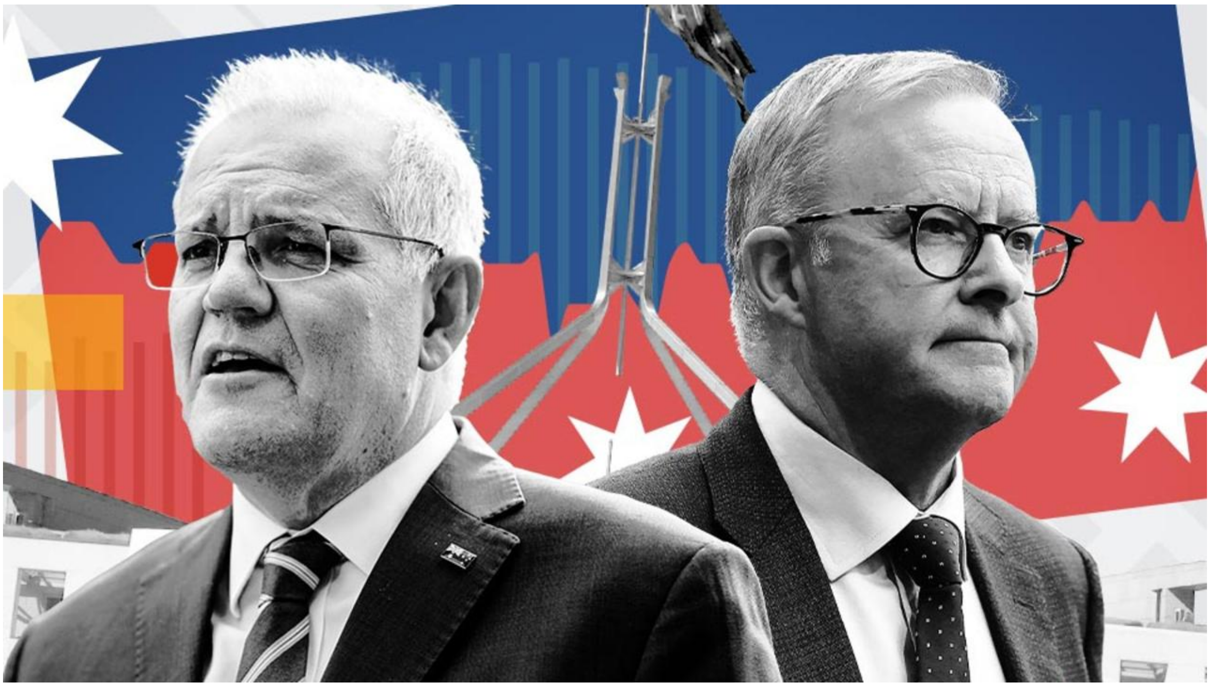
Capstone Project

"Mapping the Mind of the Electorate: A Machine Learning Analysis of the 2022 Australian Federal Election”



by I. Font

Table of Contents

Problem statement

Australia's major political parties — the Labor Party and the Coalition — have experienced a steady decline in primary vote share over recent elections, as more voters turn to minor parties and independents. This shift reflects not only dissatisfaction with traditional parties but also growing diversity in voter values, identities, and policy priorities.

Political campaigns, advocacy groups, and policymakers often lack granular, data-driven insight into who these voters are and what drives their choices. While campaign strategies still lean heavily on demographics and historic voting patterns, they often miss the evolving complexity of public opinion. Similarly, party policy platforms are frequently shaped through top-down agendas rather than grounded in the real, expressed concerns of voters. As a result, both engagement efforts and policy development may overlook key voter groups, leading to missed opportunities for representation and trust-building.

This project explores how survey-based clustering and machine learning classification can be used to better understand and predict voting behaviour. The current state is one of limited targeted precision and poor alignment between policy messaging and voter priorities. The desired state is a more nuanced, evidence-based understanding of voter segments that can improve both campaign effectiveness and the responsiveness of political platforms.

Previous work using Australian Election Study (AES) data has largely focused on sociological or academic insights. This project applies a machine learning lens to uncover patterns in voter attitudes, priorities, and demographics — translating complex survey data into insights that could support more inclusive campaigning and data-informed policy development.

Industry/ domain

This project sits within the space where data science meets politics, focusing on how we can use data to better understand voter behaviour. In Australia, as in many countries, the way people engage with politics is changing — traditional party loyalty is declining, and voters are more diverse and harder to predict. At the same time, there's growing interest in using data and technology to make sense of these changes.

Political organisations face several challenges, including voter distrust, fragmented data, and increasing electoral unpredictability. While this project mainly looks at elections, the techniques used — like clustering and classification — can be applied in other areas too, such as marketing, social research, and public policy, where it’s just as important to understand different groups and anticipate their behaviour.

Although this work focuses on vote choice, it also highlights how survey data like the AES can be used to go beyond basic demographics and help uncover what actually matters to different types of voters — something that’s relevant not only for campaigns, but also for policy design, community engagement, and broader democratic participation.

Stakeholders

- **Political campaign teams** (e.g. Labor, Liberal/National): need to understand which groups are likely to vote for them or swing.
- **Pollsters and political consultants**: want better predictive models than simple polling margins.
- **Policy makers**: can tailor policies based on voter segment priorities.
- **Public interest technology organisations**: may use insights to promote informed voter engagement.

These stakeholders care because **targeted engagement and policy alignment** can influence election outcomes, shape public trust, and reduce campaign waste. They expect clear voter segments, interpretable models, and actionable predictors of voting behaviour.

Business question

Which voter groups are most likely to support Labor, the Coalition, or minor parties and how do they differ in terms of demographics, values, and issue priorities?

Answering this question can help both campaign teams and policy makers better understand the changing political landscape. Campaigns can use these insights to improve how they engage voters, allocate resources, and shape messaging in targeted ways — especially in close electorates. At the same time, political parties and advocacy groups can use this information to design policies that are more aligned with what voters actually care about, rather than relying on assumptions or outdated segments.

While there's no fixed accuracy threshold, it's important that the models offer clear, interpretable results, particularly in distinguishing Labor, Coalition, and Other voters. Misclassifying voter groups could lead to missed opportunities, ineffective communication, or policies that fail to connect with key segments of the electorate.

Data question

Can voter support for Labor, Coalition, or Other be predicted using survey data on demographics, political attitudes, and issue priorities?

To answer this, the project uses the 2022 Australian Election Study (AES), a nationally representative survey with 350+ variables covering, amongst other topics:

- Demographics (age, gender, education, income, region)
- Political attitudes (trust, satisfaction, ideology)
- Issue salience (top concerns, spending preferences)
- Voting behaviour (first preference vote)

The analysis involves subsetting relevant variables, cleaning and encoding data, applying clustering to identify voter segments, and building classification models to predict vote choice.

Data

The dataset used comes from the 2022 Australian Election Study (AES), which is publicly available through the Australian Data Archive (ADA) at the Australian National University. It's a large, nationally representative survey designed to capture people's views on politics, voting, and key social issues. The dataset includes over 2,500 responses and more than 350 variables covering things like party preferences, demographics, political engagement, trust in institutions, and issue-based attitudes. 29 features were chosen for this analysis. It's a really rich source for exploring patterns in voter behaviour.

The data is considered highly reliable and is widely used in political science research in Australia. It's collected using proper sampling techniques, including stratified sampling and weighting, so it reflects the broader Australian population quite well. The raw data is generally clean and well-documented, which made it easier to work with. The AES is conducted after every federal election, so the data is available on a regular basis, making it useful for tracking changes in public opinion over time.

Data science process

Data analysis

The data pipeline used to wrangle the raw data involved a series of steps to clean, simplify, and prepare the 2022 Australian Election Study dataset for analysis. First, columns were renamed to more descriptive labels. Then, non-voters and rows with too many missing values were removed, reducing the total number of samples from 2508 to 2430. Missing values were then handled based on the type and distribution of the variable. Columns with low missing counts were filled using the mode for categorical features or the median for ordinal features. For variables with higher counts of missing values, such as party_support_strength, missing values were treated as a valid category ("missing") to retain potentially meaningful patterns. In the case of household_income, missing values were imputed using the median within grouped subsets based on age_group, has_uni_degree, and social_class, to improve accuracy.

Feature engineering was also part of the pipeline. Variables like party_id and first_pref_vote were recoded to simplify the analysis, reducing the categories to just three: the Coalition, Labor, and Other, as shown in the table below:

New code	Party
1	Coalition
2	Labor
3	Other

The income variable was re-mapped from 13 original brackets into 5 broader income groups that better align with Australian income tax bands, as shown in the table below:

New code	Bracket description
1	Under $15000
2	$15000 - $45000
3	$45000 - $120000
4	$120000 - $200000
5	Over $200000

Label dictionaries were created to map numerical codes into descriptive labels, aiding interpretation throughout the analysis.

The highlights of the exploratory data analysis (EDA) include the identification of demographic and political engagement trends among voters. For example, the dataset is heavily skewed towards older individuals, with 61.4% of the sample over 55 years old—significantly higher than their representation in the general adult population (see Fig. 1).

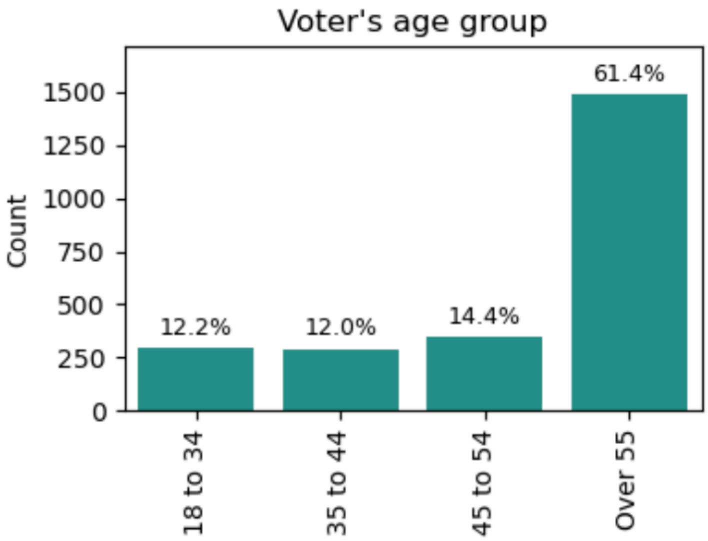


Fig. 1

Voter engagement levels are high, with over 84% of respondents expressing interest in politics, and issues such as cost of living, health, and global warming ranking as top concerns. Interestingly, while more respondents identified with the Coalition, Labor received a higher share of first preference votes—suggesting disillusionment with the government at the time.

Bivariate analysis highlighted meaningful associations between voter preferences and demographic or opinion-based variables. A Cramér's V heat-map showed the strongest association between the target variable first_pref_vote and party_id, which aligns with expectations (see Fig. 2). This variable was retained for clustering but was excluded from supervised modelling to avoid data leakage. Moderate associations were also observed between variables such as party_id and party_support_strength, and between several spending priority variables.

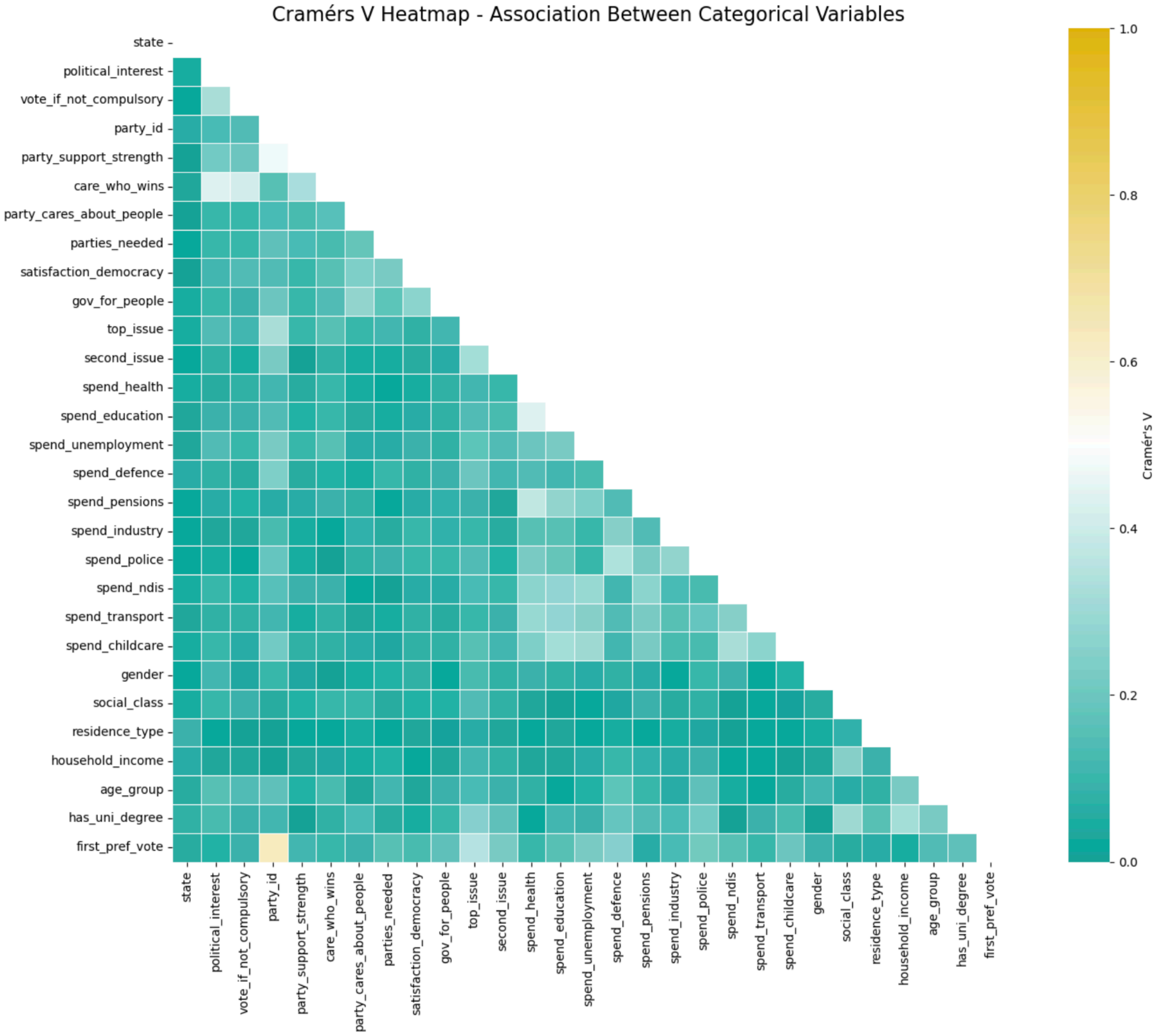


Fig. 2

Chi-squared tests confirmed that all features were statistically significant in relation to first_pref_vote ($p < 0.05$), with the top features including party_id, top_issue, and several spending priorities like spend_defence, spend_unemployment, and spend_childcare.

Further breakdowns revealed that Coalition voters prioritise the economy and national security, while Labor and Other voters are more concerned with issues like health, education, global warming, and social services (see Fig. 3).

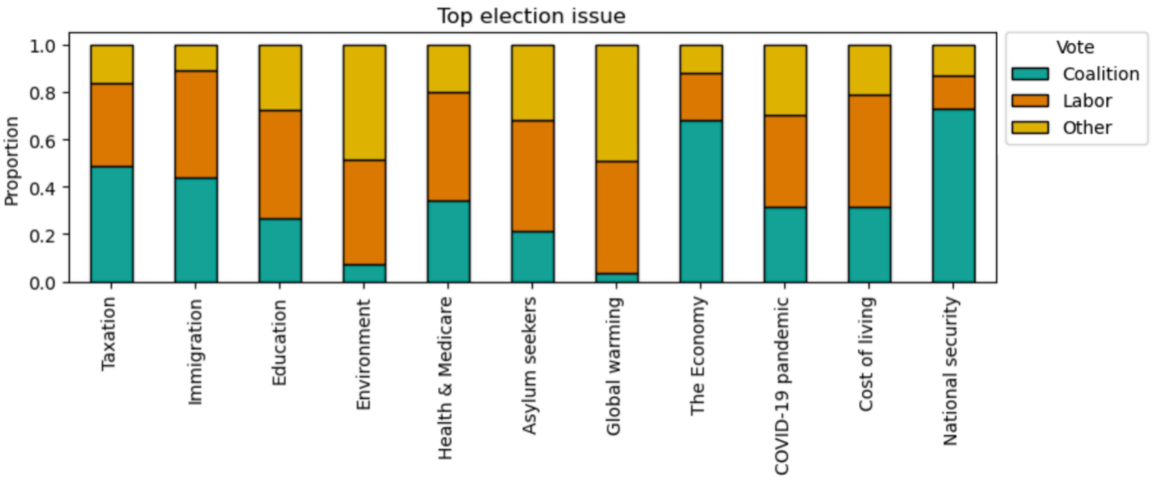
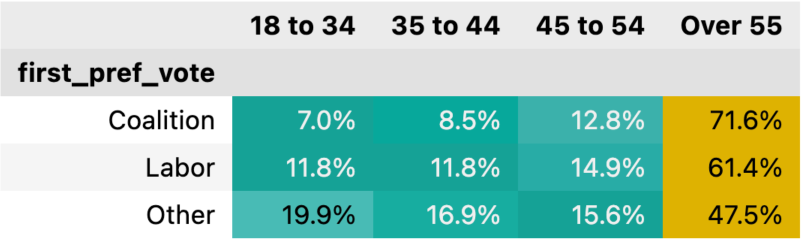


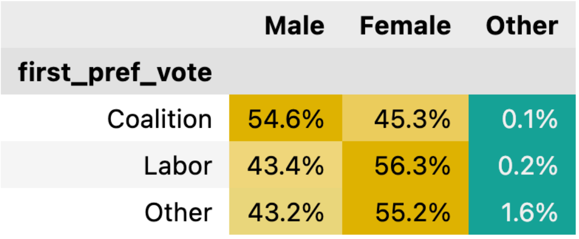
Fig. 3

Distinct voting patterns were also observed across demographics. Younger voters leaned more toward Labor and Other parties, while older voters preferred the Coalition (see Fig. 4). Female and university-educated voters were more likely to support Labor or Other parties, whereas male and less-educated voters leaned more toward the Coalition (see Fig. 5 and Fig. 6). Geographic differences also emerged, with Coalition support higher in rural areas and Labor and Other support higher in major cities (see Fig. 7).



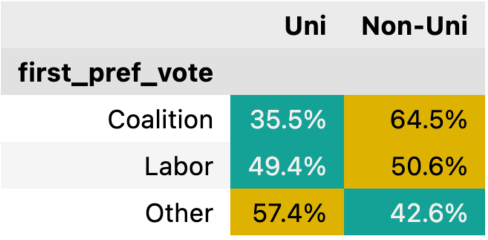
	18 to 34	35 to 44	45 to 54	Over 55
first_pref_vote				
Coalition	7.0%	8.5%	12.8%	71.6%
Labor	11.8%	11.8%	14.9%	61.4%
Other	19.9%	16.9%	15.6%	47.5%

Fig. 4



	Male	Female	Other
first_pref_vote			
Coalition	54.6%	45.3%	0.1%
Labor	43.4%	56.3%	0.2%
Other	43.2%	55.2%	1.6%

Fig. 5



	Uni	Non-Uni
first_pref_vote		
Coalition	35.5%	64.5%
Labor	49.4%	50.6%
Other	57.4%	42.6%

Fig. 6

These insights guided feature selection and interpretation in the clustering and modelling phases.

The pipeline is largely reusable. The sequence of preprocessing steps was structured in a clear, modular way using Jupyter Notebook cells. Functions were defined for tasks like calculating Cramér's V and generating visualisations, and label mappings were stored in a dictionary — making it easier to apply the same pipeline to updated versions of the AES dataset or other similar survey data.

Several intermediary data structures were used to support the workflow. These included cleaned or filtered DataFrames for specific parts of the EDA, such as contingency (crosstab) tables, grouped summaries, and subsets of features used for visualisations or bivariate analysis. Lookup dictionaries were also used to apply readable labels across plots and tables, improving clarity throughout the process.

Modelling – Clustering Analysis

To explore natural groupings among voters, clustering analysis was applied across four subsets of features: demographics, political attitudes and engagement, government spending preferences, and issue priorities. Each subset was selected to reflect a distinct dimension of voter behaviour and identity. The features were chosen based on thematic coherence and relevance to political segmentation, rather than statistical feature selection.

Given the nature of the data—mostly label-encoded categorical and Likert-scale ordinal variables— MinMaxScaler was applied to the first three subsets to normalise feature ranges. For the issue priority subset, which included two one-hot encoded categorical variables representing voters' top two election issues, scaling was not applied to preserve interpretability.

The Fuzzy C-Means (FCM) algorithm was used because it is well-suited to social science survey data where group membership can be ambiguous. FCM allows each voter to belong partially to multiple clusters, providing a more realistic picture of the electorate. To determine the number of clusters for each subset, models were evaluated with 2 to 9 clusters and the configuration that offered the highest interpretability and a Fuzzy Partition Coefficient (FPC) near 0.6 — indicating moderate but meaningful separation — was selected. Training time was very fast, with each model completing in under two seconds due to the small dataset size (n ≈ 2430) and low dimensionality per subset.

Model confidence was assessed by examining the distribution of maximum membership values across clusters, and hard labels were assigned based on the highest membership score. Then each cluster was profiled by comparing the mean values of features per cluster and linking them to voting behaviour where relevant. The resulting clusters revealed clear and interpretable segments such as rural, older, working-class Coalition voters, and younger, urban, university-educated voters leaning toward Labor or Other parties. Social and economic preferences further separated voters into groups prioritising social services versus national security.

Although FCM does not explicitly model interactions between features, some interesting patterns emerged across subsets. For instance, the demographic profile of urban, middle-class voters aligned closely with those prioritising climate change and supporting increased investment in health and education. Among all subsets, demographics produced the clearest structure, with the highest FPC (0.62), followed by political engagement and spending preferences (both around 0.56), while issue priority clusters showed more overlap (FPC ≈ 0.50).

All clustering was performed in a local Jupyter Notebook environment using Python 3.11 with libraries including scikit-learn, scikit-fuzzy, and pandas.

Modelling – Supervised learning

To prepare for supervised modelling, first recursive feature elimination (RFE) was used with multinomial logistic regression to rank the 27 input features by importance. The analysis suggested that all features contributed meaningfully to the model. However, to reduce complexity and improve interpretability, a two-sided t-test was performed to compare accuracy using all 27 features versus the top 20. Since the difference in performance was not statistically significant, the top 20 features were selected for the final models.

These features included key policy attitudes (e.g., spend_defence, spend_police, spend_unemployment), political engagement variables (party_support_strength, satisfaction_democracy), and demographic indicators (age_group, gender, social_class, has_uni_degree). Several interactions were observed between features — for example, younger, urban voters with progressive spending attitudes tended to prefer Labor or Other parties, while older voters prioritising defence and police spending were more likely to vote for the Coalition.

While the models did not rely on a single dominant feature, a core subset — particularly spend_defence, top_issue, spend_unemployment, and party_support_strength — consistently ranked highest across all classifiers and captured a significant portion of the model's discriminative power.

Feature engineering involved remapping categorical variables (e.g., vote choice, income brackets), scaling ordinal variables using MinMaxScaler, and ensuring all variables were numeric and consistently encoded. No one-hot encoding was used, as tree-based models handled label-encoded categorical data effectively and one-hot encoding significantly increased dimensionality.

Several models were tested:

- Multinomial Logistic Regression (interpretable baseline)
- Random Forest Classifier (with hyper-parameter tuning)
- XGBoost Classifier (with hyper-parameter tuning)

Additional tests included KNN, SVC, and ensemble methods (Stacking, AdaBoost), though these were not retained due to similar or lower performance.

Models were trained and evaluated on a stratified train-test split (80/20), and training time was minimal for all models due to the small dataset size (2,430 samples × 20 features). All work was performed locally using Python, Jupyter Notebook, and libraries including scikit-learn, xgboost, and matplotlib.

The main performance metrics used were accuracy, precision, recall, F1-score, and ROC AUC, calculated per class. Although XGBoost only marginally outperformed other models with an accuracy of 61%, it showed the most balanced classification across all three classes and the highest AUC scores — 0.86 for Coalition, 0.74 for Labor, and 0.75 for Other. For this reason, the XGBoost model was selected as the final predictive model.

Outcomes – Clustering Analysis

In all data subsets, two clusters were identified. Fuzzy Partition Coefficients (FPC) were moderate, with lowest value of 0.50 (issues - categorical variables) and highest value of 0.62 (demographics). This suggests that there is significant overlap between clusters, but the clusters are not completely indistinguishable. Data points may have partial membership in multiple clusters, which is expected in fuzzy clustering. The clusters are not perfectly distinct, but there is still some structure in the data that the algorithm has captured. This is quite common when analysing survey data or behavioural data.

Demographics clusters: Membership Confidence skewed towards higher values with most of them between 0.68 and 0.82.

> - Cluster 0 (1298 samples): This group is largely made up of older voters living in smaller towns or rural areas. They tend to have lower incomes, no university qualifications, and are more likely to identify as working class. Voters in this cluster showed stronger support for the Coalition.
>
> - Cluster 1 (1132 samples): This cluster primarily consists of younger, urban voters with higher incomes and university degrees. They are more likely to identify as middle class and tended to vote for Labor or Other parties.

Political engagement and attitudes clusters: Membership Confidence is bimodal, with most values between 0.50 and 0.58 or between 0.67 and 0.76.

> - Cluster 0 (1164 samples): This group consists mostly of voters who identify as moderate to strong supporters of Labor or Other parties. They tend to believe that political parties do not care much about ordinary people and are less likely than Cluster 1 voters to see political parties as essential to democracy.
>
> - Cluster 1 (1266 samples): This cluster is primarily made up of moderate Coalition supporters. Compared to Cluster 0, they report higher satisfaction with democracy, stronger belief in the importance of political parties, and are more likely to feel that parties care about the public.

Government spending clusters: Membership Confidence skewed towards lower values with most of them between 0.58 and 0.70.

> - Cluster 0 (1254 samples): Voters in this group favour increased government spending on social services such as health, education, unemployment benefits, the NDIS, and childcare. They were more likely to vote for Labor or Other parties.
>
> - Cluster 1 (1176 samples): This cluster prefers greater government investment in areas like defence, police, and business and industry. These voters showed stronger support for the Coalition.

Important issues to voters clusters: Membership Confidence values are all around 0.5. These are the clusters with the most overlaps.

> - Cluster 0 (866 samples): Voters in this group are primarily concerned with cost of living, health, and Medicare. Their priorities tend to focus on immediate, day-to-day issues.
>
> - Cluster 1 (1564 samples): This cluster is more focused on broader or long-term concerns such as global warming and the economy, while also ranking cost of living highly. Compared to Cluster 0, they also place slightly more importance on issues like the environment, immigration, and asylum seekers.

Although there is considerable overlap between voter segments, the clustering results suggest broad patterns in the types of voters attracted to each party in the 2022 election:

Voters attracted to the Coalition in the 2022 election were more likely to be older Australians living in rural or regional areas, with lower levels of formal education and household income, and often identifying as working class. They tended to express stronger trust in the party system and greater satisfaction with the way democracy is working. These voters prioritised government investment in areas like defence, police, and business and industry, and were generally less concerned about environmental and social issues. Their issue focus leaned toward national security, economic stability, and cost of living pressures.

Voters drawn to Labor and Other parties, including the Greens and independents, tended to be younger, more urban, university-educated, and middle-class, with higher income levels. They were often more skeptical of political parties and more dissatisfied with the state of democracy, yet showed higher engagement in political issues. These voters supported increased public spending on health, education, the NDIS, childcare, and social safety nets. Their top concerns included global warming, the environment, and social equity, while also acknowledging economic issues like cost of living. This group represented a more progressive and socially-oriented segment of the electorate.

Outcomes – Supervised learning

Multinomial Logistic Regression:

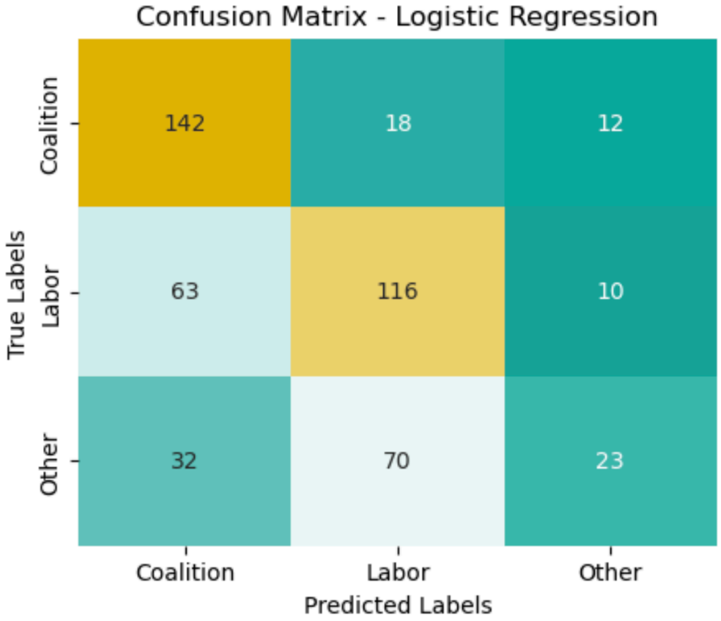
The multinomial logistic regression model was the most interpretable of the classifiers tested. Its coefficients revealed clear ideological, policy, and demographic distinctions between voter groups. Coalition voters were more likely to support increased defence and police spending, and showed less support for welfare, education, health, and childcare funding. They also tended to trust government more, believed political parties care about ordinary people, and expressed stronger party identification.

In contrast, Labor voters were more supportive of increased spending on social services, more likely to be working class, female, and in favour of programs like the NDIS and pensions. Other voters were typically younger, urban, and more concerned about climate change, public transport, and environmental issues, while showing greater dissatisfaction with democracy and skepticism toward political parties.

Notably, Coalition voters skewed older and male, while Labor and Other voters were more mixed across demographics but often shared progressive spending attitudes. Labor voters also expressed the strongest interest in which party won the election, whereas Coalition voters were the least concerned, highlighting differences in political engagement.

While optimising the classification thresholds improved the model's performance somewhat, predictive accuracy remained limited. The model achieved an overall accuracy of 58%, with strong recall for Coalition voters, moderate performance for Labor, and poor recall for Other voters — a segment that proved more difficult to capture across all models. Despite this, logistic regression provided valuable explanatory insight, making it well suited for profiling voter segments and understanding the key factors that drive political preferences.

A detailed breakdown of coefficients and their interpretation is provided in the table on next page.



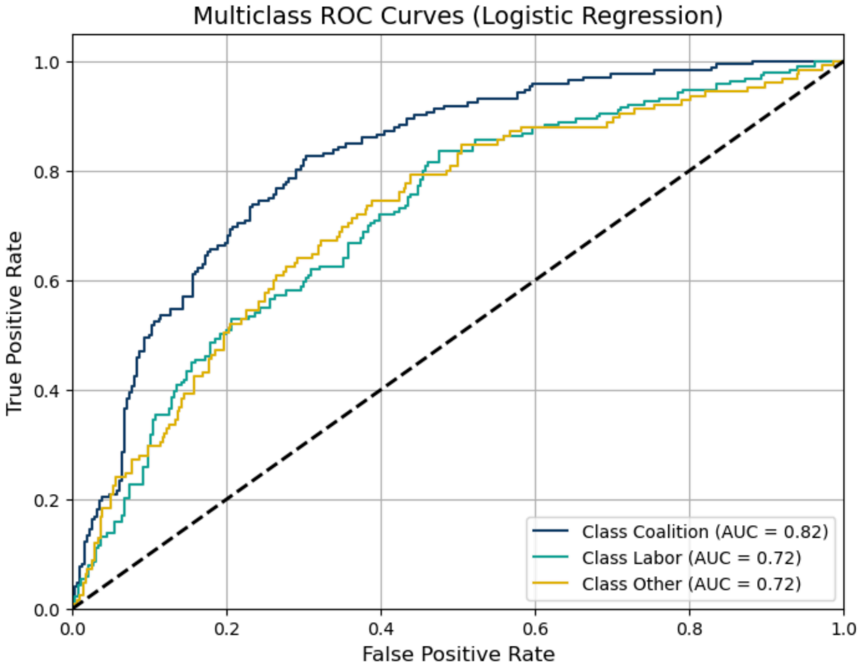


Table of coefficients

Feature	Coalition	Labor	Other	Interpretation
spend_defence	–1.37	0.42	0.95	Coalition voters support **more** defence spending; Other voters support **less**.
spend_police	–1.36	0.26	1.10	Strong **pro-police spending** sentiment among Coalition; Other voters oppose it.
spend_unemployment	1.33	–0.17	–1.18	Coalition voters prefer **less** welfare spending; Other voters strongly support it.
gov_for_people	1.03	–0.28	–0.75	Coalition voters more likely to believe government is run for the people.
spend_education	0.89	–0.29	–0.59	Coalition voters less supportive of increased education spending.
party_support_strength	–0.75	0.31	0.43	Coalition voters show stronger party identification.
spend_transport	0.55	0.18	–0.73	Other voters more supportive of increased public transport investment.
spend_childcare	0.73	–0.72	–0.01	Labor voters favor childcare spending; Coalition voters do not.
spend_industry	–0.71	0.58	0.14	Coalition voters support more industry spending; Labor and Other voters less supportive.
satisfaction_democracy	–0.07	–0.52	0.59	Other voters are most dissatisfied with democracy; Labor voters are the most satisfied.
spend_pensions	–0.24	–0.28	0.52	Labor and Coalition voters favour increased spending on pensions.
party_cares_about_people	–0.17	–0.24	0.41	Coalition voters more likely to believe parties care about ordinary people. Other voters think political parties do not care about them.
spend_health	0.43	–0.22	–0.22	Coalition voters less supportive of increased health spending.
age_group	0.42	0.07	–0.48	Coalition voters are older; Other voters skew younger.
spend_ndis	0.34	–0.38	–0.03	Labor voters more supportive of NDIS funding increases.
gender	–0.35	0.01	0.34	Coalition voters more likely to be male; Other voters more likely female.
social_class	–0.11	0.36	–0.26	Labor voters more likely to identify as working class. Other voters more likely to identify as upper class
parties_needed	–0.37	0.08	0.30	Other voters are more skeptical about the need for political parties.
top_issue	0.41	–0.12	–0.29	Coalition voters prioritize different issues than Labor/Other.
care_who_wins	0.38	–0.44	0.07	Labor voters are the most concerned about which party wins the election and Coalition voters are the least concerned.

Random Forest Classifier:

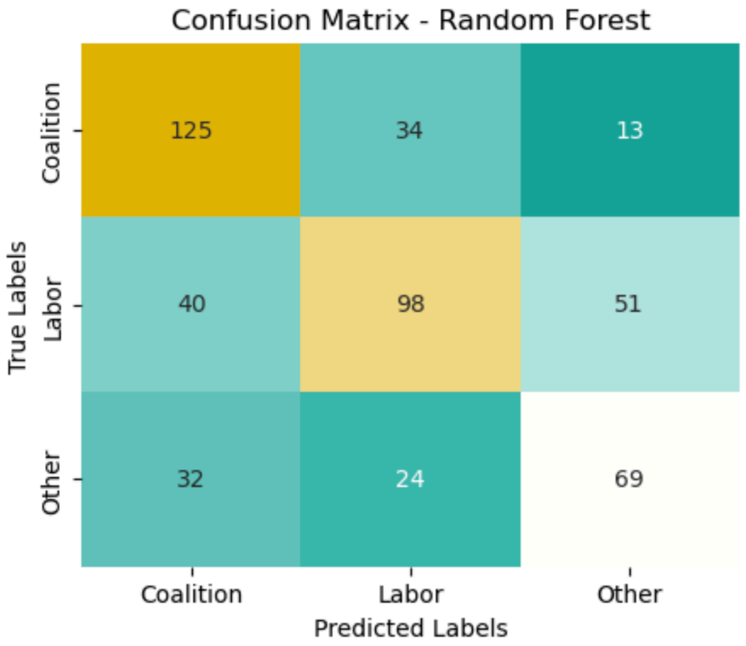
The Random Forest model provided a modest improvement in predictive performance over logistic regression, achieving an overall accuracy of 60% and a more balanced classification across all three voter groups. It produced F1 scores of 0.68 for Coalition, 0.57 for Labor, and 0.53 for Other voters. Most notably, it significantly improved recall for Other voters — reaching 55% compared to only 18% with logistic regression — suggesting that it was more effective at capturing the nuanced distinctions between Other and Labor voters, which the linear model struggled to detect.

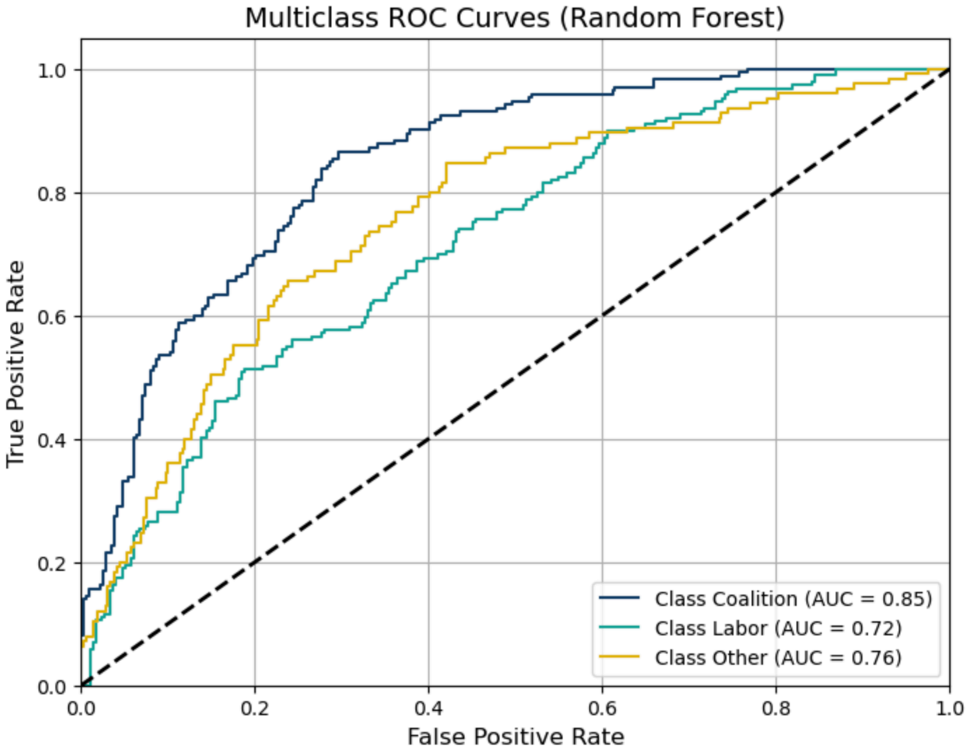
The model's ROC AUC scores also reflected this improvement in class separability, with values of 0.85 for Coalition, 0.72 for Labor, and 0.76 for Other, indicating strong discriminative ability for Coalition voters and moderate separability for the other two classes.

In terms of feature importance, Random Forest ranked several of the same features highly as logistic regression — particularly spend_defence, spend_police, and spend_unemployment. However, it placed significantly more emphasis on categorical issue-based variables such as top_issue and second_issue, which were less prominent in the logistic model. This difference likely stems from Random Forest's ability to model non-linear relationships and interactions between features, which are common in survey data.

The confusion matrix further supported these findings. While the model tended to misclassify some Coalition voters as Labor, it correctly reclassified many voters that logistic regression had misclassified as Labor into the "Other" category, providing a clearer distinction between those two groups.

Overall, the Random Forest model showed stronger performance on more complex or overlapping segments, particularly for minor party and independent voters. While it remains limited as a predictive tool, its flexibility and improved class balance make it a valuable complement to the more interpretable logistic regression model.





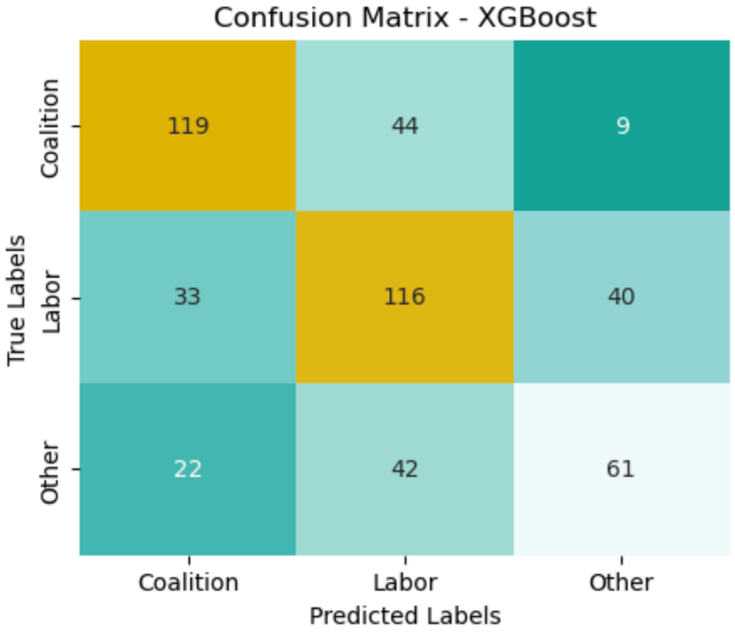
XGB Boost classifier:

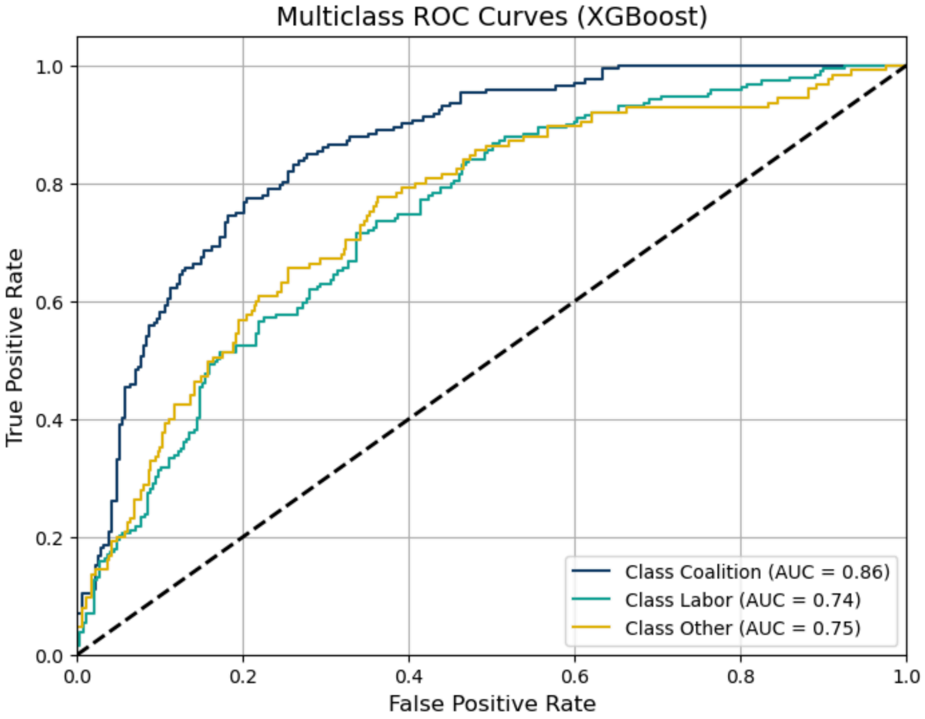
The XGBoost classifier delivered the strongest overall performance among all models tested. After hyper-parameter tuning, it achieved an accuracy of 61%, with balanced F1 scores of 0.69 for Coalition, 0.59 for Labor, and 0.52 for Other voters. The model demonstrated improved recall and precision across all classes compared to previous models, particularly in distinguishing Other voters from Coalition — a challenge for both logistic regression and Random Forest. The confusion matrix confirmed this improvement, showing that fewer Other voters were misclassified as Coalition than in the Random Forest model (22 vs. 32).

XGBoost also offered the best ROC AUC scores, with 0.86 for Coalition, 0.74 for Labor, and 0.75 for Other, indicating strong overall separability and model confidence. Although the model still misclassified some Labor voters as Other (and vice versa), the shape of the ROC curves showed better alignment with the ideal classification threshold than any other model.

In terms of feature importance, XGBoost aligned closely with logistic regression on the top three predictive features, though in a different order. Like Random Forest, it also placed greater emphasis on top_issue, confirming that tree-based ensemble models are better suited to capturing complex interactions and categorical variable effects — particularly in survey data. Unlike Random Forest, however, XGBoost ranked second_issue as less important, suggesting that issue salience may play out differently across model architectures.

While the XGBoost model produced only a modest improvement in predictive accuracy, it offered the most balanced and reliable performance overall. However, its results also reinforce a key insight from the project: voter behaviour is difficult to predict from survey data alone. XGBoost's greatest value lies in its ability to uncover underlying structure and segmentations, rather than deliver precise individual-level predictions.





Implementation

This model wasn’t designed to be deployed in a live or production environment, but it’s still useful to think about how the results could be applied in practice. In a real-world setting, insights from this kind of analysis could support campaign strategy, voter engagement, or even policy design, by helping parties or advocacy groups understand what different types of voters care about.

That said, there are a few important considerations. First, political data is sensitive, so any use of this kind of model would need to be handled with care — especially around privacy, ethics, and fairness. Also, the predictive accuracy of the models isn’t high enough to make individual-level decisions. The models are better suited to helping identify broad voter segments and explore trends, rather than predicting how someone will vote.

Another important consideration is that the dataset used in this project was heavily skewed toward older voters, which likely influenced the results. For the models to be more reliable and generalisable, future efforts should aim to collect more responses from younger voters, who were significantly underrepresented in the sample.

If this approach were to be used more formally, it would need access to updated and more representative data, and the models would have to be retrained regularly to reflect changes in voter attitudes.

Data answer

The data question was answered satisfactorily. The AES 2022 dataset provided rich, high-quality survey data that made it possible to explore the relationships between demographics, attitudes, issue priorities, and vote choice. The results are based on a robust sample size, but the dataset was skewed toward older voters, which introduces some limitations. Despite this, the confidence level in the patterns and insights is moderate to high, particularly when interpreting broad trends across the voter base.

Business answer

The business question — understanding which voter groups support Labor, the Coalition, or minor parties, and what distinguishes them — was answered successfully through both clustering and classification models. These approaches helped reveal meaningful differences in values, priorities, and demographics across voter segments. While the predictive models were not accurate enough for high-stakes targeting, the findings are still valuable for segmenting the electorate and guiding messaging or policy development. Confidence in the overall business insight is moderate, with the caveat that it should be used as a strategic guide, not a forecasting tool.

Response to stakeholders

Stakeholders — including campaign teams, policy analysts, and political strategists — can use the insights from this project to better understand how different types of voters think and what matters to them. Key recommendations include: focusing on issue-based segmentation, investing in data-driven voter profiling, and improving engagement with underrepresented groups, particularly younger voters. While prediction should not be the goal, these insights can inform more inclusive and targeted communication and policy decisions.

End-to-end solution

The project delivered an end-to-end pipeline: from data cleaning and exploratory analysis to clustering and predictive modelling. The process is reusable and could be applied to future waves of the AES, or extended to other surveys or datasets. If updated and expanded with more representative data — especially from younger voters — this pipeline could help political teams better segment the electorate, track shifts in public sentiment, and test policy alignment over time. The models work best as decision support tools, offering insight rather than precise prediction.

References

Data and Code:

Dataset: 2022 Australian Election Study
Source: Australian Data Archive Dataverse

Notebooks:

- Ivonne_Capstone1.ipynb – Data cleaning, feature selection, exploratory data analysis
- Ivonne_Capstone2.ipynb – Clustering analysis, supervised modelling, evaluation

Original data: aes22_unrestricted_v3_cleaned.csv (used in first notebook)

Cleaned data: aes22_v3_cleaned.csv (used across both notebooks)

Tools and Libraries:

Programming Environment: Python 3.11, Jupyter Notebooks in VS Code (local machine)

Libraries:

Data manipulation: pandas, numpy
Visualisation: matplotlib, seaborn, plotly
Modelling & evaluation: scikit-learn, xgboost, skfuzzy
Statistics: scipy, statsmodels
Hyperparameter tuning: GridSearchCV from scikit-learn

Methods and Techniques:

Data preprocessing: Label encoding, MinMaxScaler, median/mode imputation, handling missing values

Unsupervised learning: Fuzzy C-Means Clustering for voter segmentation

Supervised learning:

Multinomial Logistic Regression
Random Forest Classifier (with hyperparameter tuning)
XGBoost Classifier (with hyperparameter tuning)

Feature selection: Recursive Feature Elimination (RFE), two-sided t-test

Model evaluation: Accuracy, Precision, Recall, F1-score, ROC AUC, Confusion Matrix, Classification Report

Website links:

https://www.displayr.com/understanding-cluster-analysis-a-comprehensive-guide/

https://www.geeksforgeeks.org/ml-fuzzy-clustering/

https://libstore.ugent.be/fulltxt/RUG01/003/008/293/RUG01-003008293_2021_0001_AC.pdf

https://statisticsbyjim.com/regression/multinomial-logistic-regression/

Apendix - Data Dictionary

Variable	New name	Variable Label	Value	Value Label
STATE	state	State	1	New South Wales
			2	Victoria
			3	Queensland
			4	South Australia
			5	Western Australia
			6	Tasmania
			7	Northern Territory
			8	Australian Capital Territory
			999	Item Skipped
A1	political_interest	A1. Generally speaking, how much interest do you usually have in what's going on in politics?	1	A good deal
			2	Some
			3	Not much
			4	None
			999	Item skipped
A10	vote_if_not_compulsory	A10. Would you have voted in the election if voting had not been compulsory?	1	Definitely would have voted
			2	Probably would have voted
			3	Might / might not have voted
			4	Probably would not have voted
			5	Definitely would not have voted
			97	Not eligible to vote
			999	Item skipped
B1	party_id	B1. Generally speaking, do you usually think of yourself as Liberal, Labor, National or what?	1	Liberal
			2	Labor
			3	National Party
			4	Greens
			5	Other party (please specify)
			6	No party
			7	Australian Democrats
			8	Christian Democratic Party
			9	Citizens Electoral Council
			10	Family First Party
			11	Pauline Hanson's One Nation
			12	Republican Party (replaced by Republican Party of Australia)
			13	Shooters, Fishers and Farmers Party
			14	Fishing Party
			15	United Australia Party (formerly Palmer's United Party)
			16	Katter's Australia Party
			17	Liberal Democrats
			18	Motoring Enthusiasts Party
			19	Australian Sports Party (dissolved in 2015)
			20	Reason Party (formerly The Australian Sex Party)
			21	The Wikileaks Party (dissolved in 2015)
			22	Australian Christians
			23	Derryn Hinch's Justice Party
			24	Centre Alliance (formerly Nick Xenophon Team)
			25	Rise Up Australia
			26	Science Party
			27	Australian Liberty Alliance
			28	Pirate Party
			30	Jacquie Lambie Network
			31	Arts Party
			32	Animal Justice Party
			33	Australian Cyclists Party
			34	Health Australia Party
			35	Affordable Housing Party
			36	Australia First Party
			37	Australian Better Families
			38	Australian Conservatives
			39	Australian People's Party
			40	Australian Progressives
			41	Australian Workers Party
			42	Child Protection Party
			43	Climate Action! Immigration Action! Accountable Politicians!
			44	Country Liberals (NT)
			45	Democratic Labour Party
			46	Fraser Anning'S Conservative National Party
			47	Help End Marijuana Prohibition (HEMP) Party
			48	Independents For Climate Action Now
			49	Involuntary Medication Objectors (Vaccination/Fluoride) Party
			50	Labour DLP
			51	Liberal National Party of Queensland
			52	Love Australia or Leave
			53	Non-Custodial Parents Party (Equal Parenting)
			54	Secular Party of Australia
			55	Seniors United Party of Australia
			56	Socialist Alliance
			57	Socialist Equality Party
			58	Sustainable Australia
			59	The Australian Mental Health Party

Variable	New name	Variable Label	Value	Value Label
			60	The Great Australian Party
			61	The Small Business Party
			62	The Together Party
			63	Victorian Socialists
			64	VOTEFLUX.ORG | Upgrade Democracy!
			65	WESTERN AUSTRALIA PARTY
			66	Yellow Vest Australia
			67	Australian Citizens Party
			68	Australian Federation Party
			69	Australian Values Party
			70	David Pocock
			71	Drew Pavlou Democratic Alliance
			72	FUSION: Science, Pirate, Secular, Climate Emergency
			73	Federal ICAC Now
			74	Indigenous - Aboriginal Party of Australia
			75	Informed Medical Options Party
			76	Rex Patrick Team
			77	TNL
			78	The Local Party of Australia
			95	Swing Voter
			96	Independent
			97	Other party (not specified)
			997	Does not apply
			999	Item skipped
B2	party_support_strength	B2. Would you call yourself a very strong, fairly strong, or not very strong supporter of that party?	1	Very strong supporter
			2	Fairly strong supporter
			3	Not very strong supporter
			0	Missing (added in analysis)
			999	Item skipped
B3	care_who_wins	B3. Would you say you cared a good deal which party won the Federal election or that you did not care very much which party won?	1	Cared a good deal
			2	Did not care very much
			3	Did not care at all
			999	Item skipped
B9_1	first_pref_vote	B9_1. In the Federal election for the House of Representatives on Saturday 21 May, which party did you vote for first in the House of Representatives?	1	Liberal
			2	Labor
			3	National Party
			4	Greens
			5	Other party (please specify)
			6	Voted informal / Did not vote
			7	Australian Democrats
			8	Christian Democratic Party
			9	Citizens Electoral Council
			10	Family First Party
			11	Pauline Hanson's One Nation
			12	Republican Party (replaced by Republican Party of Australia)
			13	Shooters, Fishers and Farmers Party
			14	Fishing Party
			15	United Australia Party (formerly Palmer's United Party)
			16	Katter's Australia Party
			17	Liberal Democrats
			18	Motoring Enthusiasts Party
			19	Australian Sports Party (dissolved in 2015)
			20	Reason Party (formerly The Australian Sex Party)
			21	The Wikileaks Party (dissolved in 2015)
			22	Australian Christians
			23	Derryn Hinch's Justice Party
			24	Centre Alliance (formerly Nick Xenophon Team)
			25	Rise Up Australia
			26	Science Party
			27	Australian Liberty Alliance
			28	Pirate Party
			30	Jacquie Lambie Network
			31	Arts Party
			32	Animal Justice Party
			33	Australian Cyclists Party
			34	Health Australia Party
			35	Affordable Housing Party
			36	Australia First Party
			37	Australian Better Families
			38	Australian Conservatives
			39	Australian People's Party
			40	Australian Progressives
			41	Australian Workers Party
			42	Child Protection Party
			43	Climate Action! Immigration Action! Accountable Politicians!
			44	Country Liberals (NT)
			45	Democratic Labour Party
			46	Fraser Anning'S Conservative National Party
			47	Help End Marijuana Prohibition (HEMP) Party
			48	Independents For Climate Action Now
			49	Involuntary Medication Objectors (Vaccination/Fluoride) Party

Variable	New name	Variable Label	Value	Value Label
			50	Labour DLP
			51	Liberal National Party of Queensland
			52	Love Australia or Leave
			53	Non-Custodial Parents Party (Equal Parenting)
			54	Secular Party of Australia
			55	Seniors United Party of Australia
			56	Socialist Alliance
			57	Socialist Equality Party
			58	Sustainable Australia
			59	The Australian Mental Health Party
			60	The Great Australian Party
			61	The Small Business Party
			62	The Together Party
			63	Victorian Socialists
			64	VOTEFLUX.ORG | Upgrade Democracy!
			65	WESTERN AUSTRALIA PARTY
			66	Yellow Vest Australia
			67	Australian Citizens Party
			68	Australian Federation Party
			69	Australian Values Party
			70	David Pocock
			71	Drew Pavlou Democratic Alliance
			72	FUSION: Science, Pirate, Secular, Climate Emergency
			73	Federal ICAC Now
			74	Indigenous - Aboriginal Party of Australia
			75	Informed Medical Options Party
			76	Rex Patrick Team
			77	TNL
			78	The Local Party of Australia
			94	No party
			95	Swing Voter
			96	Independent
			97	Other party (not specified)
			997	Does not apply
			999	Item skipped
B16	party_cares_about_people	B16. Some people say that political parties in Australia care what ordinary people think. Others say that political parties in Australia don't care what ordinary people think. Where would you place your view on this scale from 1 to 5?	1	1 - Political parties in Australia care what ordinary people think
			2	2
			3	3
			4	4
			5	5 - Political parties in Australia don't care what ordinary people think
			999	Item skipped
B17	parties_needed	B17. Where would you place your view on this scale from 1 to 5, where 1 means that political parties are necessary to make our political system work, and 5 means that political parties are not needed in Australia?	1	1 - Political parties are necessary to make our political system work
			2	2
			3	3
			4	4
			5	5 - Political parties are not needed in Australia
			999	Item skipped
C5	satisfaction_democracy	C5. On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the way democracy works in Australia?	1	Very satisfied
			2	Fairly satisfied
			3	Not very satisfied
			4	Not at all satisfied
			999	Item skipped
C7	gov_for_people	C7. Would you say the government is run by a few big interests looking out for themselves, or that it is run for the benefit of all the people?	1	Entirely run for the big interests
			2	Mostly run for the big interests
			3	About half and half
			4	Mostly run for the benefit of all
			5	Entirely run for the benefit of all
			999	Item skipped
D3_1	top_issue	D3_1. Still thinking about the same 10 issues, which of these issues was the most important to you and your family during the election campaign?	1	Taxation
			2	Immigration
			3	Education
			4	The environment
			6	Health and Medicare
			7	Refugees and asylum seekers
			8	Global warming
			10	Management of the economy
			11	The COVID-19 pandemic
			12	The cost of living
			13	National security
			999	Item skipped
D3_2	second_issue	D3_2. Which was the second most important issue to you and your family?	1	Taxation
			2	Immigration
			3	Education
			4	The environment
			6	Health and Medicare
			7	Refugees and asylum seekers
			8	Global warming
			10	Management of the economy
			11	The COVID-19 pandemic
			12	The cost of living

Variable	New name	Variable Label	Value	Value Label
			13	National security
			999	Item skipped
D8_1	spend_health	D8_1. Should there be more or less public expenditure in the following area? Health	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_2	spend_education	D8_2. Should there be more or less public expenditure in the following area? Education	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_3	spend_unemployment	D8_3. Should there be more or less public expenditure in the following area? Unemployment benefits	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_4	spend_defence	D8_4. Should there be more or less public expenditure in the following area? Defence	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_5	spend_pensions	D8_5. Should there be more or less public expenditure in the following area? Old-age pensions	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_6	spend_industry	D8_6. Should there be more or less public expenditure in the following area? Business and industry	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_7	spend_police	D8_7. Should there be more or less public expenditure in the following area? Police and law enforcement	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_8	spend_ndis	D8_8. Should there be more or less public expenditure in the following area? The National Disability Insurance Scheme	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_9	spend_transport	D8_9. Should there be more or less public expenditure in the following area? Public transport infrastructure	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
D8_10	spend_childcare	D8_10. Should there be more or less public expenditure in the following area? Child care	1	Much more than now
			2	Somewhat more than now
			3	The same as now
			4	Somewhat less than now
			5	Much less than now
			999	Item skipped
H1	gender	H1. Are you male or female?	1	Male
			2	Female
			3	Other
			999	Item skipped
J4	social_class	J4. Which social class would you say you belong to?	1	Upper class
			2	Middle class
			3	Working class
			4	None
			999	Item skipped
J5	residence_type	J5. Would you say you now live in ... ?	1	A rural area or village
			2	A small country town (under 10,000 people)
			3	A larger country town (over 10,000 people)
			4	A large town (over 25,000 people)
			5	A major city (over 100,000 people)
			999	Item skipped
J6	household_income	J6. What is the gross annual income, before tax or other deductions, for you and your family living with you from all	1	Less than $15,000 per year
			3	$15,001 to $25,000 per year
			5	$25,001 to $35,000 per year

Variable	New name	Variable Label	Value	Value Label
		sources? Please include any pensions and allowances, and income from interest or dividends.	7	$35,001 to $45,000 per year
			9	$45,001 to $60,000 per year
			11	$60,001 to $80,000 per year
			13	$80,001 to $100,000 per year
			15	$100,001 to $120,000 per year
			17	$120,001 to $140,000 per year
			19	$140,001 to $160,000 per year
			21	$160,001 to $200,000 per year
			24	$200,001 to $250,000 per year
			25	More than $250,000 per year
			999	Item skipped
H2_AGE_(	age_group	H2_AGE_GRP1. Age group derived from year of birth.	1	18 to 34
			2	35 to 44
			3	45 to 54
			4	55 years and over
			999	Item skipped
G3_EDU	has_uni_degree	G3_EDU. Whether have a University-level education	1	University
			2	Non-University
			999	Item skipped