Predicting Population Growth Categories Using Environmental Indicators: A KNN-Based Classification Approach

1.Introduction

Predicting population growth is important for effective planning in areas like resource allocation, infrastructure, and sustainability. While traditional models rely on demographic factors, this study explores the use of environmental indicators—such as water availability, water stress, and natural disaster frequency—to classify countries into two categories: rapid and slow population growth. Using a K-Nearest Neighbors (KNN) classifier, the project demonstrates how multi-year environmental data can be used to predict growth trends. This approach highlights the potential of integrating environmental data into population modeling, offering an alternative perspective for understanding global growth patterns.

2. R Packages Used

Package	Purpose
tidyverse	Data Manipulation and visualization
caret	Model training and validation framework
MLmetrics	Evaluation metrics such as F1 score, precision, recall
VIM	Missing value imputation using KNN
ggplot2	Model result visualization

3. Variable and Country selection

3.1. Target Variable: Population growth category

The target variable in this study is a binary label representing the population growth trend of each country, categorized as either Rapid Growth or Slow Growth. This label is derived from the percentage change in population between the years 2002 and 2021, using data from the World Bank. Countries with less than 30% population growth over the period were labeled as “Slow Growth,” while those with 30% or more were labeled as “Rapid Growth.” This classification provides a simplified yet meaningful way to differentiate demographic trends and serves as the output variable for the K-Nearest Neighbors (KNN) model.

3.2. Input Features: Environmental Indicators

To predict the population growth category, a diverse set of environmental and water-related indicators was used as input features. These include:

- Water Productivity
- Available water resources
- Water stress levels
- Water withdrawals (total, agricultural, domestic and industry)
- Precipitation levels
- Natural disaster occurrences

These features were selected based on the hypothesis that environmental conditions and resource availability influence demographic patterns over time. Each indicator was collected annually and reshaped into a wide format suitable for machine learning models.

3.3. Country inclusion criteria

Countries were included in the modeling dataset only if they had sufficient data across the chosen feature set and a valid population growth label. To ensure data completeness and avoid excessive imputation, countries missing most environmental indicators were excluded from the final dataset. This filtering step ensured that the model was trained on a consistent and reliable dataset, enabling more accurate classification and generalization.

4. Classification Framework

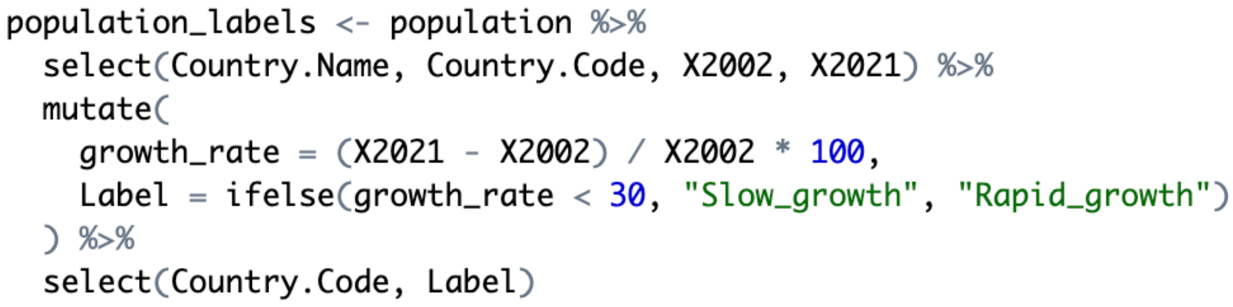
This study adopts a classification approach to predict population growth categories based on environmental indicators. The classification task involves distinguishing countries with rapid population growth from those with slow growth, using features derived from water-related metrics and natural disaster occurrences. The process involved multiple stages: feature engineering, label generation, data preparation, and model training using the K-Nearest Neighbors (KNN) algorithm.

4.1. Feature Engineering and Preprocessing

All input features were collected as time-series data from 2002 to 2021. These included water productivity, renewable water resources, total water withdrawals by sector (agriculture, domestic, industry), water stress levels, precipitation, and disaster-related records. To ensure consistent dimensionality, each feature was transformed from long format to wide format using the year as a suffix (e.g., water_stress_X2005). For natural disaster data, missing values were imputed using region-wise k-nearest neighbors based on the 2000 data point and replicated across all years.

4.2. Label generation

Population growth labels were derived using the relative percentage change in population between 2002 and 2021. Countries with a growth rate below 30% were assigned the Slow Growth label, and those above or equal to 30% were labeled as Rapid Growth. These binary labels served as the response variable for classification.

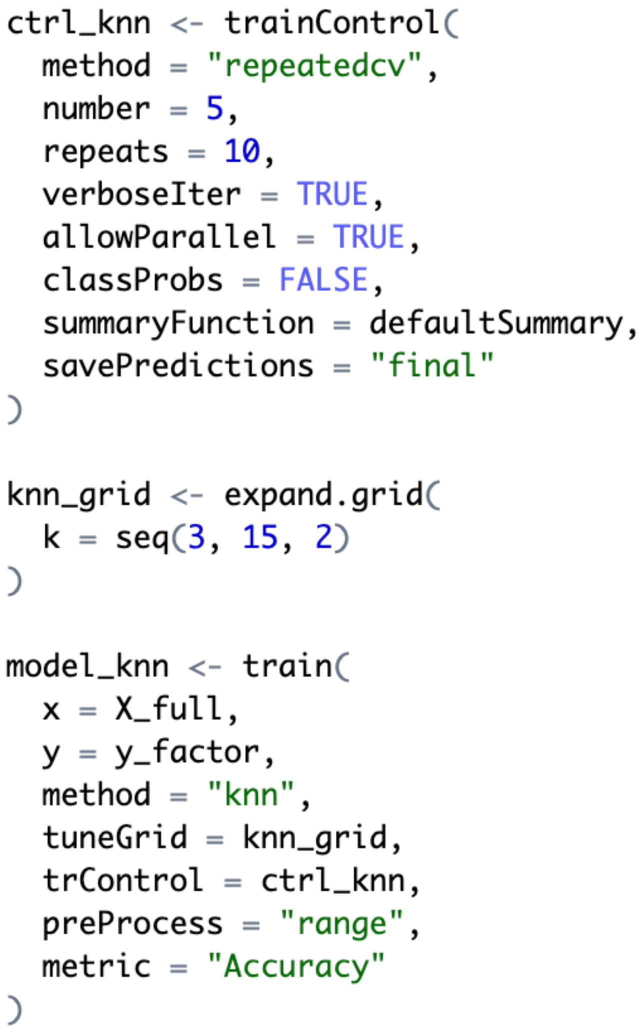


```r
population_labels <- population %>%
  select(Country.Name, Country.Code, X2002, X2021) %>%
  mutate(
    growth_rate = (X2021 - X2002) / X2002 * 100,
    Label = ifelse(growth_rate < 30, "Slow_growth", "Rapid_growth")
  ) %>%
  select(Country.Code, Label)
```

4.3. Model training: K- Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm was selected due to its simplicity and interpretability in high-dimensional tabular data. KNN makes no parametric assumptions and classifies instances based on the majority class among the k closest training points, measured via Euclidean distance. Prior to training, all features were normalized using min-max scaling to ensure fair distance-based comparisons.

A repeated k-fold cross-validation strategy was employed to tune the number of neighbors (k). The training process involved evaluating performance across multiple values of k (ranging from 3 to 15), with the optimal value selected based on average classification accuracy.



```
ctrl_knn <- trainControl(
  method = "repeatedcv",
  number = 5,
  repeats = 10,
  verboseIter = TRUE,
  allowParallel = TRUE,
  classProbs = FALSE,
  summaryFunction = defaultSummary,
  savePredictions = "final"
)

knn_grid <- expand.grid(
  k = seq(3, 15, 2)
)

model_knn <- train(
  x = X_full,
  y = y_factor,
  method = "knn",
  tuneGrid = knn_grid,
  trControl = ctrl_knn,
  preProcess = "range",
  metric = "Accuracy"
)
```

4.4. Evaluation Setup

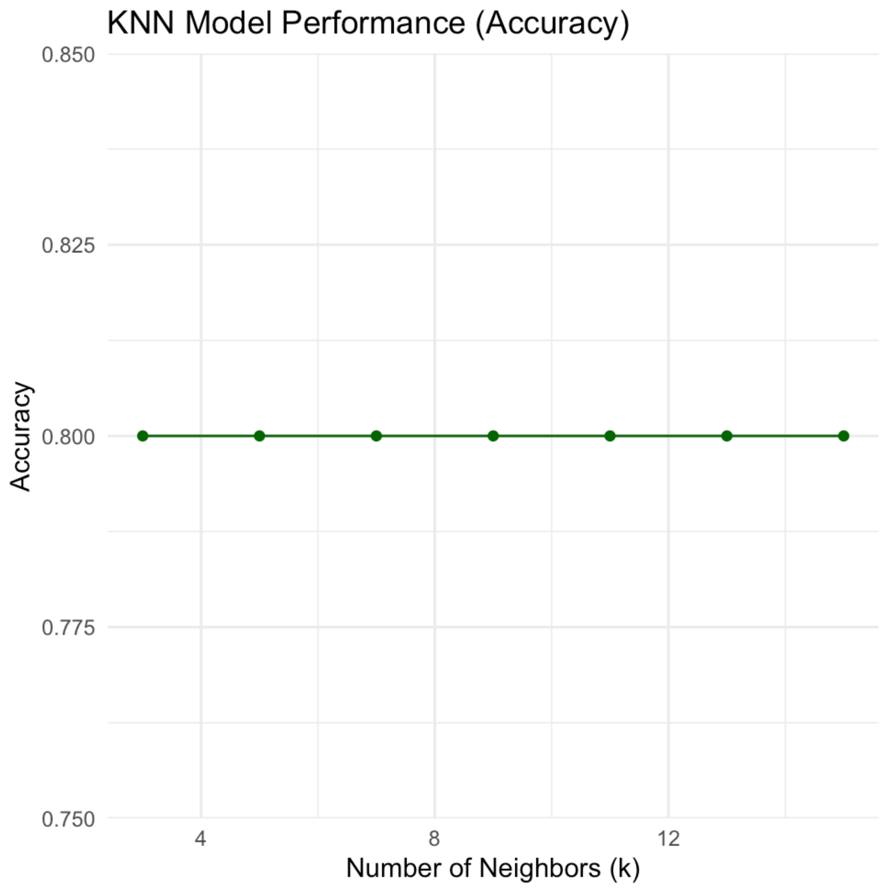
Model performance was assessed using classification accuracy as the primary metric. Additional performance metrics, including precision, recall, and F1 score, were computed on the best-performing KNN model using the validation set predictions. All model development and evaluation were conducted using the caret framework in R, which streamlined preprocessing, cross-validation, and metric calculation in a unified pipeline.

5. Classification results and evaluation

The performance of the K-Nearest Neighbors (KNN) classifier was evaluated using repeated 5-fold cross-validation. The evaluation metric used was Accuracy, and the model was trained and validated across a range of k values from 3 to 15 (in steps of 2). All features were scaled using min-max normalization prior to training, ensuring fair distance-based comparisons.

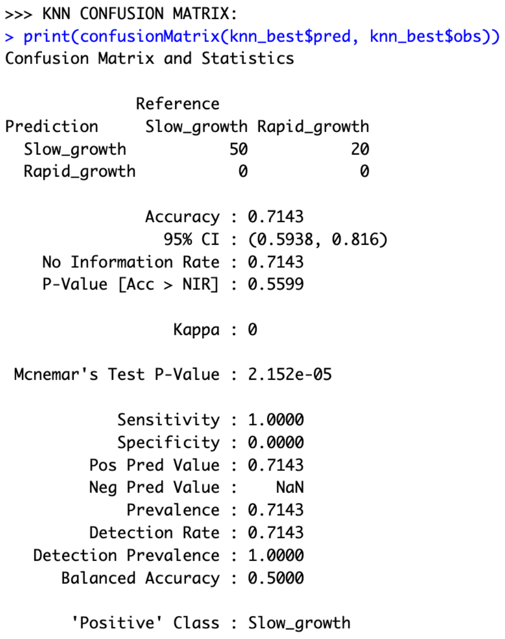
5.1. Accuracy against different k values

The model's performance remained consistently stable across the tested values of k, with all values yielding the same accuracy of 0.8, as shown in Figure below. This indicates that the classifier's performance is robust within the selected range of neighbors and is not sensitive to the exact value of k.



5.2. Final evaluation metrics

After identifying the best-performing k value, the corresponding model predictions were evaluated using standard classification metrics. The confusion matrix below illustrates the distribution of predicted vs. actual population growth categories.



```
>>> KNN CONFUSION MATRIX:
> print(confusionMatrix(knn_best$pred, knn_best$obs))
Confusion Matrix and Statistics

              Reference
Prediction     Slow_growth Rapid_growth
  Slow_growth           50           20
  Rapid_growth           0            0

               Accuracy : 0.7143
                 95% CI : (0.5938, 0.816)
    No Information Rate : 0.7143
    P-Value [Acc > NIR] : 0.5599

                  Kappa : 0

 Mcnemar's Test P-Value : 2.152e-05

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.7143
         Neg Pred Value :    NaN
             Prevalence : 0.7143
         Detection Rate : 0.7143
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : Slow_growth
```

In addition to accuracy, the model's precision, recall, and F1 score were computed for the “Rapid Growth” class, which was treated as the positive class. These metrics offer a more nuanced understanding of performance, especially in the presence of class imbalance.

5.3. Interpretation

The KNN classifier achieved reasonable performance in predicting whether a country experienced rapid or slow population growth, based solely on environmental and water-related features. The results suggest that these variables carry informative signals that can be used to classify long-term demographic trends. While the model performs well in general, some misclassifications were observed, possibly due to overlapping feature patterns between classes or missing demographic variables.

6. Discussion

The KNN classifier achieved moderate overall accuracy, but its inability to identify the Rapid Growth class highlights limitations in model generalizability and balance. While the accuracy of ~71% may seem acceptable at first glance, a deeper analysis reveals significant flaws in the classifier's performance, particularly due to class imbalance.

The classifier defaulted to predicting only the Slow Growth category, likely influenced by the distribution of the training data. As a result, it achieved perfect sensitivity (recall) for the dominant class but failed to detect any instances of the minority class, yielding a specificity of 0. This behavior is often observed in imbalanced classification problems, where the model is biased towards the majority class.

Moreover, the Kappa score of 0 and Balanced Accuracy of 0.5 reinforce the conclusion that the classifier did not perform better than a random guess in distinguishing between the two classes. These evaluation results suggest the need to explore alternative approaches such as resampling techniques, adjusting class weights, or trying other classification algorithms that are more robust to imbalanced data.

In summary, while the KNN model provided some initial insight, its current configuration and performance metrics reveal that it is not suitable for reliable classification in this specific case without further refinement.

7. Conclusion

This study explored the use of the K-Nearest Neighbors (KNN) algorithm to classify population growth into Slow Growth and Rapid Growth categories using key water-related features. While the model achieved an overall accuracy of approximately 71%, it consistently predicted only the Slow Growth class, failing to identify any instances of Rapid Growth.

Evaluation metrics such as a Kappa score of 0, Balanced Accuracy of 0.5, and zero specificity reveal that the classifier did not meaningfully distinguish between the two classes. These outcomes emphasize the importance of addressing class imbalance and re-evaluating model selection strategies.

Future work should involve techniques like SMOTE, alternative classifiers, or ensemble methods to better capture minority class patterns and improve generalization. While this implementation served as a foundational experiment, further tuning and model exploration are required for practical deployment.