Supervised Learning – Part II: Predicting Population Growth Categories from Water-Related Features

1. Introduction

Population-growth dynamics influence economic development, resource allocation, and environmental sustainability worldwide. This study classifies countries as “slow growth” or “rapid growth” based on demographic changes from 2002 to 2021, using water-related indicators to represent socio-economic and ecological pressures.
Each indicator was summarized in terms of its mean, variability, extremes, and long-term trend. Three machine-learning models—random forest, XGBoost, and GLMNET—were evaluated using nested cross-validation, with a random-baseline model included for comparison via hypothesis testing. Interpretability was enhanced through variable-importance rankings, odds-ratio estimates, and partial-dependence plots, and a global choropleth map was used to observe spatial patterns of misclassification.

2. R Libraries and Tools Used

Library	Purpose
tidyverse	Data manipulation and visualization
caret	Training and tuning of machine learning models
MLmetrics	Provides the F1-score, precision, recall, and other classification metrics
pROC	Computes ROC curves and AUC (Area Under the Curve)
pdp	Generation of partial-dependence plots for model interpretation
rnaturalearth	Retrieval of global country geometries for mapping
sf	Handling and plotting of spatial data

3. Labels and Features

3.1. Label Construction

The population percentage growth rate for each country was calculated as:

$$growth\,rate\ = \frac{X2021 - X2002}{X2002} \times 100$$

where X2002 and X2021 are the population counts in those years. Countries with a growth rate below 30% were labeled Slow growth, and those at or above 30% were labeled Rapid growth.

.

The resulting label counts are:

- Rapid growth: 58
- Slow growth: 57

(total 115 countries), yielding an almost perfect 50/50 split. The 30% cutoff was deliberately chosen to achieve this balance, preventing majority-class bias during model training and ensuring that performance metrics such as accuracy and F1-score remain reliable. By balancing the classes up front, model comparisons focus on genuine predictive performance instead of compensating for label skew.

3.2. Feature Selection and Summary Statistics

A total of nine water-related time-series indicators were initially considered for inclusion—water_productivity, available_resources, water_stress, withdrawals, agriculture, domestic, industry, plus precipitation and natural_disasters. However, several were dropped for the following reasons:

1. Functional redundancies:
   Because agriculture, domestic, and industry sum to 100, the “domestic” feature was removed; likewise, since water_productivity and water_stress are calculated as GDP / withdrawals and withdrawals / available_resources respectively, the “withdrawals” feature was also excluded to avoid functional redundancy and prevent multicollinearity.

2. Lack of temporal dynamics:
   Precipitation and natural_disasters offered only long-term averages without year-to-year variation—precluding the computation of slopes, extrema, and standard deviations—so they were excluded for lacking the dynamic information required for feature engineering.

This left five core datasets—water_productivity, available_resources, water_stress, agriculture, and industry—each spanning 2002–2021. These indicators collectively capture distinct facets of a country’s water dynamics: water_productivity relates economic output to withdrawal volume, water_stress measures withdrawal pressure against renewable supply, available_resources quantifies the total endowment, and the agriculture and industry percentages reveal how water use is allocated between food production and industrial activity. It was deemed that no additional exclusions were necessary, since each feature seems to contribute complementary information beyond the functional redundancies already removed (domestic and withdrawals). Domain knowledge suggests these variables can indeed vary independently—for instance, high productivity in water-rich nations may coexist with low stress, and sectoral allocation often differs even among countries with similar overall stress—indicating a reasonably comprehensive, non-redundant feature set for modeling.

For each retained indicator, five summary statistics were calculated per country—mean, standard deviation, minimum, maximum, and the linear trend (slope) over the twenty-year period. By including these statistics, the model can account for extreme events and long-term shifts as well as average behavior—insights that would be unavailable if only mean values were used.

4. Classification Methodology

4.1. Candidate Model Selection

Model	Description
Random Forest	Ensemble of decision trees using bagging; captures non-linear interactions; robust to noise and overfitting
XGBoost	Gradient-boosted decision trees; emphasizes errors from prior trees; strong performance with good tuning
GLMNET	Regularized logistic regression (L1/L2); interpretable coefficients; automatic feature selection
Linear SVM	Maximizes class margin with a linear boundary; effective in high-dimensional and small sample settings; limited interpretability

Three models were chosen to span distinct learning paradigms: a tree-based ensemble (random forest), a gradient-boosted method (XGBoost), and a penalized linear model (GLMNET). Random forests excel at capturing high-order interactions without overfitting, XGBoost refines those interactions through sequential boosting and built-in regularization, and GLMNET offers a transparent, coefficient-based view of linear relationships with automatic feature selection. Though only one model is ultimately selected for deployment, evaluating complementary models ensures that the decision is grounded in a comprehensive and well-contextualized framework.

Linear support vector machine could similarly provide a robust linear decision boundary and would likely perform competitively, given the modest size (`115 countries, 25 features`) of the dataset. However, it falls into the same linear paradigm as GLMNET and does not natively yield readily interpretable coefficients or odds ratios without additional post-hoc procedures. By contrast, GLMNET's regularization path explicitly shrinks and selects features, making its internal mechanics directly transparent to practitioners. Thus, GLMNET was chosen over SVM not due to anticipated performance advantages but to its interpretability and transparency.

4.2. Nested Cross-Validation Procedure

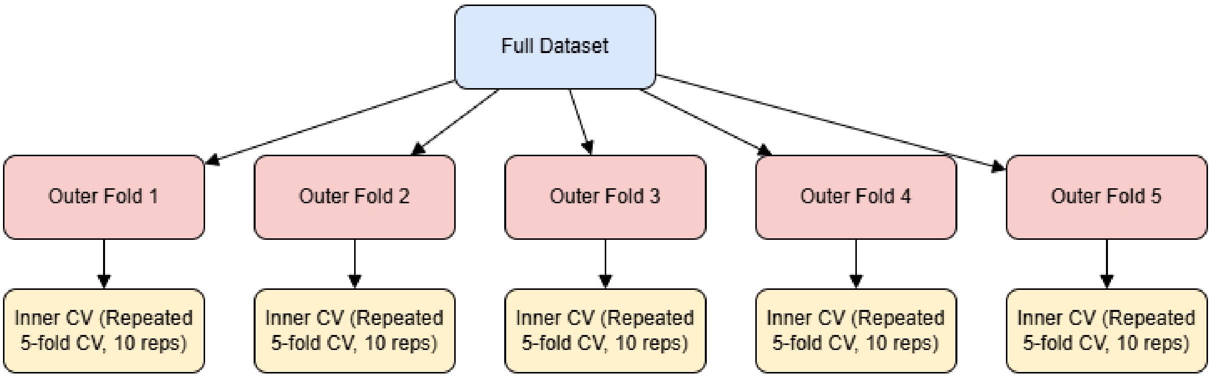

Full Dataset
Outer Fold 1
Outer Fold 2
Outer Fold 3
Outer Fold 4
Outer Fold 5
Inner CV (Repeated 5-fold CV, 10 reps)
Inner CV (Repeated 5-fold CV, 10 reps)
Inner CV (Repeated 5-fold CV, 10 reps)
Inner CV (Repeated 5-fold CV, 10 reps)
Inner CV (Repeated 5-fold CV, 10 reps)

Figure - Structure of the nested cross-validation framework used in this study.

To robustly evaluate model performance while minimizing overfitting risk, a nested cross-validation framework was employed. The outer loop consisted of a 5-fold split of the full dataset. For each outer fold, the data was partitioned into training and testing subsets, with the testing portion held out entirely for final evaluation. This structure ensures that hyperparameter tuning occurs independently of the evaluation data, yielding more reliable performance estimates.

Within each outer training fold, three models—random forest, XGBoost, and GLMNET—were tuned using repeated 5-fold cross-validation with 10 repetitions. A random search approach was used for hyperparameter tuning, with F1-score selected as the primary optimization metric, as further discussed in Section 4.3. Preprocessing steps were applied according to model requirements: for GLMNET, features were centered and scaled to ensure stable coefficient estimation and effective regularization. Range scaling was applied for random forest and XGBoost. However, it proved unnecessary: results remained unchanged without scaling, likely because these algorithms rely on threshold-based splits and are inherently insensitive to feature magnitude.

After hyperparameter tuning within the inner loop, each model was used to generate predictions on the outer test data. Performance was assessed using both F1-score and accuracy, and results were aggregated across all outer folds. As will be discussed in Section 6, GLMNET was ultimately selected as the final model; thus, additional metrics including AUC, sensitivity, specificity, and out-of-fold predictions were specifically recorded for GLMNET to support further interpretation and mapping.

Note on Feature Selection:
During preliminary modeling, feature selection procedures based on variable importance were explored for the Random Forest and XGBoost models. However, these steps did not yield consistent performance improvements, and any potential benefit appeared to be masked by variability introduced from different seed initializations. Therefore, these procedures were ultimately discarded to streamline computation and reduce unnecessary complexity.

4.3. Evaluation Metric Selection

Metric	Description
Accuracy	The proportion of total correct predictions (both positive and negative) out of

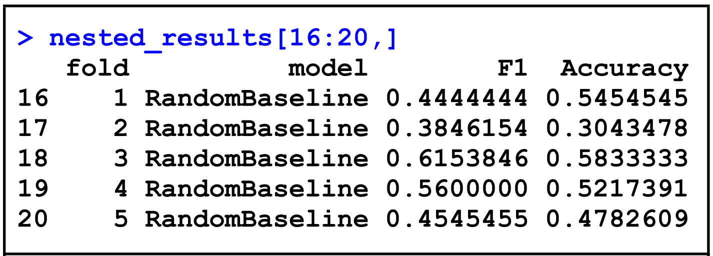
	all predictions. Used for model comparison but not for tuning in this study.
F1-score	The harmonic mean of precision and recall. Prioritizes balance between false positives and false negatives. Used for both hyperparameter tuning and model comparison.
AUC	Area Under the ROC Curve. Evaluates a model's ability to distinguish between classes by ranking predicted probabilities. Not used for tuning or comparison, but provides complementary insight into ranking ability.

As previously stated, a random search approach was used for hyperparameter tuning, with F1-score selected as the primary optimization metric. Although the dataset was nearly balanced and accuracy was initially considered as the optimization metric, it was later replaced by F1-score due to an observed imbalance between precision and recall in the result. This choice reflects the recognition that both false positives and false negatives may carry significant and distinct costs in the classification context. However, when comparing the final performance of the three models, both accuracy and F1-score were reported to provide a more comprehensive evaluation.

While AUC provides valuable insight into a model's ability to rank predictions, it was not used for either tuning or final model comparison, as the objective of this study was to produce discrete class predictions rather than probability-based rankings.

5. Model Evaluation and Comparison

5.1. Baseline Model Introduction



```
> nested_results[16:20,]
   fold          model        F1  Accuracy
16    1 RandomBaseline 0.4444444 0.5454545
17    2 RandomBaseline 0.3846154 0.3043478
18    3 RandomBaseline 0.6153846 0.5833333
19    4 RandomBaseline 0.5600000 0.5217391
20    5 RandomBaseline 0.4545455 0.4782609
```

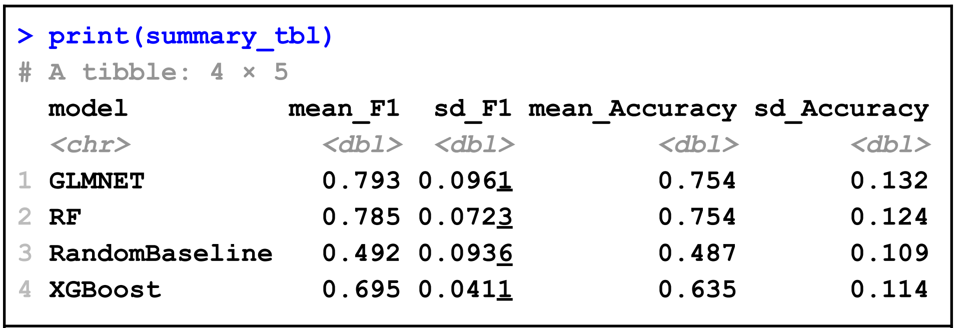
Table - Cross validation performance of the random baseline model

To provide a meaningful performance reference, a random baseline classifier was implemented. This model predicts each class label ("Slow growth" or "Rapid growth") uniformly at random, simulating the performance of a no-skill model in a balanced binary classification setting.

The baseline model was evaluated across the same five outer folds as the candidate models. As shown in the table above, the baseline yielded a mean F1-score of 0.49 (±0.09) and a mean accuracy of 0.49 (±0.10). These results reflect the expected performance level of a naive classifier and serve as a lower bound for meaningful model performance.

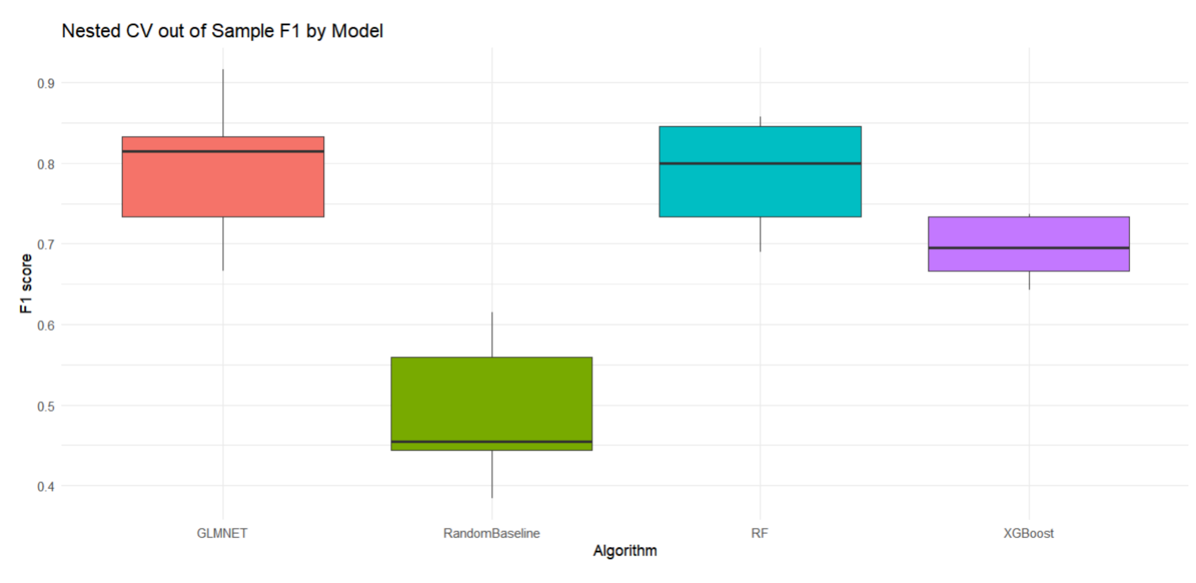
5.2. Model Performance Summary and Visualization

F1-score and accuracy were calculated across the five outer folds to assess model performance. The table below reports the mean and standard deviation for each model, while the two boxplots visualize their distribution:

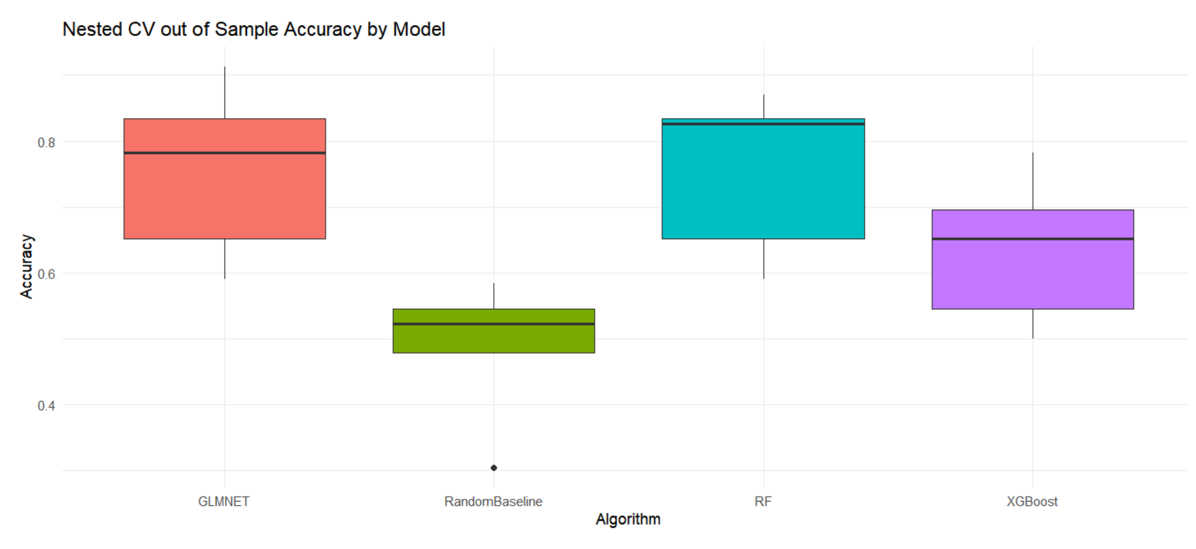


```
> print(summary_tbl)
# A tibble: 4 × 5
  model          mean_F1  sd_F1 mean_Accuracy sd_Accuracy
  <chr>            <dbl>  <dbl>         <dbl>       <dbl>
1 GLMNET           0.793 0.0961         0.754       0.132
2 RF               0.785 0.0723         0.754       0.124
3 RandomBaseline   0.492 0.0936         0.487       0.109
4 XGBoost          0.695 0.0411         0.635       0.114
```

Table - Mean and Standard Deviation of F1-score and Accuracy across Outer Folds

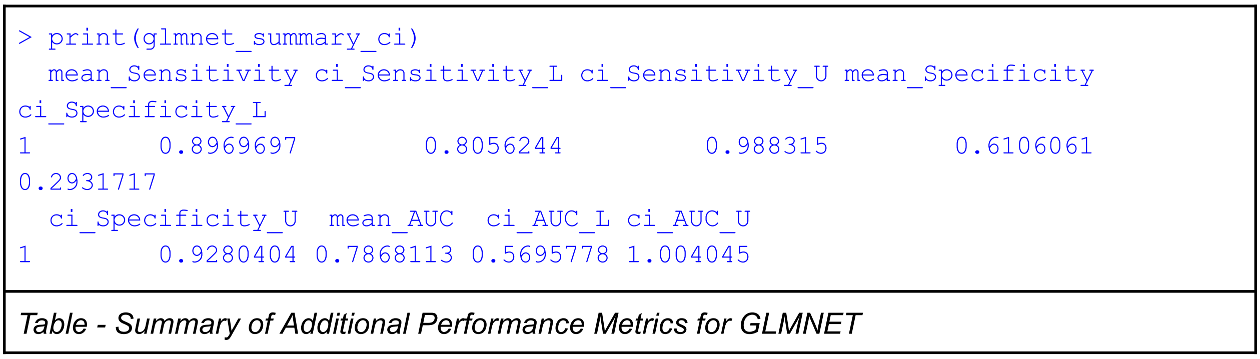


Boxplot - Distribution of F1 score across Outer Folds



Boxplot - Distribution of Accuracy across Outer Folds

Both GLMNET and Random Forest achieved comparable performance, with nearly identical mean accuracy and F1-scores. XGBoost lagged slightly behind, in terms of both F1-score and accuracy. However, all three models substantially outperformed the random baseline, suggesting that the water-related features provided meaningful signal for classification. The standard deviations (SD) of F1-score across models were moderate—0.096 for GLMNET, 0.072 for Random Forest, and 0.041 for XGBoost—indicating reasonably consistent performance across folds. While the limited dataset size could have introduced greater variability, the use of nested cross-validation appears to have mitigated this risk to the present extent.



```
> print(glmnet_summary_ci)
  mean_Sensitivity ci_Sensitivity_L ci_Sensitivity_U mean_Specificity
ci_Specificity_L
1         0.8969697        0.8056244         0.988315        0.6106061
0.2931717
  ci_Specificity_U  mean_AUC  ci_AUC_L ci_AUC_U
1         0.9280404 0.7868113 0.5695778 1.004045
```

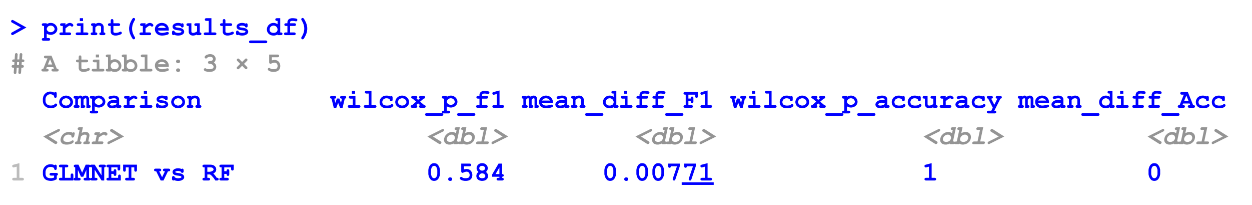
Table - Summary of Additional Performance Metrics for GLMNET

As GLMNET was selected as the representative model for interpretation (see Section 6), its sensitivity, specificity, and AUC across the outer folds were additionally evaluated to provide a more comprehensive understanding of model performance. The results were:

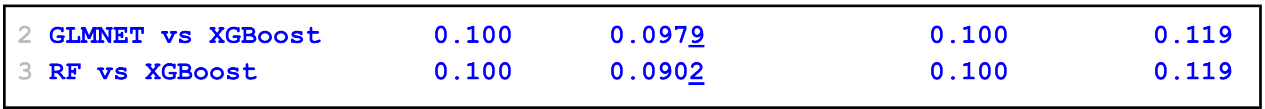
- Sensitivity: 0.897 [95% CI: 0.806, 0.988]
- Specificity: 0.611 [95% CI: 0.293, 0.928]
- AUC: 0.787 [95% CI: 0.570, 1.004]

These results suggest that the model performs strongly in identifying rapid-growth countries, as indicated by the high sensitivity with a relatively narrow confidence interval. However, the broader confidence interval for specificity reflects less stable performance in correctly identifying slow-growth cases, primarily due to the modest dataset size and associated variability across folds. The AUC value indicates a solid overall discriminative ability, although the upper bound slightly exceeding 1 is a statistical artifact resulting from normal approximation; AUC is theoretically bounded between 0 and 1 and should be interpreted as approaching but not exceeding 1.

5.3. Pairwise Model Comparison via Statistical Testing



```
> print(results_df)
# A tibble: 3 × 5
  Comparison       wilcox_p_f1 mean_diff_F1 wilcox_p_accuracy mean_diff_Acc
  <chr>                  <dbl>        <dbl>             <dbl>         <dbl>
1 GLMNET vs RF           0.584      0.00771                 1             0
```



```
2 GLMNET vs XGBoost      0.100      0.0979           0.100         0.119
3 RF vs XGBoost          0.100      0.0902           0.100         0.119
```

Table - Wilcoxon Signed-Rank Test Results and Mean Differences in F1-score and Accuracy Between Models

Pairwise Wilcoxon signed-rank tests were conducted to assess whether differences in F1-score and accuracy between models were statistically significant. These non-parametric tests were chosen over paired t-tests due to the small sample size and potential violations of normality and homogeneity of variances in the performance metrics. The results are summarized in the above Table.

The comparison between GLMNET and Random Forest yielded high p-values (0.584 for F1-score and 1.000 for accuracy), indicating no statistically significant difference in performance between these two models. In contrast, comparisons involving XGBoost showed lower p-values (0.100 for both F1-score and accuracy), suggesting a possible trend toward better performance by GLMNET and Random Forest. The mean differences in F1-score were approximately 0.098 (GLMNET vs. XGBoost) and 0.090 (RF vs. XGBoost), while the corresponding differences in accuracy were both around 0.119.

Although these differences did not reach conventional thresholds for statistical significance ($p < 0.05$), the consistency in direction across both metrics indicates that GLMNET and Random Forest may have offered better performance than XGBoost in this context. Nevertheless, due to the limited sample size and moderate variance, these findings should be interpreted with caution.

6. GLMNET Interpretation and Error Analysis

For further interpretation, GLMNET was chosen due to its strong and stable performance shown in earlier testings, as well as its clear advantage in interpretability. While both GLMNET and Random Forest achieved similar performance, GLMNET provides direct access to model coefficients and supports straightforward techniques such as odds ratio analysis and partial dependence plots. The following subsections therefore focus on interpretive analysis of the GLMNET model and explore misclassification patterns across countries.

6.1. Variable Importance, Coefficients, and Feature Effects

6.1.1. Variable Importance Rankings

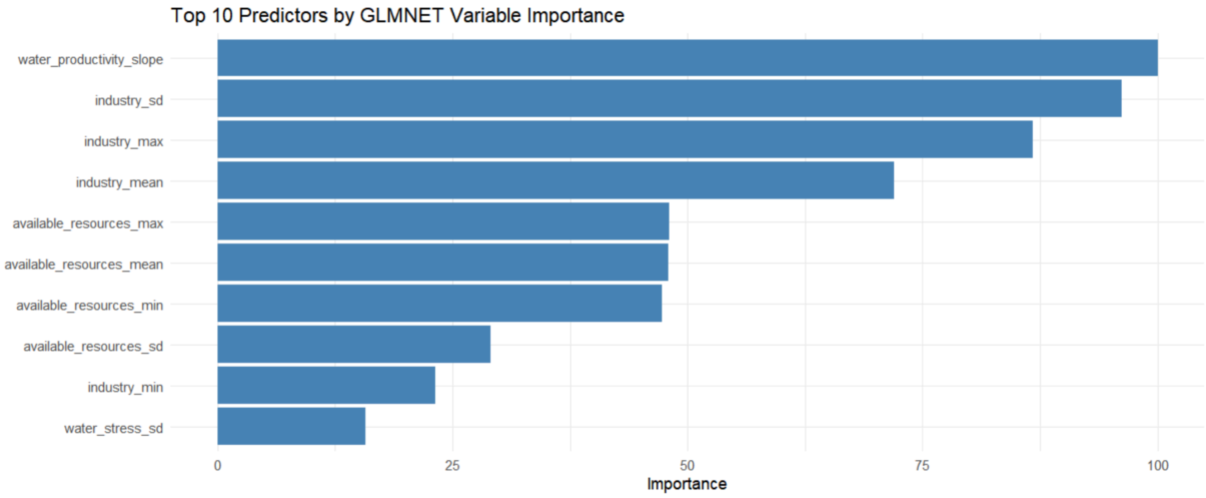


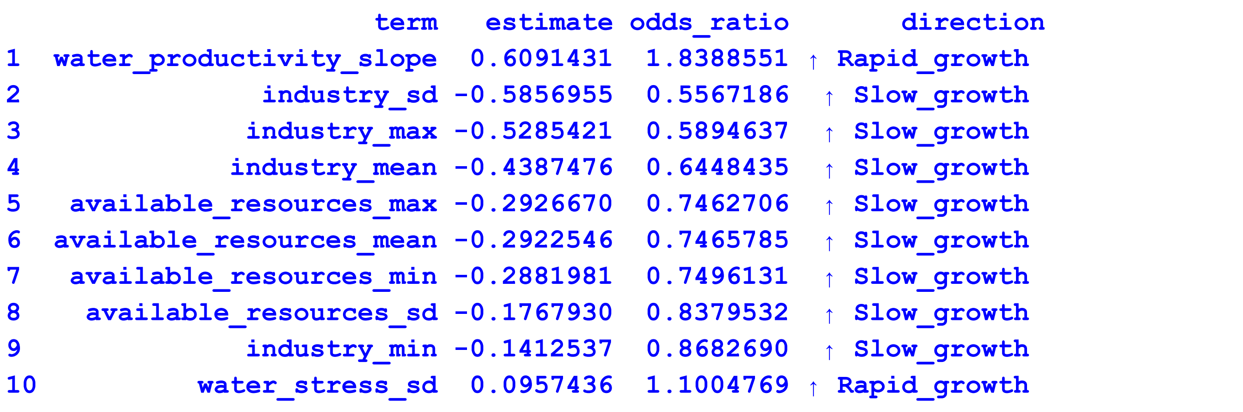
Figure - Top 10 predictors by GLMNET variable importance
(Higher values indicate stronger contribution to model predictions.)

To better understand which features contributed most strongly to the classification decision, the GLMNET model was first re-tuned on the full dataset using repeated cross-validation, and then refit using the best hyperparameters for interpretation. Variable importance scores were subsequently computed from this final model to highlight the most influential predictors. As shown in the figure above, the top-ranked feature was the slope of water productivity over the 2002–2021 period. This feature captures the direction and rate of long-term change in how efficiently water is used for economic output, rather than a static snapshot. Its prominence suggests that sustained improvements or declines in water productivity are strongly associated with a country's population growth category.
Several indicators related to industrial water use followed closely, reflecting the influence of both magnitude and variability in this sector. Available water resources also emerged as important predictors (with importance scores around 50%), suggesting that countries with greater or more stable access to natural water resources may follow distinct demographic trajectories. Lastly, variability(standard deviation) in water stress appeared as the tenth most influential feature, although its importance score was below 20%—considerably lower than that of the leading predictors. This suggests that while it was not a dominant factor, fluctuations in water system pressure may still contribute meaningfully to the model's decision-making, particularly when considered alongside more influential variables.

6.1.2. Coefficient Interpretation

In addition to variable importance, GLMNET coefficients were examined to assess the direction and magnitude of each feature's effect. Coefficients were converted to odds ratios and labeled to indicate whether higher values increased the likelihood of a country being classified as "Rapid growth" or "Slow growth," offering interpretable insight into how water-related trends relate to demographic patterns. A summary of the top 10 coefficients is provided in the table below.

```
> print(coefs_df)
```



```
                      term   estimate odds_ratio       direction
1  water_productivity_slope  0.6091431  1.8388551 ↑ Rapid_growth
2               industry_sd -0.5856955  0.5567186  ↑ Slow_growth
3              industry_max -0.5285421  0.5894637  ↑ Slow_growth
4             industry_mean -0.4387476  0.6448435  ↑ Slow_growth
5   available_resources_max -0.2926670  0.7462706  ↑ Slow_growth
6  available_resources_mean -0.2922546  0.7465785  ↑ Slow_growth
7   available_resources_min -0.2881981  0.7496131  ↑ Slow_growth
8    available_resources_sd -0.1767930  0.8379532  ↑ Slow_growth
9              industry_min -0.1412537  0.8682690  ↑ Slow_growth
10           water_stress_sd  0.0957436  1.1004769 ↑ Rapid_growth
```

Table - Top 10 GLMNET Coefficients with Odds Ratios and Classification Direction. Positive coefficients indicate association with the "Rapid growth" class; negative values indicate association with "Slow growth".

The coefficient patterns reveal a clear contrast between features associated with rapid versus slow growth. All indicators related to industrial water use and available water resources had negative coefficients and odds ratios below one, pointing to a consistent association with the “Slow growth” category. In contrast, water productivity slope had the largest positive coefficient, with an odds ratio of approximately 1.84, linking sustained improvements in water productivity to rapid population growth. A smaller positive effect was observed for water stress variability, suggesting that fluctuating pressure on water systems may also play a role in characterizing fast-growing countries. Collectively, these results suggest that both long-term trends and resource variability are relevant factors in demographic outcomes.

6.1.3. Partial Dependence Analysis

Partial dependence plots (PDPs) were generated for the three most influential predictors: water_productivity_slope, industry_sd, and industry_max. These plots illustrate the isolated effect of each feature on the predicted probability of a country being classified as “Rapid growth,” by marginalizing over the joint distribution of all other features. This approach enables an interpretable approximation of how the model responds to changes in a single variable, independent of interactions with other predictors.

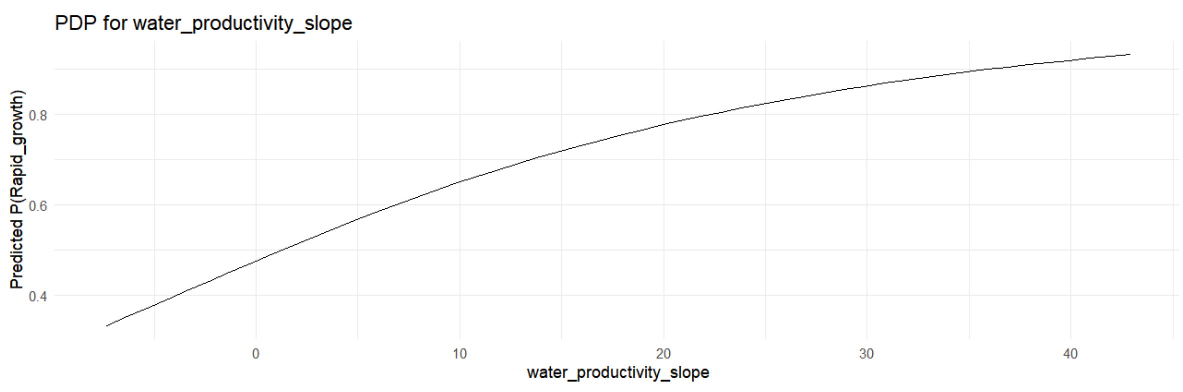


Figure - Partial dependence plot for water productivity slope

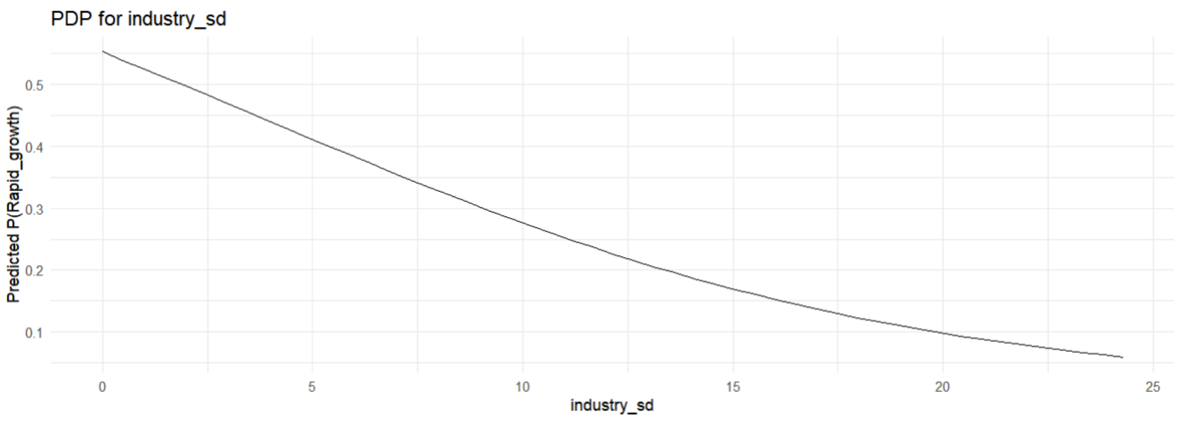


Figure - Partial dependence plot for industrial withdrawals sd(standard deviation)

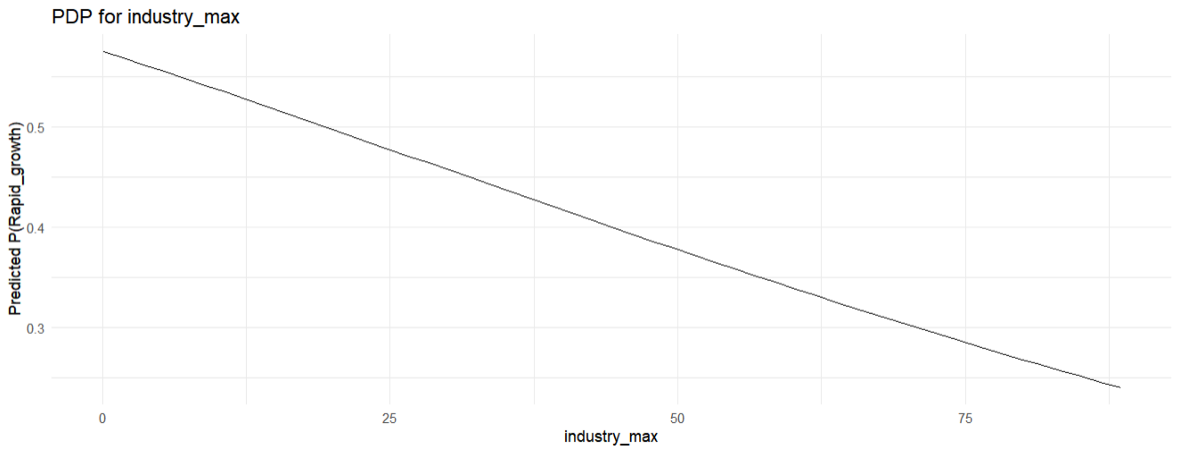


Figure - Partial dependence plot for industrial withdrawals max

The first figure above shows that higher values of water_productivity_slope are associated with a steadily increasing probability of rapid growth. This aligns with previous findings that countries experiencing consistent improvements in water productivity are more likely to fall into the high-growth category. The effect appears smoothly positive and nonlinear, with a gradually decreasing slope.

In contrast, both industry_sd and industry_max exhibit strong negative relationships with the probability of rapid growth (second and third plots). Also, the effect of industry_max is nearly linear, suggesting a consistent downward influence across its range. These trends support the interpretation that excessive or unstable industrial water demand may be more characteristic of slower-growing populations.

Together, these PDPs illustrate how the GLMNET model integrates both static and dynamic features into its decision-making, offering interpretable insights into the relationships between water dynamics and population growth.

Note: The interpretations provided here are specific to the GLMNET model. Other algorithms, such as Random Forest or XGBoost, which capture nonlinear interactions and complex feature hierarchies, may prioritize different features and yield different patterns of importance. While GLMNET offers a transparent and interpretable view of the data, its *perspective should be understood as one among several possible representations learned by different model classes.*

6.2. Country-Level Misclassification Analysis

To complement the interpretation of model behavior, performance was also examined at the country level. The following map visualizes which countries were misclassified by the GLMNET model during cross-validation.

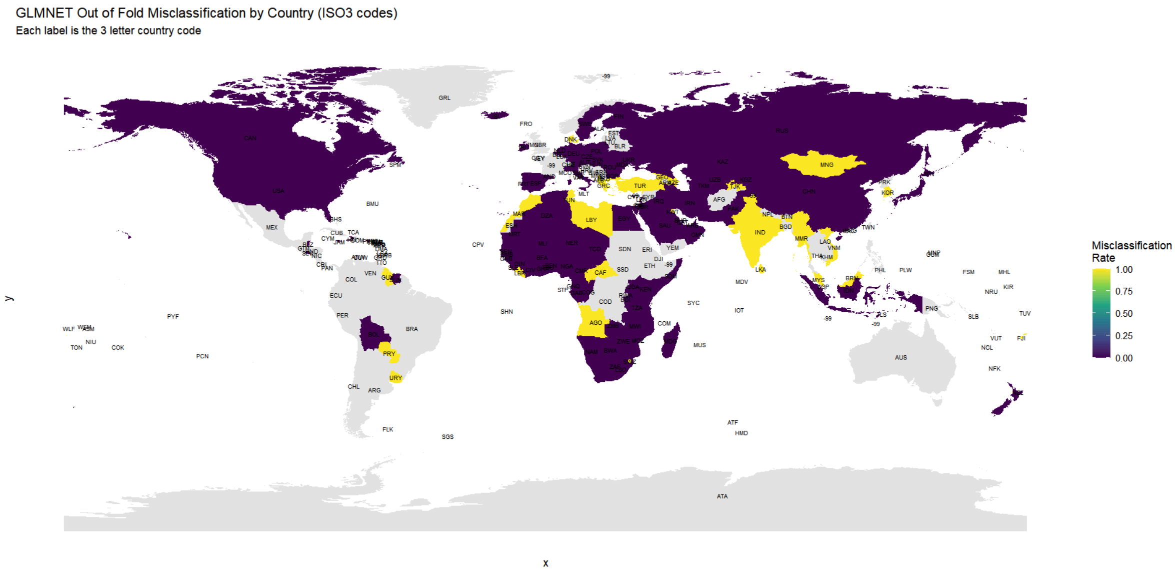


Figure – Country-Level Classification Outcomes from GLMNET (Out-of-Fold Predictions)

This choropleth map highlights spatial patterns in misclassification based on out-of-fold predictions from the GLMNET model. Each country is shaded according to whether its single out-of-fold prediction was correct or incorrect during nested cross-validation. While most countries were classified correctly (shown in purple), a subset—particularly in Sub-Saharan Africa, Latin America, parts of the Middle East and Asia—were misclassified. These geographic patterns may reflect regional limitations in the feature set, data quality inconsistencies, or contextual factors not captured by the model. Such spatial diagnostics can help identify areas where predictive performance is less reliable and warrant further investigation.

7. Reflections and Possible Applications

While the analysis yielded interpretable results and moderately strong predictive performance, several limitations warrant cautious interpretation. Chief among these is the modest dataset size: with only 115 countries and a limited number of temporally aggregated features, the models faced inherent constraints in generalizability. Although the standard deviations of F1-score and accuracy across folds were not excessively high, some variability was still present. This was further reflected in the wide confidence intervals of secondary metrics such as specificity and AUC, highlighting not only variability across folds but also the model's limited reliability in identifying slow-growth countries.

Moreover, the interpretation of model behavior was specific to the GLMNET classifier and may not generalize across algorithms. Models such as Random Forest or XGBoost—which capture nonlinear interactions and hierarchical structures—would likely have yielded different feature rankings and response patterns.

Future analyses could benefit from expanding the dataset both in terms of size and the variety of features. This might involve including more countries, using longer time periods, or adding more detailed variables. For example, more complex patterns could be captured by using lagged inputs (e.g., previous year's values), or by retrieving dynamic precipitation and natural disaster data from alternative sources, as these were previously excluded due to being available only as long-term averages.

On the methodological side, techniques such as averaging across multiple models (ensemble methods) may help improve the reliability and generalizability of the results, while repeating the training process with different random seeds (a seed sensitivity test) would ensure that the findings are not overly dependent on chance.

Lastly, since the current threshold was chosen primarily to ensure a balanced split between the two classes, refining the labeling strategy—based on theoretical reasoning or expert guidance—could lead to a more meaningful and interpretable classification of population growth.

Despite its limitations, the current modeling framework offers a tentative foundation for exploring the relationship between water-related characteristics and demographic trends. While the results are not definitive, they suggest a few possible directions for practical reflection:

1. Early signals of demographic pressure: Countries showing consistent improvements in water productivity may be more likely to sustain rapid population growth. Monitoring long-term gains in water productivity could thus offer an early indication of developmental momentum.

2. Industrial water use as a potential constraint: High variability or peaks in industrial water withdrawals appeared to be associated with slower-growing populations. While the causal relationship is unclear, such patterns may signal inefficiencies or infrastructural limitations that could impact demographic outcomes over time.

3. Resource availability and planning: While not the most dominant predictors, the static measures of available water resources occupied a solid mid-range in importance. Their consistent association with slower growth may suggest underlying structural dependencies—such as the potential for complacency or inefficiencies in settings where water is relatively abundant.

While these insights are exploratory and model-dependent, they may nonetheless contribute to broader conversations in areas such as water resource governance, urban planning, and development strategy. That said, the model's difficulty in consistently identifying slow-growth countries warrants particular caution when interpreting results for more demographically stagnant regions, where predictive uncertainty may be higher. Ultimately, this study illustrates that even relatively simple water indicators, when systematically modeled, can provide valuable perspectives on the complex interplay between environmental factors and population dynamics—though more robust, theory-driven approaches would be necessary to support confident, real-world decision-making.

8. Conclusion

This study explored whether long-term water-related indicators could help classify countries into distinct population growth categories. By leveraging three supervised learning models—GLMNET, Random Forest, and XGBoost—and applying a nested cross-validation framework, the analysis aimed to capture the most comprehensive model capacity, especially in terms of both predictive performance and interpretability.

Among the models tested, GLMNET emerged as the most interpretable and comparably effective, offering insights into how variables such as water productivity trends and industrial water use variability relate to demographic dynamics. The findings suggested that countries with sustained improvements in water productivity were more likely to exhibit rapid growth, while high variability or peaks in industrial water use tended to align with slower growth. Available water resources—although not the most influential predictors—were also consistently associated with slower-growing populations, hinting at underlying resource management or infrastructure patterns.

Nevertheless, these findings should be interpreted cautiously. The modest dataset and limited set of aggregated features could affect generalizability. Additionally, since interpretation was based solely on the GLMNET model, the results may not reflect the behavior of more complex algorithms. Future studies could enhance robustness by expanding the dataset, incorporating more diverse or dynamic features, applying ensemble methods, and refining the classification threshold based on theoretical or expert input.

Despite its exploratory nature, this analysis illustrates that even modest datasets and simple indicators—when analyzed systematically—can reveal suggestive patterns worth further investigation. While not definitive, the framework may offer a useful starting point for future work in water governance, resource planning, and environmental-demographic modeling.