Predicting Political Leaning Using AITA questions

Juo Ru Faustina Chong

Link to code: https://github.com/Faustch/AITA-decision-tree

Introduction

This report investigates whether decision trees can predict individuals' self-identified political leanings based on response to various Am I the Asshole questions (AITA) from these 2 data sets - Dataset Generation Fall2025 and Dataset Generation (Max). To assess whether decision trees can predict political leanings, the questions that most strongly correlate with political orientation were identified. A second analysis was conducted where respondents with unclear political self-identification were removed from the dataset.

Overview

Data Cleaning Max and Fall.

After performing data cleaning, the dataset size was 96 rows for the Max dataset and 148 rows for the Fall dataset. Rows which were completely empty or had less than 10 answered AITA questions were dropped. Some responses in the AITA scenario questions were missing. For each missing AITA question response, the distribution of responses within the respondent's political group was calculated. For example if 70% of Strong Liberals answered "Not a jerk" and 30% answered "Mildly a jerk" for "split the rent 50/50" question, and a Strong Liberal respondent had a missing value for that question, the imputation would be to randomly select between these 2 options with those probabilities. To ensure reproducibility, a fixed seed of 42 was used. This maintains the variability in responses while maintaining each group's distribution.

Data Makeup- Max

Dataset size: 96 entries after clean up.
Political make up:

- 34.38% Mildly liberal
- 28.13% Strongly liberal
- 16.67% Neutral
- 12.5% Mildly conservative

- 4.1667% Strongly conservative
- 4.1667% Don't know / It's complicated

Features: 14 AITA questions
Target: self politics (self-identified politics) which was simplified into the classes/ labels Liberal, Conservative, Neutral and Don't know/ It's complicated
Versions tested:

- Max (original with all categories)
- Max_Removed (excludes "Don't know/ It's complicated")
- Random Forest with Max_Removed

Dataset Makeup- Fall

Dataset size: 148 entries after clean up.
Political make up:

- 33.78% Mildly liberal
- 25.0% Strongly liberal
- 24.32% Neutral
- 10.81% Mildly conservative
- 3.38% Strongly conservative
- 2.70% Don't know / It's complicated

Features: 14 AITA questions
Target: self politics simplified into Liberal, Conservative, Neutral and Don't know/ It's complicated
Versions tested:

- Fall (original with all categories)
- Fall_Removed (excludes "Don't know/ It's complicated" entries)

Method for Decision Tree

Categorical variables were one-hot-encoded. Models were trained with 70/30 splits. The decision tree used Gini Impurity with depth limits of either 3 or 4 to prevent overfitting. The Random Forest model combined 200 trees with Gini criterion.

Findings

The decision tree and random forest models were evaluated using accuracy, precision, recall and F1 scores. Figures 1-5 show the confusion matrixes for each dataset and model.

Max Dataset Decision Tree

The accuracy of this model was 62.5%. Again, accuracy in classifying liberals was high while neutral and unaffiliated groups classification was poor.

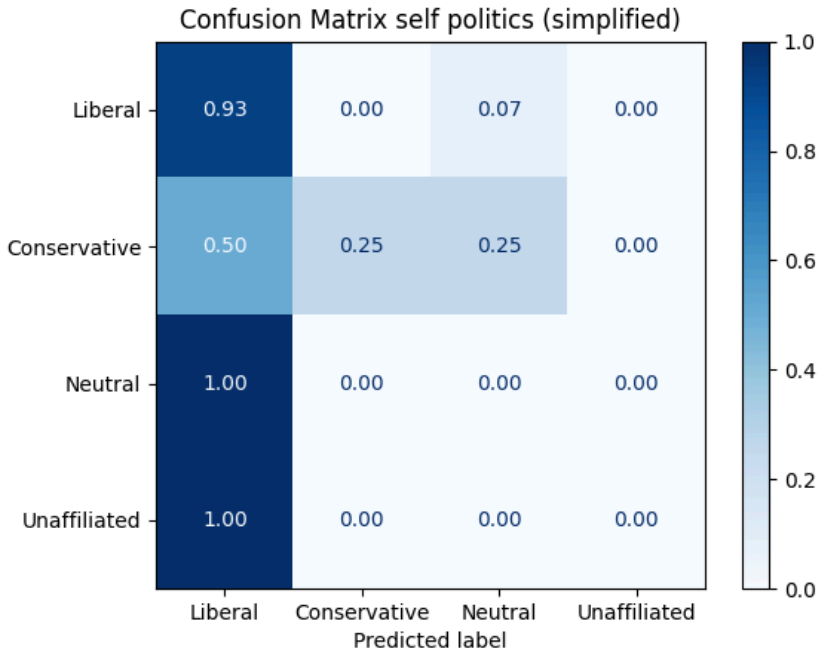

Fig. 1: Max dataset confusion matrix

Max_Removed Dataset Decision Tree

Accuracy of this decision tree trained on the Max_Removed dataset is 71.4%. After removing unknown politics respondents, accuracy improved by nearly 10%. There is a clearer separation between liberals and conservatives with liberals correctly classified 94% of the time. Neutral respondents remain the hardest to classify correctly. This is likely because their response incorporates elements from both ideological sides, resulting in less distinguishable patterns in their responses.

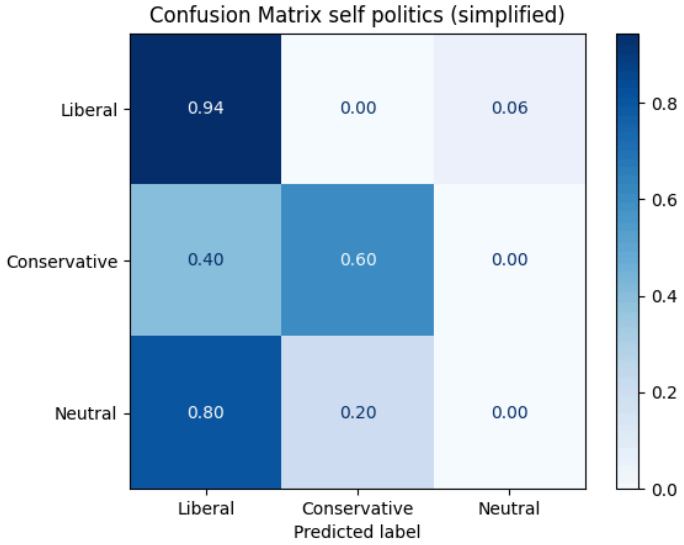

Fig. 2: Max_removed dataset confusion matrix

Fall dataset Decision tree

The fall dataset decision tree had an accuracy of 48.9%. The model predicted liberals the most accurately 18/27 but struggles with Conservatives and neutrals which are frequently misclassified as liberals. The large number of Liberal samples may have contributed to bias towards classifying samples as liberals.

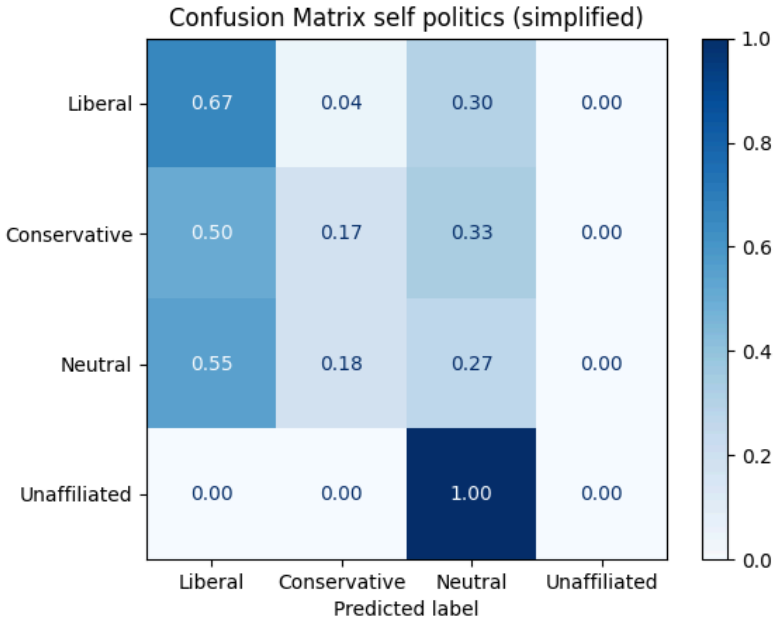

Fig. 3: Fall dataset confusion matrix

Fall Removed dataset Decision Tree

The accuracy of this model is 50.0%. Removing respondents who marked “Don’t know / It’s complicated” slightly improved model stability but made minimal improvements to accuracy. Some conservatives and Neutral respondents were still classified as liberal but this decreased as compared to the Fall dataset Decision tree.

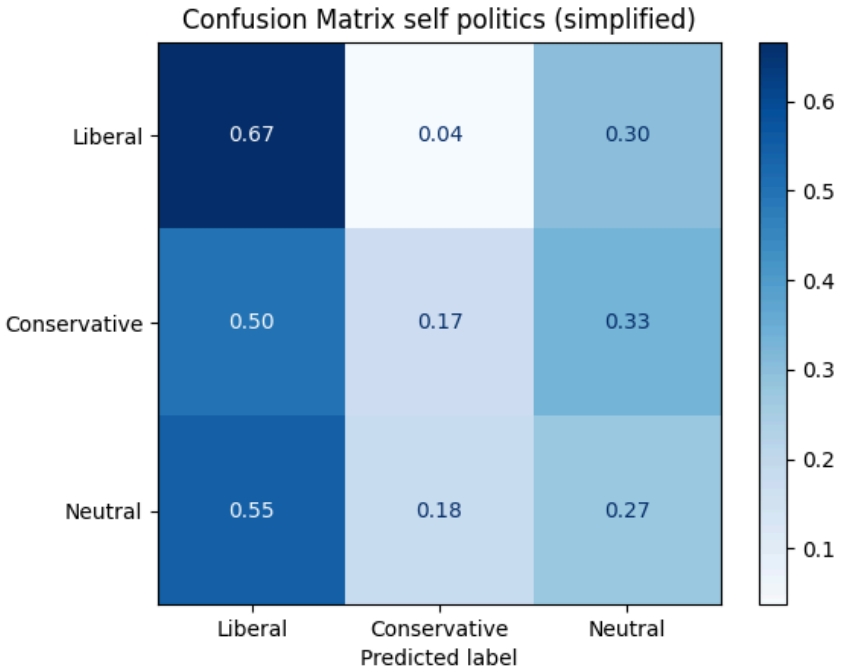

Fig. 4: Fall_removed dataset confusion matrix

Max_removed Dataset Random Forest

The Random forest’s model predicted with 64.3% accuracy. The random forest model was more robust overall but overfitted towards the liberal class and predicted liberal for all samples. It completely failed to predict conservatives or neutrals correctly.

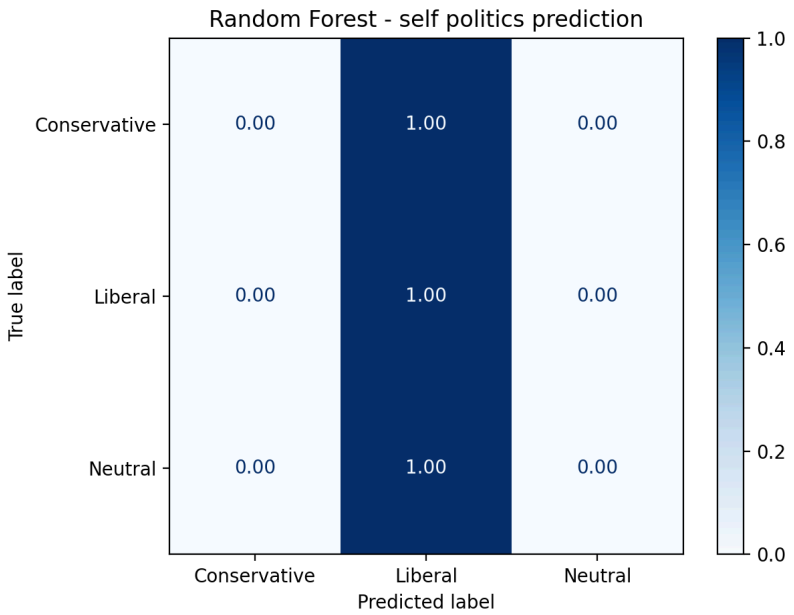

Fig. 5: Max_removed dataset random forest confusion matrix

Summary table of model F1 and accuracy scores:

Model	Accuracy	Liberal F1	Conservative F1	Neutral F1	Macro Avg
Fall	0.49	0.67	0.20	0.24	0.28
Fall removed	0.5	0.67	0.20	0.25	0.37
Max	0.625	0.78	0.40	0.00	0.29
Max removed	0.714	0.83	0.67	0.00	0.50
RF Max removed	0.64	0.78	0.00	0.00	0.26

The decision tree model for the Max_removed dataset achieved the highest accuracy of 71.4%. Precision and recall were highest for the liberal class, indicating that the model predicted liberals with high confidence. However, the neutral and conservative classes showed recall values near 0.

Top predictor AITA questions

- Donating to LGBTQ organization
- Falling out with mother-in-law's boyfriend
- Trust fund splits rent 50/50 with girlfriend
- Not walking daughter down the aisle

Conclusions

Problems with the decision tree

Most data points were liberal, which led the model to be biased towards classifying an entry as liberal. The Max dataset sample size was small (n=96) and for both datasets, there was a big difference in the number of liberals and conservatives. More balanced sampling could improve non-liberal class recall.

Simplifying political categories and removing respondents with “Don’t know / It’s complicated” as their political opinion improved performance. Though random forests provided slightly higher stability, decision trees offered higher accuracy.

Figures

Decision trees

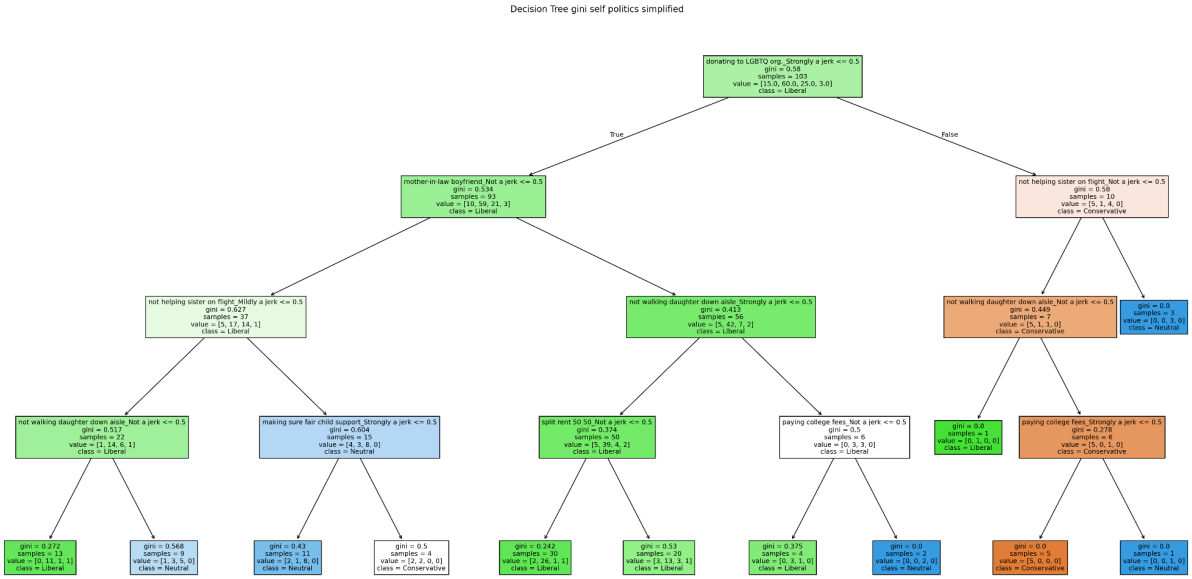

Fig.6 :Decision tree based on fall dataset

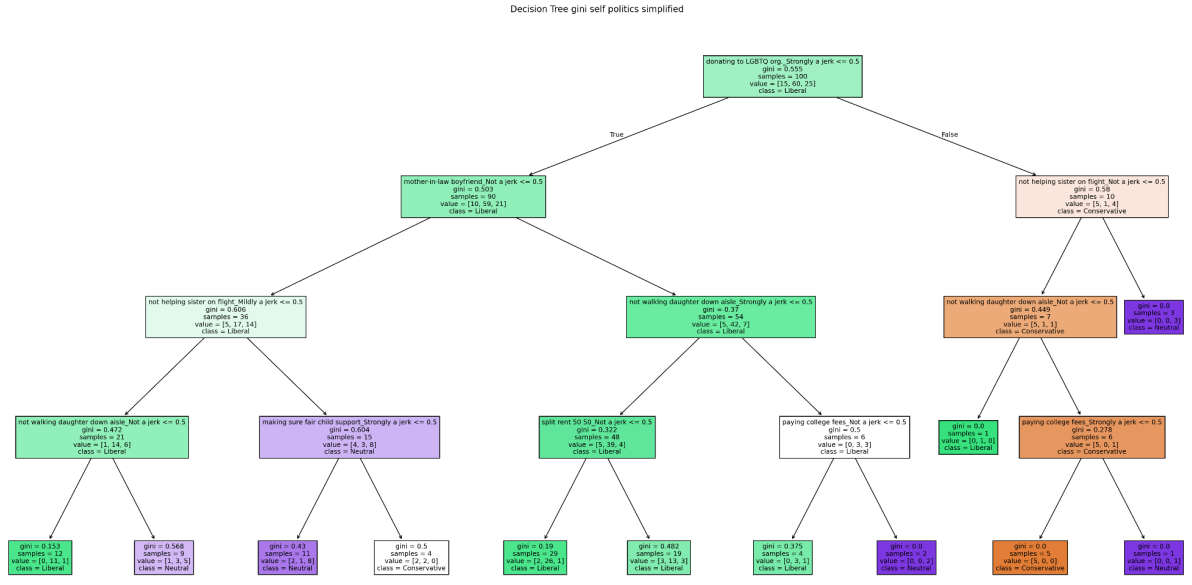

Fig. 7: Decision tree based on Fall_removed dataset

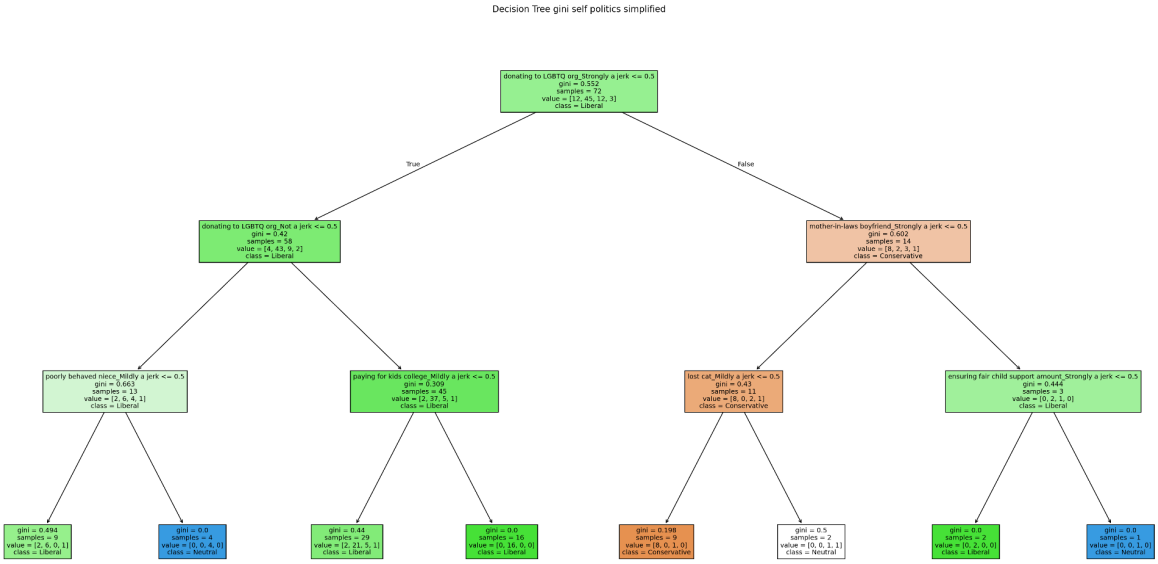

Fig. 8: decision tree based on Max dataset

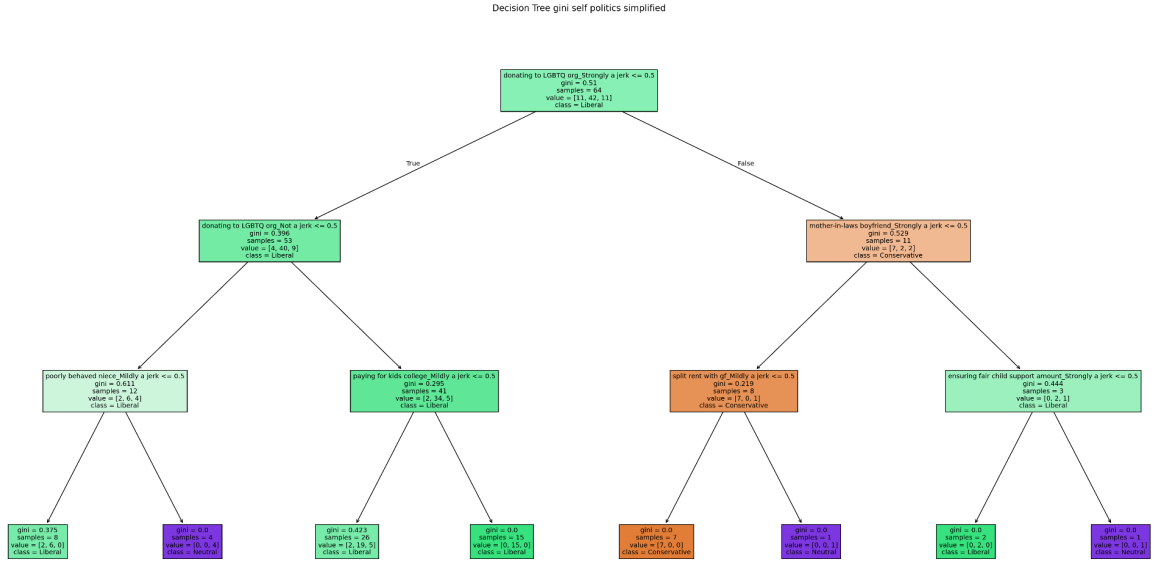

Fig. 9: Decision tree based on Max_removed dataset