conducted a panel analysis to explore the causal linkages between HDI and economic complexity in selected OECD countries. Their study found a bidirectional relationship between economic complexity and human development, indicating that improvements in one area can lead to advancements in the other. This interdependence suggests that policies aimed at enhancing economic complexity should also focus on improving human development outcomes, and vice versa. The study provides a comprehensive framework for understanding how these factors interact and influence each other, highlighting the importance of integrated policy approaches (Arica and Kurt, 2021).

Zulham (2021) examined the nexus of HDI, economic growth, and environmental degradation in Aceh Province, Indonesia. The study found that rapid economic and population growth can lead to environmental degradation despite improvements in human development. This underscores the need for balanced approaches that enhance economic complexity and human development while mitigating environmental impacts. The study suggests that sustainable development strategies should consider the specific challenges and contexts of different regions to be effective (Zulham, 2021).

Hickel (2020) further elaborates on the need for a sustainable development paradigm that integrates human development with ecological efficiency. The SDI developed by Hickel provides a valuable framework for aligning human development with ecological sustainability, promoting a model of development that is both sustainable and equitable. The SDI can guide policymakers in balancing the trade-offs between development and sustainability, ensuring that progress in human development does not come at the expense of the environment (Hickel, 2020).

Numan (2022) also examines the interlinkages between HDI, ECI, and ecological footprint (EF) per capita, using pairwise linear regression analysis for 114 countries between 1995 and 2017. His findings show that while economic complexity and human development both increase ecological footprint, they are positively correlated, supporting the idea that complex economies with higher human development tend to have larger environmental impacts. This study revises the Environmental Kuznets Curve (EKC) hypothesis, testing it for EF as HDI and ECI change, finding partial support for an inverted U-shaped relationship for EF and ECI but not for HDI (Numan, 2022).

2.4 Role of Machine Learning in Environmental Sustainability

Machine learning (ML) can presents significant opportunities for enhancing environmental sustainability analysis. Rao (2021) conducted a study on the role of natural resources in environmental sustainability, utilizing ML techniques to analyze large datasets and predict environmental impacts. The study found that ML can optimize resource use, predict environmental impacts, and develop more efficient sustainability strategies. Despite its potential, ML remains underutilized in this field. Incorporating ML into sustainability analysis can provide more accurate and actionable insights, helping policymakers design more effective interventions (Rao, 2021).

Carneiro (2021) explored the use of ML in marine management for human development. The study reviewed two decades of scholarly evidence and highlighted how

ML can improve the management of marine resources by predicting fish populations and optimizing fishery practices. Carneiro's research underscores the potential of ML to enhance the sustainability of marine ecosystems, supporting the achievement of SDG 14 (Carneiro, 2021).

Dyck (2021) examined the economic impact of ocean fish populations on the global fishery, suggesting that ML models can enhance our understanding of marine ecosystems and support sustainable fishery management. The study highlighted the importance of accurate data and advanced analytical tools in managing marine resources effectively. By leveraging ML, policymakers can develop more precise and efficient strategies for preserving marine biodiversity and ensuring sustainable fisheries (Dyck, 2021).

The potential of ML in environmental sustainability is further emphasized by the need for more comprehensive and integrated data analysis. ML techniques can handle large and complex datasets, making them ideal for analyzing the multifaceted interactions between economic complexity, human development, and environmental sustainability. Future research should focus on developing and applying ML models to various aspects of sustainability analysis, including predicting environmental impacts, optimizing resource use, and assessing the effectiveness of sustainability policies (Rao, 2021).

2.4.1 Environmental Kuznets Curve (EKC) and Sustainability

The Environmental Kuznets Curve (EKC) hypothesis posits an inverted U-shaped relationship between economic growth and environmental degradation. Initially proposed by Grossman and Krueger in 1991 and later popularized by the World Bank's 1992 World Development Report, the EKC concept suggests that as economies evolve, environmental degradation initially worsens but then improves with economic development (Grossman and Krueger, 1991). This phenomenon is attributed to three main effects: the Scale Effect, where economic expansion leads to a proportional increase in environmental pollution; the Composition Effect, which suggests that as economies transition from agriculture to industrialization and eventually to a service-based model, environmental degradation initially rises but later declines; and the Technique Effect, wherein technological advancements result in more efficient resource use, thereby reducing environmental impact (Stern, 2012). Additionally, the EKC examines whether environmental quality behaves as a normal good, with higher incomes prompting increased spending on environmental protection, or as a luxury good, where environmental concerns become more significant only at higher income levels.

Sasse (2022) delves into the complex dynamics of sustainable development in Latin America and the Caribbean (LAC), particularly through the lens of the EKC

hypothesis. By examining the relationships between renewable energy, economic complexity, and greenhouse gas (GHG) emissions, the research provides insights into the region's environmental performance vis-à-vis its economic development trajectory. The findings indicate that while economic complexity is positively correlated with a reduction in GHG emissions and energy supply, suggesting a potential mitigating effect on environmental degradation, the EKC hypothesis's application is nuanced. Despite expectations of an inverted U-shaped curve, the study reveals that economic activity, represented by GDP, continues to exert pressure on environmental indicators, with emissions rising alongside economic growth. This underscores the need for targeted policy interventions to navigate the trade-off between economic expansion and environmental sustainability in LAC (Sasse, 2022).

Numan (2022) revises the EKC hypothesis by testing it specifically for ecological footprint (EF) as HDI and ECI change. His study employs pairwise linear regression analysis for 114 countries over the period from 1995 to 2017. Numan's findings show partial support for an inverted U-shaped relationship for EF and ECI, suggesting that as economic complexity increases, ecological footprint initially increases but eventually decreases after reaching a certain level of complexity. However, this inverted U-shaped relationship is not observed for HDI, indicating that higher human development does not necessarily lead to a reduction in ecological footprint. This revision of the EKC hypothesis highlights the complex interactions between economic complexity, human development, and environmental impacts, and suggests that different measures of development may yield different environmental outcomes (Numan, 2022).

Sasse's study contributes to the ongoing discourse on sustainable development in LAC by elucidating the intricate relationship between economic complexity, renewable energy adoption, and environmental outcomes within the framework of the EKC hypothesis. While economic complexity shows promise as a driver of sustainable development, the persistence of environmental challenges suggests the need for context-specific strategies to achieve green growth. Moving forward, the study advocates for further research to unravel the region's potential for fostering economic complexity, promoting clean energy innovation, and steering towards more sustainable development trajectories, while acknowledging the complexities and trade-offs inherent in the pursuit of environmental sustainability amidst economic growth (Sasse, 2022).

2.4.2 Achieving SDG 14 and SDG 15

Achieving SDGs 14 (Life Below Water) and 15 (Life on Land) requires a comprehensive approach that integrates economic complexity and human development with environmental sustainability.

Charfeddine (2021) argues that socio-political factors and economic policies must

align to support these goals, emphasizing the importance of regional cooperation and policy integration. The study highlights the need for governance structures that promote sustainable development and address the socio-political challenges that can hinder environmental sustainability (Charfeddine, 2021).

Rafique et al. (2021) highlight the need for global collaboration to address the environmental impacts of complex economies, particularly in managing the transboundary nature of environmental degradation. Their research suggests that international cooperation and policy coordination are essential for mitigating the negative environmental impacts of economic complexity. This includes developing global standards and regulations that ensure sustainable practices across different countries and regions (Rafique et al., 2021).

Hickel's (2020) SDI provides a valuable framework for aligning human development with ecological efficiency, promoting a model of development that is sustainable and equitable. The SDI can guide policymakers in balancing the trade-offs between development and sustainability, ensuring that progress in human development does not come at the expense of the environment. Hickel's research emphasizes the importance of redefining development metrics to include ecological considerations, which is crucial for achieving SDG 14 and SDG 15 (Hickel, 2020).

Abbas (2021) conducted a study on the integration of economic complexity, tourism, energy prices, and environmental degradation in highly complex economies. The research highlights the need for targeted policies that address the specific challenges of complex economies, including the environmental impacts of tourism and energy consumption. Abbas suggests that leveraging economic complexity to enhance sustainable practices and mitigate environmental impacts is essential for achieving SDG 14 and SDG 15 (Abbas, 2021).

Whereas, these studies do not directly address SDGs, however, they illustrate that the underscore the need for integrated global policies that consider the interactions between economic complexity, human development, and environmental sustainability. This is useful for backing up the purpose of this research which is that by adopting a holistic approach that includes advanced analytical tools such as ML, policymakers can better achieve critical global objectives, such as SDG 14 and SDG 15, in pursuit of sustainable development.

2.5 Identified Research Gaps

Despite extensive research on the impacts of economic complexity and human development on environmental sustainability, few studies examine their combined effects on SDGs 14 and 15. The literature mainly focuses on either economic complexity or human development, rarely integrating both in the context of specific environmental

goals.

While machine learning (ML) has potential for advancing sustainability analysis, its application remains limited. More research is needed to develop ML models that can predict and manage environmental impacts by analyzing the interactions between economic complexity, human development, and environmental sustainability. This would provide more accurate insights for policymakers.

Inconsistencies in findings on the contributions of economic complexity and human development to environmental goals highlight the need for standardized methodologies and comprehensive data. Comparative studies across different regions can offer robust insights into how these factors influence sustainability. Standardizing data collection and analysis methods will enhance the reliability and comparability of results.

All in all, Achieving SDGs 14 and 15 requires understanding the interplay between economic complexity, human development, and environmental sustainability. Managed effectively, higher economic complexity and human development can support sustainability, necessitating integrated policies and advanced tools like ML. Addressing research gaps and leveraging comprehensive frameworks can help policymakers balance economic growth, human well-being, and environmental protection. This review highlights the importance of integrated approaches and advanced tools for SDGs 14 and 15. By understanding these interactions and developing targeted policies, policymakers can promote sustainable development that supports both human well-being and environmental health.

3 Theoretical Framework

The conceptual foundation of this research is grounded in fact that that the evolving understanding of sustainable development has highlighted the necessity of balancing economic and social progress with environmental preservation, as stated in the introduction. This study builds on Numan's (2022) conceptual framework, which positions the environment as the sustaining dimension and economic and social progress as development dimensions. Numsn's framework highlights that true sustainable development requires analyzing the synergies and trade-offs between these dimensions to ensure that economic and social advancements do not come at the expense of environmental health and advocates for a nuanced understanding of how socio-economic development and environmental sustainability can be harmonized (Numan, M. N., 2022).

Most research has been focusing on similar understandings, however, they have focused on isolated aspects, for example Envrionmental Kuznets Curve (EKC). EKC hypothesize a concave parabolic relationship between environmental degradation

and economic growth (Prasad, M.N.V., 2024), shown in 4. However, this approach, among others, fail to fully capture the intricate interdependencies and comprehensive inter-connectivity between socio-economic development and its linkages with environmental sustainability (Numan, M. N., 2022). argued that there should be a shift from merely examining how income growth impacts the environment to a multidimensional analysis that captures the complex interactions between economic, social, and environmental factors (Numan, M. N., 2022). As mentioned in the literature review section, Numan (2022) superficially explored the impact of socio-economic development on environmental degradation, by incorporating a broader perspective that includes human welfare, equitable opportunities, the capability approach, economic complexity, and technological advancements. However, there are gaps in Numan's analysis that can be addressed using advanced methodologies.

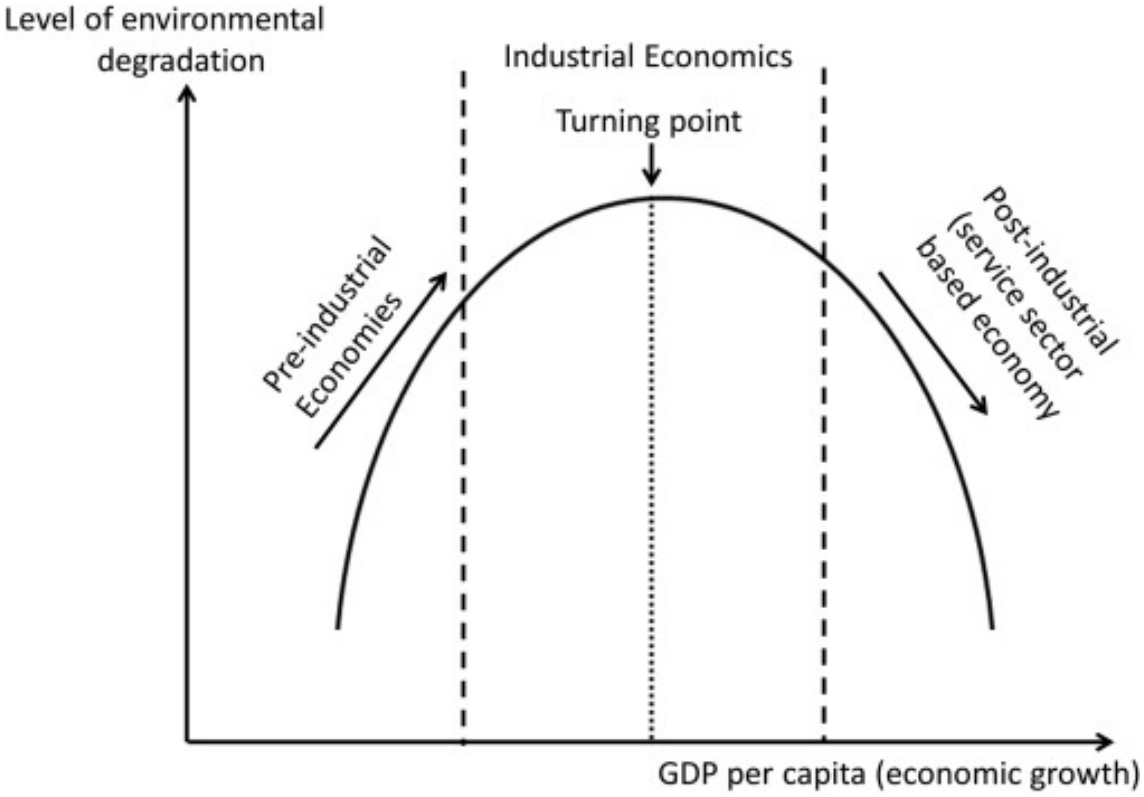

Figure 4: Environmental Kuznets Curve (Prasad, M.N.V., 2024)

This study aims to fill these gaps by employing predictive modeling rather than traditional statistical analysis. Machine learning method offers a more practical and comprehensive approach to understanding the multifaceted interrelationships between socio-economic development, economic complexity, human development, and environmental sustainability. The study deeply integrates human development theory and economic complexity to reflect inclusive socio-economic development, as both Ul Haq (1995) and Hidalgo and Hausmann (2009) emphasize the importance of better indicators to measure development.

In addition, this study views environmental sustainability through the lens of Sustainable Development Goals (SDGs) 14 and 15, providing a comprehensive perspective that encompasses both marine and terrestrial environments. This approach is particularly practical given that these goals are already being actively implemented globally. Furthermore, a detailed exposition of the key concepts and theories is provided below to establish a structured foundation for the empirical investigation.

Economic Complexity

The Economic Complexity Index (ECI), developed by César Hidalgo and Ricardo Hausmann in 2009, offers a profound insights into the capabilities of different economies by evaluating its complexity through analysis of exports. This metric builds on the notion that economic development hinges not just on what a country can produce, but on how diversified and sophisticated its production capabilities are (Hausmann, R. et al., 2014). There are many explanations behind economic output growth is not enough to meassure the true state of economic development. One of which is conceptualized by resource curse theory, which describes that economies that are well endowed with natural resources can achieve high growth but may remain underdeveloped (Collier P., 2007).

The Economic Complexity Index (ECI) is calculated by combining the diversity of products a country exports and the ubiquity of these products. Diversity refers to the range of different products a country exports, while ubiquity measures how many countries export a particular product (Hausmann, R. et al., 2014).

The diversity (k_c) of a country is the sum of the products it exports:

$$k_c = \sum_p M_{cp}$$

where k_c is the diversity of country c, and M_{cp} is a matrix element that equals 1 if country c exports product p and 0 otherwise. (Hausmann, R. et al., 2014).

The ubiquity (k_p) of a product is the sum of the countries that export it:

$$k_p = \sum_c M_{cp}$$

where k_p is the ubiquity of product p, and M_{cp} is a matrix element that equals 1 if country c exports product p and 0 otherwise. (Hausmann, R. et al., 2014).

These measures are then refined iteratively. High ECI values indicate economies that manufacture a wide array of complex products, suggesting an advanced industrial base and extensive know-how. Conversely, low ECI values are seen in economies that produce a limited range of simpler goods, typically reflecting lower levels of industrial sophistication and knowledge intensity (Hausmann, R. et al., 2014).

The processes required to sustain high economic complexity, such as resource extraction, industrial production, and technological innovation, can exert significant environmental pressures (Boleti, E. et al., 2019). These activities often lead to pollution, habitat destruction, and resource depletion (Boleti, E. et al., 2019). Therefore, while economic complexity can drive development, it also necessitates careful management to ensure environmental sustainability. The dual challenge is to leverage the benefits of economic complexity for socio-economic development while mitigating its environmental impacts. This balance is critical for achieving sustainable development goals (SDGs), particularly those related to economic growth, industry innovation, and environmental protection.

The Complexity Outlook Index (COI) complements the ECI by predicting future economic growth based on its current position and potential to diversify into more complex products (Hausmann, R. et al., 2014). The COI addresses the opportunities available for moving into new, more sophisticated products.

The COI is calculated as (Hausmann, R. et al., 2014):

$$\mathrm{COI}_c = \sum_p \Delta\mathrm{ECI}_p \cdot P_{cp}$$

where:

$$\Delta\mathrm{ECI}_p = \mathrm{ECI}_p - \mathrm{ECI}_c$$

$$P_{cp} = \frac{M_{cp}}{\sum_c M_{cp}}$$

Where, $\Delta\mathrm{ECI}_p$ represents the difference in ECI between the product p and the country's current ECI, reflecting the potential increase in complexity if the country were to start producing product p. P_{cp} denotes the proximity or likelihood of country c diversifying into product p, based on the current export structure and capabilities. (Hausmann, R. et al., 2014). A high COI value would imply that a country has significant unrealized potential to diversify into more complex products. (Hausmann, R. et al., 2014).

In essence, while the ECI provides a snapshot of where a country currently stands in terms of economic complexity, the COI offers a forward-looking perspective on how a country can enhance its economic complexity. This connection between ECI and COI is vital because it not only pinpoints the current strengths and weaknesses of an economy but also illuminates potential pathways for future growth and development.

Human Development

Mahbub Ul Haq's (1990) concept of human development theory, as encapsulated by the Human Development Index (HDI), emerged as a response to the failure of

economic indicators to capture the broader aspects of human well-being. Ul Haq (1990) emphasizing that true progress involves improving people's lives and is closely linked to Amartya Sen's capablity apporach as stated by Stanton, E.A. (2007). Concept of human development aims to shift the focus of development economics from national income accounting to people-centered policies (Stanton, E.A. (2007). The HDI, measure of human development, was first published by the United Nations Development Programme (UNDP), it takes into account development across three dimensions: health, education and standard of living (United Nations Development Programme., 1990).

The HDI is a composite index calculated by combining normalized indicators for health, education, and income (UNDP (United Nations Development Programme)., 2010)

$$\text{HDI} = \sqrt[3]{I_{\text{health}} \cdot I_{\text{education}} \cdot I_{\text{income}}}$$

where:

$$I_{\text{health}} = \frac{\text{Life expectancy at birth} - 20}{85 - 20}$$

where I_{health} is the normalized life expectancy index.

$$I_{\text{education}} = \frac{\text{Mean years of schooling}}{15} \times \frac{\text{Expected years of schooling}}{18}$$

where $I_{\text{education}}$ is the normalized education index.

$$I_{\text{income}} = \frac{\ln(\text{GNI per capita}) - \ln(100)}{\ln(75000) - \ln(100)}$$

where I_{income} is the normalized income index.

Moreover, the human development is intrinsically linked to sustainable development. However, the relationship between human development and environmental sustainability is complex. On one hand, countries with high HDI tend of have high ecological footprint, showcasing that HDI negatively influences environmental sustainability (Neumayer, E., 2016). However, some researcher have shown how HDI, especially after a certain point, improve environmental indicators (Van Tran, N. et al., 2019).

Both the Economic Complexity Index and the Human Development Index provide valuable frameworks for understanding the multifaceted nature of development and both are quite interlinked as well. While the HDI directly measures human capabilities in terms of health, education, and income, the ECI reflects the broader economic environment that supports these capabilities through industrial diversification and knowledge development. Together, these indices provide a comprehensive

understanding of development that incorporates both human well-being and economic complexity.

In the context of this study, particular attention is given to Sustainable Development Goals (SDGs) 14 (Life Below Water) and 15 (Life on Land). These goals focus on conserving marine and terrestrial ecosystems, respectively. SDG 15 is especially of significance due to its focus on terrestrial ecosystem and as highlighted in it"Sustainable Development Goals: Their Impacts on Forests and People," forests are crucial for achieving multiple SDGs by providing ecosystem services, supporting biodiversity, and contributing to human well-being (Katila et al., 2019).

Forests play a vital role in supporting SDG 15 (Life on Land) by conserving biodiversity and promoting sustainable land use. They also contribute to SDG 13 (Climate Action) as carbon sinks, and they support SDGs 1 (No Poverty) and 2 (Zero Hunger) by providing resources and livelihoods (Katila et al., 2019). Katila et al. (2019), however, describes that the pursuit of economic growth (SDG 8) and other goals often results in trade-offs with SDG 15, leading to conservation challenges. This dynamic can position development as a rival to conservation, causing SDG 15 and even SDG 14 (Life Below Water) to be disregarded as economic advancement takes precedence (Katila et al., 2019).

Nevertheless, the protection of marine and terrestrial ecosystems is of vital significance for overall sustainable development in the long run, despite the trade-offs faced (Sturesson, A., Weitz, N. and Persson, Å., 2018 and Katila et al., 2019). The integration of SDG 14 and SDG 15 into broader development strategies is essential for balancing economic, social, and environmental objectives. This approach aligns with Numan's (2022) framework, which emphasizes the need for a nuanced understanding of the synergies and trade-offs between different dimensions of sustainable development.

In conclusion, this study integrates economic complexity, human development, SDGs and environmental sustainability to provide a comprehensive approach to sustainable development. It emphasizes the synergies and trade-offs between these dimensions, using advanced methodologies to capture their intricate relationship. The Economic Complexity Index and Human Development Index, and SDG 14 and SDG 14 reflect the intertwined nature of economic development, human well-being, and environmental health. Henceforth, To gain a comprehensive understanding of the interactions, the empirical investigation will translate theoretical constructs into observable variables by considering three indicators each for SDGs 14 and SDG 15 (explained further in next sections), along with the HDI, ECI, and COI.

To empirically test these theoretical relationships, various modeling techniques will be employed. Linear regression will model the relationships between the dependent variable (environmental sustainability indicators) and the independent vari-

ables (ECI and HDI). Interaction terms will capture the combined effects of economic complexity and human development on environmental sustainability. Spline regression will provide flexibility in modeling nonlinear relationships, while Random Forest and Feed-Forward Neural Networks (FNNs) will capture complex, nonlinear interactions and patterns in the data. SHAP (SHapley Additive exPlanations) analysis will be used to interpret the results of these complex models, offering insights into the relative importance of different variables and their interactions in predicting environmental sustainability outcomes. The next section will expand on these empirical methods and their application in this study.

4 Data Collection and Source

Our World in Data (OWID) compiles information from reputable sources like academic institutions, government databases, and organizations such as the United Nations, World Bank, and World Health Organization. OWID's datasets cover topics including economics, health, education, and environmental sustainability. For this research, OWID was the primary source due to its extensive and meticulously curated datasets.

Two key OWID articles were crucial for data on SDG 14 and SDG 15. The first, "Conserve and sustainably use the oceans, seas and marine resources" (Our World in Data team, 2023), provided detailed information on marine biodiversity, ocean pollution, and the economic impact of marine activities. The second, "Sustainably manage forests, combat desertification, halt and reverse land degradation, halt biodiversity loss" (Our World in Data team, 2023), offered comprehensive data on terrestrial biodiversity, forest area, land degradation, and urbanization's impact on natural habitats. These articles supplied necessary data and valuable insights into the methodologies and definitions used, ensuring a thorough understanding of the progress and challenges in achieving SDG 14 and SDG 15.

4.0.1 The indicators collected for SDG 14:

Data for Sustainable Development Goal (SDG) 14, "Life Below Water," was gathered from OWID (Our World in Data team, 2023). Indicators included marine biodiversity, pollution levels, and the economic impact of marine tourism. These are crucial for assessing progress towards conserving and sustainably using marine resources.

- **Marine Pollution (Indicator 14.1.1)**: Measures coastal eutrophication and plastic debris density (Our World in Data team, 2023).
- **Ecosystem-Based Management (Indicator 14.2.1)**: Tracks countries using ecosystem-based approaches to manage marine areas (Our World in Data team, 2023).
- **Ocean Acidification (Indicator 14.3.1)**: Measures average marine acidity (pH) (Our World in Data team, 2023).
- **Sustainable Fishing (Indicator 14.4.1)**: Assesses fish stocks within biologically sustainable levels (Our World in Data team, 2023).
- **Protected Marine Areas (Indicator 14.5.1)**: Measures coverage of protected marine areas (Our World in Data team, 2023).

- **Combatting Illegal Fishing (Indicator 14.6.1)**: Evaluates implementation of international instruments to combat illegal fishing (Our World in Data team, 2023).
- **Economic Benefits from Marine Resources (Indicator 14.7.1)**: Measures sustainable fisheries as a proportion of GDP (Our World in Data team, 2023).
- **Marine Research (Indicator 14.a.1)**: Tracks research budget for marine technology (Our World in Data team, 2023).
- **Supporting Small-Scale Fishers (Indicator 14.b.1)**: Measures frameworks for small-scale fisheries access rights (Our World in Data team, 2023).
- **International Sea Law (Indicator 14.c.1)**: Tracks progress in ratifying and implementing ocean-related laws (Our World in Data team, 2023).

4.0.2 The indicators collected for SDG 15:

Data for Sustainable Development Goal (SDG) 15, "Life on Land," was sourced from OWID (Our World in Data team, 2023). Indicators included forest area, the Red List Index of species survival, and protected terrestrial areas. These are essential for evaluating efforts to protect and sustainably use terrestrial ecosystems.

- **Conserve and restore ecosystems (Indicator 15.1.1):** Measures forest area as a proportion of total land area (Our World in Data team, 2023).
- **Protect biodiversity areas (Indicator 15.1.2):** Proportion of important biodiversity sites covered by protected areas (Our World in Data team, 2023).
- **End deforestation (Indicator 15.2.1):** Measures progress towards sustainable forest management (Our World in Data team, 2023).
- **End land degradation (Indicator 15.3.1):** Assesses proportion of degraded land over total land area (Our World in Data team, 2023).
- **Conserve mountain ecosystems (Indicator 15.4.1):** Coverage of important mountain biodiversity sites by protected areas (Our World in Data team, 2023).
- **Vegetation coverage (Indicator 15.4.2):** Measures green vegetation in mountainous areas (Our World in Data team, 2023).
- **Species extinction risk (Indicator 15.5.1):** Tracks species extinction risk using the Red List Index (Our World in Data team, 2023).

- **Genetic resources (Indicator 15.6.1):** Number of countries with frameworks for sharing benefits from genetic resources (Our World in Data team, 2023).
- **Wildlife poaching (Indicator 15.7.1):** Measures proportion of illicitly traded wildlife (Our World in Data team, 2023).
- **Invasive species (Indicator 15.8.1):** Proportion of countries with legislation to control invasive species (Our World in Data team, 2023).
- **Ecosystem integration (Indicator 15.9.1):** Number of countries integrating biodiversity into planning (Our World in Data team, 2023).
- **Biodiversity financing (Indicator 15.A.1):** Measures development assistance for biodiversity conservation (Our World in Data team, 2023).
- **Forest management financing (Indicator 15.B.1):** Measures financial resources for sustainable forest management (Our World in Data team, 2023).

4.1 SDG Indicator Selection

The SDG indicator selection process began with assessing data availability. Not all indicators were monitored consistently, resulting in missing values. To address this, a subset of reliable indicators was chosen, focusing on three indicators for each goal based on data availability and consistency, details given below:.

SDG 14 Indicators

- **14.1.1 - Beach litter per square kilometer**: Measures coastal pollution by tracking litter on beaches (Our World in Data team, 2023).
- **14.1.2 - Chlorophyll-a deviation from the global average**: Tracks nutrient levels and water quality in marine ecosystems (Our World in Data team, 2023).
- **14.4.1 - Proportion of fish stocks within biologically sustainable levels**: Assesses sustainable fishing practices (Our World in Data team, 2023).

SDG 15 Indicators

- **15.1.1 - Forest area as a proportion of total land area (%):** Measures forest coverage (Our World in Data team, 2023).
- **15.2.1 - Above-ground biomass in forest (tonnes per hectare):** Assesses forest health and carbon storage (Our World in Data team, 2023).

- **15.4.1 - Average proportion of Mountain Key Biodiversity Areas (KBAs) covered by protected areas (%):** Measures protection of biodiversity in mountain regions (Our World in Data team, 2023).

4.2 Feature's Data

The main focus of the paper is on performing predictive modeling using socio-economic development. Since they paper lays its theoratical foundation on human development theory and economic complexity theory, hence, the paper particularly focuses on the HDI and ECI, acting as a proxy for socio-economic development. The relevant data for complexity measures (ECI and COI) was obtain from Harvard University *The Atlas of Economic Complexity* (2023) and for human development was obtained from United Nations Development Programme (2023) .

However, to enhance the robustness and accuracy of our model, we incorporate additional features such as GDP, complexity outlook index, and land area. These variables are essential as they provide a more comprehensive understanding of the multifaceted nature of environmental impacts. Detials of these variables are given below:

Country-specific geographic, demographic and development expenditure data:

- **Total Landmass**:
  - *Source*: Ortiz-Ospina & Roser, 2016; Ritchie & Roser, 2019
  - *Explanation*: The total landmass influences a country's environmental footprint and resource utilization. Larger land areas may have more diverse ecosystems and natural resources but also pose greater challenges in environmental management.
- **Population**:
  - *Source*: Ortiz-Ospina & Roser, 2016; Ritchie & Roser, 2019
  - *Explanation*: Population size affects resource consumption, waste production, and environmental impact. It also provides context for understanding the scale of socio-economic activities and development needs.
- **GDP**:
  - *Source*: Ortiz-Ospina & Roser, 2016; Ritchie & Roser, 2019
  - *Explanation*: Gross Domestic Product indicates the overall economic output of a country, reflecting its economic performance and capacity for environmental investments and innovations.

- **GDP per Capita**:
  - *Source*: Ortiz-Ospina & Roser, 2016; Ritchie & Roser, 2019
  - *Explanation*: GDP per capita measures the average economic prosperity of individuals in a country, offering insights into living standards and potential investments in sustainable development.
- **Research and Development Expenditure (% of GDP)**:
  - *Source*: Ritchie & Roser, 2020
  - *Explanation*: Investment in R&D reflects a country's commitment to innovation and technological advancement, which are crucial for sustainable economic growth and environmental solutions.
- **Education Expenditure (% of GDP)**:
  - *Source*: Ritchie & Roser, 2020
  - *Explanation*: Education expenditure indicates the emphasis on human capital development. Higher investments in education can lead to a more informed and capable population, better equipped to address socio-economic and environmental challenges.
- **Change over 5 years**:
  - *Source*: Harvard University, 2023
  - *Explanation*: Monitoring changes in the ECI ranking over five years provides valuable information on a country's economic development trajectory.

By integrating these additional features into the analysis, policymakers, researchers, and stakeholders can gain a more comprehensive understanding of the socio-economic and environmental dynamics shaping countries' development trajectories. In our research, we opted for a global focus, analyzing country-level data from 2015 to 2020 to ensure comprehensive coverage and consistency in available indicators. This time frame was chosen for its widespread data availability, facilitating robust analysis while capturing recent developments. Employing an outer join method, we integrated datasets for selected indicators based on country and year, prioritizing inclusivity to retain as much data as possible. Our multidimensional data analysis and modeling approach aims to provide a holistic understanding of global trends, identifying common patterns and disparities to inform evidence-based decision-making for sustainable development initiatives on an international scale. Our final dataset is presented in the following table 1:

Table 1: Presence of Missing Values in the Dataset

Feature	**Variable Type**	**Unit of Measurement**
Entity	Categorical	N/A
Code	Categorical	N/A
Year	Numeric	N/A
Economic Complexity Index (ECI)	Continuous	N/A
Complexity Outlook Index (COI)	Continuous	N/A
Change over 5 year	Numeric	N/A
Human Development Index (HDI)	Continuous	N/A
Land Area	Continuous	square kilometers
Annual CO2 Emissions	Continuous	metric tons
Population	Continuous	N/A
CO2 Emissions per Capita	Continuous	metric tons per capita
GDP	Continuous	US dollars
GDP per Capita	Continuous	US dollars per capita
Research and Development Expenditure (% GDP)	Percentage	percentage
Education Expenditure (% GDP)	Percentage	percentage
Beach Litter per Square Kilometer	Continuous	items per square kilometer
Chlorophyll-a Deviation from Global Average	Percentage	percentage
Proportion of Fish Stocks within Sustainable Levels	Percentage	percentage
Above-Ground Biomass in Forest	Continuous	metric tons per hectare
Average Proportion of Mountain Key Biodiversity Areas Covered by Protected Areas	Percentage	percentage

5 Exploratory Data Analysis (EDA)

5.1 Summary Statistics

For the exploratory data analysis, we focused on the summary statistics of critical indicators such as the Economic Complexity Index (ECI), Human Development Index (HDI), and GDP per capita, shown in table 2. The mean values and standard deviations of these indicators provide valuable insights into the socio-economic landscape of the dataset.

The mean ECI is 0.0116, with a standard deviation of 0.9958, signifying a diverse range of economic complexities among the countries or regions represented. This variability is highlighted by the minimum (-1.9616) and maximum (2.5501) values, underscoring the nuanced economic structures and developmental challenges observed globally.

Similarly, the Human Development Index (HDI) exhibits a mean of 0.7201, with a standard deviation of 0.1519. The HDI values range from a minimum of 0.367 to a maximum of 0.963, illustrating considerable disparities in human development levels across different geographic entities. Furthermore, the GDP per capita has a mean of 21,377.52 (PPP constant 2017 international dollar) and a standard deviation of 21,946.78, revealing substantial variations in economic prosperity among the countries or regions included in the dataset. The wide range of GDP per capita values, spanning from 711.36 to 128,437.32, highlights the vast differences in income levels and economic performance observed globally.

Table 2: Statistical Summary of Key Features

Stat	ECI	Change Over 5 Years	COI
Count	798	798	798
Mean	0.012	0.000	0.012
Std	0.996	10.820	0.999
Min	-1.962	-52.000	-2.677
25%	-0.778	-5.000	-0.913
50%	-0.134	0.000	-0.079
75%	0.729	4.000	0.812
Max	2.550	66.000	2.928

Stat	HDI	Land Area (sq. km)	Population
Count	1152	1284	1413
Mean	0.720	607214.820	32426958.360
Std	0.152	1755587.780	132523203.159
Min	0.367	2.027	1447.000
25%	0.601	10230.000	395649.000
50%	0.741	95300.000	5348285.000
75%	0.838	452860.000	20128132.000
Max	0.963	16376870.000	1424929792.000

Stat	GDP per capita	R&D expenditure	Education Expenditure
Count	1155	532	1060
Mean	21377.521	1.058	14.336
Std	21946.775	1.054	4.931
Min	711.355	0.010	0.833
25%	4860.802	0.250	11.053
50%	13633.048	0.690	13.708
75%	31038.743	1.464	17.176
Max	128437.320	5.436	38.106

Stat	Beach Litter	Chlorophyll-a Deviations	Sustainable Fisheries
Count	392	1138	313
Mean	4304443.528	3.633	0.967
Std	16249399.065	4.564	2.112
Min	100.060	0.000	0.000
25%	206955.450	0.670	0.050
50%	810260.775	1.960	0.220
75%	2387827.900	4.728	0.670
Max	179760100.000	46.260	13.600

Stat	Forest Cover	Forest Biomass	Mountain KBAs
Count	1345	1224	1098
Mean	31.945	124.210	45.288
Std	24.300	83.539	30.523
Min	0.000	0.000	0.000
25%	10.897	65.680	20.210
50%	30.900	111.995	42.100
75%	50.432	171.110	71.750
Max	97.767	500.390	100.000

These statistical findings provide a comprehensive overview of the socio-economic diversity and disparity present in the dataset, offering valuable insights into the dataset's overall socio-economic landscape.

5.2 Missing Values

The ammount of missing values in the dataset is shown in the Table 3 below. This step highlighted the extent of missing data across different variables.

Table 3: Presence of Missing Values in the Dataset

Variable	**Missing Values**
ECI	701
Change Over 5 Years	701
COI	701
Human Development Index	347
Land area (sq. km)	215
Population	86
GDP per capita	344
Research and development expenditure	967
Government expenditure on education	439
14.1.1 - Beach litter per square kilometer (Number)	1107
14.1.1 - Chlorophyll-a deviations, remote sensing (%)	361
14.7.1 - Sustainable fisheries as a proportion of GDP	1186
15.1.1 - Forest cover	154
15.2.1 - Above-ground biomass in forest (tonnes per hectare)	275
15.4.1 - Average proportion of Mountain Key Biodiversity Areas (KBAs) covered by protected areas (%)	401

The examination of missing values within the dataset revealed notable instances across various socio-economic and environmental indicators. With 701 missing values each, significant gaps were observed in crucial metrics such as the Economic Complexity Index (ECI), Change Over 5 Years, and Complexity Outlook Index (COI). Additionally, the Human Development Index (HDI) exhibited 347 missing values, suggesting potential challenges in assessing comprehensive human development levels. Furthermore, variables such as Research and Development Expenditure and Sustainable Fisheries as a Proportion of GDP displayed considerable missingness, with 967 and 1186 missing values, respectively. These findings underscore the importance of addressing missing data to ensure the integrity and reliability of subsequent analyses.

5.3 Pair Plot Analysis

Following the examination of summary statistics and identification of missing values within our dataset, we proceeded to visualize the relationships between variables through a pair plot analysis. This graphical exploration enabled us to gain deeper insights into the interdependencies and potential patterns present within the data. By plotting pairwise scatterplots for each combination of variables, we were able to visually assess correlations and identify potential trends, clusters, or outliers.

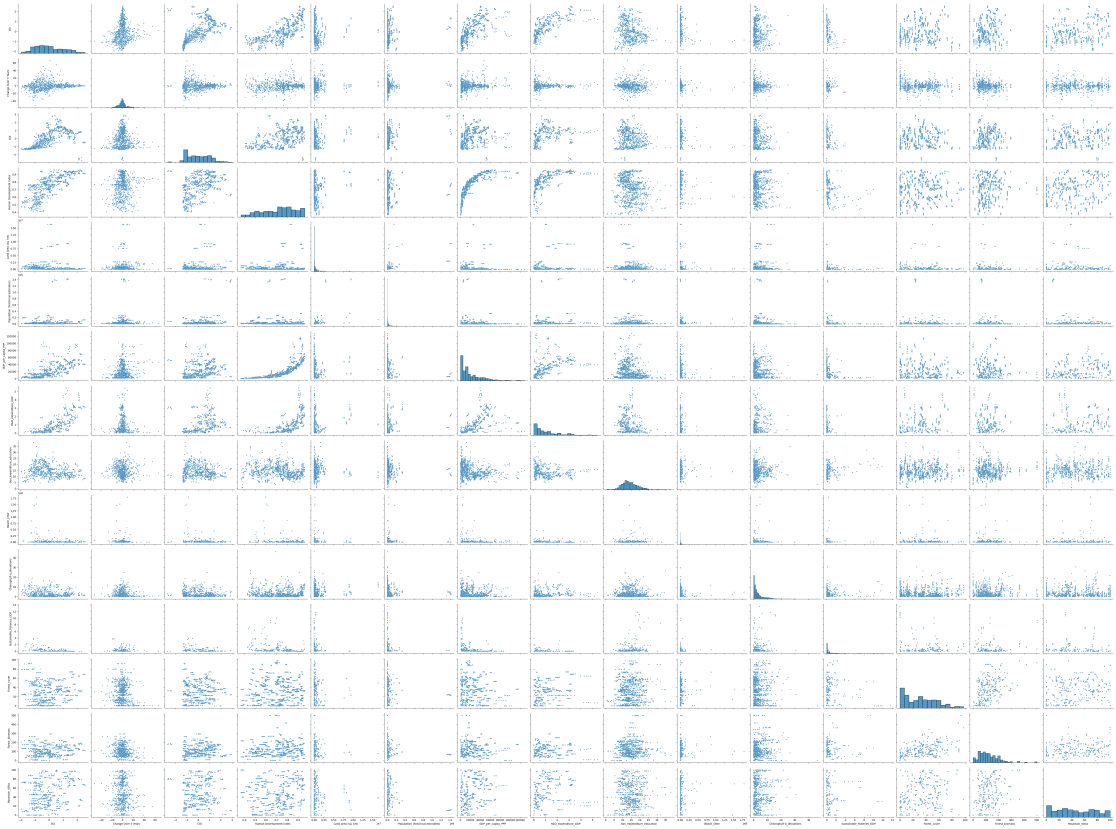

Figure 5: Original Data Pair Plot

From the visual examination of the pair plot, it is evident that certain variables exhibit distinct distributional characteristics. For instance, variables like Change Over 5 Years and Government Expenditure on Education as a percentage of GDP demonstrate a relatively normal distribution. Conversely, GDP per capita, along with Research and Development Spending as a percentage of GDP, display positive skewness, indicating a concentration of lower values with a tail stretching towards higher values.

Moreover, the pair plot analysis served as a valuable tool for uncovering potential multicollinearity issues and nonlinear relationships among variables, thereby guiding subsequent modeling and analysis decisions. While many features initially appeared to exhibit little to no linear relationship, a closer examination revealed underlying patterns in their interactions. Specifically, there appeared to be a higher-order relationship, indicating potential exponential relationships between certain variables. Notably, a discernible pattern emerged in the interaction between HDI and GDP per capita, as well as HDI and Research and Development Expenditure. These findings suggest the presence of complex dynamics and nonlinear dependencies within the dataset, highlighting the importance of considering higher-order relationships in further analyses and modeling efforts.

5.4 Correlation Matrix Analysis

The correlation matrix provides a comprehensive overview of the pairwise correlations between different variables in the dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

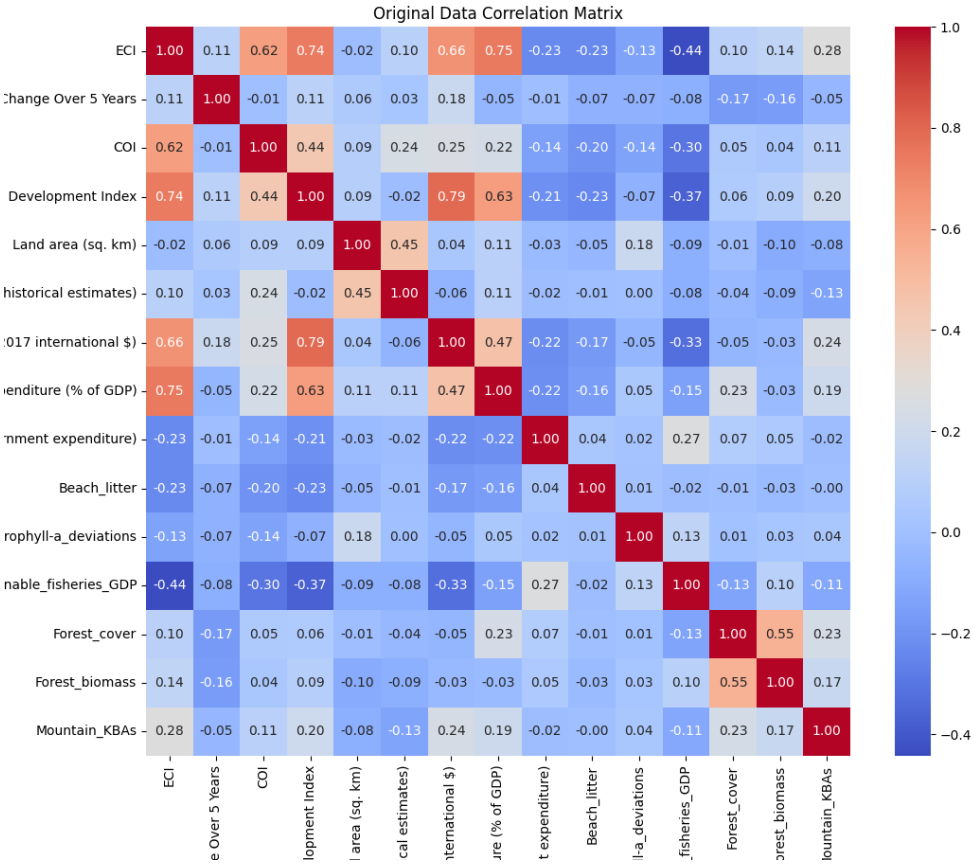

Figure 6: Original Data Correlation Matrix

In this correlation matrix:

- Variables like ECI (Economic Complexity Index), HDI (Human Development Index), GDP per capita, and Research and Development Expenditure (percentage of GDP) exhibit strong positive correlations with each other. For example, ECI shows a high positive correlation with HDI (0.739) and GDP per capita (0.663), suggesting that countries with higher economic complexity tend to have higher human development levels and GDP per capita.

- Notably, there are also significant positive correlations between HDI and other indicators such as Sustainable Fisheries as a Proportion of GDP (0.625) and Government Expenditure on Education (percentage of Government Expenditure) (0.625), indicating that countries with higher human development levels tend to invest more in education and sustainable fisheries.

- Conversely, variables like Land Area (sq. km) and Change Over 5 Years show weaker correlations with other indicators, suggesting less direct influence on the overall relationships within the dataset.

- Additionally, there are negative correlations observed between some variables, such as Forest Cover and Economic Complexity Index (-0.443), indicating potential trade-offs or inverse relationships between certain socio-economic and environmental factors.

The pair plot and correlation heat map are instrumental in understanding the data's structure, detecting multicollinearity, and identifying significant relationships that can inform further analysis and modeling efforts. Overall, the correlation matrix provides valuable insights into the interrelationships between different variables, guiding further analysis and interpretation in understanding the complex dynamics of socio-economic and environmental factors within the dataset.

6 Methodology

This section details the methodological framework employed to investigate the relationship between economic complexity, human development, and environmental sustainability, focusing on SDG 14 and SDG 15. The analysis leverages advanced data imputation techniques, various regression models, and interpretability methods such as SHAP to ensure robustness and interpretability of the results.

6.1 Data Preprocessing

In this step, non-informative variables such as entity codes, geographical coordinates, and certain economic rankings were excluded to streamline the dataset and reduce noise. This step ensured that only the most relevant variables were retained for further analysis. By removing these columns, the dataset becomes more manageable. It is also evident, stated previously in EDA, that that SDG 14 have considerably less values than SDG 15, thereby, making SDG 15 the main focus of our study. Addressing missing values is critical to prevent biases in the analysis and to ensure that the models are trained on complete and representative data, therefore, few imputation techniques were explored.

6.1.1 Imputations for SDG Data

4 different imputation techniques were employed:

MICE Imputation Multivariate imputation by chained equations (MICE) is a statistical method for handling missing data. It generates multiple imputations based on observed data, accounting for uncertainty and providing more accurate analyses than single imputation. MICE is flexible and can handle various variable types and complexities (Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J., 2011). MICE is particularly suitable for our dataset due to its ability to handle multivariate relationships, ensuring that the imputed values are consistent with the observed data structure. This technique is crucial for maintaining the integrity of complex relationships between economic and environmental indicators.

KNN Imputation K-Nearest Neighbour (KNN) imputation was also used, where missing values were imputed based on the mean values of the k-nearest neighbors (Beretta, L. and Santaniello, A., 2016). This method leverages the similarity between observations to fill in missing data. It is arugable that KNN imputation is beneficial in our context as it preserves the local structure of the data (Beretta, L. and Santaniello, A., 2016), making it useful for datasets with similar entities (e.g., countries with comparable economic profiles). However, this technique's performance depends on the choice of k and the distribution of missing values (Nti, I., yarko-Boateng, O., and Aning, J., 2021).

VAE and GAIN Imputation Variational Autoencoder (VAE) and Generative Adversarial Imputation Nets (GAIN) were also explored for imputation to enhanced the depth of analysis. VAE uses a probabilistic framework to model the data distribution (Qiu, Y., Zheng, H., and Gevaert, O., 2020), while GAIN, an advanced neural network-based imputation method, employs a generative adversarial approach

to estimate missing values (Yoon, J., Jordon, J., and Schaar, M., 2018). GAIN's loss function includes a hint mechanism to improve imputation accuracy (Yoon, J., Jordon, J., and Schaar, M., 2018):

$$\mathcal{L}_{GAIN} = E_m[\mathcal{L}_d + \lambda \mathcal{L}_h] \tag{1}$$

where $\mathcal{L}_d$ is the data loss, $\mathcal{L}_h$ is the hint loss, and λ is a hyperparameter.

These advanced techniques are particularly useful for complex datasets with intricate relationships and high levels of missing data. They leverage the power of deep learning to generate more accurate imputations, potentially improving the quality of the dataset for subsequent analysis (Qiu, Y. et al., 2020 and Yoon, J., 2018).

6.1.2 Comparative Analysis for Imputed Datasets

The imputed datasets were compared based on summary statistics and correlation matrices. This comparison is done to ensure that the imputed values preserved the underlying distribution and relationships between variables which is important for correct imputation (Beretta, L. and Santaniello, A., 2016). By assessing these metrics, we ensured that the imputation methods did not introduce significant biases or distortions in the data. The MICE imputation method closely follows the original data's distribution and effectively retains the statistical properties, making it a highly reliable approach. MICE imputation maintains the count, mean, standard deviation, and range across all columns, closely mirroring the original dataset. KNN imputation also shows promising results, closely following the original data's distribution and effectively handling missing data, albeit with slight adjustments in means and potential smoothing of variability. Conversely, the VAE and GAIN imputation methods show significant deviations from the original data. VAE imputation introduces substantial changes in summary statistics, particularly in variables like 'ECI', 'Change Over 5 Years', and 'COI', suggesting potential over-regularization or inadequate training. The GAIN method exhibits substantial issues, with an incomplete imputation process as indicated by a low count of 68 entries, and considerable deviations in mean, standard deviation, and range, highlighting the need for revisiting its implementation and training processes.

From a correlation perspective, the MICE and KNN methods are again the most effective, preserving the original relationships between variables with minor variations. The MICE imputation method particularly excels in maintaining the original data's correlations, ensuring the integrity of relationships such as the strong positive correlations between ECI and variables like Human Development Index, GDP per capita, and Research and Development expenditure. KNN imputation also re-

Table 4: Concise Report of Negative Values

Variable	Total Negative Values	Close to Zero
R&D_expenditure_GDP	198	158
Sustainable_fisheries_ of GDP	115	44

tains the general pattern of correlations, with only minor deviations. In contrast, VAE imputation introduces stronger correlations, potentially reflecting overfitting or over-regularization, while GAIN imputation results in weaker and inconsistent correlations, suggesting significant issues in the imputation process. Given these findings, MICE imputation is recommended for its superior performance in preserving the data's statistical properties and inherent distribution, making it ideal for ensuring data integrity. KNN imputation is also a viable option, albeit slightly less effective in maintaining correlation patterns. VAE and GAIN imputation methods require further tuning and validation to address their current limitations and ensure they do not distort the data's true relationships.

6.1.3 Fixing Invalid Values In Imputed Datasets

However, summary statistics pointed towards the fact that certain variables, such as percentages and counts, cannot have negative values. All thought the MICE imputed dataset has some invalid negative values that amount which is given in table 4 below. Since the amount of values, not close to zero were small, therefore, for simplicity the correction was done by setting all of negative values to zero, ensuring the imputed datasets were realistic and consistent. This step was necessary to maintain the validity and interpretability of the dataset.

Correlation matrix, pair plots and summary statistics analysis was performed again after fixation of negative values to ensure that imputated dataset preserves the original data's correlation and distribution. The alaysis shows that MICE retains the original data's correlation structure, preserving high correlations such as ECI with the Human Development Index (0.75) and GDP per capita (0.65). It also maintains moderate correlations, though with minor deviations. KNN imputation shows lower correlation coefficients for some key relationships, indicating slightly less accurate imputation.

6.1.4 Variance Inflation Factor (VIF)

Nevertheless, correlation matrices show something high correlation among certain feature variables. Addressing multicollinearity is crucial for ensuring the stability and reliability and can be achieve by performing variance inflation factor (Sheather, S., 2009). Variance Inflation Factor (VIF) is used to measure the amount of multi-

collinearity among the features in a dataset (Sheather, S., 2009). Mathematically, VIF is defined as (Sheather, S., 2009):

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 is the coefficient of determination of the regression of the i-th predictor on the remaining predictors.

The interpretation of VIF values is straightforward:

- VIF < 5: Indicates low multicollinearity.
- VIF between 5 and 10: Indicates moderate multicollinearity, which may be acceptable depending on the context.
- VIF > 10: Indicates high multicollinearity, which is problematic and needs addressing.

Our VIF analysis results are shown for the MICE and KNN imputed datasets, respectively in the following tables 5 and 5:

Table 5: VIF for MICE Imputed Data

feature	VIF
ECI	3.861
Change Over 5 Years	1.093
COI	1.956
Human Development Index	20.897
Land area (sq. km)	1.507
Population (historical estimates)	1.455
GDP_per__PPP	6.094
R&D_expenditure_GDP	4.595
Gov_expenditure_education	11.533

Table 6: VIF for KNN Imputed Data

feature	VIF
ECI	2.426
Change Over 5 Years	1.180
COI	1.769
Human Development Index	14.801
Land area (sq. km)	1.747
Population (historical estimates)	1.597
GDP_per_capita_PPP	3.835
R&D$_{e}$*xpenditure_GDP*	2.657
Gov_expenditure_education	8.920

As seen in tables 5 and 5 above, the VIF analysis for the MICE and KNN imputed datasets identified high multicollinearity in several features. HDI and Government expenditure on education has the highest VIF values in case of both datasets. However, te critical importance of HDI and HDI's comprehensive nature, which includes the aspects of education, we decided to retain HDI and drop Government expenditure on education. The updated VIF results are shown in tables below:

Table 7: VIF and Tolerance for Reduced MICE Imputed Data After Dropping

feature	VIF
ECI	3.813
Change Over 5 Years	1.093
COI	1.915
Human Development Index	6.090
Land area (sq. km)	1.503
Population (historical estimates)	1.448
GDP_per_capita_PPP	5.507
R&D_expenditure_GDP	4.594

Table 8: VIF and Tolerance for Reduced KNN Imputed Data After Dropping

feature	VIF
ECI	2.308
Change Over 5 Years	1.180
COI	1.768
Human Development Index	4.511
Land area (sq. km)	1.742
Population (historical estimates)	1.597
GDP_per_capita_PPP	3.348
R&D_expenditure_GDP	2.593

After removing the 'Government expenditure on education' variable, VIF values notably decreased, yet HDI still maintains the highest VIF among other features, reaching 6.09 for MICE-imputed data and 4.511060 for KNN imputed data, slightly surpassing the common threshold of 5 in the former but remaining well within acceptable limits in the latter. However, if only one or two features exceeds VIF value of 5 with a small margin, multicollinearity does not cause any major problem (Sheather, S., 2009). Consequently, we can proceed confidently with these datasets for further analysis, ensuring reliable outcomes. However, considering that the MICE-imputed dataset still performed better in preserving the distribution and correlation closely resembling that of the original data. Thus, we decide to use MICE imputed dataset for the actual modeling and analysis, methodology for which is given in the next section.

6.1.5 Data Preparation

The MICE imputed dataset is normalized to ensure uniformity in feature ranges, facilitating effective training of machine learning models. The datasets were split into training (2015-2019) and testing (2020) sets to evaluate model performance on unseen data. Target variables were defined, and features and targets were separated for modeling. This split ensures that the models are evaluated on their ability to generalize to new data, providing a realistic assessment of their performance.

6.2 Predictive Modelling Methodology

Achieving SDGs 14 and 15 requires understanding the interplay between economic complexity, human development, and environmental sustainability. Managed effectively, higher economic complexity and human development can support sustainability, necessitating integrated policies and advanced tools like ML. Addressing research gaps and leveraging comprehensive frameworks can help policymakers balance economic growth, human well-being, and environmental protection.

This review highlights the importance of integrated approaches and advanced tools for SDGs 14 and 15. By understanding these interactions and developing targeted policies, policymakers can promote sustainable development that supports both human well-being and environmental health.

6.2.1 Linear Regression with k-fold Validation

Linear regression is the most fundamental statistical supervised learning technique. It is used to predict numeric variable by finding a linear relation ship with set of independent features.It is a very simple model which presume a linear relation between independent and depented variable and it works better if there is little or no multi-collinearity with in the variables.(Lindholm., 2022)

Linear regression was applied with k-fold validation to provide a robust estimate of its performance. K-fold cross-validation divides the dataset into K equal-sized folds and iteratively trains the model on K-1 folds while testing it on the remaining fold, repeating this process K times. The average performance across all K iterations provides a robust estimate of the model's effectiveness. The model fits a linear relationship between features and targets, providing a baseline for comparison with more complex models.

6.2.2 Spline Regression with Hyperparameter Tuning

We've observed that polynomial regression produces smooth curve but it's trained on the full dataset, which can pose issues. And using step functions on to train different subset of data can often result in uneven fits.Regression splines offer a solution by striking a balance between these approaches, offering adaptive local smoothness.

Spline-based regression serves as a versatile framework encompassing various models. It revolves around the concept of piecewise polynomial regression, where polynomials are fitted to distinct segments or regions of the input variables, rather than across its entire range. Polynomial regression and step functions are specific instances of this broader approach to piecewise polynomial regression. Spline functions offer a means of approximating the shape of a nonlinear random function without requiring prior specification of its mathematical form (Suits et al., 1978).

Spline regression was employed to capture nonlinear relationships. Hyperparameter tuning using GridSearchCV identified the optimal number of knots, degree of splines, and regularization parameter, enhancing model flexibility and fit. The parameters used for tuning are used to configure a spline transformer and a ridge regression model. The knots parameter determines the number of knots in the spline, while degree sets the degree of the polynomial spline basis functions. Additionally, alpha specifies the regularization strength for the ridge regression. Grid search was

applied to these parameters, testing different combinations, such as 2 or 3 knots, degrees of 2 or 3, and alpha values of 0.1, 1, or 10. By systematically evaluating each combination, grid search identifies the optimal parameters that maximize the model's performance based on a specified metric.

6.2.3 Random Forest with Hyperparameter Tuning and k-fold Validation

Random Forest, an ensemble learning method that utilizes multiple decision trees to predict continuous outcomes.Random Forest, this algorithm is well-suited for regression tasks, leveraging the collective predictions of individual trees to generate robust and accurate models. By creating subsets of the data and features for each decision tree, Random Forest Regression effectively handles complex datasets while mitigating overfitting.(Lindholm., 2022)

Grid search was applied for hyperparameter tuning and k-fold validation.Grid search was utilized to determine the optimal combination of hyperparameters, specifically estimator and max depth. The estimator parameter represents the number of trees in the forest, while max depth determines the maximum depth of each tree. By defining a grid of possible values for these parameters, such as [100, 200, 500, 1000] for estimators and [10, 20, 30, 40, 50] for max depth It constructs multiple decision trees and aggregates their predictions, reducing over-fitting and improving generalization.

6.2.4 Feed-Forward Neural Network (FNN) with Iterative Hyperparameter Tuning

A Feedforward Neural Network (FNN) is a specific type of artificial neural network characterized by the unidirectional flow of information. Data travels from the input layer through any intermediate hidden layers before reaching the output layer. FNNs lack cyclic connections between units, hence the term "feedforward." This simple structure makes them particularly suitable for tasks involving straightforward data processing, such as pattern recognition and predictive modeling(Fine., 1999).

An FNN was trained with iterative hyperparameter tuning to optimize the network architecture.The number of neurons in each hidden layer of the neural network, offering flexibility in model complexity. learning rate controls the rate at which the weights of the neural network are updated during training, influencing the speed and stability of learning. The FNN's ability to capture complex, non-linear relationships makes it suitable for the dataset, providing enhanced predictive performance.

6.2.5 Creating Interaction Terms

Interaction terms, representing the combined effects of variables, were generated based on theoretical relevance. These terms were added to the dataset, scaled, and split into training and testing sets. Interaction terms allow for capturing the synergistic effects between variables, providing a more nuanced understanding of their combined impact (Ai & Norton., 2002). The theoretical framework previously discussed highlights the intrinsic connections between ECI, COI, and HDI.

ECI and COI are linked in that COI builds upon the current economic complexity measured by ECI to predict future growth potential. the interaction between the Economic Complexity Index (ECI) and the Complexity Outlook Index (COI) can reveal how a country's current industrial capabilities (ECI) and its future potential for diversification (COI) work together to influence economic growth and sustainability. On the other hand, HDI and ECI are intertwined as improvements in human development can drive economic complexity, and vice versa. For instance, a higher HDI can lead to a more skilled and educated workforce, which can then contribute to producing more complex and diverse goods, thereby increasing the ECI (Caous & Huarng., 2020). Thus, the interaction between the Human Development Index (HDI) and ECI can illustrate how improvements in health, education, and standard of living (HDI) can enhance or be enhanced by a country's industrial and knowledge complexity (ECI).

To integrate these interaction terms into our analysis involved creating new variables that multiply the values of the interacting factors:

$$\text{ECICOI} = \text{ECI} \times \text{COI} \tag{3}$$

$$\text{HDIECI} = \text{HDI} \times \text{ECI} \tag{4}$$

6.2.6 Modeling on Interaction Term Dataset

The same modeling approaches were applied to the dataset with interaction terms to evaluate whether incorporating these terms improved model performance.

6.2.7 Dimensionality Reduction and Feature Selection

*Principal Component Analysis PCA, is a method for reducing the dimensions of data by transforming it into a new feature space. In this new space, the variables are linear combinations of the original ones. The primary objective of PCA is to identify the principal components, which represent the most substantial variation in the dataset.(Bair, Hastie, Debashis, & Tibshirani., 2006)

*Partial Least Squares PLS initially computes the outer relationship of the X and Y blocks separately. Subsequently, it establishes a connection between both sets of scores through an inner mechanism It take into account how the X and Y scores interact, providing more insights to explore and comprehend.(Maitra & Yan., 2008)

*PCA and PLS were applied to reduce dimensionality. PCA transforms the data into orthogonal components capturing maximum variance, while PLS maximizes covariance between predictors and responses. These techniques enhance model efficiency and performance by focusing on the most informative features.

*Autoencoders, a type of neural network, were used to create additional features by learning a compressed representation of the data. These features were combined with PCA and PLS features to create a hybrid feature set, leveraging the strengths of different dimensionality reduction techniques.(Ran et al., 2022)

6.2.8 Modeling on Hybrid Feature Set

The same modeling approaches were applied to the hybrid feature set, with hyperparameter tuning and k-fold validation, to leverage the combined strengths of PCA, PLS, and autoencoders.

6.3 Model Interpretation

Models were evaluated using metrics such as MAE, MSE, RMSE, and R^2, and computation times were tracked. Results were compiled into well-organized tables for comparison, ensuring a comprehensive assessment of model performance.

SHapley Additive exPlanations (SHAP) Analysis for Best Models :
SHAP analysis was performed to interpret feature importance. SHAP values quantify the contribution of each feature to the model's predictions, providing insights into how each feature influences the predictions. Summary plots and dependence plots were generated, and SHAP values were summarized into a table.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{5}$$

where ϕ_i is the SHAP value for feature i, S is a subset of all features N, and v is the value function representing the model's prediction.

In technical terms, the process involves utilizing Shapley values derived from coalitional game theory (Acerta., 2022). SHAP analysis enhances the interpretability of complex models, providing insights into the influence of each feature on the

predictions. Essentially, Shapley values offer a method to illustrate the relative influence of each feature or variable being assessed on the final output of the machine learning model. This is achieved by comparing how the inputs' effects deviate from the average impact (Acerta., 2022).

7 Results and Findings

The primary objective of this study was to evaluate the predictive performance of several advanced machine learning models—Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN)—on three critical Sustainable Development Goals (SDG) indicators: Forest Cover, Forest Biomass, and Mountain Key Biodiversity Areas (KBAs). The models were tested on three different datasets: MICE imputed data, an interaction term dataset, and a hybrid feature set combining PCA, PLS, and autoencoder features. The models' performances were evaluated using metrics such as Mean Squared Error (MSE) and R^2, both for training and testing datasets.

8 SDG 15 Modelling Results and Findings

The results of the various machine learning models—Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN)—are systematically presented using tables in the appendix. These tables detail the performance metrics for each model, including Mean Squared Error (MSE) and R^2 values for both training and testing datasets. Each model's performance is evaluated across three datasets: MICE imputed data, an interaction term dataset, and a hybrid feature set. Comprehensive visual presentation of results is shown for each indicator before the model results are descriptively presented. This allows for a clear comparison of model efficacy and highlights the most effective algorithms for predicting each SDG indicator. Relevant table summarizing the result for SDG 15 is provided in the appendix A, as table 9.

8.1 Forest Cover

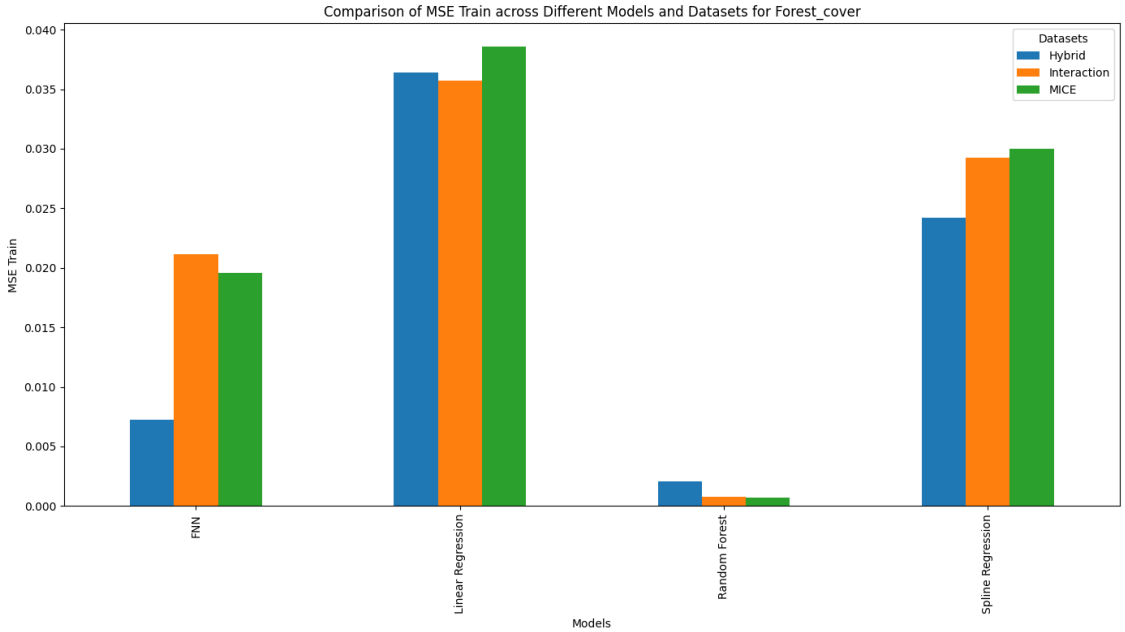

Figure 7: Comparison of MSE Train across Different Models and Datasets for Forest Cover

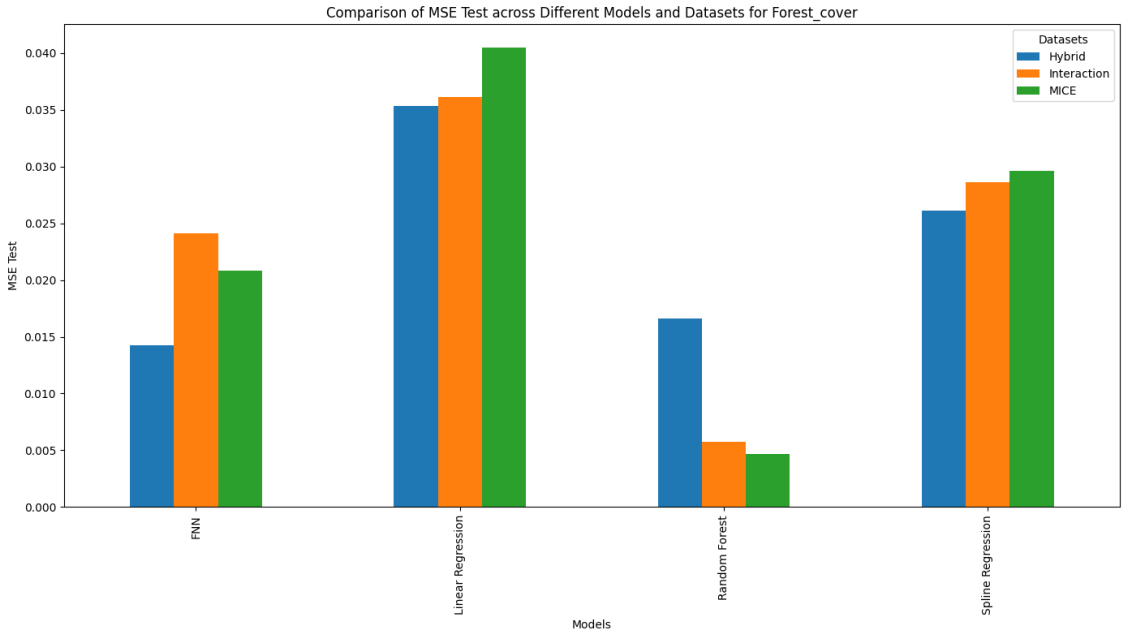

Figure 8: Comparison of MSE Test across Different Models and Datasets for Forest Cover

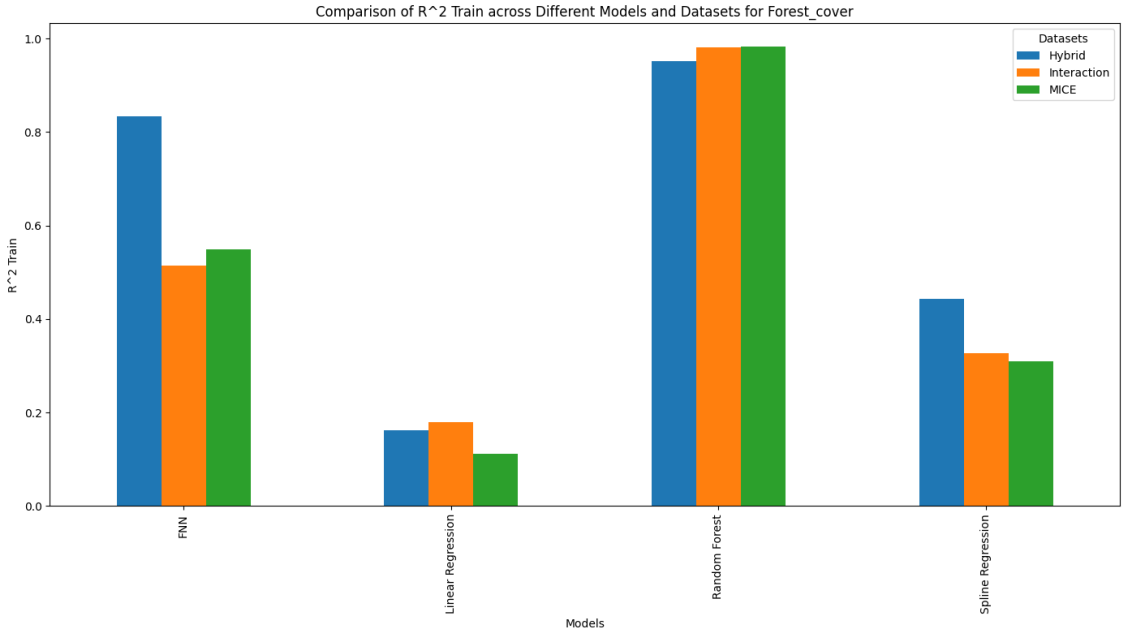

Figure 9: Comparison of R^2 Train across Different Models and Datasets for Forest Cover

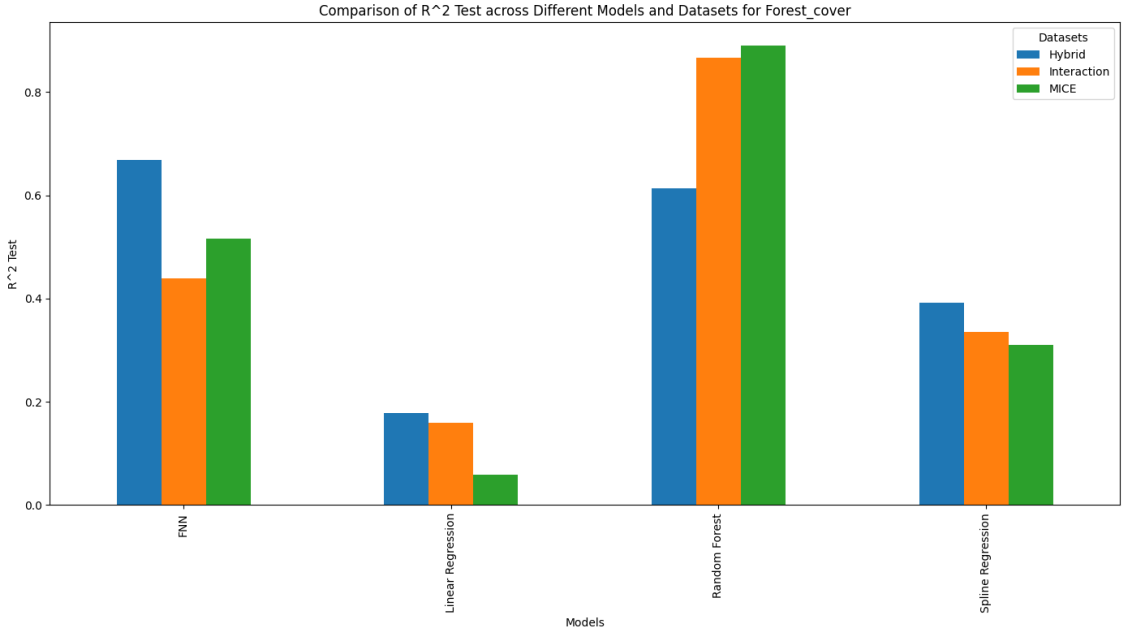

Figure 10: Comparison of R^2 Test across Different Models and Datasets for Forest Cover

8.1.1 Linear Regression:

The linear regression model demonstrated limited performance across all datasets. For the MICE imputed dataset, the model achieved an MSE Train of 0.0386 and an R^2 Train of 0.1116, indicating that only about 11% of the variance in the training data was explained. The testing performance was even lower, with an MSE Test of 0.0405 and an R^2 Test of 0.0582. With the interaction term dataset, there was a slight improvement in performance (R^2 Test of 0.1599), suggesting that the interaction terms captured additional variance. However, the hybrid feature set did not significantly enhance performance (R^2 Test of 0.1778), demonstrating the linear model's limitations in capturing complex relationships.

8.1.2 Spline Regression:

Spline regression improved the performance over linear regression by accounting for non-linear relationships. On the MICE dataset, the model yielded an MSE Train of 0.0300 and an R^2 Train of 0.3089, with an R^2 Test of 0.3106. The interaction term dataset further improved the model's performance (R^2 Test of 0.3349), indicating the significance of interaction terms. The hybrid feature set provided the highest R^2 Test of 0.3923, showcasing the model's capability to leverage advanced feature engineering techniques. The best parameters for spline regression were {`'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5`}.

8.1.3 Random Forest:

Random forest models significantly outperformed linear and spline regression models. For the MICE dataset, the random forest achieved an MSE Train of 0.0007 and an R^2 Train of 0.9835, with an R^2 Test of 0.8907, indicating that the model explained approximately 89% of the variance in the test data. The best parameters were {`'max_depth': 30, 'n_estimators': 100`}. The interaction term dataset showed slightly lower performance (R^2 Test of 0.8669). The hybrid feature set, despite an MSE Train of 0.0021 and an R^2 Train of 0.9526, yielded an R^2 Test of 0.6134, illustrating the variability in performance based on feature engineering techniques. The best parameters for the hybrid feature set were {`'max_depth': 40, 'n_estimators': 1000`}.

8.1.4 Feed-Forward Neural Networks (FNN):

FNNs exhibited moderate performance for Forest Cover. On the MICE dataset, the model achieved an MSE Train of 0.0196 and an R^2 Train of 0.5491, with an R^2 Test of 0.5158. The interaction term dataset performed similarly, with an R^2

Test of 0.4389. The hybrid feature set showed improved performance, with an R^2 Test of 0.6691, indicating that neural networks benefit significantly from complex feature sets. The best parameters for the hybrid feature set were {`'neurons': 128, 'learning_rate': 0.01`}.

8.2 Forest Biomass

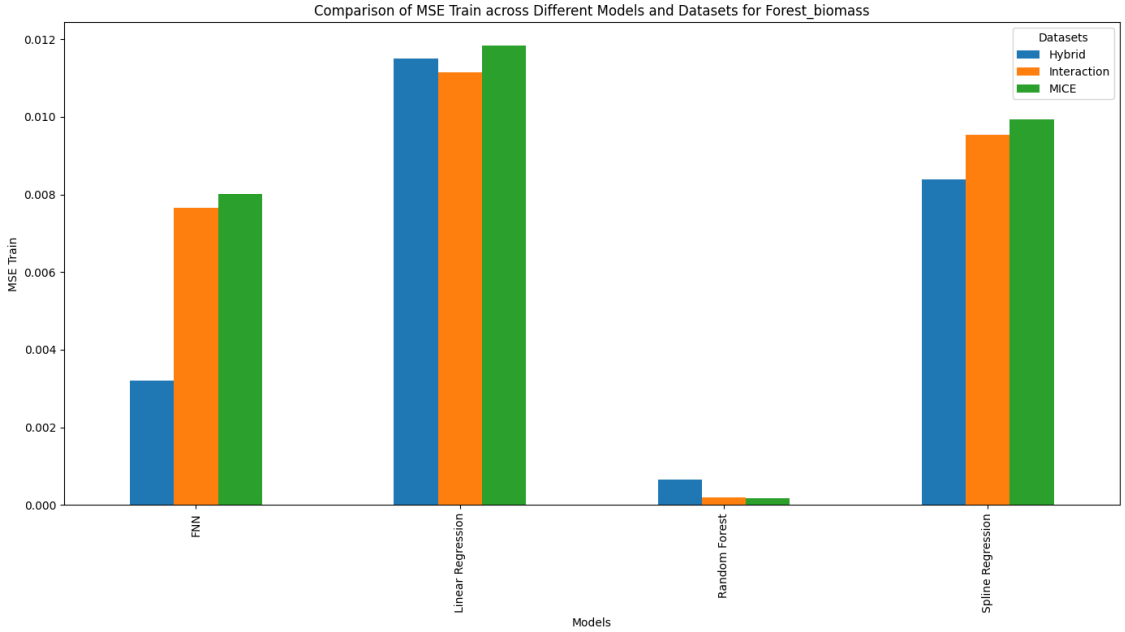

Figure 11: Comparison of MSE Train across Different Models and Datasets for Forest Biomass

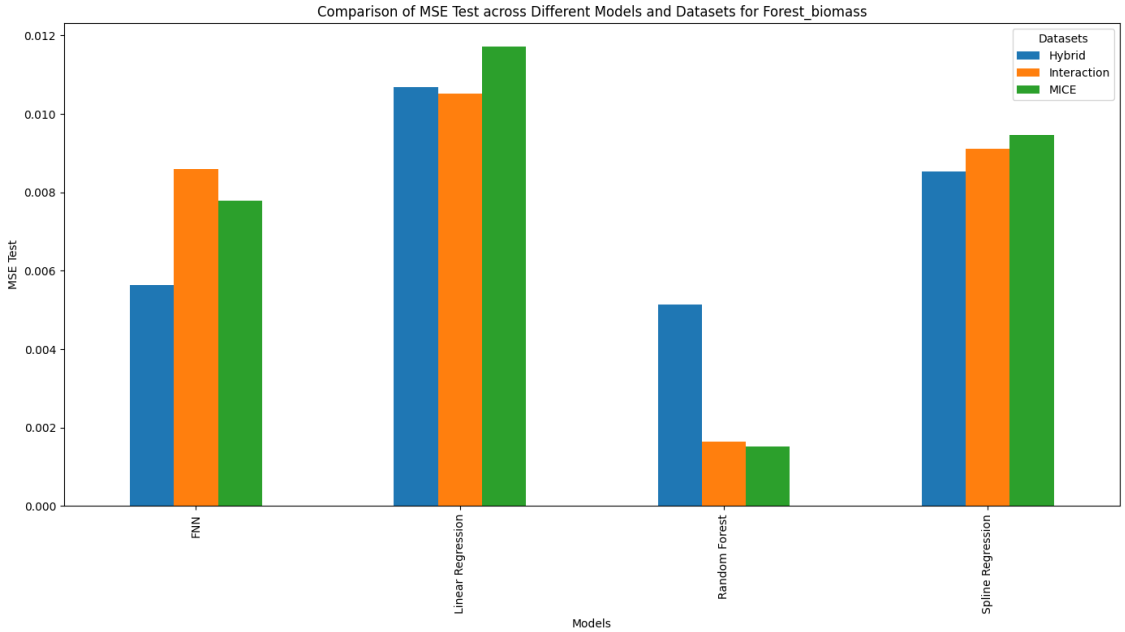

Figure 12: Comparison of MSE Test across Different Models and Datasets for Forest Biomass

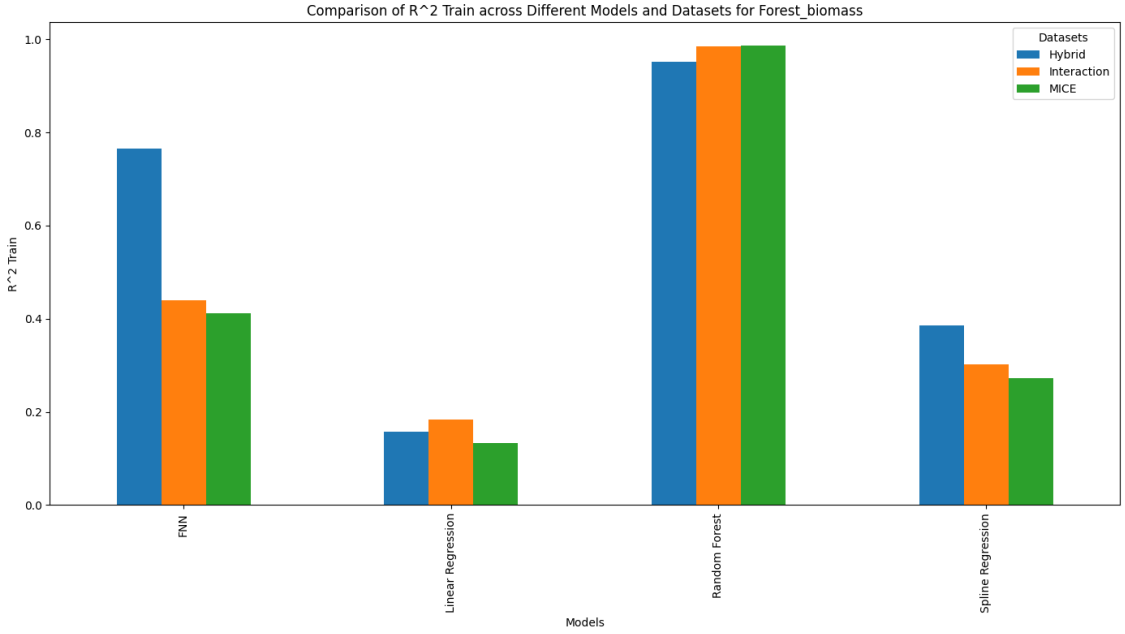

Figure 13: Comparison of R^2 Train across Different Models and Datasets for Forest Biomass

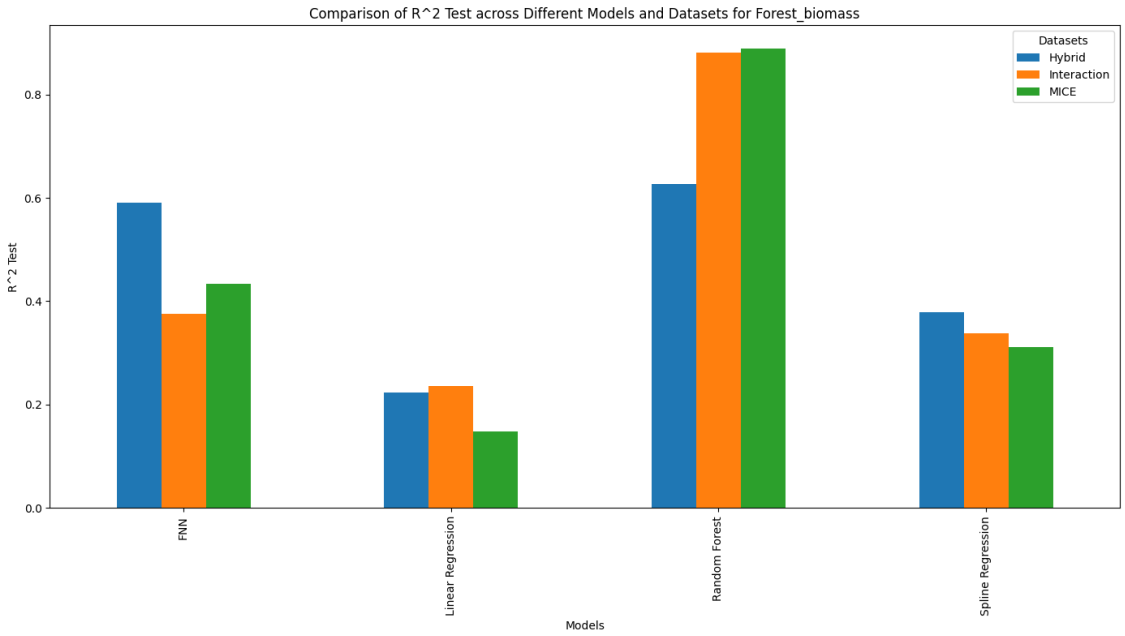

Figure 14: Comparison of R^2 Test across Different Models and Datasets for Forest Biomass

8.2.1 Linear Regression:

The linear regression model performed better for Forest Biomass compared to Forest Cover. On the MICE dataset, the model achieved an MSE Train of 0.0118 and an R^2 Train of 0.1322, with an R^2 Test of 0.1480. The interaction term dataset significantly enhanced the model's performance (R^2 Test of 0.2353), demonstrating the importance of capturing interactions. The hybrid feature set maintained a high performance (R^2 Test of 0.2241), corroborating the utility of advanced feature engineering.

8.2.2 Spline Regression:

For Forest Biomass, spline regression models consistently outperformed linear regression. The MICE dataset yielded an MSE Train of 0.0099 and an R^2 Train of 0.2718, with an R^2 Test of 0.3119. The interaction term dataset further improved performance (R^2 Test of 0.3386), while the hybrid feature set achieved an R^2 Test of 0.3795, highlighting the importance of capturing non-linear relationships and interactions. The best parameters for spline regression were {`'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5`}.

8.2.3 Random Forest:

Random forest models excelled for Forest Biomass as well. The MICE dataset resulted in an MSE Train of 0.0002 and an R^2 Train of 0.9874, with an R^2 Test of 0.8894. The best parameters were {`'max_depth': 30, 'n_estimators': 100`}. The interaction term dataset showed slightly lower performance (R^2 Test of 0.8804), while the hybrid feature set yielded an R^2 Test of 0.6272. The best parameters for the hybrid feature set were {`'max_depth': 40, 'n_estimators': 1000`}. This underscores the model's robustness in handling various feature sets and capturing complex relationships.

8.2.4 Feed-Forward Neural Networks (FNN):

For Forest Biomass, FNNs showed moderate performance across datasets. The MICE dataset resulted in an MSE Train of 0.0080 and an R^2 Train of 0.4124, with an R^2 Test of 0.4343. The interaction term dataset performed slightly lower (R^2 Test of 0.3762). The hybrid feature set showed improved performance, with an R^2 Test of 0.5909, indicating that neural networks benefit from sophisticated feature engineering. The best parameters for the hybrid feature set were {`'neurons': 128, 'learning_rate': 0.01`}.

8.3 Mountain KBAs

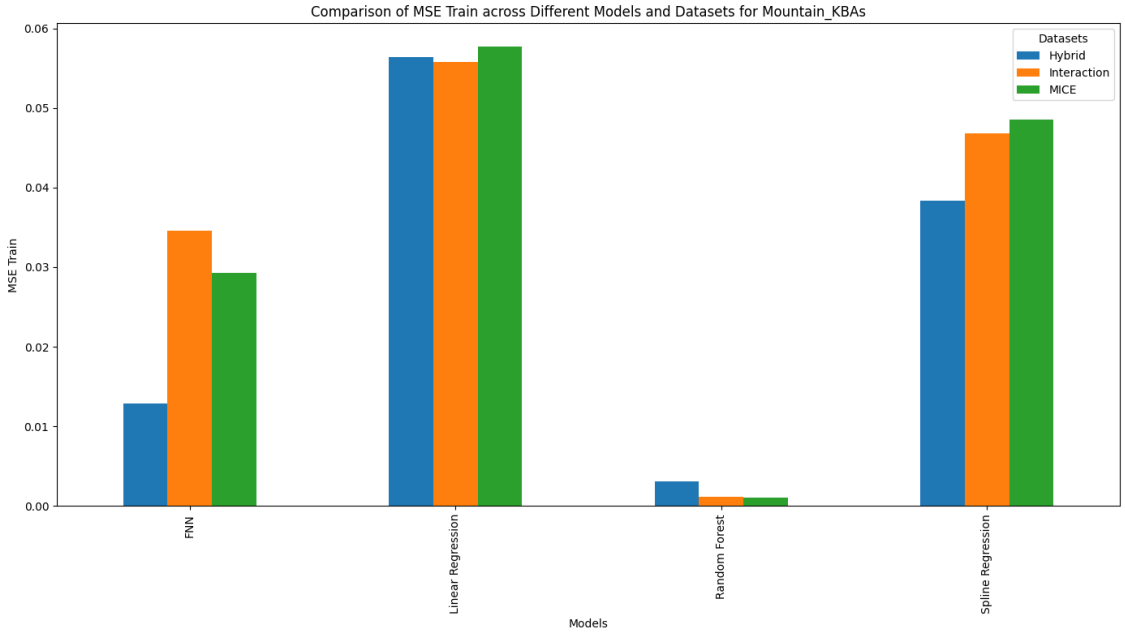

Figure 15: Comparison of MSE Train across Different Models and Datasets for Mountain KBAs

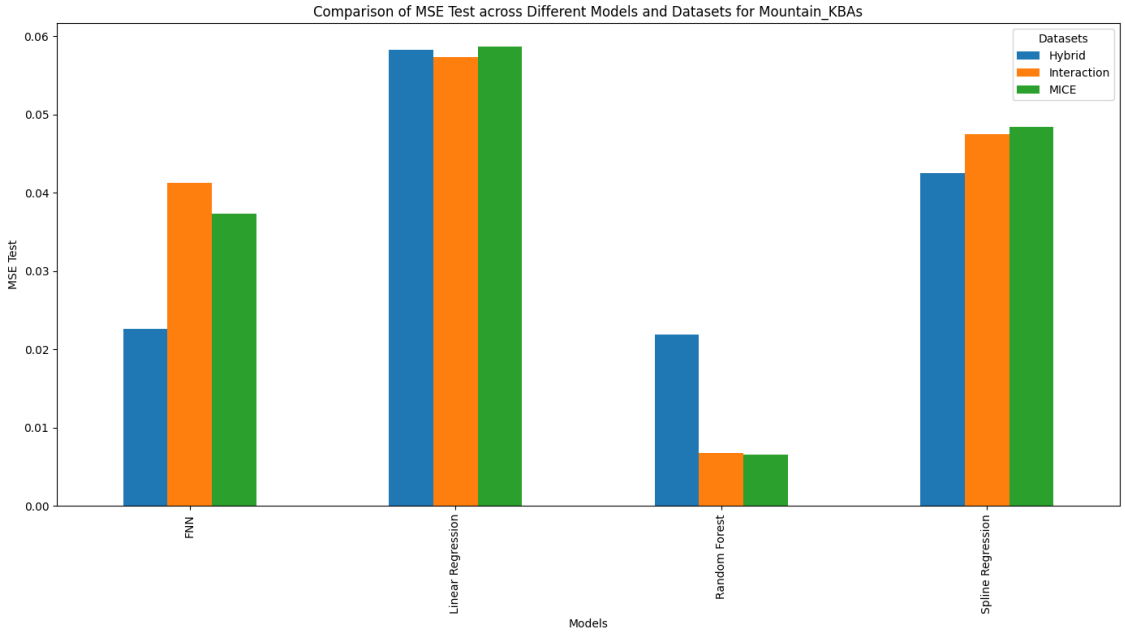

Figure 16: Comparison of MSE Test across Different Models and Datasets for Mountain KBAs

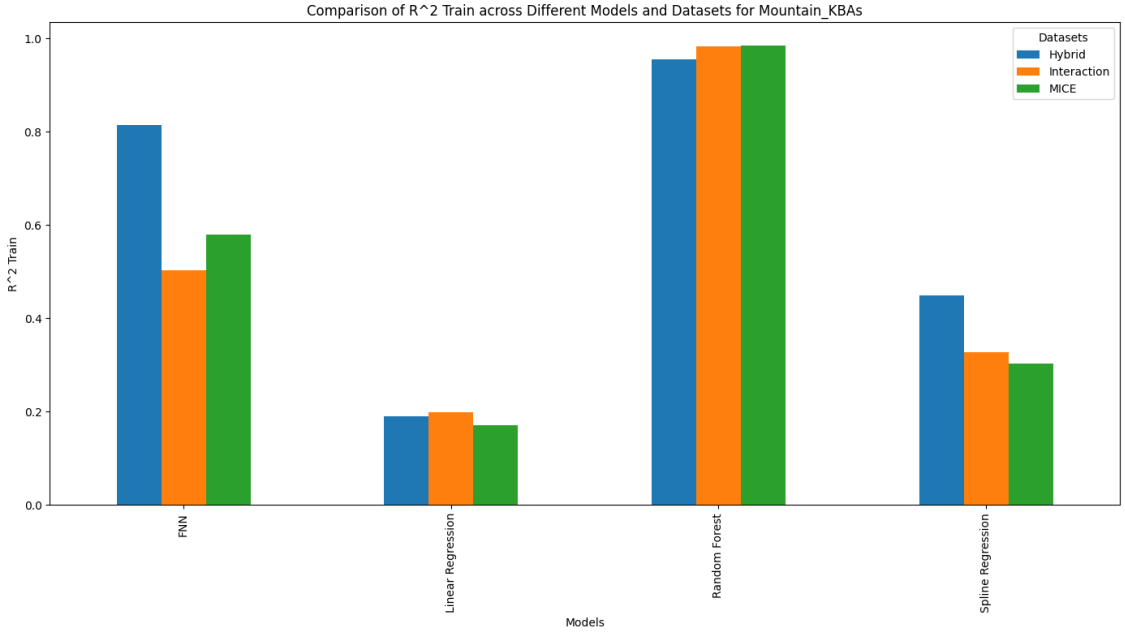

Figure 17: Comparison of R^2 Train across Different Models and Datasets for Mountain KBAs

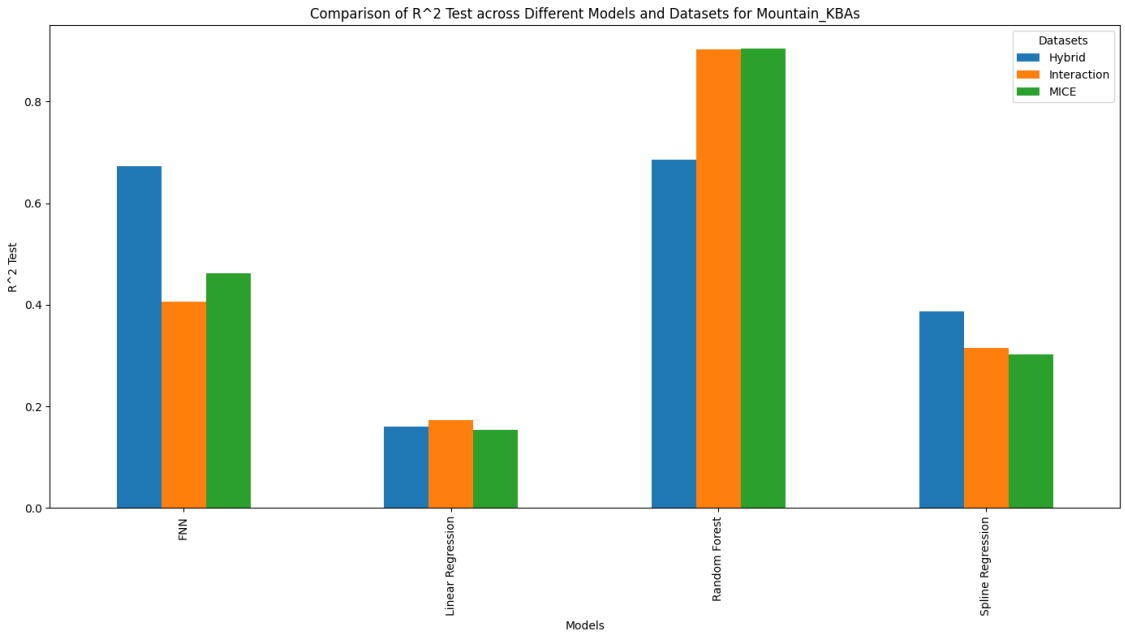

Figure 18: Comparison of R^2 Test across Different Models and Datasets for Mountain KBAs

8.3.1 Linear Regression:

The linear regression model's performance for Mountain KBAs was limited. On the MICE dataset, the model achieved an MSE Train of 0.0577 and an R^2 Train of 0.1704, with an R^2 Test of 0.1539. The interaction term dataset showed a slight improvement (R^2 Test of 0.1733), while the hybrid feature set did not significantly enhance performance (R^2 Test of 0.1602). This highlights the model's inability to capture complex ecological relationships effectively.

8.3.2 Spline Regression:

Spline regression models performed better for Mountain KBAs. The MICE dataset yielded an MSE Train of 0.0485 and an R^2 Train of 0.3030, with an R^2 Test of 0.3029. The interaction term dataset slightly improved performance (R^2 Test of 0.3150), while the hybrid feature set showed the highest R^2 Test of 0.3868. This demonstrates the importance of capturing non-linear relationships and interactions in ecological data. The best parameters for spline regression were {`'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5`}.

8.3.3 Random Forest:

Random forest models excelled in predicting Mountain KBAs. The MICE dataset resulted in an MSE Train of 0.0010 and an R^2 Train of 0.9855, with an R^2 Test of 0.9050, indicating the model explained 90% of the variance. The best parameters were {`'max_depth': 30, 'n_estimators': 100`}. The interaction term dataset showed slightly lower performance (R^2 Test of 0.9023), while the hybrid feature set yielded an R^2 Test of 0.6850. The best parameters for the hybrid feature set were {`'max_depth': 40, 'n_estimators': 1000`}.

8.3.4 Feed-Forward Neural Networks (FNN):

FNNs showed moderate performance for Mountain KBAs. On the MICE dataset, the model achieved an MSE Train of 0.0293 and an R^2 Train of 0.5792, with an R^2 Test of 0.4619. The interaction term dataset performed similarly, with an R^2 Test of 0.4060. The hybrid feature set showed improved performance, with an R^2 Test of 0.6736, highlighting the importance of rich feature sets for enhancing neural network performance. The best parameters for the hybrid feature set were {`'neurons': 128, 'learning_rate': 0.01`}.

8.4 Shap Analysis SDG 15

The SHAP analysis, supplemented by dependence plots, offers a nuanced understanding of feature importance and their relationships with target variables. The SHAP summary plots provide an overview of the importance of each feature in predicting the target indicators, while the SHAP dependence plots illustrate the relationship between individual features and the target variables. Additionally, tables summarizing the mean absolute SHAP values and rankings for each feature offer a quantitative perspective on feature importance. These visual and tabular presentations ensure a comprehensive understanding of the factors influencing each SDG indicator and support the interpretability of the machine learning models used. The section above provides a detailed analysis of the SHAP results for each target indicator, based on findings from SHAP summary and dependence plots, and SHAP summary table 11 detailing the SHAP analysis. Relevant plots and table are provided in the appendix C for a thorough review.

For Forest Cover, the Economic Complexity Index (ECI) emerges as a significant predictor. The SHAP summary plot indicates a generally positive impact of ECI on forest cover, suggesting that higher economic complexity tends to support better forest cover. This is further elucidated by the SHAP dependence plot, which reveals a U-shaped curve. Initially, as ECI increases, SHAP values decrease, indicating a negative impact on forest cover. However, after a certain threshold, further increases in ECI lead to positive SHAP values, suggesting an improvement in forest cover. This pattern aligns well with the EKC hypothesis, where environmental degradation initially increases with economic complexity but decreases after reaching a certain level of development. The GDP per capita dependence plot shows an inverted-U shape, supporting the EKC hypothesis. Initially, increases in GDP per capita result in worsening forest cover, but beyond a certain point, further economic growth improves forest cover. The color-coded feature in this context, population, though scarce in red points, hints at limited interaction between ECI and population in influencing forest cover. Notably, the SHAP summary plots for ECI show positive values, while GDP per capita shows negative values, indicating the complex interplay between these economic indicators and forest cover.

Numerically, the SHAP summary results indicate that ECI has the highest mean absolute SHAP value of 0.036989, ranking first, followed by GDP per capita (0.032868), population (0.030205), and land area (0.029983). These values quantify the average impact of each feature on the model's output, with higher values indicating greater importance. The high SHAP value for ECI corroborates the visual analysis, where ECI's positive impact is clearly visible in the SHAP summary and dependence plots. The GDP per capita's inverted-U pattern is also reflected in its

high SHAP value, signifying its substantial role in explaining forest cover variability.

For Forest Biomass, the Human Development Index (HDI) and ECI are prominent predictors. The SHAP summary plot shows that HDI has a positive impact on forest biomass, indicating that higher human development correlates with better forest biomass. This is reinforced by the SHAP dependence plot for HDI, which displays a wavy positive relationship. As HDI increases, SHAP values generally rise, albeit with fluctuations, indicating a complex but overall beneficial relationship between human development and forest biomass. ECI also exhibits a positive impact, with its dependence plot showing a U-shaped or tick mark pattern. At lower levels, increases in ECI negatively impact forest biomass, but after a certain threshold, further increases in ECI significantly boost forest biomass, again aligning with the EKC hypothesis. The color feature for ECI, with red values concentrated on the right side of the HDI dependence plot, suggests that higher economic complexity is associated with higher HDI, reinforcing the interconnectedness of these variables.

According to the SHAP summary values, HDI has the highest mean absolute SHAP value of 0.024082 for forest biomass, followed by ECI (0.019219), land area (0.016079), and GDP per capita (0.015890). The high SHAP values for HDI and ECI emphasize their significant roles in influencing forest biomass. The positive impact of HDI and the U-shaped impact of ECI are consistent with the observed SHAP dependence plots, where HDI shows a wavy positive relationship and ECI displays a U-shaped pattern.

In the case of Mountain KBAs (Key Biodiversity Areas), land area and GDP per capita are critical predictors. The SHAP summary plot highlights that land area has a significant positive impact on Mountain KBAs, indicating that larger land areas support better conservation of key biodiversity areas. The SHAP dependence plot for land area shows a vertical line with some spread at the tail and outliers, suggesting that while land area is crucial, other factors also significantly contribute to explaining the variability in Mountain KBAs. GDP per capita's dependence plot indicates a somewhat positive linear relationship with Mountain KBAs, although points are quite spread, reflecting the variability in how GDP influences conservation efforts. HDI's dependence plot, showing exponential growth, suggests that improvements in HDI lead to significant increases in SHAP values for Mountain KBAs, underscoring the importance of human development in biodiversity conservation. The color feature for GDP, with HDI points concentrated on the right side, indicates that higher human development is associated with better Mountain KBA outcomes, reinforcing the interplay between economic and human development indicators.

The SHAP summary values show that land area has the highest mean absolute SHAP value of 0.066013 for Mountain KBAs, followed by GDP per capita

(0.045928), HDI (0.038213), and ECI (0.035237). The high SHAP value for land area underscores its critical role in preserving biodiversity, consistent with the SHAP dependence plot that shows a strong positive impact. The positive impact of GDP per capita and HDI, as evidenced by their high SHAP values and dependence plots, highlights the importance of economic and human development in biodiversity conservation efforts.

The SHAP analysis, enriched with dependence plots, underscores the critical role of economic, demographic, and developmental indicators in predicting environmental outcomes. The Economic Complexity Index (ECI), GDP per capita, and Human Development Index (HDI) consistently emerge as top predictors across different target variables, highlighting their importance in environmental management and conservation. The findings support the EKC hypothesis, with ECI and GDP per capita displaying U-shaped and inverted-U patterns, respectively. Initially, economic development may lead to environmental degradation, but beyond a certain threshold, further development promotes better environmental outcomes. Population size and land area significantly influence environmental outcomes, with larger populations providing the necessary workforce for conservation efforts, and larger land areas naturally offering more space for forests and biodiversity hotspots. Investment in Research and Development (RD) is vital for developing innovative solutions for environmental conservation, as higher RD expenditure is associated with better forest cover and biomass, indicating that technological advancements and research play a significant role in environmental sustainability.

Overall, the SHAP analysis provides a comprehensive understanding of the factors influencing environmental indicators. Economic, demographic, and developmental indicators are all crucial for predicting and managing environmental outcomes. The results highlight the need for a multifaceted approach to sustainable development and conservation, emphasizing the importance of economic complexity, wealth, human development, and research investment in achieving positive environmental outcomes. The analysis underscores the importance of robust imputation methods like MICE to ensure data integrity and enhance model performance, particularly when using advanced machine learning models such as random forests and neural networks. The study demonstrates that random forests and neural networks, particularly when combined with advanced feature engineering techniques, offer superior predictive capabilities for complex datasets, providing valuable insights for policymakers and researchers aiming to promote sustainable development and environmental conservation.

9 SDG 14 Modelling Results and Findings

Similar to SDG 15 result section, the results of the various machine learning models—Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN)—are systematically presented below. These tables detail the performance metrics for each model, including Mean Squared Error (MSE) and R^2 values for both training and testing datasets. Each model's performance is evaluated across three datasets: MICE imputed data, an interaction term dataset, and a hybrid feature set. Comprehensive visual presentation of results is shown for each indicator before the model results are descriptively presented. This allows for a clear comparison of model efficacy and highlights the most effective algorithms for predicting each SDG indicator. The table summarizing these results for SDG 14 is provided in the appendix B, as table **??**results$_SDG14tab$: $final\_results_SDG14BeachLitter$

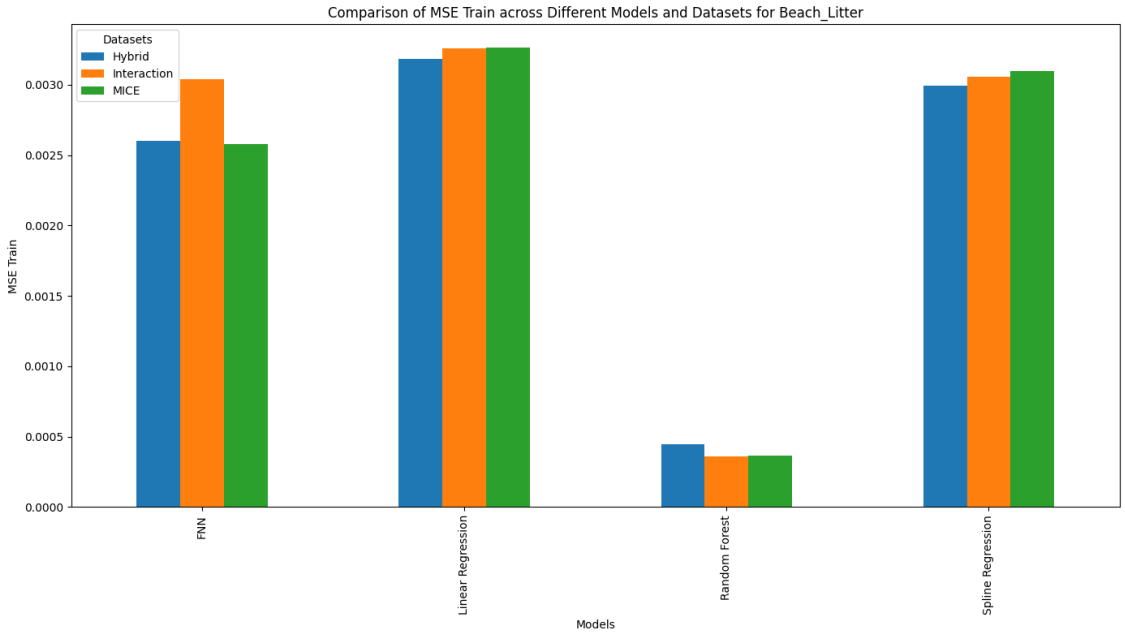

Figure 19: Comparison of MSE Train across Different Models and Datasets for Beach Litter

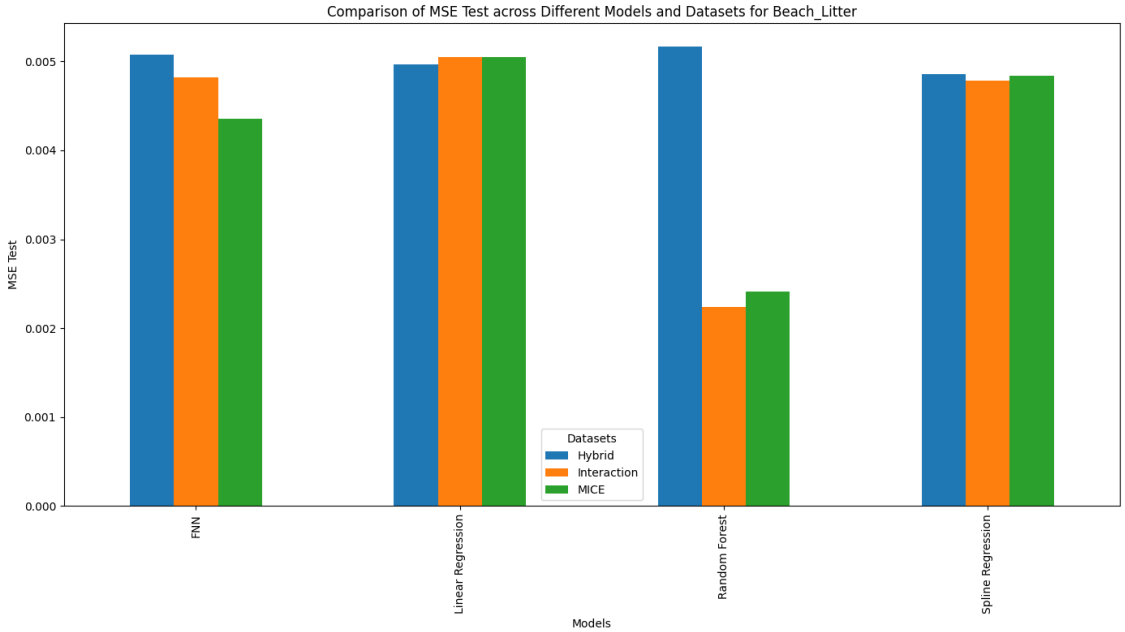

Figure 20: Comparison of MSE Test across Different Models and Datasets for Beach Litter

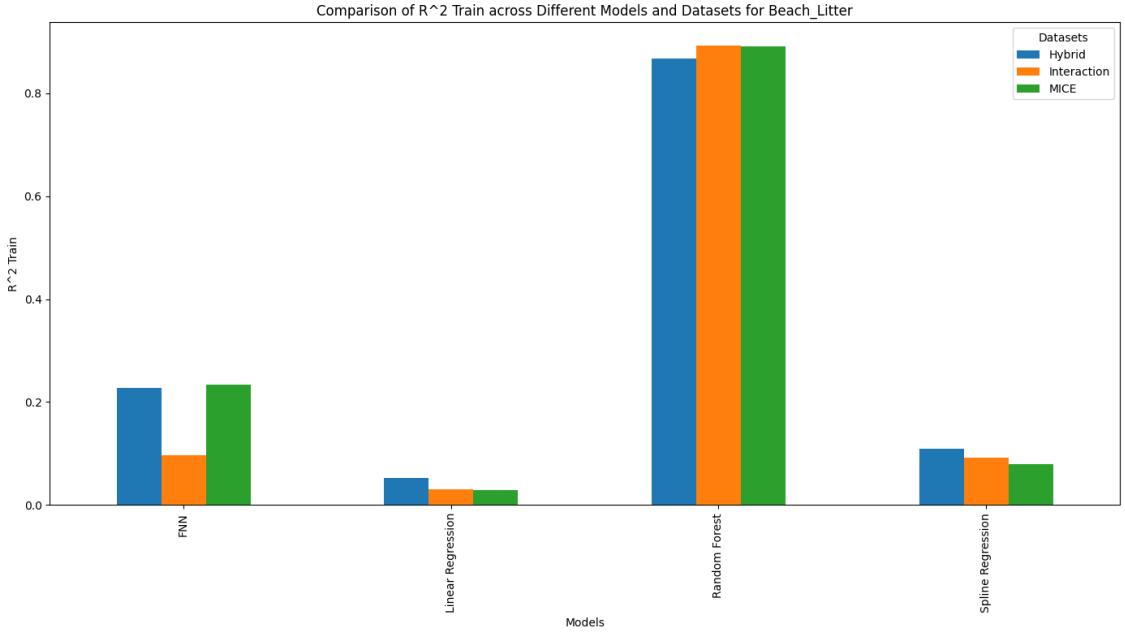

Figure 21: Comparison of R^2 Train across Different Models and Datasets for Beach Litter

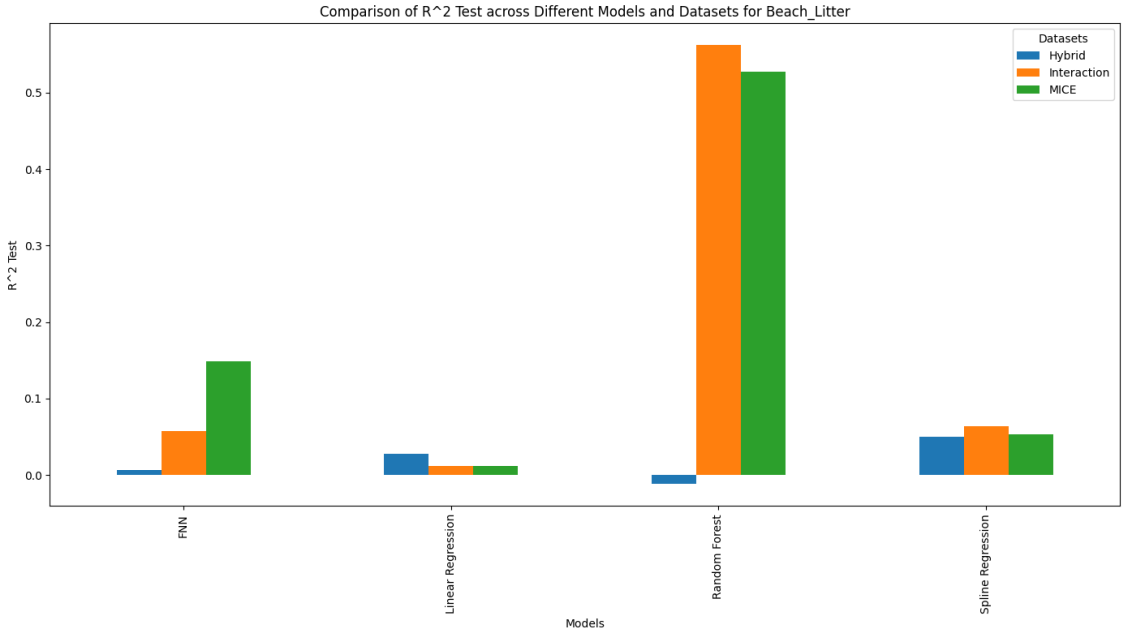

Figure 22: Comparison of R^2 Test across Different Models and Datasets for Beach Litter

9.0.1 Linear Regression

For Beach Litter, Linear Regression displayed varied performance across the different datasets. On the MICE dataset, the model achieved an MSE of 0.00327 for training and 0.00505 for testing, with corresponding R^2 values of 0.029 and 0.012, indicating limited explanatory power. The interaction term dataset yielded similar results, with an MSE of 0.00326 for training and 0.00505 for testing, and R^2 values of 0.031 and 0.012. The hybrid dataset showed slight improvements, achieving an MSE of 0.00319 for training and 0.00497 for testing, and R^2 values of 0.053 and 0.028, suggesting a marginally better fit.

9.0.2 Spline Regression

Spline Regression offered a significant improvement over Linear Regression by accommodating non-linear relationships. Using the MICE dataset, the model produced an MSE of 0.00310 for training and 0.00484 for testing, with R^2 values of 0.080 and 0.054. The interaction term dataset further improved the model's performance, with an MSE of 0.00306 for training and 0.00478 for testing, and R^2 values of 0.091 and 0.064. The hybrid dataset achieved the best performance, with an MSE of 0.00300 for training and 0.00485 for testing, and R^2 values of 0.110 and 0.050. The optimal parameters for Spline Regression were a ridge alpha of 0.1, a spline transformer degree of 3 (cubic splines), and 5 knots.

9.0.3 Random Forest

Random Forest models significantly improved performance for Beach Litter predictions. For the MICE dataset, the model achieved an MSE of 0.00037 for training and 0.00241 for testing, with R^2 values of 0.891 and 0.528, indicating a high degree of fit. The interaction term dataset yielded an MSE of 0.00036 for training and 0.00224 for testing, with R^2 values of 0.893 and 0.562. However, the hybrid dataset showed poorer testing performance, with an MSE of 0.00045 for training and 0.00517 for testing, and R^2 values of 0.868 and -0.012, suggesting potential overfitting. The best parameters included a max depth of 20 and 200 estimators.

9.0.4 Feed-Forward Neural Networks (FNN)

Feed-Forward Neural Networks (FNN) also demonstrated promising results for Beach Litter. The MICE dataset resulted in an MSE of 0.00258 for training and 0.00435 for testing, with R^2 values of 0.234 and 0.148. The interaction term dataset performed similarly, with an MSE of 0.00304 for training and 0.00482 for testing, and R^2 values of 0.097 and 0.057. The hybrid dataset resulted in an MSE of 0.00260 for

training and 0.00507 for testing, with R^2 values of 0.227 and 0.007. The optimal parameters for FNN included 128 neurons and a learning rate of 0.001.

9.1 Chlorophyll

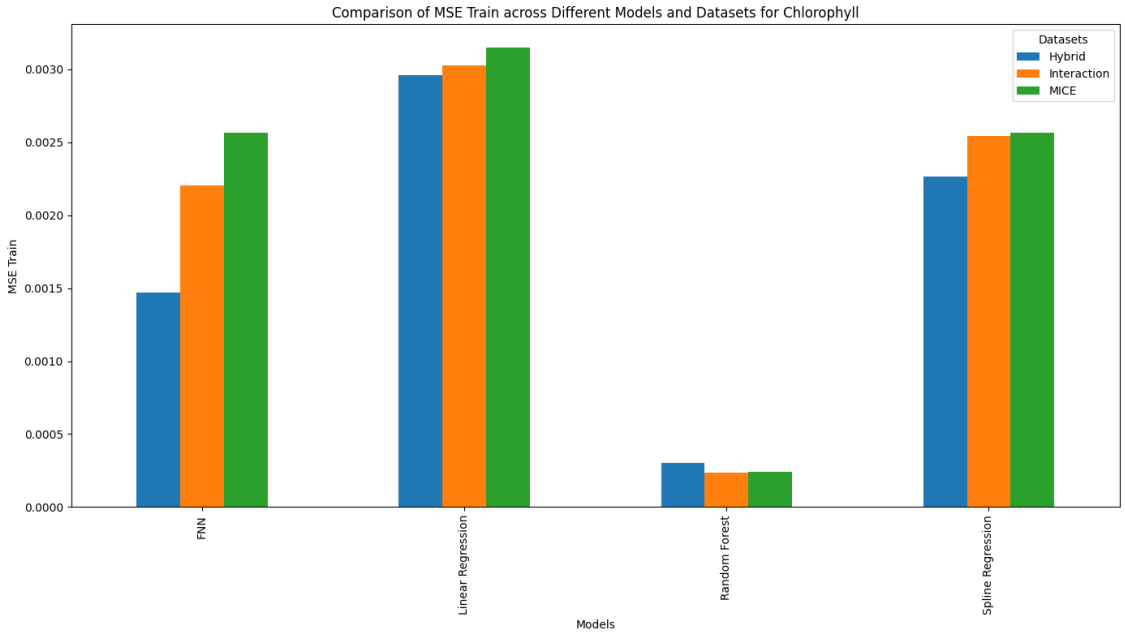

Figure 23: Comparison of MSE Train across Different Models and Datasets for Chlorophyll

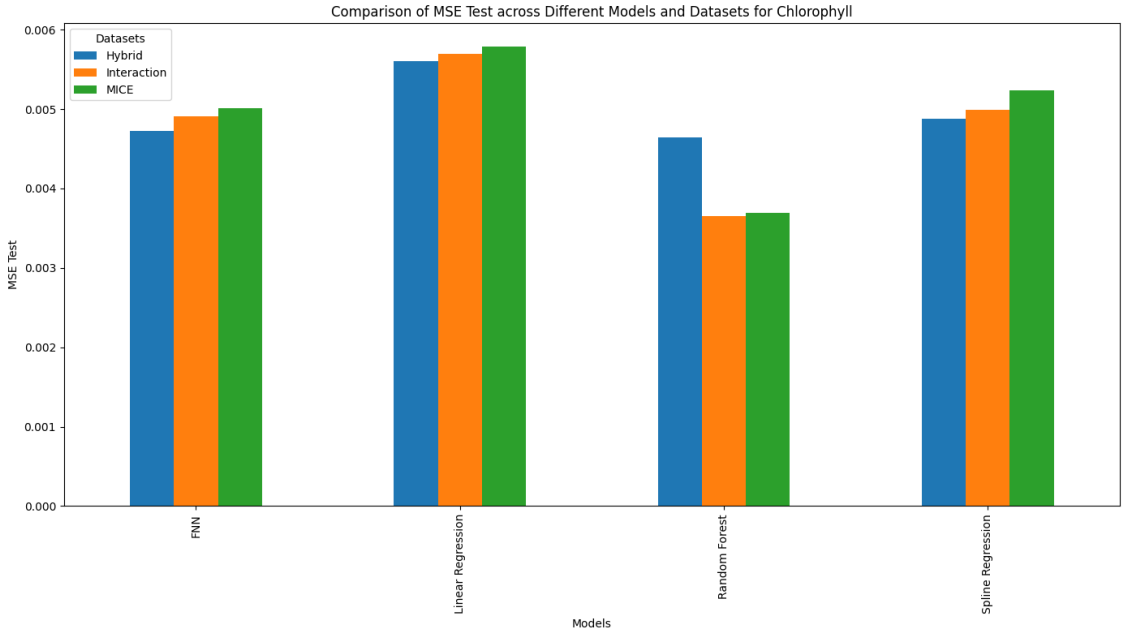

Figure 24: Comparison of MSE Test across Different Models and Datasets for Chlorophyll

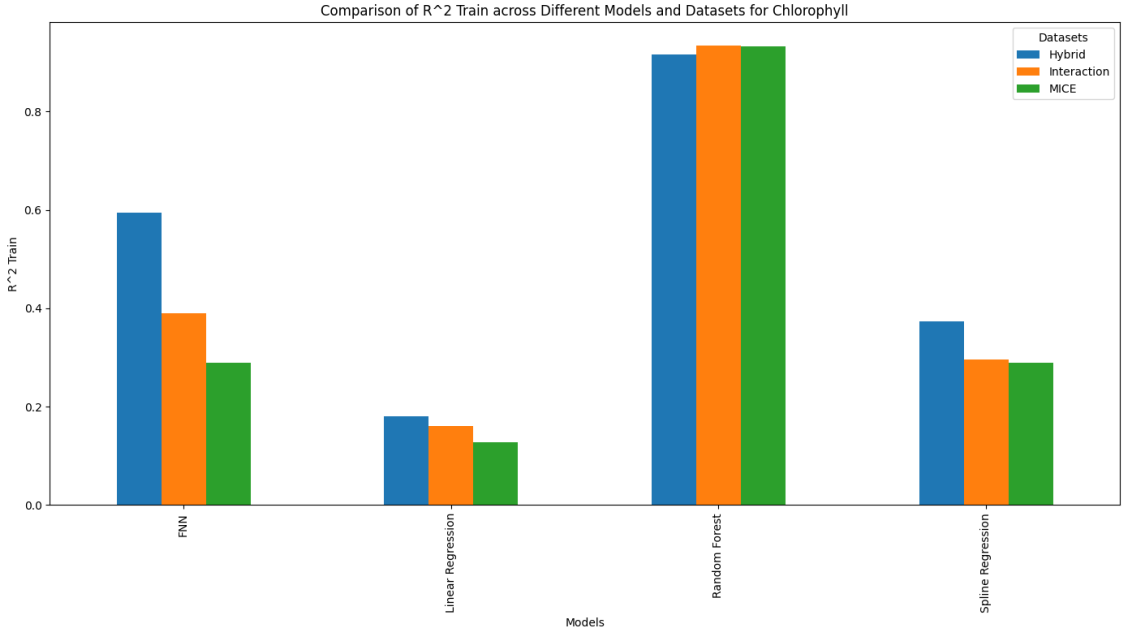

Figure 25: Comparison of R^2 Train across Different Models and Datasets for Chlorophyll

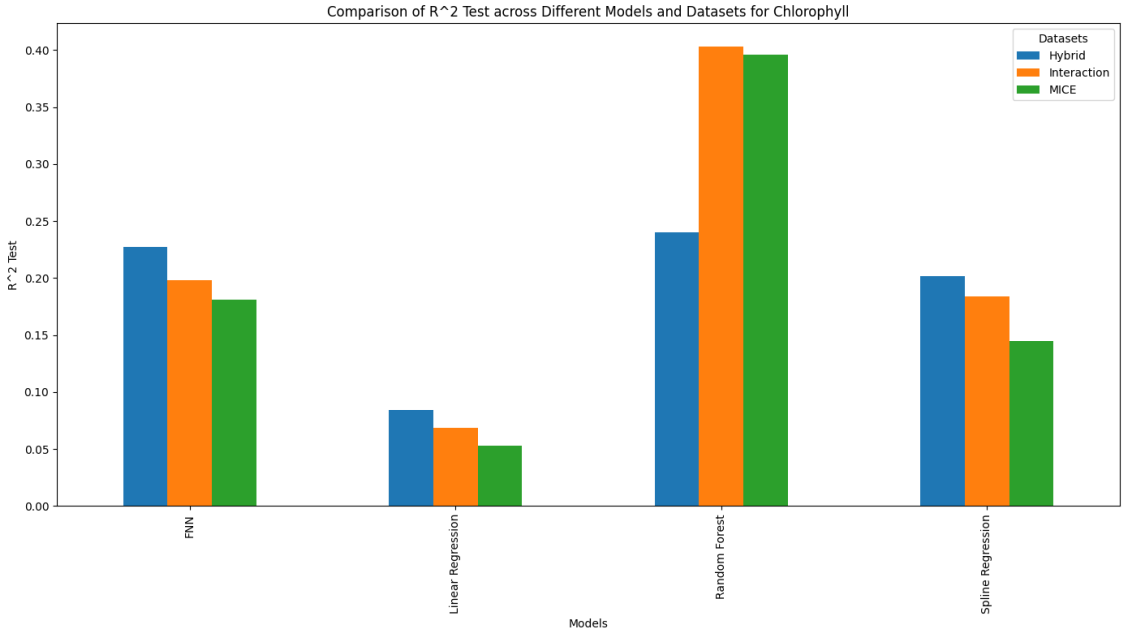

Figure 26: Comparison of R^2 Test across Different Models and Datasets for Chlorophyll

9.1.1 Linear Regression

For Chlorophyll, Linear Regression demonstrated moderate performance. On the MICE dataset, the model achieved an MSE of 0.00315 for training and 0.00579 for testing, with R^2 values of 0.127 and 0.053, indicating limited predictive power. The interaction term dataset improved slightly, with an MSE of 0.00303 for training and 0.00570 for testing, and R^2 values of 0.161 and 0.068. The hybrid dataset performed the best, achieving an MSE of 0.00296 for training and 0.00560 for testing, and R^2 values of 0.181 and 0.084, reflecting better capture of data variance.

9.1.2 Spline Regression

Spline Regression outperformed Linear Regression for Chlorophyll predictions. On the MICE dataset, the model achieved an MSE of 0.00257 for training and 0.00523 for testing, with R^2 values of 0.289 and 0.145. The interaction term dataset showed further improvement, with an MSE of 0.00254 for training and 0.00499 for testing, and R^2 values of 0.295 and 0.184. The hybrid dataset achieved the best performance, with an MSE of 0.00226 for training and 0.00488 for testing, and R^2 values of 0.373 and 0.202. The optimal parameters included a ridge alpha of 0.1, a spline transformer degree of 3, and 5 knots.

9.1.3 Random Forest

Random Forest models achieved the highest performance for Chlorophyll predictions. For the MICE dataset, the model achieved an MSE of 0.00024 for training and 0.00369 for testing, with R^2 values of 0.933 and 0.396. The interaction term dataset performed similarly, with an MSE of 0.00024 for training and 0.00365 for testing, and R^2 values of 0.934 and 0.403. The hybrid dataset also maintained high performance, with an MSE of 0.00030 for training and 0.00465 for testing, and R^2 values of 0.916 and 0.240. The optimal parameters included a max depth of 20 and 1000 estimators.

9.1.4 Feed-Forward Neural Networks (FNN)

Feed-Forward Neural Networks (FNN) showed competitive performance for Chlorophyll predictions. On the MICE dataset, the model achieved an MSE of 0.00257 for training and 0.00501 for testing, with R^2 values of 0.289 and 0.181. The interaction term dataset performed similarly, with an MSE of 0.00221 for training and 0.00491 for testing, and R^2 values of 0.389 and 0.198. The hybrid dataset showed the best performance, achieving an MSE of 0.00147 for training and 0.00473 for testing, with R^2 values of 0.594 and 0.227. The optimal parameters included 128 neurons and a learning rate of 0.001.

9.2 Sustainable Fisheries

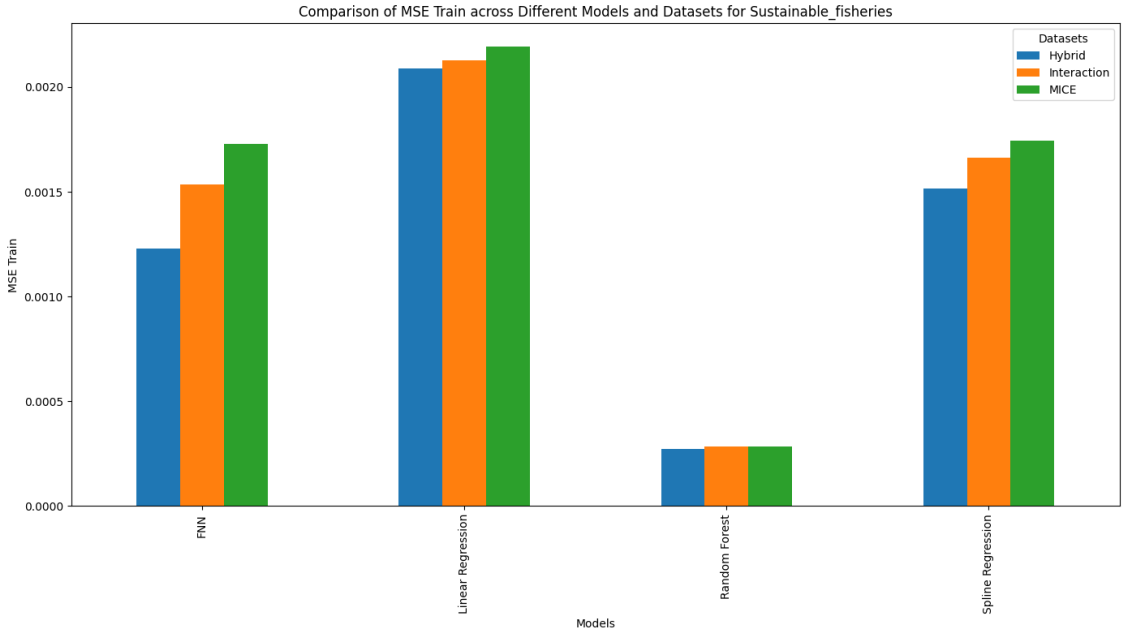

Figure 27: Comparison of MSE Train across Different Models and Datasets for Sustainable Fisheries

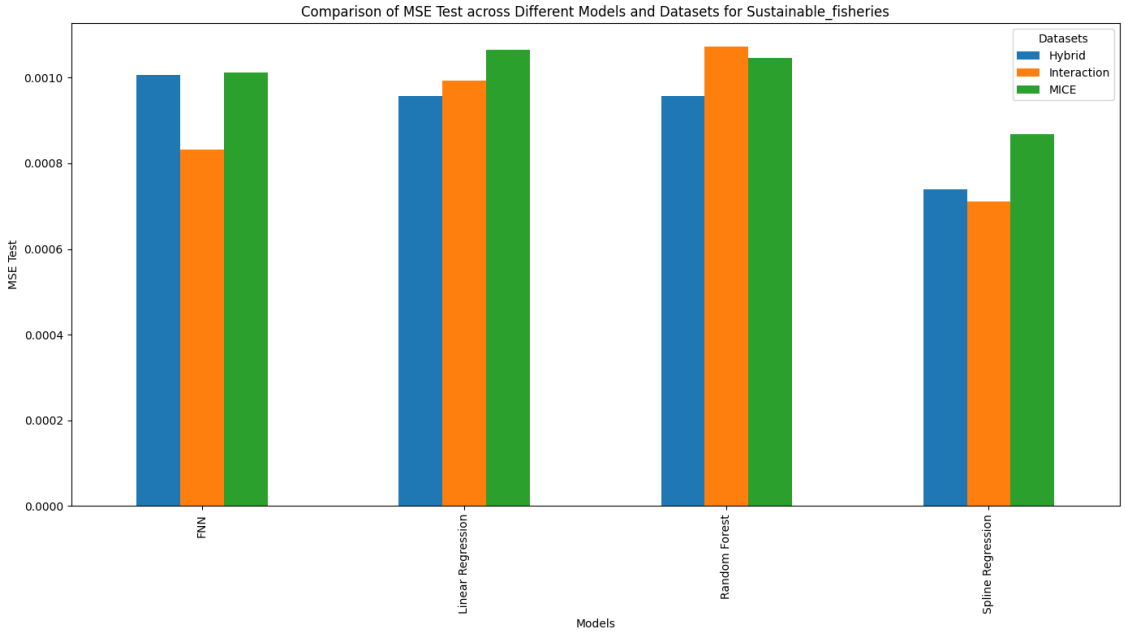

Figure 28: Comparison of MSE Test across Different Models and Datasets for Sustainable Fisheries

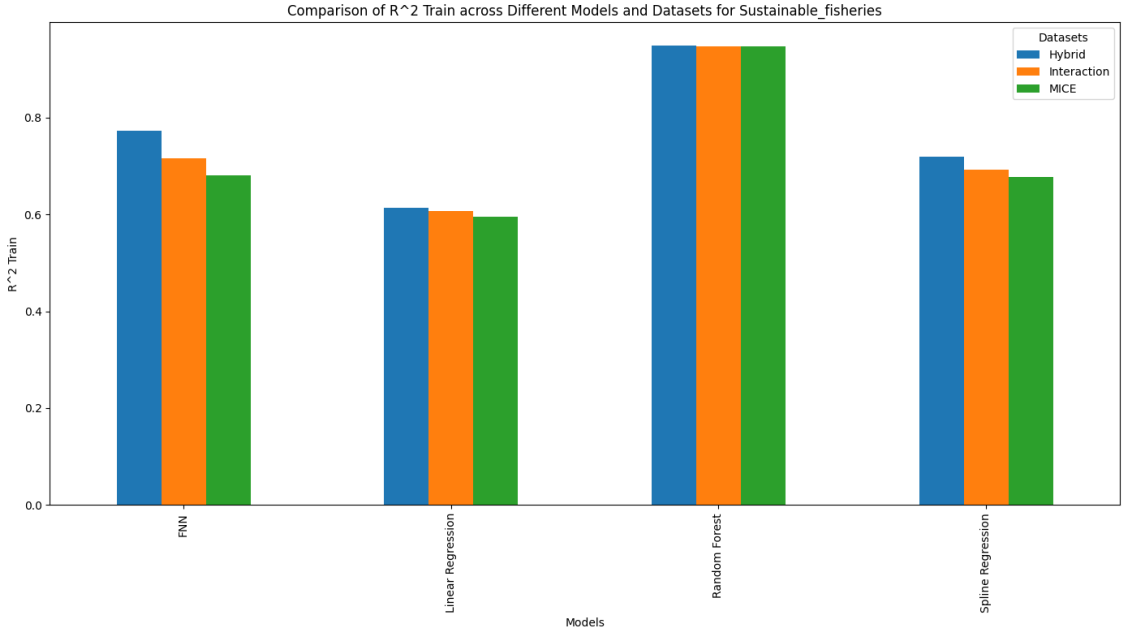

Figure 29: Comparison of R^2 Train across Different Models and Datasets for Sustainable Fisheries

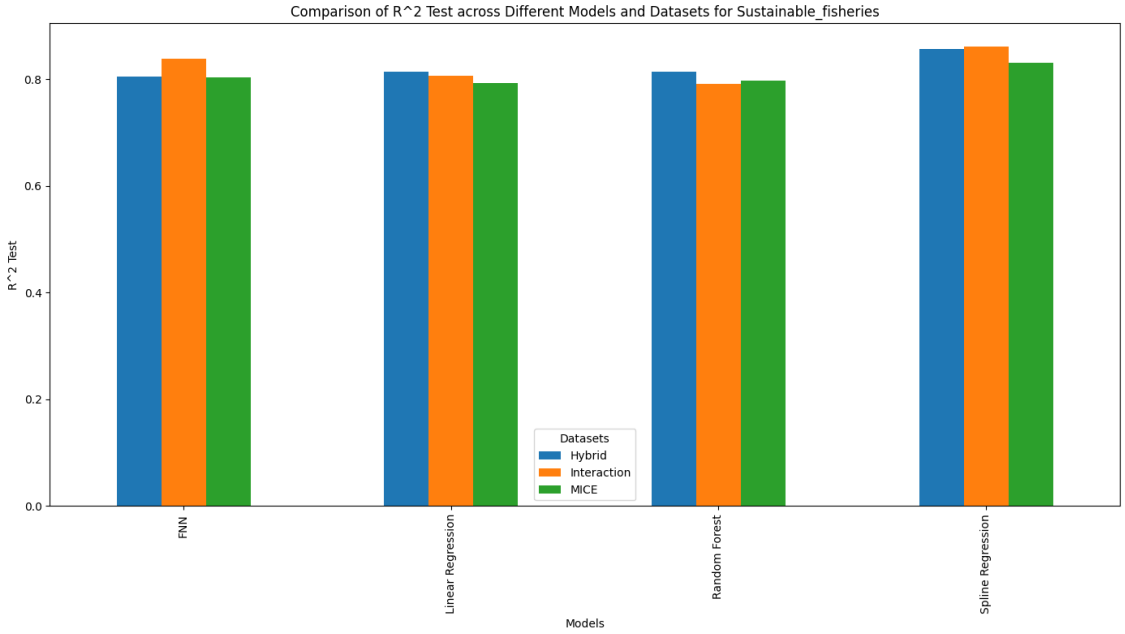

Figure 30: Comparison of R^2 Test across Different Models and Datasets for Sustainable Fisheries

9.2.1 Linear Regression

Linear Regression models performed well for Sustainable Fisheries. On the MICE dataset, the model achieved an MSE of 0.00219 for training and 0.00106 for testing, with R^2 values of 0.595 and 0.793, indicating a good fit and suggesting effective capture of variance. The interaction term dataset showed similar performance, with an MSE of 0.00213 for training and 0.00099 for testing, and R^2 values of 0.607 and 0.807. The hybrid dataset performed slightly better, achieving an MSE of 0.00209 for training and 0.00096 for testing, with R^2 values of 0.614 and 0.814.

9.2.2 Spline Regression

Spline Regression models demonstrated strong performance for Sustainable Fisheries. On the MICE dataset, the model achieved an MSE of 0.00174 for training and 0.00087 for testing, with R^2 values of 0.678 and 0.831, indicating the model's ability to capture complex non-linear relationships. The interaction term dataset performed better, with an MSE of 0.00166 for training and 0.00071 for testing, and R^2 values of 0.693 and 0.861. The hybrid dataset achieved the highest performance, with an MSE of 0.00152 for training and 0.00074 for testing, and R^2 values of 0.720 and 0.856. The optimal parameters included a ridge alpha of 0.1, a spline transformer degree of 3, and 5 knots.

9.2.3 Random Forest

Random Forest models achieved high performance for Sustainable Fisheries but did not outperform Spline Regression. On the MICE dataset, the model achieved an MSE of 0.00028 for training and 0.00105 for testing, with R^2 values of 0.948 and 0.796. The interaction term dataset showed similar performance, with an MSE of 0.00028 for training and 0.00107 for testing, and R^2 values of 0.947 and 0.791. The hybrid dataset maintained high performance, achieving an MSE of 0.00027 for training and 0.00096 for testing, with R^2 values of 0.949 and 0.814. The optimal parameters included a max depth of 20 and 1000 estimators.

9.2.4 Feed-Forward Neural Networks (FNN)

Feed-Forward Neural Networks (FNN) showed strong performance for Sustainable Fisheries across all datasets. The interaction term dataset provided the best results, with an MSE of 0.00153 for training and 0.00083 for testing, with R^2 values of 0.717 and 0.838. The optimal parameters included 128 neurons and a learning rate of 0.001.

9.3 Shap Analysis SDG 14

Similarly to the SDG 15, The SHAP analysis results for SDG 14 modelling are presented based on the visual and tabular results of SHAP, to ensure a comprehensive understanding of the factors influencing each SDG 14 indicator and support the interpretability of the machine learning models used. The relevant figures and SHAP summary table for SDG 14 (12), detailing the SHAP analysis are provided in the appendix C for a thorough review.

The SHAP analysis table of SDG 14 12 provides a quantitative summary of these insights. For Beach Litter, RD expenditure, Change Over 5 Years, and Land area are the top three features, with mean absolute SHAP values of 0.0088, 0.0044, and 0.0040, respectively. For Chlorophyll, Land area, Population, and COI are the top features, with mean absolute SHAP values of 0.012, 0.010, and 0.0055. For Sustainable Fisheries, COI, RD expenditure, and GDP per capita lead the ranking with mean absolute SHAP values of 0.032, 0.013, and 0.012. These values quantitatively confirm the qualitative insights from the dependence plots, highlighting the varying degrees of influence these features have on the target indicators.

For Beach Litter, the SHAP summary plot indicates that most features have SHAP values concentrated around zero, with predominantly blue dots, suggesting minimal influence on the target variable. The top three features, according to the mean absolute SHAP values, are RD expenditure (0.0088), Change Over 5 Years (0.0044), and Land area (0.0040). These values indicate that RD expenditure has the most substantial impact on Beach Litter, but even this influence is relatively small, as reflected by the low SHAP values. The SHAP dependence plot for RD expenditure shows a nearly horizontal line around zero on the y-axis, implying that variations in RD expenditure have a negligible effect on beach litter. Similarly, the dependence plot for Change Over 5 Years also exhibits a horizontal line, indicating minimal impact. The dependence plot for Land area shows a cluster of points in the bottom left corner, suggesting that smaller land areas are associated with less beach litter, though the overall impact is still limited. The presence of blue points and the concentration around zero in the SHAP summary plot corroborate this minimal influence, highlighting the limited predictive power of these features for Beach Litter.

In the case of Chlorophyll, the SHAP summary plot reveals a more varied distribution of SHAP values, with Land area showing both blue dots around zero and some red dots expanding towards the right, indicating a more significant impact. The top feature, Land area, has a mean absolute SHAP value of 0.012, indicating its importance in predicting chlorophyll levels. This value suggests a moderate influence, with larger land areas having a more substantial effect on chlorophyll levels. The SHAP dependence plot for Land area shows a vertical line with a cluster of points at the

top, suggesting a strong influence of larger land areas on higher chlorophyll levels. The dependence plot for Population, the second top feature with a mean absolute SHAP value of 0.010, also shows a vertical line, reinforcing the significant role of demographic factors. The COI dependence plot exhibits an inverted-U shape, indicating that moderate COI values are associated with higher chlorophyll levels, while very high or very low COI values have a lesser impact. This pattern aligns with the Environmental Kuznets Curve (EKC) hypothesis, where environmental degradation initially increases with economic complexity but decreases after a certain threshold. The red points in the summary plot for Human Development Index (HDI) and Land area expanding towards the right further illustrate their positive influence on chlorophyll levels.

For Sustainable Fisheries, the SHAP summary plot highlights COI as the most influential feature, with a mean absolute SHAP value of 0.032. This relatively high value indicates a strong impact, with COI significantly affecting sustainable fisheries. The plot shows red points on the left of zero and blue points on the right, indicating that higher COI values negatively impact sustainable fisheries. This negative relationship is confirmed by the SHAP dependence plot for COI, which shows a negative linear trend. The dependence plot for RD expenditure, the second most influential feature with a mean absolute SHAP value of 0.013, shows a positive relationship with a shallow slope, suggesting that increased RD investment slightly enhances sustainable fisheries outcomes. The color coding indicates that COI values increase with RD expenditure, as evidenced by the red points moving rightward. The dependence plot for GDP per capita, which has a mean absolute SHAP value of 0.012, shows a negative trend, with the color coding for ECI indicating that higher economic complexity correlates with reduced sustainable fisheries outcomes. This observation aligns with the notion that economic activities can adversely affect fisheries unless managed sustainably. The red points for RD expenditure spreading in the positive direction and GDP spreading in the negative direction in the SHAP summary plot reinforce the significant but divergent impacts of these features.

10 Empirical Analysis of Machine Learning Models for SDG 15 and SDG 14

The detailed analysis of machine learning models for SDG 15 and SDG 14 indicators provides an in-depth understanding of their predictive performance. This section intricately compares the performance of Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN) across three datasets: MICE imputed data, an interaction term dataset, and a hybrid feature set. This compre-

hensive comparison highlights the best-performing models and explores the nuances of each model's performance across different scenarios, integrating an in-depth exploration of SHAP findings and their implications.

For SDG 15, which includes indicators for Forest Cover, Forest Biomass, and Mountain Key Biodiversity Areas (KBAs), Random Forest models consistently exhibited the best performance. Specifically, for Forest Cover, the Random Forest model using the MICE dataset achieved an MSE Train of 0.0007 and an R^2 Train of 0.9835, with an R^2 Test of 0.8907. This high accuracy indicates the model's robustness in handling non-linear relationships and its ability to explain a significant portion of the variance in the test data. The SHAP analysis further illuminates these results, with the Economic Complexity Index (ECI) emerging as the most significant predictor. The SHAP summary plot indicates a generally positive impact of ECI on forest cover, with a mean absolute SHAP value of 0.037, suggesting that higher economic complexity tends to support better forest cover. The SHAP dependence plot reveals a U-shaped relationship, where initial increases in ECI decrease SHAP values (indicating a negative impact), but beyond a certain threshold, further increases lead to positive SHAP values. This pattern aligns with the Environmental Kuznets Curve (EKC) hypothesis, suggesting that economic development initially leads to environmental degradation, but further development improves environmental outcomes. The GDP per capita dependence plot shows an inverted-U shape, supporting the EKC hypothesis. Initially, increases in GDP per capita worsen forest cover, but beyond a certain point, further economic growth improves it. These intricate relationships captured by SHAP analysis underscore the complex interplay between economic factors and forest cover, providing a nuanced understanding of the model's predictions.

The implications of these findings are significant. The positive impact of ECI, as revealed by the SHAP values, underscores the importance of economic complexity in promoting sustainable forest management. It is arguable that countries with higher ECI have the technological and institutional frameworks necessary to implement effective forest conservation strategies. The inverted-U relationship with GDP per capita suggests that economic growth, up to a certain point, can exacerbate deforestation due to increased demand for land and resources. However, once a country reaches a higher economic status, it may invest more in environmental protection and sustainable practices. This illustrates the critical balance between economic development and environmental sustainability, emphasizing the need for targeted policies that support economic growth while mitigating its environmental impacts.

Spline Regression models, while not matching the overall performance of Random Forests, provided significant improvements over Linear Regression by effectively capturing non-linear relationships. But even for Forest Biomass, the Random Forest

model using the feature set from mice achieved an MSE Train of 0.0002 and an R^2 Train of 0.987, with an R^2 Test of 0.889. The SHAP analysis highlights the HDI and ECI as prominent predictors, with mean absolute SHAP values of 0.024 and 0.019, respectively. The dependence plots reveal a wavy positive relationship for HDI, indicating that higher HDI correlates with better forest biomass, albeit with some fluctuations. ECI shows a U-shaped or tick mark pattern, where lower ECI values negatively impact forest biomass, but beyond a certain threshold, further increases significantly boost it. This complex interaction reflects the EKC hypothesis, where economic complexity initially harms but eventually benefits environmental outcomes. These insights from SHAP analysis demonstrate the model's ability to capture the intricate dependencies between human development, economic complexity, and forest biomass.

The implications of these SHAP findings are profound. The positive impact of HDI suggests that higher human development levels, which encompass education, health, and income, are crucial for sustainable forest management. This reinforces the argument that improving socio-economic conditions can lead to better environmental stewardship. The U-shaped impact of ECI further supports the notion that economic complexity can drive both negative and positive environmental outcomes, depending on the level of development. This emphasizes the need for policies that foster economic complexity and human development concurrently, ensuring that advancements in these areas translate into positive environmental impacts.

For Mountain KBAs, Random Forest models again outperformed other approaches. The MICE dataset resulted in an MSE Train of 0.0010 and an R^2 Train of 0.9855, with an R^2 Test of 0.9050. The SHAP analysis identified land area and GDP per capita as critical predictors, with mean absolute SHAP values of 0.066 and 0.046, respectively. The dependence plot for land area shows a vertical line with some spread at the tail, indicating that while land area is crucial, other factors also significantly contribute to explaining the variability in Mountain KBAs. GDP per capita's dependence plot indicates a somewhat positive linear relationship, with points spread out, reflecting the variability in how GDP influences conservation efforts. HDI's dependence plot, showing exponential growth, suggests that improvements in HDI lead to significant increases in SHAP values for Mountain KBAs, highlighting the importance of human development in biodiversity conservation. The color-coded features in the dependence plots further illustrate the interplay between economic and human development indicators, reinforcing the multifaceted nature of biodiversity conservation.

The implications of these findings are substantial. The high SHAP value for land area underscores its critical role in preserving biodiversity, as larger land areas provide more habitats for species. The positive impact of GDP per capita suggests

that wealthier countries may have more resources to invest in conservation efforts, though the variability in points indicates that this relationship is not straightforward. The exponential growth relationship with HDI highlights the importance of human development in biodiversity conservation, suggesting that improving health, education, and income levels can significantly enhance conservation outcomes. This underscores the argument for integrated policies that address both human development and environmental conservation to achieve sustainable biodiversity outcomes.

For SDG 14 indicators—Beach Litter, Chlorophyll, and Sustainable Fisheries—the performance patterns varied. For Beach Litter, Random Forest models using the MICE dataset achieved an MSE Train of 0.00037 and an R^2 Train of 0.891, with an R^2 Test of 0.528. The SHAP analysis indicates that RD expenditure, Change Over 5 Years, and Land area are the top features, with mean absolute SHAP values of 0.0088, 0.0044, and 0.0040, respectively. The SHAP dependence plots show that variations in RD expenditure and Change Over 5 Years have negligible effects, as indicated by nearly horizontal lines around zero on the y-axis. Land area's plot clusters points in the bottom left, suggesting that smaller land areas are associated with less beach litter, though the overall impact is limited. These SHAP values and dependence plots corroborate the model's limited predictive power for these features, indicating that other, less influential factors might also play a role in beach litter outcomes.

The implications here are clear: while RD expenditure shows some influence, the overall impact on beach litter is minimal, suggesting that beach litter is influenced by a complex interplay of factors not fully captured by the current models. This implies a need for further research to identify additional predictors and to develop more comprehensive models that can account for the multifaceted nature of beach litter accumulation.

For Chlorophyll, both Random Forest and Spline Regression models demonstrated superior performance. The Random Forest model on the MICE dataset achieved an MSE Train of 0.00024 and an R^2 Train of 0.933, with an R^2 Test of 0.396. SHAP analysis highlights Land area, Population, and COI as crucial features, with mean absolute SHAP values of 0.012, 0.010, and 0.0055, respectively. The dependence plot for Land area shows a strong influence on higher chlorophyll levels, while the inverted-U shape for COI suggests that moderate COI values optimize chlorophyll levels. This pattern supports the EKC hypothesis, indicating that moderate economic activities balance environmental health and economic growth. The positive influence of HDI and Land area, illustrated by the expansion of red points in the SHAP summary plot, argues for policies that promote human development alongside environmental conservation to enhance chlorophyll levels.

For Sustainable Fisheries, the performance was more evenly distributed across

models. Spline Regression models, however, emerged as slightly superior, especially with the hybrid dataset, achieving an MSE Train of 0.00152 and an R^2 Train of 0.720, with an R^2 Test of 0.856. SHAP analysis identifies COI, RD expenditure, and GDP per capita as top features, with mean absolute SHAP values of 0.032, 0.013, and 0.012, respectively. The negative linear trend for COI in the dependence plot indicates that higher COI values adversely affect sustainable fisheries, while the positive relationship for RD expenditure suggests that increased research investment enhances fisheries sustainability. The nuanced negative trend for GDP per capita implies that higher economic complexity, without sustainable management, can harm fisheries, highlighting the need for policies that balance economic growth with sustainable resource management.

Feed-Forward Neural Networks (FNNs) displayed varied performance across different indicators and datasets. While not always the best, FNNs showed significant potential, particularly when using the hybrid dataset. For example, in predicting Sustainable Fisheries, FNNs achieved competitive performance, with an MSE Train of 0.00153 and an R^2 Train of 0.717, with an R^2 Test of 0.838. This indicates that while FNNs can capture complex relationships, they require sophisticated feature engineering to perform well. For Beach Litter and Chlorophyll, FNNs demonstrated comparable performance to Spline Regression, particularly with the hybrid dataset, suggesting their utility in scenarios where non-linear relationships are prevalent.

Overall, the comparative analysis reveals that Random Forest models consistently offer superior performance for both SDG 15 and SDG 14 indicators, particularly with the MICE imputed dataset. Spline Regression models also demonstrate significant capabilities, especially in capturing non-linear relationships. The SHAP analysis provides valuable interpretability, highlighting critical features and their impacts on the target indicators. These insights emphasize the importance of advanced machine learning techniques, robust imputation methods, and sophisticated feature engineering in achieving accurate predictions for sustainable development indicators. The findings underscore the necessity for a multifaceted approach in promoting sustainable development and environmental conservation, leveraging the strengths of different modeling techniques and comprehensive feature analysis. This integrated approach, supported by the nuanced insights from SHAP analysis, offers a pathway for policymakers and researchers to design effective strategies that balance economic development with environmental sustainability.

11 Discussion and Conclusion

The research presented in this thesis delves deeply into the intricate relationships between economic complexity (ECI), human development (HDI), and environmental sustainability, specifically within the contexts of Sustainable Development Goals (SDGs) 14 (Life Below Water) and 15 (Life on Land). This comprehensive analysis leverages advanced machine learning models to predict environmental outcomes and elucidate the dynamics driving these relationships. Our conclusion interweaves theoretical frameworks, empirical findings, and broader implications, providing an insightful synthesis of our study.

11.1 Linking Back to Theoretical Framework

Our investigation is anchored in Numan's (2022) conceptual model, which positions the environment as the sustaining dimension and economic and social progress as the developing dimensions. This model emphasizes the necessity of balancing these dimensions to achieve genuine sustainable development. Our findings substantiate this balance by revealing how economic complexity and human development interact to influence environmental sustainability.

Economic complexity, quantified by the Economic Complexity Index (ECI), encapsulates a country's production capabilities and the sophistication of its economic activities. High economic complexity often leads to diversified and technologically advanced industries, which can exert both positive and negative environmental impacts. Our study indicates that countries with higher ECI tend to have better forest cover and biomass, aligning with the Environmental Kuznets Curve (EKC) hypothesis. This hypothesis posits that environmental degradation initially escalates with economic growth but eventually diminishes as economies become more complex and capable of implementing sustainable practices.

Human development, measured by the Human Development Index (HDI), encompasses education, health, and income. Higher HDI levels are generally associated with greater environmental awareness and the capacity to implement sustainable practices. Our analysis reveals that HDI positively impacts environmental indicators such as forest biomass and mountain KBAs, suggesting that enhancements in human capital lead to superior environmental outcomes.

11.2 Empirical Findings and SHAP Analysis

Our predictive modeling approach, utilizing Linear Regression, Spline Regression, Random Forest, and Feed-Forward Neural Networks (FNN), offers nuanced insights into the relationships between ECI, HDI, and environmental sustainability. The

Random Forest models consistently exhibit superior performance, particularly for SDG 15 indicators like forest cover and mountain KBAs.

The SHAP (SHapley Additive exPlanations) analysis provides a granular understanding of the relative importance and interactions of various predictors. For forest cover, the SHAP analysis underscores the Economic Complexity Index (ECI) as the paramount predictor, with a generally positive impact. The SHAP dependence plot reveals a U-shaped relationship, where initial increases in ECI decrease SHAP values (indicating a negative impact), but beyond a certain threshold, further increases lead to positive SHAP values. This pattern aligns with the Environmental Kuznets Curve (EKC) hypothesis, indicating that higher economic complexity initially leads to environmental degradation but ultimately supports better forest cover. Similarly, the inverted-U shape relationship between GDP per capita and forest cover corroborates the EKC hypothesis. Initially, increases in GDP per capita exacerbate deforestation, but beyond a certain point, further economic growth fosters improved forest cover as countries invest in sustainable practices.

The implications of these findings are profound. The positive impact of ECI, as elucidated by the SHAP values, underscores the significance of economic complexity in promoting sustainable forest management. It is arguable that countries with higher ECI possess the technological and institutional frameworks necessary to implement effective forest conservation strategies. The inverted-U relationship with GDP per capita suggests that economic growth, up to a certain point, can intensify deforestation due to increased demand for land and resources. However, once a country attains a higher economic status, it may allocate more resources toward environmental protection and sustainable practices. This illustrates the critical balance between economic development and environmental sustainability, emphasizing the need for targeted policies that support economic growth while mitigating its environmental impacts.

The implications of these SHAP findings are profound. The positive impact of HDI suggests that higher human development levels, encompassing education, health, and income, are crucial for sustainable forest management. This reinforces the argument that improving socio-economic conditions can lead to better environmental stewardship. The U-shaped impact of ECI further supports the notion that economic complexity can drive both negative and positive environmental outcomes, depending on the level of development. This emphasizes the need for policies that foster economic complexity and human development concurrently, ensuring that advancements in these areas translate into positive environmental impacts.

The implications of these findings are substantial. The high SHAP value for land area underscores its critical role in preserving biodiversity, as larger land areas provide more habitats for species. The positive impact of GDP per capita suggests

that wealthier countries may have more resources to invest in conservation efforts, though the variability in points indicates that this relationship is not straightforward. The exponential growth relationship with HDI highlights the importance of human development in biodiversity conservation, suggesting that improving health, education, and income levels can significantly enhance conservation outcomes. This underscores the argument for integrated policies that address both human development and environmental conservation to achieve sustainable biodiversity outcomes.

The implications here are clear: while R&D expenditure shows some influence, the overall impact on all some SDG indicator is minimal, suggesting that suggesting that SDG indicators are influenced by a complex interplay of factors not fully captured by the current models. This implies a need for further research to identify additional features and to develop more comprehensive models that can account for the multifaceted nature of features like R&D expenditure. As for example R&D expenditure will leads to development and innovation which in turn can ensure further efficient use of resources which is likely to have an effect on SDG indicator but there are other nuances feature that can capture that relation more effectively and further research needs to be done in that area.

11.3 Implications and Recommendations

The findings of this study have significant implications for policymakers and stakeholders aiming to achieve SDGs 14 and 15. The positive impact of economic complexity and human development on environmental sustainability highlights the need for integrated policies that promote both economic and social development while ensuring environmental preservation.

1. **Promoting Economic Complexity:** Policies should focus on enhancing economic complexity by fostering innovation, technological advancement, and industrial diversification. This can be achieved through investments in research and development, education, and infrastructure that support the growth of complex and sustainable industries.

2. **Improving Human Development:** Investments in education, healthcare, and income generation are crucial for improving HDI and achieving better environmental outcomes. Policies should aim to provide equitable access to quality education and healthcare services, which can lead to more environmentally conscious behaviors and sustainable practices.

3. **Balancing Economic Growth and Environmental Sustainability:** The observed U-shaped and inverted-U shape relationships between economic indicators and environmental outcomes underscore the need for balanced ap-

proaches that mitigate the initial negative impacts of economic growth on the environment. This can be achieved through the implementation of stringent environmental regulations, promotion of sustainable practices, and incentives for green technologies.

4. **Leveraging Machine Learning for Policy Insights:** The application of advanced machine learning models, as demonstrated in this study, provides valuable insights into the complex interactions between economic, social, and environmental dimensions. Policymakers can leverage these models to predict environmental outcomes, assess the effectiveness of existing policies, and design more targeted and efficient interventions.

In conclusion, this study contributes to the growing body of literature on sustainable development by providing a comprehensive analysis of the relationships between economic complexity, human development, and environmental sustainability. Our findings underscore the importance of integrating economic and social development with environmental preservation to achieve the SDGs. By adopting a holistic approach that leverages advanced analytical tools such as machine learning, policymakers can better navigate the complexities of sustainable development and design strategies that promote long-term environmental health and socio-economic prosperity. The intricate interplay of economic complexity, human development, and environmental sustainability revealed by our study highlights the critical need for multifaceted and well-coordinated policy interventions.

11.4 Future Work

For our project, we encountered significant challenges in obtaining consistent data for most indicators, particularly concerning country coverage and time periods. Consequently, future work should focus on gathering more uniform and comprehensive datasets. Additionally, we should broaden our feature set by incorporating a wider range of variables. These variables may not have a direct impact on social development goals but could indirectly influence economic and social development.

To enhance the robustness of our analysis, it would be beneficial to develop models specifically tailored to each indicator. This approach would involve tuning the feature sets to align closely with the unique characteristics and requirements of each indicator. Given the number of indicators and their varying functionalities, this strategy will require a more meticulous and prolonged effort.

Moreover, we should explore the use of neural networks trained on larger and more consistent datasets. By adjusting the weights and biases specific to each indicator, we could develop specialized models that can be generalized across all indicators.

This method would allow us to leverage the power of neural networks to capture complex patterns and relationships within the data. A dedicated and systematic strategy in data collection and model development will significantly enhance the effectiveness and reliability of our future analyses.

Adopting a specialized approach for each indicator will also help us better understand the nuanced relationships and interactions within the data. This methodical approach will not only improve the accuracy of our models but also provide deeper insights into the indirect factors affecting economic and social development. Overall, incorporating these advanced techniques and methodologies will further advance our work and contribute to more robust and insightful analyses in the future.

12 References

Abbas, K. R. (2021). Economic complexity, tourism, energy prices, and environmental degradation in the top economic complexity countries. *Journal of Environmental Management*.

Acerta (2022). Understanding machine learning with SHAP analysis. [online] Available at: `https://acerta.ai/blog/understanding-machine-learning-with-shap-analysis/` [Accessed 20 May 2024].

Ai, C. & Norton, E. (2002). Interaction terms in nonlinear models. Social Science Research Network.

Arica, F. & Kurt, U. (2021). The Causal Linkages of Human Development Index and Economic Complexity Index: a Panel Analysis for Selected OECD Countries. *Economics Research International*.

Azur, M.J., Stuart, E.A., Frangakis, C. & Leaf, P.J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, 20(1), pp.40-49.

Bair, E., Hastie, T., Debashis, P. & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473), pp.119-137.

Barbier, E.B. (1987). The concept of sustainable economic development. *Environmental Conservation*, 14(2), pp.101-110. Available at: doi:10.1017/S0376892900011449. [Accessed 13 April 2022].

Boleti, E., Garas, A., Kyriakou, A. & Lapatinas, A. (2019). Economic Complexity and Environmental Performance: Evidence from a World Sample. *Environmental Modeling & Assessment*, 26, pp.251-270. `https://doi.org/10.1007/s10666-021-09750-0`.

Brundtland, G. (1987). Report of the World Commission on Environment and Development: Our Common Future. United Nations General Assembly document A/42/427. Available at: `https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf`.

Carneiro, G. (2021). Marine management for human development: A review of two decades of scholarly evidence. *Marine Policy*.

Carson, R.L. (1962). *Silent Spring*. 1st edn. Houghton Mifflin.

Charfeddine, L. (2021). The impact of economic development and social-political factors on ecological footprint: A panel data analysis for 15 MENA countries. *Ecological Economics*.

Collier, P. (2007). *The Bottom Billion: Why the Poorest Countries are Failing and What Can Be Done About It.* Oxford University Press.

Costantini, V. & Monni, S. (2008). Environment, human development and economic growth. *Ecological Economics*, 64(4), pp.867-880.

Dyck, A.J. (2021). Economic impact of ocean fish populations on the global fishery. *Marine Resource Economics.*

Fine, T.L. (1999). *Feedforward Neural Network Methodology.* Springer Science & Business Media.

Global Footprint Network (2024). Ecological Footprint. Available at: `https://www.footprintnetwork.org/our-work/ecological-footprint/` [Accessed 17 May 2024].

Grossman, G.M. & Krueger, A.B. (1991). Environmental impacts of a North American Free Trade Agreement. *National Bureau of Economic Research*, Working Paper No. 3914.

Hausmann, R. (2016). Economic development and the accumulation of know-how. *Welsh Economic Review*, 24, pp.13-16.

Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M. & Simoes, A. (2014). *The atlas of economic complexity: Mapping paths to prosperity.* MIT Press.

Hausmann, R. & Hidalgo, C.A. (2011). The network structure of economic output. *Journal of Economic Growth*, 16(4), pp.309-342.

Hickel, J. (2020). The sustainable development index: Measuring the ecological efficiency of human development in the anthropocene. *Ecological Economics.*

Hidalgo, C.A. & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26), pp.10570-10575.

Horvath, B. (2018). The recognition of resource use through industrial development from a social perspective. *Studia Mundi - Economica*, 5. doi:10.18531/Studia.Mundi.2018.05.01.68-78.

Katila, P., Colfer, C.J.P., De Jong, W., Galloway, G., Pacheco, P. & Winkel, G. eds. (2019). *Sustainable development goals.* Cambridge University Press.

Le Caous, E. & Huarng, F. (2020). Economic complexity and the mediating effects of income inequality: Reaching sustainable development in developing countries. *Sustainability.*

Lindholm, A., Wahlström, N., Lindsten, F. & Schön, T.B. (2022). *Machine Learning: A First Course for Engineers and Scientists.* Cambridge University Press.

Maitra, S. & Yan, J. (2008). Principal component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, 79, pp.79-90.

Mehrjo, A. & Yuzbashkand, S.S. (2021). Economic complexity, ICT, biomass energy consumption, and environmental degradation: evidence from Iran. *Journal of Cleaner Production*.

Neumayer, E. (2016). Human development and sustainability. *The Capability Approach and Sustainability*, pp.154-172.

Nguyen Van Tran, N., Van Tran, Q., Do, L.T.T., Dinh, L.H. & Do, H.T.T. (2019). Trade off between environment, energy consumption and human development: Do levels of economic development matter?. *Energy*, 173, pp.483-493.

Nti, I., Yarko-Boateng, O. & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*. `https://doi.org/10.5815/ijitcs.2021.06.05`.

Ortiz-Ospina, E. & Roser, M. (2016). Government spending. [online] Available at: `https://ourworldindata.org/government-spending` [Accessed 20 May 2024].

Our World in Data team (2023a). Conserve and sustainably use the oceans, seas and marine resources. [online] Available at: `https://ourworldindata.org/sdgs/life-below-water` [Accessed 20 May 2024].

Our World in Data team (2023b). Sustainably manage forests, combat desertification, halt and reverse land degradation, halt biodiversity loss. [online] Available at: `https://ourworldindata.org/sdgs/life-on-land` [Accessed 20 May 2024].

Prasad, M.N.V. (2024). Bioremediation, bioeconomy, circular economy, and circular bioeconomy—Strategies for sustainability. In *Bioremediation and Bioeconomy* (pp.3-32). Elsevier.

Qiu, Y., Zheng, H. & Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *GigaScience*, 9. `https://doi.org/10.1093/gigascience/giaa082`.

Rafique, M.Z., Nadeem, A.M., Xia, W., Ikram, M., Shoaib, H.M. & Shahzad, U. (2022). Does economic complexity matter for environmental sustainability? Using ecological footprint as an indicator. *Environment, Development and Sustainability*, 24, pp.4623-4640.

Ran, X., Chen, W., Yvert, B. & Zhang, S. (2022). A hybrid autoencoder framework of dimensionality reduction for brain-computer interface decoding. *Computers in Biology and Medicine*, 148, p.105871.

Rao, A. (2021). The role of natural resources in the management of environmental sustainability: Machine learning approach. *Sustainability*.

Ritchie, H. & Roser, M. (2019). Land use. [online] Available at: `https://ourworldindata.org/land-use` [Accessed 20 May 2024].

Ritchie, H. & Roser, M. (2020). CO2 emissions. [online] Available at: https://ourworldindata.org/co2-emissions [Accessed 20 May 2024].

Sachs, J.D. (2012). From millennium development goals to sustainable development goals. *The Lancet*, 379(9832), pp.2206-2211.

Sasse, L.M. (2022). Sustainable development in Latin America–linking renewable energy, economic complexity and GHG. [PDF document]. June. Published by von lu.se.

Sheather, S. (2009). *A modern approach to regression with R*. Springer Science & Business Media.

Stanton, E.A. (2007). *The Human Development Index: A History*. Global Development and Environment Institute, 127.

Stern, D.I. & Kander, A. (2012). The role of energy in the Industrial Revolution and modern economic growth. *The Energy Journal*, 33(3), pp.125-152.

Sturesson, A., Weitz, N. & Persson, Å. (2018). SDG 14: life below water. A Review of Research Needs. Technical annex to the Formas report *Forskning för Agenda 2030*.

Suits, D.B. et al. (1978). Spline functions fitted by standard regression methods. *The Review of Economics and Statistics*, 60(1), pp.132-139.

Swart, J. & Brinkmann, L. (2020). Economic complexity and the environment: Evidence from Brazil. In: *Economic Complexity and the Environment: Evidence from Brazil*, pp.3-45.

The Global Goals (2024). Resources. Available at: https://www.globalgoals.org/resources/ [Accessed 14 May 2024].

Ul Haq, M. (1995). *Reflections on human development*. Oxford University Press.

UNDP (United Nations Development Programme). (1990). *Human Development Report 1990: Concept and Measurement of Human Development*. New York.

UNDP (United Nations Development Programme). (2010). *Human Development Report 2010: The Real Wealth of Nations: Pathways to Human Development*. New York.

United Nations Development Programme (2023). Human Development Index. [online] Available at: https://hdr.undp.org/en/content/human-development-index-hdi [Accessed 20 May 2024].

Van Tran, N., Van Tran, Q., Do, L.T.T., Dinh, L.H. & Do, H.T.T. (2019). Trade off between environment, energy consumption and human development: Do levels of economic development matter?. *Energy*, 173, pp.483-493.

Yoon, J., Jordon, J. & Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. *ArXiv*, abs/1806.02920.

Zulham, T. (2021). The Nexus of Human Development Index, Economic and Population Growth on Environmental Degradation in Aceh Province, Indonesia. *Environmental Science and Pollution Research*.

13 Appendix A: Final Result Tables

13.1 SDG 15 Modelling Final Results

Table 9: Final Compiled Results for All Models and Datasets (SDG 15)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.039	0.112	0.041	0.058	Linear Regression	MICE	Forest_cover	NaN
0.012	0.132	0.012	0.148	Linear Regression	MICE	Forest_biomass	NaN
0.058	0.170	0.059	0.154	Linear Regression	MICE	Mountain_KBAs	NaN
0.036	0.179	0.036	0.160	Linear Regression	Interaction	Forest_cover	NaN
0.011	0.184	0.011	0.235	Linear Regression	Interaction	Forest_biomass	NaN
0.056	0.199	0.057	0.173	Linear Regression	Interaction	Mountain_KBAs	NaN
0.036	0.162	0.035	0.178	Linear Regression	Hybrid	Forest_cover	NaN
0.012	0.157	0.011	0.224	Linear Regression	Hybrid	Forest_biomass	NaN
0.056	0.191	0.058	0.160	Linear Regression	Hybrid	Mountain_KBAs	NaN
0.030	0.309	0.030	0.311	Spline Regression	MICE	Forest_cover	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.010	0.272	0.009	0.312	Spline Regression	MICE	Forest_biomass	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5

Continued on next page

Table 9: Final Compiled Results for All Models and Datasets (SDG 15)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.049	0.303	0.048	0.303	Spline Regression	MICE	Mountain_KBAs	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.029	0.327	0.029	0.335	Spline Regression	Interaction	Forest_cover	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.010	0.301	0.009	0.339	Spline Regression	Interaction	Forest_biomass	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.047	0.328	0.048	0.315	Spline Regression	Interaction	Mountain_KBAs	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.024	0.444	0.026	0.392	Spline Regression	Hybrid	Forest_cover	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.008	0.385	0.009	0.380	Spline Regression	Hybrid	Forest_biomass	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5

Continued on next page

Table 9: Final Compiled Results for All Models and Datasets (SDG 15)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.038	0.450	0.043	0.387	Spline Regression	Hybrid	Mountain_KBAs	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.001	0.984	0.005	0.891	Random Forest	MICE	Forest_cover	'max_depth': 30, 'n_estimators': 100
0.000	0.987	0.002	0.889	Random Forest	MICE	Forest_biomass	'max_depth': 30, 'n_estimators': 100
0.001	0.986	0.007	0.905	Random Forest	MICE	Mountain_KBAs	'max_depth': 30, 'n_estimators': 100
0.001	0.982	0.006	0.867	Random Forest	Interaction	Forest_cover	'max_depth': 30, 'n_estimators': 100
0.000	0.986	0.002	0.880	Random Forest	Interaction	Forest_biomass	'max_depth': 30, 'n_estimators': 100
0.001	0.984	0.007	0.902	Random Forest	Interaction	Mountain_KBAs	'max_depth': 30, 'n_estimators': 100
0.002	0.953	0.017	0.613	Random Forest	Hybrid	Forest_cover	'max_depth': 40, 'n_estimators': 1000
0.001	0.953	0.005	0.627	Random Forest	Hybrid	Forest_biomass	'max_depth': 40, 'n_estimators': 1000

Continued on next page

Table 9: Final Compiled Results for All Models and Datasets (SDG 15)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.003	0.955	0.022	0.685	Random Forest	Hybrid	Mountain_KBAs	'max_depth': 40, 'n_estimators': 1000
0.020	0.549	0.021	0.516	FNN	MICE	Forest_cover	'neurons': 64, 'learning_rate': 0.01
0.008	0.412	0.008	0.434	FNN	MICE	Forest_biomass	'neurons': 64, 'learning_rate': 0.01
0.029	0.579	0.037	0.462	FNN	MICE	Mountain_KBAs	'neurons': 64, 'learning_rate': 0.01
0.021	0.514	0.024	0.439	FNN	Interaction	Forest_cover	'neurons': 128, 'learning_rate': 0.001
0.008	0.439	0.009	0.376	FNN	Interaction	Forest_biomass	'neurons': 128, 'learning_rate': 0.001
0.035	0.503	0.041	0.406	FNN	Interaction	Mountain_KBAs	'neurons': 128, 'learning_rate': 0.001
0.007	0.833	0.014	0.669	FNN	Hybrid	Forest_cover	'neurons': 128, 'learning_rate': 0.01
0.003	0.766	0.006	0.591	FNN	Hybrid	Forest_biomass	'neurons': 128, 'learning_rate': 0.01

Continued on next page

Table 9: Final Compiled Results for All Models and Datasets (SDG 15)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.013	0.814	0.023	0.674	FNN	Hybrid	Mountain_KBAs	'neurons': 128, 'learning_rate': 0.01

14 Appendix B: Final Result Tables

14.1 SDG 14 Modelling Final Results

Table 10: Final Compiled Results for All Models and Datasets (SDG 14)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.003	0.029	0.005	0.012	Linear Regression	MICE	Beach_Litter	NaN
0.003	0.127	0.006	0.053	Linear Regression	MICE	Chlorophyll	NaN
0.002	0.595	0.001	0.793	Linear Regression	MICE	Sustainable_fisheries	NaN
0.003	0.031	0.005	0.012	Linear Regression	Interaction	Beach_Litter	NaN
0.003	0.161	0.006	0.068	Linear Regression	Interaction	Chlorophyll	NaN
0.002	0.607	0.001	0.807	Linear Regression	Interaction	Sustainable_fisheries	NaN
0.003	0.053	0.005	0.028	Linear Regression	Hybrid	Beach_Litter	NaN
0.003	0.181	0.006	0.084	Linear Regression	Hybrid	Chlorophyll	NaN
0.002	0.614	0.001	0.814	Linear Regression	Hybrid	Sustainable_fisheries	NaN
0.003	0.080	0.005	0.054	Spline Regression	MICE	Beach_Litter	'ridge__alpha': 0.1, 'splinetransformer__degree': 2, 'splinetransformer__n_knots': 5
0.003	0.289	0.005	0.145	Spline Regression	MICE	Chlorophyll	'ridge__alpha': 0.1, 'splinetransformer__degree': 2, 'splinetransformer__n_knots': 5

Continued on next page

Table 10: Final Compiled Results for All Models and Datasets (SDG 14)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.002	0.678	0.001	0.831	Spline Regression	MICE	Sustainable_fisheries	'ridge__alpha': 0.1, 'splinetransformer__degree': 2, 'splinetransformer__n_knots': 5
0.003	0.091	0.005	0.064	Spline Regression	Interaction	Beach_Litter	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.003	0.295	0.005	0.184	Spline Regression	Interaction	Chlorophyll	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.002	0.693	0.001	0.861	Spline Regression	Interaction	Sustainable_fisheries	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.003	0.110	0.005	0.050	Spline Regression	Hybrid	Beach_Litter	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.002	0.373	0.005	0.202	Spline Regression	Hybrid	Chlorophyll	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5

Continued on next page

Table 10: Final Compiled Results for All Models and Datasets (SDG 14)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.002	0.720	0.001	0.856	Spline Regression	Hybrid	Sustainable_fisheries	'ridge__alpha': 0.1, 'splinetransformer__degree': 3, 'splinetransformer__n_knots': 5
0.000	0.891	0.002	0.528	Random Forest	MICE	Beach_Litter	'max_depth': 30, 'n_estimators': 200
0.000	0.933	0.004	0.396	Random Forest	MICE	Chlorophyll	'max_depth': 30, 'n_estimators': 200
0.000	0.948	0.001	0.796	Random Forest	MICE	Sustainable_fisheries	'max_depth': 30, 'n_estimators': 200
0.000	0.893	0.002	0.562	Random Forest	Interaction	Beach_Litter	'max_depth': 20, 'n_estimators': 200
0.000	0.934	0.004	0.403	Random Forest	Interaction	Chlorophyll	'max_depth': 20, 'n_estimators': 200
0.000	0.947	0.001	0.791	Random Forest	Interaction	Sustainable_fisheries	'max_depth': 20, 'n_estimators': 200
0.000	0.868	0.005	-0.012	Random Forest	Hybrid	Beach_Litter	'max_depth': 20, 'n_estimators': 1000
0.000	0.916	0.005	0.240	Random Forest	Hybrid	Chlorophyll	'max_depth': 20, 'n_estimators': 1000

Continued on next page

Table 10: Final Compiled Results for All Models and Datasets (SDG 14)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.000	0.949	0.001	0.814	Random Forest	Hybrid	Sustainable_fisheries	'max_depth': 20, 'n_estimators': 1000
0.003	0.234	0.004	0.148	FNN	MICE	Beach_Litter	'neurons': 128, 'learning_rate': 0.01
0.003	0.289	0.005	0.181	FNN	MICE	Chlorophyll	'neurons': 128, 'learning_rate': 0.01
0.002	0.681	0.001	0.803	FNN	MICE	Sustainable_fisheries	'neurons': 128, 'learning_rate': 0.01
0.003	0.097	0.005	0.057	FNN	Interaction	Beach_Litter	'neurons': 64, 'learning_rate': 0.001
0.002	0.389	0.005	0.198	FNN	Interaction	Chlorophyll	'neurons': 64, 'learning_rate': 0.001
0.002	0.717	0.001	0.838	FNN	Interaction	Sustainable_fisheries	'neurons': 64, 'learning_rate': 0.001
0.003	0.227	0.005	0.007	FNN	Hybrid	Beach_Litter	'neurons': 128, 'learning_rate': 0.001
0.001	0.594	0.005	0.227	FNN	Hybrid	Chlorophyll	'neurons': 128, 'learning_rate': 0.001

Continued on next page

Table 10: Final Compiled Results for All Models and Datasets (SDG 14)

MSE Train	R^2 Train	MSE Test	R^2 Test	Model	Dataset	Target	Best Parameters
0.001	0.773	0.001	0.804	FNN	Hybrid	Sustainable_fisheries	'neurons': 128, 'learning_rate': 0.001

15 Appendix C: SHAP Analysis Results

15.1 SDG 15: Forest Cover

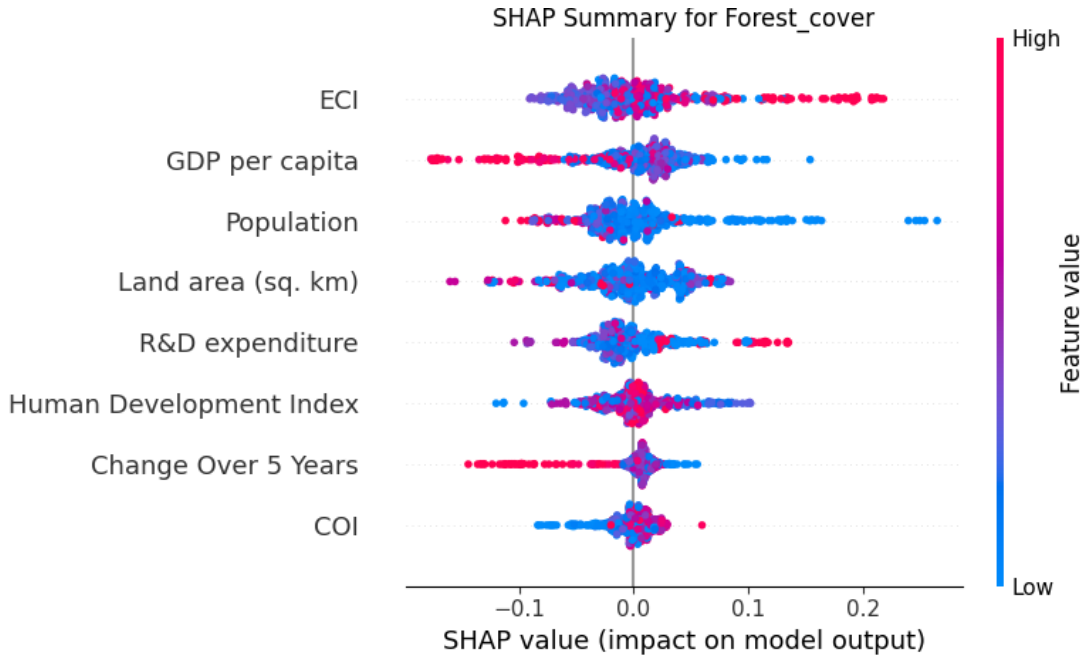

Figure 31: SHAP Summary Plot for Forest Cover

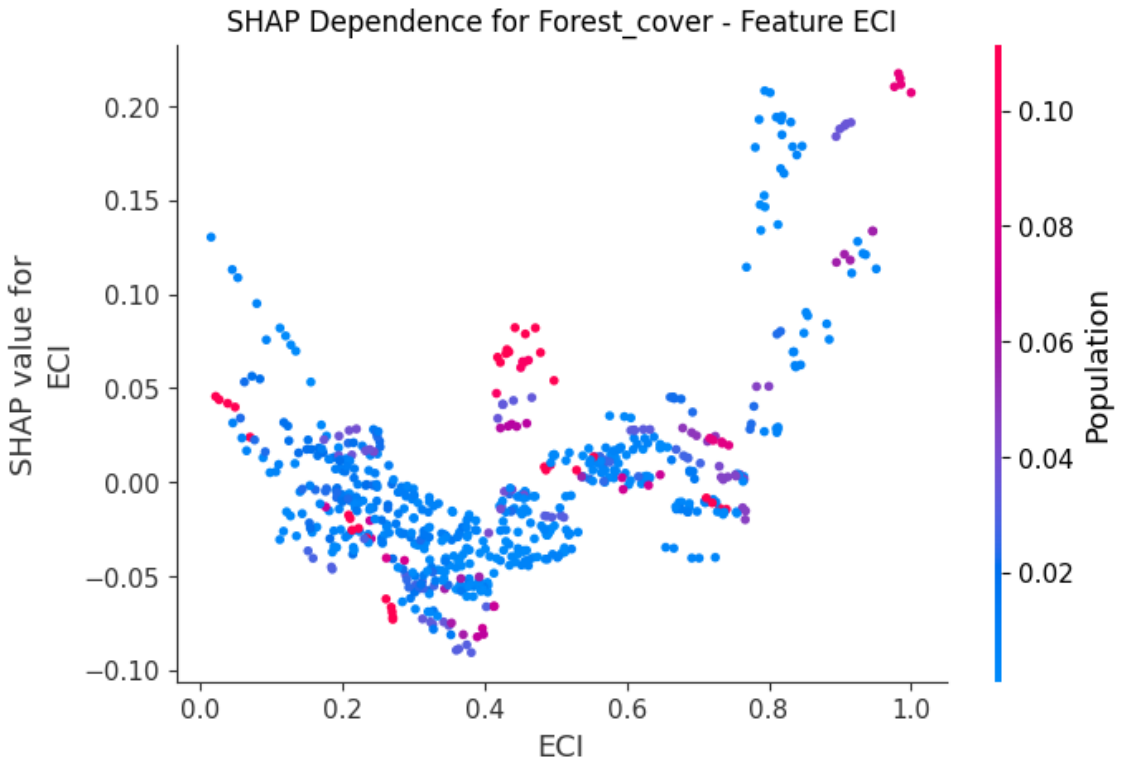

Figure 32: SHAP Dependence Plot for Forest Cover - ECI

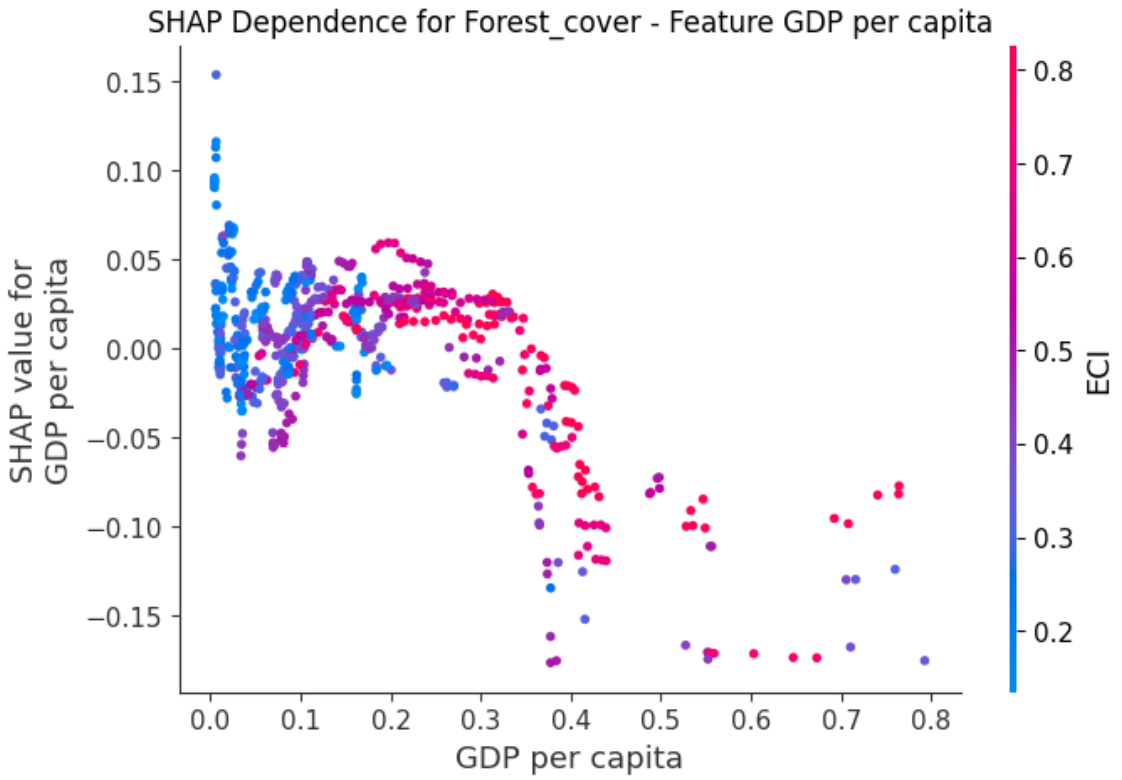

Figure 33: SHAP Dependence Plot for Forest Cover - GDP per capita

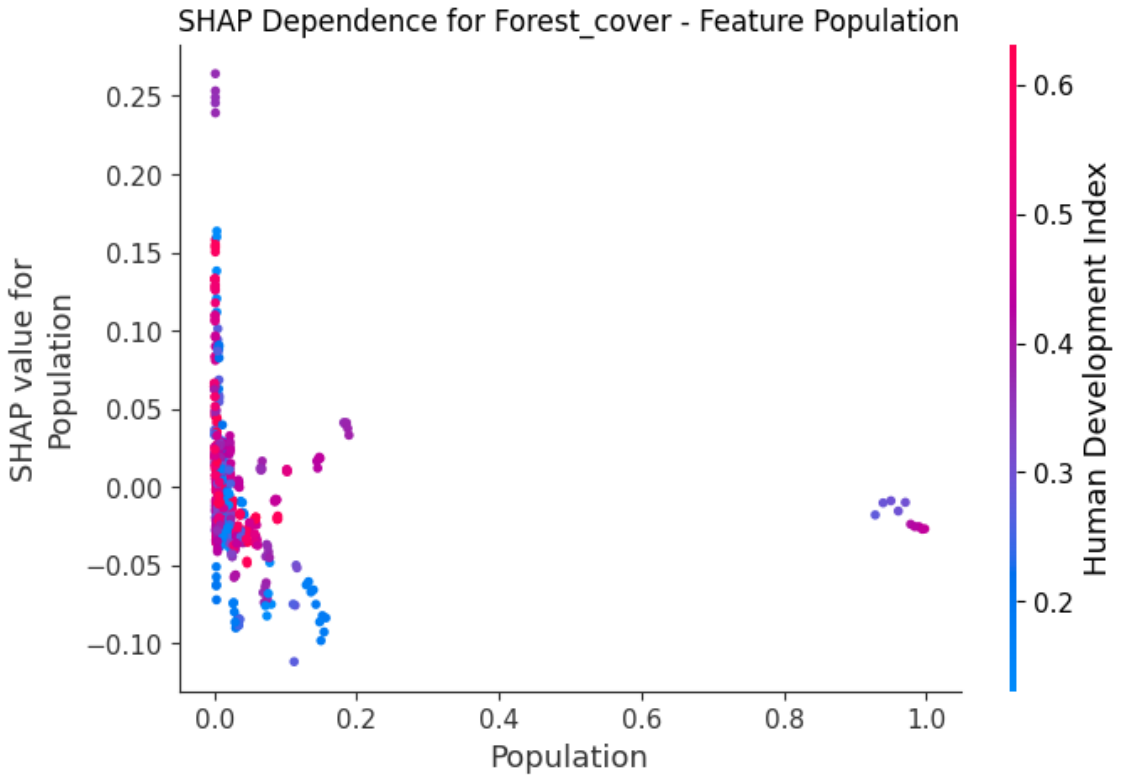

Figure 34: SHAP Dependence Plot for Forest Cover - Population

15.2 SDG 15: Forest Biomass

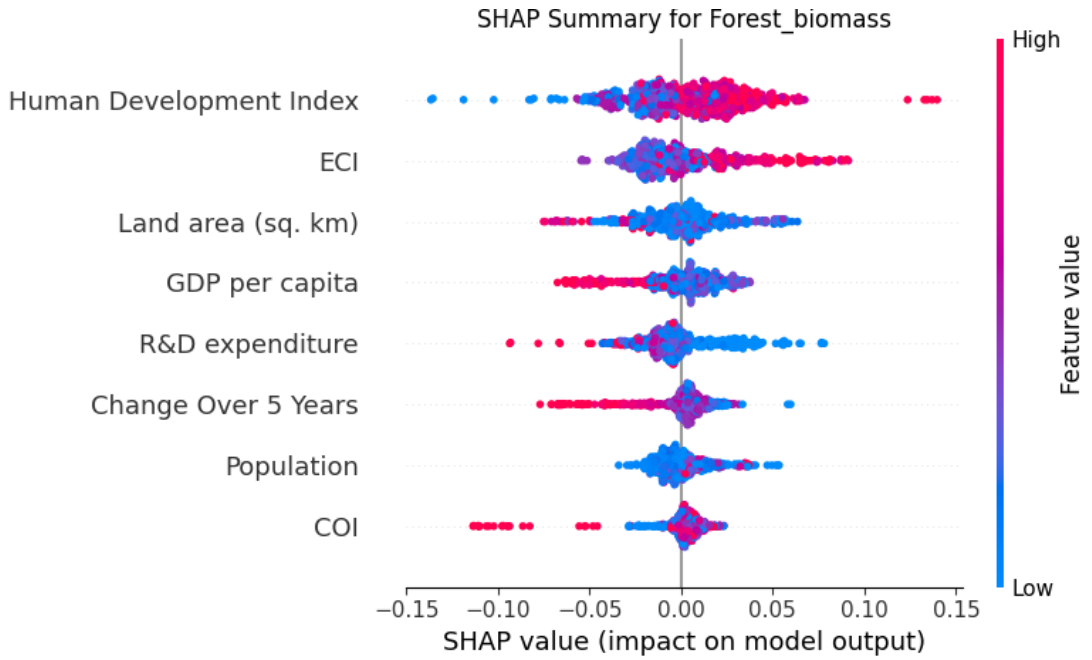

Figure 35: SHAP Summary Plot for Forest Biomass

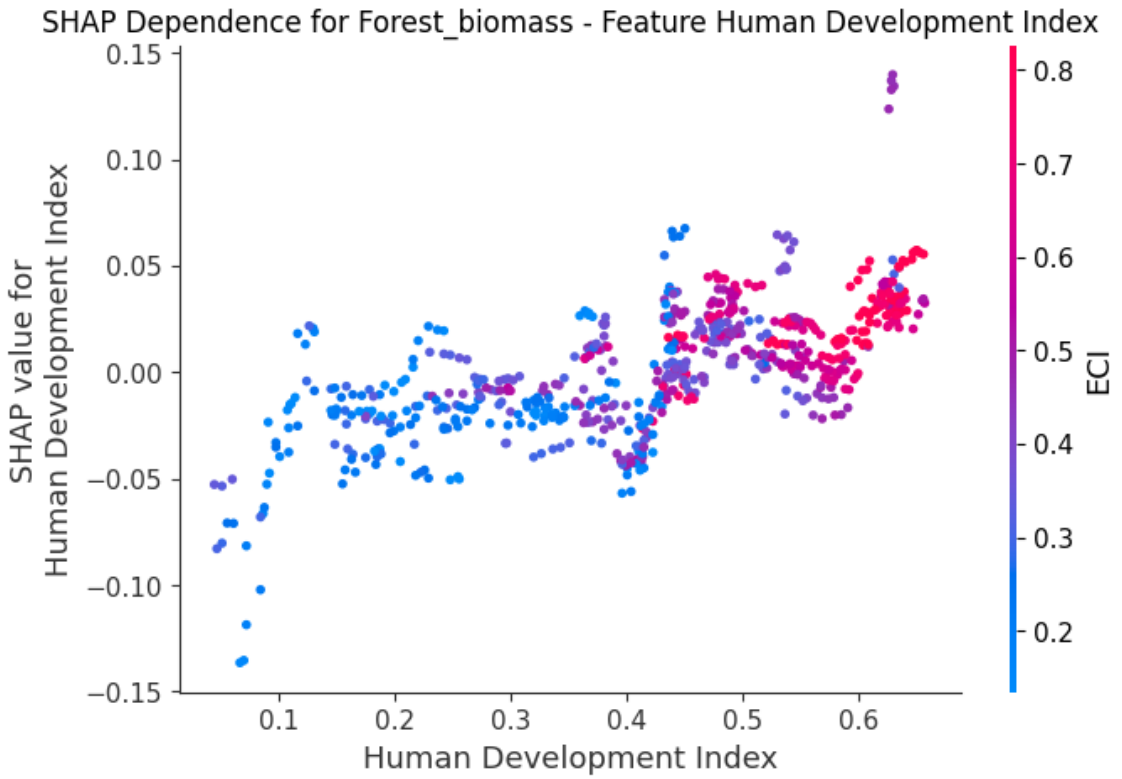

Figure 36: SHAP Dependence Plot for Forest Biomass - Human Development Index

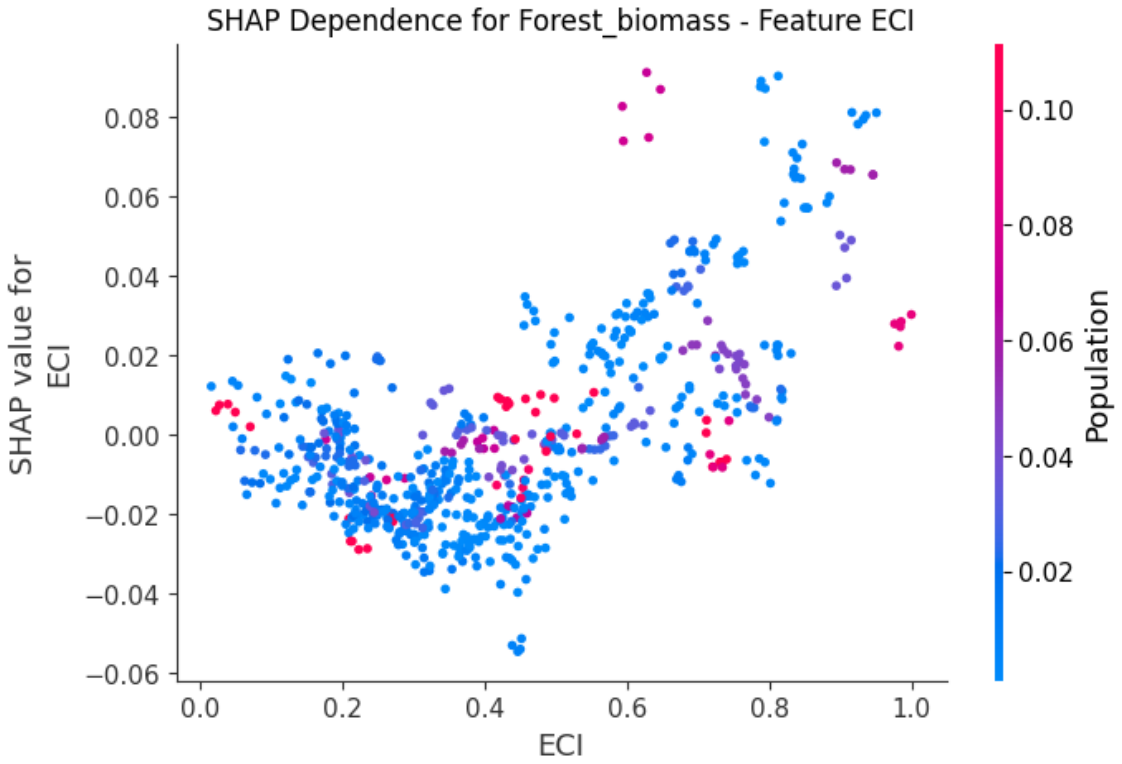

Figure 37: SHAP Dependence Plot for Forest Biomass - ECI

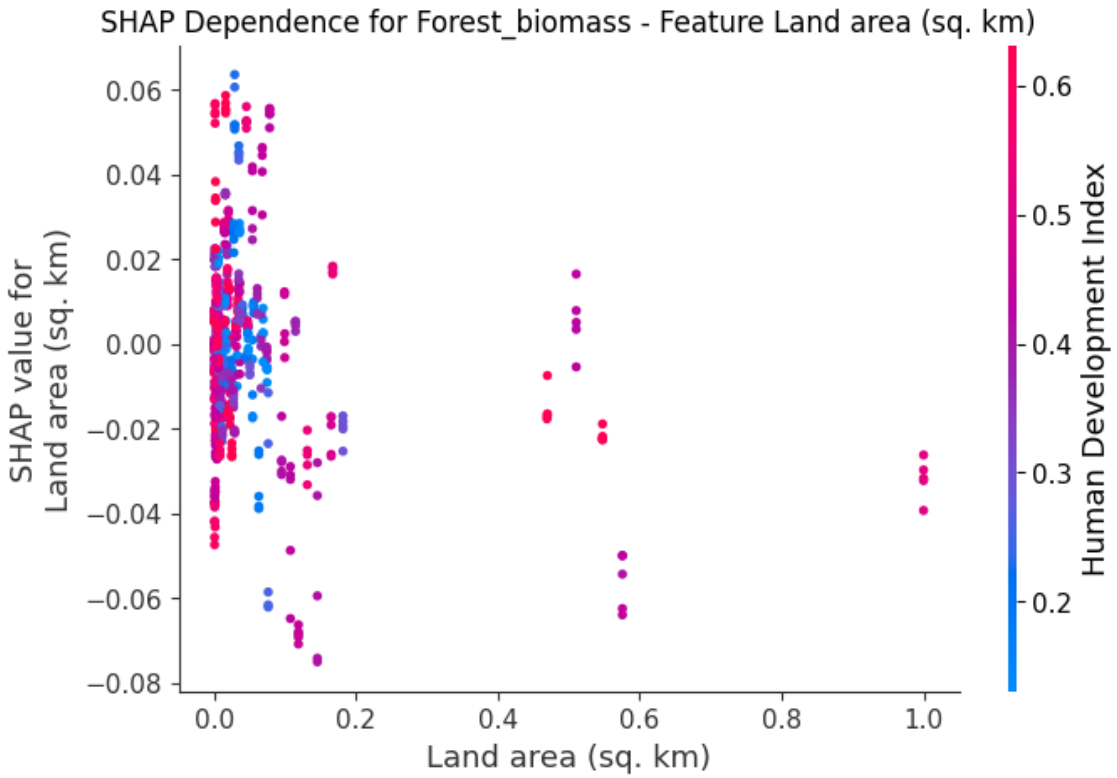

Figure 38: SHAP Dependence Plot for Forest Biomass - Land area (sq. km)

15.3 SDG 15: Mountain KBAs

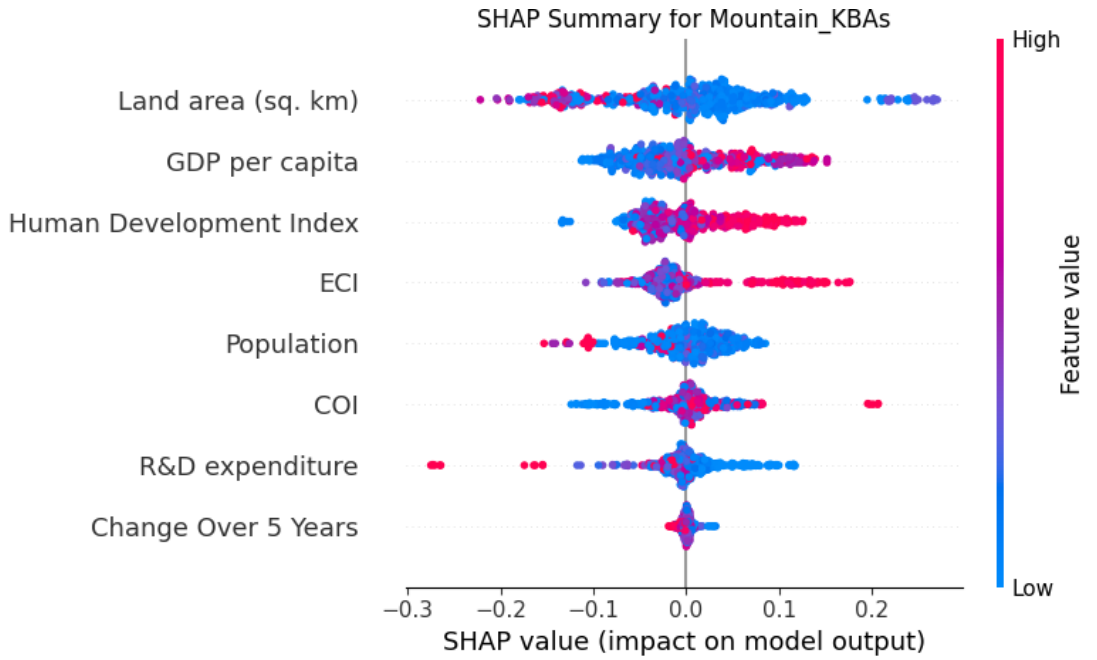

Figure 39: SHAP Summary Plot for Mountain KBAs

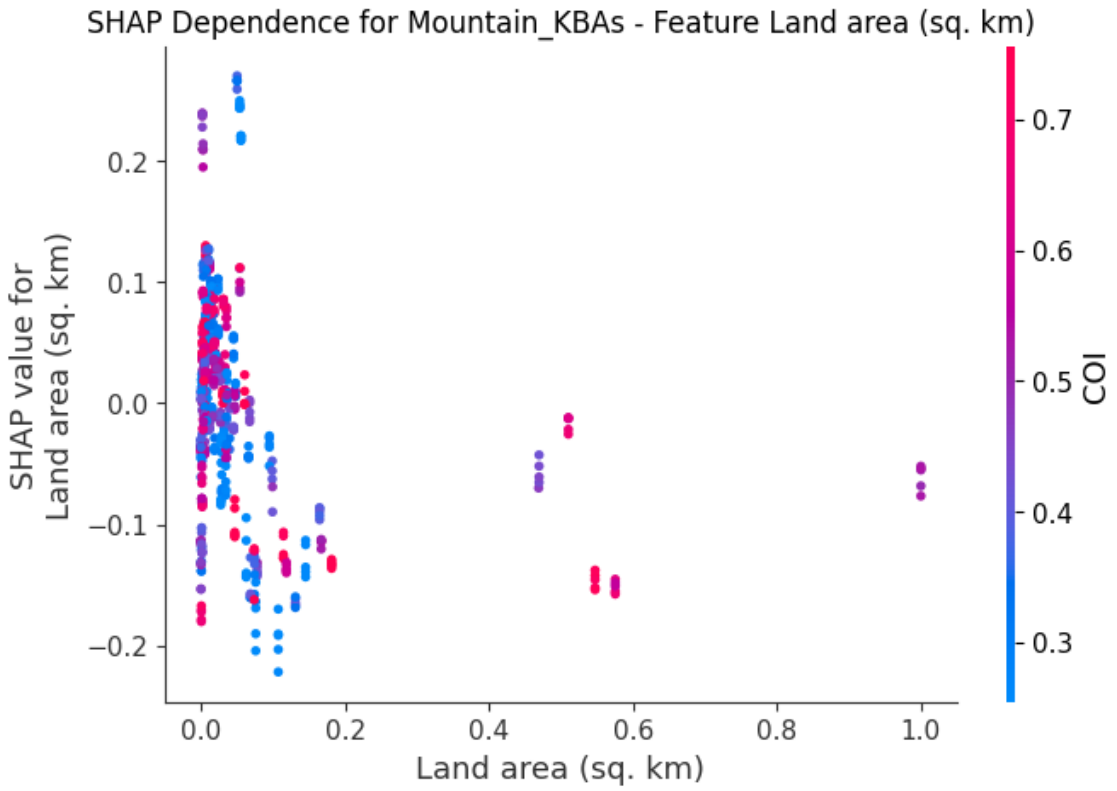

Figure 40: SHAP Dependence Plot for Mountain KBAs - Land area (sq. km)

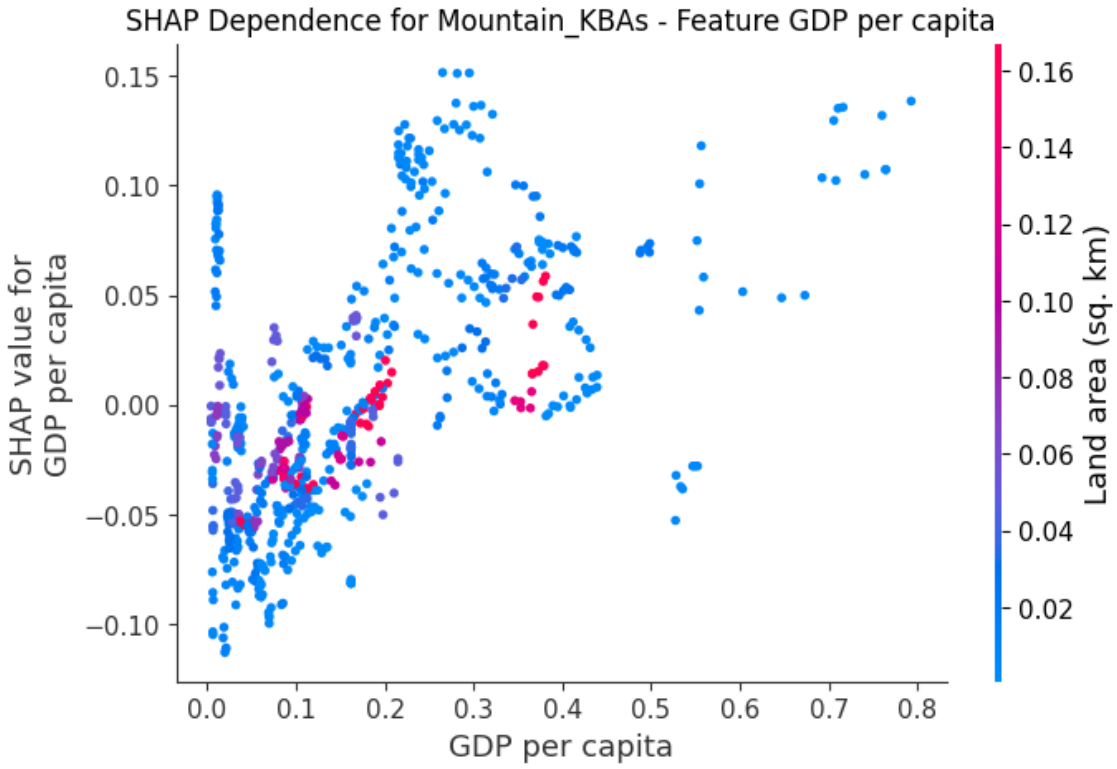

Figure 41: SHAP Dependence Plot for Mountain KBAs - GDP per capita

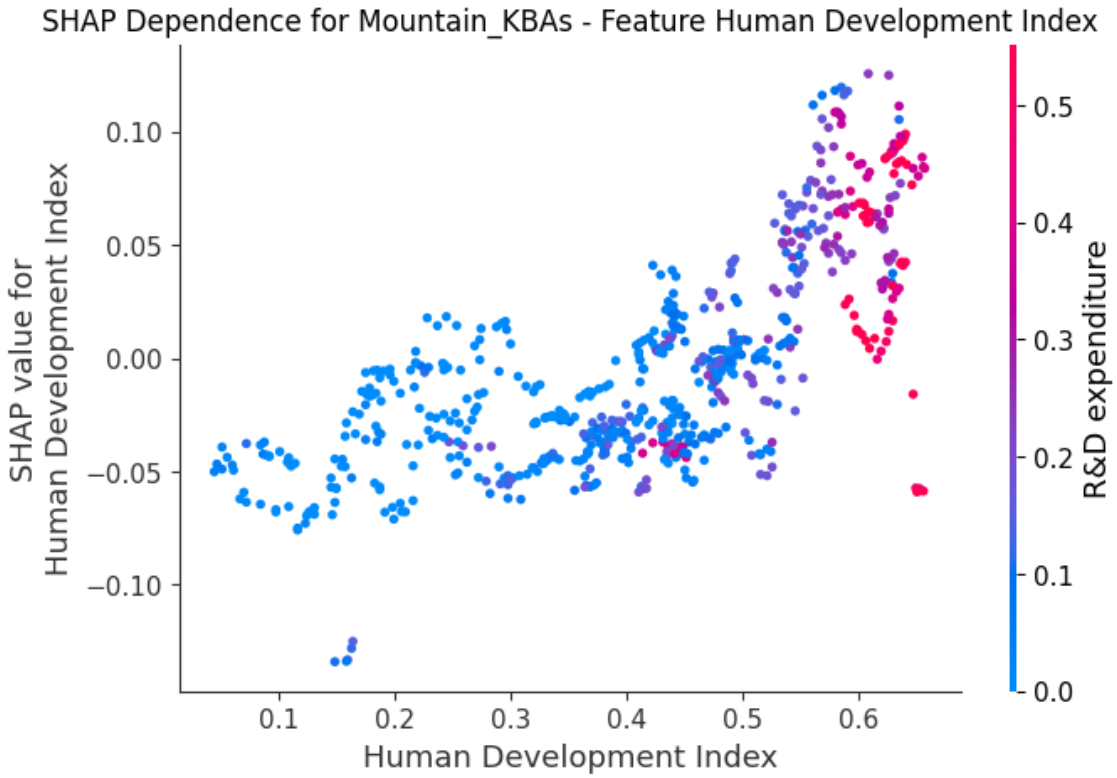

Figure 42: SHAP Dependence Plot for Mountain KBAs - human Development Index

15.4 SDG 15: SHAP Summary Table

Table 11: Combined SHAP Summary Table for All Targets

Feature	Mean Abs SHAP Value	Ranking	Target
ECI	0.037	1	Forest_cover
GDP per capita	0.033	2	Forest_cover
Population	0.030	3	Forest_cover
Land area (sq. km)	0.030	4	Forest_cover
R&D expenditure	0.023	5	Forest_cover
Human Development Index	0.020	6	Forest_cover
Change Over 5 Years	0.015	7	Forest_cover
COI	0.012	8	Forest_cover
Human Development Index	0.024	1	Forest_biomass
ECI	0.019	2	Forest_biomass
Land area (sq. km)	0.016	3	Forest_biomass
GDP per capita	0.016	4	Forest_biomass
R&D expenditure	0.016	5	Forest_biomass
Change Over 5 Years	0.011	6	Forest_biomass
Population	0.010	7	Forest_biomass
COI	0.008	8	Forest_biomass
Land area (sq. km)	0.066	1	Mountain_KBAs
GDP per capita	0.046	2	Mountain_KBAs
Human Development Index	0.038	3	Mountain_KBAs
ECI	0.035	4	Mountain_KBAs
Population	0.029	5	Mountain_KBAs
COI	0.024	6	Mountain_KBAs
R&D expenditure	0.023	7	Mountain_KBAs
Change Over 5 Years	0.004	8	Mountain_KBAs

15.5 SDG 14: Beach Litter

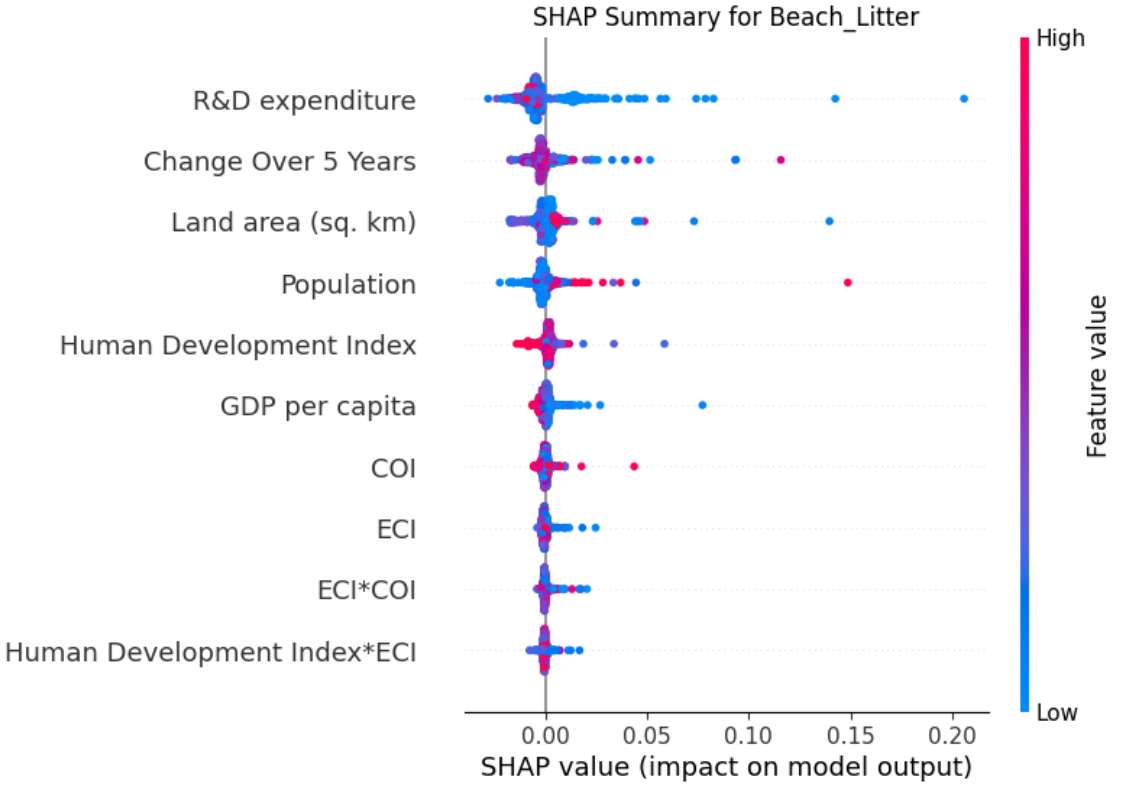

Figure 43: SHAP Summary Plot for Beach Litter

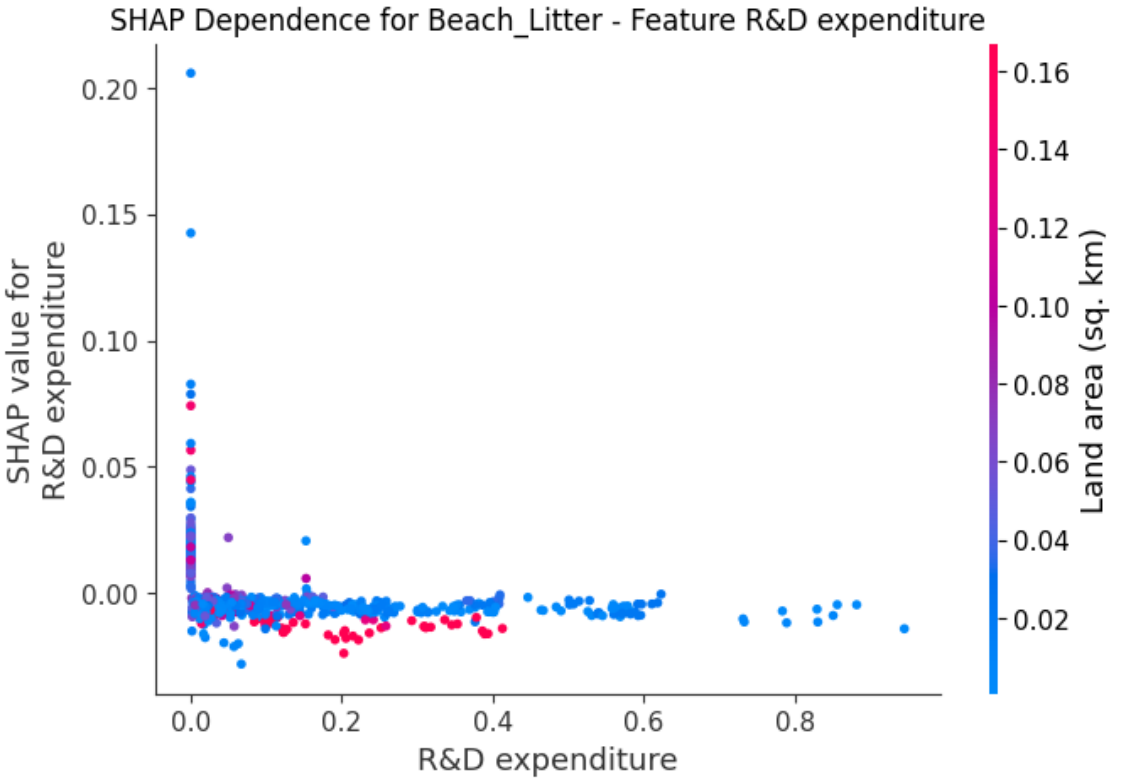

Figure 44: SHAP Dependence Plot for Beach Litter - RD expenditure

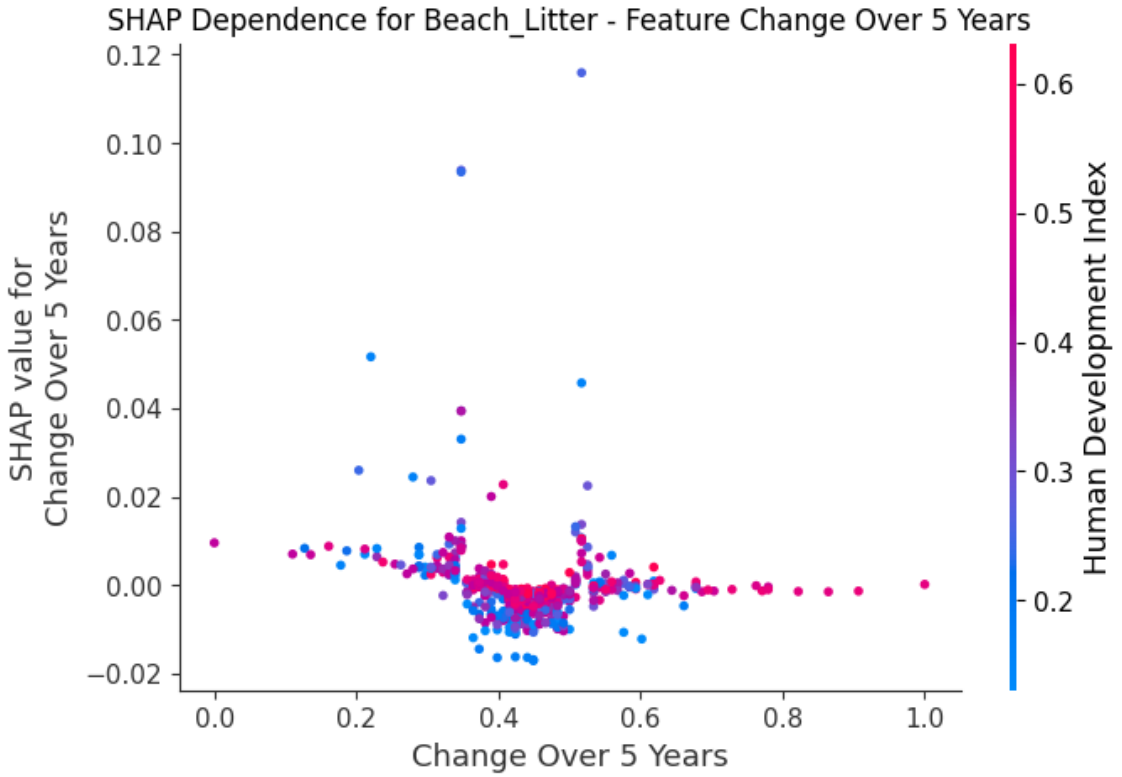

Figure 45: SHAP Dependence Plot for Beach Litter - Change Over 5 Years

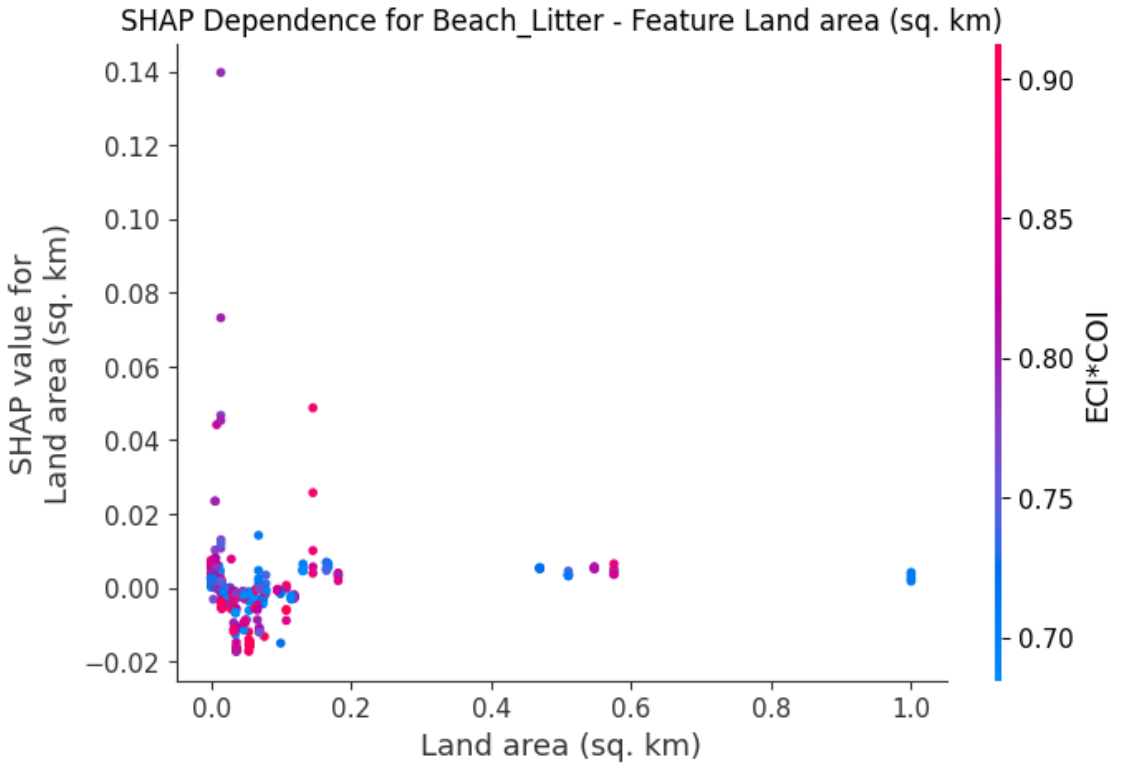

Figure 46: SHAP Dependence Plot for Beach Litter - Land area (sq. km)

15.6 SDG 14: Chlorophyll

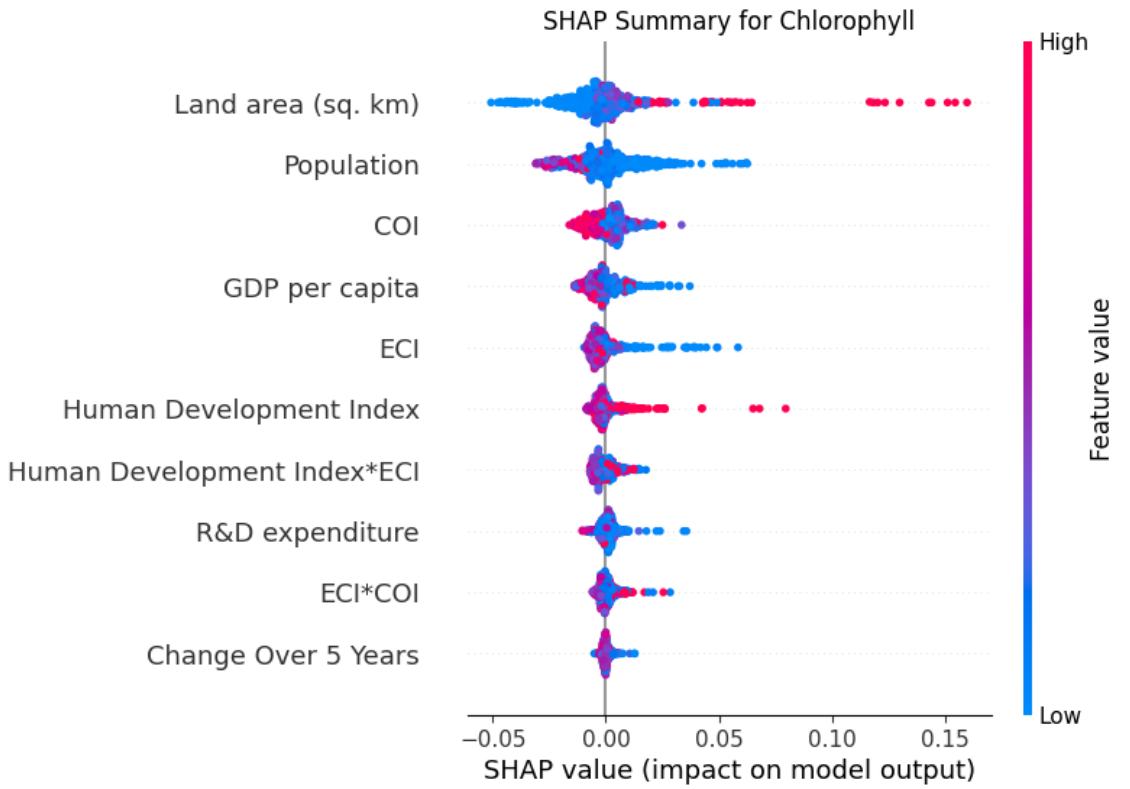

Figure 47: SHAP Summary Plot for Chlorophyll

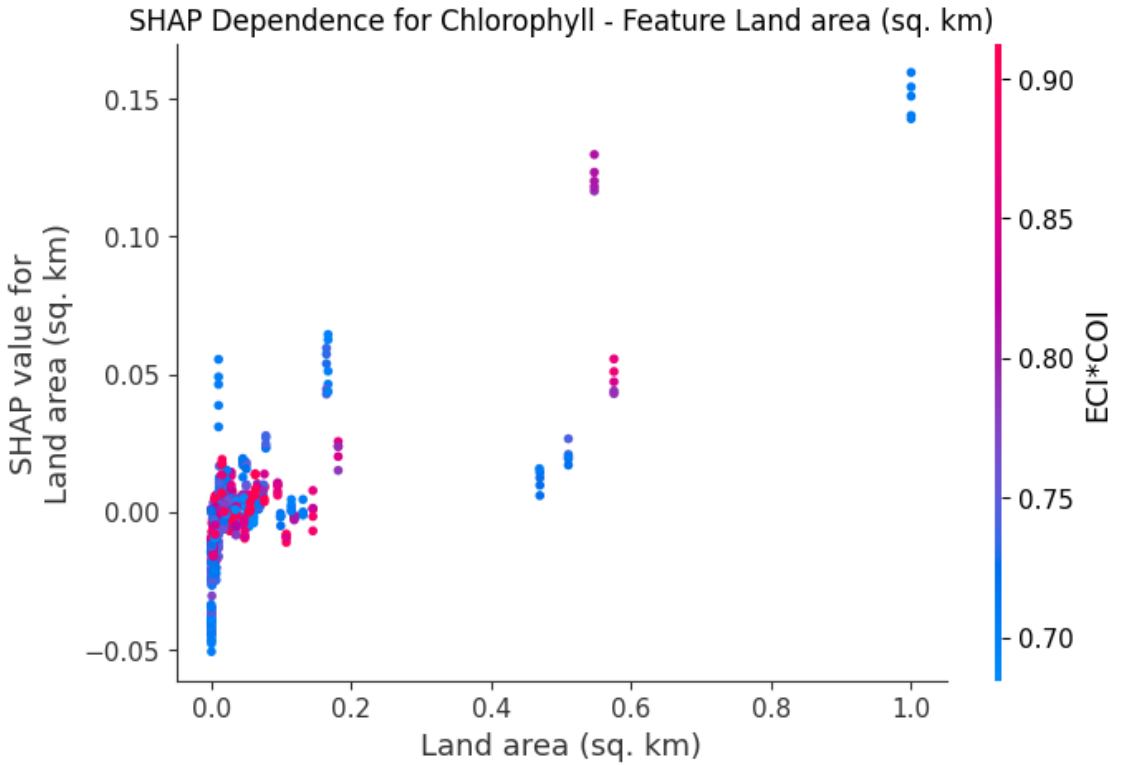

Figure 48: SHAP Dependence Plot for Chlorophyll - Land area (sq. km)

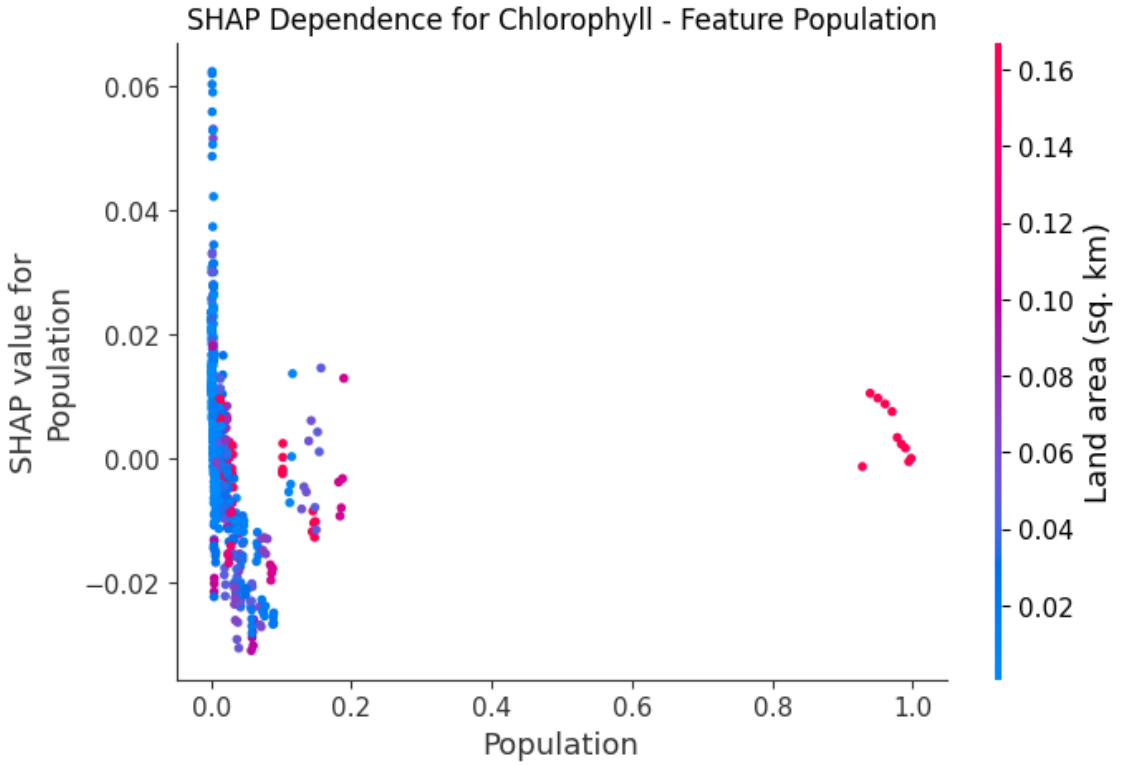

Figure 49: SHAP Dependence Plot for Chlorophyll - Population

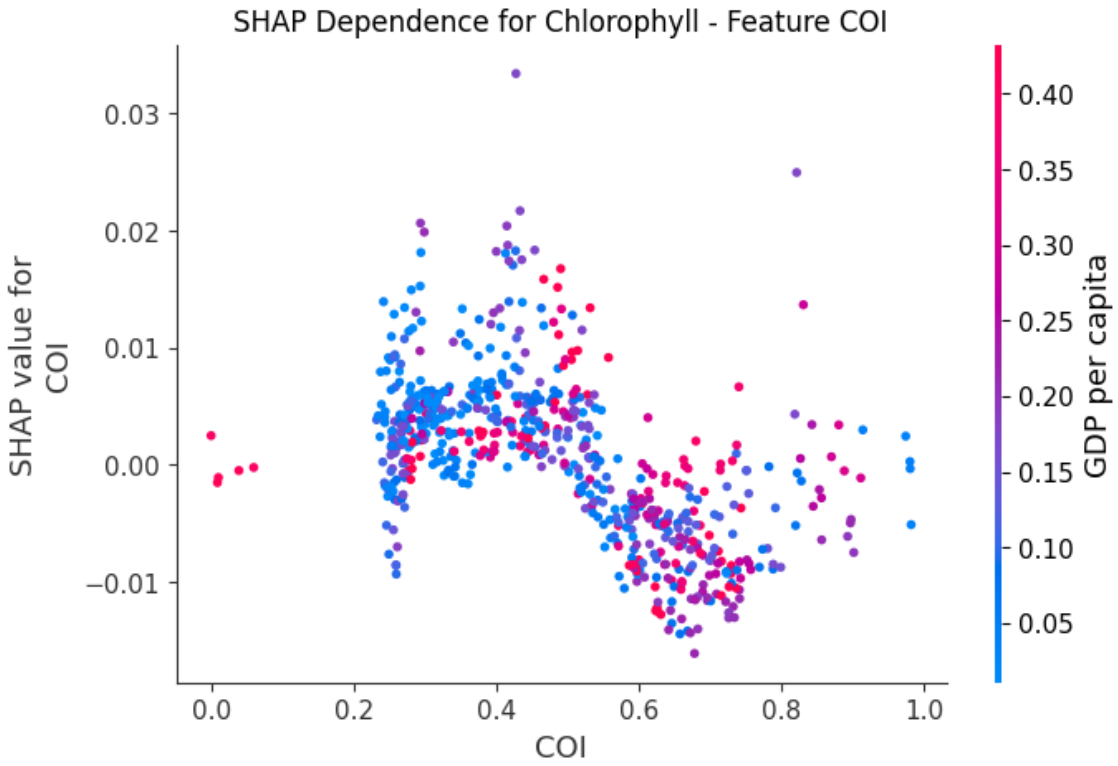

Figure 50: SHAP Dependence Plot for Chlorophyll - COI

15.7 SDG 14: Sustainable Fisheries

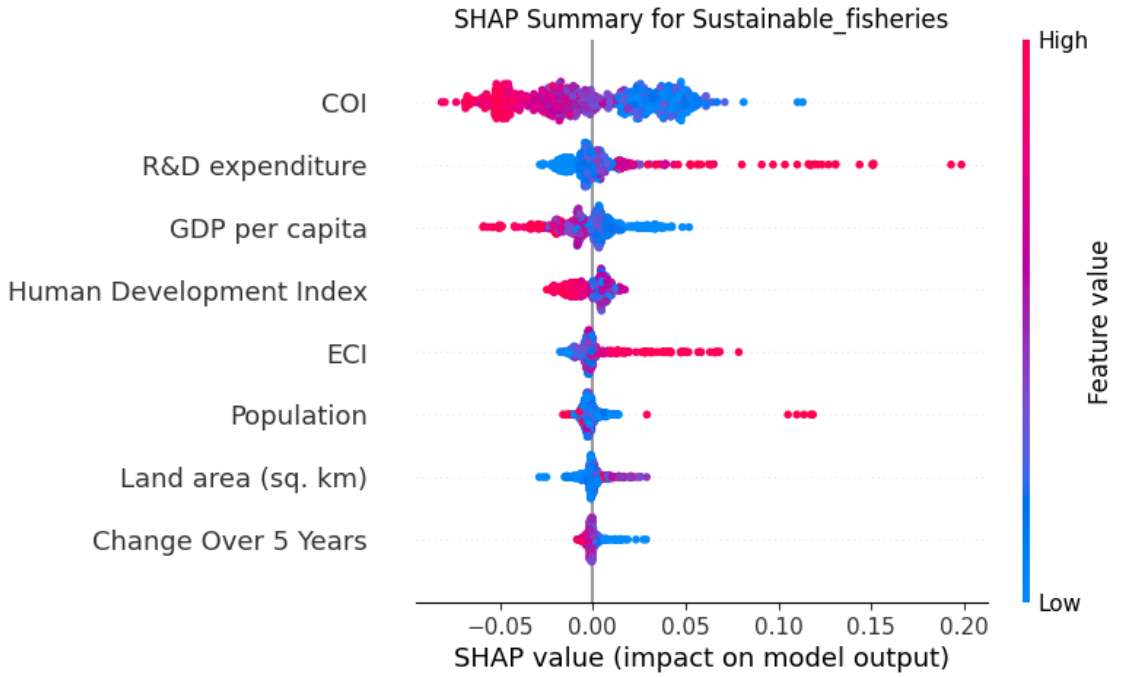

Figure 51: SHAP Summary Plot for Sustainable Fisheries

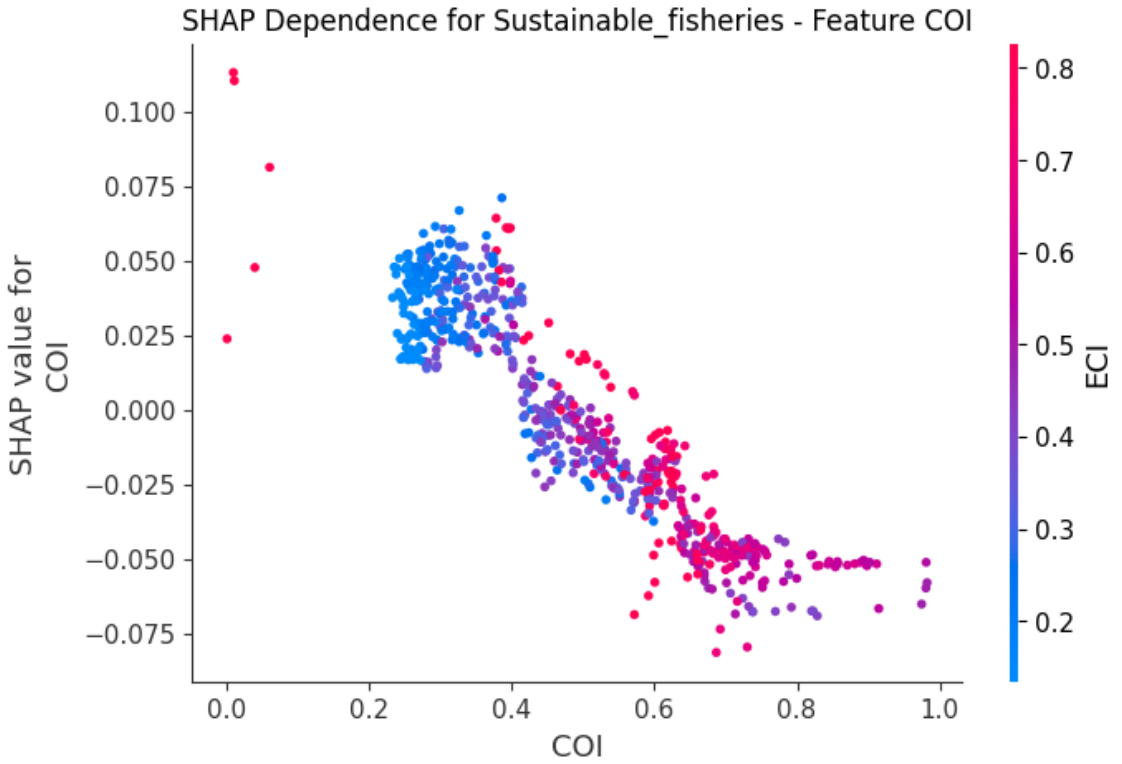

Figure 52: SHAP Dependence Plot for Sustainable Fisheries - COI

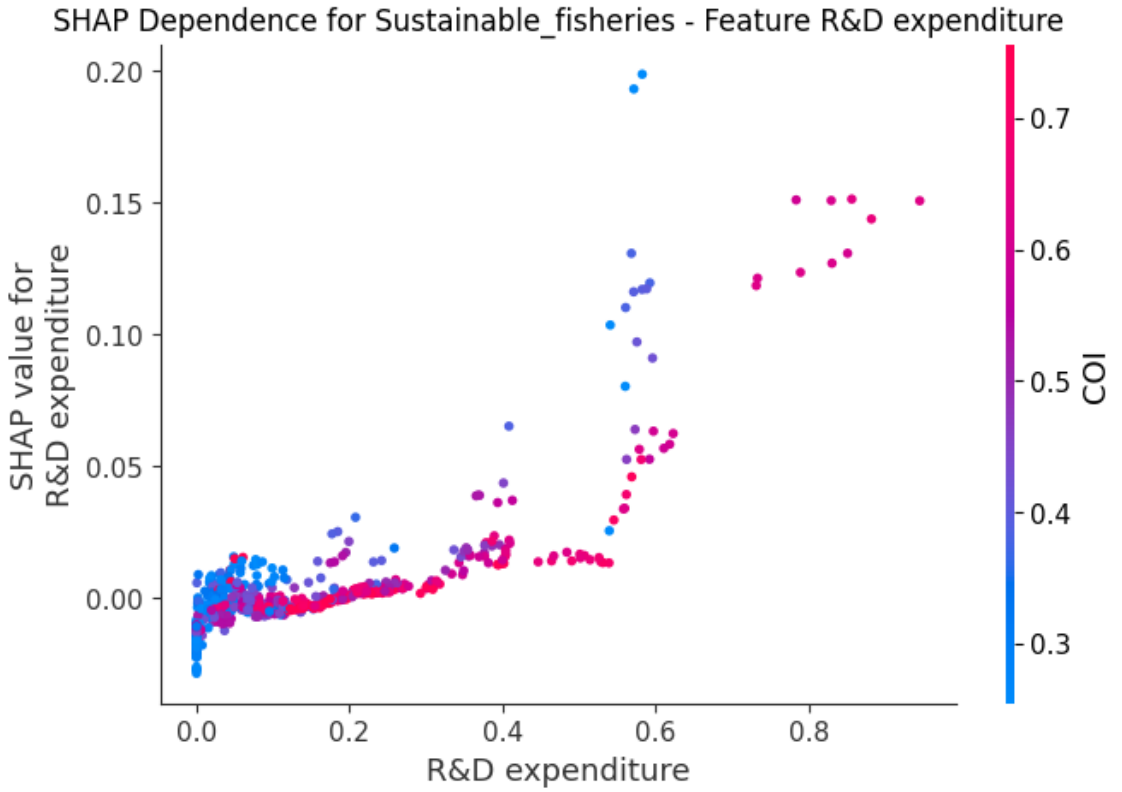

Figure 53: SHAP Dependence Plot for Sustainable Fisheries - RD expenditure

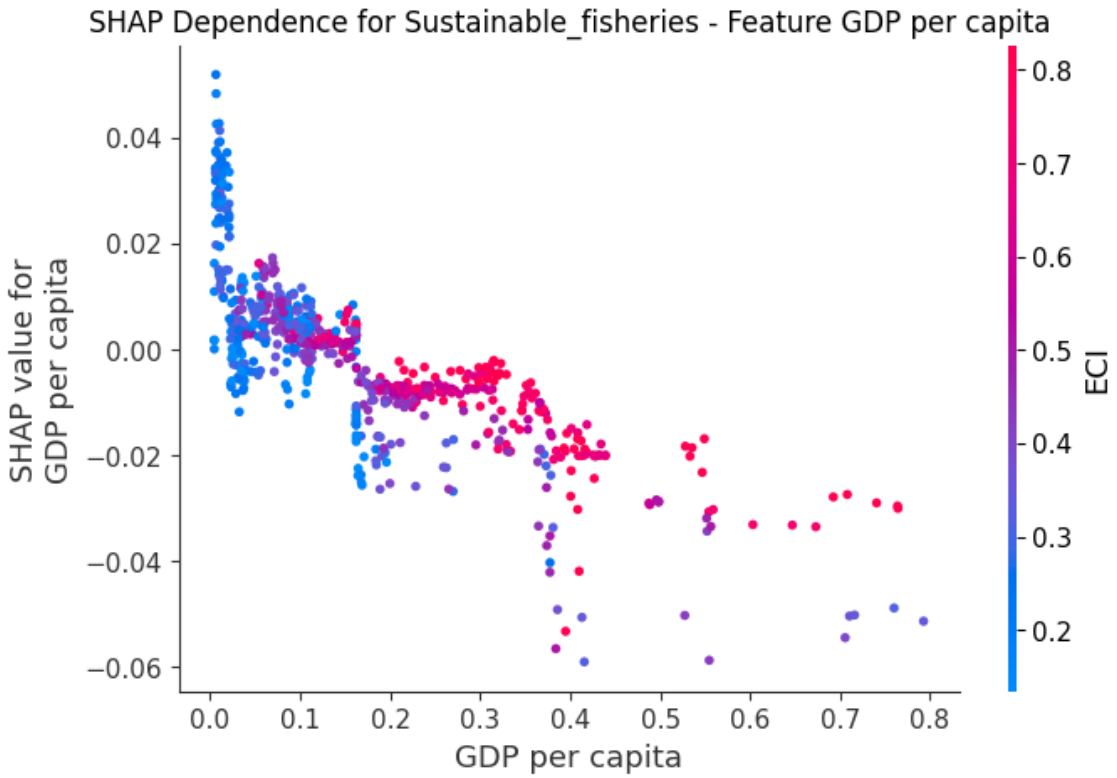

Figure 54: SHAP Dependence Plot for Sustainable Fisheries - GDP per capita

15.8 SDG 14: SHAP Summary Table

Table 12: Combined SHAP Summary Table for All Targets

Feature	Mean Abs SHAP Value	Ranking	Target
R&D expenditure	0.009	1	Beach_Litter
Change Over 5 Years	0.004	2	Beach_Litter
Land area (sq. km)	0.004	3	Beach_Litter
Population	0.003	4	Beach_Litter
Human Development Index	0.003	5	Beach_Litter
GDP per capita	0.002	6	Beach_Litter
COI	0.001	7	Beach_Litter
ECI	0.001	8	Beach_Litter
ECI*COI	0.001	9	Beach_Litter
Human Development Index*ECI	0.001	10	Beach_Litter
Land area (sq. km)	0.012	1	Chlorophyll
Population	0.010	2	Chlorophyll
COI	0.006	3	Chlorophyll
GDP per capita	0.005	4	Chlorophyll
ECI	0.005	5	Chlorophyll
Human Development Index	0.004	6	Chlorophyll
Human Development Index*ECI	0.003	7	Chlorophyll
R&D expenditure	0.002	8	Chlorophyll
ECI*COI	0.002	9	Chlorophyll
Change Over 5 Years	0.001	10	Chlorophyll
COI	0.032	1	Sustainable_fisheries
R&D expenditure	0.013	2	Sustainable_fisheries
GDP per capita	0.012	3	Sustainable_fisheries
Human Development Index	0.008	4	Sustainable_fisheries
ECI	0.007	5	Sustainable_fisheries
Population	0.004	6	Sustainable_fisheries
Land area (sq. km)	0.003	7	Sustainable_fisheries
Change Over 5 Years	0.002	8	Sustainable_fisheries