Global Mortality Analysis: A Machine Learning Approach

Abstract

This project investigates global health and governance trends by integrating multiple international datasets and applying both unsupervised and supervised machine learning models. By systematically merging all the datasets by country and year we have constructed a comprehensive file at country-disease-year level. PCA and clustering techniques are used to uncover the patterns and grouping within data, while supervised models such as SVR and Decision Tree are used to find key health and social outcomes.

First tree-based regression models were developed to predict country-level mortality rates using healthcare infrastructure and disease burden features. Random forest and Gradient boosting achieved R2 scores of 0.706 and 0.687 respectively. Next support vector regression models with different kernels were evaluated. Radial Kernel SVR outperformed all other kernels achieving a highest r2 score (0.841) and lowest RMSE (29.97). Models like linear and polynomial kernels underperformed suggesting limited power to capture patterns in the data. These findings tell us the effectiveness of non-linear models in predicting health outcomes.

Introduction

The main objective is to identify meaningful structure in country-cause-year such as a natural cluster of countries that share similar health risk, levels of investment or governance using unsupervised learning methods. We have used PCA to reduce dimensionality and highlight the main axes of variation, followed by clustering algorithms(k-means and hierarchical methods) to discover grouping of the records that exhibit similar characteristics across mortality,corruption,medical resources,financial spending and happiness measures.

We also implemented supervised models using SVR(support vector regressor) and Decision trees.These models are trained to predict outcomes such as high mortality rate or low happiness using a country's health resources, investments and governance scores.

Using the auxiliary data file with advanced machine learning, this study provides a broader view of population health and well-being around the world.This project's findings have the potential to inform policymakers about which factors most differentiate successful from struggling countries, enabling better-targeted health interventions and governance reforms.

Technical Background

Unsupervised

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique used to summarize and visualize complex, high-dimensional datasets. PCA identifies the directions (principal

components) in which the data varies the most, projecting the data into a new coordinate system where the axes are uncorrelated and ordered by the amount of variance explained.
Mathematical formula:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

We can find m-th principal component score by taking a weighted sum of all the original variables, where the weights come from the corresponding loadings

KMeans Clustering:
KMeans is a partition-based, unsupervised clustering algorithm that splits data into *K* non-overlapping clusters by minimizing within-cluster variance (inertia).
Objective Function: KMeans minimizes the sum of squared distances from each point to its assigned cluster centroid.

Hierarchical Clustering:
Hierarchical clustering builds a tree (dendrogram) representing nested groupings of samples, without requiring a predefined number of clusters. It is widely used for its interpretability and visualization.
Linkage Methods in Hierarchical Clustering

Single Linkage**:** Distance between the closest pair of points in two clusters.Sensitive to “chaining” (can form long, stringy clusters).

Complete Linkage: Distance between the farthest pair of points in two clusters.Tends to produce compact, spherical clusters.

Average Linkage: Average distance between all pairs of points in two clusters.Balances single and complete linkage behaviour.

Ward’s Method: Merges clusters that result in the minimum increase in total within-cluster variance (inertia).Favors clusters of similar size, often preferred for quantitative data.

Supervised learning

Support vector machines:
Support vector machines is a supervised learning algorithm for classification and regression. SVM aims to find the optimal hyperplane that maximally separates data points of different classes in the feature space.General form of a kernel is a measure of relationships between any two points in the space.Support vectors have 3 types of kernels Linear Kernel:Computes the dot product between two feature vectors, suitable for linearly separable data. Polynomial Kernel: Computes similarity as a polynomial function of the dot product, capturing curved boundaries. Radial Kernel: Measures similarity based on the distance between points, ideal for non-linear and localized decision boundaries.

Decision Trees:
A Decision Tree is a supervised learning algorithm that models data using a tree structure of decisions based on feature values. It splits the data recursively based on the feature that maximizes class purity (classification) or minimizes error (regression). To improve performance and generalisation, ensemble methods were like Random Forest (bagging) and Gradient boosting (boosting) which combine multiple trees.
For random forest hypertuning can be done by adjusting the features like n_estimators(number of trees), max_depth(depth of the tree), max_features(number of features considered at each split), min_samples_leaf/split (to control the growth and improve regularization)
For gradient boosting hypertuning can be done by adjusting the features like n_estimators(number of boosting rounds), learning_rate(how much each tree contributes to the model), max_depth(limit complexity of individual tree) , min_samples_split/leaf(to prune weak learners).

About the Data

This project uses a merged dataset combining mortality data from the Global Burden of Disease (IHME) with global indicators from 2010 to 2021. Each row represents a unique Country-year-disease. We added health system features like (physicians and hospital beds per 1,000), financial metrics (health spending per person and % of GDP spent on healthcare), and social indicators(corruption index and happiness score). Country names were standardized for clean joins. The final dataset helps enables a rigorous machine learning analysis that can both reveal global trends and support policy-relevant insights

Methodology

Data Preprocessing
The data preprocessing began with the IHME Global Burden of Disease dataset, providing disease-specific mortality rates by country and year. Irrelevant columns were dropped, and country identifiers were standardized for clean merging. Additional datasets covering health spending, infrastructure, corruption, and happiness were cleaned separately, with column names harmonized and country names aligned using automated and manual methods.

All datasets were merged into IHME Global base using left joins on country and year to preserve mortality records while enriching them with contextual indicators. After merging, the data was checked for duplicates and missing values. Gaps were handled using forward fill, interpolation, or imputation as appropriate. The final dataset was fully aligned and ready for analysis.

Unsupervised Learning
For the unsupervised analysis section of continuous numeric features were made here we selected mortality rate, corruption index, physicians per 1000 patients, beds per 1000 patients. Health expenditure, percentage of gdp spent on healthcare and happiness score. Since clustering methods

depend on distance metrics and are sensitive to feature scales all numeric features were standardized. Reference columns such as country name , cause name and year were preserved for post clustering interpretation but were not used in the clustering process.

Supervised Learning

We dropped unnecessary ID columns and one-hot encoded the disease column to meet the requirements of models. Instead of one-hot encoding the country which creates more columns which increases overfitting risk, we applied smoothed K-Fold target encoding to convert each country into its average mortality rate based on training data. This helped preserve predictive signal without causing overfitting or leakage of data to the model.

Decision Tree

We started with a basic random forest regressor using 100 trees and a max_depth of 10 to capture patterns in the data. We used GridSearchCV with 5- fold cross validation to tune key hyperparameters like n_estimators, max_depth, max_features, min_samples_leaf, and min_samples_split. Feature importance graph was plotted to identify the top features.

For Gradient Boosting Regressor, we trained the model with conservative learning rate and moderate depth to prevent overfitting. We applied GridSearchCV with 5-Fold cross validation to tune key hyper parameters like learning rate, max_depth and minimum samples settings. Feature importance was also calculated for this method

Support Vector Regression

Cross-validation with 5 folds is performed for all kernels via GridSearchCV to optimize hyperparameters and improve SVR generalization. Common evaluation metrics used were RMSE, R squared and MAE.

Linear Kernel: It fits a straight hyperplane in feature space to model linear relationships. The regularization parameter C balances margin width against training error. Linear kernel is computationally efficient , and gives interpretable coefficients.

Radial basis function kernel(RBF): This maps input into an infinite dimensional space, enabling SVR to capture complex nonlinear relationships.This is more flexible for non-linear data.

Polynomial Kernel: This projects data into a higher degree feature space, using specified degree to model moderate nonlinear patterns. By adjusting degree and the regularization parameter C, it balances bias and variance. It’s well-suited when interactions of features influence the target.

Results:

Unsupervised Learning

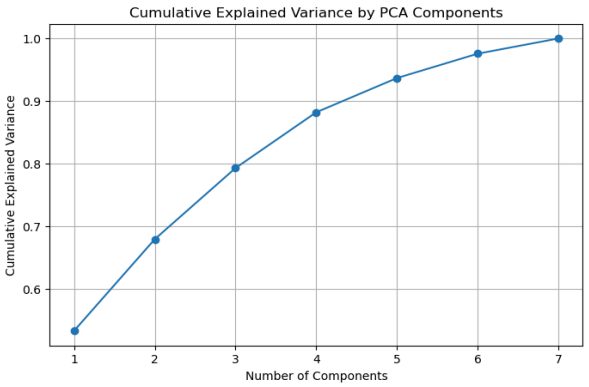


Figure: Cumulative Explained Variance Plot

The plot above shows that the first four principal components capture approximately 88% of the total variance in the dataset. Now the dataset is reduced from 7 dimensions to 4 .

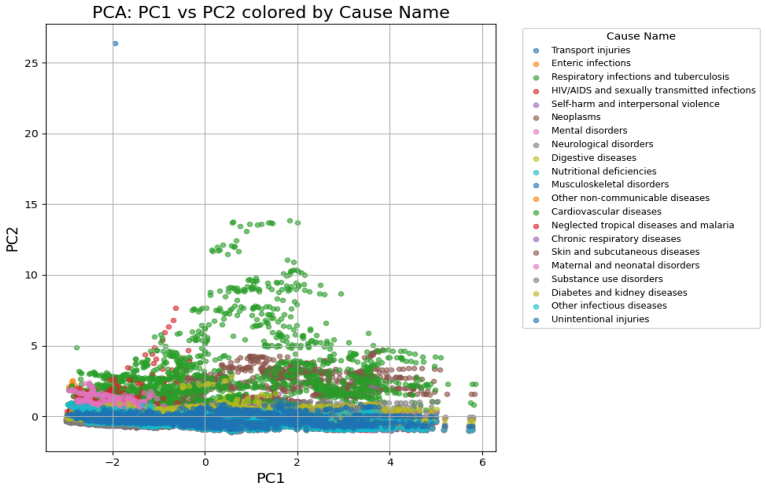


Figure: PCA plot

The PCA plot of PC1 vs PC2 provides a clear visualization of patterns in the data across different causes of death.

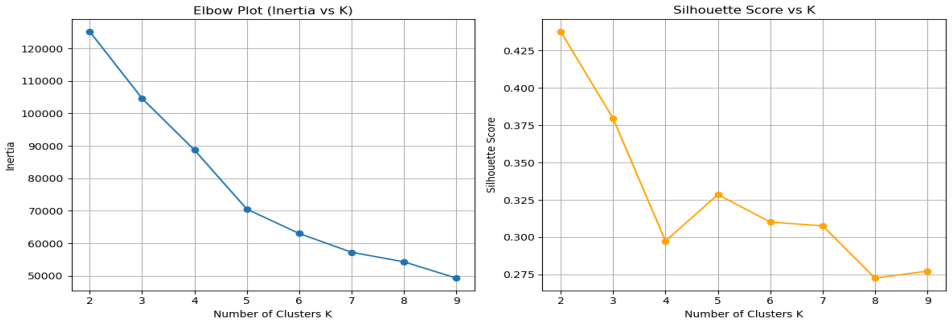


Figure: Optimal Number of cluster using Elbow Method and Silhouette Score

The elbow plot shows a clear bend at K=5, where the curve starts to flatten. The silhouette score is highest at k=2 (0.43), but that likely oversimplifies the data. The next best peak is at K=5, making it the most reasonable k value based on both methods.

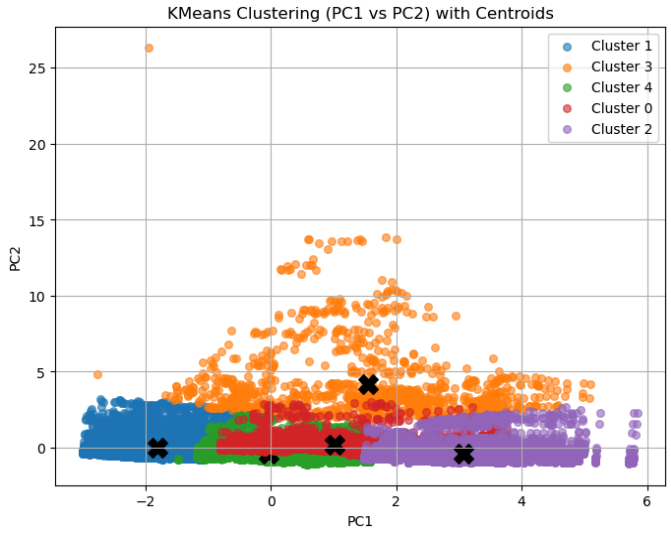


Figure: K means Clustering

The KMeans clustering on the PCA plot shows that the data is grouped into five distinct clusterClusters are separated and the centroids were positioned.

Hierarchical Clustering

Figure: Hierarchical Clustering Ward Dendrogram

The hierarchical clustering dendrogram shows five clear clusters based on the red line. The large gaps between some branches mean that the clusters are well separated. This suggests that hierarchical; clustering found meaning patterns in the data.

Decision tree

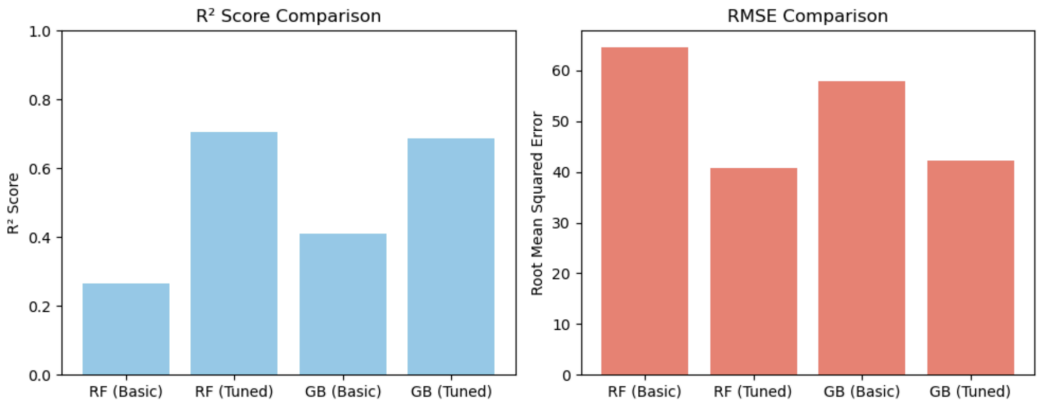


Figure: Comparison of Evaluation Metrics between the model

The plot compares the evaluation metrics like RMSE and R2 score for the four models. We can clearly see that the Tuned model for both Random Forest and Gradient Boosting outperformed the initial model.

Hyperparameter	Gradient Boosting (Tuned)	Random Forest (Tuned)
n_estimators	300	200
max_depth	7	40
Learning _rate	0.1	-
Min_samples_split	4	5
min_samples_leaf	1	1
max_features	-	'Sqrt'

The table shows the best hyperparameter selected through GridSearchCV for the tuned Gradient Boosting and tuned Random Forest.

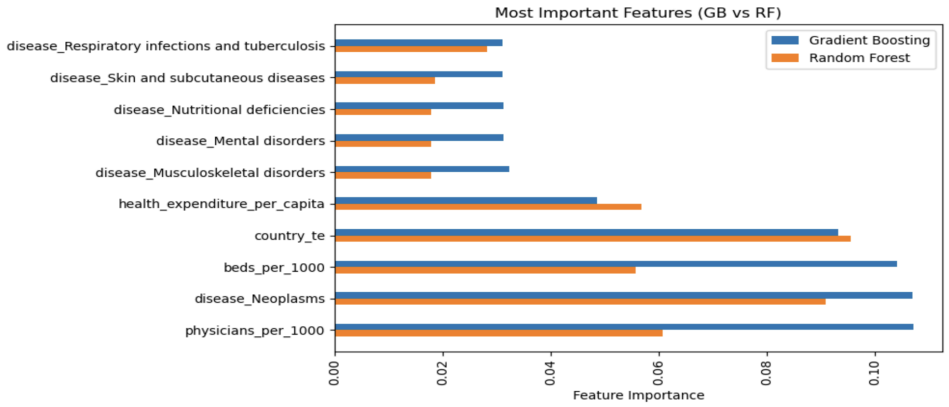


Figure: Feature Importance Graph for Random Forest and Gradient Boosting

From the graph we can tell that both models have similar important features but the degree of importance assigned to each feature differs for both the models. For example both models had physicians_per_1000, disease_Neoplasms (Cancerous) as the top feature but the Gradient Boosting tends to assign higher importance to these features when compared to Random Forest.

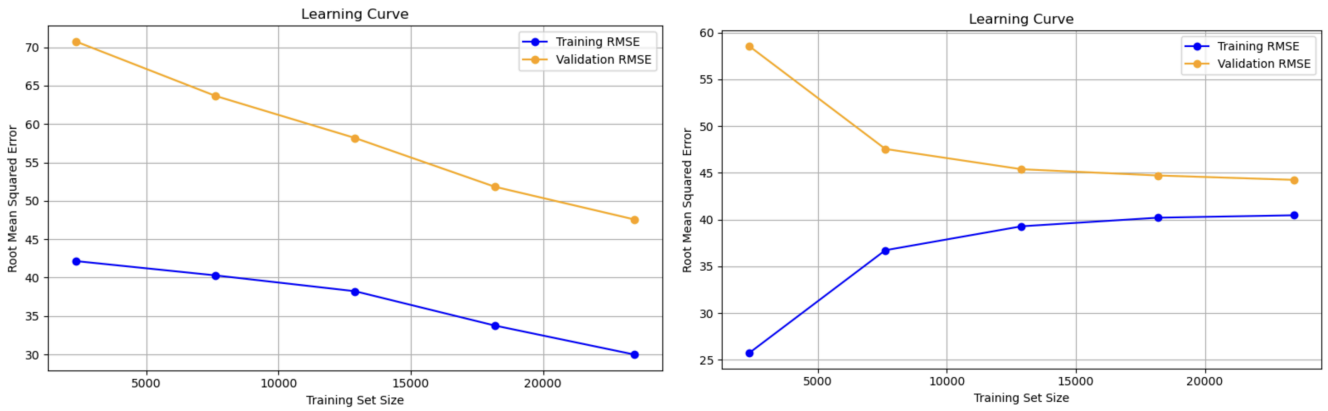
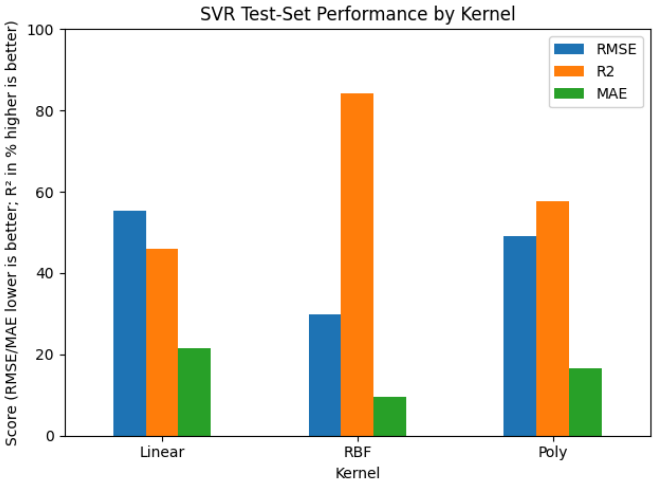


Figure: Learning Curve Comparison between the models

Random Forest (left) shows signs of overfitting, the gap between training and validation RMSE stays large. Gradient Boosting learns fast and stabilises, showing better model generalisation.

Support Vector Regressor:



The RBF kernel has the smallest errors and highest (R square =0.84), based on this it fits the data's curve best. Polynomial SVR is the second best model (R square = 0.57), the curve fits some but not all nonlinearity in the data.. The Linear SVR is the worst (R square = 0.46) because the data is nonlinear, so a straight-line model cannot fit well.

Discussion:

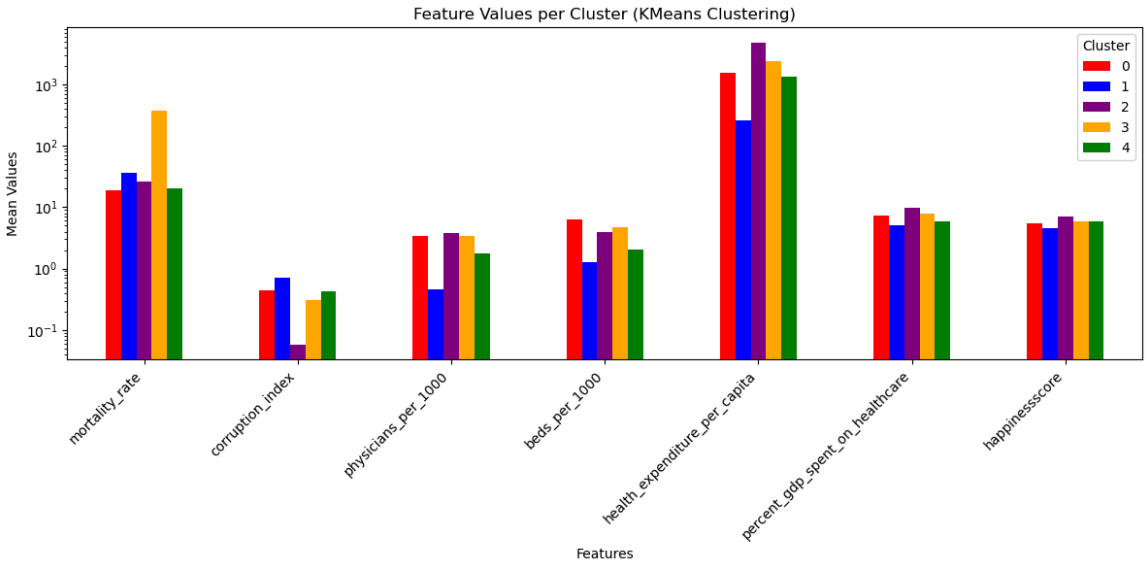


Figure: K means features for clusters

In K-Means clustering, countries grouped into five clusters with clear differences in healthcare and outcomes. Countries with high health spending, low mortality, and high happiness (Cluster 2) were mostly well-developed nations. In contrast, Cluster 1 had low spending, poor healthcare access, high mortality, and low happiness typical of low-income countries. Cluster 3 was interesting: despite good healthcare resources, it still had high mortality, suggesting other social or policy factors at play. Corruption also helped separate clusters, with lower corruption linked to better healthcare outcomes. Overall, the clusters showed that stronger healthcare systems align with better health and greater well-being.

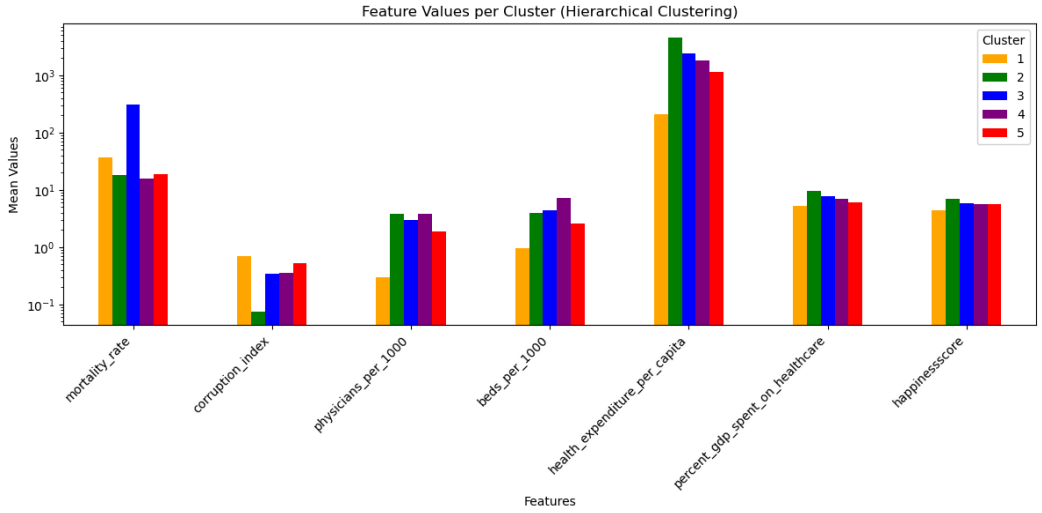


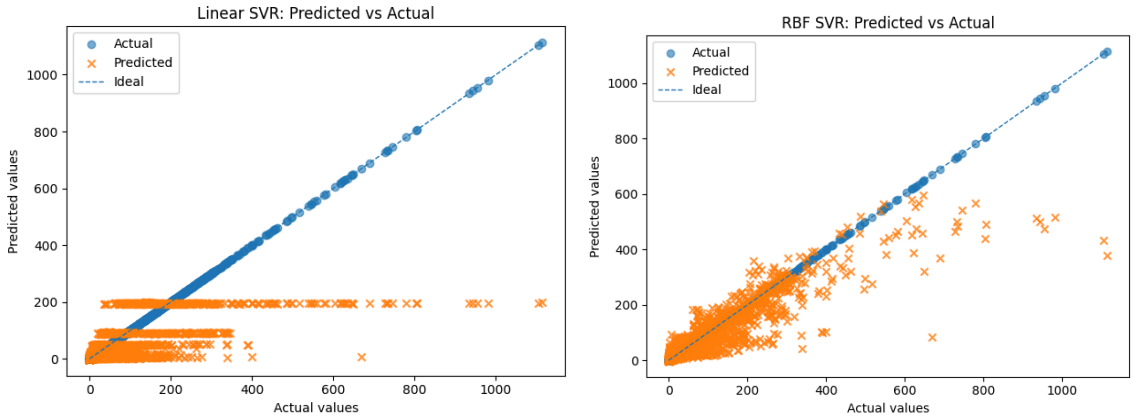
Figure: Hierarchical Clustering

In Hierarchical Clustering, countries were grouped into five clusters based on healthcare and well-being. Countries in Cluster 2 had high healthcare spending, low death rates, and high happiness,

showing strong healthcare systems. Cluster 3 included countries with very high death rates and poor healthcare, mostly low-income nations. Other clusters showed different mixes, for example, Cluster 4 had good healthcare but higher death rates, meaning healthcare spending alone isn't always enough. Corruption and happiness also played a role, with less corruption linked to better health and happier people. Overall, the clusters gave a clear picture of how countries differ in healthcare and quality of life.

Evaluation Metric	Random Forest	Gradient Boosting	Linear kernel SVR	Radial kernel SVR	Polynomial Kernel SVR
RMSE	40.824	42.1	55.38	29.97	49.03
R Square	0.706	0.687	0.459	0.841	0.57

The table shows us the comparison of models based on two key evaluation metrics RMSE and R2 score. From the table we can conclude that the Radial Kernel of Support Vector Regressor has the better performance.



Left plot shows the predicted vs actual value in the linear kernel of SVR where predictions lie along a straight trend but clearly unpredict large actual values and overpredict smaller ones, creating visible gaps at both extremes.

Right plot shows the prediction vs actual value in the RBF kernel of SVR where predictions lie around the ideal line by capturing nonlinear relationships.The result is because the kernel's high dimensional mapping lets it flexibly fit complex patterns.

Conclusion:

This project looked at how healthcare quality, investment, and governance relate to mortality outcomes. Unsupervised methods like PCA and clustering grouped countries with similar health profiles. These patterns were stable and can guide resource allocation.

In supervised learning, our best model (RBF SVR) explained **84%** of the variation in mortality. Key factors included hospital beds, doctors per 1,000, health spending, and corruption levels. The results highlight the global need to strengthen healthcare systems and improve transparency.

References

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Corinna Cortes & Vladimir Vapnik. (n.d.). Retrieved September, 1995, from

https://link.springer.com/article/10.1007/BF00994018

Principal component analysis (PCA). (n.d.).

https://scikit-learn.org/stable/modules/decomposition.html#pca

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press.