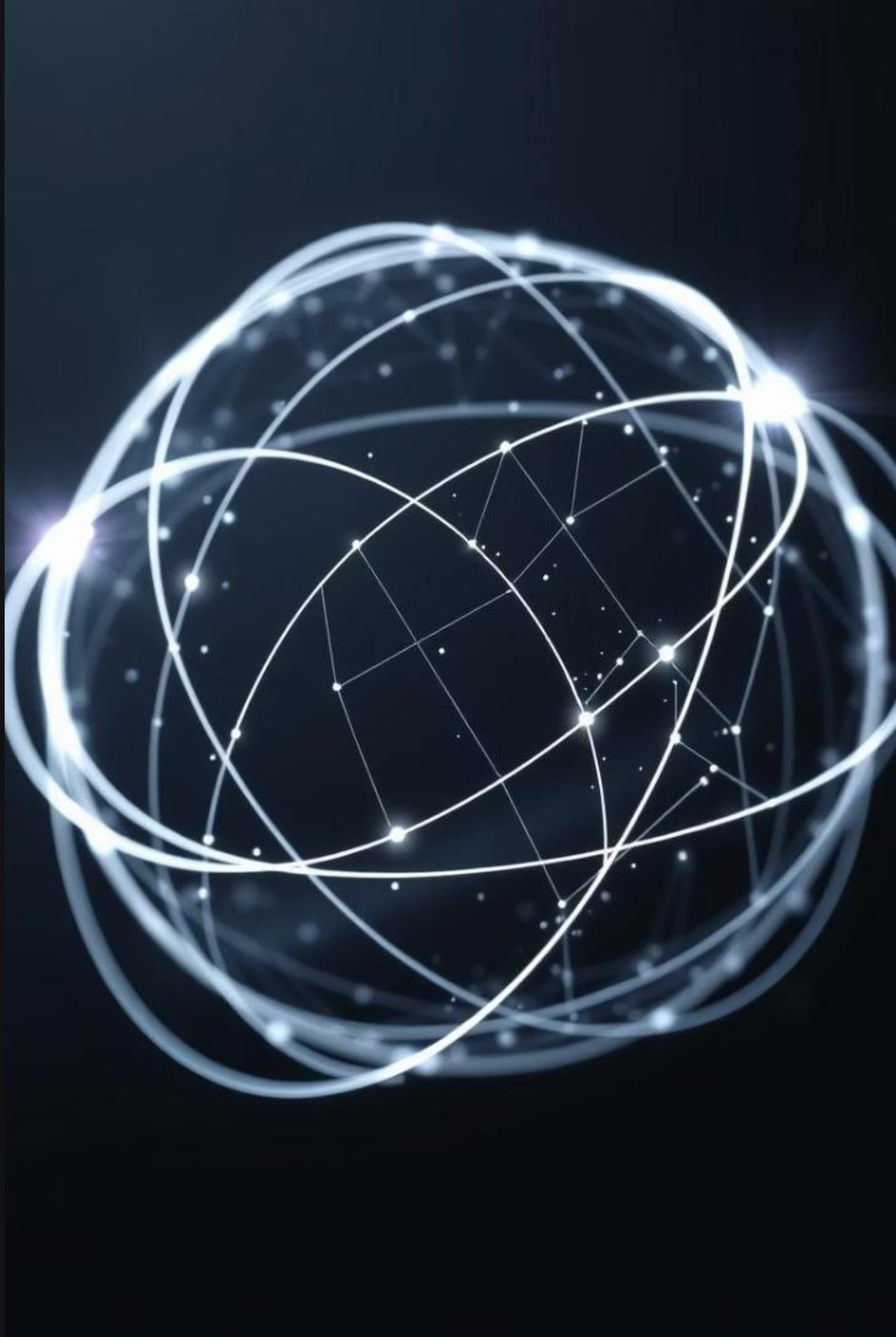
Global Mortality Analysis: A Machine Learning Approach

Understanding and predicting mortality rates across countries using advanced analytical models.

Presented by: Khaja Moinuddin Mohammed, Prasanth Gururaj, Sanjay Ramesh Kannan, Sowmya Polagoni, Venkat Saketh Kommi.



Project Goals

1

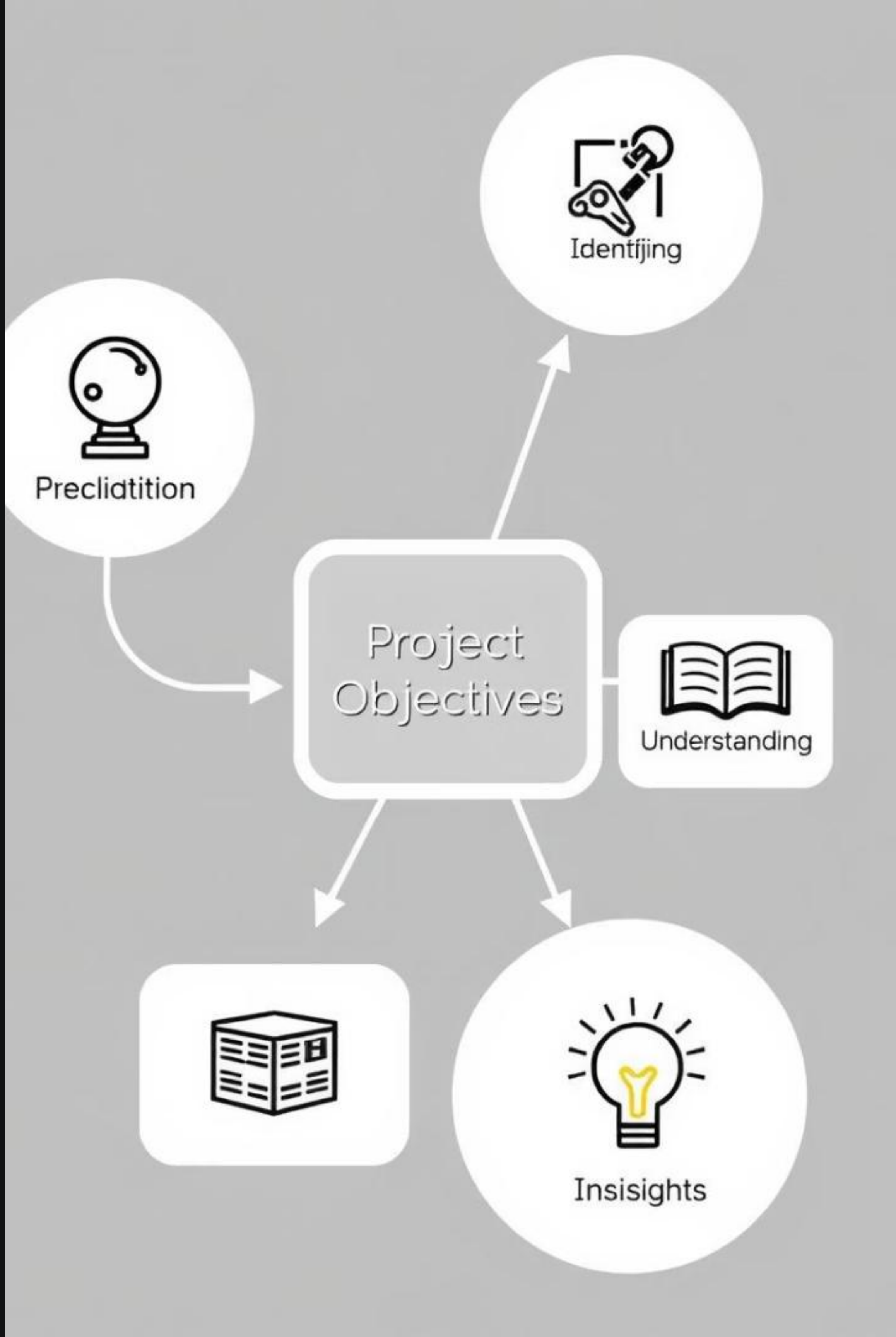
Primary Objective

Predict mortality rates using healthcare and socio-economic factors.

2

Secondary Objectives

- Identify key factors influencing mortality.
- Understand regional variations.
- Provide insights for healthcare policy.



Data Sources Overview

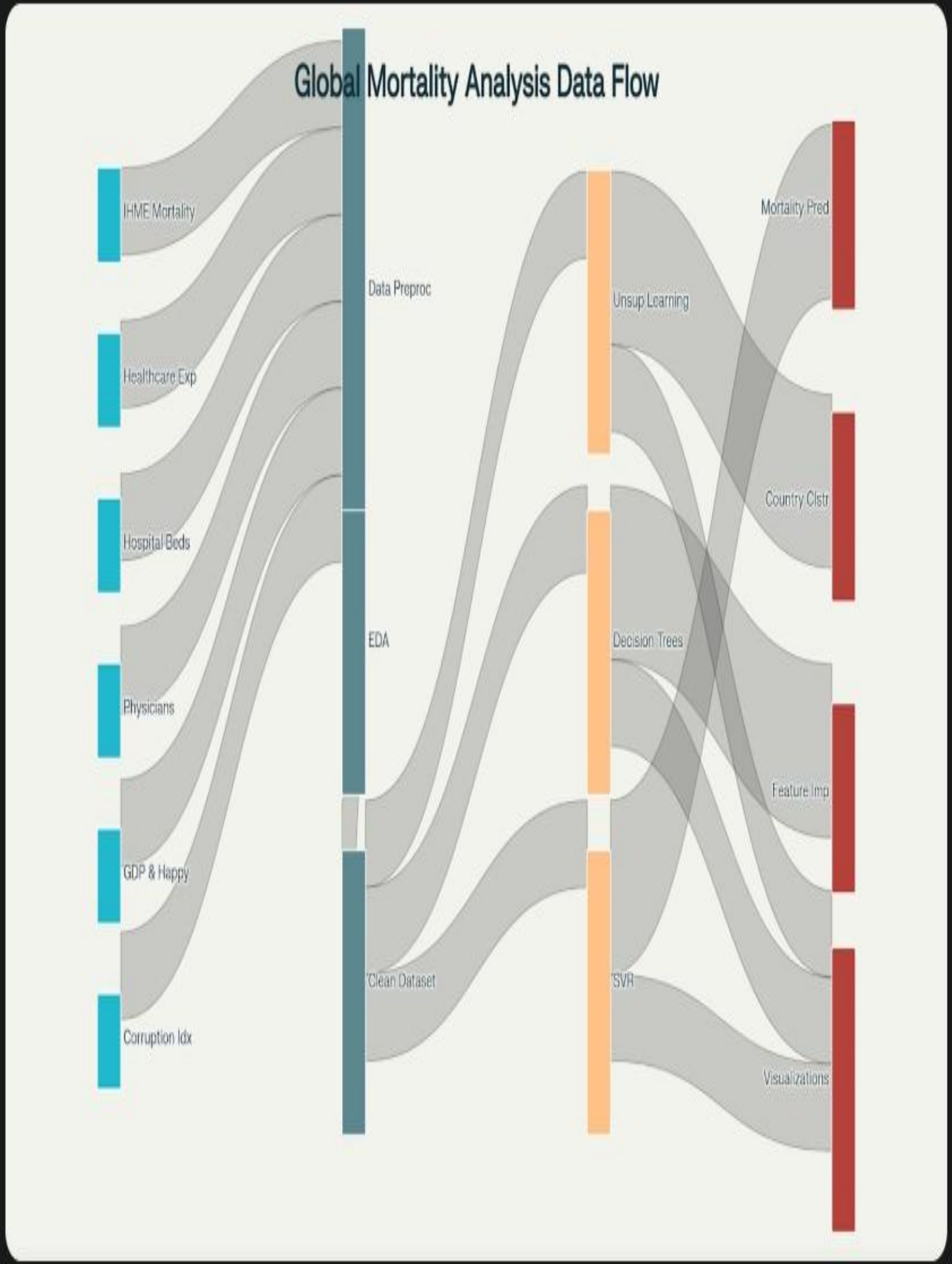
IHME Global Burden of Disease (GBD) 2021

Healthcare infrastructure data

Economic indicators

Governance metrics

Quality of life measures



Data Challenges & Preprocessing

- Missing value handling
- Country code standardization
- Feature engineering
- Data cleaning steps

Random Forest & Gradient Boosting Regressors

What is Random Forest Regressor?

An ensemble learning method for regression that operates by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees.

What is Gradient Boosting Regressor?

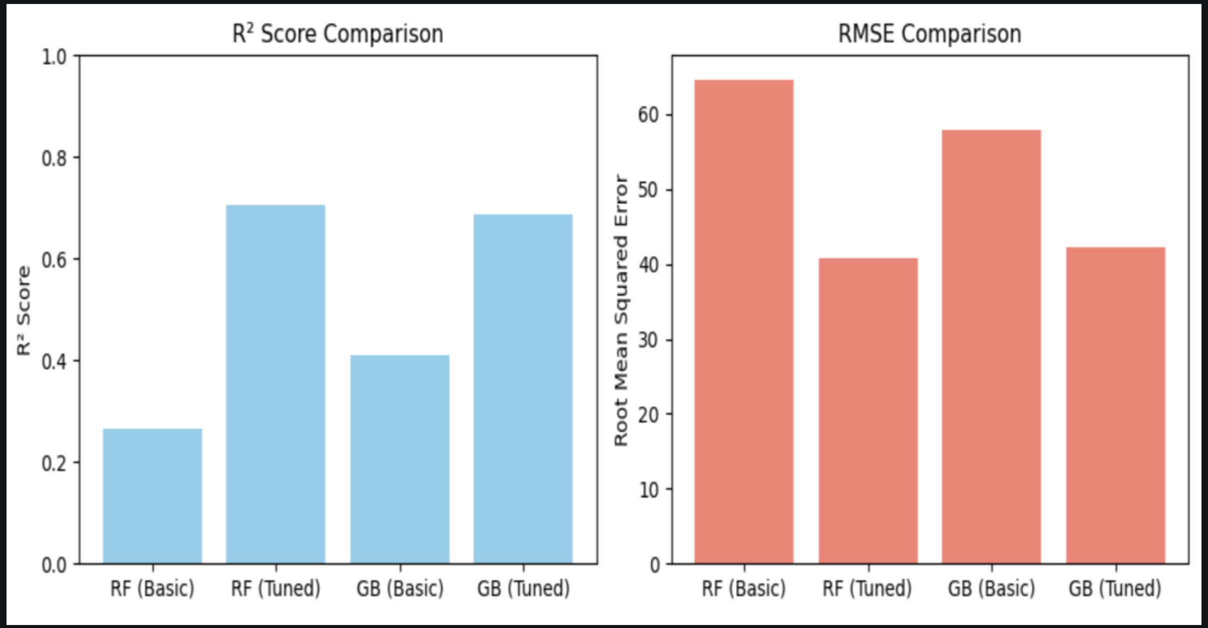
A machine learning technique for regression problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

When are they used?

Both are used for complex regression tasks, handling non-linear relationships and interactions between features. Random Forest is good for high-dimensional data, while Gradient Boosting often achieves higher accuracy.

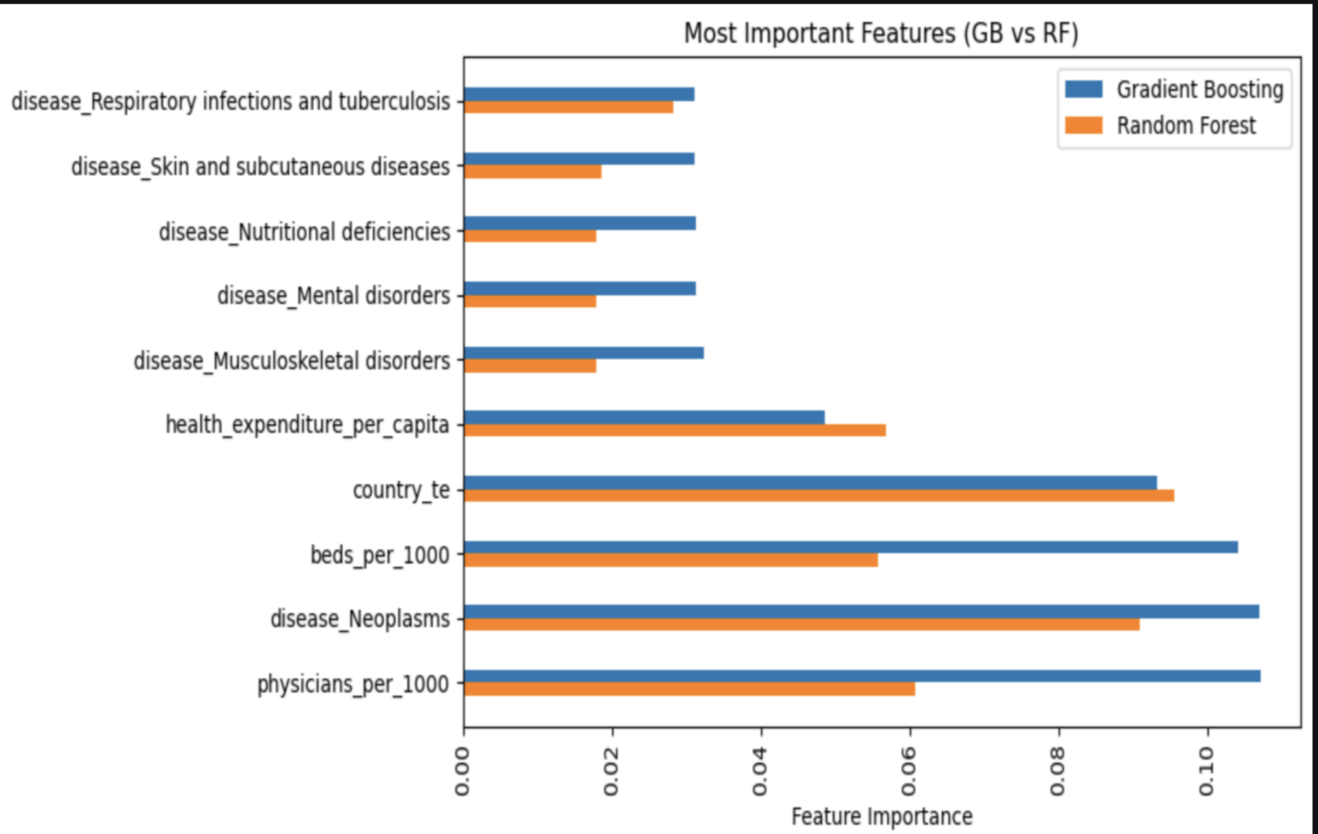
Key parameters

- n_estimators: Number of trees in the forest/ensemble.
- max_depth: Maximum depth of the tree.



Random Forest & Gradient Boosting Results

Metrics	Random Forest	Gradient Boosting	Random Forest (Tuned)	Gradient Boosting (Tuned)
RMSE	64.49	57.00	40.824	42.1
R-Squared	0.21	0.41	0.706	0.687

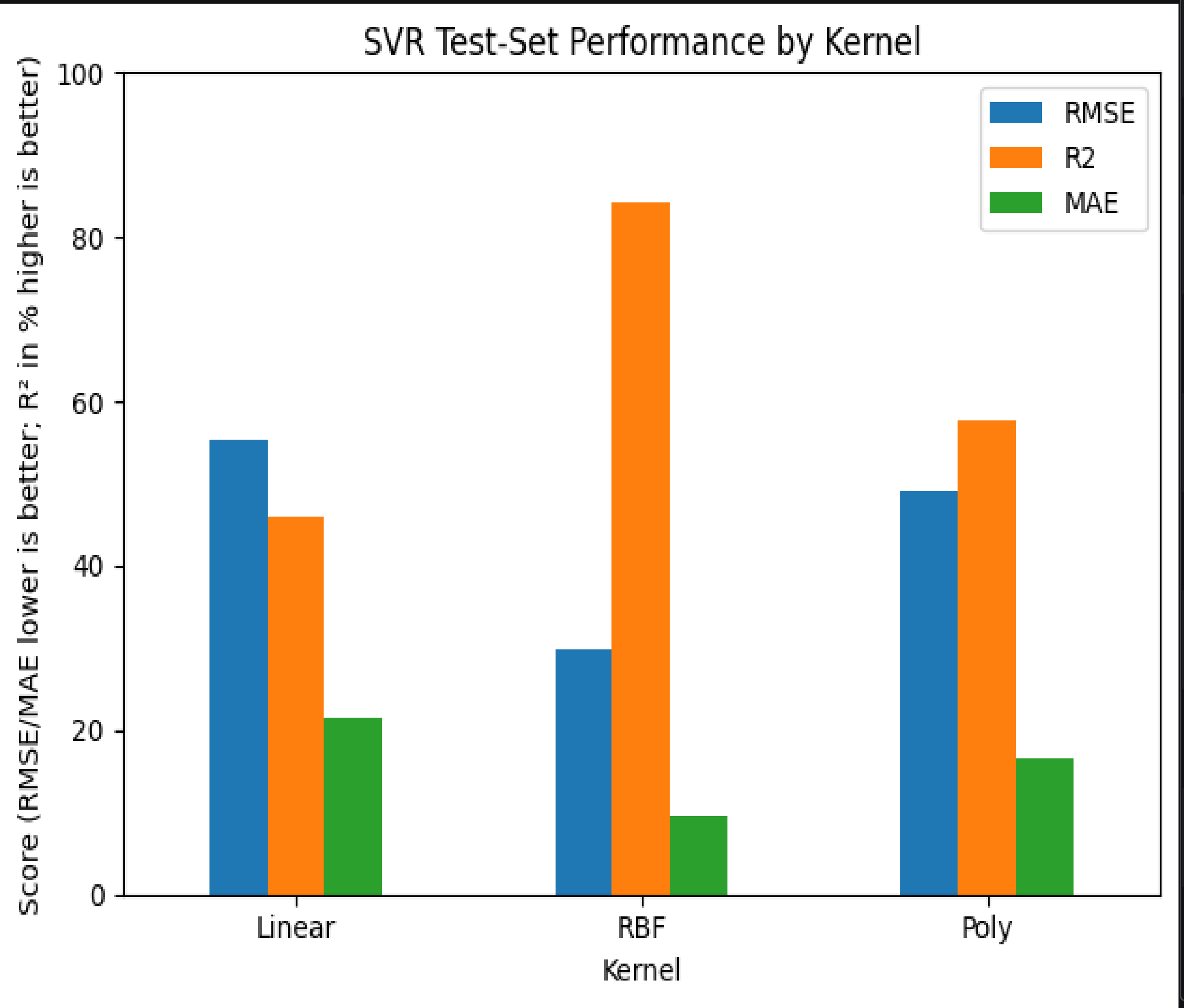


Initial and tuned model performance (Train/Test RMSE, R^2)

Best parameters after tuning

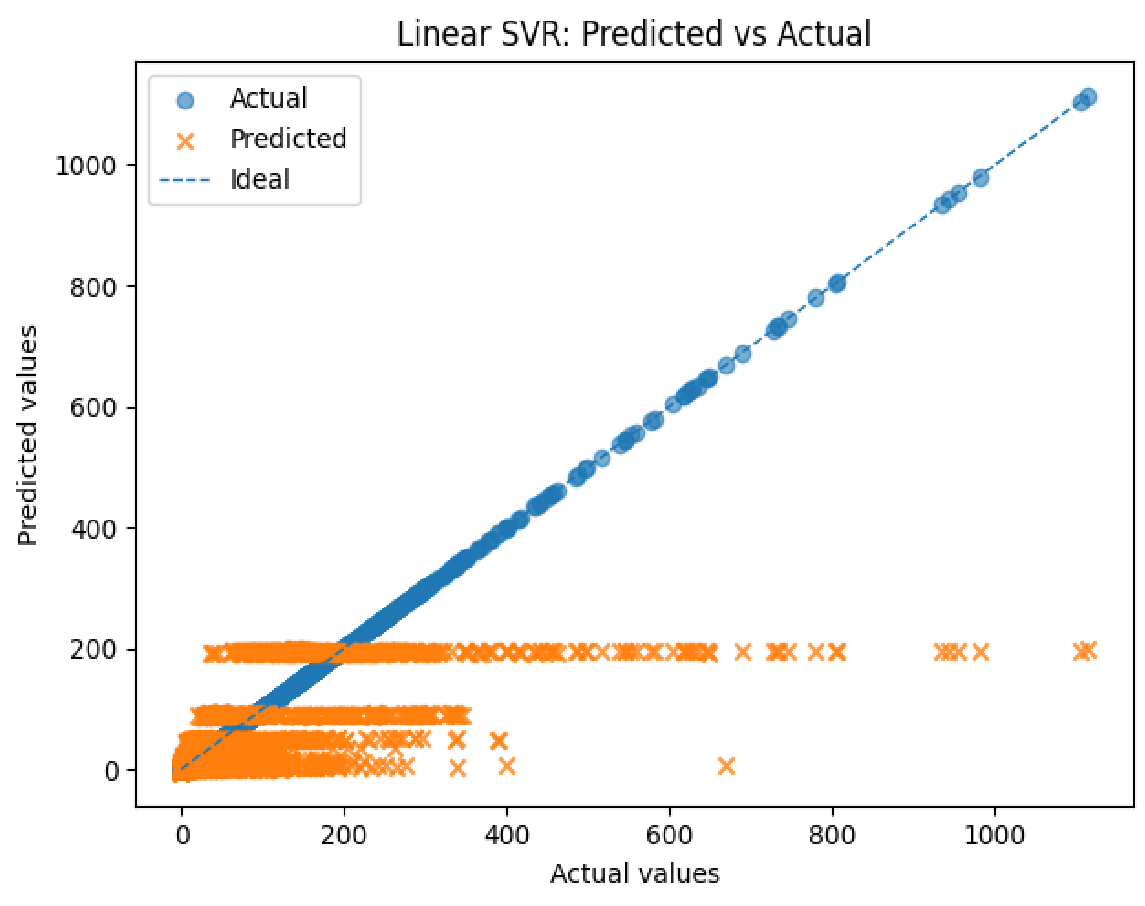
Top 10 important features

Support Vector Regressor

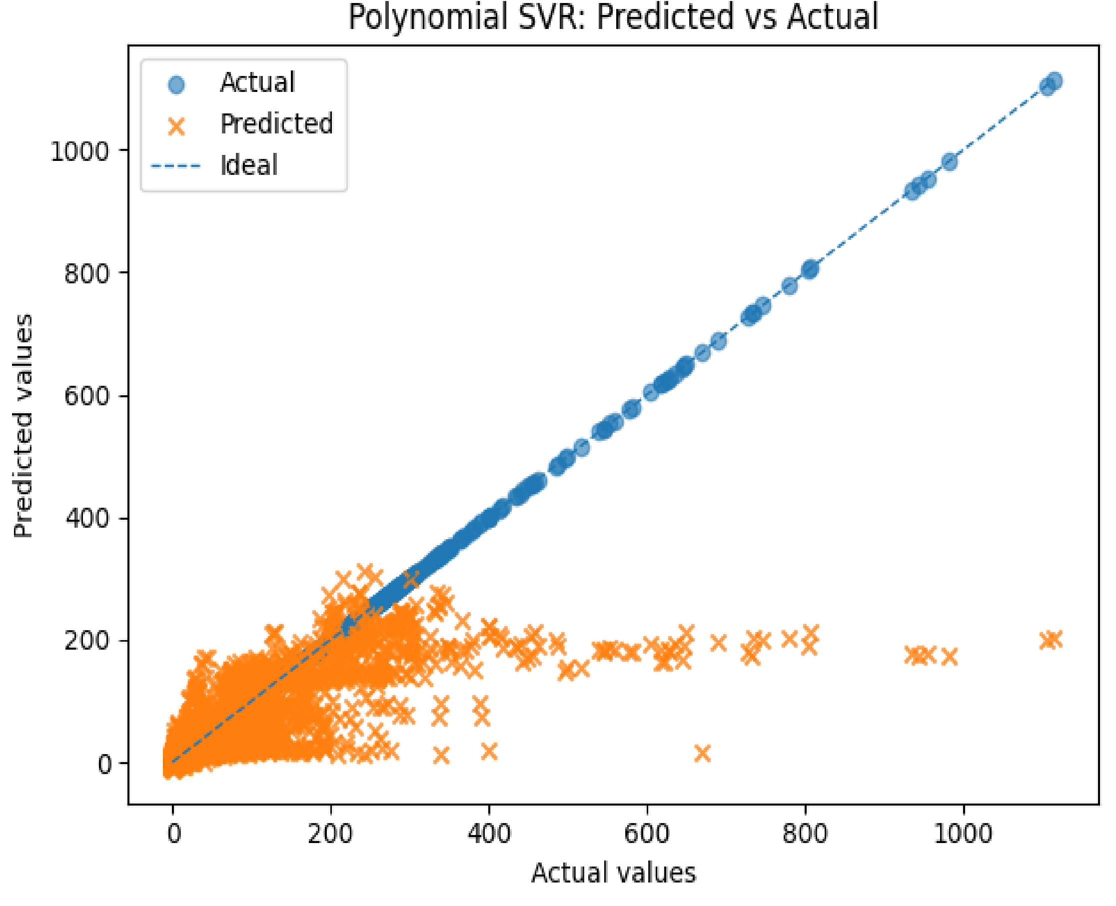


Parameter	Linear	Radial (RBF)	Polynomial
C	10	10	10
gamma	-	0.1	'scale'
degree	-	-	2
kernel	'linear'	'rbf'	'poly'
Evaluation Metrics			
RMSE	55.3810	29.9740	49.0354
R^2 Score	0.4595	0.8417	0.5763
MAE	21.4257	9.5529	16.5027
Time to run	8.00 min	16.08 min	20.24 min

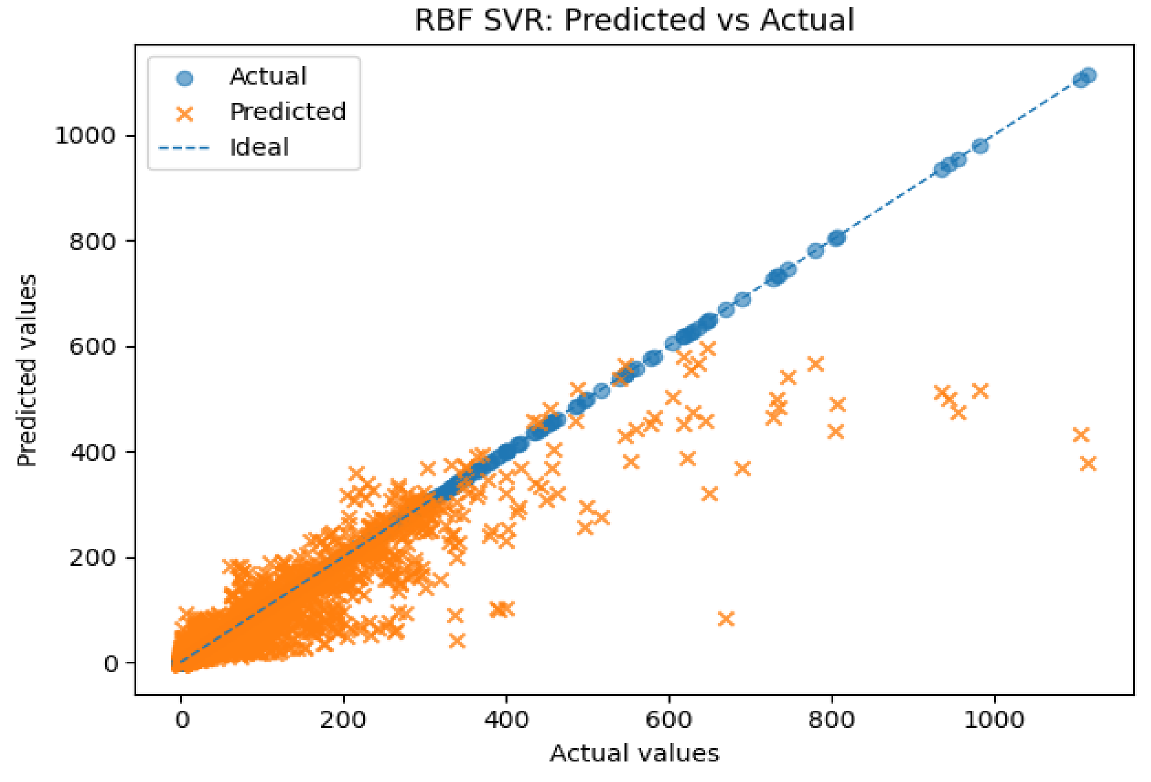
Key points: RBF best performance, polynomial slowest, linear fastest but least accurate.



Polynomial SVR: Captures some non-linear relationships but still struggles with higher values. Moderate improvement over Linear SVR but limited generalization.



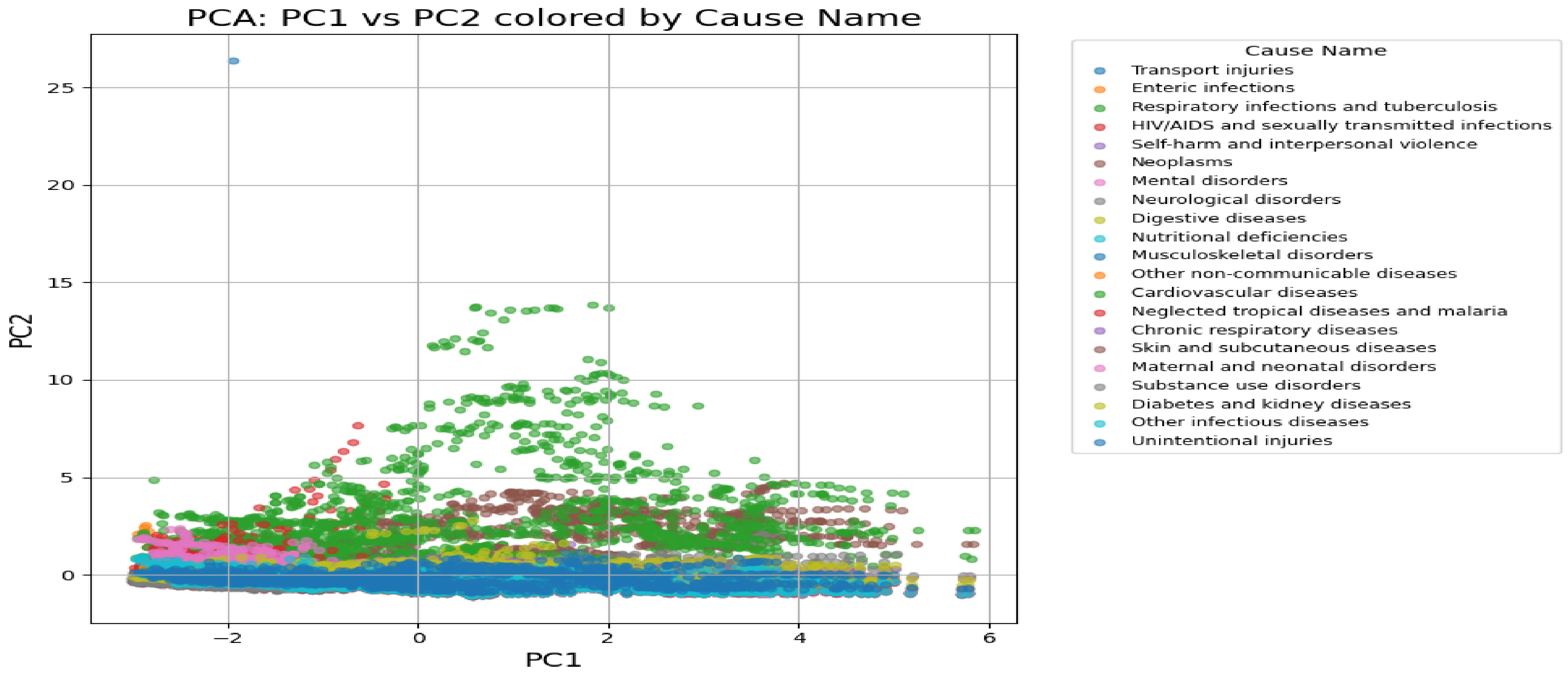
Linear SVR: Shows severe underfitting. The model fails to capture the non-linear patterns in the data, leading to significant prediction errors.

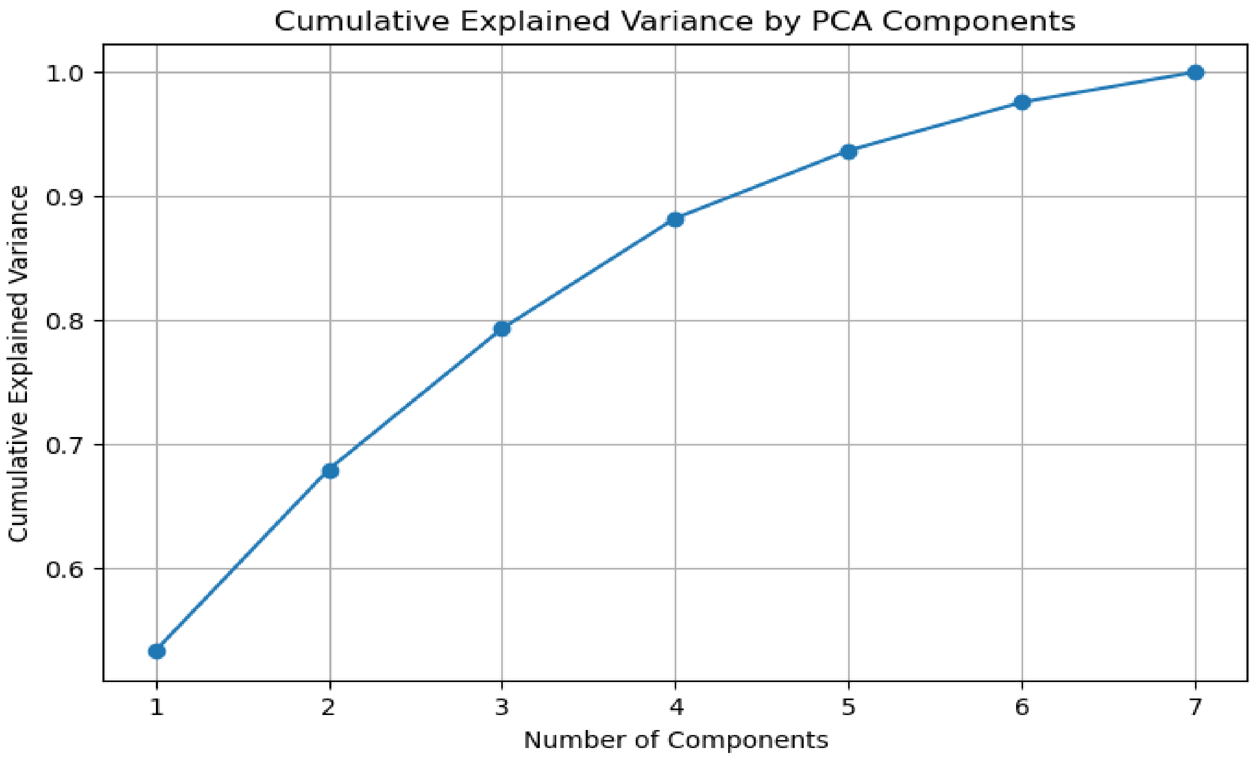


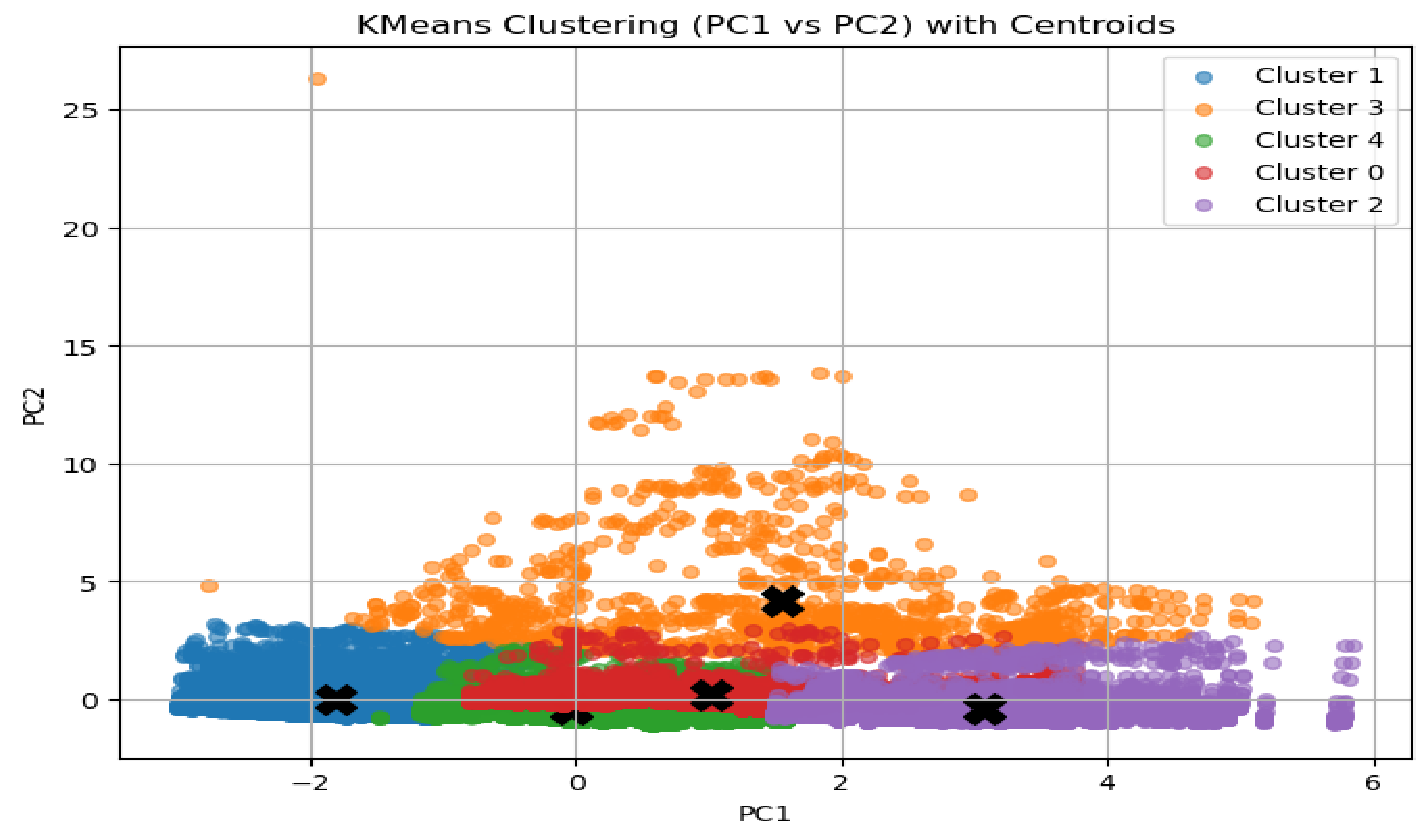
RBF SVR: Performs better than both Linear and Polynomial kernels in modeling complex patterns. However, there are still notable errors at higher target values, indicating the need for better parameter tuning or data normalization.

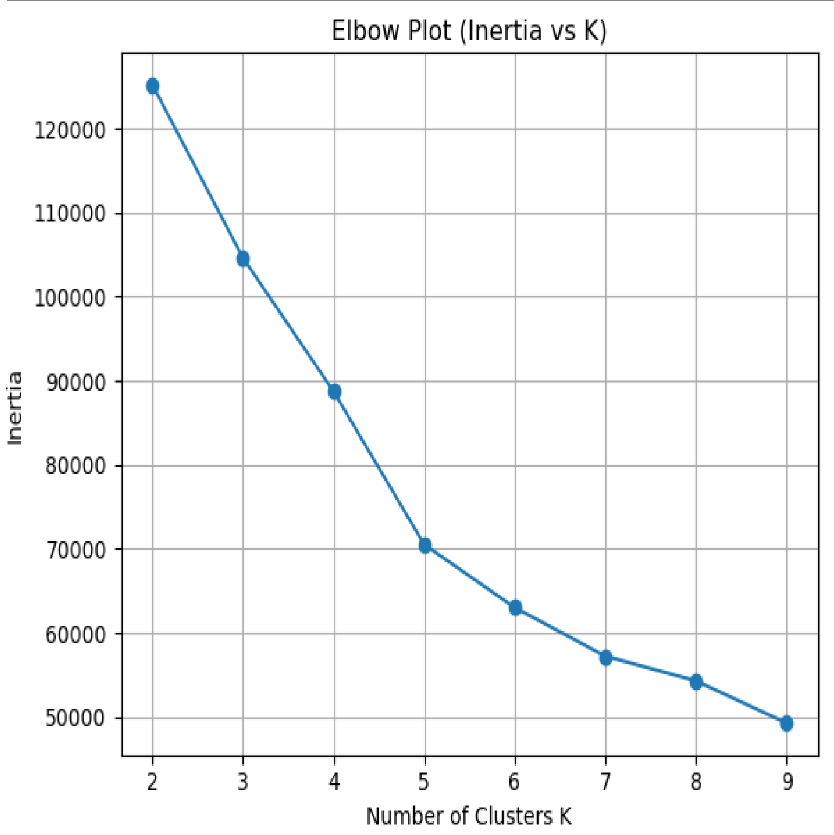
Unsupervised Learning - PCA Analysis

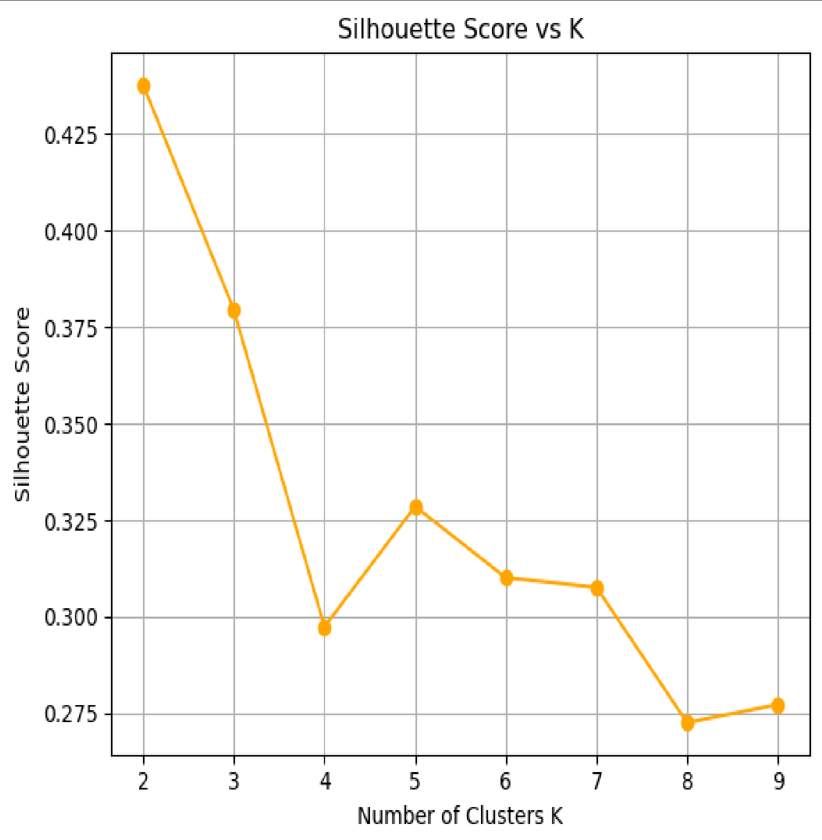
PCA 1	PCA 2
Health Expenditure per Capita (0.47)	Mortality Rate (0.95)
Physicians per 1000 (0.44)	Beds per 1000 (0.23)
Happiness Score (0.42)	Corruption Index (0.08)
Percent GDP Spent on Health Care (0.37)	Physicians per 1000 (0.07)
Beds per 1000 (0.32)	Percent GDP Spent on Health Care (0.05)
Mortality Rate (0.02)	Health Expenditure per Capita (-0.08)
Corruption Index (-0.42)	Happiness Score (-0.16)











Unsupervised Learning - Clustering Results

K-Means Clustering (K=5)

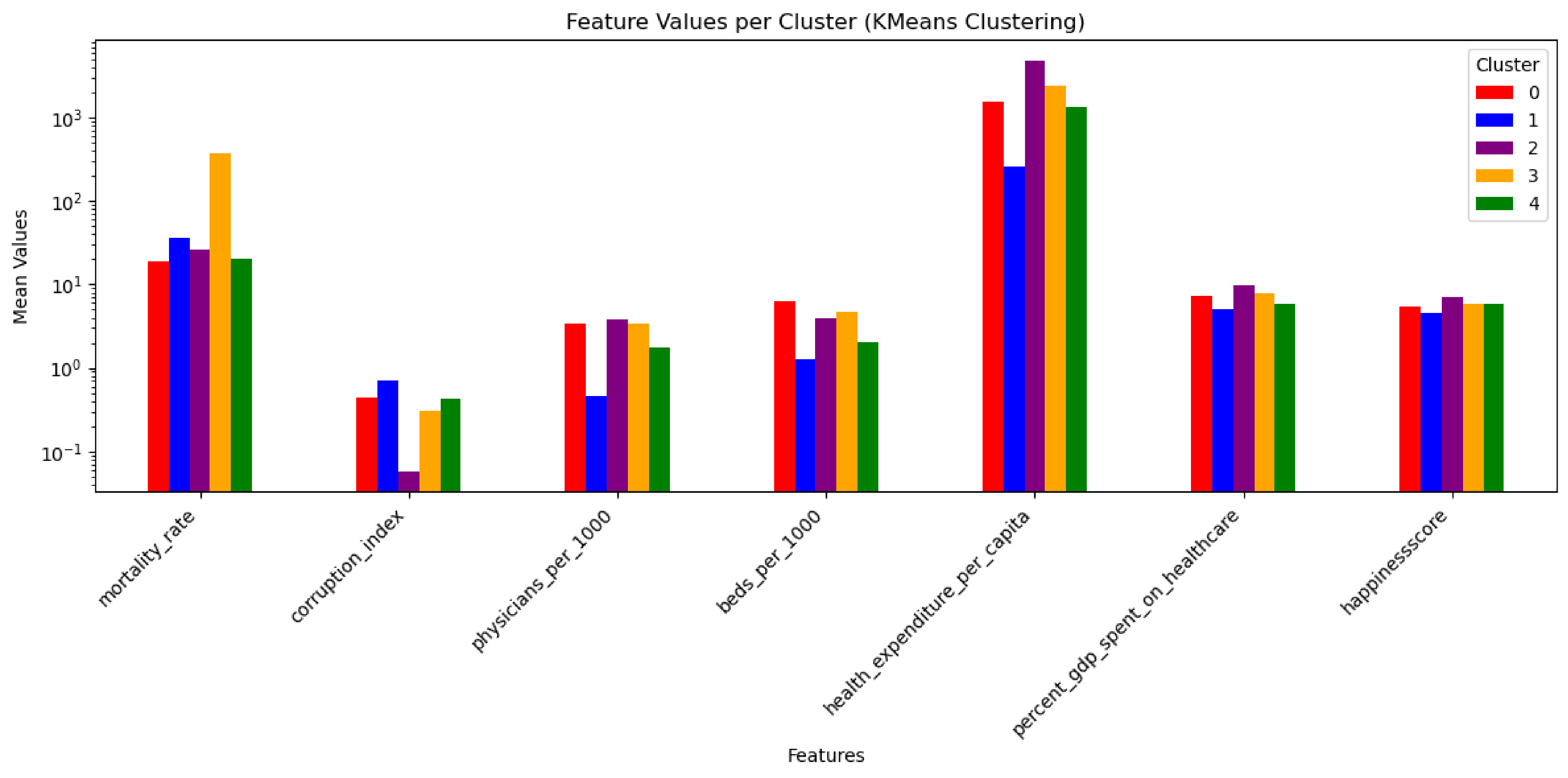
Chosen by elbow and silhouette methods, indicating optimal number of clusters.

Cluster characteristics

Distinct patterns observed in feature values across the 5 clusters.

Key findings

Identified natural groupings of countries based on their mortality-related characteristics.



Cluster 0	**Upper Middle-Income Countries(Maldives, Poland)**
Cluster 1	**Low Income Countries(India, Africa, etc.)**
Cluster 2	**High Income Countries(Western Europe, USA, UK, etc.)**
Cluster 3	**Developing Countries(Eastern Europe, Central Asia)**
Cluster 4	**Developing Countries(Peru, Southeast Asia)**

Limitations & Challenges

Data Quality & Gaps

Incomplete or inconsistent data sources were a challenge. Regional data often lacked uniformity.

Model Extrapolation

Predicting future mortality trends is complex. Models do not factor in unforeseen global events.

Evolving Factors

Socio-economic and health factors change constantly. This impacts long-term model reliability.

Future Work & Roadmap

Expand Data Sources

Integrate diverse datasets for comprehensive global coverage.

Enhance Model Accuracy

Refine algorithms, explore ensemble methods for better predictions.

Conduct Deeper Analysis

Investigate regional nuances and new demographic factors.

Inform Policy Actions

Translate insights into actionable public health strategies.

Conclusions: Key Learnings & Outlook

Actionable Insights

Healthcare access and economic stability critically influence mortality rates. Targeted interventions are essential for positive change.

Model Robustness

The RBF kernel SVR model demonstrated superior predictive performance. It offers reliable forecasting capabilities.

Policy Implications

Data-driven policies can significantly impact public health outcomes. Continuous data quality improvement is vital.

Q&A

Thank you for your attention. We welcome any questions you may have about our research.

Please feel free to contact us for further details or potential collaborations.