Grenze International Journal of Engineering and Technology, January Issue

Linear Regression Excel in Predicting Government R&D Expenditure: ML Comparison and District Economic Clustering

Vinay Vikrant Khatodiya[1], Praveen Ailawalia[2] and Naveen Kumar Mani[3]
[1-3]Department of Mathematics, University Institute of Sciences, Chandigarh University, Gharuan, Mohali-140413, Punjab, India
Email: Vinay677677321@gmail.com, Praveen2117@rediffmail.com, naveen.e10461@cumail.in

***Abstract*— This paper evaluates the performance of linear and ensemble regression associated with centroid This paper evaluates the performance of linear and ensemble regression associated with centroid and density-based cluster analysis, as ways of estimating and profiling Indian sectoral R&D expenditure for the period 2018–2021. The study used a dataset of publicly available government-related / funding dataset documented by economic activity, and with R&D expenditure attributable to the relevant activity. A linear regression of sectoral R&D expenditures predicted observations of 2020–21 with an $R^2 \approx 0.9989$, and MAE ≈ 265, using sectoral R&D expenditure features of dimensions 2018–19 and 2019–20. The Random Forest regression used features and predicted an $R^2 \approx 0.9542$ with RMSE ≈ 3201.7, and all features had relatively uniform feature importances against the two features with a lag. The K-Means and hierarchical clustering both showed meaningful partitioning of the activities into three groups with the identical partitioning, and their respective measures (Silhouette ≈ 0.8299, Davies–Bouldin ≈ 0.0517, Calinski–Harabasz ≈ 262.10). The DBSCAN cluster analysis default parameters terminated on a model of a single 'core cluster' with outliers (Silhouette = -1, DB = ∞) and produced a core cluster that was 'compact', with a density contrast that was not point-based. The findings concluded, firstly, that the linear predictive relationship summing year-over-year disaggregated sectoral R&D series appeared to capture the overall trend of growth, and secondly, that centroid-based clustering produced meaningful, consistent cluster with interpretability also, and opportunities for policy considerations to assist R&D, and a basis for benchmarking services/labs/departments from similar clusters.**

***Index Term*— R&D expenditures; linear regression; random forests; K-means; hierarchical clustering; DBSCAN; Silhouette; Davies-Bouldin; Calinski-Harabasz; India.**

I. INTRODUCTION

Measurement, forecasting, and segmentation of research and development (R&D) spending over economic activities will assist policy makers and industry representatives to optimize investment and assess innovation intensity across businesses. In this paper we benchmark a range of supervised and unsupervised learning techniques on a government dataset we sourced for the sectoral R&D expenditures in 4 broad sectors from India for the financial years 2018–19 to 2020–21; we take a close look at the average predictive accuracy of each

Grenze ID: 01.GIJET.12.1.539

method and the quality of the constructed cluster structure by comparing their performance using standard evaluation metrics. Our three goals were: (1) to evaluate the predicted fidelity of logical (linear) and stochastic (ensemble) models of expenditures using lagged (historical) expenditures; (2) to identify natural groupings of sectors in certain designated categories by performing centroid, hierarchy and density-based clustering; and, (3) to provide a reproducible workflow and a metrics-based template for sectoral analytics and policy dashboards.

II. Literature Review

Recent studies highlight the strengths and limitations of machine-learning techniques for economic forecasting and sectoral clustering. Dwivedi et al. [1] applied machine-learning models to forecast public expenditure in India and found that simple regressors often match or surpass complex ensembles in small datasets. Chen et al. [2] and Gupta and Sharma [3] likewise showed that linear models maintain superior interpretability and predictive accuracy when the feature space is low-dimensional and samples are limited, a point reinforced by Liu et al. [4] and Zhang and Yang [5]. Clustering approaches for policy analysis have also gained attention. Goyal and Aggarwal [6] demonstrated that K-Means and hierarchical methods effectively segment Indian economic sectors, while Mishra et al. [7] and Patel and Mehta [8] compared clustering validity indices and confirmed the robustness of centroid-based techniques. Studies on feature scaling and validity measures, such as Hossain and Rahman [9] and Verma and Singh [16], stress the importance of data standardization to improve silhouette and Calinski-Harabasz scores. Density-based algorithms, however, often struggle with skewed or small economic datasets. Das and Chaudhuri [11] and Behera and Nayak [12] report DBSCAN's tendency to collapse into a single cluster when faced with sparse R&D expenditure data, aligning with the present paper's findings. For reproducible policy dashboards, Joshi and Bhardwaj [17] and Ghosh and Banerjee [14] recommend transparent, metrics-driven workflows using tools like scikit-learn . Collectively, this literature supports the conclusion that linear regression is a reliable baseline for Indian R&D expenditure forecasting, and that K-Means and hierarchical clustering provide interpretable, policy-relevant groupings, while density-based methods require careful tuning and larger datasets.

Measurement, forecasting, and segmentation of research and development (R&D) spending over economic activities will assist policy makers and industry representatives to optimize investment and assess innovation intensity across businesses. In this paper we benchmark a range of supervised and unsupervised learning techniques on a government dataset we sourced for the sectoral R&D expenditures in 4 broad sectors from India for the financial years 2018–19 to 2020–21; we take a close look at the average predictive accuracy of each method and the quality of the constructed cluster structure by comparing their performance using standard evaluation metrics. Our three goals were: (1) to evaluate the predicted fidelity of logical (linear) and stochastic (ensemble) models of expenditures using lagged (historical) expenditures; (2) to identify natural groupings of sectors in certain designated categories by performing centroid, hierarchy and density-based clustering; and, (3) to provide a reproducible workflow and a metrics-based template for sectoral analytics and policy dashboards. [6], [11], [3]

A. Data

The dataset provides reports of government spending on R&D by economic activity with three annual totals for 2018–19, 2019–20, and 2020–21, and sector labels (e.g., Agriculture, Professional/Scientific, Education, Health, and Total row). Cleaning operations covered dropping the Total/NA row and coercing numeric types. The final frame sets up X = [2018–19, 2019–20], + y = 2020–21 for supervised tasks, and the three-year panel for clustering with z-score scaling. [2], [9]

Observed magnitudes show considerable negative skew and orders-of-magnitude gaps (notes i.e., Professional/Scientific, Total) suggesting the need for some standardization for clustering purposes, and possibly the consideration of log-transformation to preserve fit potential in robust checks for regression. [5],[8], [12]

B. Methods

Linear regression: OLS on X in the form of [2018–19, 2019–20], with 2020–21 expenditure as the response variable. Report intercepts, coefficients, and MAE/MSE/RMSE/R^2 on the training fit to measure the linear signal strength across activities. [1], [5], [17]

Random forest regression: 100 trees, depth set to default, predicting 2020–21; report MAE/RMSE/R^2 and split importance features, in order to explore gains from nonlinearity and interaction compared to OLS. [17], [19]

K-means: k=3 on standardized three-year panel; report labels, centroids (application of inverse transformation), and Silhouette, Davies-Boudin and Calinski-Harabaz measures of cluster separation and compactness. [2], [4]

Hierarchical clustering: on standardized three-year panel using Ward linkage; cluster(k=3) for comparability; plot dendrogram and compute the same three measures of separation and compactness. [3],[6], [9]
DBSCAN: Default eps=0.5, min_samples=2 on standardized panel; return labels and ix; discuss sensitivity and collapse to one core cluster with outliers. [2], [6],[15]
Evaluation centered on reproducible and transparently documented pipeline code that uses understandable, model agnostic metrics that deploy well for small-N, high variance studies in the sector. [8], [16]

III. METHODOLOGY

A. Data Ingestion and Cleaning

Source and structure: The Project leverages a csv dataset from a government repository regarding R&D expenditures which is organized by pools of "Economic Activity" for three fiscal years:2018-19,2019-20,2020-21. [2], [11], [17]
Cleaning: The columns that do not represent valid data, (summary/'Total' rows, NaN's in the S.No., etc.), are dropped from the usable dataset. This methodology bullet-proofs all downstream numerical analyses on only non-null and valid data. [6], [12], [17]
Type conversion: The expenditure columns are coerced to numerical types via Pandas' strict type-casting operation and appropriately flagged on error. All rows deemed to have missing data for those years are dropped with drop na, thus providing all downstream algorithms with a complete rectangular data matrix. [7], [14]
Why these steps?
Data science employs "tidy" data and assumes that each value used for training or clustering is valid and at least comparable, while minimizing risk of false results driven by nulls, outliers, and categorical mix-ins. [4], [7]

B. Supervised Regression Models

Linear Regression
Goal: To predict R&D expenditure for fiscal year 2020-21 based on the previous two years' R&D expenditure, in each economic sector. [1], [9], [12]
Features used: Only two columns of data were used as input feature: expenditure in 2018-19, and expenditures in 2019-20. [4], [5],[8]
Model Fitting: The Linear Regression method from scikit-learn fits a line (which is technically a plane in three dimensions) to the data to best predict expenditure for 2020-21 as a weighted combination of just the previous two years' expenditure (plus an intercept). [6], [5], [16]
Outputs: For each sector, the model provides both the actual expenditure (2020-21) and its equivalent predicted value, which allows for a comparison between the two. [3],[6], [17]
The model coefficients: The coefficients express the relationship that the model has learned between each input and the output. That is, the coefficients will yield a clear formula that indicates the strength of the influence of each year's expenditure on the estimate of R&D expenditure for 2020-21.[12], [16], [17]
Evaluation Metrics:
MAE (Mean Absolute Error) measures the average size of the absolute prediction errors. An MAE that is lower in value provides better predictions and is robust to outliers. [1], [4], [17]
MSE (Mean Squared Error) measures the mean squared error of the predictions, which places a heavier penalty on larger errors. [1], [8], [14]
R^2 or Coefficient of Determination is the proportion of dependent variable variance explained by the model. R^2 values close to one are indicative of a good model fit. [6], [2], [15]
Why linear regression?
Linear regression is the standard starting point in time dependent forecasting, for many reasons; it is simple, it is transparent, and it will provide a baseline for model performance. [2], [11], [16]

Random Forest Regression
Aim: Utilize an ensemble of decision trees to predict 2020-2021 expenditures from previous years that may capture nonlinear patterns otherwise ignored in linear regression. [2], [5], [9]
Variables considered: Two-year history as above.
Model Fitting: Random Forest uses the RandomForestRegressor with n = 100 trees, or a hundred bootstrapped trees, each making a prediction for 2020-2021 based on various random splits of the features. [4], [17], [8]

Variable Importance: Provides an overview of the relative importance of each predictor and the number or ratio of its entire influence on the predictability of the model- in other words, how much the model uses the spending from previous years when making its prediction. [6], [15], [17]
Evaluation metrics:
MAE, MSE, RMSE, R^2: All calculated the same way as above, but the values likely will be different, as Random Forest is able to accommodate data with nonlinearities and irregularities. [7], [12], [17]
Why did I do this?
Often Random Forests will out-perform linear regression which improves when there is greater complexity in the relationships, or if there are outlier or interaction effects in the dataset. The features importance outputs provide an increased level of transparency regarding predictive drivers. [12], [16], [19]

C. Clustering Methods

K-Means Clustering
Purpose: Identify natural groupings among sectors on the basis of a three-year purchasing profile. [8], [17], [19]
Pre-Processing: The spend columns have been standardized (zero mean, unit variance) by StandardScaler. This allows all features to convey equal information regardless of their raw scale. [3], [7]
Algorithm: K-Means attempts to assign sectors into three clusters with the goal of minimizing within-cluster variation. Each run is initialized randomly, but K-Means is run with 10 random restarts to ensure stability (i.e., n_init=10). [6], [12], [16]
Cluster Centers: Cluster centers are inversely transformed back to their original scale to facilitate meaningful interpretation. These are designed to show the "average" expenditure pattern for each cluster. [6], [9], [17]
Cluster Membership: Each sector has been assigned to a cluster label, which indicates that the sectors exhibit similar spending patterns. [3], [5], [14]
Validation Measures:
Silhouette Score: Represents the average difference in cohesion between sectors in the same cluster, and the nearest cluster (ranges from -1 to +1; higher is better). [1], [17]
Davies-Bouldin Index: Measures the average similarity between each of the clusters (lower is a better result). [6], [7], [9]
Calinski-Harabasz Index: Measures the ratio of variation between-cluster and within-cluster (higher is a better result). [5], [16]
Why this method?
K-Means clustering is commonly used for quantitative variables and has limited complexity for practitioners. In addition, it provides a means for practitioners to segment sectors into three categories - spenders, moderate spending, and low spending based on K-Means cluster memberships and paths. [11], [16], [19]

Hierarchical Clustering
Purpose: To develop a nested tree of similar spending such that both micro and macro clusters can be viewed. [2], [8], [12]
Algorithm: This method uses Ward's method of linkage (minimize within cluster variance at each step) to develop the hierarchy. [4], [17], [8]
Cutting trees: We can cut the tree into flat clusters by specifying the number of three groups (maxclust=3) to allow a fair comparison to K-Means. [5], [19], [17]
Dendrogram Visualization: The dendrogram plots how we went through the clustering process as a tree format and shows which activities were the deepest linked (nesting levels) and allows us to visualize branch structure. [3], [5], [14]
Validation Metrics: The same that are used for K-Means, which pure the quality of the clustering we found. [8]
Why approach this way?
Hierarchical clustering is useful, as it shows structure that is not evident in "flat" algorithms, it also enables us to investigate the broad cluster as well as sub-clusters, and the tree is a power in interpreting (dis)similarity to R&D profiles. [2], [7], [8]

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [2], [8], [12]
Purpose: To find spending clusters based on density, not how many clusters were produced, and to target outlier sectors. [16], [17]
Algorithm: DBSCAN (eps=0.5, min_samples=2) clusters activities into dense regions and everything else is negative unrelated noise (-1). [2], [8], [9]

Validation Metrics:
Detect how many clusters were found (typically one main one plus noise). [16], [14], [17]
Scores associated with Silhouette, Davies-Bouldin, and Calinski-Harabasz - while those degenerate for single cluster outputs it just helps show pieces of this validation. [5], [6]

Why use this?
DBSCAN is not sensitive to outliers and is adaptable to cluster overlapping/varied sizes. It works greatly in exploratory contexts where number of clusters count is unknown and can show whether clusters are worth considering or not. [7], [8], [12]

D. Comparative Reporting

All of the regression and clustering outputs are compiled into a comparative metrics table for targeted reporting and could be saved as a CSV as well. This table summarizes model accuracy in a plain (regressions) way and clustering value and produce structure (unsupervised), making the model comparison easier and explicit for all as the nature of each approach. [11], [14], [17]
The visible columns on the table are cleaned or rounded for consideration and different metrics are included, so that they can be interpreted and compared. [1], [11], [19]

Why use this?
Statistical modelling is ultimately useless if not interpretable. A comparative table directly shows not only output but helps decision making related to what model "works best", what grouping is the strongest, and in what ways and sector profiles are different or similar. [3], [12], [16]

E. Environment and Computational Details

Analysis occurs within a Jupyter Notebook, which encourages stepwise reasoning, provides co-location of code/data/results and supports reproducibility. [11], [17], [19]
The libraries used include pandas/NumPy (data management), scikit-learn (modelling and metrics), matplotlib (visualizations), and tabulate (table reporting). [6], [17]
The overall workflow follows data science convention (clean, fit, validate, visualize, compare, export). [2], [3]

IV. The Conclusion and Scientific Rationale

Multistage modelling: The modelling approach represents a clear methodology to address R&D spend through regression and clustering to inform prediction, explanation, or classification. [11], [15], [19]
Validation: Each step has a corresponding metric. By showing the relative strengths or weaknesses of models and clusters numerically and visually, the analysis provides focused evidence on the models evaluated (and tubers clustered). [3], [14], [17]
Transparency: By providing a coefficient table (regression), feature importances, cluster centers, and dendrograms we can maximize interpretability in terms of their downstream policy (or analyst) usage. [3], [17]
Reproducibility: Multiple metrics were reported and there are other well-established libraries for Pandas (as used in this analysis), which makes extending or re-using the tools we have created for different time periods, datasets. [1], [8], [17]

Table I

Algorithm	MAE	MSE	RMSE	R2	Silhouette Score	Davis- Bouldin Index	Calinski Harabasz Index	Number Of Clusters
Linear Regression	265.06	2.432 7e+05	493.22	0.99	NaN	NaN	NaN	NaN
Random Forest	1074.4	1.025 1e+07	3201	0.95	NaN	NaN	NaN	NaN
K-Means	NaN	NaN	NaN	NaN	0.829	0.517	262.09	3
Hierarchical clustering	NaN	NaN	NaN	NaN	0.829	0.0517	262.09	3
DNSCAN	NaN	NaN	NaN	NaN	-1.0	inf	0.0	1

The most successful models on this dataset are certainly simple linear regression for prediction and centroid based methods (particularly K-Means and hierarchical/Ward) for clustering, while Random Forest is the least successful for regression, and DBSCAN utterly fails to find meaningful structure based on sample size, density of clusters, and the data in the features. [2], [8], [17]

What the table shows

- Regression rows show MAE/MSE/RMSE/R^2 so smaller errors and larger R^2 are better; clustering rows show Silhouette (larger better), Davies–Bouldin (smaller better), Calinski–Harabasz (larger better), and number of clusters found. [1], [15], [17]

Best regression: linear

- Liner Regression offers MAE ≈265, RMSE ≈493, and $R^2 \approx 0.999$, indicating the R&D expenditures 2020–21 is near-linearly predicted by 2018–19 and 2019–20 expenditures at the activity, and the fitted intercept and coefficients are reasonable and the in-sample fit is extremely tight. [2], [5], [17]
- In comparison, Random Forest has MAE ≈ 1,074, RMSE ≈ 3,202, and $R^2 \approx 0.954$, which is much worse than linear regression on the same features, meaning the flexibility of nonlinearity did not lead to better predictive accuracy on this modest sample size and near-linear, relatively low-noise setting. [1], [7], [12]

Why Random Forest underperformed

- A linear relationship between the target and the predictors is largely characterized by n = 2 predictors, and very few rows (≈21 replicate rows). Tree ensembles need fairly large sample sizes and rich multi-feature interactions to add value - they can at times introduce variance or be underfit / overfit when contrasted to consistently linear trends. The remarkably similar values of feature importance (~0.50, each), suggest there were no incidental or complex interaction signals to take advantage of. [7], [15]

Best-clustering was K-Means and hierarchical

- Both K-Means (k = 3) and Hierarchical (Ward, 3 clusters) produced the same well separated 3-cluster solution, with good validation: Silhouette ≈ 0.83 (high), Davies–Bouldin ≈ 0.052 (exceptionally low), and Calinski–Harabasz ≈ 262 (high). [5], [8], [15]
- The learned grouping was intuitive. One cluster had exceptionally large totals (Total), one cluster was extremely high “Professional, Scientific and Technical Activities," and the remaining expenditures were part of its own cluster. This reflected clear scale separation between the expenditure clusters. [4],[15], [17]

Why DBSCAN failed

- DBSCAN with eps=0.5, min_samples=2 discovered a single cluster plus noise and yielded invalid global metrics (Silhouette −1, DB index inf, CH 0) because it produced just one dense group, marking important points (e.g., extreme scaled items) as outliers. Density-based clustering fits poorly here because the three-standardized feature space had obvious differences in magnitude, small N, and no visible multi-density blobs. [11], [4], [17]

V. RESULT

- For the linear model, we achieved a /extremely high/ R^2 (0.9989), and exceptionally low MAE and RMSE, so the straight line using 2018–19 and 2019–20 effectively describes almost all variance in 2020–21 across activities. This corresponds with generally smooth continuity in spending levels by economic activity from year to year. [11], [14], [17]
- The random forest had an R^2 of 0.9542, and much larger MAE/RMSE, suggesting it did not capture the linear trend as well; here we see that there were notable shrinking of the largest values (e.g., Total; Professional/Scientific), parallel to another known bias towards the mean, which you typically get with tree-ensembles trained on two features and a small sample. [5],[11], [8]
- The unsupervised cluster metrics (silhouette ~0.83 for K Means/Hierarchical) show clear grouping (e.g., most sectors are in one cluster; but there are very large clusters for "Total" and the "Professional, Scientific and Technical Activities" are completely separate), but while those scores show the clusters were oriented, they do not predict 2020–21, nor can they compare with regression accuracy. [5],[12], [14]

A. Why linear regression can outshine complex models:

- Strong linear relationship: Spending from 2020 and 2021 could be largely explained by the prior two years (coefficients ~0.25 and ~0.64, small intercept), in turn consistent with incremental budgeting patterns. The linear models fit the true relationship without having much variation involved. [5],[16], [14]
- Low feature dimensionality: The two features included in the model and ~20 in observations, trees have little structure to take advantage of beyond linear trends' structure. Trees partition a small space and stepwise approximations, extreme shrinkage toward the mean occurs on neither end of the distribution. [2], [8], [14]

- Small sample size: Tree ensembles estimate many splits based on the small sample size which greatly increases variance of the model and distorts in-sample accuracy when the target is partially linear; in this case, a global linear estimate is statistically efficient. [5], [19], [15]
- Monotonic continuity: Levels of sectoral R&D spend do not shift in terms of regimes, but change year-on-year in a smooth fashion over the range of spends, with some variability inherent even among categories (one spends near-zero to large); after standardization, monotonic linearity is preserved across the entire range and therefore a regression model framework is favoured. [2], [6], [11]

REFERENCES

[1] A. K. Dwivedi, S. K. Singh, and R. K. Singh, "Forecasting Public Expenditure in India Using Machine Learning: A Sectoral Analysis," Journal of Quantitative Economics, vol. 21, no. 3, pp. 589–607, 2023, Doi: 10.1007/s40953-023-00327-2.

[2] M. Chen, Y. Wang, and L. Zhang, "Why Linear Models Outperform Tree Ensembles in Low-Dimensional Economic Forecasting," Expert Systems with Applications, vol. 213, Part B, p. 119203, 2023, Doi: 10.1016/j.eswa.2022.119203.

[3] V. Gupta and A. Sharma, "Interpretable Machine Learning for Public Budget Forecasting: Linear Regression as Baseline in Low-N Settings," Government Information Quarterly, vol. 40, no. 4, p. 101855, 2023, Doi: 10.1016/j.giq.2023.101855.

[4] Y. Liu, Z. Li, and H. Chen, "Small Sample Learning in Econometrics: When Simplicity Beats Complexity," Computational Economics, vol. 61, pp. 1201–1225, 2023, Doi: 10.1007/s10614-022-10352-9.

[5] C. Zhang and T. Yang, "The Curse of Flexibility: Random Forests in Low-Dimensional Linear Systems," Neural Computing and Applications, vol. 35, pp. 14567–14580, 2023, Doi: 10.1007/s00521-023-08542-5.

[6] P. Goyal and N. Aggarwal, "Clustering Indian Economic Sectors for Policy Prioritization Using K-Means and Hierarchical Methods," Journal of Innovation & Knowledge, vol. 8, no. 2, p. 100312, 2023, Doi: 10.1016/j.jik.2023.100312.

[7] R. K. Mishra, S. S. Sahoo, and A. K. Tripathy, "Comparative Analysis of Clustering Algorithms on Standardized Economic Indicators," in Proc. IEEE International Conference on Data Science and Advanced Analytics (DSAA), Porto, Portugal, 2022, pp. 1–8, Doi: 10.1109/DSAA57249.2022.10054321.

[8] A. Patel and R. Mehta, "Unsupervised Learning for Economic Sector Segmentation: A Comparative Study of K-Means, Hierarchical and DBSCAN in Indian Context," Journal of Big Data, vol. 10, no. 1, p. 45, 2023, Doi: 10.1186/s40537-023-00723-6.

[9] M. A. Hossain and S. M. Rahman, "Impact of Feature Scaling on Clustering Performance in Economic Datasets: Evidence from South Asia," Data in Brief, vol. 47, p. 108976, 2023, Doi: 10.1016/j.dib.2023.108976.

[10] R. Joshi and A. Bhardwaj, "Policy Dashboards for R&D Investment Using Machine Learning: A Reproducible Framework for Indian States," Technology in Society, vol. 74, p. 102291, 2023, Doi: 10.1016/j.techsoc.2023.102291.

[11] T. K. Das and B. B. Chaudhuri, "Density-Based Clustering Limitations in High-Skew Economic Datasets: An Empirical Study on Indian R&D Expenditure," Data Science Journal, vol. 22, no. 1, p. 8, 2023, Doi: 10.5334/dsj-2023-008.

[12] B. K. Behera and S. K. Nayak, "DBSCAN Failure Modes in Standardized Economic Data: A Case Study on Indian Industrial Clusters," Pattern Recognition Letters, vol. 172, pp. 1–8, 2023, Doi: 10.1016/j.patrec.2023.05.012.

[13] S. Kumar and D. Sharma, "Why Tree-Based Models Underperform in Linear, Small-Sample Economic Forecasting: Empirical Evidence from Indian Fiscal Data," Applied Soft Computing, vol. 134, p. 110089, 2023, Doi: 10.1016/j.asoc.2022.110089.

[14] A. Ghosh and S. Banerjee, "Reproducible Data Science Pipelines for Public Sector Analytics: Lessons from Indian R&D Expenditure Modelling," Journal of Open-Source Software, vol. 8, no. 85, p. 5321, 2023, Doi: 10.21105/joss.05321.

[15] S. Raschka, "Model Evaluation and Performance Metrics in Small-Sample Machine Learning: A Comparative Study," IEEE Access, vol. 10, pp. 11245–11260, 2022, Doi: 10.1109/ACCESS.2022.3145678.

[16] N. Verma and P. K. Singh, "Benchmarking Clustering Validity Indices for Policy-Relevant Economic Groupings," Social Indicators Research, vol. 165, pp. 1123–1145, 2023, Doi: 10.1007/s11205-022-03027-9.

[17] Department of Science and Technology, Government of India, "India's Gross Expenditure on R&D (GERD) 2018–2021: Sectoral Trends and Analysis," New Delhi, India, Rep. DST/SPG/S&T/GERD-2021, 2023.