# Predicting Party Preference through Bayesian networks

Project for Probabilistic Modelling course

Marco Cazzola (matr. 964573)

## Introduction

The present work is inspired by the study of Pippa Norris and Ronald Inglehart investigating support for populist parties in Western democracies. In their book *Cultural Backlash* (2019), the authors argue that while the younger generations (and society in general) are getting more and more progressive, older people become feeling as "stranger in their own land", since the values according to which they were educated are no longer dominant. As a result, a *tipping point* has emerged, where conservatives have become increasingly resentful at finding themselves becoming the minority. The authors' hypothesis is that the tipping point in public opinion can catalyze social conservatives into voting for authoritarian-populist leaders.

This paper wants to test Norris and Inglehart's hypothesis in the Italian context, by building a discrete Bayesian network that predicts party preference based on individual's characteristics, both in terms of sociodemographics and moral values. As argued in the book Cultural Backlash, the expectations are that, as we condition on characteristics like age, education or urbanization, party preference should change substantially; in particular, we expect the older people, low educated and living in the countryside to be far more likely to support conservative values. On the contrary, younger people, highly educated and living in a big city are expected to support progressive values and vote accordingly.

## Data set description

The data we will use come from the 9th wave of the *European Social Survey* (ESS)[1] and regards exclusively the Italian context. The variables that we will use throughout the analysis are:

- `prtvtcit`: Party voted for in last national election (2018, in case of Italy). Party preference has been recoded in just three groups: `M5S`, for those who voted *Movimento 5 Stelle*; `CSX` for those who voted for center-left parties (PD, +Europa, Civica Popolare Lorenzin, LeU); `CDX` for those who voted for center-right parties (Forza Italia, Lega Nord, FdI, UDC). Those who did not vote for any of these parties were dropped from the analysis.
- `euftf`: Support for European Union, on a scale from 0 to 10. The numerical scale has been recoded into three categories: `LOW` for values lower than 4; `MODERATE` for values between 4 and 6 (included); `HIGH` for values greater than 6.
- `imwbcnt`: Support for immigration, on a scale from 0 to 10. The numerical scale has been recoded into three categories: `LOW` for values lower than 4; `MODERATE` for values between 4 and 6 (included); `HIGH` for values greater than 6.
- `rlgdgr`: Level of religiosity, on a scale from 0 to 10. The numerical scale has been recoded into three categories: `LOW` for values lower than 4; `MODERATE` for values between 4 and 6 (included); `HIGH` for values greater than 6.

[1] Since the variables' names have been left unchanged, you can refer to the ESS9 codebook for further details.

- `stflife`: Level of life satisfaction, on a scale from 0 to 10. The numerical scale has been recoded into three categories: `LOW` for values lower than 4; `MODERATE` for values between 4 and 6 (included); `HIGH` for values greater than 6.
- `gincdif`: Support for redistributive policies, on a scale from 1 to 5. The numerical scale has been recoded into three categories: `HIGH` for values lower than 3; `MODERATE` for values equal to 3; `LOW` for values greater than 3.
- `hmsacld`: Support for gay rights, on a scale from 1 to 5. The numerical scale has been recoded into three categories: `HIGH` for values lower than 3; `MODERATE` for values equal to 3; `LOW` for values greater than 3.
- `edlveit`: The 21 categories describing the level of education have been recoded into just three classes: `Low edu`, for people having a level of education below the diploma (classes from 1 to 7); `College edu`, for those having a diploma (classes from 8 to 10); `University edu`, for those having a level of education above the diploma (classes from 11 to 21). Observations having `Other` as level of education were discarded.
- `domicil`: The level of urbanization where the individual lives. There are four classes: `Big city`, `Suburbs`, `Small city`, `Countryside` (the latter obtained by joining *Country village* and *Farm* original categories).
- `hincfel`: The level of subjective economic insecurity, measured as how the individual is living with current household's income. There are three possible categories: `Comfortably`, `Coping` and `Difficult` (the latter obtained by joining the original `Difficult` and `Very difficult` categories).
- `lrscale`: Ideological self-placement on a scale from 0 to 10, where 0 represents extreme left and 10 extreme right. The numerical scale has been recoded into three categories: `LEFT` for values lower than 4; `MODERATE` for values between 4 and 6 (included); `RIGHT` for values greater than 6.
- `yrbrn`: The respondent's year of birth. The numerical variable has been recoded into four categories, as done by Norris and Inglehart. In particular, the transformation shown in the table has been applied. Individuals that were born after 1996 have been discarded from the analysis (due to the limited number of cases).

| **Original year of birth** | **Categorization** |
|---|---|
| $< 1945$ | Interwar |
| 1946 - 1964 | Boomers |
| 1965 - 1979 | Generation X |
| 1980 - 1996 | Millennials |

The resulting data set, cleaned from any missing value, was composed of 1028 observations for each of the 12 variables; their distribution is shown below.

```
##  prtvtcit       euftf          imwbcnt         rlgdgr        stflife
##  M5S:395   LOW     :344   LOW     :379   LOW     :254   LOW     : 56
##  CSX:302   MODERATE:356   MODERATE:502   MODERATE:353   MODERATE:244
##  CDX:331   HIGH    :328   HIGH    :147   HIGH    :421   HIGH    :728
##
##      gincdif        hmsacld              edlveit           domicil
##  LOW     : 44   LOW     :523   Low edu       :465   Countryside:434
##  MODERATE:115   MODERATE:223   College edu   :387   Small city :384
##  HIGH    :869   HIGH    :282   University edu:176   Suburbs    : 70
##                                                     Big city   :140
##         hincfel            yrbrn         lrscale
##  Difficult  :219   Interwar   :130   LEFT    :272
##  Coping     :498   Boomers    :380   MODERATE:396
```

```
##  Comfortable:311   Gen X      :292   RIGHT   :360
##                    Millennials:226
```

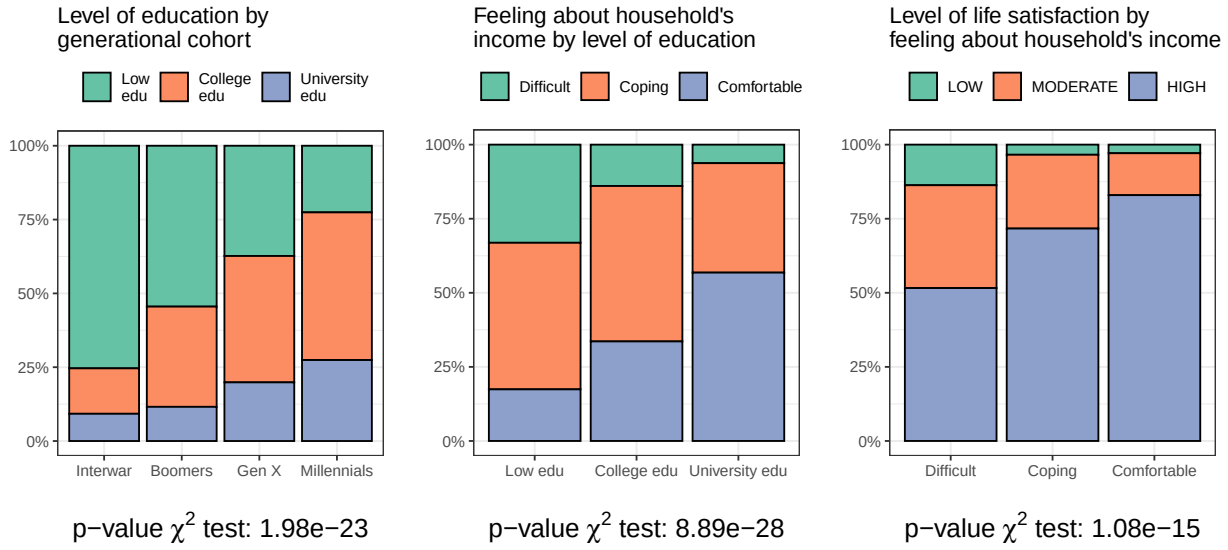
Notice how our target variable, `prtvtcit`, is quite nicely distributed; all the other variables, perhaps with the exception of `euftf`, are instead skewed: for example, there are few people showing high support for immigrants, and the majority of the people show low support for gay rights.

# Methodology

As we said in the Introduction, the goal of the paper is to build a discrete Bayesian network that predicts party preference based on the individual's characteristics reported in the data set. The obtained network will then be used to study how the variables interact with one another and make inference.
The first step is therefore to learn the structure of the graph. We will do so in two ways: one time relying exclusively on learning algorithms, another time by taking into account prior knowledge about variables' relationships. In particular, we will assume that:

- `yrbrn` → `edlveit`: as argued also in the book *Cultural Backlash*, high-income Western societies have been experiencing growing access to highest levels of education, so that the year when one is born can influence their level of education;
- `edlveit` → `hincfel`: it is a well known and established fact that the higher the level of education of an individual, the higher their level of income;
- `hincfel` → `stflife`: one of the most relevant predictors of life satisfaction is the level of income.

Notice how all these assumptions are perfectly met by the data, as shown by the plots below.




Moreover, we will also assume that:

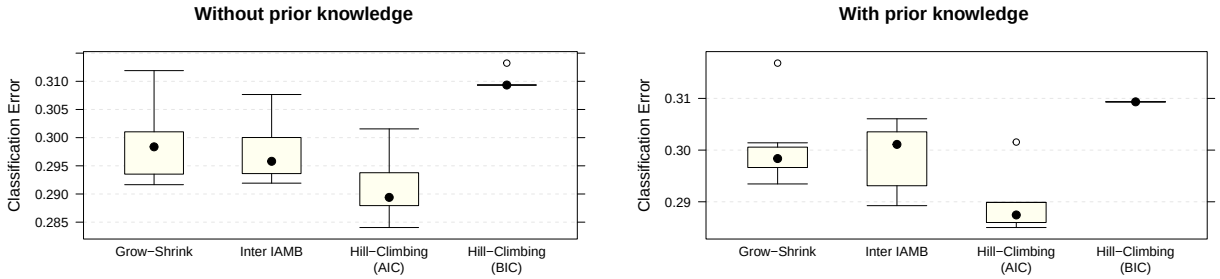
- `yrbrn` cannot be (of course) influenced by any variable;
- `prtvtcit` cannot influence any variable, since we are assuming that `prtvtcit` is our target variable that is explained by the other variables in the data set;
- `rlgdgr` cannot be influenced by `imwbcnt`, `hmsacld`, `euftf` and `gincdif` (while the reverse may be plausible);
- `hincfel` cannot be influenced by `stflife`, `lrscale`, `euftf`, `imwbcnt`, `gincdif` and `hmsacld` (while the reverse may be plausible);

- `edlveit` cannot be influenced by `gincdif`, `euftf`, `imwbcnt`, `hmsacld` and `hincfel` (while the reverse may be plausible).

The methods we will use to learn the graph structure are:

1. *Grow-Shrink*: a constraint-based method based on iteratively testing conditional independencies;
2. *Interleaved Incremental Association*: a variant of traditional IAMB algorithm which is more robust to false positives;
3. *Hill-Climbing with AIC score*: a score-based learning method;
4. *Hill-Climbing with BIC score*: a score-based learning method with a larger penalty than AIC.

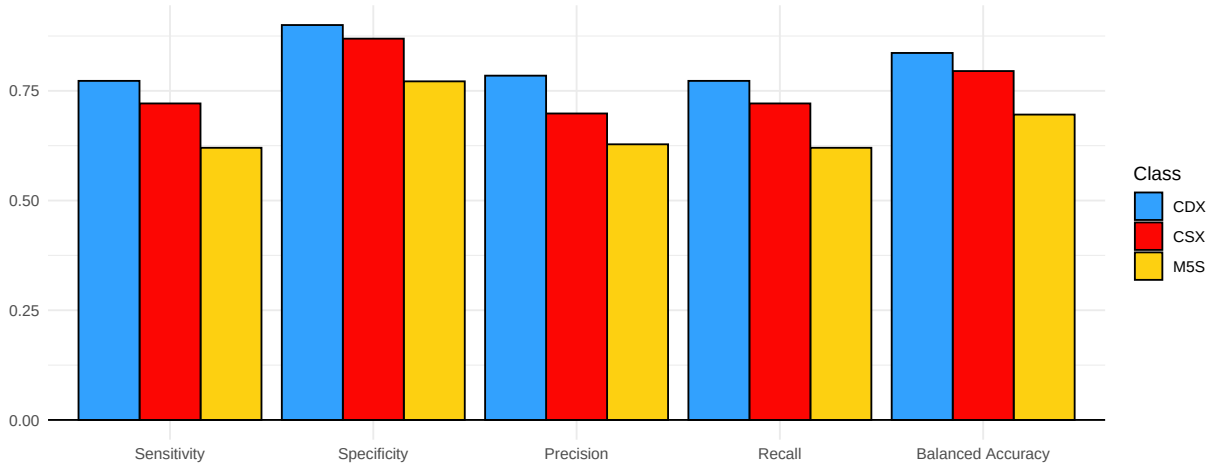
Each of these algorithms will be run with and without prior knowledge assumptions. In order to select the optimal one, we run 10 times cross-validation on the original data set and evaluate the results according to the traditional misclassification error, where our target variable is `prtvtcit`, the party preference. The results are plotted below.



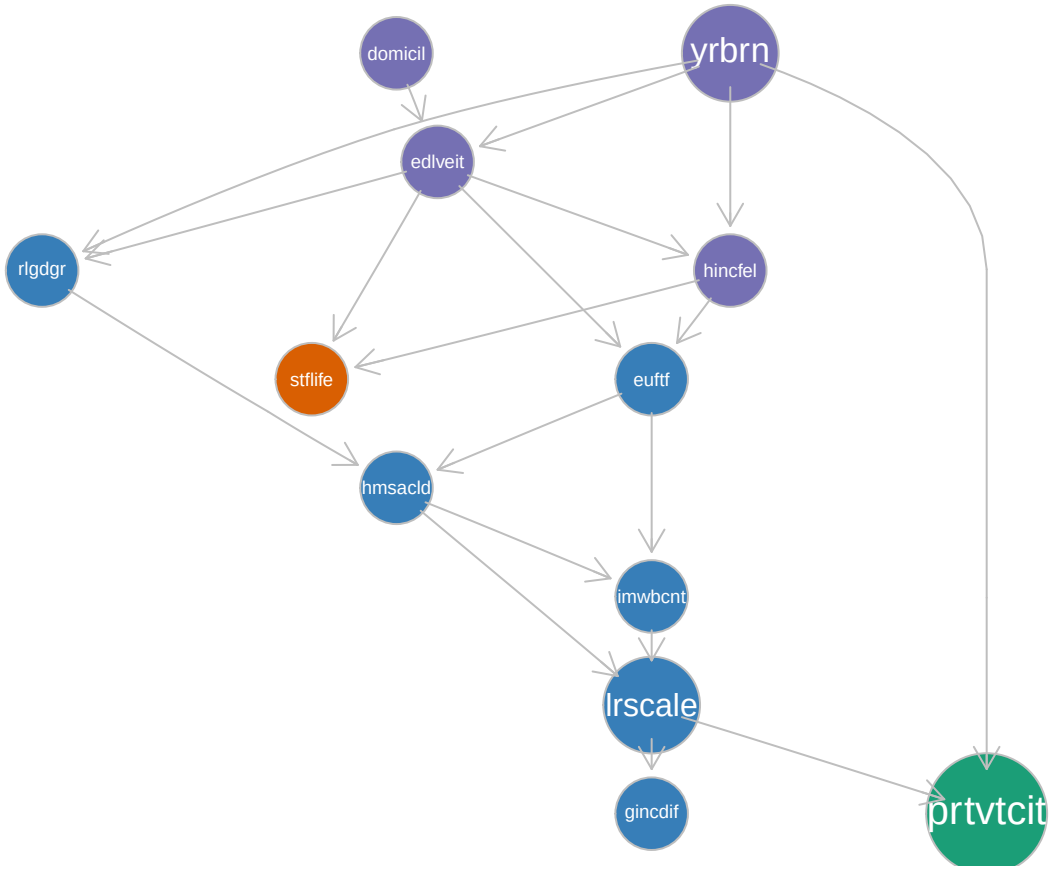

Both with and without prior knowledge, the best result is achieved by Hill-Climbing with AIC score. Moreover, the error rate of the two models obtained with that algorithm are not significantly different, and they all lie around 30%. In order to have a more interpretable network, we select the one obtained using prior knowledge assumptions.

## Building the network

After splitting the original data set into a training set (80% of the observations) and a test set (20% of the observations), we train the algorithm on the training set and test the resulting network on the test set to predict party preference. The results are shown below.

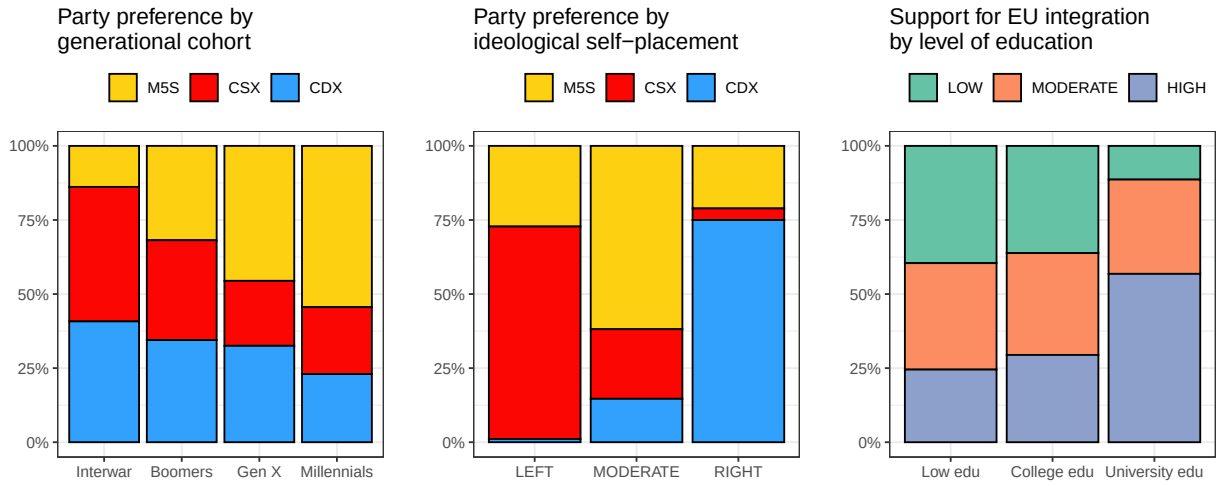


The plot shows how, irrespective of the metric we consider, `CDX` has always the highest score, while `M5S` has the lowest. This seems to suggest that `CDX` electors are easier to recognize than `M5S` supporters. The performance of the model are coherent with the ones observed through cross-validation, as the misclassification rate on the test set is around 30%. We now plot the graph learned through the Hill-Climbing algorithm with AIC score.



### Dependencies

The only two factors directly influencing party preference are `lrscale` (ideological self-placement) and `yrbrn` (year of birth). The hypothesis made by Norris and Inglehart is only partially met: in fact, as we can see from the graphs below, while it is true that older generational cohort tend to support conservative values and vote for center-right and right-wing parties, younger generations are indeed attracted by progressive values, but their electoral support goes to populist parties like M5S, rather than to traditional progressive parties like PD and LeU.
The ideological self-placement is of course a strong predictor of party preference, and it is interesting to notice how, according to the graphs below, a person who poses herself in the `LEFT` political spectrum has more or less the same probability of being a M5S supporter with respect to another person having the opposite placement. On the contrary, those who see themselves as `MODERATE` are far more likely to vote M5S than any other party.
Finally, another interesting dependency is the one between support for EU integration and level of education: while there is no big difference between those having lower degree of education and those with a college degree, the real gap is between those having a university degree and those who have not, with the former category showing a substantially higher support for EU integration.

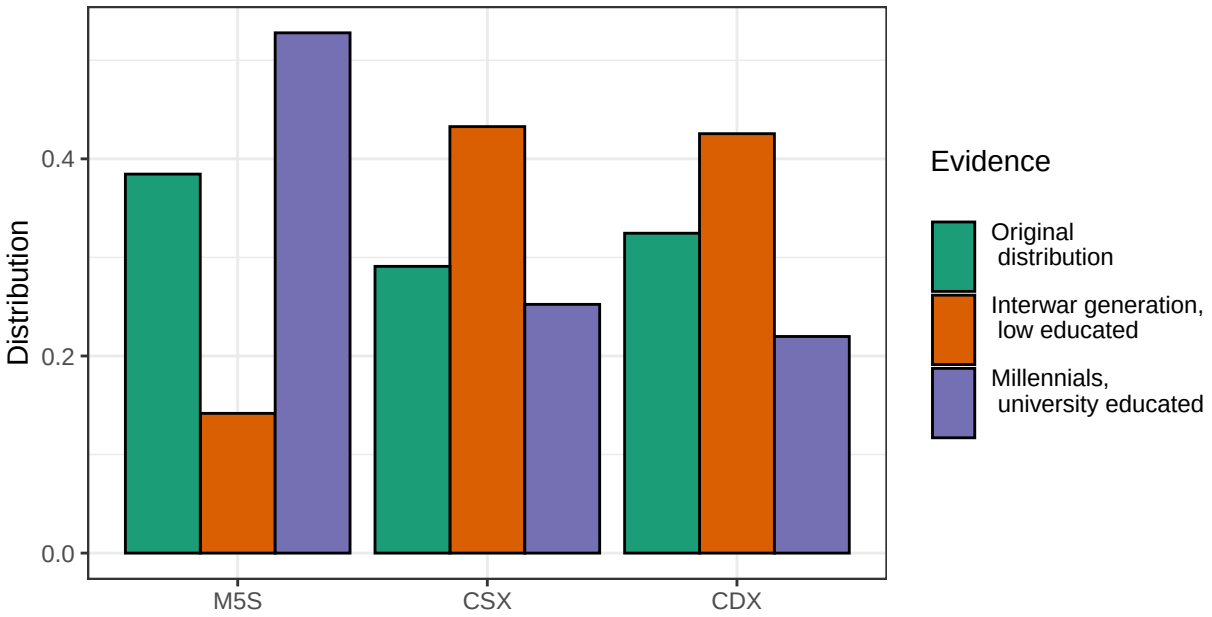



### Independencies

- `domicil` $\perp$ `euftf` | `edlveit`, `yrbrn`: the aversion or support for the European Union does not seem to depend on the place where one lives, once we take into account the level of education and the year of birth. This seems to suggest that big city are more cosmopolitans because they are inhabited by young and highly educated people, rather than because of some inner characteristic.
- `edlveit` $\perp$ `hmsacld` | `rlgdgr`, `euftf`: the support for gay rights does not seem to depend on education, once we take into account religiosity and support for EU integration. This seems to suggest that support for gay rights has mainly to do with someone's personal values, rather than with his/her education.
- `edlveit` $\perp$ `lrscale` | `euftf`, `hmsacld`: the same reasoning may be applied to this evidence, according to which the level of education is independent from the ideological self-placement, once we take into account some personal values like support for the EU integration and gay rights.

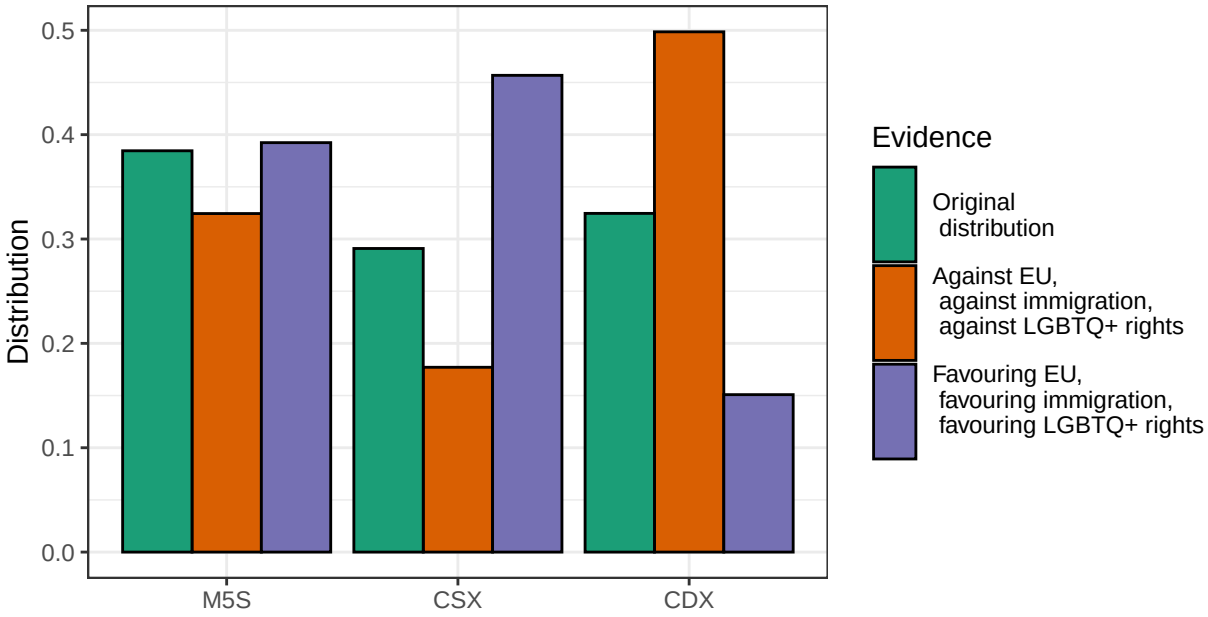
## Inference

We now exploit the graph structure to make some inference about how the distribution of votes changes as we condition on some electorate's characteristics. For example, we may be looking at how the year of birth and the level of education influence the party preference.




Interestingly, parties categorized as `CSX` and the ones from `CDX` collect approximately the same consensus among the older and lower educated generation: this means that, at least in Italy, older and low educated people are not so unbalanced in favor of right-wing and conservative parties. Another interesting point is that, for what concerns university educated Millennials, the support for the `CDX` area is only slightly lower with respect to the support for `CSX` parties, and the real difference is between these two blocks (which we can think of as "traditional" parties) and M5S, a populist party which, at least partially, supports progressive values, especially from a welfare state policy point of view.

Another potentially relevant inference in which we could be interested in is understanding how personal opinions about LGBTQ+ rights, immigration and European integration influence the party preference.




In this context, differences between `CDX` and `CSX` become quite more evident: indeed, the two distributions are quite specular, with those favoring EU, immigration and LGBTQ+ rights supporting mainly parties from `CSX`, while those having opposite position on all these issues are much more likely to vote for `CDX` parties. Once again, we find very interesting evidence about M5S, with those supporting all the issues having only a slightly larger probability of voting M5S with respect to those that oppose all these issues. This confirms the heterogeneous nature of M5S' electorate.

## Conclusion

In this paper, we investigated the role of sociodemographic characteristics and personal values in determining one's party preference. Using the data from the 9th Wave of the European Social Survey, we built a Bayesian network and exploited the particular structure of this model for better understanding the relationships between small subsets of variables.

Among sociodemographic variables, the one having the strongest impact on party preference is the year of birth: in contrast with our starting expectations, we found out that the older generation is not so unbalanced in favor of right-wing parties, and it is indeed quite balanced between the `CDX` and `CSX` area. On the contrary, younger generation tends to support parties that are outside those two traditional blocks, like M5S.

If we move from sociodemographic characteristics to personal values and political opinions, we would instead see a much clearer separation between `CDX` and `CSX`, while M5S seems to attract both those having conservative ideas and those with more progressive attitudes.