Predictive and Exploratory Analysis of Social Attitudes Towards Controversial Issues

Computational Social Science
https://github.com/filipponardi17/Controversial-Attitudes-Analysis

Author: Filippo Nardi

1 Introduction

In recent years, societal attitudes towards controversial topics such as abortion, same-sex marriage, and immigration have garnered significant attention due to their profound implications on public policy, social cohesion, and individual rights. However, understanding the intricate interplay between demographic characteristics and the formation of these attitudes remains a complex and pressing challenge.

This research seeks to delve into this complexity by examining how demographic factors, including age, gender, education, income, and cultural background, shape individuals' perspectives on contentious social issues. Leveraging the rich dataset provided by the World Values Survey, this study employs advanced explorative techniques such as Latent Class Analysis (LCA) and latent factor analysis alongside predictive machine learning algorithms (such as SVM, Random Forest, KNN, and Logistic Regression) to uncover underlying patterns and predict social attitudes.

By examining the complex relationship between demographic variables and attitudes toward abortion, same-sex marriage, and immigration, this research aims to provide valuable insights into the drivers of public opinion formation. These findings can inform policymakers, social scientists, and advocacy groups in crafting more effective strategies to address societal divisions and promote inclusive dialogue on these contentious issues.

2 Methodology

2.1 Dataset

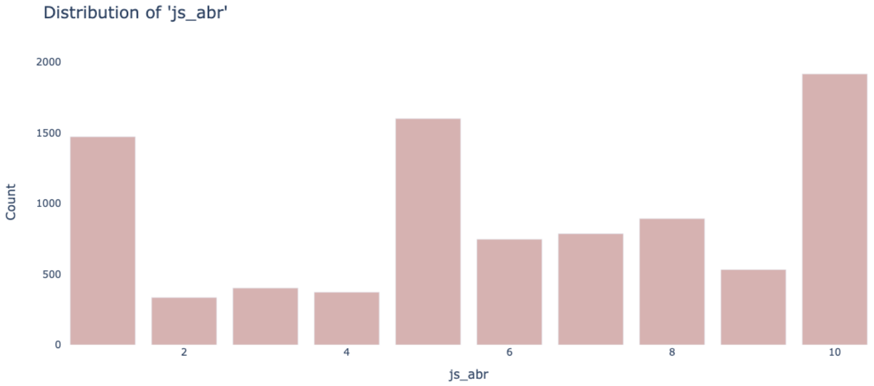
For this study, I utilized data from the joint European/World Values Survey (EVS-WVS) 2017-2020, which covers a broad dataset collected between 2017 and 2022 EVS (2017-2020). This comprehensive survey provides insights from 90 countries worldwide, with a total of 153,716 respondents. However, my analysis focuses on the four largest European economies by GDP—Germany, France, Spain, and Italy. I selected these countries due to their large respondent base and their relative comparability in terms of economic and social structures. The data was gathered through face-to-face interviews conducted in respondents' homes between 2017 and 2021, ensuring a high level of data quality and consistency across the sample.

The codebook for the survey will be in the reference link (World Values Survey Association, 2024), as well as a table to see what the variable name is in the codebook (Filippo Nardi, 2024).

2.2 Data Cleaning and Pre-Processing

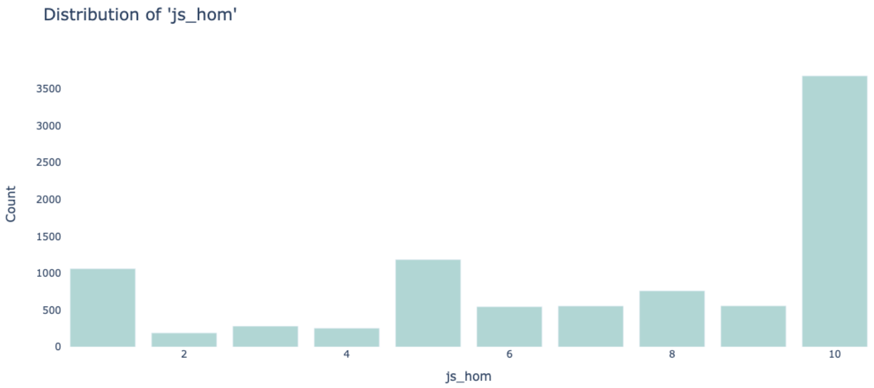
The controversial issues I selected as target variables for this study, and their distribution are as follows:

- **js_abr** which relates to this question in the WVS: Justifiable: Abortion; Please tell me for each of the following whether you think it can always be justified, never be justified, or something in between, using this card. (1 Never justifiable, 10 Always justifiable)



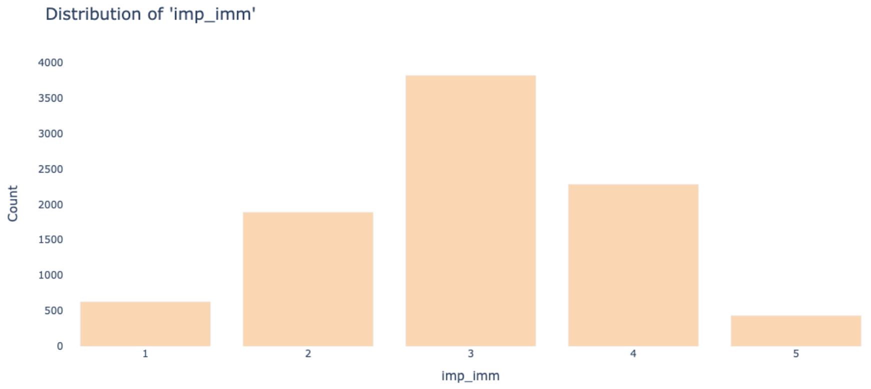
Distribution of **js_abr**

- **js_hom** which relates to this question in the WVS: Justifiable: Homosexuality; Please tell me for each of the following whether you think it can always be justified, never be justified, or something in between, using this card. (1 Never justifiable, 10 Always justifiable)



Distribution of **js_hom**

- **imp_imm** which relates to this question in the WVS: Now we would like to know your opinion about the people from other countries who come to live in [your country] - the immigrants. How would you evaluate the impact of these people on the development of [your country]? (1 Very bad, 5 Very good)

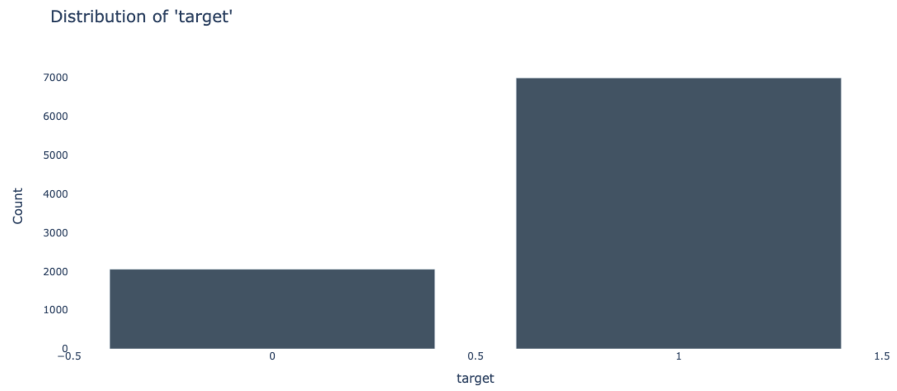


Distribution of **imp_imm**

To investigate the controversial issues, I retained all questions and key demographic features as potential predictors. However, missing data posed a challenge for the algorithms I employed, prompting me to consider several imputation techniques. I tested three primary methods: Classification and Regression Trees (CART), Predictive Mean Matching (PMM), and Miss Forest. After conducting preliminary trials on data from five countries, I found that all methods produced similar results and did not affect the outcomes of the analysis. Therefore, I opted to use PMM due to its simplicity and speed. PMM works by identifying the nearest neighbors of data points with missing values and imputing values from other data points with similar characteristics.

In the data cleaning and preprocessing phase, outliers were removed during the to improve the reliability of the results. I first created individual sub-dataframes for each country. I then combined the data from France, Italy, Germany, and Spain into a single dataframe, referred to as `dfEU`. A new variable, `target`, was created by summing the variables `dfEU['js_abr']`, `dfEU['js_hom']`, and `dfEU['imp_imm']`. To simplify the analysis, this `target` variable was converted into a binary format, with values of either zero or one. Specifically, if the sum of these three variables was less than 12, the target variable was set to zero, indicating a higher prevalence of controversial opinions (as the maximum possible sum was 25). This transformation helped classify individuals based on the predominance of controversial attitudes. This variable was only utilized in the predictive portion of the paper, ensuring that its binary classification served as a key component for the predictive analysis. After this passage, scaling was applied only to ensure the comparability of the variables.

The analysis highlighted a significant imbalance in the distribution of the `dfEU['target']` variable (as detailed in the figure below, with controversial opinions appearing more than three times less frequently than non-controversial ones. This imbalance skewed the predictive models towards non-controversial outcomes.



Distribution of **target**

To address this issue, I explored various rebalancing techniques for the training dataset while preserving the integrity of the validation set. These methods included oversampling techniques such as SMOTE and ADASYN, hybrid oversampling, and undersampling. Interestingly, undersampling proved to be the most effective approach.

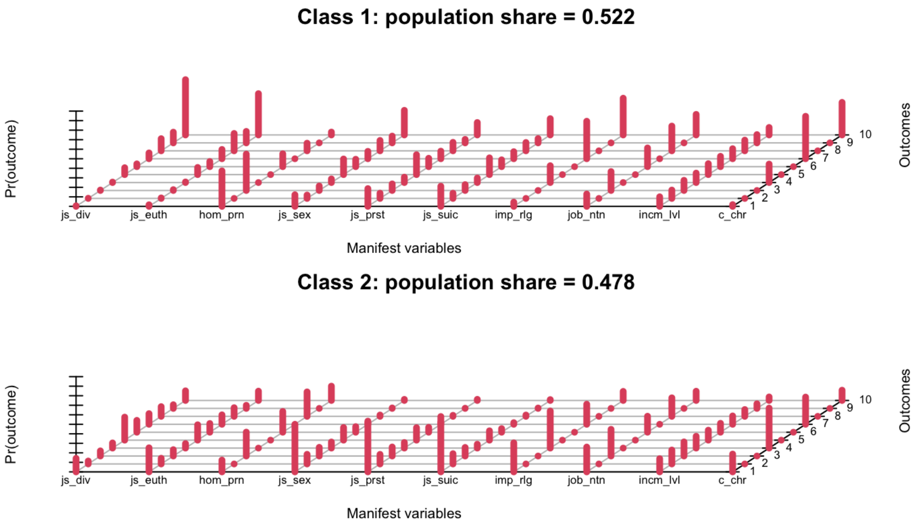
Subsequent trials with different undersampling ratios revealed that ratios of 0.7 and 0.8 yielded the best performance. These results highlight the importance of rebalancing techniques when dealing with imbalanced datasets, with undersampling emerging as a particularly effective strategy for improving predictive accuracy.

3 Exploratory Analysis

In this section, I outlined the Exploratory Analysis conducted using several key techniques. Latent Class Analysis (LCA) was applied to identify hidden groups or patterns within the data, while Principal Component Analysis (PCA) was used to examine the loadings of the first two principal components, capturing the majority of the variance and revealing the dataset's underlying structure. Random Forest Feature Importance further helped evaluate the significance of each feature in predicting the target variable, offering insights into key drivers and guiding feature selection and model refinement. Additionally, Exploratory Factor Analysis (EFA) was performed to uncover latent factors explaining the correlations among variables, providing a deeper understanding of the data structure and enhancing the overall model development.

3.1 Latent Class Analysis

I applied the Latent Class Analysis (LCA) algorithm to gain insights into the subpopulations within my dataset. Given that the target variable was categorized as binary, I opted to use LCA to identify two distinct classes, providing a clear distinction between the primary subgroups in the data. This approach allowed me to capture the underlying heterogeneity within the dataset and focus on the two major classes that played a pivotal role throughout the analysis. By examining these subpopulations, I could better understand their characteristics and how they contributed to the overall findings of the study. The results of the analysis were as follows:


Class 1: population share = 0.522

Class 2: population share = 0.478

Latent Class Analysis

The two subpopulations are relatively similar in size, with one comprising 48 percent of the data and the other 52 percent. The characteristics of these two groups reveal distinct social and economic viewpoints. By examining each class in detail, we can better understand their attitudes and values, which provide valuable insights into the broader patterns present in the data. Below is a breakdown of the defining features and ideologies of each class, offering a deeper exploration of how these subpopulations diverge on key social and economic issues.

- **Class 1: Socially Conservative and Economically Protectionist** This group displays very low acceptance of divorce (`js_div`), along with similarly restrictive attitudes towards euthanasia (`js_euth`) and homosexuality (`hom_prn`). Their stance on sexual norms (`js_sex`) and prostitution (`js_prst`) reflects traditional values, while their opposition to suicide (`js_suic`) is also pronounced. Religion (`imp_rlg`) plays a significant role in shaping their worldview, strongly influencing their conservative views. Economically, this class prefers national over foreign workers (`job_ntn`) and demonstrates a preference for income equality, indicating a protective stance on economic issues. Confidence in religious institutions (`c_chr`) is high, further reflecting their traditionalist values.

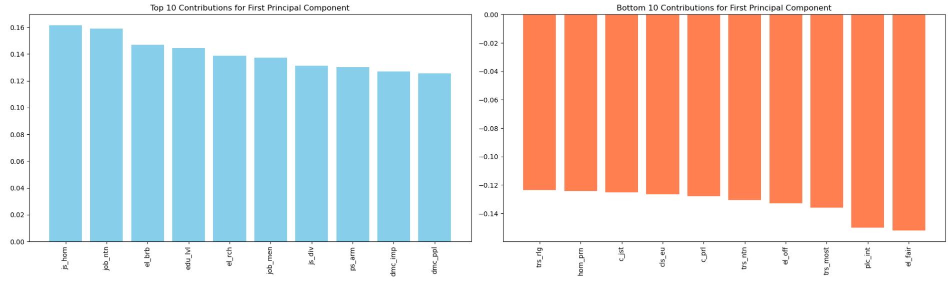
- **Class 2: Moderately Progressive** This group is more accepting of divorce and displays moderate openness towards euthanasia and homosexuality. While their views on sexual norms and prostitution are less conservative than Class 1, some reservations remain. They show more permissiveness regarding suicide compared to Class 1. Despite their progressive tendencies, they still place significant importance on religion, suggesting that traditional values influence their outlook. Economically, they are less protectionist, showing more openness to foreign workers and a balanced view on income distribution. Confidence in religious institutions is present but lower than in Class 1.

These results suggest that while both classes adhere to traditional values, they differ in their willingness to embrace social change and their economic priorities. Understanding these distinctions can provide valuable insights for policymakers and researchers, helping to design more effective interventions in public policy and social programs.

3.2 Principal Components Loadings Analysis

To further explore the underlying structure of the dataset, Principal Component Analysis (PCA) was applied. This method reduces the dimensionality of the data while retaining most of its variability, making it an effective tool for identifying key patterns. By examining the loadings of the first two principal components, we gain insights into the most influential variables and their contributions to the overall variance, enhancing our understanding of the dataset's composition.

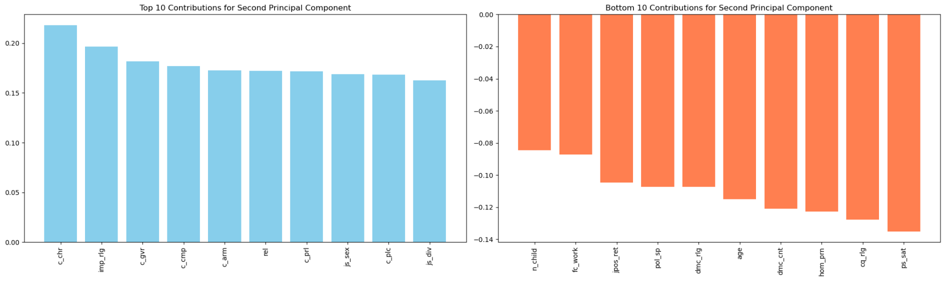
The loadings of the first components reveal significant data structure insights as follows:



Principal Component 1 Loadings

- **Component 1: Perceptions of Justice and Equity:** This factor centers on the societal norms and values that govern ideas of justice, fairness, equality, and societal roles. The variables with the highest loadings, such as the justifiability of homosexuality (*js_hom*) and attitudes towards job opportunities for different nationalities (*job_ntn*), underscore a community's perspective on social fairness, inclusivity, and the equitable treatment of diverse groups. These variables reflect how societies are grappling with critical issues surrounding the rights of marginalized groups and the distribution of opportunities across social lines. For example, the emphasis on justifiability of divorce (*js_div*) and homosexuality (*js_hom*) reveals a collective reflection on personal freedoms and moral flexibility, while political representation measures (*dmc_imp*, *dmc_ppl*) highlight concerns around civic participation and the fair representation of voices within the political system. In contrast, the variables with the lowest loadings—such as trust in people of another religion (*trs_rlg*), confidence in the judicial system (*c_jst*) and Electoral Integrity (*el_fair*) suggest a more complex and perhaps fragmented perception of institutional fairness. This could indicate that while individuals may support abstract principles of justice and equity, they remain skeptical about whether these ideals are effectively upheld in practice. The disparity between theoretical commitment to justice and the lived experience of inequities within judicial or societal systems points to underlying tensions.

Presented below is the plot for the second principal component loadings, highlighting the key variables driving its variance and revealing additional data patterns



Principal Component 2 Loadings

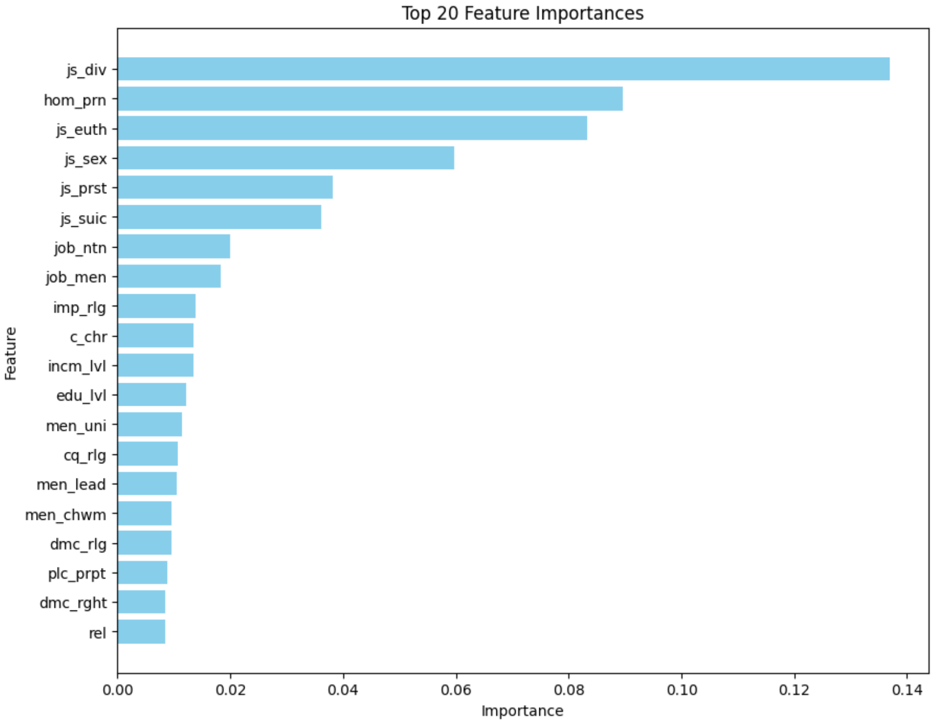
- **Component 2: Societal Values and Religious Influence:** This factor emphasizes the impact of religious values, governance, and societal norms. High-loading variables like confidence in churches (`c_chr`) and the importance of religion in life (`imp_rlg`) indicate a strong influence of religious beliefs on shaping societal attitudes and behaviors. Trust in governmental and social institutions (`c_gvr`, `c_cmp`, `c_prl`) further suggests a close connection between religious influence and confidence in these entities.

  The presence of variables related to personal behavior and moral judgments, such as the justifiability of casual sex (`js_sex`) and divorce (`js_div`), highlights the ethical guidance that religion and societal values provide in shaping both personal life decisions and public morality.

  In contrast, the lowest loadings—such as the number of children (`n_child`), future importance of work (`fc_work`), and political self-positioning (`pol_sp`)—indicate that these aspects are less influenced by religious and ethical frameworks. Negative loadings on variables like political system satisfaction (`ps_sat`) suggest these areas may be seen as less relevant within the dominant religious and societal value systems shaping this component.

3.3 Random Forest Features Importance

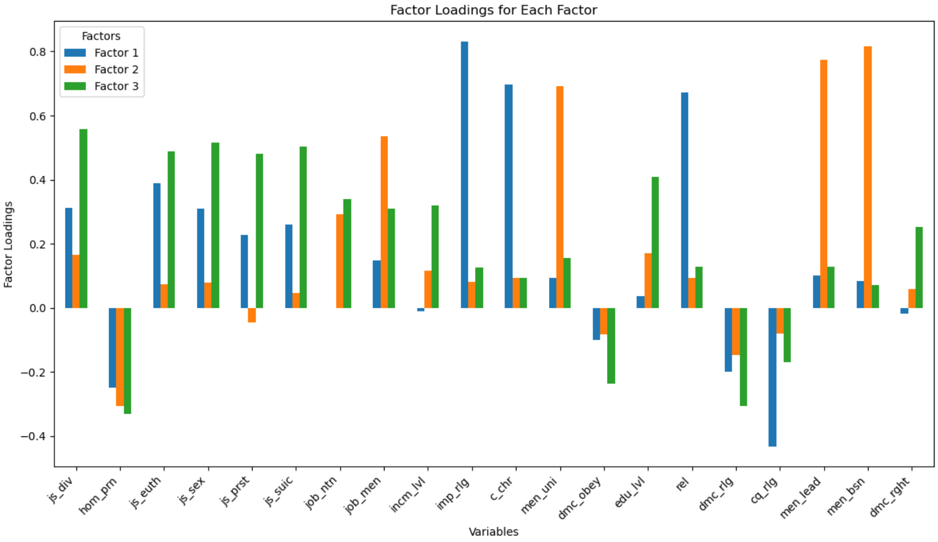
Here I started using the Random Forest alghoritm, using as target variable the variable "target" I created, which is just the sum of `dfEU['js_abr']`, `dfEU['js_hom']`, and `dfEU['imp_imm']`. I gathered the 20 highest variables for Mean Decrease Index for the feature importance. Doing this ensured me to understand what are the 20 most important features for this study. Top20 variable as follow:



Top 20 Feature Importance

3.4 Factor Analysis

I conducted an exploratory factor analysis (EFA) with the aim of identifying underlying patterns within the data, using only the first 20 variables found in Random Feature Importance. Given that the variables were not independent, I employed an oblimin rotation, which allows for factors to correlate and provides a more realistic representation of complex social data where relationships between variables are expected. After examining the eigenvalues, I chose to retain three factors, as these had eigenvalues noticeably greater than 1, indicating that they explain a significant proportion of variance in the data compared to other factors.



Exploratory Factor Analysis

Here I suggest an explanation of the 3 factors:

- Factor 1: Personal and Ethical Judgments
  This factor had strong positive loadings for variables like js_div (justifiable: divorce), js_euth (justifiable: euthanasia), and js_sex (justifiable: casual sex), indicating its focus on Individual Morality and Personal Judgments. Additionally, edu_lvl (education level) and job_ntn (jobs for nationals) loaded onto this factor, connecting education and employment views to personal ethics.

- Factor 2: Gender Roles and Leadership
  This factor captured attitudes toward traditional gender roles, with strong loadings from men_lead (men better political leaders) and men_bsn (men better business executives). It reflects support for traditional Gender Roles and Leadership.

- Factor 3: Religious and Moral Convictions
  High loadings from imp_rlg (importance of religion) and rel (religious person) show a focus on Religiosity and Spiritual Convictions. Notably, cq_rlg

(critical view of religion) had a negative loading, opposing critical religious perspectives.

4 Predictive Analysis

In this section, I conducted predictive analysis using various machine learning algorithms to predict the controversial respondents. These algorithms were selected for their capacity to handle classification tasks and analyze complex relationships within the data. The methods used include:

1. **Random Forest:** An ensemble learning method that constructs multiple decision trees during training. It combines the output of these trees (majority vote) to improve classification accuracy and reduce overfitting.

2. **K-Nearest Neighbors (KNN):** A simple, non-parametric algorithm that classifies a data point based on how closely it resembles the data points in its nearest neighbor set. It predicts the label by majority voting among the k-nearest points.

3. **Support Vector Machine (SVM) with Polynomial Kernel:** This SVM variation uses a polynomial function to transform the data, allowing the algorithm to find complex relationships between the input features and the target variable by fitting non-linear decision boundaries.

4. **SVM with Radial Basis Function (RBF) Kernel:** The RBF kernel allows the SVM to create a more flexible decision boundary by mapping data points into higher dimensions, making it effective for data that is not linearly separable.

5. **SVM with Linear Kernel:** This is the simplest form of SVM, where the data is assumed to be linearly separable. It constructs a straight-line decision boundary between classes.

6. **Gradient Boosting:** An ensemble technique that builds models sequentially, with each new model correcting the errors of the previous one. It combines the predictions of multiple weak learners (usually decision trees) to create a strong predictive model.

7. **Logistic Regression:** A generalized linear model used for binary classification. It predicts the probability that a given input belongs to one of two classes by applying a logistic function to a weighted sum of the input features.

To optimize the performance of each algorithm, I applied a grid search with cross-validation to tune the hyperparameters. This approach systematically evaluates combinations of hyperparameters to identify the best configuration for each algorithm, ensuring that the models are optimized for predictive accuracy before final evaluation.

To assess the performance of each algorithm, I employed the following evaluation metrics:

- **Area Under the ROC Curve (AUC):** AUC quantifies the ability of the model to distinguish between positive and negative classes, where a value closer to 1 indicates better performance.
- **Sensitivity (Recall):** Measures the proportion of actual positives correctly identified by the model, i.e., the true positive rate.
- **Specificity:** Refers to the proportion of actual negatives that are correctly identified, or the true negative rate.
- **Accuracy:** Represents the overall correctness of the model's predictions, calculated as the ratio of correctly predicted instances to the total number of instances.
- **Precision:** Defines the proportion of true positive predictions among all positive predictions, focusing on the accuracy of positive class identification.
- **F1 Score:** A harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. It is particularly useful in cases of imbalanced datasets.

By employing a diverse set of algorithms and evaluation metrics, I aimed to capture a comprehensive understanding of the predictive power of the features related to controversial issues. The use of these various metrics ensures a thorough assessment of each model's performance across different dimensions, enabling a more robust comparison and selection of the most effective model for this task.

5 Results

In this section, I present the results of the predictive analysis, where the machine learning algorithms were evaluated based on their ability to predict the controversial respondents. The performance of each model was assessed using several key metrics, including Area Under the ROC Curve (AUC), accuracy, sensitivity, specificity, precision, and F1 score. The table below summarizes the performance of all algorithms, highlighting the predictive power of each approach in capturing the complex relationships between the demographic features and the target variable for each different performance metric.

Method	AUC	Overall Accuracy	Sensitivity	Specificity	Precision	F1 Score
Random Forest	0.83	0.86	0.89	0.77	0.92	0.91
KNN	0.78	0.83	0.87	0.70	0.90	0.88
Linear SVM	0.82	0.85	0.87	0.77	0.92	0.90
RBF SVM	0.84	0.86	0.88	0.80	0.93	0.91
Polynomial SVM	0.84	0.85	0.86	0.81	0.93	0.90
Logistic Regression	0.84	0.85	0.86	0.82	0.94	0.90
Gradient Boosting	0.84	0.85	0.86	0.81	0.93	0.90

Table 1: Performance metrics of different methods

6 Conclusions

The analysis presented in this paper offers an exploration of how demographic factors influence societal attitudes toward controversial issues such as abortion, same-sex marriage, and immigration in major European countries. Using the rich dataset from the World Values Survey and employing statistical and machine learning techniques, I identified significant patterns and predictive relationships within the data.

The exploratory analysis provided multiple key insights into the underlying structure of societal attitudes. Latent Class Analysis (LCA) revealed two distinct subpopulations: one with socially conservative views, particularly opposing homosexuality, euthanasia, and divorce, and the other group being moderately progressive, more open to these issues yet still influenced by traditional values. This distinction highlights the deep divide in societal views across Europe.

Further insights were gained from Principal Component Analysis (PCA), which revealed that the primary drivers of societal attitudes relate to perceptions of justice, fairness, and equity, particularly around issues like the justifiability of homosexuality, divorce, and gender roles. Additionally, the influence of religious values and confidence in institutions such as churches played a significant role in shaping public opinion, as shown by the high loadings on components related to religious influence.

The Random Forest feature importance analysis highlighted that the justifiability of divorce, homosexuality, euthanasia, casual sex, prostitution, and suicide are the most significant predictors of controversial opinions. These findings emphasize that personal moral and ethical judgments strongly influence attitudes toward these contentious issues.

The Exploratory Factor Analysis (EFA) identified three key underlying factors: (1) Personal and Ethical Judgments, strongly linked to the justifiability of actions like divorce, euthanasia, and casual sex; (2) Gender Roles and Leadership, reflecting traditional beliefs about male leadership in politics and business; and (3) Religious and Moral Convictions, with high loadings on the importance of religion and moral perspectives, illustrating the deep influence of religious values on societal attitudes.

On the predictive side, Random Forest and Logistic Regression were identified as the best-performing algorithms for predicting controversial attitudes, achieving high levels of accuracy. Key features driving these predictions included the justifiability of divorce, homosexuality, euthanasia, casual sex, prostitution, and suicide. These findings highlight the importance of moral and ethical judgments in shaping societal views on contentious issues.

Understanding the demographic factors that influence public opinion offers valuable insights into how societal divisions can be addressed more effectively. This analysis has revealed the intricate interplay of moral, religious, and ethical influences that shape attitudes toward contentious issues, providing an understanding of the social landscape. Recognizing which demographic groups are more likely to hold certain views enables more targeted communication strategies and interventions.

References

Evs/wvs european values study and world values survey: Joint evs/wvs 2017-2021 dataset. 2017-2020. doi: DatasetVersion1.1.0,doi:10.14281/18241.11.

Filippo Nardi. Variable renames, 2024. URL https://github.com/filipponardi17/Controversial-Attitudes-Analysis/blob/main/Original_Data_%26_Codebook_Questionnaire/Survey_Questions_Table.pdf.

World Values Survey Association. Joint evs/wvs 2017-2022 dataset, 2024. URL https://www.worldvaluessurvey.org/WVSEVSjoint2017.jsp.