Mini-Project 1: Demographic and Ideological Influence on Voter Intent in Spanish Elections

Group 4 - Sara Hamidi, Shirley Augustin, Jenna Brooks, Allison Park

Introduction and Hypothesis

Spain's political system is highly pluralistic and decentralized, with over 20 active political parties. The major political parties reflect various stances on issues such as immigration and sexism which differentiate them across the political spectrum, with far right parties typically promoting anti-feminist and anti-immigrant rhetoric in much of their political discourse (Anduiza and Rico 2024) while other far left parties claim the opposite. We aim to answer the question: Can the beliefs of Spanish citizens regarding sexism and immigration be used to predict their voting intentions in the Spanish elections?

In this project, we use Spanish Political Attitudes data to test methods for classifying respondents' intentions to vote for 6 of the major Spanish political parties spanning from the far left to far right political spectrum (see Appendix A for details on Spain's political spectrum). We hypothesize that methods such as k-nearest neighbors or random forest classifiers will be able to predict which party a respondent intends to vote for, using the survey response variables regarding political beliefs (nativism, sexism, and participation in Women's Day protests) and demographic features (income and sex).

We expect parties on the far ends of the political spectrum to be easier to classify than center parties due to the mobilization of issues like sexism and immigration by more ideologically extreme parties to distinguish themselves and galvanize their voter base. By exploring the link between social attitudes and voting intentions, this study sheds light on the extent to which ideological polarization around issues like immigration and sexism influences party alignment in Spain's increasingly fragmented political landscape.

Methods

We use data from the Spanish Political Attitudes dataset (Hernández Pérez et al. 2021). The survey uses a quota sampling method to ensure a representative sample of the Spanish adult

population aged 18 to 56, with quotas based on gender, age, educational background, geographic region, and municipality size. It also includes respondents' answers to questions on sexism, voting intention, participation in feminist protests, and beliefs surrounding immigration. The raw data comprises 7,850 observations and the unit of analysis is individual voters in Spain. When cleaned to the parties of interest and after removing NA's, the data contains 3,034 observations.

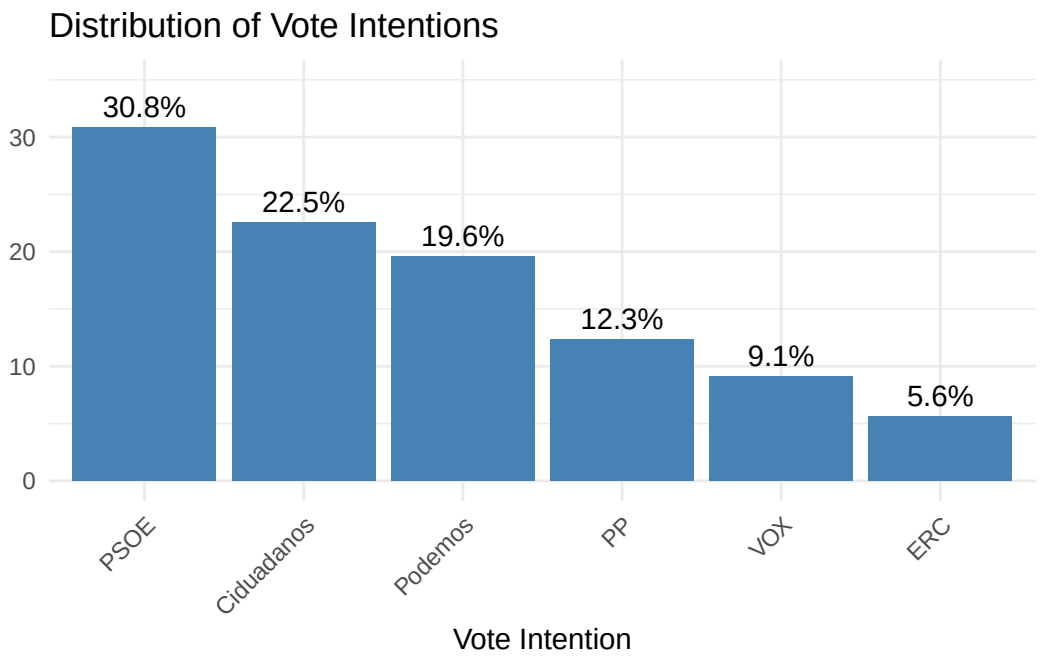
We focused on the following covariates: `dincome_all`, `female`, `nativism`, `msexism`, and `femdemonstrate` with `voteintentionspain` of the 6 parties of interest as the dependent variable which we aim to classify. Our motivation behind including these covariates in particular align with our hypothesis that there is a correlation between people's ideological beliefs and their political party affiliations. Moreover, we also include income as a demographic variable to assess whether gender and economic status influence party affiliation.

Within the `voteintentionspain` variable, the six political parties that we chose are: the following:

(numbers 1-23 indicate how they are coded in the data)

- 1 - PSOE (Center Left)
- 2 - PP (Center Right)
- 3 - Podemos (Far Left)
- 4 - Ciduadanos (Centrist)
- 7 - ERC (Catalonia)
- 23 - Vox (Far Right)

Next, we explored the data distribution and it became clear that upon evaluation of the data distribution, there is a significant class imbalance among the political parties, with the PSOE party having 30% of the observations, followed closely by Ciduadanos (22.5%) and Podemos (19.6%) with the other three trailing far behind. Therein lies a challenge in producing a successful classification model: having a class that reflects 30% of the observations but another class that represents only 5.6%. Inevitably, this would lead to higher prediction rates for the majority class versus significantly less predictions for the minority classes.



Our main objective is to predict which party a person intends to vote for, based on their demographics and other ideological standpoints. Our main computational methods involve experimenting with various classifiers such as K-Nearest Neighbors, random forests, and multinomial regression. KNN was chosen because it is non-parametric and can easily capture nonlinear relationships, since political attitudes often cluster spatially in feature space as people with similar beliefs vote similarly. Random forests similarly handle non-linearity well and offer feature importance metrics, which are helpful for interpreting which beliefs/demographics most influence vote intention. Multinomial regression is used as a baseline approach, assuming a linear relationship between variables.

To gain preliminary insights into how well we may be able to classify a voter's political party affiliation based on the covariates above, we chose to run a multinomial regression. After fitting this model, we then conducted both an in-sample evaluation using a train vs. test set partition and calculated a correlation matrix and prediction accuracy, which we found to be about 40%. However, the issue with just fitting a multinomial logistic regression was that the output is probabilistic, rather than classification based. For our motives, we want a model that is robust, to avoid overfitting and that is able to identify complex relationships between variables. Because of this, we chose to run a Random Forest with the same covariates.

We hoped that the robustness of a random forest and the decision tree algorithm would be more insightful in identifying which variables are most influential in making predictions — the key to our questions regarding the influence of ideology on political party affiliation.

Once we ran our random forest, we then looked at its prediction accuracy in a confusion matrix, and to go even further, we plotted a variable importance plot to identify which features are

most important in classifying voter's intent. Through these insights, it became clear that the variables most influential are: sexism and nativism – aligning with our hypothesis that ideology influences voter intentions.

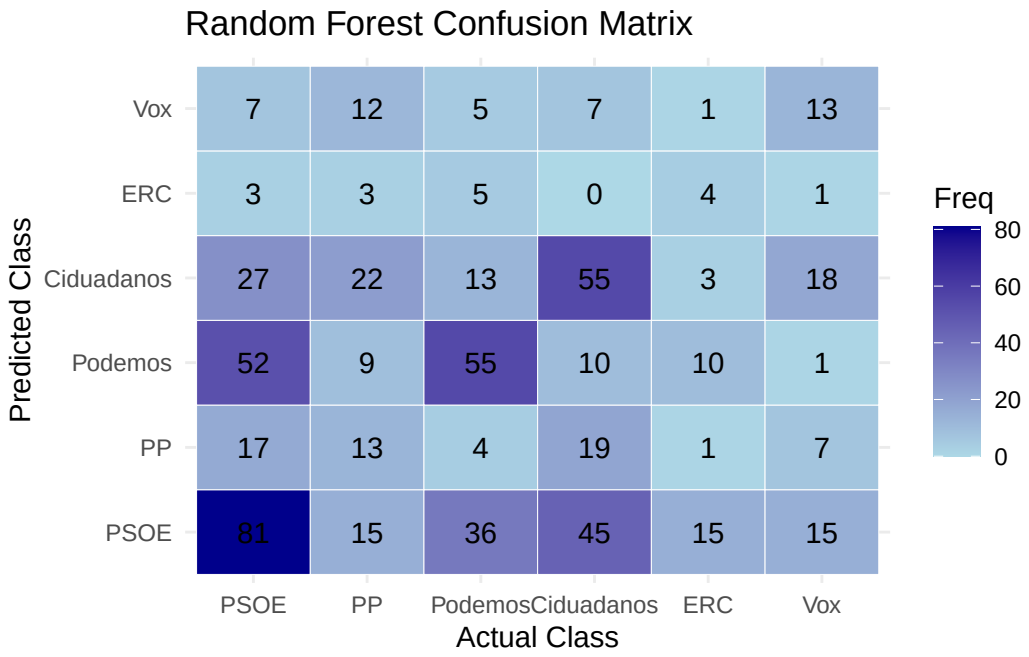
Although the random forest performed similarly to the multinomial regression model in terms of prediction accuracy, we also decided to try a K-NN model to see whether the proximity-based logic of k-nearest neighbors would pick up on the potential clustering of voters with similar ideological preferences. If our hypothesis is true and the do covariates cluster around the same voter intentions, this model would be able to leverage these insights and make better classification predictions. This model was also favorable because it provided class-based insights into sensitivity and specificity.

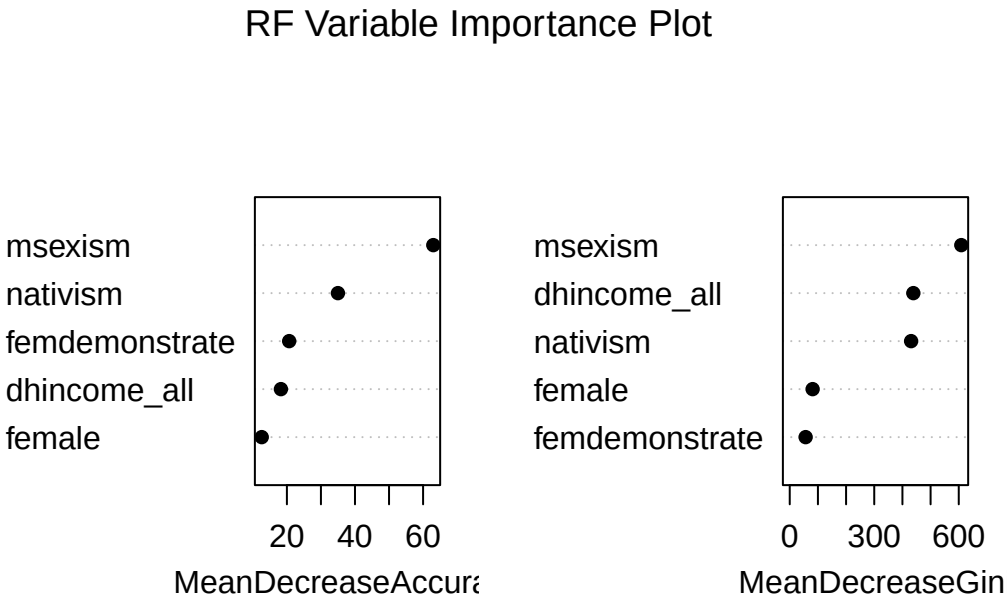
Results

Random Forest

The random forest model, with 500 trees and considering three variables at each split, achieved an accuracy of 37.5% on the portion of our data reserved for testing the models. This accuracy is at least twice better than chance (random decision between one of six parties being approximately 16%). We then examined the relative importance of the variables included in the random forest using the randomForest package's variable importance plot. This plot visualizes both the mean decrease in accuracy when a variable is randomized into noise and the mean decrease in Gini impurity index when a variable is considered at split. The variable measuring sexism is rated the highest importance for both these measures; the variable measuring nativism is rated second most important considering the mean decrease in accuracy and third most important when considering Gini impurity decreases. Sexism and nativism being the most important pieces of information when categorizing voter intention reflects the original authors' study, where they examined the effect of sexism on intent to vote for the far right party Vox.

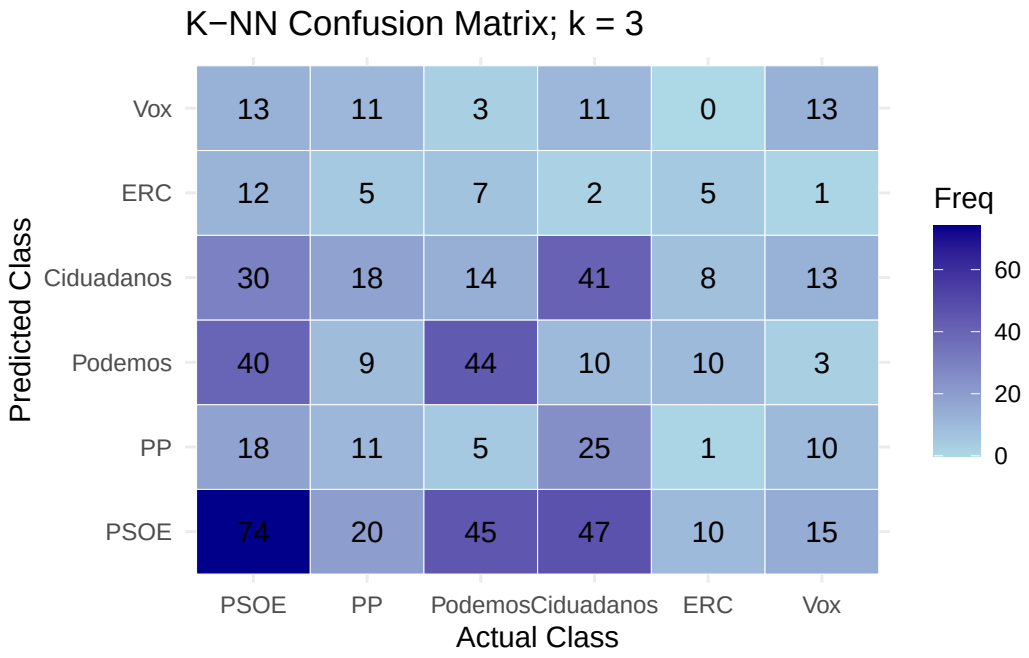
While the random forest model is not reliably accurate, its information regarding importance of variables is still useful when continuing to analyze the KNN models.

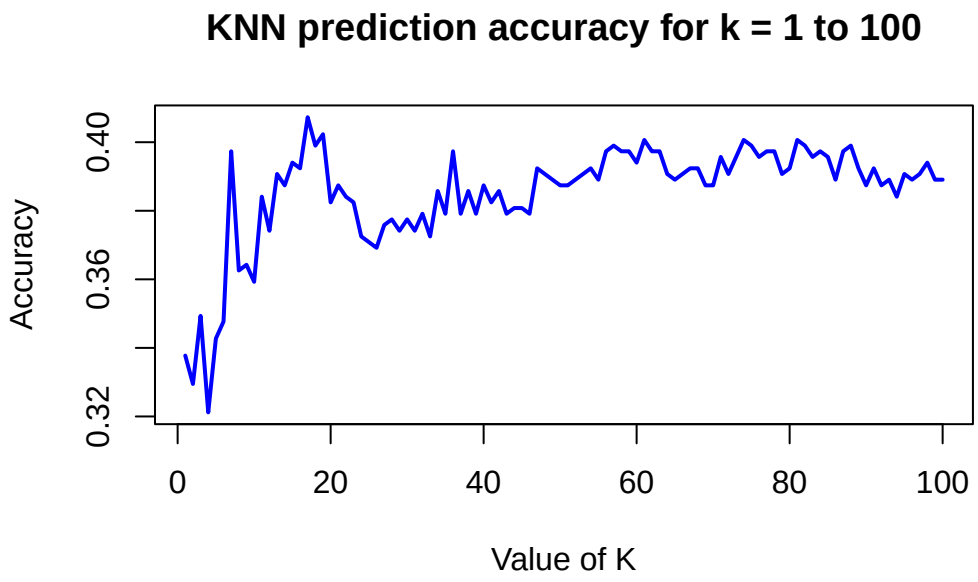

Random Forest Confusion Matrix
Predicted Class
Vox
ERC
Ciduadanos
Podemos
PP
PSOE
7 12 5 7 1 13
3 3 5 0 4 1
27 22 13 55 3 18
52 9 55 10 10 1
17 13 4 19 1 7
81 15 36 45 15 15
PSOE
PP
Podemos
Ciduadanos
ERC
Vox
Actual Class
Freq
80
60
40
20
0

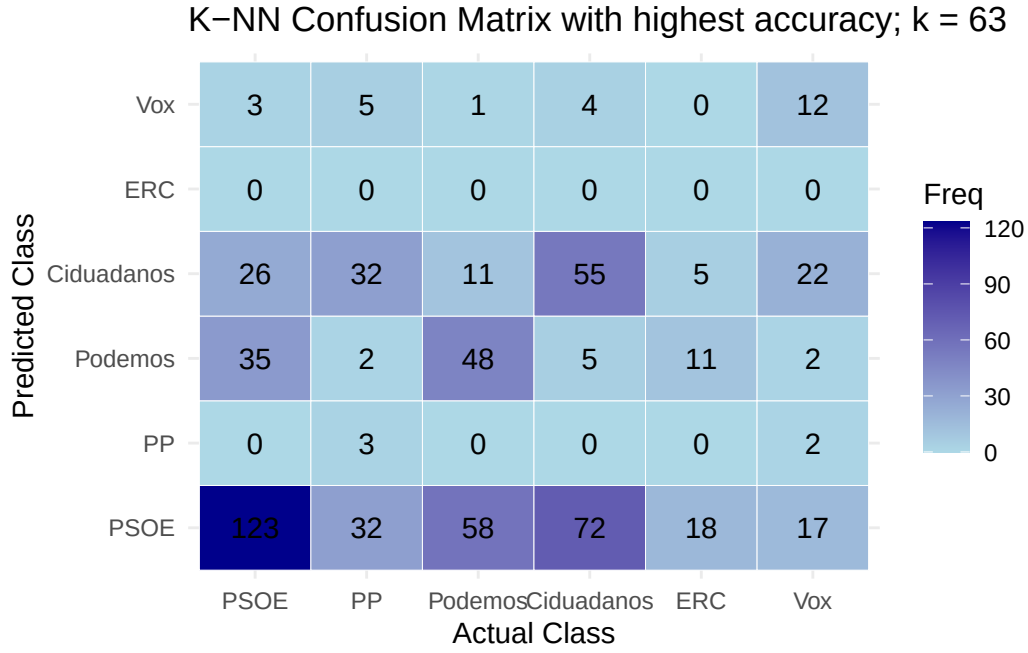

RF Variable Importance Plot
msexism
nativism
femdemonstrate
dhincome_all
female
20
40
60
MeanDecreaseAccura
msexism
dhincome_all
nativism
female
femdemonstrate
0
300
600
MeanDecreaseGin

K-Nearest Neighbors

We first modeled a "baseline" KNN where k = 3, achieving an accuracy of 33.3% on the test set, before tuning the hyperparameter k. The best k was defined as the k that produces the highest overall accuracy on the test data; after iterating over k = 1 to 100, the best value of k was found to be 63, with an accuracy of 40.2%. Similarly to the random forest, both models perform about twice as well as chance, but not so well that they could be considered accurate when predicting vote intention.



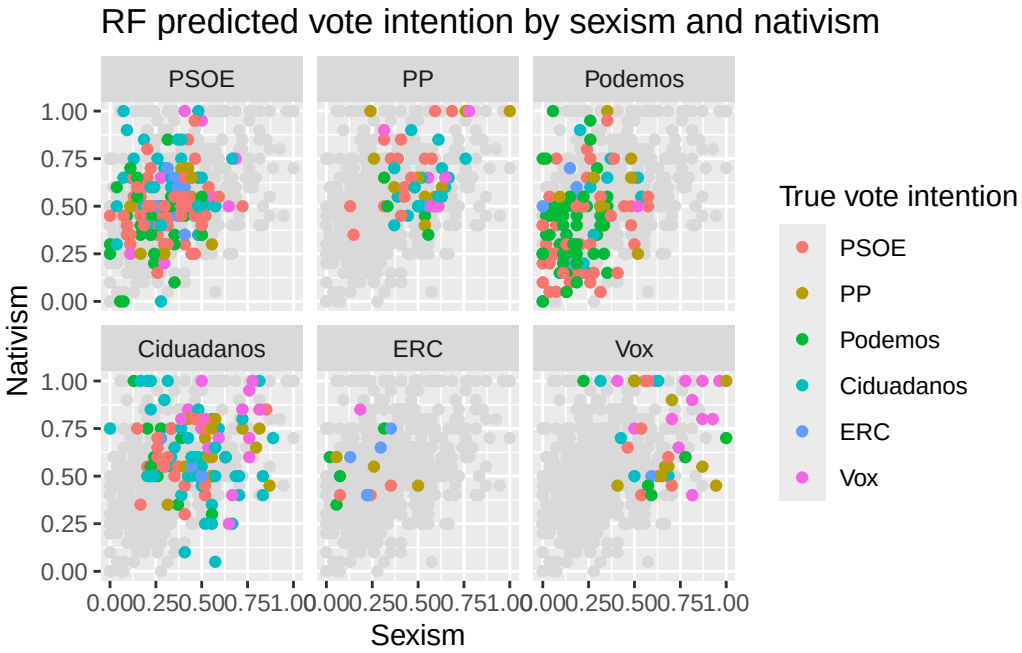


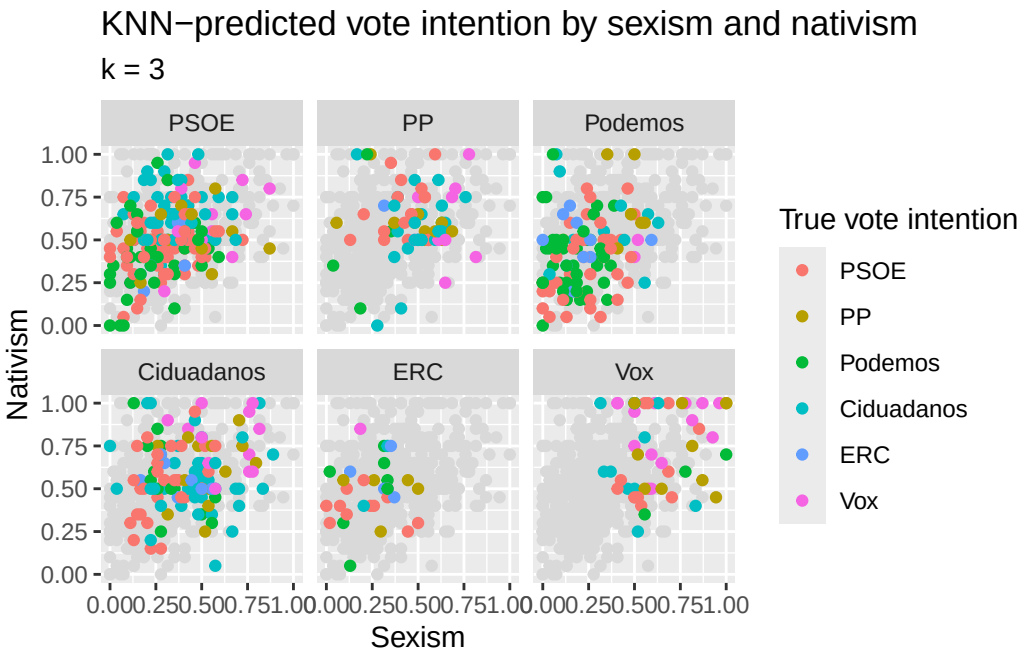


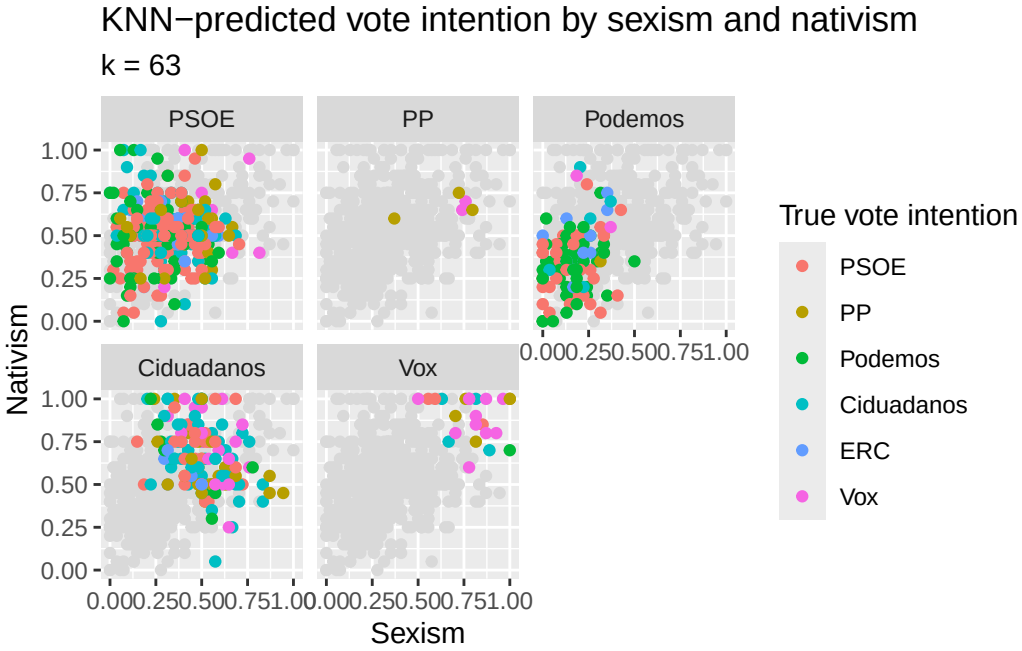
The confusion matrices give us an opportunity to analyze the behavior of the different k-nn models. When comparing the "baseline" KNN, k = 3, to the random forest model, we see similar patterns in three parties: PSOE, Podemos, and Ciudadanos. These classes are the top three parties that participants intend to vote for, comprising 30.8%, 22.5%, and 19.6% of the

data, respectively. The KNN model with the best k, k = 63, also displays bias toward the majority classes. This model never predicts the ERC class and heavily favors PSOE instead, predicting PSOE in 259 out of 604 cases. All models showing bias toward majority classes likely resulted in their similar misclassification patterns.

We can visualize the models' predictions with scatterplots, faceted by predicted vote intention. These scatterplots place voters on the axes of sexism and nativism, which are the most important as judged by the random forest's variable importance plot and variables relevant to identifying placement on political left-to-right scales.





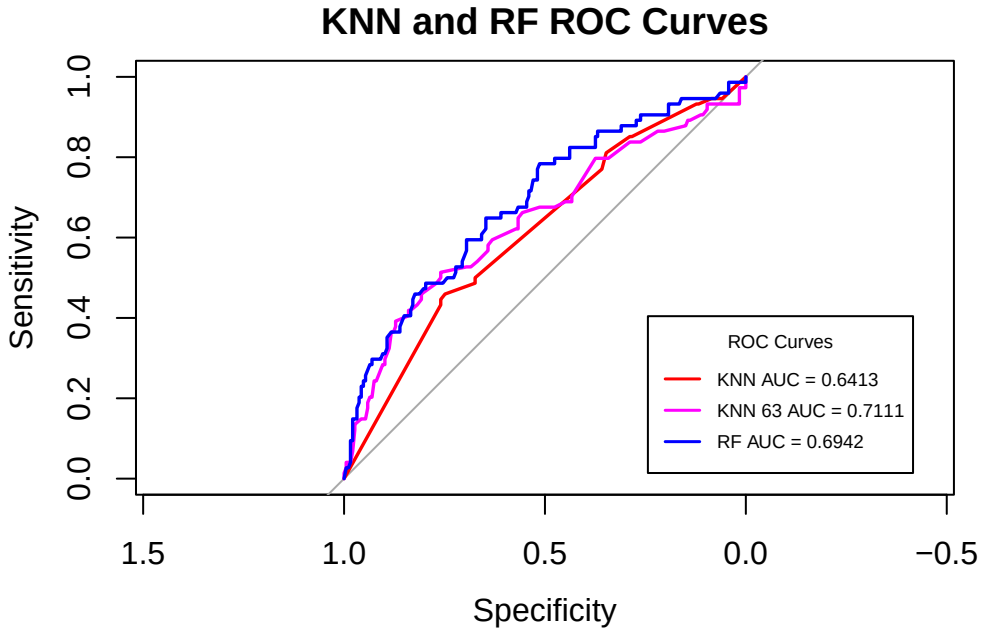


The high variability of colors in each facet, across all models, highlights visually that none of the models are achieving high accuracy. The lack of consistent colors also shows that the models are not consistently miscategorizing one specific party as another. The plot showing the categorization for k = 3 has the most scattered, unclustered patterning across all facets,

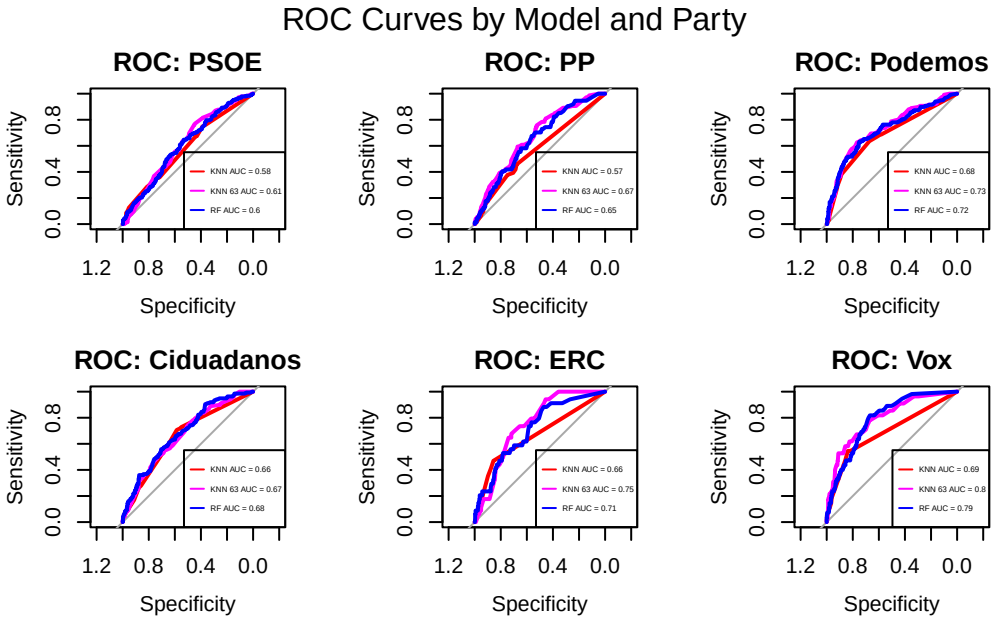
which means that this model has not picked up on strong patterns for party predictions relative to sexism or nativism. However, this is likely due to the k = 3 and not an essential fault of the k-nearest neighbors algorithm, as the model where k = 63 does show clustering for the different facets. The participants predicted to vote for the far-right party Vox are comprised entirely of the participants with the highest nativism and sexism, clustered in the top right; the participants predicted to vote for the far left party Podemos are clustered in the bottom left, with low scores on the measures of sexism and nativism. The center left party PSOE and the centrist party Ciduadanos also show relatively interpretable patterns; predicted PSOE voters tend toward the center on both measures, and predicted Ciduadanos voters tend toward middling sexism with higher nativism. These patterns in measures of sexism and nativism aligning with political party affilation suggests that the KNN model where k = 63 has learned some sort of pattern that reflects reality despite only achieving roughly 40% accuracy. The presence of similar patterns in the random forest scatterplot, although each facet is less clustered than that of the KNN model where k = 63, reflects a similar outcome.

Comparison of Random Forest and KNN

When observing the predictive models through the use of the multi-model ROC plot, we can see more clearly that when k = 63, the performance is most similar to that of the random forest model in predicting the different parties. The AUC score, which gives a numerical evaluation of the ability of the models to distinguish between different parties, also shows that the KNN where k = 63 and the random forest perform quite similarly.



To get further insight on how each model performs, we decided to plot an ROC curve for each party. This confirmed that both KNN when k = 63 and the Random Forest model consistently outperformed the original KNN model.



Despite being the two parties with the lowest numbers of cases, ERC and Vox are the two parties in which all models perform best. The independent variables we have chosen may be more helpful in determining votes for certain parties as opposed to others, especially considering that Vox is a far right party with strong sexist and nativist views. On the other hand, the fact that the KNN k = 63 models performs well on the ERC party's ROC curve suggests that the party is such a minority that always predicting non-ERC vote intention makes for good accuracy.

Moreover, despite a large number of cases, the models perform badly on the center-aligned parties. The center left PSOE party holds the most intended votes and the worst performance on its specific ROC plot; the center right PP party holds 12.3% of intended votes and the second worst performance on its ROC plot; and the centrist Ciudadanos party holds 22.5% of intended votes and the third-worst performance on its ROC plot. These patterns do not align with bias for or against majority classes. Instead, the independent variables analyzed here may not be strong predictors for relatively center political parties.

Discussion

Every model achieves similar accuracy performance on the test data, ranging only from 37.5% to 40.2%; a difference of approximately only 3%. Although all models perform at least twice as well as chance, given that there are six different parties to classify, the overall lack of accuracy suggests that the conceptual space cannot be so easily divided or mapped by the classification algorithms we used. The random forest approach's inflexible decision boundaries may not be suited to predicting vote intention out of these six independent variables. The k-nearest neighbors approach, which asserts that new data points can be categorized according to those nearest it, performs poorly when k is set very low, and best when k = 63, which is approximately 10% of the test dataset size. Visually observable patterns at k = 63, versus the lack of visual patterns at k = 3, suggests that the k-nearest neighbors model can pick up patterns relative to measures of nativism and sexism when casting a wider net than just the three closest neighbors, which makes sense; vote intention is complex, and cannot be accurately predicted just from the three closest datapoints.

However, the patterns that can be predicted align with parties that have the strongest views. Vox and Podemos, on the far right and far left respectively, show the most obvious clustering relative to measures of nativism and sexism across all models, implying that these variables reflect patterns that the models have learned about which voters are predicted to vote for which party, despite the models' relative inaccuracy. Similarly interpretable patterns appear for Ciduadanos and PSOE, which are centrist and center left parties. Fascinatingly, this is in spite of the models' low performance on predicting these parties; recall that these parties were the third-worst and worst performance as measured by AUC, for all models.

These patterns regarding nativism and sexism in the models' predictions, despite the models' relatively low accuracy, suggests that other variables still contribute to the models' categorizations, even if these two variables are highest on measures of importance.

References

Anduiza, Eva, and Guillem Rico. 2024. "Sexism and the Far-Right Vote: The Individual Dynamics of Gender Backlash." American Journal of Political Science 68 (2): 478–493.https://doi.org/10.1111/ajps.12759.

Hernández Pérez, Enrique, Eva Anduiza Perea, Carol Galais González, Guillem Rico Camps, Jordi Muñoz Mendoza, María José Hierro Hernández, Roberto Pannico, Berta Barbet Porta, and Dani Marinova. 2021. "POLAT Project: Spanish Political Attitudes Panel Dataset (Waves 1–6)." Universitat Autònoma de Barcelona.https://ddd.uab.cat/record/243399 (accessed September 13, 2021).

Data

Anduiza, Eva, and Guillem Rico. 2022. Replication Data for: Sexism and the Far-Right Vote: The Individual Dynamics of Gender Backlash. Harvard Dataverse.https://doi.org/10.7910/DVN/A11CD5.

Appendix

Baseline Multinomial Regression Model

```
# weights:  42 (30 variable)
initial  value 5436.198230
iter  10 value 4567.895403
iter  20 value 4416.496490
iter  30 value 4409.141891
final  value 4409.085197
converged


Call:
multinom(formula = voteintentionspain ~ dhincome_all + female +
    nativism + msexism + femdemonstrate, data = df, Hess = T,
    model = T, maxit = 200)

Coefficients:
   (Intercept) dhincome_all    female1   nativism   msexism femdemonstrate1
2    -4.067055   0.71410557  0.1412445  1.9688404  4.263394      -0.8240724
3     1.070342  -0.04126417 -0.7534162 -0.8972243 -2.656777       0.2883177
4    -2.444808   0.84123595 -0.1501845  1.8797754  2.245535      -0.6566397
7    -1.918678   2.15332199 -0.5757352  0.3195155 -3.001782       0.0826683
23   -6.963914  -0.12535354  0.1005077  4.1716956  7.171629      -0.6808231

Std. Errors:
   (Intercept) dhincome_all   female1  nativism   msexism femdemonstrate1
2    0.3140095    0.2651451 0.1351833 0.3533736 0.3886637       0.2154174
3    0.2250991    0.2217404 0.1143460 0.2910473 0.3473124       0.1245083
4    0.2475053    0.2161349 0.1095732 0.2856450 0.3168082       0.1533539
7    0.3870583    0.3903637 0.1787889 0.4512022 0.5509253       0.1984703
23   0.3984747    0.3189850 0.1668008 0.4137660 0.4688717       0.2824882

Residual Deviance: 8818.17
AIC: 8878.17
```

```
# weights:  36 (25 variable)
initial  value 4353.975510
iter  10 value 3663.265876
iter  20 value 3551.775705
iter  30 value 3548.662804
final  value 3548.658744
converged
```

Predicted	Actual PSOE	PP	Podemos	Ciduadanos	ERC	Vox
1	127	26	61	66	19	10
2	1	0	0	1	0	0
3	29	0	43	4	10	1
4	25	35	13	58	5	25
7	0	0	0	0	0	0
23	5	13	1	7	0	19

```
[1] "Accuracy: 0.409"
```

```
Penalized Multinomial Regression

3034 samples
   4 predictor
   6 classes: '1', '2', '3', '4', '7', '23'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2730, 2729, 2731, 2729, 2731, 2730, ...
Resampling results across tuning parameters:

  decay  Accuracy   Kappa
  0e+00  0.3674934  0.1494128
  1e-04  0.3674934  0.1494128
  1e-01  0.3681491  0.1497701

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was decay = 0.1.
```

Predicted	Actual PSOE	PP	Podemos	Ciduadanos	ERC	Vox
1	127	26	61	66	19	10

```
       2     1   0        0          1   0   0
       3    29   0       43          4  10   1
       4    25  35       13         58   5  25
       7     0   0        0          0   0   0
       23    5  13        1          7   0  19
```

```
[1] "Accuracy: 0.409"
```