SIADS 696 Milestone II Project Report

Machine Learning for SDG6: Forecasting Safe Water Access

Aabir (aabir@umich.edu), Edith (etache@umich.edu), Kajal kdraut@umich.edu)

Introduction

In 2015, the UN launched the 2030 Agenda with 17 Sustainable Development Goals (SDGs) [1] to address global issues like poverty, inequality, and climate change. SDG 6 focuses on ensuring access to safe water and sanitation for all, recognizing its importance for health, dignity, and development. It includes targets such as universal access to clean drinking water, improved water quality, and protection of water ecosystems. While progress has been made, billions still lack safely managed water and sanitation services, making SDG 6 a critical priority.

This project aims to identify and predict regions at risk of water stress and poor sanitation using data-driven methods, helping uncover patterns and predictors of water-related inequality. The goal is to understand which factors contribute most to water insecurity, enabling targeted interventions.

While Brown et al. (2019) focus on predicting water supply using primarily population and climate variables, our model extends this approach by also incorporating socio-economic indicators. Similarly, Dolan et al. (2021) and Yao et al., (2023) assess the economic and electricity impacts of water scarcity respectively, whereas our model treats the economy and electricity as contributing factors to scarcity. [9][10] By integrating a broader range of variables, our approach aims to offer a more holistic view of the drivers of water availability.

Data Source, Scope and Preprocessing

The two data sources used for this project were UNSDG and the World Bank. The UNSDG data was downloaded in CSV format directly from the official website. In contrast, World Bank data was accessed using the WBGAPI Python library, which provides modern, Pythonic access to the World Bank's data API.

UNSDG Data

We utilized indicator-level data from the United Nations Sustainable Development Goals (SDG) Global Database, specifically focusing on:

- SDG 6.1: Population using safely managed and basic drinking water services
- SDG 7, 9, 11, and 13: Clean energy adoption, infrastructure development, sustainable cities, responsible consumption, and climate action

These indicators offer country-wise measurements aligned with UN-defined methodologies and are essential for evaluating multidimensional progress toward sustainability.

World Bank Open Data

This project utilizes country-level indicators from the World Bank Open Data platform to provide socio-economic, environmental, and infrastructure context for analyzing water access and sustainability. Key indicators cover areas such as GDP, electricity access, education, poverty levels, water usage, environmental stress, and sanitation investment. These variables support a comprehensive analysis of the drivers and disparities in access to safely managed water services (see **Appendix A** for exact attribute list).

The scope of the project was limited to the **latest 10 years of available data** from both sources, specifically from **2011 to 2020**, covering **217 countries worldwide**.

Data Preparation and Transformation

To ensure the reliability and relevance of our dataset, we applied the following data cleaning and transformation steps:

1. **Filter World Bank Dataset:** Selected only those features with data available for more than 100 countries, reducing the attribute list from 28 to 9 to retain globally representative indicators.
2. **Filter UNSDG Dataset:** The UN SDG dataset contains data by location. Filtered for *ALLAREA* to select only national-level data. For indicators with breakdowns by activity type, we selected *TOTAL* to capture overall performance. This yielded 15 SDG features relevant to the study.
3. **Dataset Merging:** Concatenated the filtered World Bank and UN SDG datasets to build a unified dataset for analysis.
4. **Reshaping and Formatting:** Transformed data from long format to wide format, where each row represents a country-year pair with all corresponding features as columns.
5. **Missing Values Treatment:**
    a. Dropped features with more than 60% missing values or those considered redundant due to overlap with other indicators (12 features dropped).
    b. Dropped 1 country/region with missing values across all features.
    c. For remaining missing values, applied mean imputation at the regional level, filling missing values with the average of all countries in the same region.

The resulting dataset, with 2170 records, provides a clean, structured, and regionally representative foundation for supervised and unsupervised machine learning tasks. A table of our final attributes with description can be found in the **Appendix B** below.

Additional Feature Engineering Steps for Supervised Learning

The feature engineering pipeline for our supervised learning task included:

1. **1 year lag of the socioeconomic features -** The relationship between changes in our features and the target (Proportion of population using safely managed drinking water services) are often not instantaneous in a real-world setting. For example, the effect of improving electricity access in 2011 might not be observed until 2012. To account for this, we engineered and used only a lagged version of our independent variables for supervised learning.
2. **Lagged target as model input -** We explore both the static and dynamic approaches to modelling panel data. The dynamic approach incorporates the lagged target as a model input. The advantages and drawbacks of these two different approaches will be explored in detail in the following sections.

Supervised Learning

Our goal is to build a supervised machine learning model to forecast the proportion of the population using safely managed drinking water services in future years. Our dataset combines both cross-sectional data (observations on multiple countries) and time-series data (observations on multiple years). This type of dataset structure is referred to as panel data. Forecasting using panel data is often used in the field of economics to forecast indicators such as GDP. [2]

There are two common modelling approaches when using panel data:

1. Forecasting with static panel data: This is the basic approach where the model is trained on historical data and tested on future data. This approach only considers the independent variables as inputs into the model. [3]
2. Forecasting with dynamic panel data: In some instances, the target variable you wish to forecast may be highly correlated with its past value. For example, GDP in 2025 is highly correlated with the GDP value in 2024. Capturing the autocorrelation between observations of the same variable at different points in time can be done using dynamic panel data modelling, which combines the basic model with a dynamic component using the lagged target variable as an additional input into the model. [2]

For our supervised learning exploration, we will develop models using both the static and dynamic approach and compare the strengths and drawbacks of each.

Methods and Evaluation

Our prediction problem is a regression task. As part of our first steps, we investigated the performance of three different machine learning models that utilize different algorithmic approaches:

1. Linear Model: Lasso Regression - Selected over OLS to better handle multicollinearity
2. Tree-Based Model: Random Forest - Better for handling nonlinear relationships and feature interactions
3. Nearest Neighbor: KNN Regression - Baseline model that requires no assumptions about the data (ie. works for both linear and nonlinear problems)

In order to avoid data leakage issues that can arise when future data is used to predict past values, for each model, the training set was developed using the observations that occurred on or before 2018, while the test set consisted of all observations after 2018. For model training, a traditional k-fold cross validation would also introduce the same data leakage concerns. To mitigate this issue, we created scikit-learn's TimeSeriesSplit class object for proper handling of train/test splits for time-ordered data.[4] For our analysis, we used 5 fold cross validation for our TimeSeriesSplit. Lastly, as part of our feature selection for preprocessing, we dropped features that showed no correlation with our target variable. The correlation matrix can be found in the unsupervised learning section below.

To measure our model performance, we used Mean Absolute Error (MAE) to determine the size of the errors between our actual value for safe drinking water access (%) and the predicted value determined by the model. Our goal is to develop a model that minimizes the MAE so that we can effectively forecast safe drinking water access for future years. Figure 1 below summarizes the performance of our three dynamic panel data models using just the default hyperparameters:

Model	Average MAE	Standard Deviation (+/-)
RandomForest	0.5004	0.0867
Lasso	0.9556	0.0843
KNN	2.6247	0.6556

Figure 1: Model performance using 5 fold cross validation using dynamic panel data modelling assumptions

Even without hypertuning, both the lasso and random forest models are able to forecast very well, yielding a less than 1% mean average error. This is likely due to the influence of our lagged target on the model, which we will explore further in the feature importance section when we apply the static modelling approach. Unlike the lasso and random forest models, the KNN regression model performs significantly worse, with a ~2.6% mean average error. Given this, we stopped pursuing the KNN regression model as a viable option. To further boost model performance, we applied the following hypertuning strategies using a grid search:

Model	Hyperparameter	Description	Set
Lasso	alpha	Constant that multiplies the L1 term [5]	[0.1, 1.0, 5.0, 10.0]
RandomForest	n_estimators	# of trees [6]	[50, 100, 200]
RandomForest	max_depth	Max depth of trees [6]	[5 ,10 , 15, 20, 25]
RandomForest	min_samples_split	Min number of samples required to split an internal node [6]	[2, 4]
RandomForest	min_samples_leaf	Min number of samples required to be at a leaf node [6]	[1, 2]

Figure 2: Hyperparameter tuning summary

Figure 3 below summarizes the performance of the two models after applying the optimal hyperparameters (scoring for MAE) as determined by the grid search:

Model	Average MAE	Standard Deviation (+/-)
RandomForest	0.4977	0.0826
Lasso	0.3817	0.0893

Figure 3: Model performance using 5 fold cross validation using dynamic panel data modelling assumptions

After hypertuning, we see improvement in both models, with the lasso regression model performing the best (~0.4% mean average error vs ~0.5 mean average error) but overall, both models yield a near perfect ability to forecast. For the remainder of this exploration, we conducted deeper analysis into our lasso regression model. Figure 4 below shows how well our lasso model was able to generalize to our unseen test data set. As a secondary evaluation metric, we computed the R^2 score to understand the proportion of variance in the target variable explained by the model.

Model	Test MAE	Test R^2 score
Lasso	0.2925	0.9996

Figure 4: Model performance on our test set (years > 2018)

When it comes to the ability to forecast safe drinking water access, a dynamic approach appears to be yielding near perfect prediction capability on our test set.

Training Curve Analysis

Before diving into our lasso regression model, we first conducted a training data curve analysis to determine how much data is needed for optimal performance and whether the model can generalize well. Figure 5 below shows how our model performs using varying training sizes, starting with the 434 records on or before 2013, to 1736 records on or before 2019. Figure 5 reveals substantial improvement in mean average error on the test set when training using the 1519 records on or before 2018. The lack of a plateau however suggests that the model hasn't saturated with our limited dataset and more data would likely further improve its ability to generalize on unseen data.

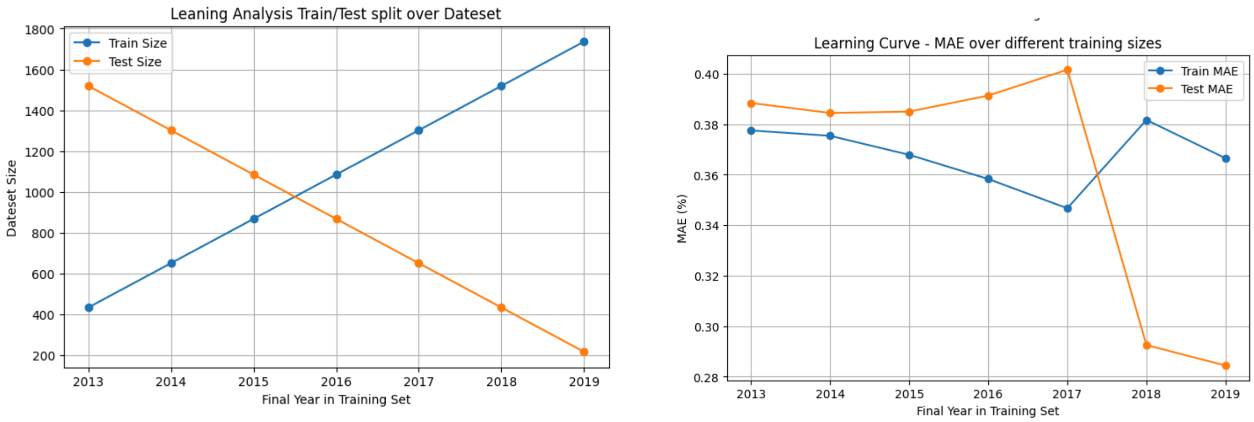


Figure 5: Mean absolute error over varying training sizes

Feature Importance and Ablation Tests

Figure 6 below shows the feature importance of our dynamic model in the form of a SHAP beeswarm plot. Unsurprisingly, the lagged target dominates feature importance in our model, suggesting that although the lasso model is able to forecast safe drinking water access with incredibly high accuracy, it exclusively relies on the lagged target to make predictions and fails to learn from the temporal patterns of the other features.

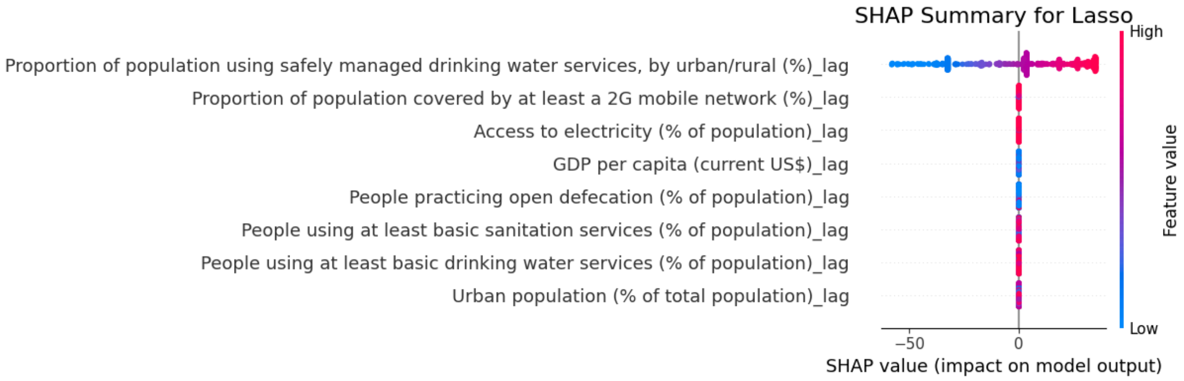


Figure 6: SHAP beeswarm plot for feature importance of dynamic model

The effect of the other features can be better understood using a static approach to modelling panel data. We modelled this effect by conducting a feature ablation analysis where we removed the lagged target variable prior to modelling. The table below summarizes the effects of this ablation on model performance using the same model, hypertuning and 5 fold cross validation.

Approach	Model	Ablated Feature	Avg. MAE	SD	Test MAE
Dynamic	Lasso	None	0.3817	0.0893	0.2925
Static	Lasso	Proportion of population using safely managed drinking water services, by urban/rural (%)_lag	12.06	0.4359	11.62

Ablating the lagged target drastically degrades model performance from an average mean absolute error of ~0.38% to ~12.06%. Figure 7 shows the feature importance of our static model.

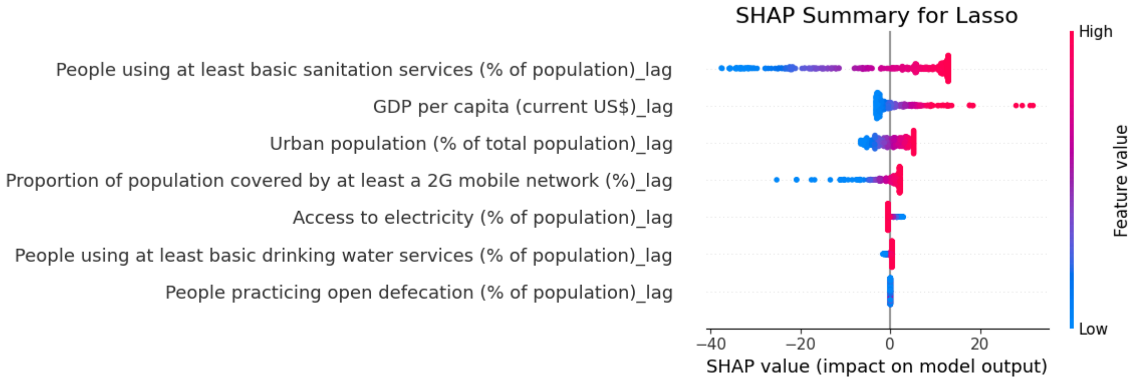


Figure 7: SHAP beeswarm plot for feature importance of static model

By excluding our lagged target, the model has to learn the temporal patterns of our features to make a prediction. From the SHAP plot, we observe that access to basic sanitation services and GDP per capita are the top two most important features for our static model. Higher values of sanitation services and GDP increase the model's predicted target outcome, suggesting that targeted improvements in those features are positively associated with greater access to safe drinking water services.

Sensitivity analysis

We want to assess model sensitivity to hyperparameter tuning for both our static and dynamic models. For both models, hyperparameter tuning was done using a range of alpha parameters for the lasso regression model. Model sensitivity is captured in figure 8 below.

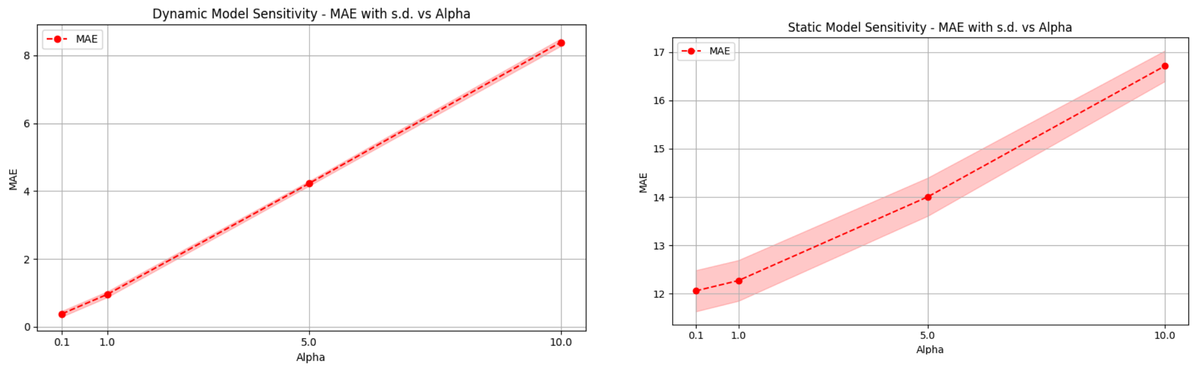


Figure 8: Model sensitivity to tuning alpha parameter

For both approaches, we observed that an alpha value closer to zero resulted in a lower mean absolute error during training. We primarily elected to use a lasso regression model in order to handle the multicollinearity we observed between our independent features, but our findings in our hypersensitivity analysis suggest that the L1 regularization used is oversimplifying the model at higher values of alpha, leading to the observed degradation in mean absolute error.

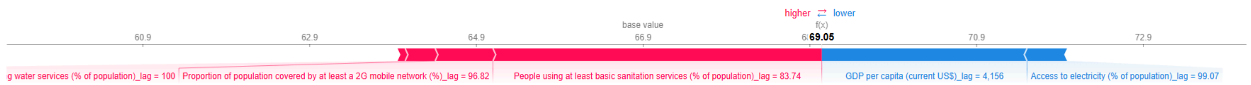
Failure Analysis

Given the poor performance of our static model, a failure analysis was conducted to better understand where this model is lacking. As part of this analysis, we chose 3 distinct instances where our static model made incorrect predictions.

Instance #1:

Year	Country	True Value	Prediction	Residual
2019	Tuvalu	8.5833	69.0512	-60.4679

This instance corresponds to the largest prediction error and overestimates the true value by 60.47%. Upon closer examination, all Tuvalu records in our dataset exhibited large prediction errors. The SHAP force plot below helps visualize what drives this prediction.

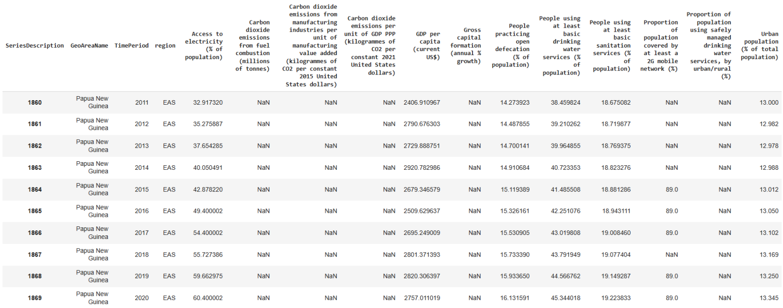


The force plot indicates that access to basic sanitation services is a key driver, significantly contributing to the increase in the predicted value. After exploring the dataset, the Tuvalu records appear to be an edge case in our dataset, where the country reported a high percentage of population with access to basic sanitation services while also reporting a low proportion of the population with safely managed drinking water services. This suggests that our static model likely omits key features that contribute to low access to safe drinking water in outlier countries like Tuvalu, thus failing to generalize well to edge cases such as this.

Instance #2:

Year	Country	True Value	Prediction	Residual
2019	Papua New Guinea	67.2782	22.7013	44.5769

This instance corresponds to the largest underestimation by our model. Upon closer examination of the Papua New Guinea records in our dataset, we discovered that all values for the target variable were missing in the raw data (Figure 9). Our region-based mean imputation strategy was an inappropriate technique to patch over the missing data for this record and contributed to the large observed error.



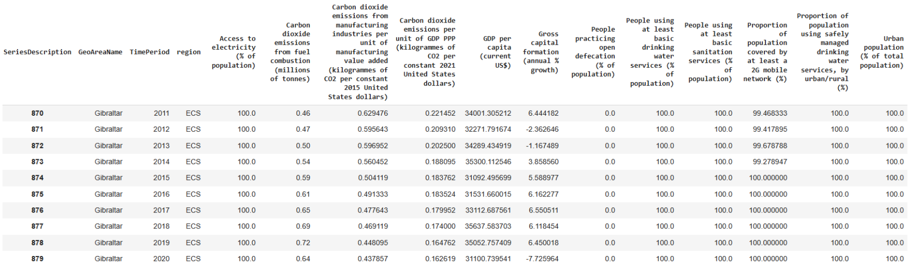
SeriesDescription	GeoAreaName	TimePeriod	region	Access to electricity (% of population)	Carbon dioxide emissions from fuel combustion (millions of tonnes)	Carbon dioxide emissions from manufacturing industries per unit of manufacturing value added (kilogrammes of CO2 per constant 2015 United States dollars)	Carbon dioxide emissions per unit of GDP PPP (kilogrammes of CO2 per constant 2021 United States dollars)	GDP per capita (current US$)	Gross capital formation (annual % growth)	People practicing open defecation (% of population)	People using at least basic drinking water services (% of population)	People using at least basic sanitation services (% of population)	Proportion of population covered by at least a 2G mobile network (%)	Proportion of population using safely managed drinking water services, by urban/rural (%)	Urban population (% of total population)
1860	Papua New Guinea	2011	EAS	32.917320	NaN	NaN	NaN	2406.910967	NaN	14.273923	38.459824	18.675082	NaN	NaN	13.000
1861	Papua New Guinea	2012	EAS	35.275887	NaN	NaN	NaN	2790.676303	NaN	14.487855	39.210262	18.719877	NaN	NaN	12.982
1862	Papua New Guinea	2013	EAS	37.654285	NaN	NaN	NaN	2729.888751	NaN	14.700141	39.964855	18.769375	NaN	NaN	12.978
1863	Papua New Guinea	2014	EAS	40.050491	NaN	NaN	NaN	2920.782986	NaN	14.910684	40.723353	18.823276	NaN	NaN	12.988
1864	Papua New Guinea	2015	EAS	42.878220	NaN	NaN	NaN	2679.346579	NaN	15.119389	41.485508	18.881286	89.0	NaN	13.012
1865	Papua New Guinea	2016	EAS	49.400002	NaN	NaN	NaN	2509.629637	NaN	15.326161	42.251076	18.943111	89.0	NaN	13.050
1866	Papua New Guinea	2017	EAS	54.400002	NaN	NaN	NaN	2695.249009	NaN	15.530905	43.019808	19.008460	89.0	NaN	13.102
1867	Papua New Guinea	2018	EAS	55.727386	NaN	NaN	NaN	2801.371393	NaN	15.733390	43.791949	19.077404	NaN	NaN	13.169
1868	Papua New Guinea	2019	EAS	59.662975	NaN	NaN	NaN	2820.306397	NaN	15.933650	44.566762	19.149287	89.0	NaN	13.250
1869	Papua New Guinea	2020	EAS	60.400002	NaN	NaN	NaN	2757.011019	NaN	16.131591	45.344018	19.223833	89.0	NaN	13.345

Figure 9: Raw Papua New Guinea data prior to region-based mean imputation

Instance #3

Year	Country	True Value	Prediction	Residual
2020	Gibraltar	100	90.1960	9.8041

This instance corresponds to a situation where our model underpredicts the true value by close to 10%. Upon a closer look at the cleaned data for the Gibraltar records (Figure 10), we noticed that the Gibraltar data showed no variation in many features over the 10 years span in our dataset, including our target feature.



SeriesDescription	GeoAreaName	TimePeriod	region	Access to electricity (% of population)	Carbon dioxide emissions from fuel combustion (millions of tonnes)	Carbon dioxide emissions from manufacturing industries per unit of manufacturing value added (kilogrammes of CO2 per constant 2015 United States dollars)	Carbon dioxide emissions per unit of GDP PPP (kilogrammes of CO2 per constant 2021 United States dollars)	GDP per capita (current US$)	Gross capital formation (annual % growth)	People practicing open defecation (% of population)	People using at least basic drinking water services (% of population)	People using at least basic sanitation services (% of population)	Proportion of population covered by at least a 2G mobile network (%)	Proportion of population using safely managed drinking water services, by urban/rural (%)	Urban population (% of total population)
870	Gibraltar	2011	ECS	100.0	0.46	0.629476	0.221452	34001.305212	6.444182	0.0	100.0	100.0	99.468333	100.0	100.0
871	Gibraltar	2012	ECS	100.0	0.47	0.595643	0.209310	32271.791674	-2.362646	0.0	100.0	100.0	99.417895	100.0	100.0
872	Gibraltar	2013	ECS	100.0	0.50	0.596952	0.202500	34289.434919	-1.167489	0.0	100.0	100.0	99.678788	100.0	100.0
873	Gibraltar	2014	ECS	100.0	0.54	0.560452	0.188095	35300.112546	3.858560	0.0	100.0	100.0	99.278947	100.0	100.0
874	Gibraltar	2015	ECS	100.0	0.59	0.504119	0.183762	31092.495699	5.588977	0.0	100.0	100.0	100.000000	100.0	100.0
875	Gibraltar	2016	ECS	100.0	0.61	0.491333	0.183524	31531.660015	6.162277	0.0	100.0	100.0	100.000000	100.0	100.0
876	Gibraltar	2017	ECS	100.0	0.65	0.477643	0.179952	33112.687561	6.550511	0.0	100.0	100.0	100.000000	100.0	100.0
877	Gibraltar	2018	ECS	100.0	0.69	0.469119	0.174000	35637.583703	6.118454	0.0	100.0	100.0	100.000000	100.0	100.0
878	Gibraltar	2019	ECS	100.0	0.72	0.448095	0.164762	35052.757409	6.450018	0.0	100.0	100.0	100.000000	100.0	100.0
879	Gibraltar	2020	ECS	100.0	0.64	0.437857	0.162619	31100.739541	-7.725964	0.0	100.0	100.0	100.000000	100.0	100.0

Figure 10: Gibraltar cleaned data

The lack of variation over time, especially for the target variable, introduces a form of systematic error as there is nothing meaningful to learn. Machine learning requires variation in the features in order to learn the relationships between them to make predictions, and cases where there is no variation will result in a degradation of model performance.

Given our findings, the following improvements can be explored to try and improve the prediction capability of our static model:

1. Model Adjustments: Tree-based modelling approaches such as random forest are more robust to outliers. In addition, unlike a lasso regression model, they are able to handle non-linear relationships. Trying the static approach using a tree based model may lead to better model performance.
2. Data Cleaning: Instead of region based imputation, dropping countries with more than 60% missing data for the target should also improve model accuracy. Dropping countries that show no variation in the target feature over time would also lead to improved accuracy.
3. Model Training: Our current dataset has a limited sample size and a small number of features, which has resulted in an overly simple model that fails to generalize well. Periodically updating the model with new data and additional features could enhance its performance.

Tradeoffs

The tradeoff between the static and dynamic approaches explored in our analysis depends on the specific goal for the model. If the goal is to build a model that achieves high accuracy (low MAE) in its ability to forecast safe drinking water access, then a dynamic approach which leverages the lagged target as a model input yields near perfect forecasting ability. This comes at a cost however, as the dynamic model fails to provide any deep insight into what features drive changes in safe drinking water access, instead learning exclusively from the lagged target to make accurate predictions due to its autocorrelation with the target. For more meaningful policy discussion, a static modelling approach provides better insights into the associations between key socioeconomic features and access to safe drinking water. This comes at the cost of a degradation in the models' ability to forecast for our limited data set. It should be noted that advanced techniques do exist to handle the endogeneity problem in dynamic modelling of panel data, which we touch on in the supervised learning discussion section.

Unsupervised Learning

An unsupervised exploration is valuable. Governments and development agencies rarely have the resources to design one-off strategies for every country; instead, they look for "archetypes" that can guide bundled interventions. The unsupervised phase translates a noisy spreadsheet of numbers into a digestible set of country archetypes, giving policymakers a coherent framework for prioritizing funds, benchmarking progress, and tailoring future supervised forecasts to the realities of each group.

Methods Description

We began our analysis with a comprehensive correlation audit (Figure 11)to assess interdependencies among developmental indicators. We rescaled all numerical indicators to a standardized [0,1] interval to ensure comparability, thereby neutralizing unit-derived biases.

The Pearson correlation analysis revealed two dominant structural patterns:

- A tightly coupled positive block comprising basic drinking water services, basic sanitation services, and electrification variables, with near-perfect pairwise correlations:
    - Basic drinking-water services ↔ basic sanitation services (*r* = 0.90)
    - Access to electricity ↔ basic drinking water services (*r* = 0.90)
    - Access to electricity ↔ basic sanitation services (*r* = 0.89
- A strong negative association between open defecation practices and electrification (*r* ≈ −0.71).

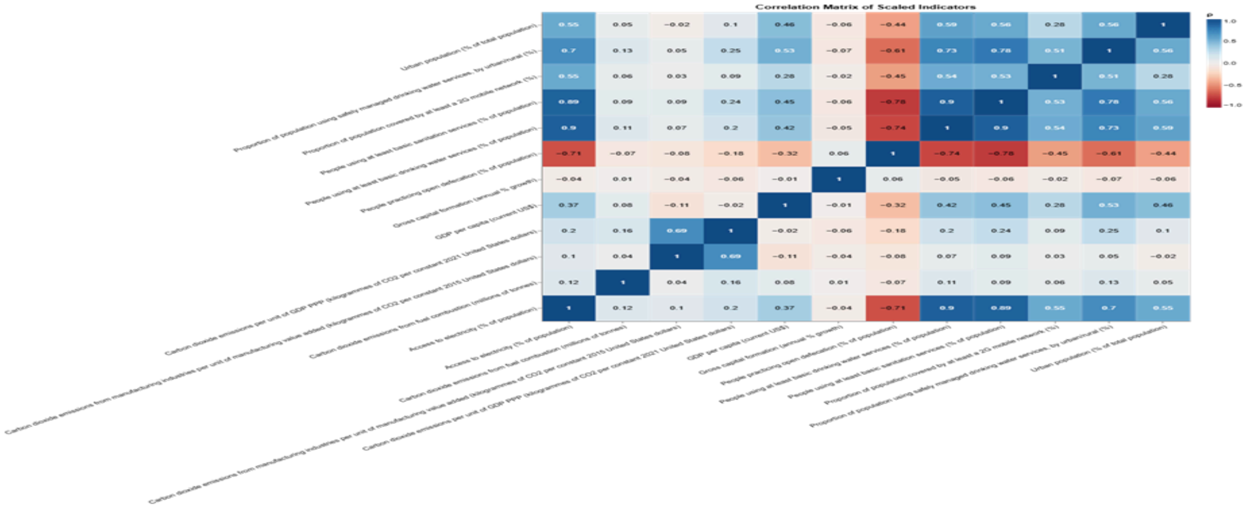


Figure 11: Correlation Matrix of Scaled Indicators

The correlation analysis revealed strong linear dependencies among several indicators. When many variables move together, Euclidean-distance measures such as those used by K-Means can overstate similarity along the redundant directions and mask structure in the rest of the space. To avoid this distortion, we opted for dimensionality reduction, which would rotate the data into orthogonal, uncorrelated axes.

We employed Principal Component Analysis (PCA) because it satisfies three practical requirements: Linear transparency – each component is an explicit weighted blend of the original variables, so we can inspect loadings to understand what the axis represents; Decorrelation – the transformation produces orthogonal (uncorrelated) components, eliminating the multicollinearity that can skew distance-based models; Variance preservation – the cumulative-variance curve (Figure 12) shows that the first five components capture 83 % of total variance, comfortably above our 80 % retention threshold and suggesting that the discarded dimensions are largely noise. Accordingly, we retained the first five PCs and conducted all subsequent distance-based analyses in this reduced orthogonal space.

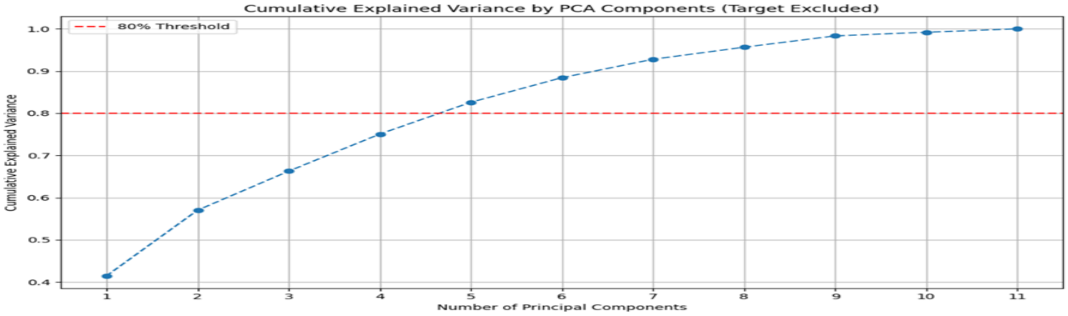


Figure 12: Cumulative Explained Variance by PCA Components

Principal-component loadings lend each axis an intuitive theme: PC 1 (“Water-Access & Infrastructure”) bundles basic drinking water, sanitation, and electrification; PC 2 (“Industrial Carbon Intensity”) tracks environmental pressure; PCs 3–5 capture, respectively, growth-versus-emissions tension, energy-investment trade-offs, and an urban-wealth gradient. Thus, PCA functions both as an unsupervised model that reveals latent structure and as a preprocessing step that neutralizes multicollinearity before clustering.

For the clustering stage, we selected K-Means and ran it on the five-dimensional PCA space. The sole hyper-parameter of interest - the number of clusters k, was scanned systematically from two through ten, with ten random initializations at each setting. Silhouette coefficients rose monotonically to 0.495 at k=4 and declined sharply thereafter, identifying four clusters as the most coherent partition (Figure 13).

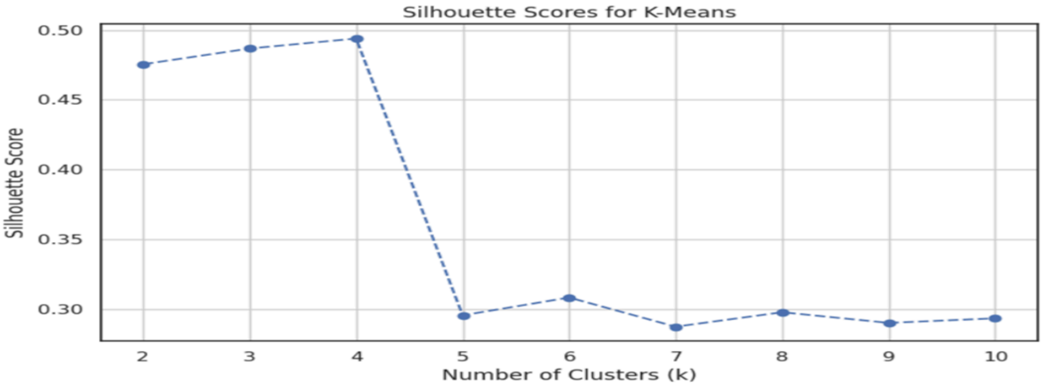


Figure 13: Silhouette Scores for K-Means

Why K-Means? In an orthogonal feature space, exactly what PCA supplies, centroid distance behaves reliably, convergence is fast, and the silhouette metric offers a clear, industry-standard quality check.

Unsupervised Evaluation

The five-component PCA embedding explains 83 % of the variance and gives us an orthogonal space where Euclidean distances are meaningful. A silhouette sweep over *k* = 2–10 shows that scores climb from about 0.47 at *k* = 2 to a peak of ≈ 0.495 at *k* = 4, then drop sharply—falling to ≈ 0.31 at *k* = 6 and settling just under 0.30 for *k* ≥ 7. A silhouette score just under 0.5 means that, on average, each country is roughly twice as close to the center of its own cluster as to the nearest competing cluster. That provides statistical confirmation that the four-cluster solution is both cohesive and well-separated in this PCA space.

Four well-separated clusters emerge when the data are plotted in the space of the first two principal components (Figure 14). On the far left, a red band marks countries as being in an “Early infrastructure, low-water-access” stage. A dense blue cloud in the lower-right quadrant represents “Developed, high-water-access” nations. Two points sit apart from these main groups: green, whose high-water access is coupled with exceptionally heavy industrial emissions, and purple, whose carbon-intensive profile lifts it to the top of the industrial-carbon axis.

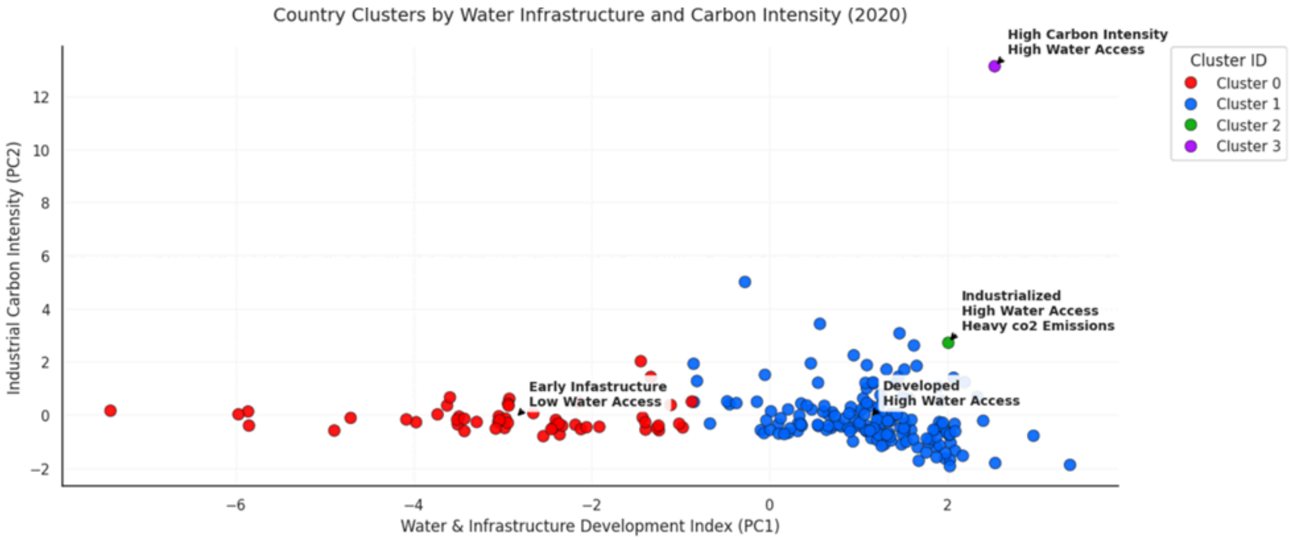


Figure 14: Country Clusters by Water Infrastructure and Carbon Intensity (2020)

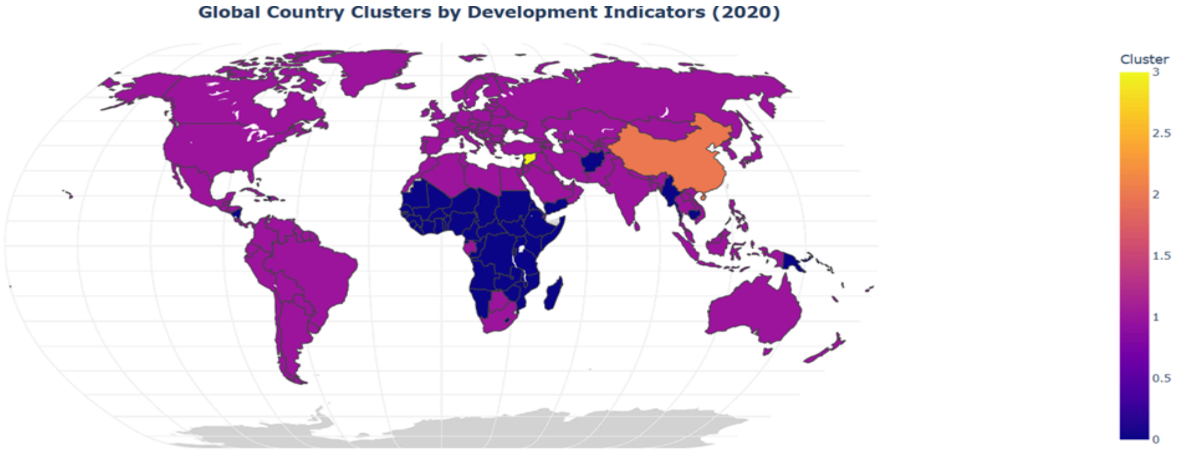


Figure 15: Global Country Clusters by Development Indicators (2020)

Mapping the same labels back to geography (Figure 15) reveals an intuitively plausible pattern: the infrastructure gap concentrates in sub-Saharan Africa and part of south Asia, the developed tier spans North America, Europe and much of Latin America, and the two outliers stand alone, China and Syria. The tight alignment between numerical quality (silhouette), planar separation, and geopolitical logic suggests that both the PCA feature set and the four-cluster assignment capture genuine structure rather than algorithmic artifact.

Sensitivity analysis:

To gauge how sensitive the solution is to its only hyper-parameter, *k*, we swept *k* = 2–10 and plotted the silhouette scores (Figure 13). The curve climbs gently from *k* = 2 (≈ 0.48) to a peak at *k* = 4 (≈ 0.495). Compressing four clusters into three trims the score by only ~0.006, but splitting one cluster to reach *k* = 5 slashes it by ~0.20. Beyond five clusters, the scores hover in a narrow 0.29–0.31 band, showing no further benefit from finer partitioning. This sharp one-step drop and the flat plateau on either side indicate that the four-cluster configuration is a well-resolved local optimum: even minor deviations in *k* erode cohesion far more than they improve separation. Because the PCA embedding is fixed throughout the sweep, we conclude that the underlying structure is stable across reasonable granularities while still flagging clear signs of over- or under-partitioning.

Discussion

Supervised Learning

The supervised learning phase offered several insights about modelling panel data.

a. What was most surprising?

Although we strongly suspected that the dynamic modelling approach would perform well, we did not expect that the presence of the lagged target would yield a near perfect ability to forecast safe drinking water service access and completely dominate the feature importance. This confirms that past values of safe drinking water access are highly correlated with what its value will be in the future. In hindsight, this was a logical finding as safe drinking water access is often a metric that changes very little over time. By adding the lagged target as an input, our model is able to quickly learn that the past value is a very strong predictor of what the future value for the target should be.

Although the dynamic approach works incredibly well in our dataset as a forecasting tool, it relies only on the lagged target to make a prediction and ignores the other features. It fails to provide any insights on the underlying drivers that most influence safe drinking water service access, which exposes a big weakness with using this approach and limiting its usefulness for policy discussion.

b. Key challenges and how they were handled

The key challenge with our supervised learning, especially for the static approach, was the lack of data available from the UN and World Bank databases. We handled this by dropping features with a high number of missing data, and using regional based mean imputation techniques to handle missing data.

This strategy introduces bias in our dataset. To handle this for supervised learning, we explored approaches that penalize noise/irrelevant features (Lasso instead of OLS for a linear approach), as well as algorithms that can handle non-linear data with outliers (Random Forest), with varying results.

c. Possible extensions with additional time or resources

1. **Adding additional features:** The limited number of features in our dataset is a key limiting factor in our static model, which is why it struggles to generalize well on unseen data. With more time and resources, we would focus on increasing the number of features in order to develop a more robust model.
2. **Exploring more advanced techniques for static modelling:** With a more robust dataset, we would explore other modelling approaches that are better suited to modeling non-linear relationships, handling multicollinearity, and handling outliers in the data. This would include tree-based models that incorporate gradient boosting methods.
3. **Exploring more advanced techniques for dynamic modelling:** Modelling with dynamic panel data is a deeply researched field, thus there are a lot of existing methods designed specifically to handle the endogeneity that arises with using the lagged target as an input. Given more time, we

would explore modelling using these approaches, such as the Arellano-Bond method, which uses an instrumental variable strategy to correct for biases introduced by the lagged target. [11]

Unsupervised Learning

The unsupervised phase offered several analytical and practical insights into the structure of global development indicators.

a. What was most surprising?

The dominance of basic infrastructure variables in the first principal component was striking. GDP per capita, which is often treated as a shorthand for development, is hardly registered in the leading axes; instead, access to electricity, basic drinking water services, and basic sanitation services determine the primary gradient in the data. Equally unexpected was the emergence of two single-country clusters: China and Syria. Their isolation shows that a few nations can deviate so strongly, whether through exceptional industrial emissions (China) or extreme carbon intensity amid conflict-damaged infrastructure (Syria) that they form their own statistical archetypes.

b. Key challenges and how they were handled

Multicollinearity was the first obstacle. The near-perfect correlations among basic drinking-water, sanitation, and electrification indicators meant that raw Euclidean distances would over-emphasize a single development dimension. We addressed this with Principal Component Analysis, which rotates the data onto orthogonal axes and captures 83 % of the total variance in the first five components.

Choosing the cluster count posed a second challenge. Rather than accept an arbitrary k, We scanned k=2–10 and used the silhouette curve as an objective yardstick. The unmistakable peak at k=4 and the rapid deterioration on either side gave a robust, data-driven rationale for the final partition.

c. Possible extensions with additional time or resources

1. **Temporal clustering:** Treat each country's 2010-to-2020 indicator series as a trajectory and cluster those paths with dynamic time warping. That would reveal development pathways rather than static snapshots.
2. **Mixed-type distance functions:** Incorporate categorical governance indicators (e.g., regime type, water-pricing policy) and switch to a Gower-distance k-medoids framework to accommodate both numeric and categorical features.
3. **Explainable boundaries:** Train a shallow decision tree in the PCA space so that each cluster can be distilled into a few easy "IF → THEN" rules - far more digestible for policy audiences than abstract centroid coordinates.

Ethical Considerations

To address missing values in the combined dataset, we applied a **region-wise mean imputation** technique. This method fills gaps using average values from the same geographic region, helping preserve socio-economic and geographic patterns more effectively than global means. However, this data imputation technique can carry ethical considerations, especially when applied in global development contexts.

Risk of Reinforcing Regional Stereotypes

Region-wise mean imputation assumes that countries within a region are similar, but this can mask important differences - particularly for outliers or marginalized populations. For example, a country with significantly worse access to water than its regional average may appear better off due to imputation, which can misinform policy. Some countries or regions may systematically lack data due to conflict, poverty, or weak infrastructure. Imputing their values from regional means can erase their unique challenges.

A model-based imputation technique that takes multiple variables into account, like economic similarities, not just region can be used to solve the above concerns. In our case, we ensured the overall correlation between features and target variables did not change much with the region based imputation technique.

A major limitation of this analysis is that many of the data points are survey-based, estimated, or imputed, rather than directly observed or measured. While such data is often the best available for global, cross-country comparisons, it raises below ethical concerns:

Risk of Misrepresentation

Estimated or modeled values can create a false sense of precision or accuracy, especially when used in high-stakes decisions or policy discussions. Low-income or politically unstable countries often lack reliable data collection infrastructure, leading to greater dependence on imputation or outdated surveys. This creates a data equity problem, where countries with the least capacity to provide accurate data may also be the most vulnerable, yet receive less accurate representation in the analysis.

Given that no single authoritative source provides fully concrete and consistently measured values across all countries and indicators, we have relied on the best-available data from trusted providers, with transparent documentation of sources and limitations throughout the analysis.

Statement of Work

Aabir	Edith	Kajal
Feature engineering, Random Forest, KNN and Lasso Regression Training, Learning curve analysis, Ablation and Failure analysis, report writing	PCA dimensionality reduction, KMeans Clustering Analysis, Sensitivity analysis, report writing	Data collection and preprocessing, Data analysis, Feature engineering, Ethical issues research, report writing

References:

[1] Nations, U. Sustainable Development Goals: 17 Goals to Transform Our World. https://www.un.org/sustainabledevelopment/ (2015).

[2] APTECH. Introduction to the Fundamentals of Panel Data. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-panel-data/#:~:text=What%20Is%20Panel%20Data?,%2C%20T

[3] Science Direct. Panel Data Models. https://www.sciencedirect.com/topics/social-sciences/panel-data-model#:~:text=The%20static%20panel%20data%20model,explanatory%20variables%20or%20instrument%20variable.

[4] Scikit Learn. Time Series cross-validator. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

[5] Scikit Learn. Lasso. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[6] Scikit Learn. RandomForestRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[7] UNSTATS. SDG Indicators. https://unstats.un.org/sdgs/metadata/

[8] World Bank Group. Sustainable Development Goals. https://databank.worldbank.org/source/sustainable-development-goals-(sdgs)#

[9] Dolan, F., Lamontagne, J., Link, R. *et al.* Evaluating the economic impact of water scarcity in a changing world. *Nat Commun* 12, 1915 (2021). https://doi.org/10.1038/s41467-021-22194-0

[10] Yao An and Lin Zhang, 2023. "The Thirst for Power: The Impacts of Water Availability on Electricity Generation in China," The Energy Journal, International Association for Energy Economics, vol. 0(Number 2). https://www.iaee.org/ej/ejexec/ej44-2-Zhang-exsum.pdf

[11] Sarah Lee. A Intro to the Arellano-Bond Estimation Method. https://www.numberanalytics.com/blog/intro-arellano-bond-estimation-method (2025)

Appendix A - Data Sources

UNSDG Data Attributes

1. Proportion of population using safely managed drinking water services, by urban/rural (%)
2. Proportion of the rural population who live within 2 km of an all-season road
3. Carbon dioxide emissions from fuel combustion (millions of tonnes)
4. Carbon dioxide emissions per unit of GDP PPP (kg of CO_2 per constant 2021 US$)
5. Carbon dioxide emissions from manufacturing industries per unit of manufacturing value added (kg of CO_2 per constant 2015 US$)
6. Research and development expenditure as a proportion of GDP (%)
7. Proportion of population covered by at least a 2G mobile network (%)
8. Proportion of population covered by at least a 3G mobile network (%)
9. Proportion of population covered by at least a 4G mobile network (%)
10. Proportion of population covered by at least a 5G mobile network (%)
11. Proportion of urban population living in inadequate housing (%)
12. Proportion of urban population living in slums (%)
13. Municipal solid waste collection coverage, by cities (%)
14. Total greenhouse gas emissions without LULUCF for Annex I Parties (Mt CO_2 equivalent)
15. Total greenhouse gas emissions without LULUCF for non-Annex I Parties (Mt CO_2 equivalent)

World Bank Data Attributes

1. GDP per capita (current US$)
2. Access to electricity (% of population)
3. Urban population (% of total population)
4. Educational attainment, at least completed upper secondary, population 25+, total (%) (cumulative)
5. Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)
6. Poverty headcount ratio at national poverty lines (% of population)
7. Rural poverty headcount ratio at national poverty lines (% of rural population)
8. Urban poverty headcount ratio at national poverty lines (% of urban population)
9. Literacy rate, youth total (% of people ages 15–24)
10. Literacy rate, adult total (% of people ages 15 and above)
11. Level of water stress: freshwater withdrawal as a proportion of available freshwater resources
12. Water productivity, total (constant 2015 US$ GDP per cubic meter of total freshwater withdrawal)
13. Annual freshwater withdrawals, total (% of internal resources)
14. Annual freshwater withdrawals, total (billion cubic meters)
15. Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)

16. Annual freshwater withdrawals, domestic (% of total freshwater withdrawal)
17. Annual freshwater withdrawals, industry (% of total freshwater withdrawal)
18. People using at least basic sanitation services (% of population)
19. Investment in water and sanitation with private participation (current US$)
20. Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)
21. People using at least basic drinking water services (% of population)
22. Gross capital formation (annual % growth)
23. Adolescents out of school (% of lower secondary school age)
24. CO_2 emissions (metric tons per capita)
25. PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)
26. People practicing open defecation (% of population)
27. Population living in slums (% of urban population)
28. Water productivity, total (constant 2015 US$ GDP per cubic meter of total freshwater withdrawal)

Appendix B - Final List of Features:

Column Name	Description	Data Source
Access to electricity (% of population)	Percentage of population with access to electricity	World Bank
Carbon dioxide emissions from fuel combustion (millions of tonnes)	Carbon dioxide emissions from fuel combustion (millions of tonnes)	UNSDG
Carbon dioxide emissions from manufacturing industries per unit of manufacturing value added (kilogrammes of CO2 per constant 2015 United States dollars)	Carbon dioxide emissions from manufacturing industries per unit of manufacturing value added (kilogrammes of CO2 per constant 2015 United States dollars)	UNSDG
Carbon dioxide emissions per unit of GDP PPP (kilogrammes of CO2 per constant 2021 United States dollars)	Carbon dioxide emissions per unit of GDP PPP (kilogrammes of CO2 per constant 2021 United States dollars)	UNSDG
GDP per capita (current US$)	GDP per capita is gross domestic product divided by midyear population.	World Bank
Gross capital formation (annual % growth)	Annual growth rate of gross capital formation based on constant local currency	World Bank

People practicing open defecation (% of population)	People practicing open defecation refers to the percentage of the population defecating in the open, such as in fields, forest, bushes, open bodies of water, on beaches, in other open spaces or disposed of with solid waste.	World Bank
People using at least basic drinking water services (% of population)	The percentage of people using at least basic water services. This indicator encompasses both people using basic water services as well as those using safely managed water services.	World Bank
People using at least basic sanitation services (% of population)	The percentage of people using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households.	World Bank
Proportion of population covered by at least a 2G mobile network (%)	The percentage of inhabitants that are within range of at least a 2G mobile-cellular signal, irrespective of whether or not they are subscribers.	UNSDG
Proportion of population using safely managed drinking water services, by urban/rural (%) (*Target Variable for modelling*)	The proportion of population using an improved drinking water source which is accessible on premises, available when needed and free from faecal and priority chemical contamination.	UNSDG
Urban population (% of total population)	Urban population refers to people living in urban areas as defined by national statistical offices.	World Bank