Congressional Voting Pattern Classification Report

1 Dataset Overview

1.1 Dataset Description

Congressional voting records (1984 House of Representatives) containing 457 initial samples across 17 features representing votes on key legislative issues.

1.2 Data Quality Issues

The dataset presented several challenges requiring preprocessing:

- **Missing Data:** Crime column entirely null (dropped); Handicapped-infants column only 5% complete (25/457 values, dropped)
- **Inconsistent Labeling:** Party labels varied (rep, repub, Republican; dem, demo, Democrat) - standardized to lowercase + Replaced with proper labels
- **Missing Values:** Represented as '?' in voting columns - replaced with mode imputation
- **Null Party Labels:** 43 records with missing party affiliation were removed

1.3 Final Dataset

After cleaning: 414 samples, 14 voting features. Class distribution: Democrat (249, 60.1%), Republican (165, 39.9%). Train/test split: 331/83 (80/20, stratified).

2 Methodology

2.1 Problem Formulation

Binary classification task predicting party affiliation (Democrat vs Republican) from congressional voting records.

2.2 Data Preprocessing

Challenges Addressed:

- **Feature Reduction:** Dropped Crime (100% null) and Handicapped-infants (95% null, only 25/457 values)
- **Label Standardization:** Unified inconsistent party labels (rep/repub/Republican → republican; dem/demo/Democrat → democrat)

- **Missing Values:** Replaced '?' with mode imputation (mean inappropriate for binary 0/1 data)
- **Encoding:** Votes encoded as binary (y→1, n→0); Party encoded via LabelEncoder
- **Data Reduction:** 457→414 samples after removing 43 records with null party labels

2.3 Model Architecture (Neural Network)

Proposed architectures: (16→7→1) for compact representation vs (14→8→1) after feature reduction. Final choice: simpler (14→8→1) network - dataset complexity doesn't justify larger capacity.

Training Configuration:

- Loss: Binary Cross-Entropy with Logits (BCEWithLogitsLoss - numerically stable)
- Optimizer: Adam (adaptive learning rates)
- Epochs: 200

2.4 Comparison Models

To benchmark neural network performance, simpler interpretable models were evaluated: Logistic Regression, Random Forest, SVM, XGBoost, Gradient Boosting, and ensemble methods (Stacking/Voting). All models used cross-validation and hyperparameter tuning.

3 Results

3.1 Model Performance

Table 1: Model Comparison Results

Model	**Test Acc (%)**	**CV Mean (%)**	**CV Std (%)**
Logistic Regression	**97.59**	95.15	3.10
Stacking Ensemble	**97.59**	95.40	1.44
Voting Ensemble	96.39	95.16	1.35
Tuned XGBoost	96.39	94.20	1.63
Tuned Random Forest	95.18	95.16	2.57
Gradient Boosting	95.18	94.92	1.81

4 Best Model: Logistic Regression

Test Accuracy: 97.59% (tied with Stacking Ensemble)

Table 2: Classification Report

Class	Precision	Recall	F1-Score	Support
Democrat	0.98	0.98	0.98	50
Republican	0.97	0.97	0.97	33
Accuracy			**0.976**	**83**

Confusion Matrix:

```
           Predicted
          Dem  Rep
Actual Dem  49   1
       Rep   1  32
```

5 Feature Importance

Top predictive features (Pearson correlation with party):

1. Physician Fee Freeze (0.90) - strongest predictor
2. Adoption of Budget Resolution (-0.72)
3. Education Spending (0.66)
4. El Salvador Aid (0.66)
5. Anti-Satellite Test Ban (-0.60)

Key Voting Pattern: On Physician Fee Freeze, Republicans voted yes 96.3% of the time while Democrats voted yes only 5.6% of the time, demonstrating clear partisan division.

6 Conclusion

Logistic Regression achieved the highest test accuracy of 97.59% (tied with Stacking Ensemble) in predicting congressional party affiliation from voting patterns. The simpler Logistic Regression model is preferred due to its interpretability and equal performance to more complex ensemble methods. Average top-3 model accuracy was 97.19%.