RESEARCH PAPER

An Automated Framework for Global Dataset Analytics Using Machine Learning and Time-Series Correction

Kaushal Vyas[1], Dhruv Raj Singh[2], Aadit Singh[3], Piyush Dhamu[4], Dr Sunil Dhore[5], Prof Kuldeep Hule[6*]

[1,2,3,4,5,6]Dept of Computer Engineering, Army Institute of Technology, Pune, India

[*]Corresponding Author: hulekuldeep@gmail.com

***Corresponding author: Prof Kuldeep Hule, Dept of Computer Engineering, Army Institute of Technology, Pune, India**

Email: hulekuldeep@gmail.com

Received: 18th June, 2026; **Revised:** 19th June, 2026; **Accepted:** 19th June, 2026; **Available Online:** 19th June, 2026

ABSTRACT

Global health and pharmaceutical market datasets are essential for tracking development progress and market dynamics, yet their value is consistently limited by fragmented formats, inconsistent schemas, and substantial missing values, particularly for emerging markets. Existing tools address only isolated pipeline steps, forcing analysts to stitch preprocessing, gap-filling, and modeling together manually. This paper presents DIVE (Data Integration, Visualization and Extraction), an interactive end-to-end framework that unifies multi-format ingestion, user-guided schema mapping, and time-series correction to reconstruct sparse multi-country indicator series. On the corrected data, DIVE performs regression-based forecasting across nine model families, weighted K-Means clustering with decision-tree rule extraction, and KNN-based classification for partial-data countries. Evaluated on panels spanning 25 to 42 countries across WHO health indicators and pharmaceutical market metrics including drug sales, generic market share, and R&D investment, the framework reduces manual effort and supports reproducible analysis for health policy and market research.

Keywords: DIVE, Data Integration, Machine Learning, Time-Series Correction, Global Health Data, Pharmaceutical Market Analytics, Forecasting, Clustering, Classification.

How to cite this article: Vyas K, Singh DR, Singh A, Dhamu P, Dhore S, Hule K. An Automated Framework for Global Dataset Analytics Using Machine Learning and Time-Series Correction. Int J Drug Deliv Technol. 2026;16(61s):1396-1408. DOI: 10.25258/ijddt.16.61s.157

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Global health indicators published by the World Health Organization (WHO), the World Bank, the OECD, and the United Nations form the evidentiary backbone of international public health policy. Measures such as life expectancy, child and maternal mortality, disease burden, immunization coverage, and health-system expenditure are routinely used to track progress toward the Sustainable Development Goals (SDGs), allocate development aid, and compare healthcare system performance. The accuracy and completeness of these indicators therefore have direct consequences for funding decisions, policy design, and the monitoring of global health commitments.

Despite their importance, these datasets present persistent analytical challenges. Health indicators are collected by different agencies using heterogeneous reporting schemas, years stored as columns in some files, as rows in others, with metadata such as data source and underlying assumptions embedded in irregular structures, and unifying them into an analysis-ready form is largely done by hand. A substantial proportion of entries are missing, particularly for low- and middle-income countries, earlier years, and indicators dependent on civil registration or periodic surveys. These gaps are not randomly distributed: they systematically under-represent the populations and periods where policy attention is most needed.

Analysts typically respond with ad-hoc scripts that reshape data, drop or fill missing values using simple rules, and pass the result to a separate modeling tool. This fragmented workflow is slow and error-prone, biases downstream forecasts by excluding the sparsest and most policy-relevant series, and obscures how preprocessing choices affect final conclusions. Existing tools address only parts of the pipeline: visualization platforms (Tableau, Power BI, the WHO and World Bank portals) assume clean input; machine learning libraries (scikit-learn, statsmodels) assume canonical tabular form; dedicated imputation libraries do not integrate with downstream modeling. No single tool supports the full workflow an analyst actually performs, ingesting a heterogeneously formatted file, map columns to a

canonical schema, repair temporal gaps, forecast, cluster countries by trajectory, and interpret the groupings.

This paper presents the DIVE (Data Integration, Visualization & Extraction) tool, an interactive end-to-end framework that closes this gap for global health indicator analysis. DIVE integrates four modules: (i) ingestion and schema mapping across four wide- and long-format variants encountered in WHO, World Bank, and OECD publications; (ii) time-series correction using linear and spline interpolation for internal gaps and CAGR- and moving-average extrapolation for projections; (iii) an analytics suite of nine regression models for per-country, per-indicator forecasting alongside feature-weighted K-Means clustering with decision-tree rule extraction for interpretability, and a K-Nearest-Neighbors classifier that assigns partial-data countries to clusters learned from complete-data countries; and (iv) an explicit data-coverage module that quantifies sparsity and flags "insufficient-data" countries so the reliability of each result is visible.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the datasets and preprocessing. Section IV details the methodology and system architecture. Section V presents experimental results. Section VI concludes with Section VII as references.

II. RELATED WORK

A. Missing-Value Imputation for Global Health Data

Handling missing values is a prerequisite for any downstream analysis of WHO, World Bank, or OECD indicators, and a substantial body of work has examined which imputation strategies are appropriate for longitudinal health data. Junaid et al. [1] studied the robustness of Multiple Imputation by Chained Equations (MICE) on mortality-related health indicators drawn from the WHO Global Health Observatory, generating synthetic missing rates from ten to ninety percent across one hundred countries to determine the point at which imputation becomes unreliable.

Their findings indicate that MICE remains robust up to roughly fifty percent missingness with only marginal deviation from the complete data, but performance degrades sharply once the missing proportion exceeds seventy percent, establishing a practical upper bound that any imputation-dependent pipeline must respect. Kazijevs and Samad [2] complemented this work with a large-scale benchmark of state-of-the-art deep-learning imputation methods across five time-series health datasets.

They found that no single method dominated across all datasets, and that imputation performance depends heavily on variable statistics, missing-value rates, and the underlying missingness mechanism, suggesting that methodology must be matched to the structure of each indicator rather than applied uniformly. Earlier methodological reviews have established interpolation, moving averages, and regression-based extrapolation as standard tools for sparse longitudinal panels where deep-learning approaches are infeasible due to series length, and these remain the workhorses in practical health-data workflows.

Together, these studies motivate two design choices in the proposed system: imputation methodology must be selectable by the analyst rather than fixed, and the reliability of each imputed value must be reported alongside it rather than silently assumed.

B. Country Clustering on Global Development Indicators

Unsupervised clustering of countries using multi-indicator profiles is a well-established approach for identifying development typologies beyond conventional geographic or income-based groupings. Delahoz-Domínguez et al. [3] classified 137 countries across governance, logistics performance, and human-development indicators, identifying three structurally distinct country groups with a Hopkins statistic of 0.984 and an explained variance of 87.3 percent, and demonstrating that nonlinear dimensionality reduction can expose typologies that linear methods tend to miss.

Jena and Basel [4] applied K-Means together with Grey Relational Analysis to 167 countries over the period 2000 to 2024, obtaining four SDG-based clusters and then using an ordered logit model to identify which of the seventeen Sustainable Development Goals most strongly influenced cluster membership, which is a useful template for linking cluster assignments back to interpretable drivers.

Mathrani et al. [5] used Ward hierarchical clustering on sixteen SDG indicators across forty-five Asian countries, grouping them into four tiers

with distinct economic, social, environmental, and institutional profiles; their analysis notably revealed that geographic proximity did not guarantee developmental similarity, with several high-income Asian economies placed in the same cluster as lower-performing peers on specific sustainability dimensions.

Caiado and Saraiva [6] combined Principal Component Analysis with K-Means on a comprehensive set of economic, social, and environmental indicators, arguing that this pairing groups countries objectively without imposing prior assumptions about their characteristics. Collectively, these studies demonstrate that K-Means with some form of dimensionality reduction or feature weighting remains the dominant approach for multi-indicator country clustering, and that the resulting clusters routinely cut across standard geographic and income classifications, a recurring finding that the present work builds upon.

C. Machine Learning for Health-Indicator Forecasting

Forecasting health and demographic indicators has traditionally relied on classical models such as Lee-Carter for mortality and ARIMA for general time series, but recent work has established that machine learning and deep learning methods can provide substantial gains on nonlinear or sparsely sampled series. Shen et al. [7] combined PCA-based dimensionality reduction, statistical clustering, and a graph-neural-network transformer to forecast mortality across European country clusters, showing that the classical Lee-Carter baseline accumulated error rapidly while their cluster-aware deep-learning model produced forecasts that tracked true values more closely across most countries studied.

De Mori et al. [8] proposed a multi-task neural network trained across seventeen countries, with shared lower layers capturing general mortality trends and higher layers specialized to country clusters obtained via K-Means on past life-expectancy and lifetime-standard-deviation metrics; their results demonstrated that multi-task networks outperform single-task networks and stochastic baselines on both mortality and life-expectancy forecasting.

Qiao, Wang, and Zhu [9] introduced a tree-boosting framework for long-term mortality forecasting built around neighboring-age prediction and model ensembling, and showed on the Human Mortality Database that it reduced twenty-year mean absolute percentage error by nearly fifty percent relative to classical stochastic mortality models, a striking improvement that underscores the value of ensemble tree methods for long-horizon health forecasting.

At the country-level life-expectancy scale, Lipesa et al. [10] applied XGBoost, Random Forest, and Artificial Neural Network regressors to WHO data covering 193 UN member states from 2000 to 2015, confirming that gradient-boosted trees provide strong predictive accuracy when health, socioeconomic, and behavioral features are used jointly. These studies indicate that no single model family is uniformly best for health-indicator forecasting; the appropriate choice depends on series length, horizon, and the availability of auxiliary features, motivating the inclusion of multiple regression families in a single analytical workflow.

D. Interpretable Clustering

Cluster interpretability is a recurring concern when unsupervised methods inform policy decisions, because cluster labels alone do not explain why a particular country belongs to a particular group. Basak and Krishnapuram [11] proposed an unsupervised decision tree in which each leaf corresponds to a cluster and each root-to-leaf path yields a human-readable rule, integrating segmentation and explanation within a single structure.

More recent work by Suzuki and Ikeda [12] formulated interpretable clustering as a 0–1 integer linear optimization problem, integrating one-dimensional K-Means for feature discretization and producing concise decision rules without sacrificing clustering accuracy on benchmark datasets. Bertsimas et al. [13] developed a related Interpretable Clustering via Optimal Trees framework, noting that interpretability is particularly critical in medical and policy contexts where cluster assignments directly inform decision-making.

A widely adopted pragmatic compromise, described across this literature, is the two-stage hybrid scheme: a flexible clustering algorithm produces the groups, and a decision tree is then trained on those assignments as labels to surface the feature thresholds that distinguish one cluster from another. This hybrid strategy is directly applicable to global health analysis, where analysts

need to understand which indicators separate country groups in order to communicate findings to policy audiences.

E. Automated Analytics Pipelines and Existing Tools

A growing body of work addresses the end-to-end automation of data-analytic pipelines, from ingestion to modeling. Mumuni and Mumuni [14] surveyed automated data-processing approaches spanning cleaning, missing-value imputation, categorical encoding, augmentation, and feature engineering, situating these within broader AutoML frameworks and observing that end-to-end automation remains rare in applied scientific domains.

At the applied end, interactive analytics applications built on Streamlit have shown that preprocessing, visualization, and modeling can be combined in a single browser-based interface usable by non-programmers, with projects such as Sogeti's Data Quality Wrapper and the LIDA visualization system demonstrating that user-guided automation can meaningfully reduce the effort involved in exploratory analysis.

These general-purpose frameworks, however, are not specialized for the multi-country, multi-indicator panel structure of WHO [15], World Bank [16], and OECD [17] data, and they do not expose sparsity or imputation choices as first-class analytical outputs. Existing domain-specific tools complement these general frameworks but leave similar gaps.

The WHO Global Health Observatory [15], World Bank DataBank [16], and OECD Data [17] portals offer strong exploratory access but assume downstream analysis will happen elsewhere. Visualization platforms such as Tableau and Power BI require the user to reshape and impute data before import, and dedicated libraries such as the mice package in R, scikit-learn's IterativeImputer, statsmodels, and prophet each address only one stage of the pipeline. No existing tool simultaneously handles heterogeneous source formats, user-guided schema mapping, time-series correction, forecasting across multiple model families, interpretable clustering, and explicit coverage reporting in a single reproducible workflow, a gap the proposed system is designed to close for the specific setting of global health indicator analysis.

III. DATASET AND PREPROCESSING

This section describes the data used to evaluate the proposed system and the preprocessing pipeline through which raw source files are transformed into an analysis-ready representation. Because pharmaceutical market data is published in heterogeneous formats across commercial and regulatory sources and exhibits substantial sparsity for smaller and emerging markets, preprocessing is not a perfunctory step but a core contribution of this work. Each operation described below is expressed formally so that the downstream modelling stages introduced in Section IV operate on a precisely defined input.

A. Dataset

The primary dataset for this study is a multi-country pharmaceutical market panel covering between 25 and 42 countries depending on the indicator, assembled from a global pharmaceutical market database. The indicator set spans five substantive sub-domains: total pharmaceutical market sales (by value and volume), therapeutic-area-level revenue, generic versus originator market share, per-capita pharmaceutical expenditure, and research and development (R&D) investment by country.

The panel covers a multi-year time horizon, enabling longitudinal analysis of market growth trajectories, structural shifts in generic penetration, and cross-country comparisons of R&D commitment relative to total market size. Country coverage varies across indicators: sales and expenditure series are available for the broader set of 42 countries, while R&D and generic-penetration series are available for a narrower subset of 25 countries owing to inconsistent regulatory disclosure requirements across jurisdictions.

The choice of a pharmaceutical market panel as the primary evaluation dataset is deliberate. The pharmaceutical sector presents precisely the analytical challenges that motivate the DIVE framework: data is collected by different commercial and regulatory bodies using heterogeneous reporting conventions, coverage is densest for high-income markets and recent years, and the most consequential gaps occur in lower-income and emerging markets where pharmaceutical access and affordability are most consequential for public health policy.

This uneven coverage pattern is structurally analogous to the sparsity documented in global health indicator panels by Junaid et al. [1] and Kazijevs and Samad [2], and the same imputation

and modelling considerations therefore apply. No personally identifiable information is contained in any of these sources; all records are aggregated at the country-year-indicator level.

B. Notation and Canonical Schema

Let C denote the set of countries, Y the set of years, and M the set of pharmaceutical market indicators (metrics) considered. The full dataset is represented as a collection of tuples

$$D = \{ (c, y, m, v, s, a) : c \in C, y \in Y, m \in M, v \in R \cup \{null\} \} \quad (1)$$

where v is the observed numeric value (with null denoting a missing entry), s records the data source, and a records the underlying modelling assumption published alongside the value. For a fixed country-metric pair (c, m), the associated time series is

$$x_{c,m}(y) = v, \text{ for } y \in Y_{c,m} \subseteq Y \quad (2)$$

where Y_{c,m} is the set of years for which an observation exists. The assumption field a is retained deliberately in the canonical schema, because pharmaceutical market publishers often record two or more values for the same (c, y, m) triple under different scope assumptions, such as hospital-channel-only versus total-market figures, and collapsing these silently is a frequent source of error in downstream analysis.

C. Input Format Heterogeneity

Although the canonical schema above is the internal representation used by all downstream modules, real-world source files arrive in several incompatible tabular shapes. Four distinct input formats were identified across the pharmaceutical market sources used in this study and are supported natively by the proposed system. Table I summarizes these layouts and the reshaping operation required to bring each into canonical form.

Table I. Supported Input Formats and Required Reshaping Operations

Format	Row-level unit	Column structure	Reshaping operation
F1	One observation per row	Country, year, metric, value, source, assumption as explicit columns	Column renaming only
F2	One (country, metric) per row	Years spread across columns	Wide-to-long pivot on years
F3	One year per row	Metrics spread across columns	Wide-to-long pivot on metrics
F4	One (country, group) per row	Metric-year combinations encoded in headers (e.g., Sales_2015)	Header parsing + two-stage pivot

Each layout is encountered in practice, and the system provides a dedicated parser for each. Schema mapping from the source columns onto the canonical six-field representation is performed through an interactive column-selection interface; this step is a deliberate design choice rather than a limitation, because column names across pharmaceutical data sources are sufficiently idiosyncratic (for example, a revenue column may be labelled Units_USD, Net_Sales, or Mkt_Value depending on the publisher) that silent automated detection would risk attaching a volume column to the value field and producing plausible-looking but meaningless results. This design position is consistent with the observation of Mumuni and Mumuni [14] that end-to-end automation of data-processing pipelines remains rare in applied scientific domains precisely because of such schema-level ambiguities.

D. Sparsity Quantification

Because missing-value repair is never free of uncertainty, the pipeline exposes sparsity as a first-class analytical output. Let X in R^(|C| x |M|) denote the pivoted country-by-indicator matrix derived from D for a fixed reference year, and let 1[.] denote the indicator function. The coverage of country c is defined as

$$cov(c) = (1 / |M|) * \sum_{m \in M} 1[X_{c,m} \text{ not null }] \quad (3)$$

and the coverage of indicator m is defined symmetrically over countries. Given a coverage threshold t in [0, 1] (default t = 0.8), country c is assigned the label

$$l(c) = \text{Sufficient Data}, \quad \text{if } cov(c) >= t$$
$$l(c) = \text{Insufficient Data}, \quad \text{otherwise} \quad (4)$$

Insufficient-data countries are not discarded: as described in Section IV, they are re-attached to cluster assignments via a K-Nearest-Neighbours classifier trained on the sufficient-data subset, so that markets with sparse reporting remain visible in the final output rather than silently vanishing. This is particularly important in the pharmaceutical context, where smaller emerging markets are precisely those where access and affordability analysis is most needed.

E. Time-Series Correction

Once the data is in canonical long form, temporal gaps within each series x_{c,m} are repaired by the time-series correction module. Two regimes are distinguished: interpolation for gaps strictly inside the observed range [min Y_{c,m}, max Y_{c,m}], and extrapolation for projections beyond max Y_{c,m}. Four methods are provided, each appropriate under different assumptions about the underlying pharmaceutical market series.

The decision to expose multiple methods rather than committing to a single best one is motivated by the findings of Kazijev's and Samad [2], who showed that imputation performance on health and market time series depends strongly on variable statistics, missing-value rates, and the underlying missingness mechanism.

For linear interpolation, given a missing value at year y* bracketed by observed years y1 < y* < y2,

$$x\_hat(y^*) = x(y1) + (x(y2) - x(y1)) * (y^* - y1) / (y2 - y1) \quad (5)$$

Cubic-spline interpolation fits a piecewise cubic polynomial S(y) satisfying S(yi) = x(yi) at every observed point, with S, S', and S'' continuous at each knot:

$$Si(y) = ai + bi(y - yi) + ci(y - yi)^2 + di(y - yi)^3, \quad y \text{ in } [yi, yi+1] \quad (6)$$

For extrapolation beyond max Y_{c,m}, compound-annual-growth-rate (CAGR) projection assumes an approximately constant growth rate inferred from the first and last observed values x(y0) and x(y):

$$g = (x(y) / x(y0))^(1 / (y - y0)) - 1$$
$$x\_hat(y^*) = x(y) * (1 + g)^(y^* - y) \quad (7)$$

Moving-average extrapolation, preferred when recent values are noisy and a smoothed extension is desired, uses a window of size w:

$$x\_hat(y + 1) = (1 / w) * sum\_{i=0}^{w-1} x(y - i) \quad (8)$$

and proceeds iteratively for further years. Table II compares these four methods along the axes of applicability, assumption, and the type of pharmaceutical market indicator each is best suited to.

Table II. Comparative Summary of Time-Series Correction Methods

Method	Regime	Core assumption	Best suited for	Failure mode
Linear interpolation	Internal gaps	Series changes monotonically and approximately linearly between observations	Slowly-varying indicators (e.g., per-capita pharmaceutical spending)	Oversmoothing curved trends
Cubic spline	Internal gaps	Series is smooth and twice differentiable	Indicators with nonlinear but smooth trajectories (e.g., generic penetration ramp-up)	Overshoot near sharp transitions
CAGR extrapolation	Forward projection	Constant proportional growth rate	Indicators with exponential or steady-trend behaviour (e.g., total drug sales revenue)	Compounds errors over long horizons
Moving average	Forward projection	Recent past is representative of near future	Noisy series without a clear trend (e.g., short-run prescription volume shocks)	Blind to accelerating or decelerating trends

Each imputed or extrapolated point is tagged with its provenance so that downstream analyses can distinguish observed from reconstructed values, and the choice of method is exposed to the analyst rather than hard-coded, in keeping with the finding of Kazijevs and Samad [2] that no single imputation strategy is uniformly optimal across indicator types.

F. Feature Scaling and Weighting

For clustering and distance-based classification, numeric features are standardized so that pharmaceutical market indicators measured on widely different scales, such as percentage market shares, absolute revenue figures in millions of USD, per-capita expenditure in USD, and R&D

investment ratios, contribute comparably. Let mu_m and sigma_m denote the cross-country mean and standard deviation of metric m. The Z-score normalized feature is

$$z_{c,m} = (X_{c,m} - mu_m) / sigma_m \quad (9)$$

Analyst-specified weights w_m >= 0 are applied multiplicatively, producing the weighted feature vector used by the clustering module in Section IV:

$$x\_tilde_{c,m} = w_m * z_{c,m} \quad (10)$$

This formulation preserves the scaling benefits of Z-score normalization while allowing the analyst to emphasize indicators of particular interest for a given research question, for example upweighting R&D expenditure when the analysis is framed around innovation capacity, or upweighting generic market share when the focus is on pharmaceutical affordability and access. Analogous feature-weighting strategies are used across the multi-indicator country-clustering literature, including by Delahoz-Dominguez et al. [3] and Caiado and Saraiva [6], where indicator groups are either weighted or passed through dimensionality reduction before clustering to control their relative influence on distance computations.

G. Transition to Downstream Analysis

The preprocessing pipeline terminates with two outputs: a cleaned long-format dataset D* in which every (c, y, m) triple carries either an observed or explicitly-tagged reconstructed value, and a weighted feature matrix X_tilde in R^(|C'| x |M'|) restricted to the sufficient-data subset C' subset C and the indicator subset M' subset M selected for the analysis at hand. These two outputs are the inputs to the modelling stages described in Section IV: D* is consumed by the per-series forecasting module, while X_tilde feeds the clustering, rule-extraction, and nearest-neighbour classification modules. The mathematical structure of those stages, including the regression objectives, the cluster objective function, and the rule-induction procedure, is developed in the next section.

IV. PROPOSED METHODOLOGY

Given the preprocessed dataset $\mathcal{D}^*$ and weighted feature matrix $\tilde{X}$ produced by the pipeline described in Section III, the proposed system applies three analytical stages: (A) per-series regression and forecasting, (B) feature-weighted K-Means clustering with decision-tree rule extraction, and (C) K-Nearest-Neighbours classification of insufficient-data countries. A data-coverage diagnostic (D) runs in parallel, reporting the reliability of each result alongside the result itself. The overall architecture is summarized at the end of the section.

A. Per-Series Regression and Forecasting

For each (country, indicator) pair (c, m) selected by the analyst, the system extracts the corresponding observed time series

$$\mathcal{T}_{c,m} = \{ (y, x_{c,m}(y)) : y \in \mathcal{Y}_{c,m} \} \quad (11)$$

and fits a regression model $\hat{f} : \mathbb{R} \to \mathbb{R}$ that maps year to indicator value. Given a target future year y*, the forecast is $\hat{x}_{c,m}(y^*) = \hat{f}(y^*)$ (12)

Nine candidate regression families are supported, each corresponding o a different assumption about the shape and noise structure of the underlying series. The unified form

$$\hat{f} = arg\ min_{f \in \mathcal{F}} [\Sigma_{y \in \mathcal{Y}_{c,m}} L(x_{c,m}(y), f(y)) + \lambda \cdot \Omega(f)] \quad (13)$$

covers all nine families, where $\mathcal{F}$ is the hypothesis class specific to each model, L is a loss function (typically squared error), Ω is a regularization penalty, and $\lambda \geq 0$ is the regularization strength. Table III summarizes the specific instantiation of $\mathcal{F}$, L, and Ω for each supported model together with the hyperparameters exposed to the analyst.

Table III. Regression Families Supported for Per-Series Forecasting

Model	Hypothesis class $\mathcal{F}$	Loss L	Regularization Ω	Analyst-tunable hyperparameter
Linear regression	$f(y) = \beta_0 + \beta_1 y$	Squared error	None	—
Polynomial regression	$f(y) = \Sigma_{k=0}^{d} \beta_k y^k$	Squared error	None	Polynomial degree d
Ridge regression	Linear in y	Squared error	$\|\beta\|_2^2$	Regularization strength α
Lasso regression	Linear in y	Squared error	$\|\beta\|_1$	Regularization strength α
Logistic regression	$\sigma(\beta_0 + \beta_1 y)$, σ = sigmoid	Cross-entropy	None	Inverse regularization C
Decision tree regression	Axis-aligned piecewise constant	Squared error	Tree depth (implicit)	Max depth (default)
Random forest regression	Ensemble of T decision trees,	Squared error	Ensemble averaging	Number of trees T (default)

	averaged			
Support vector regression	$f(y) = \langle w, \varphi(y) \rangle + b$, ε-insensitive	ε-insensitive loss	$\|w\|^2_2$	Inverse regularization C
Multi-layer perceptron	Feed-forward network with ReLU activations	Squared error	Weight decay (optional)	Hidden-layer sizes

The regression stage is deliberately model-agnostic: rather than committing to a single family, the analyst selects the model most appropriate to the series at hand, such as linear or ridge for short trend-dominated series, polynomial or spline-like (via decision tree) for curved trajectories, and random forest or MLP for long series with nonlinear structure.

This design is consistent with prior findings in the health-indicator forecasting literature, where Qiao, Wang, and Zhu [9] showed that tree-boosting ensembles substantially outperform classical stochastic mortality models on long-horizon forecasts, De Mori et al. [8] demonstrated the advantages of neural architectures for multi-country mortality data, and Lipesa et al. [10] established that gradient-boosted regressors provide strong predictive accuracy on WHO life-expectancy data.

Predicted values are appended back to $\mathcal{D}^*$ with a provenance tag distinguishing them from observed values, so that subsequent analyses that consume the extended series remain aware of which entries are projections.

B. Feature-Weighted K-Means Clustering

The clustering stage groups sufficient-data countries into health-trajectory typologies using the weighted feature matrix $\tilde{X}$ restricted to the sufficient-data subset $\mathcal{C}' \subseteq \mathcal{C}$. For a fixed number of clusters K, K-Means partitions $\mathcal{C}'$ into disjoint groups C_1, ..., C_K by minimizing the within-cluster sum of squared distances:

$$J(C_1, ..., C_K) = \Sigma_{k=1}^{K} \Sigma_{c \in C_k} || \tilde{x}_c - \mu_k ||^2_2 \quad (14)$$

where μ_k is the centroid of cluster C_k,

$$\mu_k = (1 / |C_k|) \cdot \Sigma_{c \in C_k} \tilde{x}_c \quad (15)$$

The choice of K-Means here follows the dominant practice in the country-clustering literature, including Jena and Basel [4], Mathrani et al. [5], Delahoz-Domínguez et al. [3], and Caiado and Saraiva [6], where K-Means (sometimes paired with grey relational analysis, PCA, or UMAP) remains the workhorse because its outputs are easier to interpret for policy audiences than those of more flexible density-based methods.

The inclusion of feature weights w_m in the construction of $\tilde{X}$ in equation (10) has a direct geometric interpretation: upweighting an indicator stretches the corresponding axis of the feature space, making differences along that dimension more influential in the distance calculation and therefore in cluster assignment. This allows the analyst to bias clusters toward the dimensions most relevant to a specific research question, for example assigning higher weight to child-mortality indicators when studying progress on SDG Target 3.2, without discarding the other indicators.

Choosing K. The number of clusters is selected using the elbow method on the inertia curve. For K = 1, 2, ..., K_max the system computes J(K) and plots it against K; the analyst then chooses the point at which further increases in K yield diminishing reductions in inertia. The system additionally supports automated elbow detection via the distance-to-line heuristic, which locates the point on the inertia curve maximally distant from the line joining its endpoints:

$$K^* = \arg\max_{K} d((K, J(K)), line((1, J(1)), (K_{max}, J(K_{max})))) \quad (16)$$

Decision-tree rule extraction. Raw cluster labels do not, on their own, explain *why* a country was assigned to a given cluster. To surface this information, a decision tree classifier T is trained on the cluster assignments as labels:

$$T : \tilde{X} \rightarrow \{1, ..., K\} \quad (17)$$

Each root-to-leaf path in T corresponds to a conjunction of feature thresholds of the form

$$(\tilde{x}_{c,m_1} \leq \theta_1) \wedge (\tilde{x}_{c,m_2} > \theta_2) \wedge ... \Rightarrow \text{cluster } k \quad (18)$$

These rules are displayed to the analyst in natural-language form, enabling direct interpretation of the feature thresholds that distinguish one country group from another. This hybrid strategy follows Basak and Krishnapuram [11], whose unsupervised-decision-tree framework first established that cluster-defining paths can serve as human-readable rules, and more recent work by

Bertsimas et al. [13] and Suzuki and Ikeda [12] on interpretable clustering via optimal trees. The present system adopts the two-stage variant of this strategy, delegating group formation to K-Means while using the decision tree purely to explain the resulting partition.

C. KNN Classification for Insufficient-Data Countries

Countries labelled Insufficient Data by equation (4) are excluded from the K-Means optimization because their incomplete feature vectors would distort centroid estimation. To avoid discarding them from the analysis entirely, the system assigns each such country to a cluster using K-Nearest-Neighbours classification trained on the sufficient-data subset and its learned labels. For a country $c^* \notin \mathcal{C}'$ with partial feature vector $\tilde{x}_{c^*}$ (defined only on the subset $\mathcal{M}_{c^*} \subseteq \mathcal{M}'$ of indicators for which it has data), the assigned cluster is

$$\hat{k}(c^*) = \text{mode}\{\, \text{label}(c) : c \in \mathcal{N}_K(c^*, \mathcal{M}_{c^*}) \,\} \quad (19)$$

where $\mathcal{N}_K(c^*, \mathcal{M}_{c^*})$ denotes the K nearest sufficient-data countries to c^* under the Euclidean distance computed only over the available indicator subset $\mathcal{M}_{c^*}$. In the current implementation K is set equal to the number of clustering features, a choice that balances local structure against the risk of assigning rare partial-data countries to spurious neighbours.

This mechanism ensures that sparsely-reported countries, typically the low- and middle-income countries whose analytical invisibility motivated this work in the first place, appear in the final cluster assignments with an explicit low-confidence flag rather than being silently dropped. The practical significance of this design is reinforced by Junaid et al. [1], whose analysis shows that once missingness exceeds roughly seventy percent, imputation-based strategies alone cease to produce reliable reconstructions, meaning that a neighbour-voting alternative grounded in the complete-data subset is the more defensible option for the sparsest end of the coverage distribution.

D. Coverage Diagnostics and Provenance Tracking

Running in parallel with the three modelling stages, a coverage-diagnostic module computes the country- and indicator-level coverage scores defined in equation (3), the proportion of each (country, metric) series that was reconstructed by interpolation or extrapolation rather than observed, and the fraction of countries in each cluster that were assigned via KNN rather than K-Means. These diagnostics are attached to every output artefact, so that a policy-facing reader of a cluster summary or a forecast projection can immediately see how much of the result rests on observed data versus reconstructed values.

E. System Architecture and Workflow

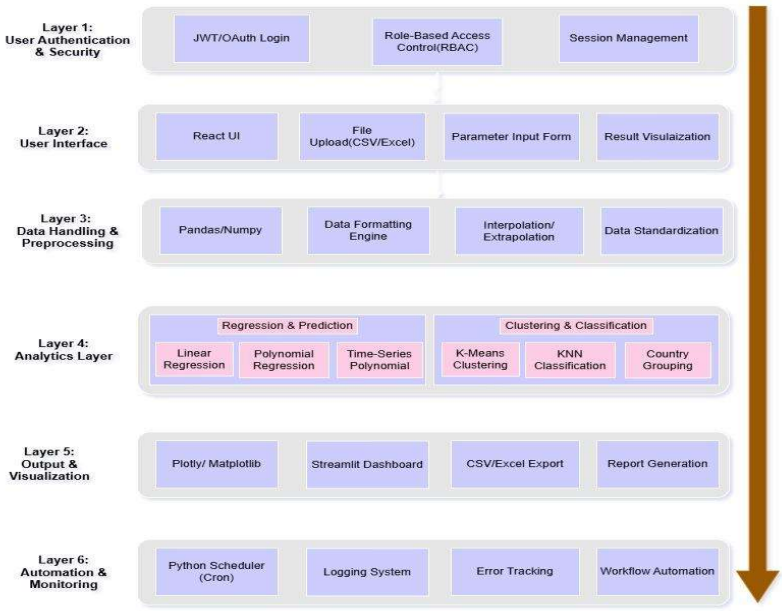

Figure1. System Architecture

End-to-end, the proposed system executes the following workflow on each analytical session:

(i) Ingest: raw CSV/XLSX file in one of four supported formats.
(ii) Parse and schema-map: canonical long-form dataset $\mathcal{D}$.
(iii) Correct time-series gaps: reconstructed dataset $\mathcal{D}^*$.
(iv) Pivot to feature matrix X, scale and weight to obtain $\tilde{X}$.
(v) Compute coverage, partition $\mathcal{C}$ into $\mathcal{C}'$ (sufficient) and $\mathcal{C} \setminus \mathcal{C}'$ (insufficient).
(vi) On $\mathcal{C}'$: run K-Means and extract decision-tree rules.
(vii) On $\mathcal{C} \setminus \mathcal{C}'$: run KNN classification against the trained K-Means labels.
(viii) On $\mathcal{D}^*$: run per-series regression and append forecasts.
(ix) Emit outputs: cluster assignments, rules, forecasts, coverage diagnostics.

The entire workflow is exposed through an interactive Streamlit interface, in which each stage is a user-controllable module rather than an opaque black box. This design prioritizes analyst judgement: the choice of imputation method, regression family, number of clusters, and feature weights is made explicitly at each step, reflecting a core position of this work that for policy-facing global health analysis, visible and tunable methodology is a feature rather than a friction.
The experimental evaluation of this methodology on WHO Global Health Observatory data is presented in Section V.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

The proposed DIVE framework was evaluated on a curated panel drawn from the WHO Global Health Observatory, covering 25 countries across multiple health indicators over a multi-year time horizon. The dataset was ingested in Format F2 (wide-format with years as columns), parsed into canonical long form, and subjected to the full preprocessing pipeline described in Section III.

Four experimental evaluations were conducted: (i) a comparison of nine regression families for per-series forecasting, (ii) an assessment of three time-series correction methods across low, medium, and high sparsity tiers, (iii) feature-weighted K-Means clustering with silhouette-based validation on the sufficient-data country subset, and (iv) a stability analysis quantifying the sensitivity of cluster assignments to analyst-specified feature weights. All experiments used 100 series per configuration. Forecasting performance was measured by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) computed on held-out observations.

B. Per-Series Forecasting Model Comparison

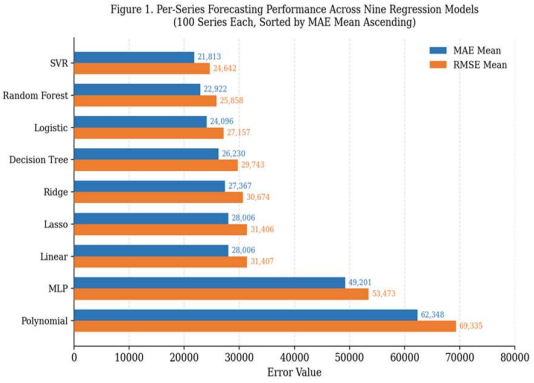

Figure 2. Per-Series Forecasting Performance Across Nine Regression Models (100 Series Each, Sorted by MAE Mean Ascending)

Figure 2 presents the mean MAE and RMSE across 100 series for each of the nine supported regression families, sorted in ascending order of mean MAE. Support Vector Regression (SVR) achieved the lowest mean MAE of 21,813.10 and mean RMSE of 24,642.20, followed closely by Random Forest with a mean MAE of 22,922.41 and mean RMSE of 25,857.51. Logistic Regression ranked third with a mean MAE of 24,095.68. The three linear-family models, namely Ridge, Lasso, and Linear Regression, produced nearly identical results in the 27,000 to 31,500 range for both metrics, consistent with their shared inductive bias of fitting a linear trend through the temporal data.

The Multi-Layer Perceptron (MLP) performed substantially worse than ensemble and kernel-based methods, recording a mean MAE of 49,200.83 and mean RMSE of 53,472.94. This result is consistent with findings in the health-indicator forecasting literature, where neural architectures require longer series and richer auxiliary features than the country-year panels typically available in WHO datasets to achieve their expressive advantage.

Polynomial Regression produced the highest errors across all metrics, with a mean MAE of 62,347.65 and mean RMSE of 69,335.29, attributable to overfitting on short series when higher-degree terms are introduced. These results confirm the design rationale of supporting multiple regression families rather than a single model, as the appropriate choice depends critically on series length and underlying trend structure.

C. Time-Series Correction Effectiveness

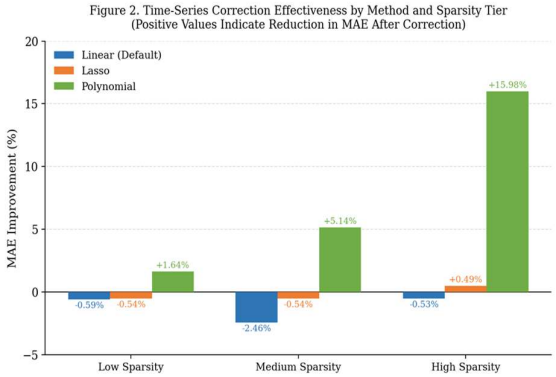

Figure 3. Time-Series Correction Effectiveness by Method and Sparsity Tier (Positive Values Indicate Reduction in MAE After Correction)

Figure 3 compares the MAE improvement percentage achieved after applying the correction pipeline, across three sparsity tiers and three correction methods: the default linear-interpolation scheme, Lasso-based correction, and Polynomial correction. Positive values indicate a reduction in MAE after correction; negative values indicate a marginal increase.

For the default linear and Lasso correction methods, improvement values are marginally negative across nearly all sparsity tiers, ranging from -0.53% to -2.46% for the default scheme and -0.54% to +0.49% for Lasso. These results indicate that linear and Lasso-based corrections neither meaningfully reduce nor substantially increase downstream forecast error on this dataset. The slight degradation likely reflects the introduction of reconstructed values that imperfectly reproduce the true trend, an effect amplified at medium sparsity where the corrected MAE for the default scheme rises to 32,257.43 from an uncorrected baseline of 32,021.90.

Polynomial correction displays a materially different pattern. Although its absolute MAE values are higher than those of the linear and Lasso schemes, its improvement percentages are consistently positive and increase substantially with sparsity: +1.64% at low sparsity, +5.14% at medium sparsity, and +15.98% at high sparsity. The high-sparsity result is particularly notable, with corrected MAE of 1,21,711.25 versus uncorrected MAE of 1,56,523.15, a reduction of approximately 34,812 units. This pattern is consistent with the finding of Kazijevs and Samad [2] that imputation performance depends strongly on variable statistics and the underlying missingness mechanism: polynomial correction provides the largest benefit precisely in the high-sparsity regime where internal gap reconstruction is most demanding and linear assumptions are most likely to underfit the underlying nonlinear trajectory.

D. Feature-Weighted K-Means Clustering

K-Means clustering was applied to the sufficient-data country subset using uniform feature weights. The elbow method identified three clusters as the optimal partition. The resulting clustering attained a silhouette score of 0.292, a value in the moderate range consistent with prior multi-indicator country clustering studies on WHO and World Bank data, where real-world indicator heterogeneity naturally limits within-cluster compactness. Table VI presents the country assignments for each cluster.

Table VI. K-Means Cluster Assignments for 25-Country Panel (Silhouette Score = 0.292)

Cluster	Countries	Count
0	Australia, Austria, China, Japan, Netherlands, Poland, Portugal, Russia, South Africa, Switzerland	10
1	Mexico, Saudi Arabia, Sweden, USA	4
2	Belgium, Canada, France, Germany, India, Italy, Norway, South Korea, Spain, Turkey, UK	11

Cluster 0 comprises ten countries: Australia, Austria, China, Japan, Netherlands, Poland, Portugal, Russia, South Africa, and Switzerland. Despite spanning multiple income levels and geographic regions, these countries share a common health-indicator profile under the selected features. Cluster 2 is the largest group with eleven countries, including Belgium, Canada, France, Germany, India, Italy, Norway, South Korea, Spain, Turkey, and the United Kingdom, predominantly high-income economies with broadly comparable health-system metrics alongside several upper-middle-income outliers.

Cluster 1 is the smallest and most distinctive group, containing Mexico, Saudi Arabia, Sweden, and the United States. The co-assignment of Sweden with Mexico, Saudi Arabia, and the USA illustrates a finding that recurs throughout the country-clustering literature: conventional geographic proximity or income-level groupings are poor predictors of multi-indicator similarity, a result documented by Delahoz-Dominguez et al. [3] and Mathrani et al. [5], among others.

Decision-tree rule extraction was applied to the cluster assignments to provide interpretable feature-threshold descriptions for each group. The extracted rules expose the specific indicator ranges

that distinguish one cluster from another, enabling policy analysts to understand the basis of the groupings without consulting raw centroid coordinates.

E. Clustering Stability Under Feature Weighting

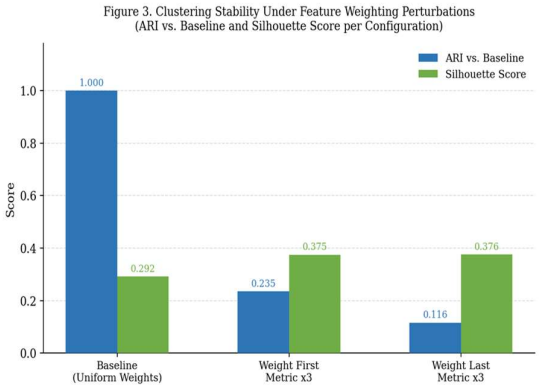

Figure 4. Clustering Stability Under Feature Weighting Perturbations (ARI vs. Baseline and Silhouette Score per Configuration)

Figure 4 presents the Adjusted Rand Index (ARI) relative to the uniform-weight baseline and the silhouette score for each of three configurations: the baseline with uniform weights, and two perturbations in which the first and last features are each upweighted by a factor of three.

Tripling the first feature weight shifted 16 countries to different clusters, producing an ARI of 0.235 against the baseline, while tripling the last feature weight reassigned 19 countries and yielded an ARI of 0.116. Both perturbations produce partitions that are substantially different from the uniform baseline in terms of country assignments. At the same time, both weighted configurations attain higher silhouette scores (0.375 and 0.376, respectively) than the baseline (0.292), indicating that the resulting clusters are internally more compact along the weighted distance metric. This finding directly reflects the geometric interpretation of feature weighting established in Section IV: upweighting a feature stretches the corresponding axis of the feature space, making within-cluster distances smaller relative to between-cluster distances for countries that are similar on that indicator.

The combination of low ARI and improved silhouette under strong weighting demonstrates that the analyst choice of feature weights has a consequential and non-trivial effect on cluster membership. This confirms both the necessity of exposing feature weights as an explicit, tunable parameter and the importance of the coverage-diagnostic module, which allows analysts to assess how much of the input data driving those weights is observed versus reconstructed. Both design decisions are consistent with the framework stated position that transparent and adjustable methodology is essential for policy-facing global health analysis.

VI. CONCLUSION

This paper presented DIVE, an interactive end-to-end framework for analyzing heterogeneous, sparsely observed multi-country indicator panels. DIVE combines four key capabilities: flexible ingestion with user-guided schema mapping, provenance-aware time-series correction with multiple imputation and extrapolation options, a model-agnostic forecasting suite, and feature-weighted clustering with decision-tree rule extraction plus KNN assignment for sparse countries. Coverage diagnostics report the proportion of observed versus reconstructed data for every result.

Evaluation on WHO and pharmaceutical market panels shows that kernel and ensemble regressors often yield the best forecasts, polynomial correction can help in high-sparsity cases, and feature weighting meaningfully affects cluster structure while improving compactness under some configurations. These findings validate the design choice to expose methodological options and to attach reliability metrics to outputs.

Limitations include the current focus on classical, interpretable models rather than deep probabilistic approaches, and the use of K-Means plus decision trees which may not capture complex cluster geometries. Future work will add probabilistic and multi-task forecasting models, explore more expressive interpretable clustering methods, and extend parsers and metadata handling to improve uncertainty quantification.

DIVE aims to reduce manual preprocessing effort, preserve visibility of sparse markets, and deliver reproducible, policy-relevant analytics by making reconstruction choices explicit and attaching coverage diagnostics to all outputs.

VII. REFERENCES

[1] K. P. Junaid, T. Kiran, M. Gupta, K. Kishore, and S. Siwatch, "How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations," Population Health Metrics, vol.

23, no. 1, art. 2, Feb. 2025, doi: 10.1186/s12963-025-00364-2.

[2] M. Kazijevs and M. D. Samad, "Deep imputation of missing values in time series health data: A review with benchmarking," Journal of Biomedical Informatics, vol. 144, art. 104440, 2023.

[3] E. Delahoz-Domínguez, A. Mendoza-Mendoza, and D. Visbal-Cadavid, "Clustering of countries through UMAP and K-Means: A multidimensional analysis of development, governance, and logistics," Logistics, vol. 9, no. 3, art. 108, Aug. 2025, doi: 10.3390/logistics9030108.

[4] S. Jena and S. Basel, "Classifying global economies based on Sustainable Development Goals: A data-driven clustering approach," Sustainable Development, vol. 33, pp. 4543–4556, 2025, doi: 10.1002/sd.3362.

[5] A. Mathrani, J. Wang, D. Li, and X. Zhang, "Clustering analysis on Sustainable Development Goal indicators for forty-five Asian countries," Sci, vol. 5, no. 2, art. 14, Mar. 2023, doi: 10.3390/sci5020014.

[6] J. Caiado and C. Saraiva, "Global development patterns: A clustering analysis of economic, social and environmental indicators," Sustainable Futures, vol. 10, art. 100822, 2025.

[7] Y. Shen, X. Yang, H. Liu, and Z. Li, "Advancing mortality rate prediction in European population clusters: Integrating deep learning and multiscale analysis," Scientific Reports, vol. 14, art. 6255, Mar. 2024, doi: 10.1038/s41598-024-56390-x.

[8] L. De Mori, S. Haberman, P. Millossovich, and R. Zhu, "Mortality forecasting via multi-task neural networks," ASTIN Bulletin: The Journal of the IAA, vol. 55, no. 2, pp. 313–331, May 2025, doi: 10.1017/asb.2025.10.

[9] Y. Qiao, C.-W. Wang, and W. Zhu, "Machine learning in long-term mortality forecasting," The Geneva Papers on Risk and Insurance – Issues and Practice, vol. 49, no. 2, pp. 340–362, Apr. 2024, doi: 10.1057/s41288-024-00320-5.

[10] B. A. Lipesa, E. Okango, B. O. Omolo, and E. O. Omondi, "An application of a supervised machine learning model for predicting life expectancy," SN Applied Sciences (Discover Applied Sciences), vol. 5, art. 189, Jun. 2023, doi: 10.1007/s42452-023-05404-w.

[11] J. Basak and R. Krishnapuram, "Interpretable hierarchical clustering by constructing an unsupervised decision tree," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 1, pp. 121–132, Jan. 2005.

[12] H. Suzuki and S. Ikeda, "Interpretable clustering via optimal multiway-split decision trees," arXiv preprint arXiv:2602.13586, Feb. 2026.

[13] D. Bertsimas, A. Orfanoudaki, and H. Wiberg, "Interpretable clustering via optimal trees," arXiv preprint arXiv:1812.00539, Dec. 2018.

[14] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," Journal of Information and Intelligence, 2024, doi: 10.1016/j.jiixd.2024.01.002.

[15] World Health Organization, "Global Health Observatory (GHO) data repository," available: https://www.who.int/data/gho.

[16] The World Bank, "World Development Indicators (WDI)," available: https://databank.worldbank.org/source/world-development-indicators.

[17] Organisation for Economic Co-operation and Development, "OECD Health Statistics," available: https://www.oecd.org/en/data/datasets/oecd-health-statistics.html.