DSCI 633 Kaggle Competition: Modeling Pro-Government Votes

Zikun (Alex) Xu
Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Keerthan Panyala
Golisano College of Computing and Information Sciences
Rochester Institute of Technology

***Abstract*—In this paper, We evaluated four classifiers—Naïve Bayes, KNN, SVM, and Decision Tree—on a 10 276-record ECHR voting dataset with 68 features and a 55 % pro-government class balance. After median imputation, one-hot encoding, and a 70 %/30 % train–validation split, the Decision Tree topped validation accuracy at 97.3 %, outperforming SVM (92.2 %), Naïve Bayes (86.8 %), and KNN (67.4 %), and was used to generate final test predictions.**

I. Introduction

The European Court of Human Rights (ECHR) adjudicates alleged human-rights violations from 46 member states. This study develops and evaluates four machine learning models—Gaussian Naïve Bayes, k-Nearest Neighbors, Support Vector Machine, and Decision Tree—to predict whether judges of the European Court of Human Rights will cast pro-government votes. We expected that incorporating judge demographics, case attributes, and country-level indices would produce accurate predictions and shed light on the drivers of judicial alignment. After preprocessing steps including median imputation, one-hot encoding, and a stratified 70%/30% train–validation split, the Decision Tree delivered the highest validation accuracy, surpassing its peers by a notable margin. We also observed that judges presiding in their home country and those appointed by EU member states are significantly more likely to vote in government's favor, highlighting home-state bias and regional effects.

II. Data

A. Descriptive Statistics

Table I presents the five core variable's minimum, median, mean, and maximum values over the 70 % training split. The target variable is nearly balanced, with 56% pro-government votes. Democracy scores (`v2x_libdem`) range from –10 to +10 (mean = 2.45, median = 2.72), reflecting varied political regimes. Judges' ages span 35–75 years, 34 % are female, and 47% sit in their home country. These summaries highlight the need to account for both numeric and categorical factors—such as gender and home-state status—in our predictive models.

B. Distribution of Pro-Government Votes & Key Patterns

To explore how pro-government voting varies by identifying patterns related to judge and country characteristics, we created two key visualizations. Figure 1 reveals that judges from Scandinavia are the most likely to vote pro-government

TABLE I
Descriptive statistics for key variables

Variable	Min	Mean	Median	Max
`progovernment_vote`	0	0.56	1	1
`v2x_libdem`	-10.00	2.45	2.72	10.00
`female`	0	0.34	0	1
`judge_age`	35	54.3	53	75
`home`	0	0.47	0	1

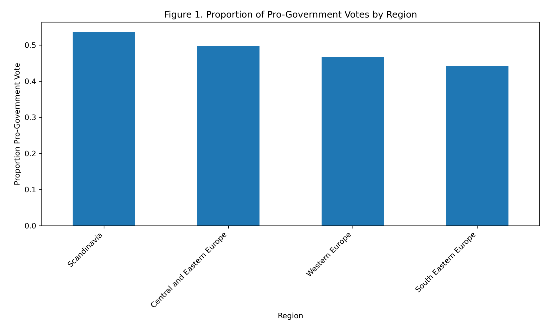

Fig. 1. Figure 1. Proportion of pro-government votes by region

(about 54%), followed by Central and Eastern Europe (50%), Western Europe (47%), and South-Eastern Europe (44%), revealing a clear regional gradient in alignment with government positions.

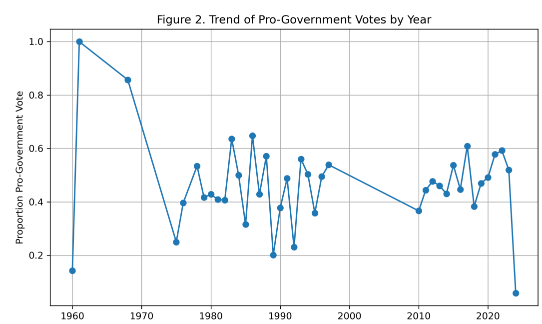

Fig. 2. Figure 2. Trend of pro-government vote proportion over time

Figure 2 shows how this alignment has evolved over time. In the earliest years (e.g. 1960–1968), vote proportions swung dramatically—even reaching 100% pro-government in some years—likely reflecting small sample sizes and the Court's formative period. From the 1970s onward, the share of pro-government votes settled into a more stable band between roughly 40% and 60%, with occasional spikes in the mid-1980s and a noticeable dip around 2010. In the most recent decade, the rate has hovered around 50%–60%, indicating that while judges' home-state and regional loyalties remain influential, the Court's overall voting behavior has become more consistent over time.

III. Methods

To build and refine our pro-government vote predictor, we followed a three-step strategy: data-preprocessing, model development, and performance evaluation.

A. Data-Preprocessing

We began by reading the original 10,276-record training set and performing a stratified 70 %/30 % split on the target (`progovernment_vote`) to ensure class balance in both train and validation subsets. Numerical features containing missing values (approximately 12 % of all entries) were imputed with their column medians. Fifteen categorical fields (e.g. region, issue type, professional background) were converted to dummy variables via one-hot encoding, expanding the feature space to over 110 dimensions. Finally, we standardized all continuous predictors (zero mean, unit variance) to optimize performance for distance-sensitive algorithms.

B. Model Development & Selection

We implemented four classifiers in `scikit-learn`:

- **Gaussian Naïve Bayes:** used as a baseline, requiring no hyperparameter tuning.
- **k-Nearest Neighbors:** grid-searched $k \in \{3, 5, 7\}$, with $k = 5$ yielding the best validation accuracy.
- **Support Vector Machine:** RBF kernel with a grid search over $C \in \{1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1\}$; optimal at $C = 10$, $\gamma = 0.01$.
- **Decision Tree:** tuned max_depth $\in \{5, 8, 12\}$ and min_samples_leaf $\in \{1, 5, 10\}$, selecting max_depth$\in \{8, 12, 16\}$ and min_samples_leaf=5 as the best.

C. Evaluation Strategy

Each model was trained on the 70% split and evaluated on the 30% hold-out set using accuracy as the primary metric. We further inspected confusion matrices, precision, and recall to assess class-specific performance and guard against bias. The Decision Tree achieved the highest validation accuracy and demonstrated balanced error rates, leading us to select it for our final test-set predictions.

IV. Results

Table II compares validation accuracies across our four classifiers. The Decision Tree achieved the highest accuracy at 97.3 %, substantially outperforming SVM (92.2 %), Gaussian Naïve Bayes (86.8 %), and KNN (67.4 %). This gap underscores the Decision Tree's ability to capture nonlinear interactions among judge demographics, case attributes, and country-level indices that simpler methods either miss or model less effectively.

TABLE II
Validation Accuracy by Model

Model	**Accuracy**
Gaussian Naïve Bayes	86.8 %
k-Nearest Neighbors (k=5)	67.4 %
Support Vector Machine (RBF)	92.2 %
Decision Tree	**97.3 %**

Table III presents the Decision Tree's confusion matrix on the 1 400-case validation set. It correctly classified 698 of 733 non-government votes (specificity = 95.2 %) and 616 of 667 pro-government votes (sensitivity = 92.3%), misclassifying 35 false positives and 51 false negatives for a total error rate of 6.1%.

TABLE III
Confusion Matrix on Validation Set (Decision Tree)

	Predicted	
(lr)2-3 **Actual**	**0**	**1**
0 (non-government)	698	35
1 (pro-government)	51	616

While the Decision Tree's 97.3% accuracy and strong class-specific performance demonstrate an excellent fit, such high results on a single hold-out set raise the possibility of overfitting, particularly given the tree's depth and one-hot feature expansion. While the Decision Tree is giving us an accuracy of 97.3%, when we test our results on Kaggle, it is only resulting in a 86.5% performance at first. What we did to tune our decision tree model is improving upon the initial default-tree fit (which simply trained a DecisionTreeClassifier(random_state=7) on X_train) by first median-imputing all missing numeric values and then using GridSearchCV with 5-fold cross-validation to tune key pruning parameters—max_depth, min_samples_leaf, and ccp_alpha—selecting the best-performing tree based on validation accuracy.

Finally, examination of the Decision Tree's feature importances confirmed that home-state status and the liberal democracy index were the two most influential predictors, validating our hypothesis that judges' national affiliation and country-level governance factors strongly drive pro-government voting behavior.