Vote choice in Germany

Assigment 2

Statistical Inference and Modeling

UPC - FIB

Jordi Espasa - Eduard Masip

December 2022

Table of Contents

Table of Contents **3**

1. Preface **4**

1.1. Data Description 4

1.2. Problem description and approach 5

2. Data Preparation **5**

2.1. Removing duplicate or irrelevant observations 5

2.2. Fix structural errors and check data types 6

2.3. Handle missing data and Outliers 6

2.4. Data balancing 6

2.5. Normality and Autocorrelation of the Target variable 6

3. Exploratory Data Analysis **6**

3.1. Data Analysis 6

3.2. Profiling Political Party 8

3.3. Profiling Political Orientation 12

4. Predictive Modeling: Political Orientation **14**

4.1. Binary model for Right and Left + Center 14

4.2. Binary Model for Center and Left 17

4.3. Model Explanation 18

5. Predictive Modeling: Political Party **20**

5.1. Binary model for Left-wing parties: Gruene and LINKE 20

5.2. Polytomous model for Center-wing parties: CDU/CSU, FDP and SPD 21

5.3. Models Explanations 22

5.3.1. Model Gruene / Linke 22

5.3.2. Model SPD / CDU_CSU / FDP 23

6. Validating the Hierarchical model **24**

7. Conclusions **27**

8. Annexes **28**

1. Preface

1.1. Data Description

The data has 1000 individual observations with personal information related to politics and personal information. These variables are the original ones in the dataset.

- **vote**: Voting decision for party into 6 levels (represented parties in the Bundestag) (**target**):
    - **AfD**: Alternative für Deutschland, right wing populist party (**right**)
    - **CDU/CSU**: Center-right Christian-democratic political alliance (**center**)
    - **FDP**: Free democratic party -- liberal party center or center-right of the political spectrum (**center**)
    - **Gruene**: Die Grünen -- "the Greens" (**left**)
    - **LINKE**: DIE LINKE the left party is a democratic socialist political party in Germany, it is the furthest left-wing party of the six represented in the Bundestag (**left**)
    - **SPD**: Social Democratic Party of Germany, center left (**center**).
- **egoposition_immigration**: Ego-position toward immigration (0 = very open to 10 = very restrictive)
- **ostwest**: Dummy for respondents from Eastern Germany (= 1)
- **political_interest**: Measurement for political interest (0 = low, 4 = high)
- **income**: Self-reported income satisfaction (0 = low, 4 = high)
- **gender**: Self-reported gender (binary coding with 1 = female)

Important things to remark:

- Is an unbalanced dataset.
- All variables are categorical.
- There are no missing values.

This repository https://github.com/waze96/SIM_2_VoteChoiceGermany contains all the scripts, the dataset and the report created for this project.

1.2. Problem description and approach

We will create three binomial and one polytomous models to create a hierarchical one in order to predict right, center and left wing voting in the political spectrum and with those results predict the party more likely to vote for each individual. With this approach we will drag the error to the next models, but probably we will obtain better results than predicting each feature separately.

2. Data Preparation

2.1. Removing duplicate or irrelevant observations

Since there is any column to identify each individual, we cannot check if there are any duplicates. The only thing that we can say is that there are 359 individuals that have the same values, but this is easy to happen because there are only 6 variables and they are categorical so the range of values are limited.

2.2. Fix structural errors and check data types

It seems that there are no coding errors, all factors are correctly defined.
We renamed the level factors with meaningful names that describe the factors to which they belong. This helps us to better understand when creating the model.

In this step we take the opportunity before converting all the variables to factors to keep a copy of the variable **egoposition_immigration** in different formats, one as intervals from [0,10], another numerical variable with the median of the interval that belongs.

2.3. Handle missing data and Outliers

There is not any missing value, so we don't need to apply any kind of imputation technique or remove individuals or features with high percentage of missing data.

As all the variables are categorical, there are no outliers. However we check if there is any strange value for any level of these factors. The only interesting thing that we found is that for **political_interest** there are only 3 individuals with the value 0, null interest, but we will not consider them outliers. Also for **income** we found that only 1.3% of the total individuals have level 0 that shows low satisfaction with income.

```
table(df$political_interest)

0   1   2   3   4
3  34 308 396 259
```

```
table(df$income)

 0   1   2   3   4
13  28 188 582 189
```

Figure 2-1: *Histogram before and after to reduce the number of levels in* ***DistanceFromHome***

2.4. Data balancing

We can observe that the dataset that we are working with is unbalanced on its target variable. This is an important thing to take into account, at least in the model validation part.

2.5. Normality and Autocorrelation of the Target variable

As this dataset is only composed of categorical variables we can't apply any kind of test, because normality only really makes sense for continuous variables.

3. Exploratory Data Analysis

3.1. Data Analysis

We did a data analysis for the explanatory variables and we extract some information related to it and relations between variables:

- The feature **egoposition_immigration** is concentrated to mid-low values. We created a new variable that groups these values in ranges.
- **Ostwest** variable is unbalanced with more than 75% from Yes.
- In general there is a medium-high level for **political_interest** and **income**. For **income**, more than 50% of individuals are at a medium level. *[Annex 1 and Annex 2]*
- The dataset contains more males than females, but the difference in **gender** is tight. *[Annex 3]*
- Bearing in mind that the **ostwest** variable is very unbalanced, it can be observed that in histogram for **ostwest:no** is more frequent low levels for income while for **ostwest:yes** it is the other way around.

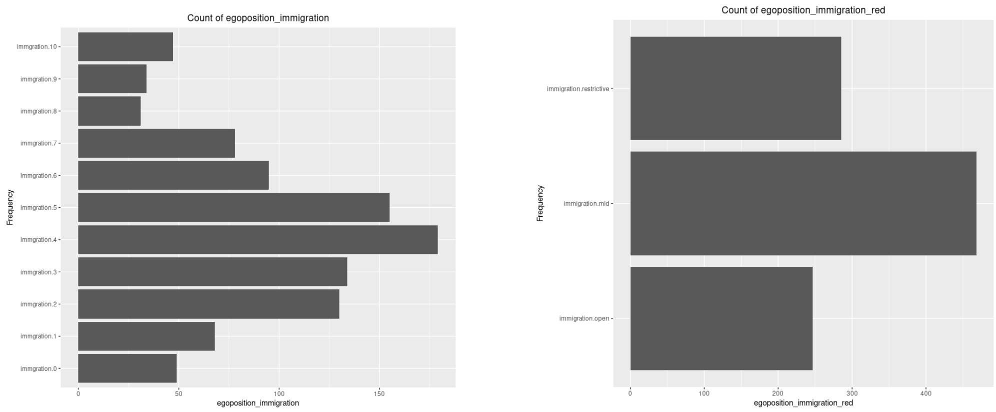


Figure 3-1:** Frequency of **egoposition_immigration** and **egoposition_immigration_red.

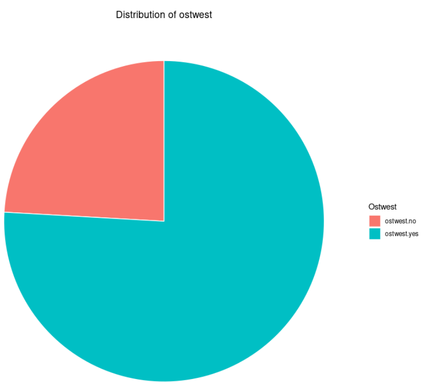


Figure 3-2:** Distribution of **ostwest

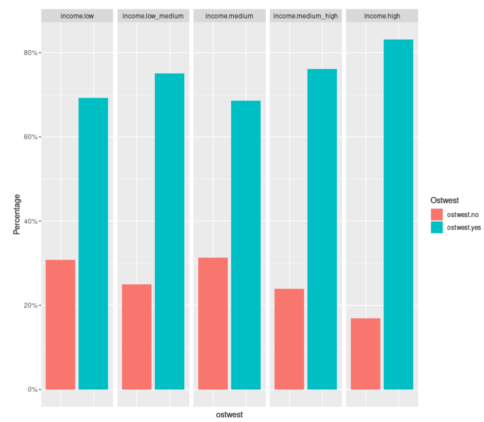


***Figure 3-3:** Distribution of **ostwest** grouped by **income** levels*

3.2. Profiling Political Party

We can see that there are two dominant political parties that both have more than 54% of all the votes (SPD and CDU/CSU).

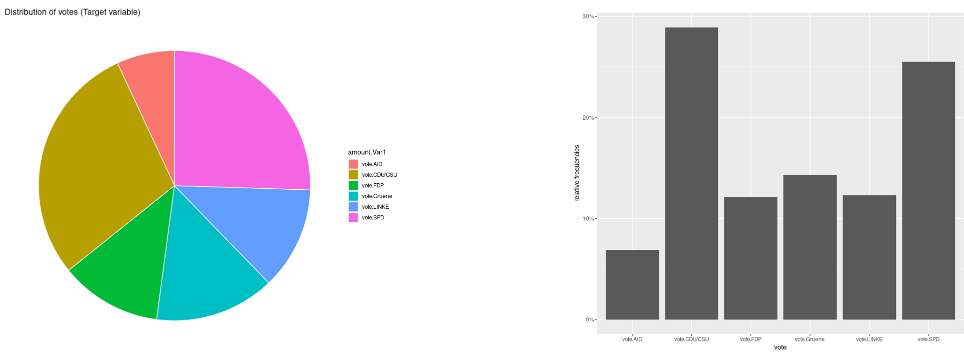


Figure 3-4: *Distribution of political parties* ***votes***

```
$vote.AfD
                                                     Cla/Mod   Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.restrictive 20.7017544 85.507246   28.5 1.463491e-24 10.229449
gender=gender.male                                  10.4089219 81.159420   53.8 1.101835e-06  4.872520
ostwest=ostwest.no                                  11.6182573 40.579710   24.1 1.731634e-03  3.132773
political_interest=political_interest.low           66.6666667  2.898551    0.3 1.377478e-02  2.463084
income=income.low_medium                            17.8571429  7.246377    2.8 4.793709e-02  1.977926
ostwest=ostwest.yes                                  5.4018445 59.420290   75.9 1.731634e-03 -3.132773
gender=gender.female                                 2.8138528 18.840580   46.2 1.101835e-06 -4.872520
egoposition_immigration_red=immigration.open         0.8097166  2.898551   24.7 4.967106e-07 -5.027579
egoposition_immigration_red=immigration.mid          1.7094017 11.594203   46.8 1.381884e-10 -6.417876
```

Figure 3-5: *Categorical description of* ***vote.AfD***

- 81.15% of the people who vote AfD are **gender:male** compared to a 53.8% overall **gender:male** share.
- 85.51% of the people who vote AfD are **immigration:restrictive** compared to a 28.5% overall **immigration:restrictive** share.
- 40.57% of the people who vote AfD are **ostwest:no** compared to a 24% overall **ostwest:no** share.

Also it is interesting to notice that ultra right-wing party AfD is more likely to be voted by male than female. *[Annex 4]*

Regarding the income, it can be seen *[Figure 3.6]* that there isn't any individual with low satisfaction related to self-income that votes ultra right-wing party AfD, but this changes for the next level where 7.25% of the people who vote AfD are low_medium satisfied with self-income compared to a 2.8 % overall low_medium share.

So if we check the last figure *[Figure 3.7]* we can check in more detail that 100% of AfD votes of low_medium incomes comes from males. For medium incomes males prevail but the difference is more tight. Finally for **income:medium_high** and **income:high** the males prevail again with significant differences.

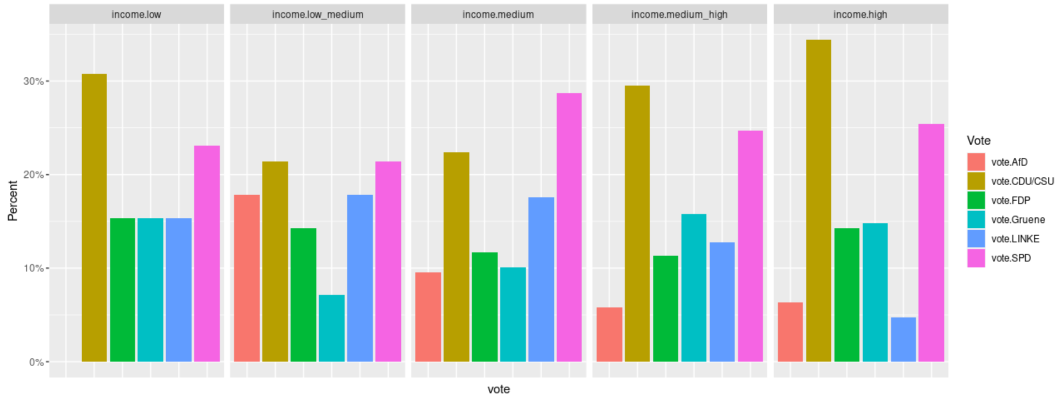


Figure 3-6: *Distribution of* ***votes*** *grouped by* ***income***

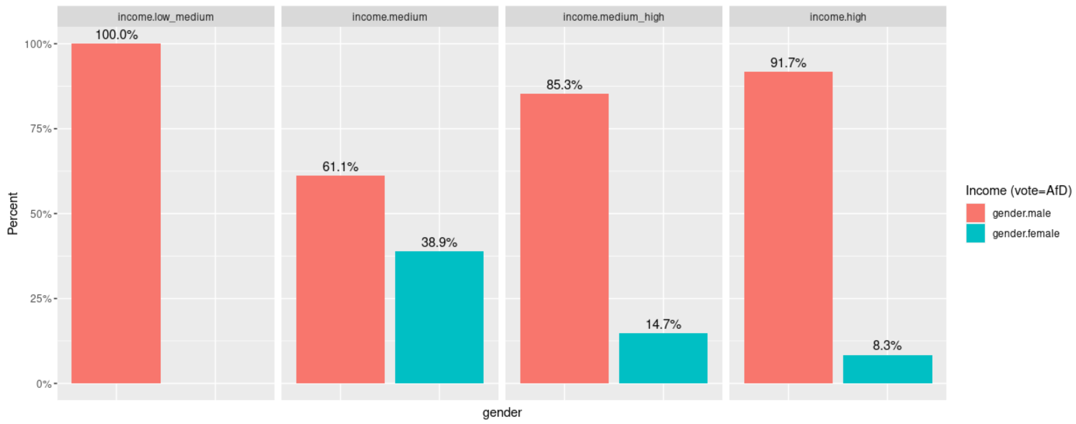


Figure 3-7: *Distribution of AfD* ***votes*** *by* ***gender*** *grouped by* ***income***

If we take a look to a immigration opinion, we can see that AfD is very related with a restrictive opinion against immigration, and there two individuals that vote AfD with an open opinion against immigration and they have high income satisfaction, so maybe AfD benefits the rich people also, that make sense because there is not any individual with low income voting AfD or are errors. They could be considered outliers because we can observe that it is a party more related with immigration opinion than income benefits. *[Figure 3.8 and Figure 3.9]*

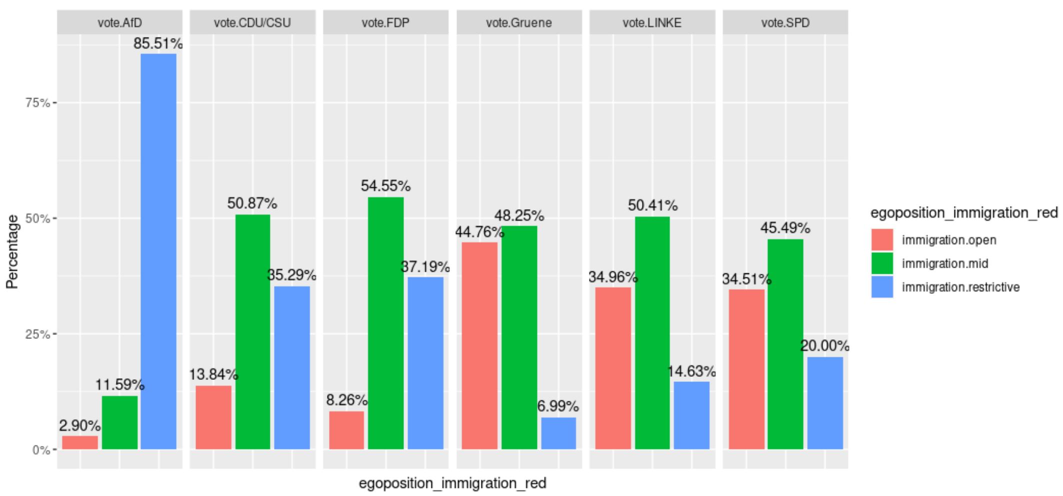


Figure 3-8: *Distribution of* ***immigration opinion*** *grouped by* ***political parties***

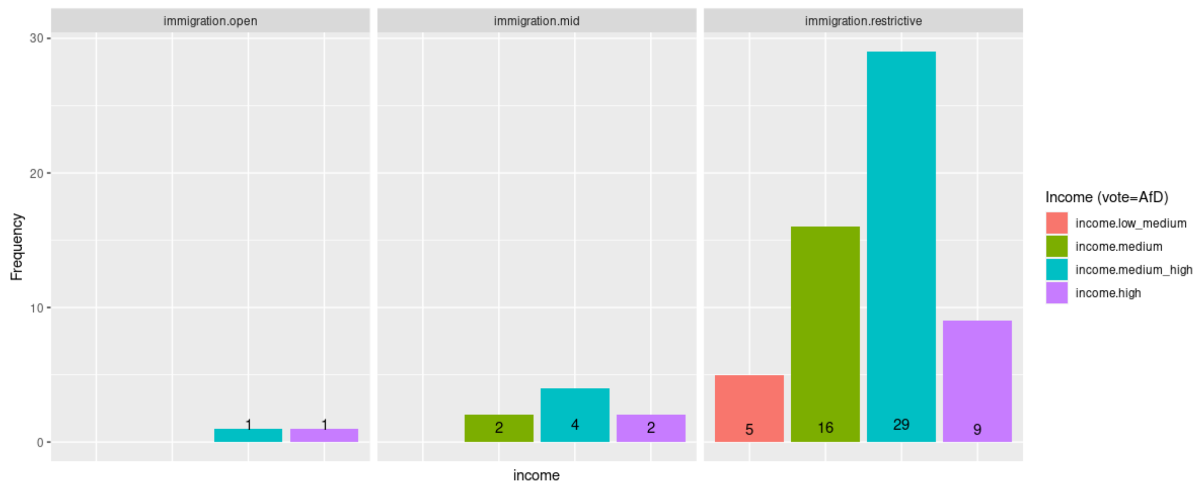


Figure 3-9: *Distribution of AfD **votes** by **income** grouped by **immigration opinion***

```
$`vote.CDU/CSU`
                                               Cla/Mod  Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.restrictive 35.78947 35.29412   28.5 2.753057e-03  2.994045
political_interest=political_interest.medium     33.44156 35.64014   30.8 3.612343e-02  2.095535
income=income.medium                             22.34043 14.53287   18.8 2.589708e-02 -2.227752
egoposition_immigration_red=immigration.open     16.19433 13.84083   24.7 1.460490e-07 -5.257472
```

Figure 3-10: *Categorical description of **vote.CDU/CSU***

- 35.29% of the people who vote Gruene are **immigration:restrictive** compared to a 28.5% overall **immigration:restrictive** share.

```
$vote.FDP
                                                     Cla/Mod   Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.restrictive 15.789474 37.190083   28.5 2.730611e-02  2.207114
egoposition_immigration_red=immigration.open         4.048583  8.264463   24.7 1.148248e-06 -4.864365
```

Figure 3-11: *Categorical description of **vote.FDP***

- 37.19% of the people who vote Gruene are **immigration:restrictive** compared to a 28.5% overall **immigration:restrictive** share.

```
$vote.Gruene
                                                     Cla/Mod   Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.open        25.910931 44.755245   24.7 1.320518e-08  5.683389
gender=gender.female                                17.748918 57.342657   46.2 4.072838e-03  2.872465
political_interest=political_interest.medium_high   17.676768 48.951049   39.6 1.459058e-02  2.442385
gender=gender.male                                  11.338290 42.657343   53.8 4.072838e-03 -2.872465
egoposition_immigration_red=immigration.restrictive  3.508772  6.993007   28.5 1.072233e-11 -6.796457
```

Figure 3-12: *Categorical description of **vote.Gruene***

- 44.75% of the people who vote Gruene are **immigration:open** compared to a 24.7% overall **immigration:open** share.
- 57.34% of the people who vote Gruene are **gender:female** compared to a 46.2% overall **gender:female** share.
- 48.95% of the people who vote Gruene are **political_interest:medium_high** compared to a 39.6% overall **ostwest:medium_high** share.
- 6.99% of the people who vote Gruene are **immigration:restrictive** compared to a 28.5% overall **immigration:restrictive** share.

It's interesting to notice that Gruene, left-wing party, is the unique party more voted by females although the dataset has more male individuals *[Annex 4]*.

Gruene is also very related to an open immigration opinion and we can observe that people that vote for this party are interested in politics.

Also the population that votes this party has a medium or more income satisfaction level.

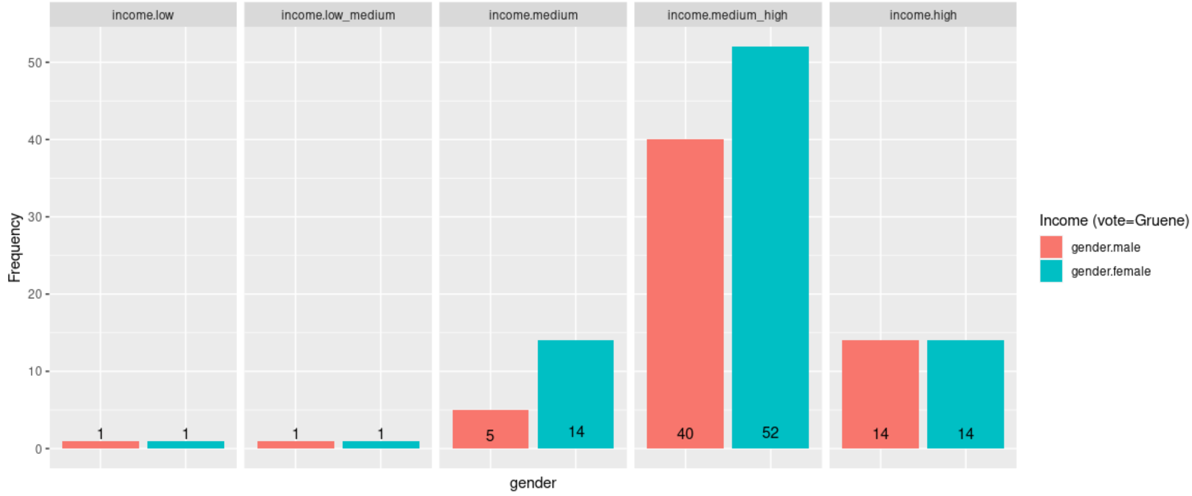


Figure 3-13: *Distribution of Gruene **votes** by **gender** grouped by **income***

```
$vote.LINKE
                                                 Cla/Mod   Mod/Cla Global      p.value    v.test
ostwest=ostwest.no                             19.087137 37.398374   24.1 0.0004265885  3.523064
egoposition_immigration_red=immigration.open   17.408907 34.959350   24.7 0.0064706238  2.722928
income=income.medium                           17.553191 26.829268   18.8 0.0192505867  2.340643
ostwest=ostwest.yes                            10.144928 62.601626   75.9 0.0004265885 -3.523064
income=income.high                              4.761905  7.317073   18.9 0.0001536281 -3.785130
egoposition_immigration_red=immigration.restrictive  6.315789 14.634146   28.5 0.0001392863 -3.809432
```

Figure 3-14: *Categorical description of **vote.LINKE***

- 37.39% of the people who vote LINKE are **ostwest:no** compared to 24.1% overall **ostwest:no** share.
- 34.96% of the people who vote LINKE are **immigration:open** compared to a 24.7% overall **immigration:open** share.
- 26.83% of the people who vote LINKE are **income:medium** compared to an 18.8% overall **income:open medium**.

```
$vote.SPD
                                                     Cla/Mod   Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.open       35.627530 34.509804   24.7 4.055236e-05  4.104310
political_interest=political_interest.low_medium    8.823529  1.176471    3.4 1.626155e-02 -2.402992
egoposition_immigration_red=immigration.restrictive 17.894737 20.000000   28.5 3.852923e-04 -3.549958
```

Figure 3-15: *Categorical description of **vote.SPD***

- 34.51% of the people who vote SPD are **immigration:open** compared to a 24.7% overall **immigration:open** share.

3.3. Profiling Political Orientation

This dataset is very unbalanced, we can observe that around 60% of votes are from center-wing political parties.

Also there is only one right-wing political party: AdF. So, in order not to add redundancy to the project, we will refer to the profiling of the AdF party.

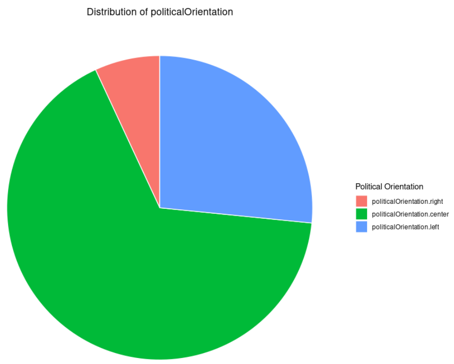


Figure 3-16: *Distribution of* ***politicalOrientation***

```
$politicalOrientation.center
                                              Cla/Mod  Mod/Cla Global      p.value    v.test
ostwest=ostwest.yes                          69.43347 79.24812   75.9 5.849772e-04  3.438486
income=income.high                           74.07407 21.05263   18.9 1.331258e-02  2.475296
egoposition_immigration_red=immigration.mid  70.29915 49.47368   46.8 1.705000e-02  2.385628
political_interest=political_interest.medium 71.75325 33.23308  30.8 1.837781e-02  2.357917
ostwest=ostwest.no                           57.26141 20.75188   24.1 5.849772e-04 -3.438486
egoposition_immigration_red=immigration.open 55.87045 20.75188   24.7 5.918386e-05 -4.016041
```

Figure 3-17: *Categorical description of* ***politicalOrientation.center***

- 21.05% of the people who vote center-wing have an **income:high** compared to a 18.9% overall **income:high** share.
- 33.23% of the people who vote center-wing have a **political_interest:medium** compared to a 30.8% overall **political_interest:medium** share.
- 20.75% of the people who vote center-wing are **ostwest:no** compared to a 24.1% overall **ostwest:no** share.
- 20.75% of the people who vote center-wing are **immigration:open** compared to a 24.7% overall **immigration:open** share.

It's interesting to notice that individuals **ostwest:no** tends to vote for a more radical political orientation. So in conclusion taking a look at *[Figure 3.18]* we can see easily that **ostwest:no** has more probability to vote parties with left or right political orientation than individuals from **ostwest:yes**.

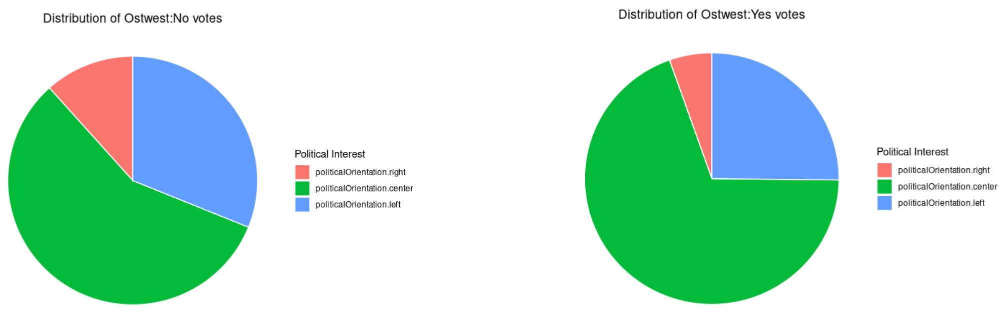


Figure 3-18: *Distribution of* ***politicalOrientation*** *grouped by* ***ostwest***

```
$politicalOrientation.left
                                               Cla/Mod  Mod/Cla Global      p.value    v.test
egoposition_immigration_red=immigration.open        43.319838 40.22556   24.7 3.339479e-11  6.630767
gender=gender.female                                30.086580 52.25564   46.2 2.119615e-02  2.304472
gender=gender.male                                  23.605948 47.74436   53.8 2.119615e-02 -2.304472
income=income.high                                  19.576720 13.90977   18.9 1.362802e-02 -2.466922
egoposition_immigration_red=immigration.restrictive  9.824561 10.52632   28.5 7.574019e-16 -8.060891
```

Figure 3-19: *Categorical description of* ***politicalOrientation.left***

- 40.22% of the people who vote left-wing are **immigration:open** compared to a 24.7% overall **immigration:open** share.
- 52.25% of the people who vote left-wing are **gender:female** compared to a 46.2% overall **gender:female** share.
- 13.91% of the people who vote left-wing have an **income:high** compared to a 18.9% overall **income:high** share.
- 10.52% of the people who vote left-wing are **immigration:restrictive** compared to a 28.5% overall **immigration:restrictive** share.

We can see at *[Figure 3.20]* the relation between **immigration:open** and **politicalOrientation:left** and also in **immigration:restrictive** and **politicalOrientation:right.**

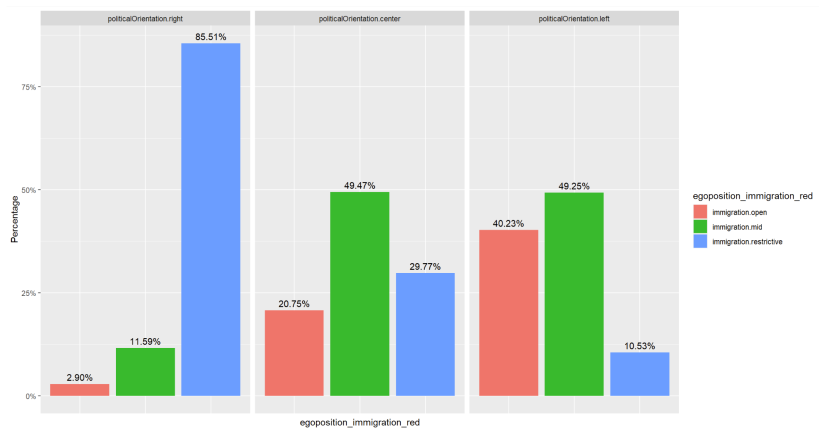


Figure 3-20: *Distribution of* ***immigration opinion*** *grouped by* ***politicalOrientation***

Other interesting things are that more females vote left-wing parties although there are less females in the dataset than males. *[Annex 5]* Also can be seen that the proportions of individuals with high income that vote left-wing parties are lower than right or center wing parties. On the other hand the proportion of individuals with low_medium income increases a lot in right-wing parties. *[Annex 6]*

4. Predictive Modeling: Political Orientation

We have done some model tests and found out that this prediction is not easy to get a great model with a generalized linear model due to the unbalanced dataset that we are facing. As we can see the model tends to predict the major category of the dataset.

	politicalOrientation.right	politicalOrientation.center	politicalOrientation.left
politicalOrientation.right	2	1	0
politicalOrientation.center	14	127	38
politicalOrientation.left	0	12	6

Figure 4-1: *Confusion matrix.*

There are several ways to takle this problem; oversampling, downsampling, changing weights… But regarding the focus of this work we will be using hierarchical modeling. Using this technique we believe the problem can be solved as we will be differentiating the unbalanced groups. Furthermore, a binary model tends to have better performance than one with more classes.

4.1. Binary model for Right and Left + Center

To find the best model we will create different variations based on the previous analysis of the data and then evaluate them by comparing their deviance, AIC, ANOVA and predicting on a subset of the dataset. The models with which we will compare are as follows.

bm0 → politicalOrientationBinary ~ egoposition_immigration
bm1 → politicalOrientationBinary ~ egoposition_immigration + ostwest
bm2 → politicalOrientationBinary ~ egoposition_immigration + political_interest + income + gender + ostwest
bm3 → politicalOrientationBinary ~ egoposition_immigration + gender + ostwest
bm4 → politicalOrientationBinary ~ egoposition_immigration * gender * ostwest

	Null-Res Deviance	**AIC**	**Confusion Matrix**
bm0	115.7823	296.3442	pred left_center right 0 184 16
bm1	118.9429	295.1836	pred left_center right 0 184 16
bm2	138.6927	293.4337	pred left_center right 0 183 15 1 1 1

bm3	127.2935	288.8330	pred left_center right 0 183 16 1 1 0
bm4	159.3413	316.7851	pred left_center right 0 183 16 1 1 0

Comparing the models we can observe that the deviance tells us that the best models are 0, 1 and 3. The AIC on the other hand indicates that the best model is 3 with a significant difference. Finally, when we apply the step function to the model with all interactions, we can see how it discards all interactions and the political interest variables and therefore stays with model 3.

```
                          Df Deviance    AIC
<none>                         262.83 288.83
- ostwest                  1   266.79 290.79
- gender                   1   271.18 295.18
- egoposition_immigration 10   367.34 373.34

Call:  glm(formula = politicalOrientationBinary ~ egoposition_immigration +
    gender + ostwest, family = "binomial", data = dfwork)
```

Figure 4-2: *Step function applied to model 4.*

With respect to the confusion matrices, it can be observed that all the models have prediction problem. The model bm3, even though it was the best model tested, it did not obtain good results in the subset of the dataset intended for testing. However, we realized that it always consistently predicted left_center which meant that the threshold could be wrongly set. By plotting the predicted probabilities we see that with a lower threshold our model could improve considerably.

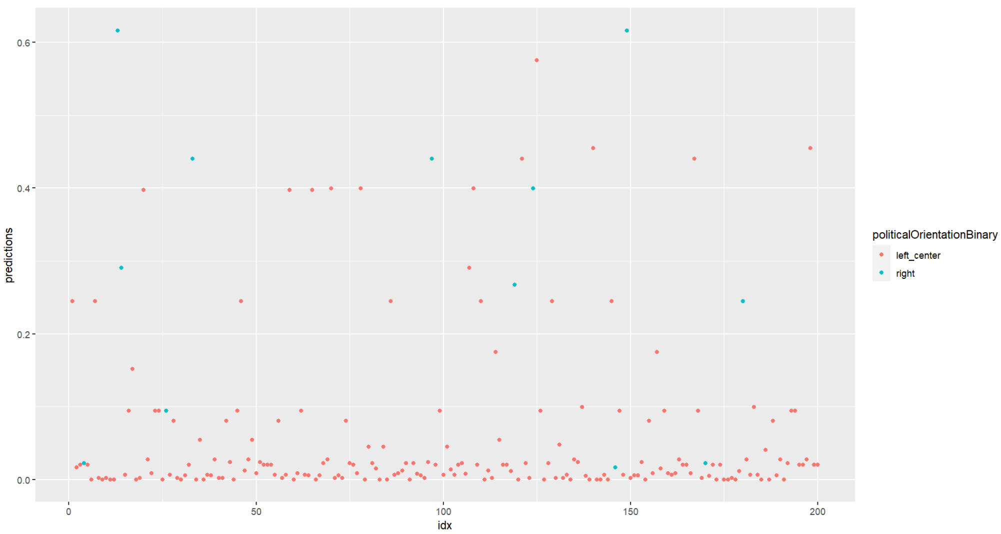


Figure 4-3: *Probability of predictions for left_center and right political orientation.*

And when we recalculate the model but this time with the new threshold we can observe much better results.

```
pred left_center right
   0         171     6
   1          13    10
```

```
> (accuracy <- sum(diag(tt)) / sum(tt))
[1] 0.905
> (precision <- diag(tt) / rowSums(tt))
        0         1
0.9661017 0.4347826
> (recall <- (diag(tt) / colSums(tt)))
left_center       right
  0.9293478   0.6250000
> (f1 <- (2*precision*recall/(precision+recall)))
        0         1
0.9473684 0.5128205
```

Figure 4-4: *Confusion matrix and metrics of the model.*

To further improve the model, we have looked at the model specifications by analyzing the coefficients and p-values of each variable. By observing it we have seen something very interesting. There are many classes of the immigration variable that had a p-value below 0.05. To improve this behavior and make use of the knowledge acquired in the data analysis sections, we decided to try with an addition of the variable to avoid irrelevant information and with the objective of increasing the p-values at a general level.

```
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.8277     1.0463  -2.703  0.00688 **
egoposition_immigrationimmgration.1  -0.3402     1.4359  -0.237  0.81272
egoposition_immigrationimmgration.2 -15.8158  1034.4552  -0.015  0.98780
egoposition_immigrationimmgration.3   0.1854     1.1756   0.158  0.87466
egoposition_immigrationimmgration.4  -1.2851     1.4306  -0.898  0.36900
egoposition_immigrationimmgration.5  -1.1024     1.4304  -0.771  0.44089
egoposition_immigrationimmgration.6   0.9120     1.1195   0.815  0.41526
egoposition_immigrationimmgration.7   2.1278     1.0701   1.988  0.04677 *
egoposition_immigrationimmgration.8   3.3316     1.0925   3.050  0.00229 **
egoposition_immigrationimmgration.9   2.0838     1.1619   1.793  0.07289 .
egoposition_immigrationimmgration.10  2.9499     1.0777   2.737  0.00619 **
gendergender.female                  -1.0477     0.3839  -2.729  0.00635 **
ostwestostwest.yes                   -0.6786     0.3373  -2.012  0.04425 *
```

Figure 4-5: *Summary of bm3.*

```
Coefficients:
                                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                                         -3.7047     0.7479  -4.953 7.30e-07
egoposition_immigration_redimmigration.mid           0.3041     0.8429   0.361  0.71826
egoposition_immigration_redimmigration.restrictive   3.1207     0.7319   4.264 2.01e-05
gendergender.female                                 -0.9723     0.3662  -2.655  0.00792
ostwestostwest.yes                                  -0.7539     0.3197  -2.358  0.01838
```

Figure 4-6: *Summary of bm3 with red immigration.*

```
pred          left_center right
  left_center         167     5
  right                17    11
>
```

Figure 4-7: *Confusion matrix and metrics of the final model.*

As we can see the result is very favorable. Both in improving the p-values and predicting right orientation. Now using this optimized model we will try to improve the prediction at three levels of left, center and right.

4.2. Binary Model for Center and Left

In this particular case, since we only have 3 classes and one class has already been predicted (right), we only need another binary model to discard the other two classes. The most relevant models we have tested are the following. As in the previous section, we will use different measures to evaluate the best model.

bm0 → politicalOrientationBinary ~ egoposition_immigration
bm1 → politicalOrientationBinary ~ egoposition_immigration + ostwest
bm2 → politicalOrientationBinary ~ egoposition_immigration + political_interest + income + gender + ostwest
bm3 → politicalOrientationBinary ~ egoposition_immigration + gender + ostwest
bm4 → oliticalOrientationBinary ~ egoposition_immigration * gender * ostwest

	Null-Res Deviance	AIC	Confusion Matrix
bm0	62.39888	841.9658	left center Center 63 10 Left 65 48
bm1	70.44865	835.9160	left center Center 72 16 Left 56 42
bm2	82.02496	842.3397	left center Center 75 18 Left 53 40
bm3	72.6956	835.6691	left center Center 76 16 Left 52 42
bm4	110.3378	858.0269	left center Center 64 12 Left 64 46

Comparing the models we can observe that the deviance tells us that the best models are 0, 1 and 3. The AIC on the other hand indicates that the best models are 1 and 3. Finally, when we apply the step function to the model with all interactions, we can see how it discards all interactions and the political interest variables and therefore stays with model 3.

```
                             Df Deviance    AIC
- gender:ostwest              1   809.67 835.67
<none>                            808.88 836.88
- egoposition_immigration 10      872.00 880.00

Step:  AIC=835.67
politicalOrientation ~ egoposition_immigration + gender + ostwest
```

Figure 4-8: *Step function applied to model 4.*

Given the balance between true positives and true negatives, we will select model 3 as the one to be used in this work.

4.3. Model Explanation

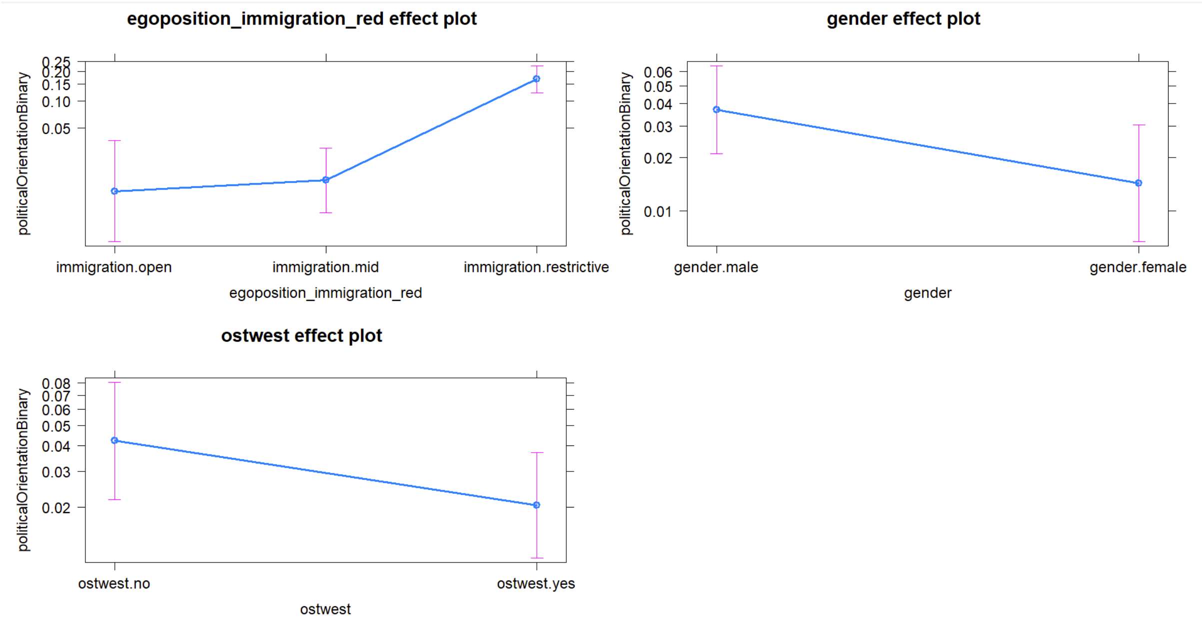


Figure 4-13: *Effects of the model.*

```
Coefficients:
                                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -3.7047     0.7479  -4.953 7.30e-07
egoposition_immigration_redimmigration.mid       0.3041     0.8429   0.361  0.71826
egoposition_immigration_redimmigration.restrictive  3.1207  0.7319   4.264 2.01e-05
gendergender.female                             -0.9723     0.3662  -2.655  0.00792
ostwestostwest.yes                              -0.7539     0.3197  -2.358  0.01838
```

Figure 4-14: *Summary of the model.*

We can observe that the lowest p-value is for the egoposition immigration variable, specifically for the restrictive class. The coefficient of the variable is also the highest coefficient and it is also positive. This indicates that when an individual belongs to this variable, he/she is more likely to belong to a right-wing party. The same can be observed in the graph of the effects. The values of immigration open and mid are close to one and on the contrary restrictive is much higher.

On the other hand, the other two variables have a much smaller effect on the prediction of the model. However, both are still significant as they have a p-value of less than 0.05.

- From the gender variable we can see that being male increases the probability of being right-wing.
- From the ostwest variable we can conclude that there is a small increase in the probability of being right wing with ostwest no.

5. Predictive Modeling: Political Party

To create models for vote prediction and following the idea of hierarchical modeling, we decide to create two models. One that predicts for center-wing parties and the other that predicts for left-wing parties. For right-wing parties, the classification of the political Orientation is enough because in the dataset there is only one left-wing party.

5.1. Binary model for Left-wing parties: Gruene and LINKE

We will create different variations based on the previous analysis of the data and then evaluate them by comparing their deviance, AIC and predicting on a subset of the dataset. The models with which we will compare are as follows.

bm3.2 → vote ~ egoposition_immigration_red + political_interest + income + gender + ostwest
bm3.3 → vote ~ egoposition_immigration_red + ostwest
bm3.4 → vote ~ egoposition_immigration_red + gender + ostwest

	Null-Res Deviance	AIC	Confusion Matrix from Train Dataset
bm3.2	22.87807	294.7055	GRUENE LINKE pred.GRUENE 72 36 pred.LINKE 44 61
bm3.3	7.469092	294.1145	GRUENE LINKE pred.GRUENE 85 59 pred.LINKE 31 38
bm3.4	10.83651	292.7471	GRUENE LINKE pred.GRUENE 63 44 pred.LINKE 53 53

We can observe that although the Null-Res and AIC indicate that the better model is 3.3 or 3.4, in the validation of the model, the one with better metrics is 3.2, but finally we decide to keep the model 3.4 because of AIC.

We found that with the default threshold at 0.5 the predictions were not very accurate so we moved down a little bit this threshold to 0.45 in order to get better results.

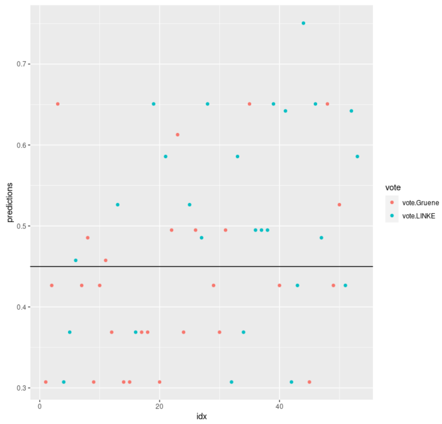


Figure 5-1: *Threshold on prediction probabilities.*

The metrics to validate this model are good enough and balanced, reaching an accuracy of 66% in the test dataset. The confusion matrix an other validation metrics can be seen at *[Figure 4.16]*

	GRUENE	LINKE
pred.GRUENE	17	8
pred.LINKE	10	18

```
> (precision <- diag(tt) / rowSums(tt))
pred.GRUENE  pred.LINKE
  0.6800000   0.6428571
> (recall <- (diag(tt) / colSums(tt)))
   GRUENE      LINKE
0.6296296 0.6923077
> (f1 <- (2*precision*recall/(precision+recall)))
pred.GRUENE  pred.LINKE
  0.6538462   0.6666667
```

Figure 5-2: *Confusion matrix and validation metrics for model bm3.4 with test dataset.*

5.2. Polytomous model for Center-wing parties: CDU/CSU, FDP and SPD

To handle this case we need to create a polytomous model in order to classify between the three center-wing parties that exist in the dataset. The principal problem has been the unbalanced dataset that we have where CDU/CSU and SPD have more than 54% of the overall votes, so it's difficult to predict the FDP party. To handle this inconvenience we add weights for the individuals that vote for the FDP party in order to force the model to predict this party more.

We try with different models in order to compare between them, and keep the better one.

mm2 → vote ~ egoposition_immigration_red + political_interest + income + gender + ostwest
mm3 → vote ~ egoposition_immigration_red + ostwest
mm4 → vote ~ egoposition_immigration_red + gender + ostwest

	AIC	Confusion Matrix from Train Dataset
mm2	1348.816	vote.CDU/CSU vote.FDP vote.SPD vote.CDU/CSU 148 63 90 vote.FDP 22 14 16 vote.SPD 61 19 99
mm3	1333.066	vote.CDU/CSU vote.FDP vote.SPD vote.CDU/CSU 199 88 133 vote.FDP 0 0 0 vote.SPD 32 8 72
mm4	1334.3315	vote.CDU/CSU vote.FDP vote.SPD vote.CDU/CSU 199 88 133 vote.FDP 0 0 0 vote.SPD 32 8 72

To decide which model to keep, we see that mm2 model is the one with higher AIC, but as the model has more parameters is able to classify some individuals in the FDP party that is the undersampled class. So we think that the trade-off is positive keeping the first one. Keeping the mm2 and setting the weights for individuals who vote FDP the model is capable of making better classifications.

The metrics to validate this model are worse than the previous model, but it's more difficult to reach higher accuracy and performance in general for polytomous models, and even more if they are unbalanced so, we think that an accuracy of 46.6% in test dataset is an acceptable result for this model. The confusion matrix an other validation metrics can be seen at *[Figure 4.17]*

	vote.CDU/CSU	vote.FDP	vote.SPD
vote.CDU/CSU	38	16	24
vote.FDP	6	2	4
vote.SPD	14	7	22

```
> (precision <- diag(tt) / rowSums(tt))
vote.CDU/CSU     vote.FDP     vote.SPD
   0.4871795    0.1666667    0.5116279
> (recall <- (diag(tt) / colSums(tt)))
vote.CDU/CSU     vote.FDP     vote.SPD
   0.6551724    0.0800000    0.4400000
> (f1 <- (2*precision*recall/(precision+recall)))
vote.CDU/CSU     vote.FDP     vote.SPD
   0.5588235    0.1081081    0.4731183
```

Figure 5-3: *Confusion matrix and validation metrics for model mm2 with test dataset.*

5.3. Models Explanations

5.3.1. Model Gruene / Linke

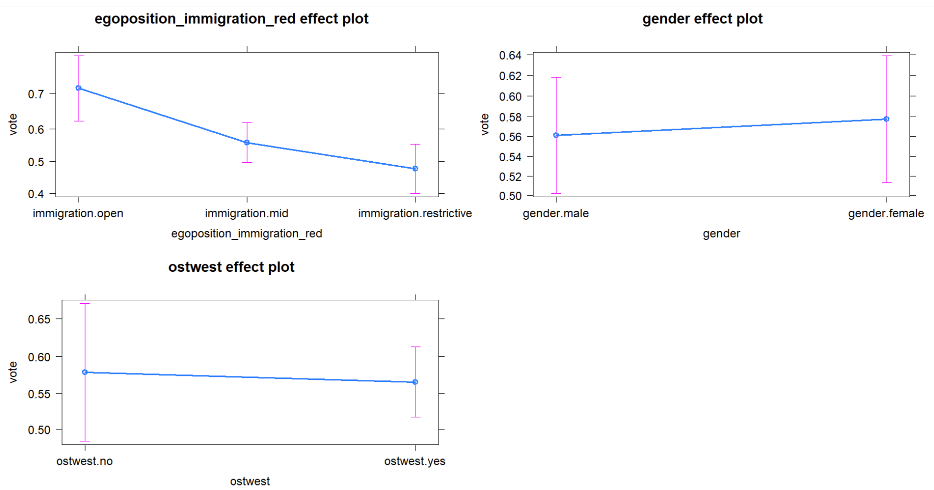


Figure 5-4: *Effects of the model.*

```
                                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                             0.92629    0.29083   3.185 0.001448 ** 
egoposition_immigration_redimmigration.mid             -0.68025    0.24417  -2.786 0.005336 ** 
egoposition_immigration_redimmigration.restrictive -1.00991    0.26172  -3.859 0.000114 ***
gendergender.female                                     0.06704    0.17910   0.374 0.708174    
ostwestostwest.yes                                     -0.05300    0.22001  -0.241 0.809648    
```

Figure 5-5: *Summary of the model.*

As we can see in the graphs of the effects of the model, the most important variable of the three used is the one related to immigration. Clearly marking a trend where the more open to immigration the more likely to kick Linke and the greater the restriction on immigration, the greater the probability of voting for Gruene.

The other two variables have a very low effect on the prediction.

- The gender seems to have a higher probability of voting Linke in women.
- Regarding the Ostwest variable, Ostwest does not have a slightly higher probability of voting for Linke.

5.3.2. Model SPD / CDU_CSU / FDP

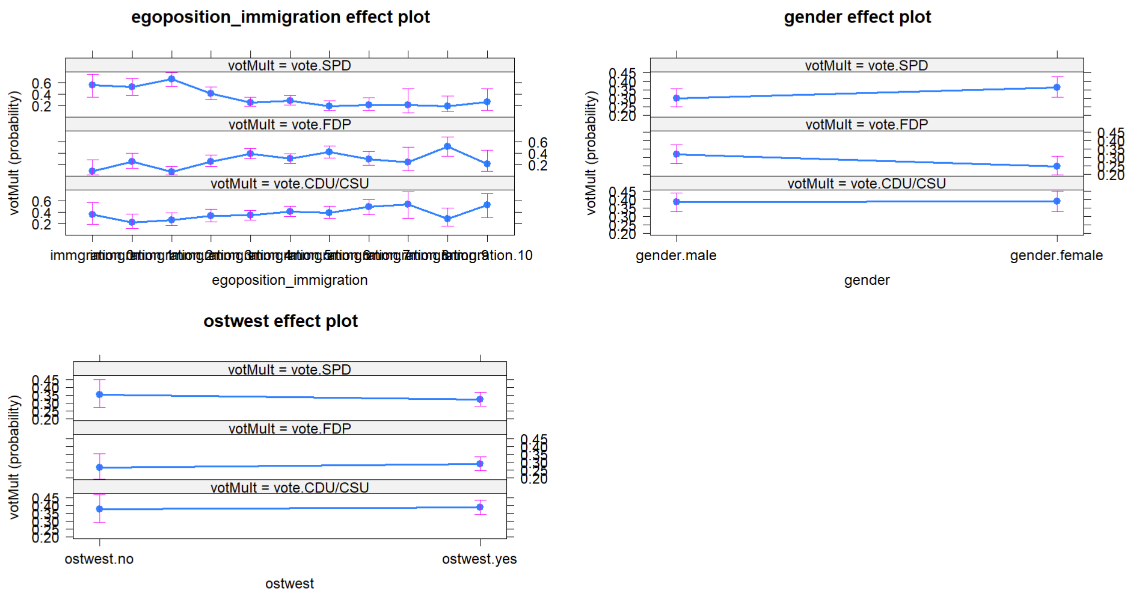


Figure 5-6: *Effects of the model.*

We can see how once again the variable with the greatest impact on the model's prediction is the position on immigration. We can see how the SDP party differs from the other two in the low immigration rates (open thoughts about it). CDU_CSU and FDP differ from each other by a specific level of immigration (level 9). It would be interesting as future work to study why this fact occurs, since it is not a trend as we would expect but a peak at a single level.

Gender also has a relevant impact with the SDP and FDP parties. The masculine gender favors FDP while the feminine favors the prediction of SDP.

Lastly, the ostwest variable has a slight impact on these same parties, with otwest yes voters being more likely to be voted SDP and otwest no for FDP voters.

6. Validating the Hierarchical model

In order to validate the results applying the models hierarchically taking into account that you accumulate the error in each layer we create the schema that follows, and we emphasize some interesting results during the execution and the final confusion matrix and metrics to evaluate correctly the whole model. To do this hierarchic process each model has been tested with the subset that was predicted for it. So at the end to validate the model we have needed to join all the datasets to evaluate the performance.

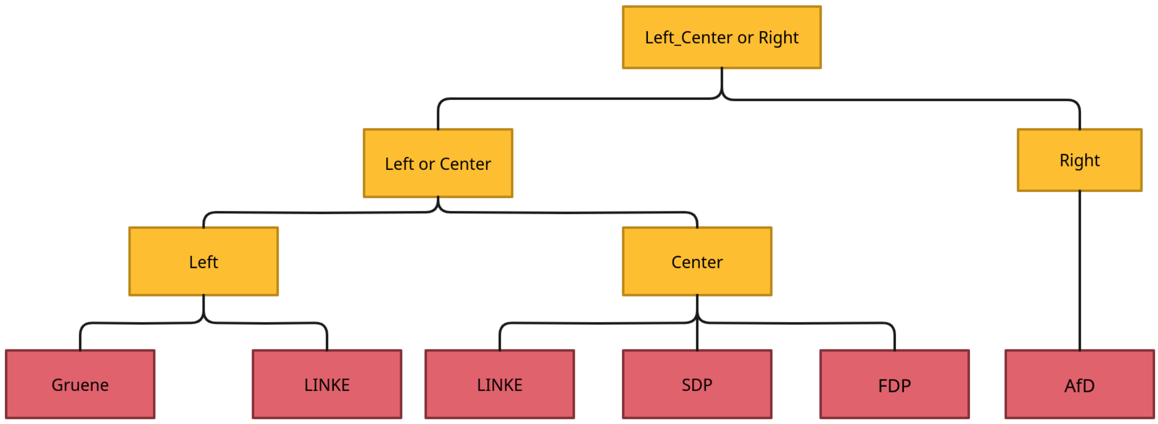


Figure 6-1: *Hierarchy schema of the models applied.*

In *[Figure 5.2]* the split of Left_center or Right can be seen that it's easy to detect the probabilities of right-wing voters, but there are a lot of misclassifications. This happens because there is a vast majority of left-center wing voters.

Figure 6-2: *Threshold on Left_center or Right model.*

In *[Figure 5.3]* the split of Left and center can be seen and it shows the misclassifications of the first model, and logically there are more classification errors of right-wing parties in the probability zone for center-wing parties, because in principle left and right wings are quite opposite. In this model we can see a lot of misclassifications, so this model is less accurate than the first one.

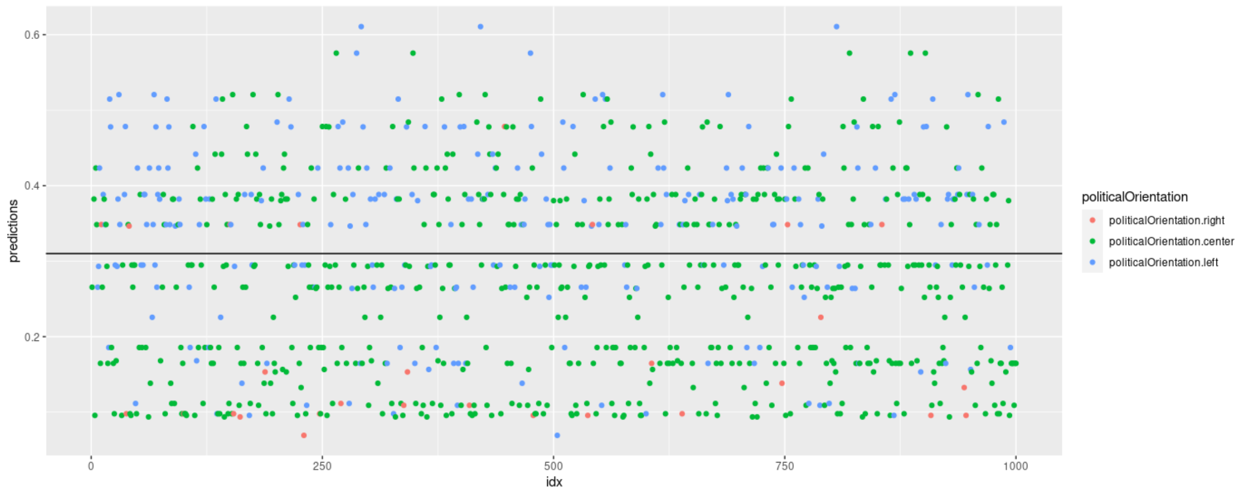


Figure 6-3: *Threshold on Left or Center model.*

From here the interpretation of the next binary model that classifies between left-wing parties is difficult because you accumulate a lot of classification errors. Ideally in this plot only light blue and dark blue should appear, but there are all the political parties in the dataset. Is difficult to detect, but more Gruene voters appear for low probabilities.

Figure 6-4: *Threshold on left-wing parties model.*

For the polytomous model the plot is not useful because there is a probability for each political party for each user so we decide to not show it.

To validate and evaluate the whole model, we joined the datasets of Right, Center and Left wing parties, all of them with the predicted labels and the original ones for **politicalOrientation** and **vote**.

For the political Orientation problem, we obtain 61.1% of accuracy, with a similar precision for each political orientation around 60%-61%. The biggest recall is for Center predictions with 77% while the others are between 40% and 43%. Finally the f1-score is bigger for Center predictions with a 68% and for Left and Right 50% and 48% respectively.

	pred.Center	pred.Left	pred.Right
politicalOrientation.center	408	205	52
politicalOrientation.left	96	162	8
politicalOrientation.right	20	8	41

Figure 6-5: *Confusion matrix for* ***politicalOrientation*** *in hierarchical model.*

```
> (accuracy <- sum(diag(tt1)) / sum(tt1))
[1] 0.611
> (precision <- diag(tt1) / rowSums(tt1))
politicalOrientation.center   politicalOrientation.left  politicalOrientation.right
                  0.6135338                   0.6090226                   0.5942029
> (recall <- (diag(tt1) / colSums(tt1)))
pred.Center   pred.Left  pred.Right
  0.7786260   0.4320000   0.4059406
> (f1 <- (2*precision*recall/(precision+recall)))
politicalOrientation.center   politicalOrientation.left  politicalOrientation.right
                  0.6862910                   0.5054602                   0.4823529
```

Figure 6-6: *Precision metrics of* ***politicalOrientation*** *in hierarchical model.*

For the vote problem, we obtain 32.9% of accuracy, with a very different precision for different political parties. AfD is the one with higher accuracy because right political orientation is well predicted, and as the dataset only contains one right-wing party. For right-wing parties Gruene is better predicted than LINKE although the percentage of voters is very similar.. Finally the worst results are for Center political parties where FDP only has a 8% of accuracy. But as in this case the votes are very unbalanced, so if we take a look at the recall and f1 the results are more positive.

	pred.AfD	pred.CDU/CSU	pred.FDP	pred.Gruene	pred.LINKE	pred.SPD
vote.AfD	41	18	2	1	7	0
vote.CDU/CSU	25	140	15	31	41	37
vote.FDP	14	58	10	13	13	13
vote.Gruene	0	28	8	57	27	23
vote.LINKE	8	22	4	45	33	11
vote.SPD	13	72	15	70	37	48

Figure 6-7: *Confusion matrix for* ***vote*** *in hierarchical model.*

```
> (accuracy <- sum(diag(tt2)) / sum(tt2))
[1] 0.329
> (precision <- diag(tt2) / rowSums(tt2))
   vote.AfD vote.CDU/CSU     vote.FDP  vote.Gruene   vote.LINKE     vote.SPD
 0.59420290   0.48442907   0.08264463   0.39860140   0.26829268   0.18823529
> (recall <- (diag(tt2) / colSums(tt2)))
   pred.AfD pred.CDU/CSU     pred.FDP  pred.Gruene   pred.LINKE     pred.SPD
  0.4059406    0.4142012    0.1851852    0.2626728    0.2088608    0.3636364
> (f1 <- (2*precision*recall/(precision+recall)))
   vote.AfD vote.CDU/CSU     vote.FDP  vote.Gruene   vote.LINKE     vote.SPD
  0.4823529    0.4465710    0.1142857    0.3166667    0.2348754    0.2480620
```

Figure 6-8: *Precision metrics of* ***vote*** *in hierarchical model.*

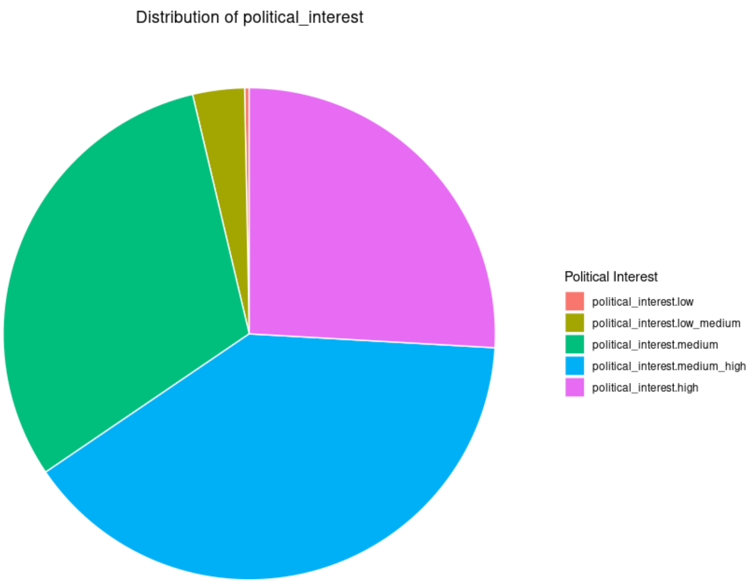
7. Conclusions

After conducting a thorough analysis of the dataset used in this study, we can reach the following conclusions:

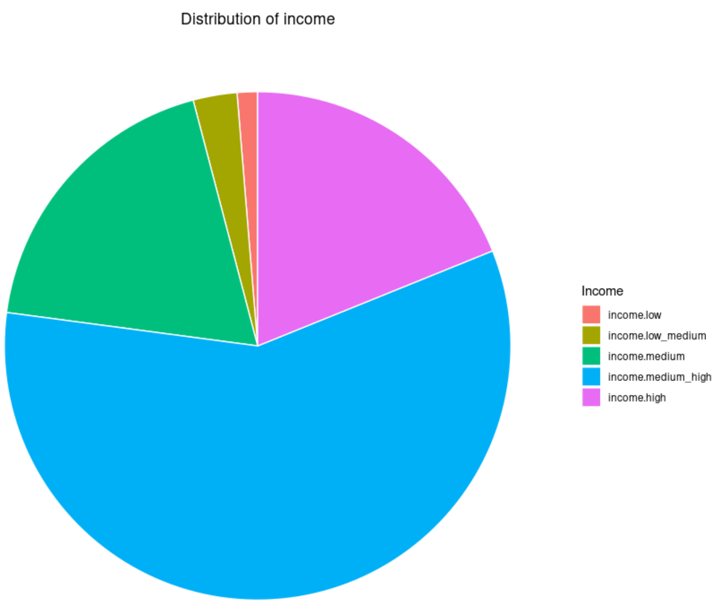
1. The dataset does not include sufficient variance or information to accurately predict the political orientation of individuals. This may be due to a variety of factors, such as a lack of data or a lack of diversity in the dataset used.
2. To address the imbalance issues in the dataset, we have tried various techniques and found that changing weights on individuals and using hierarchical models have given the best results.
3. The most important variable for predicting political orientation was found to be opinion on immigration. We believe that in future work, it could be interesting to delve deeper into this field while taking into account more information on individuals' opinions.
4. While we have not been able to obtain precise predictions using the current dataset, it has been very useful for performing statistical analysis and extracting conclusions from it. Additionally, we have identified potential improvements that could be applied in future work to obtain more accurate results.
5. It is interesting to remark that binomial models performance seems to give better predictions than multinomial models. An interesting future work could be to try to split the votes for the two big center-wing parties and the small one in order to convert the polytomous model into two binomial models. This decision has not been taken in order to add variety to the project with different model types.

In summary, although the dataset used in this study did not allow for accurate predictions of individuals' political orientation, it has been useful for performing statistical analysis and identifying potential improvements for future work.

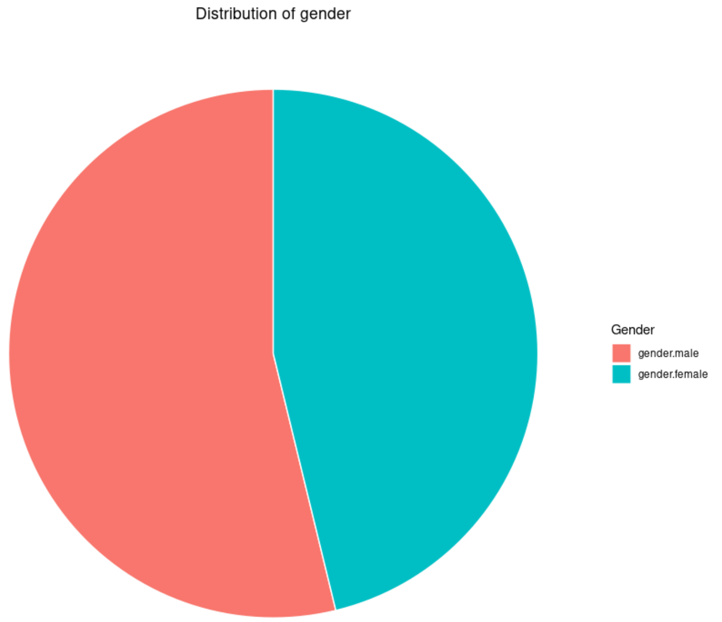
8. Annexes



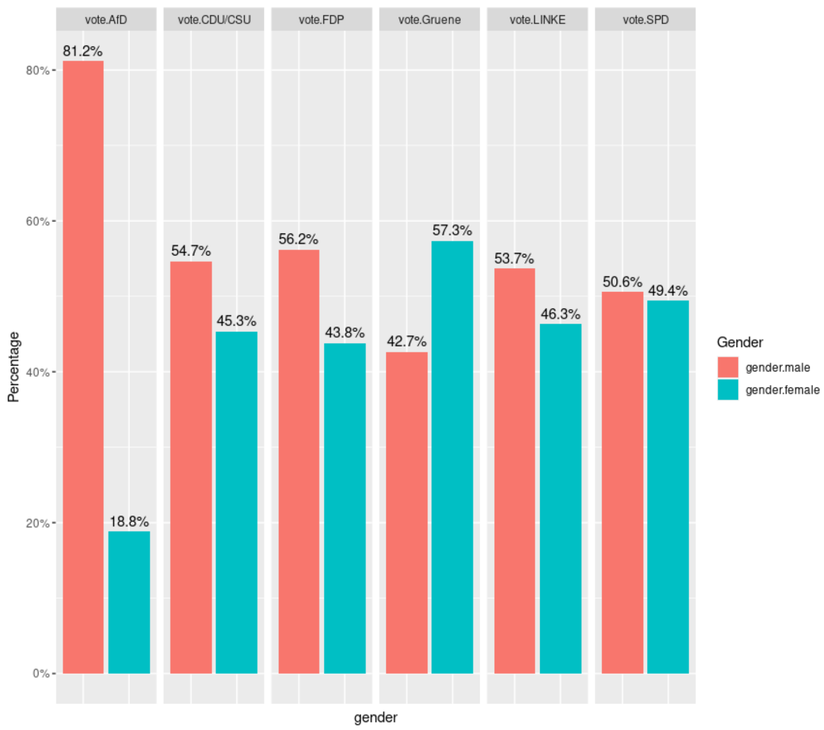
Annex 1: *Distribution of* ***political_interest****.*



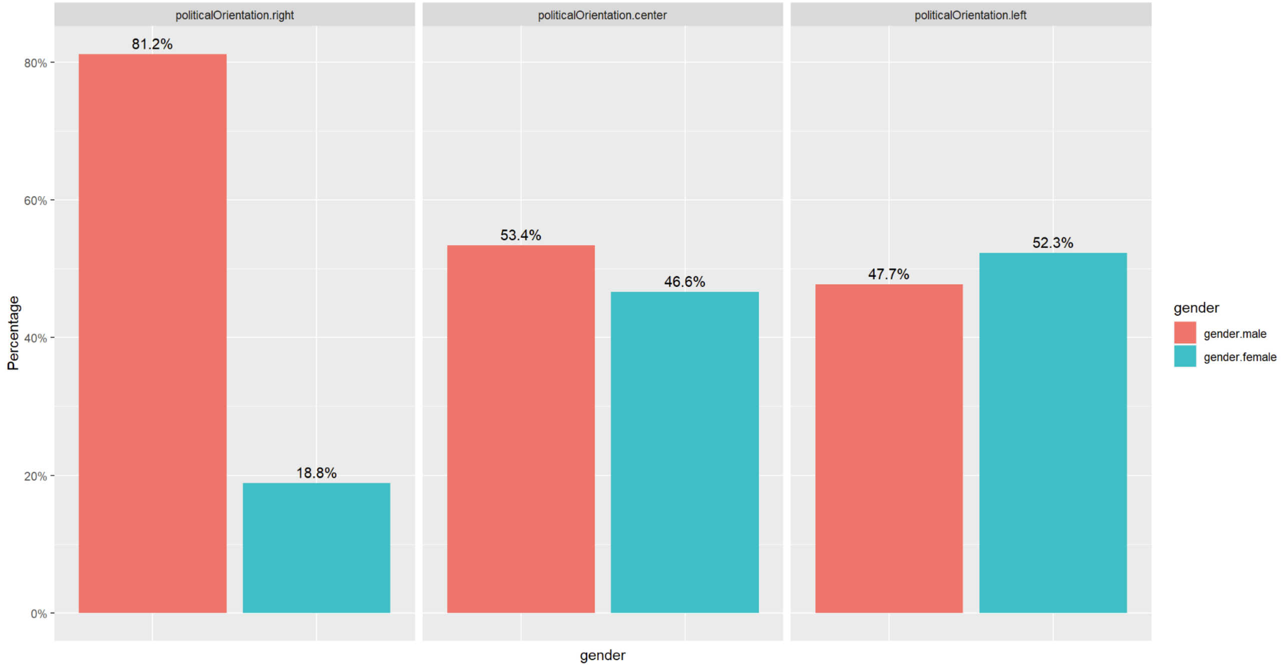
Annex 2: *Distribution of* ***income****.*



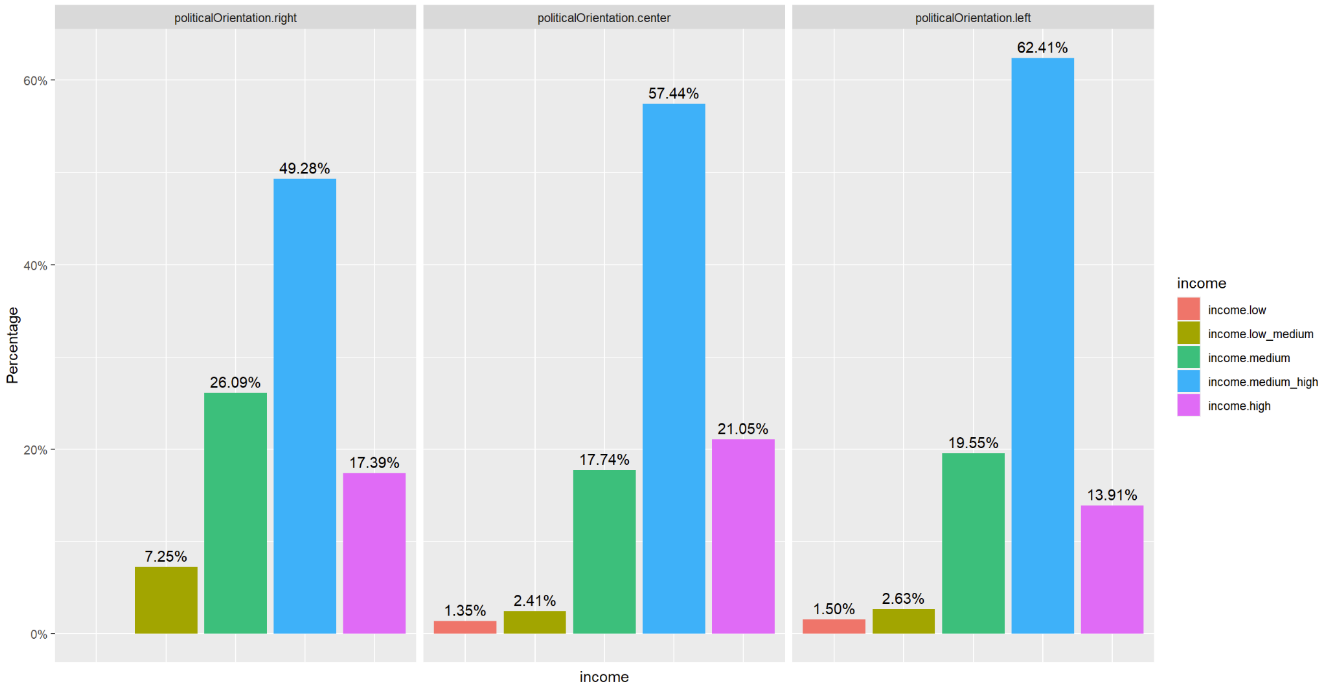
Annex 3: *Distribution of* ***gender****.*



Annex 4: *Distribution of* ***gender*** *grouped by political parties (****votes****).*



Annex 5: *Distribution of* ***gender*** *grouped by* ***politicalOrientation.***



Annex 6: *Distribution of* ***income*** *grouped by* ***politicalOrientation****.*