# Classification of Congressional Voting Records

1984 US House of Representatives Dataset

Karim Ahmed Zaki Badawi

February 10, 2026

## 1 Introduction

In this project, I developed a machine learning model to classify political party affiliation (Democrat vs. Republican) based on voting records. The dataset consists of votes on 16 different key issues. This report outlines the technical approach, the challenges faced regarding data integrity, and the final results.

## 2 Data Preprocessing

The raw data presented several challenges that required careful cleaning before modeling:

- **Label Standardization**: I mapped inconsistent target labels (e.g., *'REP', 'repub', 'REP'*) to a unified *'Republican'* and *'Democrat'* format.
- **Noise Filtering**: I identified numeric outliers such as $-999.0$ and $-998.515$ within the categorical voting columns and treated them as missing values ($NaN$).
- **Encoding**: Categorical votes ('y', 'n') were encoded into numerical values ($1, 0$) to be compatible with mathematical models.

## 3 Methodology and The Data Leakage Fix

A critical part of my workflow involved addressing **Data Leakage**. In my initial implementation, I applied imputation to the entire dataset before splitting. I realized this was a flaw because the training set would inherit statistical properties from the test set.

### 3.1 Corrected Pipeline

To ensure the integrity of the model, I updated the pipeline as follows:

1. **Split First**: I separated the data into training (80%) and testing (20%) sets.
2. **Local Imputation**: I calculated the *mode* of each voting column based on the party affiliation within the **Training Set only**.
3. **Transformation**: I used these training-derived modes to fill missing values in both the training and test sets. This ensures the test set remains truly "unseen" data.

## 4 Model Selection

I utilized an **Ensemble Voting Classifier** to provide a robust prediction. The ensemble consists of:

- **Logistic Regression**: Chosen for its efficiency in linearly separable datasets.
- **Random Forest**: Utilized to capture non-linear relationships and provide feature stability.
- **Support Vector Machine (SVM)**: Used with a linear kernel to maximize the margin between the two classes.

## 5 Results

The model achieved a perfect accuracy score. This result is attributed to the high polarization of the 98th Congress, where certain bills served as nearly perfect predictors of party identity.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Democrat | 1.00 | 1.00 | 1.00 | 50 |
| Republican | 1.00 | 1.00 | 1.00 | 33 |
| **Accuracy** | | | **100.00%** | 83 |

Table 1: Classification Report Results

## 6 Conclusion

By correcting the data leakage issue and implementing a robust ensemble strategy, I successfully built a model capable of predicting political affiliation with total accuracy for this dataset. This project highlights the importance of rigorous validation and proper preprocessing in the machine learning lifecycle.