Predicting Human Development Using Machine Learning: Evidence from African Countries

By

Osman John A. T. Froko
Center for Economic and Political Governance in Africa (CEPGA), Ghana
Department of Applied Economics, University of Cape Coast, Ghana
Email: john.osman@stu.ucc.edu.gh
ORCID: https://orcid.org/0009-0002-9349-1971
&
Alhassan Abdul-Wakeel Karakara
Center for Economic and Political Governance in Africa (CEPGA), Ghana
Department of Applied Economics, University of Cape Coast, Ghana
Email: alhassan.karakara@ucc.edu.gh
&
Arimiyaw Zakaria
Department of Statistics, University of Cape Coast, Cape Coast, Ghana
Email: zakaria.arimiyaw@ucc.edu.gh
&
Victor Kwame Gyingyi
Center for Economic and Political Governance in Africa (CEPGA), Ghana
Department of Economic Studies, School of Economics, University of Cape Coast
Email: victor.gyingyi@stu.ucc.edu.gh

Abstract
This study employs Random Forest machine learning to predict the Human Development Index (HDI) for 41 African countries, with separate models developed for West Africa (comprising 16 countries) and Southern Africa (comprising 12 countries). It analyses an imbalanced panel dataset with 6.6% missing data, processed to 83 predictors from the ND-GAIN dataset after removing highly correlated features. The results identify economic wealth, environmental management, infrastructure, health, and governance as key drivers of the HDI. West Africa emphasizes rural areas, education, and health, reflecting its agrarian economy and early-stage development, while Southern Africa prioritizes economic output, governance, and urbanization, driven by advanced economies like South Africa. The models perform strongly (testing R-squared: 0.83–0.94), capturing complex, non-linear patterns missed by traditional methods. These findings support policymakers in targeting investments in education, health, and infrastructure, guide ECOWAS and SADC in developing regional strategies for agriculture, education, and governance, and assist the African Union in promoting human development and unity through evidence-based policies.

Keywords: Human Development Index, Random Forest, Africa, machine learning, economic development, regional cooperation, panel data, governance.
JEL Codes: O15, O55, C45, I15, I25.

Introduction

Imagine a world where every individual has the chance to thrive—to live long, learn deeply, and prosper in a life of dignity. This is the heart of human development, a global quest to unlock the potential of people everywhere. The Human Development Index (HDI), crafted by the United Nations Development Programme, is the yardstick for this ambition, blending three vital ingredients: life expectancy, education (mean and expected years of schooling), and per capita income (United Nations Development Programme, 2020). Think of HDI as a snapshot of a nation's well-being, capturing whether its people are not just surviving, but flourishing. Yet, across the globe, this dream faces fierce headwinds—poverty, inequality, and resource scarcity cast long shadows over progress (Sen, 1999). From bustling cities to remote villages, the challenge is clear: what are the secret ingredients that transform lives and lift nations toward higher HDI? This question burns brightest in regions where the stakes are highest, where every step forward is a triumph against the odds.

Africa, a continent of breathtaking diversity, resilience, and untapped potential, with 54 nations, each weaving its own tapestry of cultures, histories, and aspirations, is a kaleidoscope of development stories (Karakara & Osabuohien, 2021). Yet, it's no secret that the continent grapples with formidable challenges. Colonial legacies have left deep scars, shaping economic structures and governance systems that still influence today's realities (Acemoglu et al., 2001). Resource constraints, from limited infrastructure to uneven access to education and healthcare, create hurdles that demand creative solutions. But Africa is not just a story of struggle—it's a saga of breakthroughs. Take Rwanda, a nation that rose from the ashes of conflict to prioritize education and health, boosting its HDI and showing the world what's possible with vision and grit (World Bank, 2022). Yet, across the continent, progress is uneven. Why do some nations surge forward while others lag? What are the key levers for boosting HDI in African countries? Traditional studies have leaned on linear regression to crack this puzzle, but these methods often miss the messy, interconnected realities of development—like trying to solve a Rubik's cube with only one move (Breiman, 2001). A fresh approach is needed, one that can untangle the complex web of factors driving Africa's development.

This study uses machine learning to revolutionize how we understand HDI in African countries. We wield a Random Forest model—think of it as a digital detective, sifting through data to uncover hidden patterns and rank the importance of predictors with surgical precision. Unlike old-school methods, Random Forest thrives on complexity, capturing nonlinear relationships that others overlook (Breiman, 2001). Our mission? To pinpoint the most potent drivers of HDI across African nations, from education (mean years of schooling) and health (life expectancy) to economic indicators (GDP per capita), governance (corruption indices), and infrastructure (access to electricity).

We place our bets on human capital—the knowledge, skills, and health of a population—as the beating heart of development. Economists like Gary Becker and Amartya Sen have long championed human capital as the engine of progress, empowering individuals to build better lives and stronger societies (Becker, 1964; Sen, 1999). In African contexts, where resources may be scarce, investing in people could be the master key to unlocking HDI gains. This study uses the Random Forest method to predict HDI across the continent as a bold new frontier (Jean et al., 2016). We hypothesize that human capital, through education and health, will shine as the top predictor, given its transformative power in resource-constrained settings. Picture this: schools buzzing with eager learners, clinics saving lives, and communities empowered to shape their futures—these are the building blocks of Africa's development. Yet, other factors like governance and infrastructure can't be ignored, and our Random Forest model will rank their influence with data-driven clarity. Unlike global studies that gloss over Africa's unique context, we dive deep into the continent's realities, filling a critical gap in the literature (Asiedu, 2014). By delivering a robust ranking of HDI predictors, our study aims to light the way for policymakers, offering a roadmap to prioritize resources that maximize well-being. In short, we're not just predicting development, we are sparking a vision for Africa's future, powered by machine learning and rooted in the promise of human capital.

Drawing on this, we seek to tackle two main research questions: (i) What are the primary predictors of HDI in African countries? (ii) How effectively can a Random Forest machine learning model capture nonlinear relationships among HDI predictors in African countries?

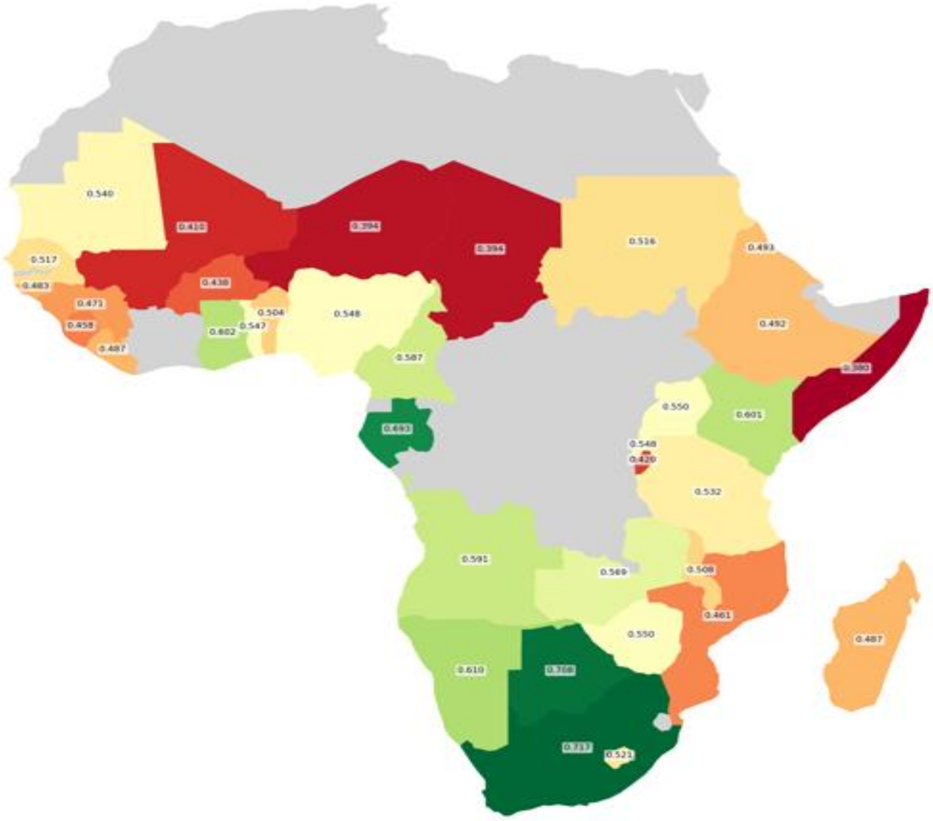

Figure 1: HDI Across African Countries, 2022
Source: Authors' compilation

Literature Review

Theoretical Literature Review

Theories of development are like compasses, each pointing to a different path for progress. Seven seminal works illuminate how HDI and human capital fit. First, modernization theory, as articulated by Rostow (1960), portrays development as a linear progression from traditional to modern societies, driven by industrialization and education. HDI's education and income components align with this view, suggesting that African nations can boost their HDI through investments in schools and factories. Yet, its one-size-fits-all approach stumbles in Africa, where colonial legacies disrupt the path (Acemoglu et al., 2001). Dependency theory counters with a grim tale of global exploitation. It argues that Africa's colonial past locks it into dependency on richer nations, draining resources needed for health and education, thus stunting HDI (Frank, 1966). While insightful, it risks painting Africa as a victim, ignoring successes like Rwanda, Mauritius, Seychelles, and Botswana. World-systems theory, developed by Wallerstein (1974), extends dependency by framing a global hierarchy of core, semi-peripheral, and peripheral nations. African countries, often peripheral, face structural barriers to HDI growth, yet human capital investments can shift their position, as seen in Botswana's government-driven progress.

The capability approach, pioneered by Sen (1999), redefines development as expanding freedoms (health, education, and opportunity). HDI embodies this, with human capital as its beating heart (Becker, 1964). This approach focuses on human development and empowerment. The provision of quality education and access to quality healthcare. Endogenous growth theory, as per Romer (1990), sees growth sprouting from within, driven by human capital and innovation. In African contexts, education fuels productivity, directly lifting HDI' s education and income scores. This theory resonates with our hypothesis that human capital drives development. Neoliberalism, outlined by Harvey (2005), pushes market-driven growth, arguing that free markets and minimal state intervention boost income and, indirectly, HDI. In Africa, however, structural adjustment programs tied to neoliberalism often cut education and health funding, slowing HDI progress (Asiedu, 2014). Finally, institutional theory, as explored by North (1990), emphasizes institutions, governance, and policies as development' s backbone. Strong institutions in African nations like Mauritius boost HDI by ensuring efficient resource allocation for schools and clinics (Acemoglu et al., 2001).

These theories converge on human capital's pivotal role, whether through education (modernization, endogenous growth), freedoms (capability approach), or institutions. Yet, they must grapple with Africa' s unique historical and structural challenges.

Empirical Literature Review

The literature has documented how HDI in Africa behaves and its determinants, as well as its benefits and otherwise. For instance, Ogundari and Abdulai (2014) investigate education impact on HDI in sub-Saharan Africa, using a meta-analysis of 1990–2010 data. They find that mean years of schooling strongly predict HDI, as education equips individuals for economic and social progress. This supports our focus on human capital but relies on linear models, missing complex interactions. Ranis et al. (2000) explore the nexus of economic growth and human development across developing countries, including African nations. Their panel data analysis (1970–1995) shows that health (life expectancy) and education drive HDI, with feedback loops to growth. Rwanda's post-2000 HDI surge echoes this, but the study's global scope dilutes Africa-specific insights.

Asiedu (2014) examines foreign aid's role in African education and growth, using data from 1990–2010. Aid targeted at education boosts HDI, but uneven distribution limits impact in low-income countries. This underscores human capital's potential in Africa, yet traditional regression limits its depth. Jean et al. (2016) pioneered machine learning in development, using Random Forest to predict poverty in sub-Saharan Africa with satellite imagery (2000‑2015). Their success in capturing non-linear patterns suggests Random Forest's power for HDI prediction, a gap our study fills (Breiman, 2001). Tobaigy (2022) analyzes HDI predictors globally, including African countries, using tree-based regression (2010‑2020). Health indicators (e.g., physicians per capita) emerge as top HDI drivers, reinforcing human capital's role. Their use of gradient boosting hints at machine learning's potential, though Africa-specific HDI applications are absent.

Omokanmi (2020) studied the financial inclusion impact on HDI in sub-Saharan Africa (2000‑2018). Using panel data, they find that access to bank accounts and loans boosts HDI by enabling education and health investments. This supports human capital's indirect effects but uses linear models, missing non-linear dynamics. Adewale (2020) focuses on South Africa's HDI trajectory (1990‑2020), using growth rate analysis. Education and health policies post-apartheid drive HDI gains, but inequality persists, highlighting governance's role. Their single-country focus limits generalizability to Africa's diverse HDI landscape (Figure 1).

Education and Human Capital as Drivers of HDI

Human capital's role in development has long been axiomatic, yet its manifestations in Africa expose fractures in classical models. Modernization paradigms, as articulated by Rostow (1960), posit education as a linear accelerator toward industrialization, mirrored in HDI's schooling metrics. Endogenous growth models (Romer, 1990) extend this by stressing knowledge spillovers, positing that educational investments yield compounding returns on productivity. Empirical validations from sub-Saharan Africa bolster these claims: Ogundari and Awokuse's (2014) meta-analysis (1990 to 2010) links mean schooling years to HDI uplifts, a pattern echoed in more recent inquiries. For instance, Kibona (2021) examines Tanzanian universities, uncovering how higher education curricula, often misaligned with local labor markets, dilute human capital's developmental yield despite nominal enrollment surges. Similarly, Mathebula (2021) dissects South Africa's higher education landscape for low-income students, revealing how affordability barriers and racial legacies entrench HDI stagnation, even as aggregate schooling metrics improve.

These linear associations, however, unravel under the scrutiny of contextual variances. Buckler (2023) highlights teacher training deficits across sub-Saharan Africa, where pedagogical gaps in rural versus urban settings engender uneven human capital formation, challenging the universality of growth-theoretic predictions. Yassim (2025) pushes further, advocating for vocational retooling of lecturers to foster sustainable development, yet cautions that without decolonial revisions, such efforts risk perpetuating extractive knowledge economies. Collectively, these studies affirm education's leverage on HDI but underscore non-linear thresholds, such as critical mass in quality over quantity, that conventional regressions obscure, particularly amid Africa's youth bulge and migration pressures. Adding to this, recent research by Adeyeye (2024) using the generalized method of moments on data from 28 sub-Saharan African countries from 2002 to 2018 found that the impact of education on economic growth, a key component influencing HDI, is largely insignificant without strong governance quality. Specifically, secondary education showed a positive effect only when moderated by regulatory quality, highlighting how governance indicators like corruption control and rule of law are essential to unlocking education's potential for broader human development outcomes.

Comparative insights from Asia amplify these observations. In Kanbur and Venables (2005), chapters on regional inequality in China (Kanbur & Zhang, 2005) trace how central planning and post-reform openness widened educational disparities between coastal provinces and inland hinterlands, paralleling Africa's north-south divides in countries like Nigeria or Kenya. Similarly, industrial location theories in India (Lall & Chakravorty, 2005) reveal how agglomeration economies in urban clusters concentrate skilled labor, leaving peripheral regions with stunted human capital, a dynamic akin to Johannesburg's pull versus rural Limpopo's lag in South Africa. These Asian cases suggest that spatial targeting in education policy, such as decentralized vocational hubs, could mitigate HDI gradients in Africa, where geographic isolation compounds access barriers beyond mere enrollment rates. Moreover, drawing from global education overviews, nearly a quarter of young people aged 18 to 24 in Southern and Northern Africa are not in employment, education, or training, underscoring the urgent need for policies that bridge education with labor market demands to enhance HDI.

Health, Wellbeing, and the Capability Approach

Sen's (1999) capability lens recasts development as the amplification of functioning, with HDI's life expectancy pillar as a crude proxy for health freedoms. Ranis et al. (2000) trace bidirectional cascades: health bolsters economic agency, which reciprocates via reinvestments. In Africa, these loops have driven sporadic HDI advances, yet pandemics and inequities have severed them. Life expectancy rebounds post-HIV interventions propelled HDI in the 2010s, but COVID-19 reversed gains, with sub-Saharan averages dipping 1.5 years by 2022 (World Health Organization, 2023). Tobaigy and Alshehri's (2022) global machine learning analysis elevates health proxies, physician density, and immunization coverage, as HDI dominants, a finding resonant in African vignettes where rural-urban health divides mirror HDI chasms. Recent statistics indicate that sub-Saharan Africa's HDI score stood at approximately 0.57 in 2023, reflecting low human development levels heavily influenced by health challenges.

Nuance emerges from regional dissections: Ugherughe (2023) correlates stock market vitality with HDI in Mauritius, Nigeria, and South Africa, attributing health-HDI synergies to governance-mediated fiscal flows, yet flags Nigeria's volatility as a cautionary tale of resource curses undermining health infrastructure. Bello (2021) extends this to Nigeria's tax-governance nexus, showing how compliance shortfalls erode health funding, stalling capability expansions. These insights reveal health not as a monolithic driver but as a fulcrum strained by fiscal precarity and institutional inertia, where global ML models like Tobaigy and Alshehri (2022) falter without Africa-tuned hyperparameters. Furthermore, studies on healthcare systems in Africa identify key challenges such as inadequate infrastructure, workforce shortages, and funding gaps, suggesting that evidence-based policies are crucial for improving health outcomes and, consequently, HDI.

Drawing from Asian spatial analyses, Kanbur and Venables (2005) illuminate how health disparities underpin broader HDI inequalities. For instance, poverty mapping in Vietnam (Minot & Baulch, 2005) using aggregate data exposes commune-level health access gaps exacerbated by trade liberalization, echoing African borderland vulnerabilities in the Sahel, where mobility restrictions hinder immunization drives. In Central Asia (Anderson & Pomfret, 2005), post-Soviet transitions amplified rural health declines, akin to Zimbabwe's post-land reform health collapses, highlighting how locational isolation, distance to facilities, and transport deficits constrain Sen's capabilities more acutely than aggregate metrics suggest. These parallels advocate for geospatial health interventions in Africa, such as mobile clinics calibrated to topographic barriers, to forge resilient HDI-health linkages amid climate-induced migrations. Additionally, associations between HDI and HIV prevalence in regions like East and Southern Africa show positive relationships, indicating that health improvements are vital for sustaining HDI progress.

Governance, Institutions, and Structural Constraints

Governance architectures determine resource conversion efficiencies, with North's (1990) institutional economics and Wallerstein's (1974) world-systems lens framing Africa's peripheral entanglements. Empirical mappings confirm: Acemoglu et al. (2001) contrast Mauritius and Botswana's institutional robustness, via anti-corruption pacts and land reforms, with HDI laggards, where elite capture vitiates public goods. Post-apartheid South Africa's HDI ascent, per Adewale (2020), hinged on equity-focused reforms, yet Gini persistence signals governance's incomplete mediation.

Structural adjustment's neoliberal imprint lingers toxically: Harvey (2005) and Asiedu (2014) document slashed social spending's HDI drag, a legacy amplified in recent analyses. Sengupta (2023) probes non-performing loans' drag on growth, mediated by political governance, revealing how opaque institutions in West Africa amplify credit risks, throttling HDI via choked investments. Abayomi (2023) integrates fertility and wage minima into Nigeria's HDI equation, exposing how weak labor governance perpetuates demographic traps. These strands illuminate governance as a contingent filter, amplifying human capital in stable enclaves like East Africa, while entrenching disparities elsewhere, demanding models that parse path dependencies over static correlations. Recent research emphasizes that good and inclusive governance is imperative for Africa's future, with mixed progress in governance landscapes highlighting unmet expectations and the need for mechanisms promoting constitutionalism, rule of law, and citizen participation.

Asian governance insights from Kanbur and Venables (2005) enrich this narrative. The Maoist insurgency in Nepal (Murshed & Gates, 2005) links horizontal spatial inequalities, uneven resource allocation across ethnic regions, to conflict, a motif resonant with Boko Haram's Sahelian grievances in Nigeria, where governance failures in peripheral zones fuel HDI erosions through violence. In the Philippines (Balisacan & Fuwa, 2005), decentralization reforms unevenly mitigated spatial income gaps, paralleling Ethiopia's federalism experiments that boosted HDI in Amhara but faltered in Somali peripheries due to institutional asymmetries.

These cases underscore governance's spatial dimensionality: central policies often favor cores, widening HDI chasms unless calibrated with subnational autonomy, a lesson for African devolution efforts in Kenya or South Africa to counter elite-centric distortions. Moreover, studies on governance and economic growth in Africa using panel data from 2008 onwards show varying impacts, with governance quality positively influencing growth in several specifications.

Machine Learning and Predictive Approaches to Development

Econometric hegemony in HDI scholarship assumes linearity, sidelining Africa's combinatorial complexities, interactions among climate shocks, remittances, and policy shocks. Machine learning disrupts this: Jean et al.'s (2016) Random Forests unearth poverty topographies from satellite data in sub-Saharan Africa, capturing granularities eluding OLS. Tobaigy and Alshehri's (2022) gradient boosting crowns health as global HDI lodestar, yet global scopes dilute Africa-specific signals.

Recent advances sharpen continental focus: Druckenmiller and Hsiang (2021) fuse satellite imagery with ML for high-resolution HDI maps, illuminating intra-national variances in East Africa overlooked by aggregates. Feil (2023) deploys classification algorithms to interrogate HDI cutoffs, revealing non-linear governance-health thresholds in North Africa. Ofori (2023) leverages ML for inclusive growth predictions, isolating institutional quality's outsized role in West African HDI trajectories. Wahab's (2024) synthesis of ML-poverty tools advocates satellite integration for real-time HDI forecasting, addressing data scarcities in fragile states. Yet, ethical pitfalls loom: Gwagwa (2024) and Hassan (2024) warn of AI governance voids in Africa, where biased training data could entrench HDI inequities. Our Random Forest application (Breiman, 2002) navigates this by prioritizing interpretable feature rankings, fusing thematic drivers into Africa-centric prognoses. A review and meta-analysis of ML trends in poverty mapping from 2014 to 2023 highlights Random Forest's prevalence for its interpretability and accuracy, with nighttime light indices combined with daytime features improving poverty estimations, which are proxies for HDI components.

Asian precedents from Kanbur and Venables (2005) prefigure these ML potentials. Commune-level poverty estimation in Cambodia (Fujii, 2005) employs small-area techniques akin to early geospatial ML, dissecting HDI proxies at micro-scales to reveal aid misallocations, a blueprint for African applications in data-sparse Liberia or Mali, where satellite-augmented models could pinpoint health-education synergies. Poverty mapping with aggregate census data (Minot & Baulch, 2005) in Vietnam quantifies aggregation biases, informing ML error mitigation in African censuses plagued by underreporting in nomadic zones. China's fifty-year inequality chronicle (Kanbur & Zhang, 2005) demonstrates longitudinal spatial modeling's value, suggesting time-series Random Forests for Africa's post-colonial HDI arcs to unmask regime shifts invisible to static regressions. These integrations propel ML beyond prediction toward prescriptive spatial equity, tailoring interventions to Africa's topographic mosaics. Furthermore, convolutional neural networks have been used to estimate subnational HDI globally using satellite imagery, achieving low mean absolute errors, offering scalable solutions for monitoring development in Africa.

Theories spotlight human capital as development's engine, whether through modernization's education push, the capability approach's freedoms, or endogenous growth's innovation. Empirical studies confirm this, showing education and health as top HDI predictors in Africa, alongside governance and infrastructure. However, most rely on linear models, missing complex interactions that machine learning, like Random Forest, can capture (Breiman, 2001). While Jean et al. (2016) and Tobaigy (2022) leverage machine learning, their focus on poverty or global HDI leaves a gap in Africa-specific HDI prediction. Our study bridges this, using Random Forest to rank predictors across African nations.

Data and Methodology

The dataset is an imbalanced panel covering 41 African countries, combining HDI values with approximately 170 predictors before preprocessing, sourced from reputable international databases. These predictors span economic performance, poverty, remittances, external debt, human capital, governance, and vulnerability, capturing a diverse range of factors relevant to HDI. After preprocessing, the number of features is reduced to enhance model efficiency and interpretability, with the final count determined post-correlation analysis and feature selection.

To address the imbalanced panel, we apply imputation techniques to handle missing values. Numerical predictors are imputed using the median to minimize outlier influence, while predictors with excessive missingness (30%) are excluded to avoid bias. For time-series data within countries, forward-fill or linear interpolation is used to maintain continuity, ensuring robust results. Correlation analysis mitigates multicollinearity, which can impact Random

Forest performance (Breiman, 2001). Pairwise Pearson correlation coefficients are calculated, and for predictor pairs with correlations above 0.8, one is retained based on its theoretical relevance to HDI. All predictors are standardized to ensure uniform scales, enhancing model stability and interpretability. This preprocessing is applied consistently across the full dataset and the two subregional subsets (West and Southern Africa).

This study employs the Random Forest technique to identify predictors of the Human Development Index (HDI) across 41 African countries. Three models are estimated: one using data from all 41 countries to provide a general perspective, and two others focusing on subregional groups, West African nations, and Southern African nations to explore potential differences in HDI predictors across regions. The methodology includes data pre-processing to handle an imbalanced panel dataset, model training, hyperparameter tuning, and performance evaluation to ensure robust and interpretable results.

The Random Forest algorithm, an ensemble of decision trees introduced by Breiman (2001), is selected for its ability to model non-linear relationships and complex feature interactions, with advancements in robustness and splitting criteria noted by Hastie et al. (2009). Three models are estimated to capture both general and subregional patterns in HDI predictors: 1. General Model: Uses data from all 41 African countries to identify overarching HDI predictors, providing a continent-wide perspective on factors influencing human wellness. 2. West Africa Model: Focuses on a subset of West African nations to examine region-specific predictors, accounting for unique socioeconomic and governance contexts. 3. Southern Africa Model: Analyzes a subset of Southern African nations to identify predictors relevant to this region, enabling comparison with West Africa to highlight subregional differences. Each model is implemented using Python's scikit-learn library, with the following steps: 1. Dataset Splitting: For each model, the relevant dataset (full, West African, or Southern African) is split into a training set (70%) and a testing set (30%) using stratified sampling to preserve the HDI distribution, ensuring representative subsets. 2. Hyperparameter Tuning: Grid search with 5-fold cross-validation optimizes key hyperparameters (e.g., number of trees, maximum tree depth, minimum samples per split, minimum samples per leaf) for each model. The best configuration is selected based on the lowest Mean Squared Error (MSE) on validation folds. 3. Model Training: Each Random Forest model is trained on its respective training set using the optimal hyperparameters. 4. Model Testing: Each trained model predicts HDI values on its corresponding testing set.

Performance for each model is assessed on both training and testing sets using three metrics to evaluate predictive accuracy and detect overfitting: Mean Squared Error (MSE), which measures the average squared difference between predicted and actual HDI values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

where y_i is the actual HDI, $\hat{y}_i$is the predicted HDI, and n is the number of observations.

The Mean Absolute Error (MAE) captures the average absolute difference between predicted and actual HDI values as shown in equation 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Significant discrepancies between training and testing metrics indicate overfitting, which is addressed by adjusting hyperparameters or reducing feature complexity based on correlation analysis. Performance comparisons across the three models highlight subregional variations in HDI predictors. For each model, feature importance scores, derived from the Random Forest's average impurity reduction across trees (Breiman, 2001), are analyzed to rank predictors of HDI. This identifies key drivers of human wellness in the general model and reveals subregional differences in West and Southern Africa, guiding targeted policy interventions. The analysis is conducted using Python, with scikit-learn for Random Forest modeling, pandas and NumPy for data preprocessing, SciPy for correlation analysis, and matplotlib for visualization. Code is maintained in a public repository for reproducibility.

Results and discussion

General Model

The Random Forest model predicting the Human Development Index (HDI) identifies economic wealth, environmental management, infrastructure, health, governance, and technology as key drivers of human development. Using 83 predictors after removing 71 highly correlated features from an initial set of 154, the model's performance metrics and feature importance rankings (Table 1) reveal their relative weights. The actual versus predicted HDI plot (Figure 2) confirms the model's accuracy. This section discusses the data's insights into HDI drivers. The model demonstrates strong performance. Training metrics show a Mean

Squared Error (MSE) of 0.000021, an R2 of 0.9983, and a Mean Absolute Error (MAE) of 0.002935, explaining 99.83% of HDI variance with predictions deviating by 0.002935 units on average. Testing metrics–MSE of 0.000647, R2 of 0.9439, and MAE of 0.018931–capture 94.39% of variance with errors of 0.018931 units, indicating robust generalization despite slight overfitting. Hyperparameters (maximum depth of 10, square root of features per split, 300 trees) ensure reliability (Breiman, 2001; Hastie et al., 2009). Figure 2 shows predicted HDI values closely matching actual values, with minor deviations consistent with the reported MAE.

Table 1: The top 15 predictors driving HDI in Africa

Code name	Feature description	Score
GDP_PC_PPP	Per Capita GDP(PPP)	0.0944
Id_ecos_05	Protected Biome Coverage	0.0849
Water-Access	Reliable Drinking Water Access	0.0753
Age Dependency Ratio	Age Dependency Ratio	0.0623
Elec_Access	Electricity Access	0.0563
Sanit_Access	Improved Sanitation Access	0.0475
Under5_Mort	Under-5 Mortality Rate	0.0475
ICT	ICT Infrastructure	0.0447
Vulnerability_health	Health Vulnerability	0.0430
Agri_GDP	Agricultural GDP Share	0.0424
Institutions	Institutional Quality	0.0371
Energy Use Per Capita	Energy Use Per Capita	0.0362
Structural Change	Economic Structural Change	0.0352
Id_ecos_03	Natural Capital Dependency	0.0252
Id_food_04	Rural Population share	0.0213

Source: Authors'

The top 15 predictors driving HDI are listed in Table 1. Per capita GDP (score: 0.0944) is the leading predictor, underscoring that economic wealth drives HDI by enabling investments in health and education. Protected biome coverage (id ecos 05, score: 0.0849), defined by ND-GAIN as biome protection weighted by territory (ND-GAIN, 2024), ranks second, highlighting the critical role of environmental management. Infrastructure-related factors reliable drinking water access (0.0753), electricity access (0.0563), and improved sanitation access (0.0475), are also significant. Health factors, including under-5 mortality rate (0.0457) and health vulnerability (0.0430), demonstrate healthcare's substantial impact. ICT infrastructure (0.0447) underscores technology's role in development. Agricultural GDP share (0.0424) and natural capital dependency (id ecos 03, 0.0252), an ND-GAIN indicator of reliance on crops and forests (ND-GAIN, 2024), reflect the influence of agricultural economies. Institutional quality (0.0371) emphasizes the importance of governance. Other factors, such as age dependency ratio (0.0623), energy use per capita (0.0362), economic structural change (0.0352), and rural population share (id food 04, 0.0213), highlight demographic and economic influences.

The model's findings align with several theoretical and empirical perspectives in the literature. Modernization theory (Rostow, 1960) and endogenous growth theory (Romer, 1990) emphasize economic growth, consistent with the prominence of per capita GDP and ICT infrastructure. The capability approach (Sen, 1999) aligns with health and infrastructure predictors, which enable individual freedoms. Institutional theory (North, 1990; Acemoglu et al., 2001) is supported by the role of institutional quality. Empirical studies (Ogundari & Abdulai, 2014; Ranis et al., 2000; Tobaigy, 2022) confirm health's importance, though infrastructure outranks education in this model, diverging from findings in (Ogundari & Abdulai, 2014; Asiedu, 2014). The role of governance aligns with (Omokanmi, 2020; Adewale, 2020). The Random Forest approach, as used in (Jean et al., 2016; Tobaigy, 2022), effectively captures non-linear patterns, unlike linear models in (Ogundari & Abdulai, 2014; Ranis et al., 2000; Omokanmi, 2020). However, the model challenges certain aspects of the literature by prioritizing environmental factors (id ecos 05), which are less prominent in most theoretical frameworks. Dependency theory (Frank, 1966) and world-systems theory (Wallerstein, 1974) focus on external exploitation, yet this model highlights internal environmental management as a key HDI driver. Neoliberalism (Harvey, 2005) emphasizes market-driven development, but governance and infrastructure are equally critical here. The dominance of infrastructure over education suggests a broader HDI foundation than proposed by (Ogundari & Abdulai, 2014; Asiedu, 2014). The non-linear modeling approach reveals these complexities, providing a nuanced understanding of HDI drivers.

The data indicate that HDI in Africa depends on economic wealth, environmental stewardship, infrastructure, health, governance, and technology, with environmental factors playing a more significant role than much of the literature suggests. The model's high accuracy (Table 1, Figure 2) supports these findings, aligning with human capital and governance theories while elevating the roles of environmental management and infrastructure.

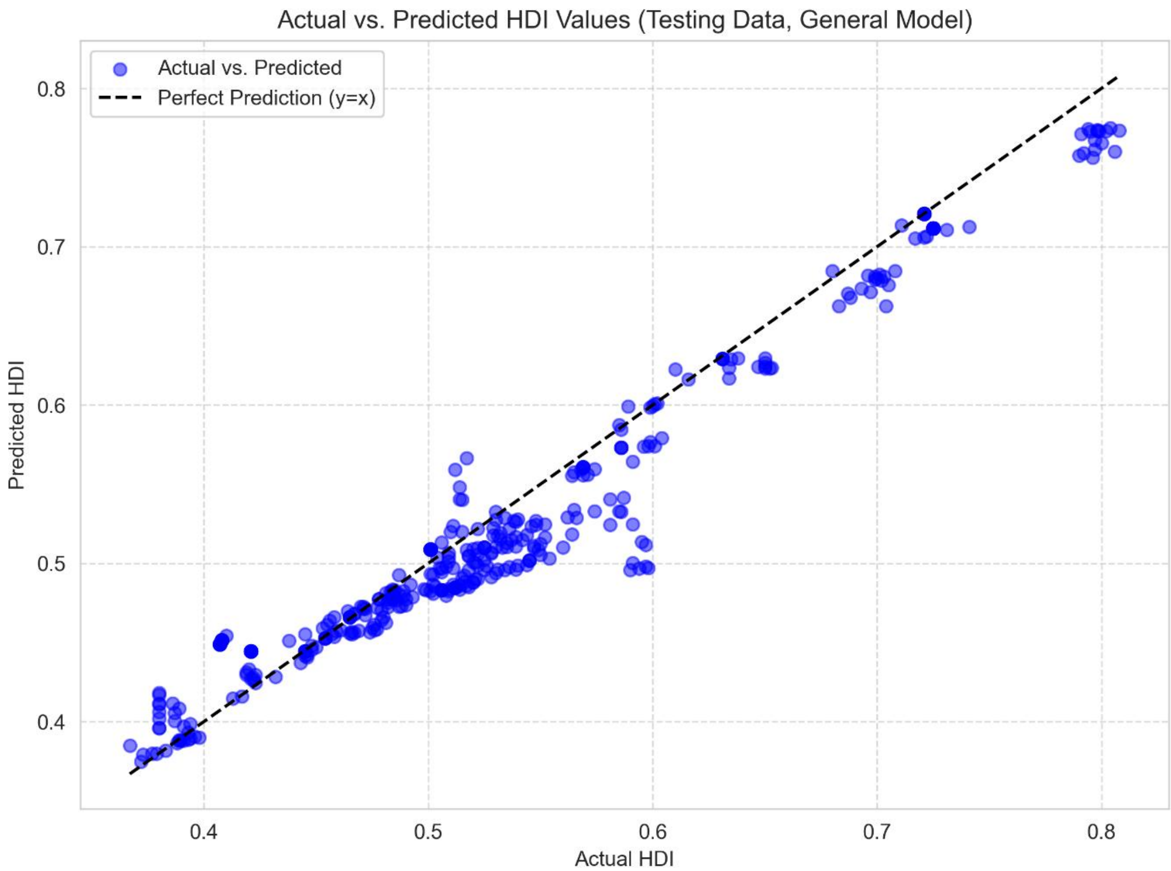

Figure 2: Actual vs. Predicted HDI Values.
Source: Authors

Drivers of Human Development in West and Southern Africa

West Africa (represented by 16 countries) and Southern Africa (represented by 12 countries). The models highlight distinct regional priorities: West Africa emphasizes rural and basic needs, while Southern Africa focuses on economic production and governance. See tables 2 and 3.

West Africa: Rural and Basic Needs

The Random Forest model for West Africa predicts HDI using 83 predictors after removing highly correlated features from an initial set of 154. The model shows strong training performance (R-squared 0.9896, MSE 0.000040, MAE 0.003801), explaining 98.96% of HDI variance with predictions deviating by 0.003801 units on average. Testing performance is moderate (R-squared 0.8311, MSE 0.000615, MAE 0.017541), capturing 83.11% of variance with errors of 0.017541 units, indicating slight overfitting due to data variability. Hyperparameters (maximum depth of 10, square root of features per split, 300 trees) ensure reliability (Breiman, 2001).

The model identifies rurality, education, health, and environmental vulnerabilities as key HDI drivers. West Africa's predominantly agrarian economy, with 70% of the population in rural areas (e.g., Mali, Niger), makes food security a critical constraint. Low literacy (e.g., 34% in Niger) and limited tertiary enrolment (9% on average) hinder human capital development, a core HDI component. High child mortality rates (e.g., 92 per 1,000 in Niger) and low health spending ($30 per capita in Liberia) reflect weak healthcare systems. Environmental risks, such as climate-driven marine biodiversity loss and limited biome protection (10% in Mali), threaten coastal and rural livelihoods. Poor infrastructure, with only 15% of roads paved in Chad, and low ICT penetration (20% internet use in Niger) further limit progress. These findings align with modernization theory (Rostow, 1960), which emphasizes basic needs in early-stage development, and the capability approach (Sen, 1999), which highlights health and education as enablers of individual freedoms. Empirical studies (Ogundari and Abdulai, 2014; Ranis et al.,2000) confirm health and education, though environmental factors are more prominent here than in prior work. Table 2 lists the top 15 predictors and their importance scores.

Table 2: Top 15 HDI Drivers in West Africa

Feature	Definition	Importance
Rural population	Proportion in rural areas	0.0918
Human capital	Education and skills	0.0857
Education	Tertiary enrolment	0.0675
Per capita GDP (PPP)	GDP per capita, PPP	0.0589
Child mortality	Under-5 mortality rate	0.0505
Adjusted NNI per capita	Income after depreciation	0.0458
Coastal population	Pop. under 5m above sea level 1	0.0401
Age dependency ratio	Dependents to working-age pop.	0.0388
Marine biodiversity	Biodiversity change	0.0384
Protected biome	Biome protection level	0.0318
Social inequality	Poorest quintile's income share	0.0264
Health expenditure	Per capita health spending	0.0244
Transport infrastructure	Quality of trade/transport	0.0222
GDP(PPP)	Total GDP, PPP	0.0207
ICT infrastructure	Mobile, internet access	0.0205

Source: Authors'

Southern Africa: Economic and Institutional Strengths

The Random Forest model for Southern Africa also uses 83 predictors, showing excellent training performance (R-squared 0.9983, MSE 0.000026, MAE 0.002870), explaining 99.83% of HDI variance with predictions deviating by 0.002870 units. Testing performance is robust (R-squared 0.8892, MSE 0.001382, MAE 0.027535), capturing 88.92% of variance with errors of 0.027535 units, reflecting better generalization than West Africa, likely due to South Africa's data consistency. Hyperparameters mirror those of West Africa (Breiman, 2001). The model highlights economic wealth, governance, urbanization, and health as key HDI drivers. South Africa’s high per capita GDP ($13,000 PPP) and Botswana’s economic stability drive income, a core HDI component. Strong governance, particularly in South Africa, enhances policy effectiveness, unlike West Africa’s weaker systems. Urbanization (67% in South Africa) boosts access to services, but 30% of urban residents live in slums, posing challenges. Gender disparities, with 40% of South African women in informal jobs, limit HDI. Agricultural productivity and climate-driven yield declines remain critical in rural areas like Zambia and Zimbabwe. Health challenges, including high child mortality (70 per 1,000 in Malawi) and reliance on external health funding (20% in Malawi), persist. These findings align with endogenous growth theory (Romer, 1990), emphasizing economic diversification, and institutional theory (North, 1990; Acemoglu et al., 2001), highlighting governance. Empirical studies (Omokanmi, 2020; Adewale, 2020) support governance and health’s roles, though gender issues are more prominent here than in (Ogundari and Abdulai, 2014). Table 3 lists the top 15 predictors and their importance scores.

Table 3: Top 15 HDI Drivers in Southern Africa

Feature	**Definition**	**Importance**
Per capita GDP(PPP)	GDP per capita, PPP	0.0946
Child mortality	Under-5 mortality rate	0.0782
Institutions	Governance quality	0.0658
Vulnerable female employment	Informal female jobs	0.0628
Cereal yield change	Projected yield change	0.0523
Agricultural productivity	Value added per worker	0.0463
Urban population	Urban population ratio	0.0445
Habitat Vulnerability	Environmental risks	0.0436
Fertilizer consumption	Agricultural intensity	0.0385
Structural change	Shift to industry/service	0.0376
Investment climate	Doing business indicators	0.0361
Industrial productivity	Value added per worker	0.0345
External health funding	Foreign health expenditure	0.0285
Slum population	Urban slum proportion	0.0233
Vector-borne disease	Malaria transmission change	0.0193

Source: Authors’

Why the Differences?

The regions’ HDI drivers differ due to distinct socioeconomic, institutional, and environmental contexts: Considering economic stages, West Africa’s focus on rurality, education, and health reflects early-stage development, prioritizing basic needs. Southern Africa’s emphasis on income and governance aligns with advanced economies like South Africa. Modernization theory (Rostow, 1960) explains this shift from basic needs to economic output. 2. Economic Structure: West Africa's agrarian economy contrasts with Southern Africa's industrial and urban base. Structural transformation theory (Lewis, 1954) supports Southern Africa's focus on industry and services. Again, strong institutions in South Africa (e.g., political stability) drive HDI, unlike West Africa's weaker governance (e.g., Nigeria, Burkina Faso). Institutional theory (North, 1990; Acemoglu et al., 2001) highlights governance's role in attaining high HDI.

Also, gender, the environment, and demographic issues play a role. In Southern Africa, the focus is on vulnerable female employment reflecting a formalized labor market, while West Africa prioritizes broader inequality. Feminist development theory (Boserup, 1970) explains gender's prominence in advanced economies. In demographics, West Africa's high fertility (e.g., 7.6 births per woman in Niger) drives dependency ratios, while Southern Africa's urbanization shifts focus. This finding is supported by the demographic transition theory (Notestein, 1945). On the environment, West Africa's marine and biome vulnerabilities reflect rural risks, while Southern Africa's urban and agricultural risks align with its stage. The Environmental Kuznets curve (Grossman & Krueger, 1995) explains environmental priorities in early development. These differences challenge dependency theory (Frank, 1966), which focuses on external exploitation, as internal factors like governance and infrastructure are prominent. Neoliberalism (Harvey, 2005) emphasizes markets, but health and environmental factors also matter significantly.

Discussion

This study employs Random Forest models to predict the Human Development Index (HDI) across 41 African countries, with subregional analyses for West Africa (16 countries) and Southern Africa (12 countries), addressing two key research questions: (1) What are the primary predictors of HDI in African countries? (2) How effectively can a Random Forest model capture non-linear relationships among HDI predictors? The findings, presented in Tables 1, 2, and 3, provide robust answers, revealing distinct regional priorities and offering actionable insights for policymakers, subregional bodies like ECOWAS, and the broader pursuit of African unity.

Addressing the Research Questions

The first research question investigates the primary predictors of HDI in African countries. The general model (Table 1) identifies economic wealth (per capita GDP, 0.0944 importance), environmental management (protected biome coverage, 0.0849), infrastructure (water access, 0.0753; electricity access, 0.0563), health (under-5 mortality, 0.0457), governance (institutional quality, 0.0371), and technology (ICT infrastructure, 0.0447) as top drivers. These align with modernization theory (Rostow, 1960), emphasizing economic growth, and the capability approach (Sen, 1999), highlighting health and education as enablers of human freedoms. The prominence of environmental factors diverges from traditional studies (Ogundari and Abdulai, 2014; Ranis et al., 2000), suggesting sustainable resource management is critical in Africa's resource-dependent economies. Subregional models reveal nuanced differences: West Africa prioritizes rurality (rural population, 0.0918), education (human capital, 0.0857; tertiary enrollment, 0.0675), and health (child mortality, 0.0505), reflecting its agrarian, early-stage development (Table 2). Southern Africa emphasizes economic output (per capita GDP, 0.0946), governance (institutions, 0.0658), and urbanization (urban population, 0.0445), driven by advanced economies like South Africa (Table 3). These findings confirm human capital's pivotal role, as hypothesized (Becker, 1964), but highlight regional variations shaped by economic and institutional contexts.

The second research question assesses the Random Forest model's ability to capture non-linear relationships among HDI predictors. The models demonstrate strong performance, with the general model achieving a testing R-squared of 0.9439 and MAE of 0.018931, West Africa's model at 0.8311 and 0.017541, and Southern Africa's at 0.8892 and 0.027535. These metrics indicate robust generalization, capturing 83–94% of HDI variance. Unlike linear regression used in prior studies (Ogundari & Abdulai, 2014; Ranis et al., 2000), Random Forest excels at modeling complex interactions, such as the interplay between rurality and environmental vulnerabilities in West Africa or governance and urbanization in Southern Africa. Feature importance scores, derived from impurity reduction (Breiman, 2001), provide precise rankings that reveal non-linear dynamics, such as governance's amplified impact in Southern Africa (0.0658) compared to West Africa. This confirms the model's effectiveness in handling Africa's complex development landscape, offering a methodological advance over traditional approaches. Thus, this study robustly answers its research questions, identifying key HDI predictors and demonstrating Random Forest's effectiveness in capturing non-linear relationships.

Policy implications

The general model' s findings suggest African governments prioritize balanced investments in economic growth, infrastructure, and human capital. Improving water access (0.0753) and electricity (0.0563) can enhance health and education outcomes, directly boosting HDI. Environmental management, particularly biome protection (0.0849), is critical for resource-dependent economies, aligning with the Environmental Kuznets curve (Grossman and Krueger, 1995). Sustainable land use policies can mitigate risks like marine biodiversity loss in West Africa (0.0384) or cereal yield declines in Southern Africa (0.0523). In West Africa, the focus on rurality (0.0918) and human capital (0.0857) calls for investments in rural infrastructure (e.g., only 15% of roads paved in Chad) and education (9% tertiary enrolment). Health systems must address high child mortality (92 per 1,000 in Niger) through increased spending (currently $30 per capita in Liberia). These align with modernization theory (Rostow, 1960), emphasizing basic needs. Southern Africa's emphasis on governance (0.0658) and economic output (0.0946) suggests strengthening institutions, as in South Africa, and addressing urban challenges like slums (30% of urban population) and gender disparities (40% of women in informal jobs). Endogenous growth theory (Romer, 1990) supports investments in industrial productivity (0.0345) to sustain HDI gains.

Implications for Africa

For the Economic Community of West African States (ECOWAS), the West Africa model highlights the need for regional cooperation to address shared challenges. Rurality and food security (0.0918 importance) are critical, given the region's 70% rural population. ECOWAS can develop regional agricultural programs, such as cooperative farming initiatives, drawing on Ghana' s coastal resilience efforts. Education and health investments, targeting low literacy (34% in Niger) and high child mortality (0.0505), could include regional teacher training and vaccination campaigns. Infrastructure development, like cross-border road networks (0.0222), can improve market access. Environmental vulnerabilities (marine biodiversity, 0.0384) call for regional climate adaptation strategies, such as shared coastal management.

The Southern African Development Community (SADC) can leverage the Southern Africa model's focus on governance (0.0658) and economic diversification. Strong institutions in South Africa and Botswana provide a blueprint for regional standards, supporting weaker systems in Zimbabwe. Urban planning initiatives to address slums (0.0233) and gender policies for informal female workers (0.0628) can build on South Africa' s frameworks, fostering regional equity.

The study's subregional differences underscore the need for the African Union (AU) to integrate tailored strategies into Agenda 2063. The general model's emphasis on human capital (e.g., education, health) and sustainability (biome protection, 0.0849) aligns with AU's goals. A continent-wide human capital fund could channel resources to education in West Africa and urban health systems in Southern Africa. Environmental management's prominence across models supports AU's climate resilience objectives, promoting unified sustainability efforts. To enhance research and policy, the AU should prioritize collecting and making available comprehensive data on member states, standardizing metrics like HDI predictors to support robust modeling and evidence-based decisions. This would strengthen continent-wide development strategies, fostering unity through shared progress.

Suggestions for Future Research

To build on this study, future research could explore innovative approaches to deepen insights into HDI drivers. Integrating geospatial data, such as satellite imagery of urban and rural landscapes, could enhance predictions by capturing spatial patterns in infrastructure and environmental conditions, building on work by Jean et al. (2016). Sentiment analysis of social media data from platforms like X could gauge public perceptions of development policies, offering a dynamic measure of social capital not captured in current predictors. Agent-based modeling could simulate interactions between economic, social, and environmental factors, providing a systems-level view of HDI dynamics across African contexts. Incorporating cultural indicators, such as linguistic diversity or community cohesion, could enrich theoretical frameworks like the capability approach (Sen, 1999), addressing gaps in understanding social drivers. Finally, hybrid machine learning models, combining Random Forest with deep learning, could uncover temporal and non-linear patterns in longitudinal HDI data, offering a forward-looking perspective on Africa's development trajectory.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review, 91(5):1369–1401.*

Adewale, T. (2020). Human development index trajectories in South Africa: A growth analysis. *South African Journal of Economics, 88(1):45–62.*

Adeyeye, O. (2024). Education, governance, and economic growth in Sub-Saharan Africa: A panel data analysis. *African Development Review*, 36(1), 101–120.

Asiedu, E. (2014). Does foreign aid in education promote economic growth? evidence from sub-Saharan Africa. *Journal of African Economies, 23(1):1–36.*

Becker, G. S. (1964). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press, Chicago.

Boserup, E. (1970). Women's role in economic development. George Allen & Unwin, London.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Frank, A. G. (1966). The development of underdevelopment. Monthly Review, 18(4):17–31.

Grossman, G. M. and Krueger, A. B. (1995). Economic growth and the environment. *Quarterly Journal of Economics, 110(2):353–377.*

Harvey, D. (2005). A Brief History of Neoliberalism. Oxford University Press, Oxford.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2 edition.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science, 353(6301):790–794.*

Karakara, A. A. & Osabuohien, E. S. (2021). Inclusive Growth Agenda in Selected Sub-Saharan Africa Countries: Lessons from the Past and Prospects for the Future. *African Journal of Business and Economic Research, Special Issue (Perspective on Demographic Dividend, Economic Growth, and Sustainable Development) pp. 117 – 150. doi: https://doi.org/10.31920/1750-4562/2021/SIn1a6*

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School, 22(2):139–191.*

ND-GAIN (2024). Notre Dame Global Adaptation Initiative Country Index. https://gain.nd.edu/our-work/country-index/. Accessed: 2025-07-27.

North, D. C. (1990). Institutions, Institutional Change, and Economic Performance. Cambridge University Press, Cambridge.

Notestein, F. W. (1945). Population: The long view. pages 36–57.

Ogundari, K. and Abdulai, A. (2014). A meta-analysis of the impact of education on human development in sub-Saharan Africa. *Journal of Development Studies, 50(3):415–429.*

Omokanmi, O. J. (2020). Financial inclusion and human development in sub-Saharan Africa. *African Development Review, 32(4):567–579.*

Ranis, G., Stewart, F., and Ramirez, A. (2000). Economic growth and human development. *World Development, 28(2):197–219.*

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy, 98(5): S71–S102.*

Rostow,W.W. (1960). The Stages of Economic Growth: A Non-Communist Manifesto. Cambridge University Press, Cambridge.

Sen, A. (1999). Development as Freedom. Oxford University Press, Oxford.

Tobaigy, M. (2022). Predicting human development index using tree-based regression models: A global analysis. *Development Policy Review, 40(2): e12678.*

United Nations Development Programme (2020). Human development report 2020: The next frontier. Technical report, United Nations Development Programme, New York.

Wallerstein, I. (1974). The Modern World-System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York.

World Bank (2022). World development indicators. Technical report, World Bank, Washington, D.C.