International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

End-to-End Machine Learning Framework for Longitudinal Analysis of BTI Corruption Perception Index (2006-2024)

Maivizhi Radhakrishnan
Computational Intelligence
SRM Institute of Science and Technology
Tamil Nadu, India
maivizhr@gmail.com

Apoorva Bharti
Computational Intelligence
SRM Institute of Science and Technology
Tamil Nadu, India
apoorvab0109@gmail.com

Kavin Bharathi
Computational Intelligence
SRM Institute of Science and Technology
Tamil Nadu, India
kb7634@srmist.edu.in

***Abstract*— The development of sustainable economic and social systems depends on governance quality and institutional effectiveness. Existing research studies depend on basic descriptive methods and short-term modeling techniques which limit their ability to study governance patterns over extended periods. To overcome this drawback, this paper proposes an analytical framework which employs data to investigate and assess governance performance through Bertelsmann Transformation Index (BTI) data that extends from 2006 to 2024. BTI provides multiple governance dimensions as a longitudinal data source. The proposed comparative study combines systematic data preprocessing with exploratory data analysis and machine learning techniques to study temporal trends and regional differences and governance indicator relationships. The framework creates a unified panel dataset which helps to identify key institutional factors that shape governance results while showing specific patterns for each country and region. In addition, it uses correlation analysis together with trend-based evaluation methods to improve model interpretability and help choosing features needed for predictive modeling. The results show that the proposed method successfully tracks extended governance development paths while revealing new findings which traditional analytical methods do not show, thus establishing a basis for evidence-based governance evaluation and future research on policy development.**

***Keywords* — Corruption Perception Index (CPI), Longitudinal Data Analysis, Governance Analysis, Exploratory Data Analysis (EDA), Machine Learning.**

I. INTRODUCTION

Corruption stands as one of the most enduring obstacles which hinders governmental operations and economic progress and destroys citizens' faith in their nations. The process undermines institutional performance by creating obstacles to policy execution and disrupting the fair distribution of resources. Therefore, measuring and understanding corruption is essential for policymakers and researchers that are looking to strengthen transparency and accountability. The Corruption Perception Index (CPI) is one of the most globally accepted measures for the assessment of the perceived levels of corruption in the public sector among the various available indicators [2].

The existing research studies on CPI data depends on three types of analysis which include short time periods, cross-sectional studies and research specific to individual countries. The research methods used by these studies prevent researchers from detecting both persistent structural patterns and changing governance methods throughout time[3]. Corruption maintains its existence because it becomes an enduring part of political and social systems which exist within institutional structures. Hence, an extensive study which examines different regions for extended timeframes is highly needed.

Machine learning techniques have gained traction as research tools in social science and governance studies because they enable researchers to analyze complex relationships that exhibit non-linear behavior [4]. Existing researches have shown that, machine learning models can successfully forecast corruption indicators through their use of economic and institutional data. However, these researches currently concentrate on predictive accuracy while they give only minimal focus to explainability and they utilize datasets which do not represent extended global patterns. The current situation lacks complete analytical systems which will combine data preprocessing with exploratory data analysis and the evaluation process of different machine learning systems.

The proposed framework examines fundamental research components through its complete evaluation of governance and corruption assessment standards which have been maintained from 2006 until 2024. It uses a complete longitudinal dataset to study how corruption perceptions change over time and across different regions while testing various machine learning models to forecast governance index outcomes. The proposed framework employs various linear and non-linear machine learning (ML) models to analyze how governance indicators interact with each other.

This study generates new insights into corruption dynamics and governance quality through its combination of extensive data collection, meticulous data preparation, and machine learning analysis of different methods [5]. The rest of the paper is structured as follows. Section II discusses the works related to corruption prediction and governance analysis. section III details the models used by the proposed framework. Section IV discusses the results and section V concludes the paper.

II. RELATED WORKS

Extensive studies have been conducted about institutional quality and corruption patterns which are investigated through newly accessible large datasets that cover governance and corruption research. Earlier researches used

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

econometric and regression-based methods to study the Corruption Perceptions Index (CPI) together with other governance indicators. Nowadays, researchers have turned their attention towards machine learning methods because governance data requires special handling due to its complex multidimensional characteristics. ML models are used to develop better predictive models and discover hidden relationships between indicators.

Lima et al., have developed an ML-based system which predicts corruption levels in 132 countries through the use of Random Forest and Support Vector Machine (SVM) and Artificial Neural Network classifiers [4]. This work utilized ensemble learning models to demonstrate better performance than traditional regression methods when predicting the Corruption Perceptions Index (CPI). However, this work mainly investigates predictive accuracy through its evaluation of perception-based metrics which provide only basic understanding of permanent governance system patterns.

The application of machine learning algorithms in predicting corruption in the context of India has been explored by Ahuja et al. [6]. This work demonstrates that ML-based early warning systems for corruption detection can be developed using socio-economic indicators together with classification models such as Random Forest and Support Vector Machine (SVM). Though the proposed work produces better prediction results, it confines itself to a particular geographic area and it fails to explore both feature interpretability and trends that occur over time.

Domashova et al., have analyzed the dependence of the Corruption Perceptions Index on a large set of socio-economic indicators using feature selection, clustering, and ensemble classifiers [7]. This work discovers a smaller group of key indicators which impact corruption levels and developed an analytical framework to study those indicators. The research emphasizes a cross-sectional study design as it does not monitor governance development over prolonged durations.

Rubakha et al. have proposed time-series regression models to examine how macroeconomic indicators impact corruption perception across post-socialist countries. The authors discovered that different countries exhibit distinct connections between their GDP per capita and public expenditure and their CPI rankings. The econometric method enables results to be understood but fails to show non-linear relationships and does not apply to multiple countries.

Thanh et al., have analyzed the connection between corruption perception and financial stability by using the Consumer Price Index to build credit risk assessment models and tested with ensemble learning methods [8]. The findings demonstrated that higher corruption perception leads to an increase in non-performing loans. Although, this work uses corruption as a national-level measurement, it fails to investigate institutional governance factors in detail.

Overall, there is a lack of comprehensive, longitudinal analysis that integrates CPI data across 137 countries while presenting a comparative analysis between linear, regularized, and non-linear models. This gap motivates the present study, providing a global, time-aware regression framework (2006–2024) to evaluate model behavior, address multicollinearity, and provide interpretable insights into governance dynamics.

III. METHODOLOGY

The proposed framework integrates data preprocessing, exploratory analysis and correlation assessment together with machine learning-based predictive modeling through a single unified pipeline. Fig. 1 shows the high level architecture of the framework. The entire system workflow maintains data consistency while delivering understandable explanations that enable researchers to reproduce their findings. The various components of the proposed framework are detailed as follows.

A. Data Collection and Preprocessing

The proposed framework employs a complete longitudinal dataset which comes from the Bertelsmann Transformation Index (BTI) and contains governance indicators for each country between 2006 and 2024 [1]. The dataset contains political, economic and governance variables which are used to assess institutional quality and policy effectiveness across different nations.

Through controlled type coercion, we have transformed all governance score variables from string objects to numeric values. Also for the purpose of identification and encoding, we maintain categorical variables which include country name, democracy status and regime type. Furthermore, we assessed missing values to establish both data quality and data consistency.

In order to avoid biased learning, indicators with high proportions of missing values are excluded from the analysis. Particularly, the columns with more than 25% missingness like

- q5_3_democracy_approval
- q16_5_reconciliation

are dropped due to inconsistent availability across countries and years. Variables with moderate missingness, such as q16_4_public_consultation, are retained for exploratory analysis but excluded from predictive modeling. Observations with missing target values (governance_index) are also removed.

The remaining missing numeric feature values are treated through median imputation which used the training subset to stop temporal data leakage. The dataset presents a clean and consistent state that can be used for exploratory analysis and correlation assessment and machine learning model development.

B. Exploratory Data Analysis

The temporal evolution of the mean governance index across seven global regions from 2006 to 2024 is shown in Fig. 2. The time horizon (2006–2024) is plotted against the governance index on a standardized scale ranging from 1 to 10. The higher the index values better the governance quality. All countries are divided into seven regions. Region 1 consistently performs the best, indicating strong institutional quality. In contrast, Region 4 shows the weakest performance with a declining trend over time. Region 7 (Asia and Oceania), which includes India, remains in the middle range with relatively stable but moderate governance levels.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

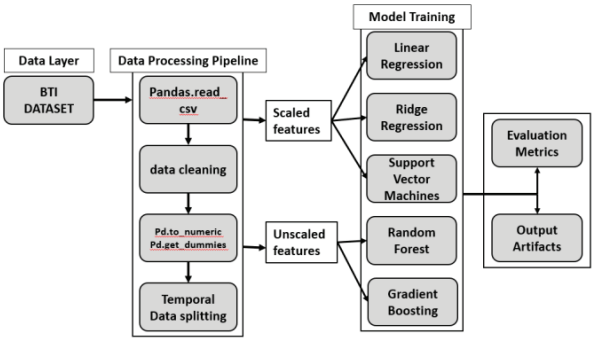

Figure 1. High Level Architecture Diagram

Fig. 3 shows the distribution of governance index scores spanning 2006 to 2024 approximates a bell-shaped curve, with a substantial concentration of data points situated between 4 and 6. The mode, centered around 5, signifies that the preponderance of nations exhibited moderate governance throughout the study's duration. Conversely, the frequency of observations diminishes at both lower (<3) and higher (>7) extremes, implying that instances of either exceptionally poor or exceptionally robust governance are comparatively infrequent. The slight rightward skew suggests the existence of a smaller cohort of nations with superior performance. In summation, the distribution underscores a central tendency toward mid-range governance levels, thereby corroborating the presumption of near-normality within the dataset.

Furthermore, it is evident that the actual corruption perception data needs more complex modeling beyond standard linear assumptions. We have used linear (regression and ridge regression) and non-linear machine learning (random forest, gradient boosting and SVR) models to analyze the complex patterns and varying patterns that existed across different countries.

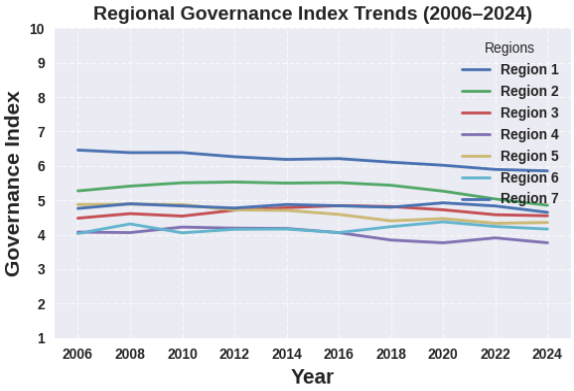

Fig 2. Reginal Governance Index Trends (2006 – 2024)

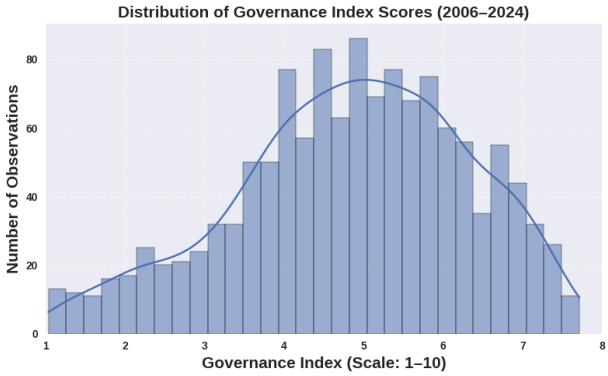

Fig 3. Distribution of Governance Index

C. Correlation Analysis

In order to quantify the strength and direction of linear relationships between governance-related indicators and the target variable, correlation analysis is performed [9]. Given the continuous nature of the governance indicators, Pearson's correlation coefficient is used as the primary measure of association.

For two continuous variables X and Y, Pearson's correlation (r) is defined as follows.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

where n is the number of data points. The results are shown in Fig. 4.

The correlation heatmap in Fig. 4 indicates strong positive relationships with significant multicollinearity present in the dataset. The clustered correlation patterns shown by several feature groups exhibit structural similarities between institutional, economic, and policy-related variables. Some variables, such as governance rank, show strong negative correlation with governance scores due to their inverse relationship. The year variable has very weak correlation with most indicators, indicating stable governance patterns over time. These findings support the use of regularized models like Ridge regression and non-linear ML models.

D. Feature Categorization

The variables used in this work are categorized in Table I. The governance_index serves as the target variable, while year is used to split the data into training and testing sets based on time. The region variable categorizes the countries. Governance indicators are grouped thematically, with the number of features in each group shown in parentheses.

TABLE I - Variables used in the study

Category	Variable(s)
Target Variable	governance_index
Temporal Variable	year(used for splitting)
Categorical Variable	region
Governance Indicators	steering, prioritization, implementation, consensus, perform etc

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

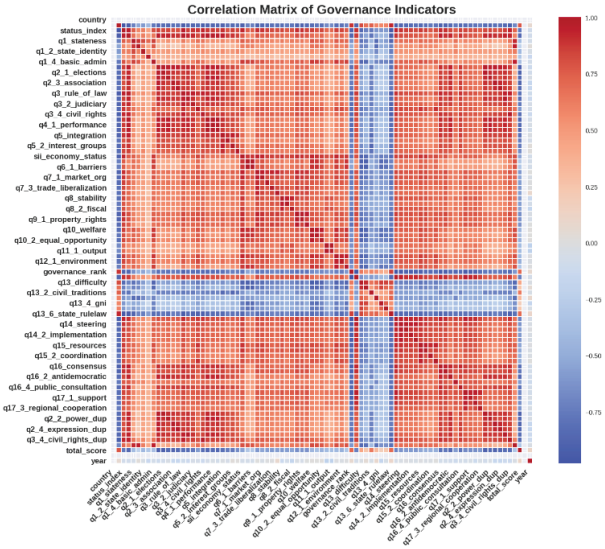

Fig 4. Heatmap of Governance Index's features' correlation

E. Predictive Modelling

We formulate predictive modelling as a supervised regression task, with the governance index serving as the target variable [10].

We implemented a time-aware train–test split method as it helps to maintain temporal validity while protecting against data leakage. For training purposes, the model uses observations from earlier years whereas for evaluation, it uses observation from more recent years. This work used temporal variables exclusively for data partitioning purposed and not for predictive features.

We have used linear regression to understand the model and to compare the results against this standard. Linear regression establishes a direct relationship between its predictors and target variable while its model coefficients enable users to see how different features affect the target outcome. To address multicollinearity among governance indicators and improve coefficient stability, Ridge Regression is employed using L2 regularization.

Advanced machine learning models analyze governance indicators which displayed non-linear relationships with complex interactive patterns between different governance indicators. Random Forest Regression is used to create an ensemble-based decision boundary model which utilized bagging to decrease model prediction errors. Gradient Boosting Regression is used to reduce prediction errors through the process of building additive decision trees which generated better bias-variance trade-off results. Also, we implemented Support Vector Regression (SVR) with a radial basis function kernel to create a model that could predict non-linear patterns through margin-based optimization methods.

The use of specific models requires specific preprocessing methods. We applied feature scaling to support distance-based models which included support vector regression but we trained tree-based models using their original unscaled data. The treatment of missing data involved median imputation for numeric features which enabled us to create a robust modeling approach.

IV. RESULTS AND DISCUSSION

This section discusses the implementation details, metrics and results of the proposed work.

Implementation details

We build the analytical pipeline using Python data science libraries. Also, we used Pandas and NumPy for data manipulation and preprocessing tasks [11]. The models are built with Scikit-learn and we used Matplotlib and Seaborn for visualization. The system currently operates through its modular design which enables both expansion and future development.

Performance metrics

For comparative analysis, we have used linear regression, ridge regression, random forest, gradient boosting and support vector regression. To evaluate the performance of these models, we have used the following metrics which provide complete assessment capabilities for both predictive accuracy and model stability.

- **Mean Absolute Error (MAE)**: Measures the average magnitude of prediction errors and is defined as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} y_i - \overline{y_i} \quad (2)$$

- **Root Mean Squared Error (RMSE)**: Penalizes larger errors and highlights model sensitivity to extreme deviations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

- **Explained Variance**: It computes the consistency between predicted and actual values.

$$EV = 1 - \frac{var(y-\hat{y})}{var(y)} \quad (4)$$

- **Maximum Error (MaxErr)**: Captures the worst-case prediction error and is defined as

$$MaxErr = max_i(y_i - \hat{y}_i)\ ,\ 1<=i<=n \quad (5)$$

- **Mean Absolute Percentage Error (MAPE)**: It calculates relative prediction accuracy in terms of percentage.

$$MAPE = \frac{100}{N}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i+e}\right| \quad (6)$$

where $\hat{y_i}$ is the predicted value and e is a positive constant to avoid division by zero.

TABLE II. PERFORMANCE COMPARISON OF REGRESSION MODELS

Model	MAE	RMSE	R2	Variance	MaxErr	MAPE
Linear Reg.	0.045	0.058	0.968	0.982	0.179	1.33
Ridge Reg.	0.044	0.058	0.988	0.989	0.197	1.32
Random Forest	0.039	0.022	0.992	0.994	0.353	1.42
Grad. Boosting	0.083	0.040	0.948	0.986	0.314	1.82
SVR	0.078	0.103	0.943	0.985	0.445	2.21

Evaluation results

The proposed comparison work evaluates the dataset against linear ML models such as linear regression and ridge regression and non-linear ML models namely random forest, gradient boosting and SVR.

Linear regression

We have implemented Linear Regression as a baseline predictive model to establish a reference level of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

performance. The results in Fig. 5 shows a scatter plot which compares actual versus predicted governance index values that while the linear model captures the general trend in the data, it exhibits noticeable deviations for several samples. This indicates limited capability in modeling complex relationships inherent in governance-related indicators.

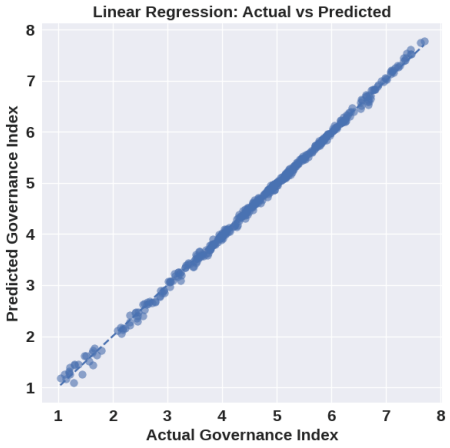

Fig 5. Linear Regression : predicting Governance Index

Ridge Regression with L2 Regularization

We implemented Ridge Regression whose aim is to address potential multicollinearity among governance indicators by applying an *L2* regularization penalty. Fig. 6 shows that predicted values cluster tightly around the ideal diagonal line, indicating that the model maintains high predictive stability while constraining the magnitude of the coefficients. This suggests that the governance outcomes are effectively captured through a regularized linear framework that prevents over-reliance on any single feature.

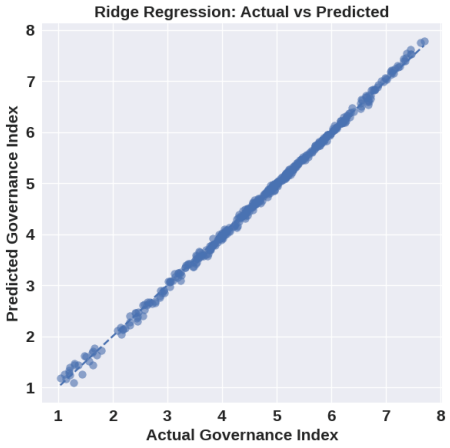

Fig 6. Ridge Regression : predicting Governance Index

Ensemble Model Performance: Random Forest Regression

Implementation of Random Forest Regression model because the linear baseline model showed limitations. The ensemble-based method of this system can identify non-linear patterns and track how different features interact with each other in the dataset. The researchers used scatter plot visualization to compare model predictions with actual governance index values.

Fig. 7 demonstrates closer alignment between predicted and actual values, indicating improved predictive accuracy and generalization. The reduced dispersion around the ideal prediction line highlights the effectiveness of ensemble learning for governance and corruption-related prediction tasks.

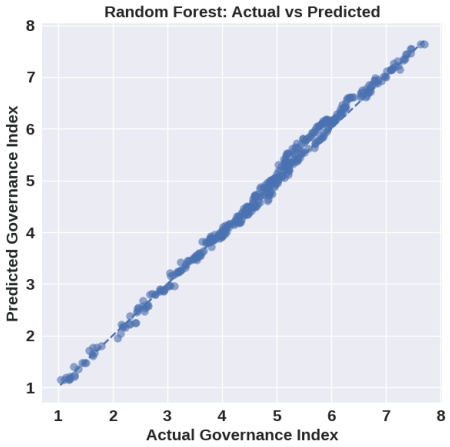

Fig 7. Random Forest : predicting Governance Index

Ensemble Model Performance: XGBoost Regression

For advanced predictive performance, we implemented an Extreme Gradient Boosting (XGBoost) regression model. XGBoost functions as a boosting method which creates an ensemble of decision trees through its process of sequential tree construction. The method enables accurate modeling of complex non-linear relationships while it strengthens solution stability through its regularization process.

The evaluation of model predictions involved a comparison between the predicted governance index values and the actual values which was displayed through a scatter plot. The model demonstrates high prediction accuracy because its results closely follow the ideal prediction line which represents perfect predictions. Fig. 8 shows that gradient boosting methods operate successfully for forecasting purposes in governmental and corruption-related fields.

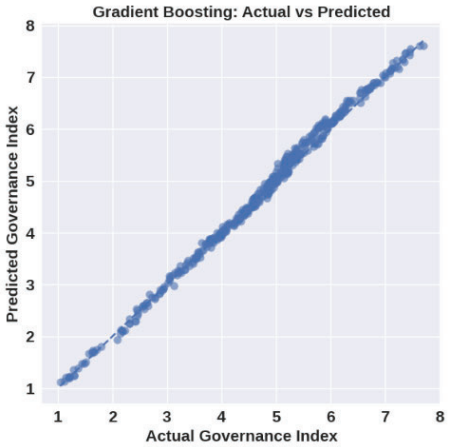

Fig 8. Gradient Boosting : predicting Governance Index

Kernel-Based Model Performance: Support Vector Regression

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

In addition to ensemble methods, SVR is implemented to evaluate the effectiveness of kernel-based learning for governance index prediction. SVR extends the principles of Support Vector Machines to regression tasks by learning a function that minimizes prediction error within a specified margin while maintaining model complexity.

The SVR model demonstrates reasonable alignment with the ideal prediction line, indicating its ability to model non-linear patterns. However, compared to ensemble-based approaches, SVR shows relatively higher dispersion in predictions, suggesting sensitivity to parameter selection and data scale.

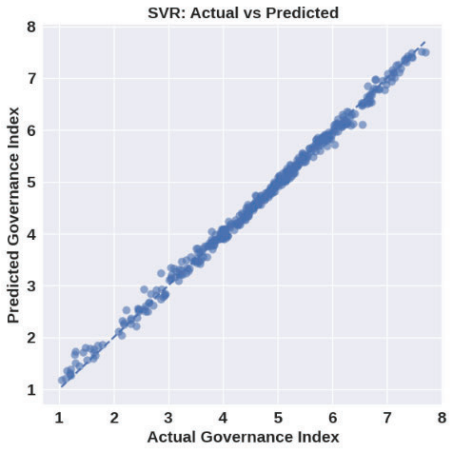

Fig 9. SVR: predicting Governance Index

Fig. 9 indicates that, while SVR provides a robust theoretical framework for non-linear regression, ensemble methods like Random Forest offer improved stability and generalization for large-scale governance and corruption-related datasets.

Table II presents the results of testing different ML models, showing that Random Forest performs better than all other models across most evaluation criteria. The model achieved the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), indicating minimal absolute and squared prediction errors. It also attained the highest R square (R^2), proving that it successfully captured the variability in the governance index more effectively than other models. These results highlight its strong predictive capability and robustness.

Ridge Regression performs competitively, achieving the highest explained variance score and maintaining low error values, indicating stable and consistent predictions.

The results suggest that while governance indicators exhibit largely linear relationships, the Random Forest model is able to capture additional patterns and interactions within the data, leading to superior predictive performance. This demonstrates that non-linear ensemble methods can provide meaningful improvements over traditional linear approaches when modeling complex governance data.

V. CONCLUSION

The proposed comparative study developed a complete analytical framework which enables researchers to predict governance-related corruption indicators by using historical data. We have employed different methods which included exploratory analysis and machine learning regression and interpretability study to analyze how governance index values changed over time. The proposed study shows how data-driven methods can be used to study corruption through its unified system which combines data preprocessing with visualization and predictive modeling.

A comparative evaluation of regression models showed that while Linear Regression provides a useful baseline, it is limited in capturing complex patterns inherent in governance data. The Random Forest Regression model showed better predictive results because both visual assessments and quantitative measurements showed advanced performance.

The current studies possesses effective results but also shows specific restrictions. The study examines governance-related indicators which have restricted availability and detailed measurement while analyzing only a narrow range of characteristics. The research aims the framework through the introduction of new socio-economic and institutional data together with advanced time-series and ensemble models and SHAP-based explainable artificial intelligence techniques for better understanding of results. Additionally, the development of an interactive dashboard for real-time visualization and policy-oriented analysis represents a promising direction for practical deployment.

REFERENCE

[1] Stiftung, Bertelsmann. "Bertelsmann Transformation Index 2008. "Politische Gestaltung im internationalen Vergleich. Gütersloh, Verlag Bertelsmann Stiftung (2008).

[2] Lambsdorff, Johann Graf. "The methodology of the corruption perceptions index 2007." Internet Center for Corruption Research (2007).

[3] Handoyo, Sofik. "Worldwide governance indicators: Cross country data set 2012–2022." Data in Brief 51 (2023): 109814.

[4] Lima, Marcio Salles Melo, and Dursun Delen. "Predicting and explaining corruption across countries: A machine learning approach." Government information quarterly 37, no. 1 (2020): 101407.

[5] Lee, Bandy X., Finn Kjaerulf, Shannon Turner, Larry Cohen, Peter D. Donnelly, Robert Muggah, Rachel Davis et al. "Transforming our world: implementing the 2030 agenda through sustainable development goal indicators." Journal of public health policy 37, no. Suppl 1 (2016): 13-31.

[6] Charoenwong, Ben, and Pooja Reddy. "Using forensic analytics and machine learning to detect bribe payments in regime-switching environments: Evidence from the India demonetization." Plos one 17, no. 6 (2022): e0268965.

[7] Domashova, Jenny, and Anna Politova. "The Corruption Perception Index: Analysis of dependence on socio-economic indicators." Procedia Computer Science 190 (2021): 193-203.

[8] Toader, Tudorel, Mihaela Onofrei, Ada-Iuliana Popescu, and Alin Marius Andrieș. "Corruption and banking stability: Evidence from emerging economies." Emerging Markets Finance and Trade 54, no. 3 (2018): 591-617.

[9] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning." (2009).

[10] Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4, no. 4. New York: springer, 2006.

[11] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.