Global Development Analysis:

A Data-Driven Look at Economic, Health, and Climate Progress (2000–2020)

---

1. Abstract

This report investigates the progression of countries between 2000 and 2020 using a wide range of development indicators, including economic, health, environmental, and infrastructural variables. We apply machine learning techniques for predictive modeling (XGBoost) and unsupervised learning (K-Means clustering). Our work reveals key drivers of human development, clusters countries with similar development trajectories, and highlights potential directions for policy design and future analysis.

---

2. Introduction

Understanding development trajectories across countries is essential for informed policymaking. This work presents a data-driven approach to track global development through multidimensional indicators and advanced machine learning methods. Specifically, we aim to:

- Predict Human Development Index (HDI) using interpretable ML.
- Segment countries into meaningful development groups.
- Explore correlations across economic, health, and environmental domains.

We build upon open-source development datasets and focus on interpretable results to aid policy-level decision making.

---

3. Dataset and Indicators

Source:
Global Development Indicators dataset (Kaggle, World Bank, WHO, UN, Climate databases)

Global Development Indicators (2000–2020)

Time Frame:
2000–2020 (Annual data)

Indicator Categories:

Category	Example Indicators
Economic	GDP per capita, Inflation, Unemployment
Health	Life expectancy, Mortality rate, Health expenditure
Environment & Energy	CO_2 emissions, Energy use per capita, Renewable energy share

Infrastructure & Access	Electricity access, Internet use, Mobile subscriptions
Education	School enrollment, Education-Health ratio
Composite & Indexes	HDI, Climate vulnerability, Governance index
Social & Demographic	Region, Income group, Pandemic flag
Temporal Features	Years since 2000, Years since the century

---

4. Data Preprocessing

4.1 Missing Values

- Dropped columns with >30% missing.
- Linear interpolation within the country.
- Median regional imputation for remaining missing values.

4.2 Outlier Detection and Handling

- Applied the **IQR method** to flag extreme values.
- Winsorized top/bottom 5% in numeric features.
- Preserved outliers to retain real-world variance.

4.3 Feature Tiers

- **Tier 2:** Raw indicators (GDP, CO_2, Life expectancy)
- **Tier 3:** Composite indices (HDI, Resilience, Ecological Index)

4.4 Scaling

- Applied **RobustScaler** to normalize features while minimizing outlier influence.

---

5. HDI Prediction using XGBoost

5.1 Why HDI?

HDI combines health, education, and income into a single composite score. Predicting it enables a better understanding of what drives human development.

5.2 Why XGBoost?

XGBoost was chosen for its optimization for tabular data, its ability to handle missing data well, and its support for feature importance and SHAP. It is also recognized for being robust and fast.

5.3 Model Setup

- Features: Selected Tier 2 and Tier 3 indicators
- Target: Human Development Index
- Tuning: RandomizedSearchCV with 5-fold cross-validation

Best Hyperparameters:

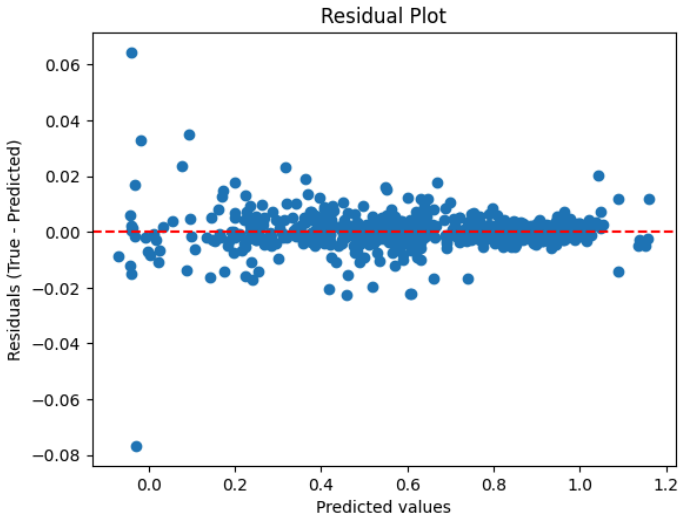
n_estimators=300, max_depth=7, learning_rate=0.1, subsample=0.7, colsample_bytree=1.0

5.4 Performance Metrics

Metric	Value
MAE	0.0033
RMSE	0.0062
R^2 Score	0.9995

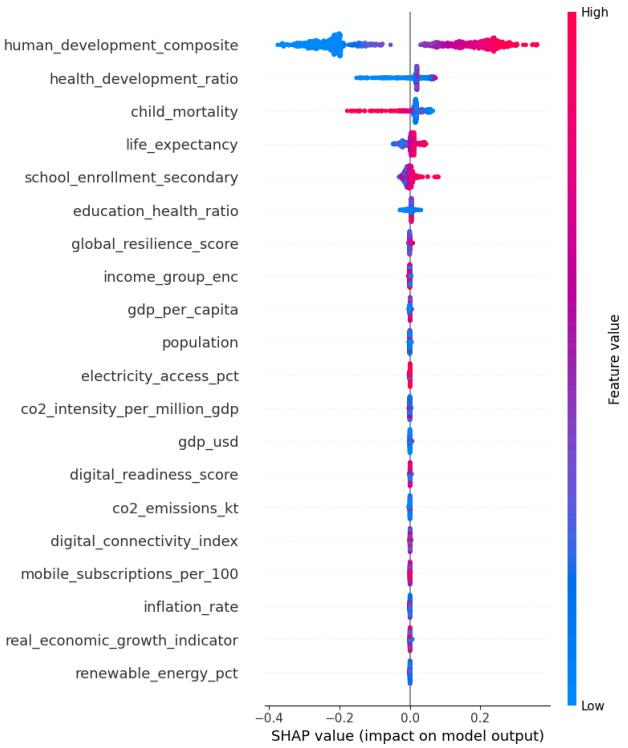
The model explains 99.95% of the variance in HDI — excellent predictive performance.

5.5 Residual Analysis



- Residuals are tightly distributed around zero.
- No evidence of heteroscedasticity or bias.
- The model generalizes well without overfitting.

5.6 SHAP Value Analysis



- Top predictive features include:
  - Life expectancy
  - GDP per capita
  - Education access
  - Renewable energy %
  - Digital readiness score

- SHAP enables understanding of **both global and local** prediction behavior.

---

6. Clustering Countries using K-Means

6.1 Objective

To identify development-based clusters of countries using a mix of raw and composite indicators.

6.2 Preprocessing Recap

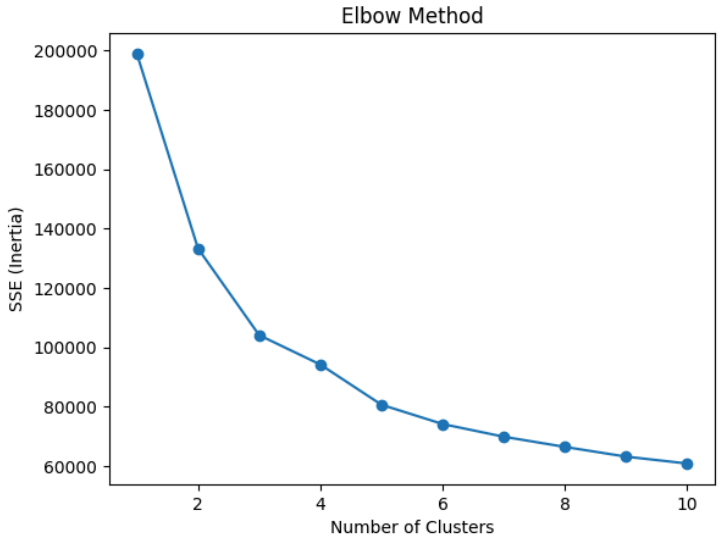
- Used the same cleaning as in regression.
- Applied Winsorization + RobustScaler.
- PCA is used for 2D visualization.

6.3 Feature Set

- Tier 2: GDP, emissions, child mortality, internet use, life expectancy
- Tier 3: HDI, Resilience Index, Ecological Index

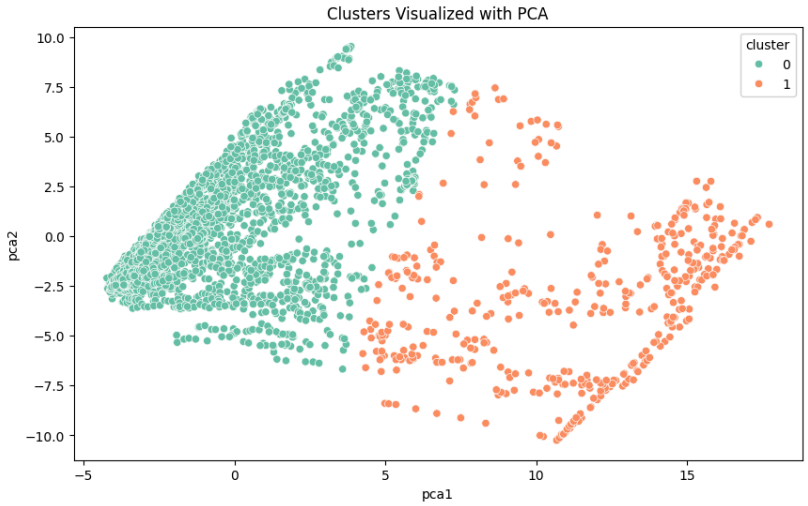
6.4 Cluster Evaluation

- **Elbow method** and **Silhouette scores** used to select k=2.



k	Silhouette Score
2	**0.5195**
3	0.3305
4	0.2973
5	0.2604

6.5 PCA Visualization



Clear separation between the two clusters. PCA preserves relative distances and aids interpretation.

6.6 Cluster Profiles

Cluster 0

- Mostly low-/middle-income countries
- Lower GDP, health access, and infrastructure

Cluster 1

- Advanced and emerging economies
- Higher HDI, stronger digital & health infrastructure

---

7. Insights & Discussion

- HDI is **highly predictable** using core indicators.
- SHAP reveals that **life expectancy and education** are among the most impactful.
- Clustering confirms two broad development pathways, enabling peer benchmarking.

---

8. Limitations

- Missing data may bias estimates in countries with sparse records.
- HDI itself may overlook key development nuances (e.g., inequality, environmental justice).
- Clustering can oversimplify diverse trajectories.

---

9. Future Work

- **Interpretability:** Use SHAP to explain clustering behavior.
- **Visualization:** Add radar charts for cluster medians.
- **Alternative Algorithms:** Test DBSCAN or Agglomerative Clustering.
- **Temporal Stability:** Extend to test cluster shifts and HDI predictability over time.
- **Model Generalization:** Validate XGBoost across continents and income groups.
- **Advanced Tuning:** Apply Bayesian Optimization for better XGBoost performance.
- **Interaction Effects:** Explore feature interactions via SHAP dependence plots.

---

10. References

1. UNDP Human Development Reports
2. World Bank Open Data
3. Kaggle Global Development Indicators Dataset
4. Chen, T., & Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System”