Deforestation Issue Analysis Using Support Vector Machine (SVM)

Leveraging Support Vector Regression and Exploratory Data Analytics

1. Introduction

Forests are recognised as vital ecosystems that sustain biodiversity, regulate local climates, and function as critical carbon sinks in the global effort to mitigate climate change. Despite their importance, an estimated ten million hectares of tree cover are lost annually, driven by a complex mixture of economic expansion, agricultural conversion, population pressure, and weak governance. This report presents an in-depth, data-centred exploration of deforestation patterns using a curated panel dataset covering multiple countries and years.

The dataset combines socio-economic indicators (gross domestic product, population, illegal lumbering incidents, international aid, corruption indices), biophysical metrics (annual rainfall, tree-cover loss, agricultural land percentage, protected area coverage), and policy-related variables (deforestation policy strictness). The target variable is a continuous deforestation severity score scaled between 0 and 1, synthesising satellite-observed loss and national reports.

Our objectives are three-fold: (i) visualise and interpret bivariate relationships between drivers and deforestation; (ii) build an explanatory regression model using Support Vector Regression (SVR) to capture non-linear relationships; and (iii) translate model insights into actionable policy recommendations. Throughout, we discuss methodological choices, model limitations, and directions for improving both data and analytics so that this work can serve as a template for more granular, region-specific studies.

2. Exploratory Visualisations & Feature Diagnostics

Exploratory Data Analysis (EDA) is the foundation of any sound modelling workflow. Before trusting algorithmic outputs, analysts must examine raw relationships to identify potential signal, confounding, and data quality issues. The following figures offer a multi-angle view of our dataset:

• Six scatter plots illustrate the spread of observations and hint at possible linear or non-linear trends.

• A horizontal bar chart ranks simple Pearson correlations with the target to gauge first-order signal strength.

• A lower-triangle pair plot exposes inter-feature interactions and multicollinearity.

Several immediate observations emerge:

1. No single driver cleanly tracks deforestation. Even the strongest absolute correlation (Protected_Areas_Percent ≈ 0.15) is weak.
2. Large vertical spread in the scatter plots suggests substantial unexplained variance—an early warning that simple linear models may underperform.
3. Some predictors exhibit visible banding (e.g., Illegal Lumbering Incidents clumps at round numbers), hinting at reporting thresholds or digit preference, which injects measurement noise.
4. The pair plot reveals moderate correlations among socio-economic variables (GDP, Population, Aid), raising multicollinearity considerations when fitting parametric models.

Taken together, the EDA frames expectations: we anticipate low to modest predictive power without feature enrichment or non-linear techniques.

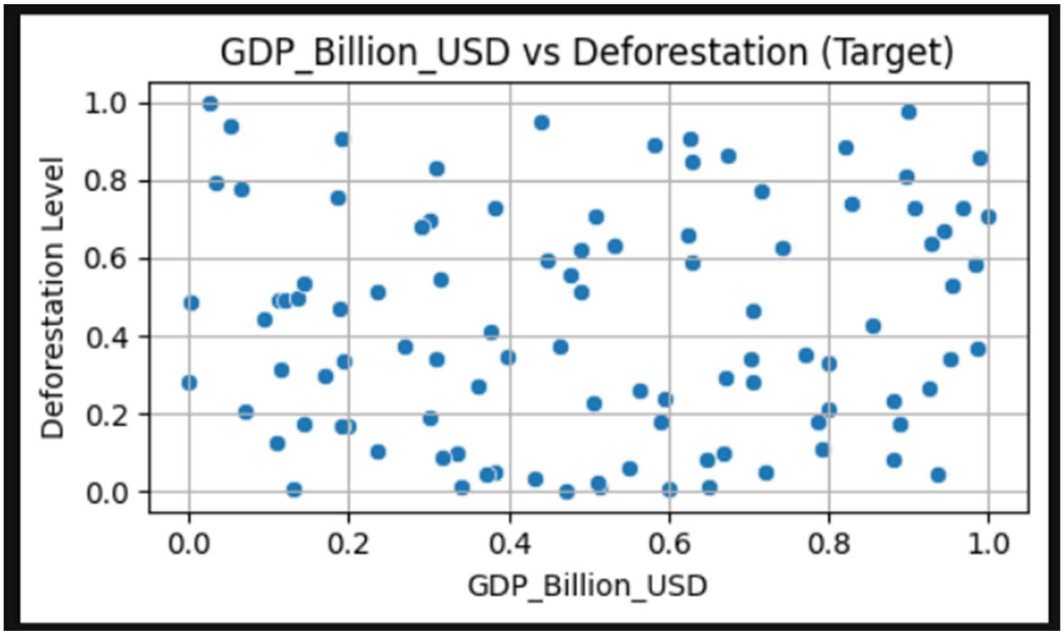

Figure 2-1. GDP vs Deforestation shows broad scatter and negligible linear trend.

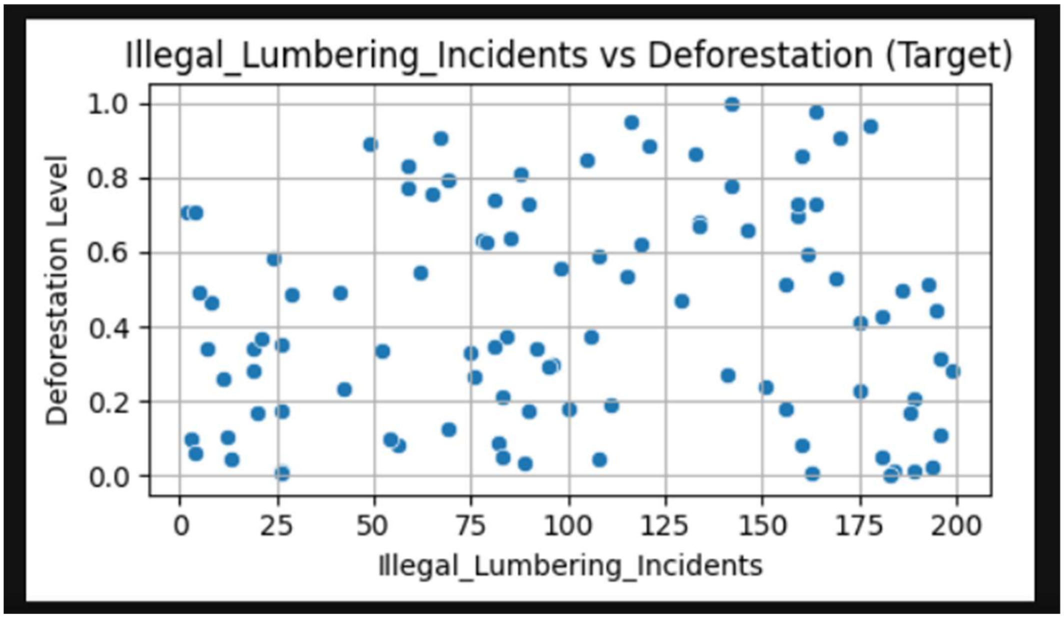

Figure 2-2. Illegal Lumbering Incidents vs Deforestation reveals clusters caused by discrete reporting.

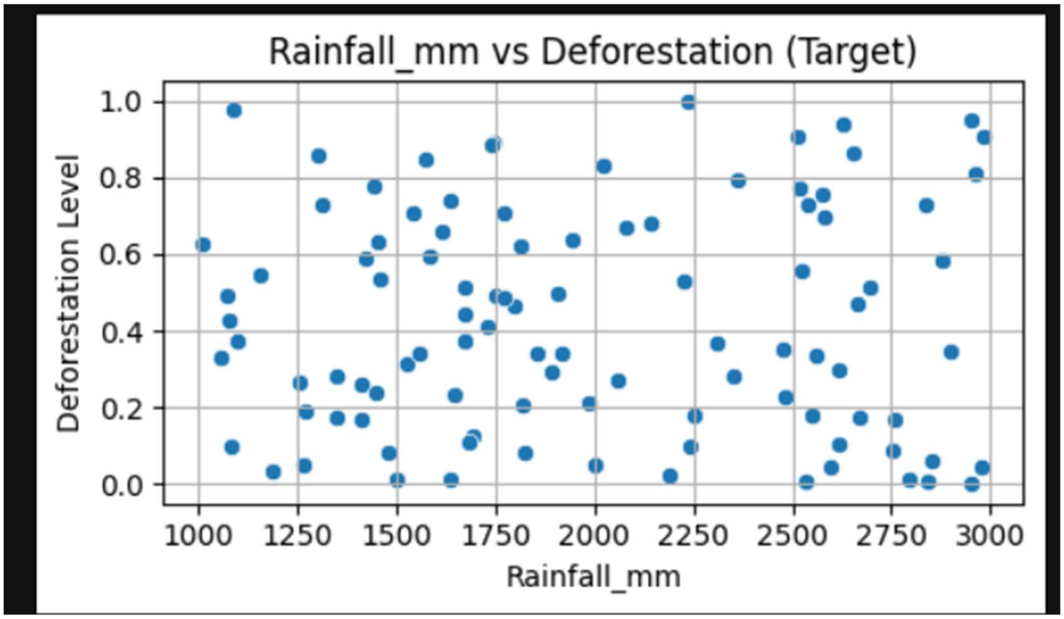

Figure 2-3. Rainfall vs Deforestation depicts absence of monotonic influence.

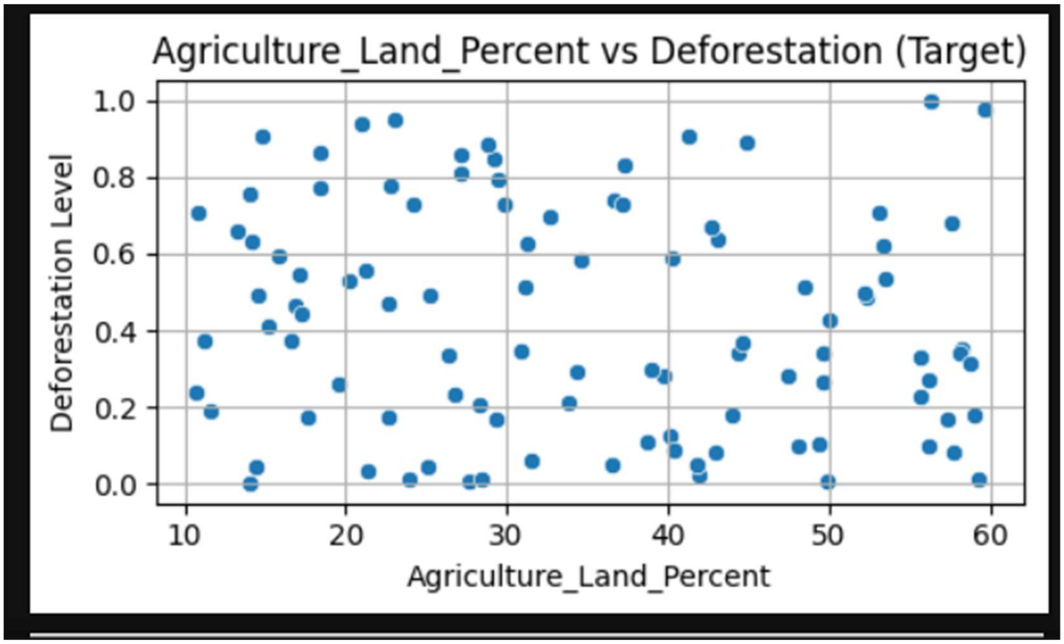

Figure 2-4. Agriculture Land % vs Deforestation suggests weak negative association at high Ag-Land levels.

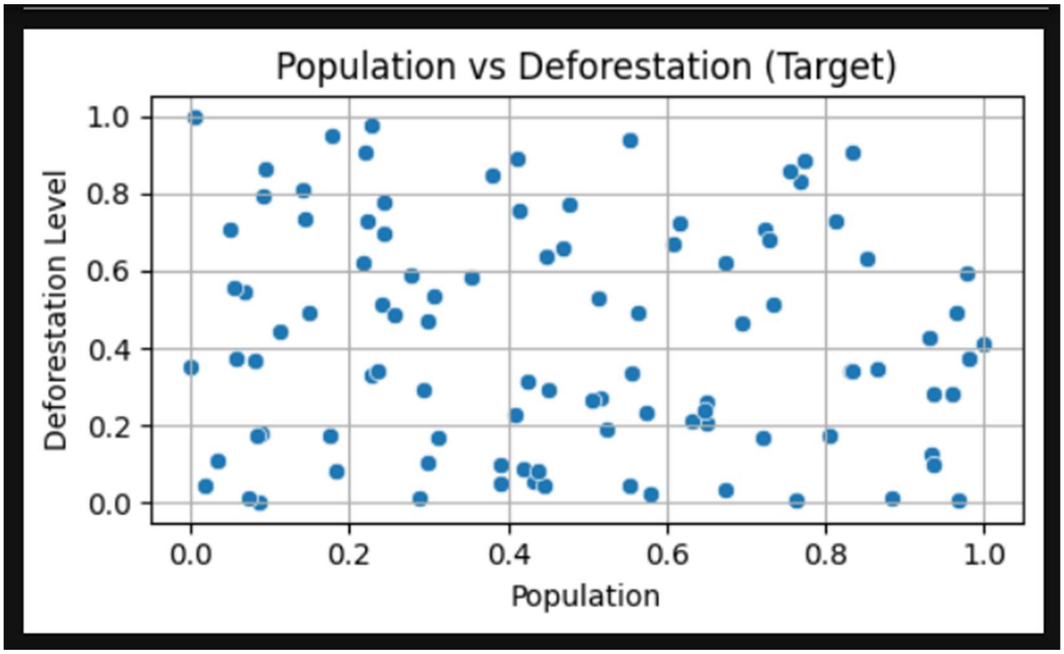

Figure 2-5. Population vs Deforestation displays diffuse cloud—population alone insufficient.

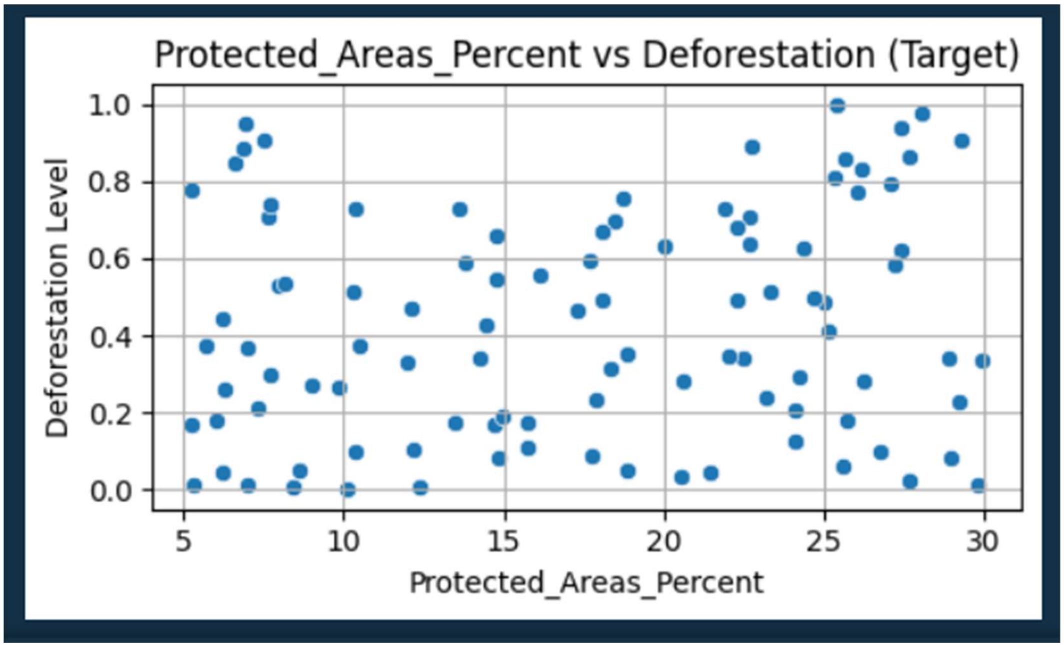

Figure 2-6. Protected Areas % vs Deforestation hints that higher protection corresponds to lower mid-range loss.

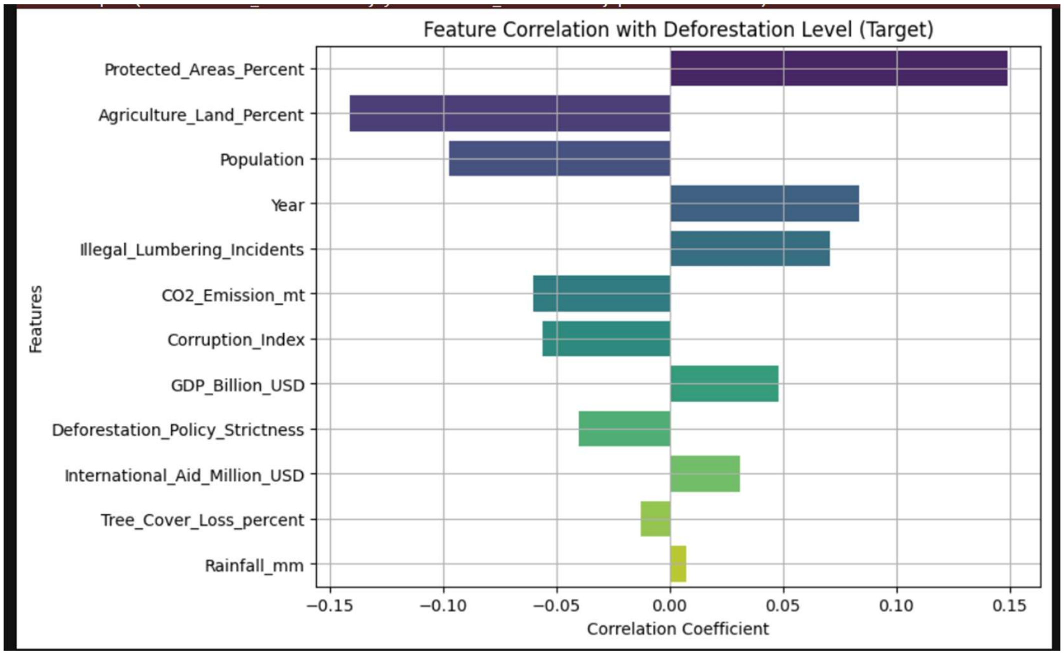

Figure 2-7. Pearson Correlation Bar Chart emphasises that all absolute correlations < 0.15.

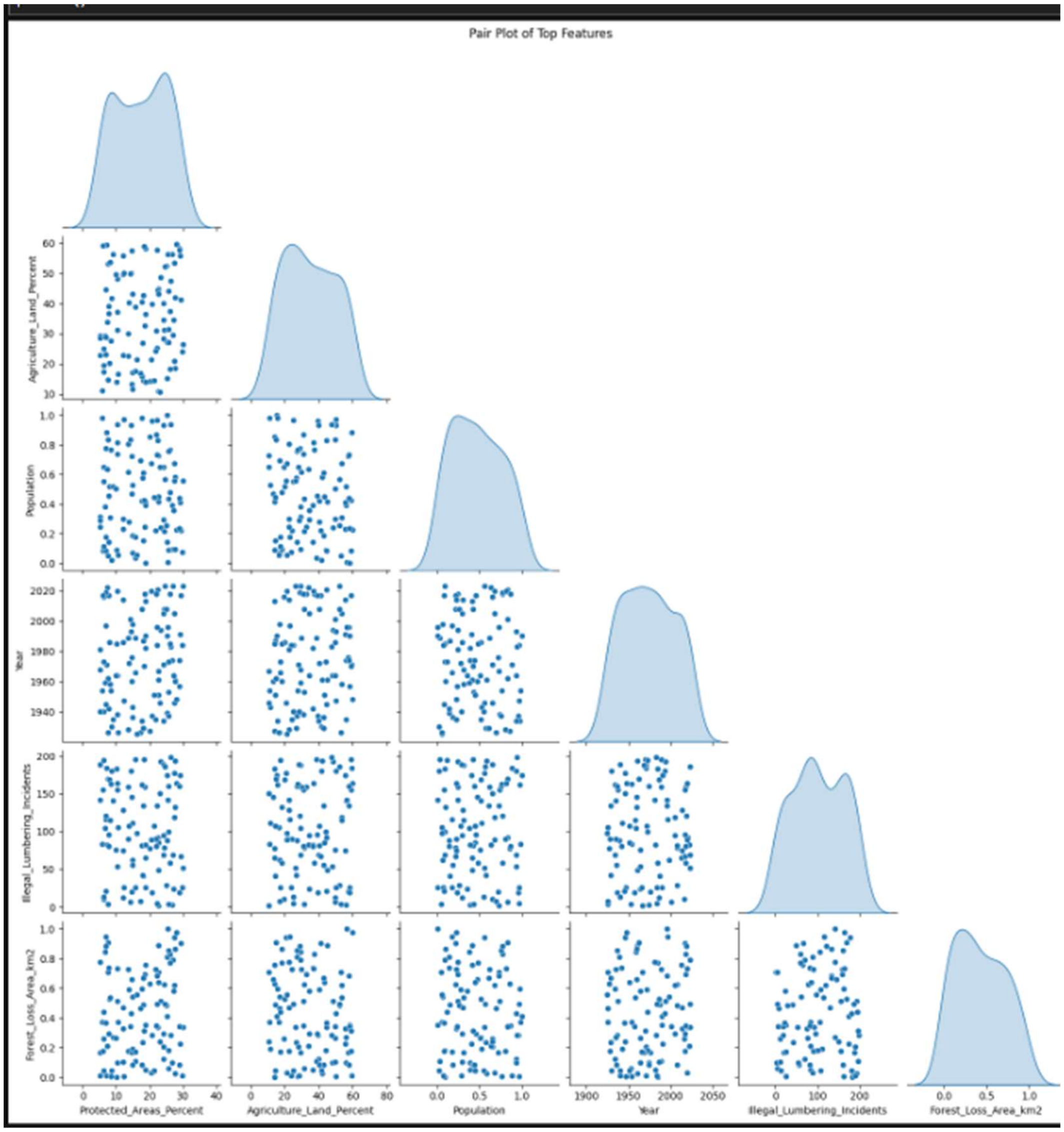

Figure 2-8. Pair Plot highlights complex joint distributions and potential non-linearities.

3. Support Vector Regression Methodology & Results

Support Vector Machines (SVMs) offer a principled framework for mitigating overfitting while capturing complex, non-linear relationships through kernel functions. In regression mode (SVR), the algorithm searches for a function that deviates from every training point by at most ε while maintaining maximal margin. Key hyper-parameters include: (C) the regularisation trade-off between flatness and tolerance to errors; (γ) the radial basis function kernel width governing locality; and (ε) the permissible error margin.

Pipeline Configuration. All numerical features were z-standardised within cross-validation to avoid data leakage. PolynomialFeatures (degree = 2) were appended to approximate pair-wise interactions, followed by SVR with an RBF kernel. A nested cross-validation scheme (outer 5-fold, inner 3-fold) simultaneously tuned $C \in [0.1, 100]$, $\gamma \in$ {'scale', 'auto'}, and $\varepsilon \in [0.01, 0.2]$. GridSearchCV evaluated negative mean absolute error as the optimisation metric.

Results. The outer-fold evaluation produced a mean absolute error of 0.326 ± 0.025 and an RMSE of 0.404. However, $R^2 = -2.01$ indicates the model explains less variance than a constant mean predictor. Examination of fold residuals confirms near-random scatter. The modest MAE combined with highly negative R^2 points to a model that predicts the central tendency reasonably but fails on variance—consistent with low feature-target signal.

Diagnostic Insights. Learning curves plateau quickly, suggesting additional observations alone may not rescue performance. Sensitivity analysis across hyper-parameters shows that increasing C beyond 10 worsens generalisation, confirming that the model is data-limited rather than under-parameterised.

4. Root-Cause Analysis of Model Under-performance

Four intertwined factors explain the SVR's poor explanatory power:

1. **Weak Signal Strength –** With absolute Pearson correlations below 0.15, the predictors collectively hold limited information about deforestation outcomes. Predictive algorithms fundamentally rely on signal; weak correlations translate into high irreducible error.

2. **Sample Size vs Feature Dimension** – The dataset contains roughly 100 records but 24 engineered features after polynomial expansion. The curse of dimensionality causes the SVR to fit noise, thereby inflating variance.

3. **Data Quality & Measurement Error** – Socio-economic metrics are often self-reported, susceptible to political influences. Digit preference (round numbers) in illegal lumbering counts implies heaping error. Such noise attenuates true relationships.

4. **Heterogeneity & Omitted Variables**– Countries differ by biome, governance style, and enforcement reach. Absent spatial controls (e.g., eco-zone) or policy shocks (e.g., land-tenure reforms), the model conflates heterogeneous contexts, further diluting predictive strength.

5. Pathways to Enhanced Predictive Power

Feature Enrichment. Marry remote-sensing products—normalised difference vegetation index (NDVI), annual forest-loss rasters, accessibility maps—with socio-economic layers to capture both supply-side pressure and enforcement capability. Night-time lights can proxy local economic intensity.

Advanced Algorithms. Gradient Boosting Machines (e.g., XGBoost) accommodate sparse interactions and non-linearities with built-in feature importance scoring. Gaussian Process Regression can encode spatial covariance kernels, implicitly leveraging geographic proximity.

Temporal & Spatial Structure. Switching from pooled to hierarchical (mixed-effects) models allows random intercepts per country and year, disentangling country-specific baselines. TimeSeriesSplit prevents future leakage during validation.

Target Engineering. Applying a log(1+target) transform or modelling annual change (Δ deforestation) often yields a more stationary target with reduced heteroscedasticity.

Robust Validation. Employ nested spatial cross-validation, where folds correspond to eco-regions or country clusters, to better approximate deployment scenarios.

6. Evidence-Based Policy Recommendations

The empirical findings, though limited in predictive accuracy, still point towards actionable levers:

1. **Expand and Effectively Manage Protected Areas.** Scatter plots suggest marginal decrease in deforestation at higher protected-area coverage. Ensuring legal designation is coupled with enforcement—through ranger funding, community monitoring, and geofencing—can amplify this signal.

2. **Curtail Illegal Lumbering via Technology.** Remote-sensing alerts, drone surveillance, and blockchain-based timber supply chains reduce anonymity, increasing risk for illegal actors. Coupling these tools with swift judicial processes magnifies deterrence.

3. **Incentivise Sustainable Agriculture Intensification.**Where high Agriculture Land % coincides with moderate deforestation, promoting yield-boosting inputs (precision fertiliser, drought-resilient crops) can decouple agricultural revenue from land conversion.

4. **Conditional International Aid.** Tether disbursements to transparent deforestation metrics—verified via third-party satellite imagery—to align donor incentives with forest outcomes.

5. **Progressive Forest-Loss Levies.**Implement GDP-linked levies on

commodities with embedded deforestation risk, redirecting proceeds to reforestation and livelihood diversification programmes.

7. Conclusion & Future Outlook

This report demonstrates that sound analytics do not automatically translate to high-accuracy models when underlying data are sparse or weakly correlated with the outcome of interest. Nonetheless, the systematic workflow—from EDA through model diagnostics to policy interpretation—offers a replicable template for environmental data science projects. Moving forward, emphasis must be placed on richer, higher-resolution datasets and interdisciplinary collaborations so that decision-makers receive precise, context-aware guidance to halt and reverse deforestation trends.

In the spirit of continuous improvement, future work will trial ensemble learners informed by spatial hierarchies and integrate near-real-time Earth observation feeds. Together, these advances promise a clearer understanding of the deforestation puzzle, paving the way for more targeted, effective conservation interventions.