ML Final Project Report

Alvaro Guijarro May
226883@students.hertie-school.org

Niklas Pawelzik
221791@students.hertie-school.org

Justus v. Samson-Himmelstjerna
222918@students.hertie-school.org

Abstract

European women face different abortion regimes, depending on their country of residence. Some countries have a more permissive approach to this issue and offer a safer environment to the women in need, but unfortunately, this is not the case everywhere [1]. Due to that fact, women in many European countries leave their countries of residence and endure cross-border travels in order to obtain better conditions for an abortion in neighboring European countries. However, there is still little scientific understanding of these cross country abortion travels, starting with difficulties to even quantify the phenomenon.

The paper at hand aims to provide a pioneering approach on how to use machine learning models in order to estimate such abortion travels for the case of Germany. In the case of German States, distance to neighbouring countries, abortion policies in neighbouring countries, the fraction of the population with a migrant background, and the foreign population in the area could be some predictors to identify future pull factors for foreign abortion-seekers to come to Germany, for the procedure. In 2021, there were approximately 100,000 registered abortion procedures performed throughout the 16 federal states, with less than 1,000 registered to be from foreign women.

In the process of this research project, balanced and unbalanced Logistic Regression, Random Forrest, and Support Vector Machine classification models are fitted with a combined dataset in order to determine which one was the most accurate at predicting our variables of interest. Out of the three groups of examined variables, the models seem to suggest that the distance to other European countries has a higher incidence in the prediction of whether abortion was performed on a German or a foreign abortion-seeker in a German state. [1]

[1]Here's a link to Alvaro's GitHub account `https://github.com/Alvaroguijarro97/ML_Group_Project`, Justus's `https://github.com/jvsamson` and Niklas's `https://github.com/nikpaw`

1. Introduction

There is a lot to be learnt from the Polish abortion laws about morality policies in Europe. Alongside the Irish referendum on gay marriage, or French reforms regarding the country´s regulation of prostitution, it is a recent example of moral issues appearing on the European political agenda. Once these topics appear in the public eye, they are often discussed vigorously: What is at stake is the question of what society we want to live in; what is debated often concerns fundamental values. And even though the EU has been a key factor in understanding political processes in Europe during the last decades, regarding morality policies the Union has remained invisible for most of its existence and was largely ignored by the respective scholar community. This has started to change when scholars began to discuss e.g. the potential of such policies to develop a shared European identity, based on common values.

Despite the scholarly perceived lack of official EU action, several factors make the issue of Europeanized morality policies pressing, and at the core of the research project at hand, there is a functional argument to be made. Open Schengen borders and Europeanization of related policies (like the Directive on Patients' Rights in Cross-Border Healthcare) ease the access for European citizens to other morality policy regimes (and i.e., their medical ethical regulation frames): Cross-border patient movements are a fact of our time.

In turn, this means that *national capacities to effectively govern their population can be undermined by the access to more liberal regimes in fellow EU-countries*: If the Polish government further restricts access to abortion care, Polish women will seek this care in more permissive neighbouring countries. To quantify this effect however is a difficult task, given the absence of official statistics or border controls, and the desire of abortion-seeking women for discretion. The current project, therefore, proposes to approach in an indirect manner, based on the assumption that, *if a tighter abortion regime in country A causes women to cross borders to seek abortion care in country B, this should result*

in a higher number of abortions performed on foreigners in country B.

In line with this thought, we present a Logistic Regression Model for Binary Classification model, Random Forrest Classifier, and a Support Vector Machine Model that predicts labels for abortions regarding whether they were performed on foreigners or members of the national population. The labelling is based on state-level information about abortion policies in neighbouring countries, population share with a migrant background, and foreign population in the respective state. Given the feasibility limits in terms of time and data access, this project developed an initial framework based on German data on abortion statistics of the year 2021. The paper concludes with several promising avenues how to expand this approach in future research.

2. Related Work

Arguably the most influental definition [12] of Morality policies by Mooney establishes that "the important distinction between morality and nonmorality policy is that at least a significant minority of citizens has a fundamental, first-principled conflict with the values embodied in some aspect of a morality policy"[11]. Given this definition, it cannot come as a surprise that even among the divided literature on how to define morality policies[16], abortion and its regulation can be considered as a common ground: Abortion policies with their high moral stakes are morality policies.

Exploring morality policymaking in an Europeanized context contributes to a strand of research that is considered to be underrepresented within the respective literature[4] [3]. This is partly due to the fact that the EU historically focused on the common market and thus, for a long time did not get directly involved with morality policies – that are seen as often non-economic in nature [6] [14]. Another crucial limitation comes into play when it comes to studies regarding cross country abortion travels: Abortion seeking women are in general often afraid to be exposed to social stigma and seek discretion. This can be expected to be the case even more so when the respective abortion procedure is illegal in their country of origin. Furthermore, there are no Polish public records on how many citizens seek abortions abroad, nor are there official numbers on the German side on how many Polish women are taken in into German institutions that conduct abortions.

Nevertheless, there have been a number of studies on cross country abortion travels. A particuarly valuable contribution comes from the "Europe Abortion Access Project", a coordinated research group between the University of Barcelona and the European Research council. Reviewing the existing literature, the scholars conclude that the purpose of seeking abortion care abroad is a "phenomenon [that] remains poorly understood".[1] The researchers argue that the lack of scientific insight in this aspect of European abortion regimes is linked to a general lack of sufficient "quantitative and qualitative data" on the matter. In addition to these few, often ethnographic studies, there are only estimations by NGOs and practicioners. Our research aims to contribute to a closure of this gap by developing a data based model on abortion numbers from 2021 on the state-level within Germany. By aiming to contribute to the body of quantitative studies on cross country abortion travels, this paper therefore aims to model human migration. Given the complexity entailed in the subject of human migration patterns, studies can benefit greatly from the predictive power of computational models. To understand how an issue as intricate can benefit from using machine learning models, previous attemps to tackle it had to be evaluated first. In "A Machine Learning Approach to Modeling Human Migration" for instance, machine learning is used in order to predict human migration with several exogenous variables such as distance between countries, population of origin and destiny [13], etc. Here, the authors used machine learning models to outperform the traditional human mobility models by using "extreme" gradient boosting regression (XGBoost model) and an artificial neural network model (ANN model), which could be later used with hyperparameter tunning in order to obtain more precise predictions. Due to time and especially data limitations described in greater detail further below, these models were not implemented in our project. Instead, this paper focused on wrangling the nececssary data into a form that allow a sumber of basic models that can be further refined subsequently. These basic models showed - even on the very limited data sources at our hands - the potential of applications of classification models in the described context, that should be further explored.

3. Proposed Method

The research interest of abortions performed on foreigners allows for two different approaches. Either one conceptionalizes on a macro-level (either municipality, state, or federal state) a total amount of abortions performed within a unit of observation and focuses the model on share of abortions performed on foreigners. Or, alternatively, abortions are operationalized on a inidividual level in a binary manner as either "foreign" or "domestic", which opens avenues for classification models. Both approaches seem to be valid but have implications with regard to the data wrangling required in order to perform the model training. Given the state level data at our hands, both approaches would operate for the moment on the same (state) level input and the decision between them makes no difference with regard to the explanatory power of the final model. However, possible future access to data on indidivual abortions might be a promising way aspect to (drastically) increase the explanatory power of these approaches. In order to ease the possible

future incorporation of individual level data, we therefore decided to pursue the classification avenue. This decision has no downside for the current paper and only increases the options for follow-up projects.

The data available to us contains information on abortions per year per state. It distinguishes 'foreign' abortion-seekers (without further information on the distribution of countries of origin), which enables us to create a binary classification model between 'foreign' and 'local'. Independent variables can be used to enable a Logistic Regression Model (LRM) to estimate the probability that an event belongs to a specific class[9]. In LRM, a statistical approach is used for classification problems and is used when the dependent variable (target) is categorical[2]. The probability of an event being of the intended class (binomial in this case) is calculated by the fitted model, trained with a portion of the available data, and if the probability is under or over a set threshold, the model will assign the corresponding class.

Nevertheless, a Random Forrest Classifier (RFC) is also a powerful tool for this type of class prediction. With this model, many decision trees built with a random subset of features are analyzed and the most frequent prediction result is selected as the predicted class[5].

In the given context, both a Logistic Regression Model and a Random Forrest Classifier are therefore great fitting tools to approach to solve our research question: Can independent variables containing information on population composition and neighbouring countries be used by our model in order to predict labels of 'foreign' or 'domestic' to abortions in the states of Germany? The independent variables we have decided to evaluate for each German State with this classification models are the following:

1. *Proximity of neighboring countries (geographical distance)* / Restrictiveness of respective countries (using abortion atlas scores)

2. *Foreign population per state (share of overall population)* / Restrictiveness of respective countries of origin (using abortion atlas scores). Also interesting for network effects (foreigners going for their abortion to states where they have relatives/social network; medical and social professionals speaking the abortion seeker´s native language)

3. *Population with a migrant background per state.* / Restrictiveness of respective countries of origin (using abortion atlas scores). Interesting for network effects (foreigners going for their abortion to states where they have relatives/social network; medical and social professionals speaking the abortion seeker´s native language)

The binary dependent variable we are trying to predict is the type of abortion seeker (foreigner / german citizen) that have a procedure performed on them. In order to predict this variable, we developed the following Linear Regression Model:

$$ln(\frac{p}{1-p})) = \beta_0 + \sum_{i}^{26} \beta_{ia} X_{ia} + \sum_{i}^{62} \beta_{ib} X_{ib} + \sum_{i}^{62} \beta_{ic} X_{ic} + \epsilon$$

The *a* group of variables involves the fraction of population of every German State with a migrant background (13) and the abortion atlas policy index (13) of European countries. Only migrant groups from 13 European countries surpass the statistically significant size to be included in the micro census data, i.e. smaller communities with migrant background could not be inlcuded. The *b* group of variables involves the fraction of foreign population residing in each German State (31) and the abortion atlas policy index of the European country the group came from. The *c* group of variables involve the average distance betweet the German states' capital cities and the primary cities of each European country belonging to group *b* (31), as well as the abortion atlas policy index of of those countries (31). Every German State will have a different ranking for the information obtained for variable group *a*, *b*, and *c*, so taking this into consideration the order in which the variables appear for each group is sorted in ascending order, e.i: $X_{1a} < X_{2a} < X_{3a} < ... < X_{3n}$.

For the Random Forrest Classifier, the algorithm will select from the dataset random subsets of data, create a decision tree out of each one of them and predict a result. After all the predictions are made, the algorithm will choose the most frequent answer as the final result of the RFC.

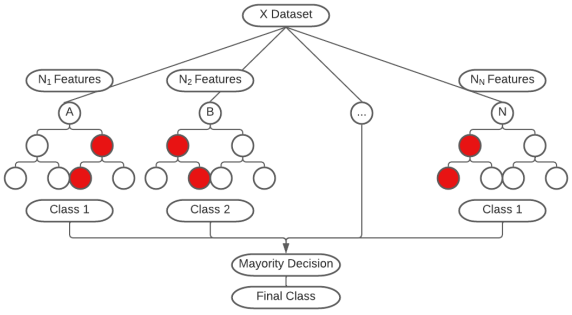

Figure 1. Random Forrest Explication

The bias caused by the imbalanced nature of our dataset can influence Random Forrest Classifier in such a way, that in the worst case scenario the minority class is ignored entirely. One approach to addressing the problem of class imbalance is to randomly resample the training dataset. The two main approaches to randomly resampling an imbalanced dataset are to delete examples from the majority class, called under sampling, and to duplicate examples

from the minority class, called oversampling. In order to address the imbalance in our dataset we decided to use the undersampling method and thereby reduce the number of examples in the majority class. Since we not only have enough examples in the minority class but already a considerable number of values in the majority class with close to no explanatory value. Therefore, it is not a limitation to use that this method might be deleting measurements that may be useful, important, or perhaps critical to fitting a robust decision boundary.

Another method that we used in order to modify our Random Forrest Classifier in such a way that it could handle the imbalanced classification is to change the weight that each class has when calculating the "impurity" score of a chosen split point. By doing so we ensured that the mixture of samples would be selected in favour of the minority class, allowing some false positives for the majority class. The BalancedRandomForestClassifier which we used can implement this directly from the imbalanced-learn library and performs random undersampling of the majority class in the meantime.

4. Experiments

Data: We developed our model based on data on abortions performed in Germany in 2021. In line with our independent variables described above, we also took into account data sets containing census data on the population with migrant backgrounds as well as foreign populations. In order to consider the geographical distance to neighbouring countries, we created a dataframe that contains information on the estimated distances between each German state and neighbouring European countries. Lastly, we included an Abortion Index Score that describes access to abortion treatment on a national level for European countries. Below, we provided a brief description of the five data sets we wrangled in order to ultimately create a sixth, combined dataframe. This dataframe was then used to train and test our classifiers.

Even though we invested a great share of our project time in successfully wrangling these data sets into a coherent data frame, the structure of the data used has **three** serious flaws that affect the explanatory power of our models considerably. Below, you, therefore, find a first brief discussion of these shortcomings and how we tackled their implications to the extent possible. We will reflect further on the insights we gained when working with these data sources in subsequent chapters, along with a number of suggestions on how to correct these issues in further research and to benefit from the promising foundations of this project.

1. **Abortion case numbers by the foreign and local population in the German States for the year 2021.** The Federal Statistical Office of Germany published a data set containing information on abortions performed annually at the federal level. This data is subdivided into each respective annual quarter. Furthermore, information on the origin of the patient is listed for the individual procedures, whereby a distinction is made between the various federal states and the category "foreigners". The dataset contains information for all 16 federal states with an annual number of about 100,000 abortions performed in Germany from 1996 up to 2021. We have wrangeld that data in such a way, that we are able to access for each federal state within Germany the fraction of foreign abortions in comparison to the abortions performed overall within Germany. [1]

2. **Abortion Index Score of European Countries.** This Index is an initiative by the European Parliamentary Forum for Sexual and Reproductive Rights (EPF) and International Planned Parenthood Federation European Network (IPPF EN). It is an in-depth analysis of abortion policies for the Year 20221 across Europe, which scores 53 European countries and territories in accordance with their respective legal frameworks in reference to their access to safe abortion care.

   The questions and structures for the Atlas were designed by a group of experts in sexual and reproductive health and rights. They came up with an overall score on a scale from 0-100 composed of four sections with several sub-categories: "access", "legal status of abortion care", "clinical care and service delivery", and "information and online information". Since this data set is only published as a PDF-Document, so we needed to extract the data points. Next, we filtered the data for rows and columns that held significance in the given context and created a dataframe that could be compared to our other data. [2]

3. **Foreign Population Distribution of German States.** The statistics on foreigners registered in Germany done by the Federal Statistical Office contain different representations of the regional distribution of the foreign population. Among other things, they list the federal states, sex, years of age and country groupings/citizenship. With the help of this dataset, we hope to identify driving factors for higher shares of abortions performed on foreigners, checking for variables such as the composition of foreign populations in comparison to the proximity to neighboring countries with stricter abortion policies. [3]

4. **German Population with migrant background in the German States.** The Federal Office for Migration and Refugees conceived in 2020 a census on a representation basis (Mikrozensus) to approximately iden-

tify the different migrant groups German citizens have as background, derived from the country of origin of their respondents or parents. For each of the federal states in Germany, this dataset contains the number of citizens that have a migrant background from one of the EU 27 countries, and only the most significant groups were kept for further research since some migrant population groups were so small that their share in the overall population of the state was negligible.[4]

5. **Geographical distance of German States to European Countries.** In order to calculate the distances between the 16 German States and the different countries in the European continent, we created a dataset with the information needed. This was done by collecting data on City Name, Longitude, Latitude, Country, State, Region, City Status, Population Density, Population, and Id of more than 4 million unique cities around de world. With this dataset, we calculated the average distance between the German State's Capitals and all of the cities available in the dataset from each one of the European countries, to sequential determine the respective distance between them.[5]

6. **Final combined dataframe.** In order to create the dataset needed to run the classification model, the transformed versions of foreign population, migrant background, distance to European countries, and abortion case numbers, all organized by the 16 federal states, were combined and cleaned. This was done by creating a naming convention that could be merged with every single dataset. We were then able to unpivot all dataframes from wide to long format, making each observation a state/country combination, instead of each observation a state with several countries as variables per single observation. We then created placeholder values within a new column "Country Score" in every dataset, which allowed us to insert for each observation the corresponding abortion score for the respective country. Then we sorted those in such a way that they would list for each state the individual foreign population, as well as the migrant population, replacing the placeholder country names with the value from the abortion atlas. In the end, we were able to combine all these datasets with the geographical data we had compiled for each federal state concerning the 31 surrounding countries. The resulting dataset contained a data point for each one of the abortion procedures performed in 2021 in the different German states, segregated by an identifier variable that differentiates between procedures performed on foreigners or locals.

Studying these datasets, a number of limitations become apparent. First and as expected, the distribution of abortions between "locals" and "foreigners" is uneven. Abortions performed in a country will under most cicrumstances be performed primarily its citizens, and Germany is no exception to that assumption:

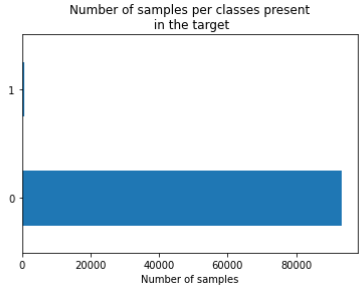

Figure 2. Dataset Distribution of Abortions performed on locals (0) and foreigners (1) in Germany

99% of the abortion procedures performed in German States were for domestic patients, so in order to be able to identify if we could explain the remaining 1%, a balancing of the dataset had to be performed. With the help of class weights, the skewed distribution of the classes for this dataset can be modifyed in the classification models. This way, the training models can penalize the misclassification made towards the minority class (foreign patient) by setting a higher class weight and simultanously reduce the weight for the majority class (domestic patient).

Secondly, all data sources used contain information exclusively on the state-level. This is true with regard to the dependent variable (abortions per state), and to all independent variables (migrant share, share of foreigners, distance to neighboring countries). While we perform a number data wrangling steps as described in the final combined dataframe section above to pivot our dataframe to have individual abortions per row, these observations contain information on state level factors. For each observation, we have a state-level based combination of explanatory variables, and an outcome of either 0 (abortion conducted on a domestic woman), or 1 (abortion conducted on a foreign woman). For the 16 German states and based on data from one year, this results in 32 different value combinations of independent and dependent variables distributed across 100.000 abortions. This small data set of distinct observations constitutes a de-facto reduction of value combinations and drastically reduces the ability of models to make any robust classficiation predictions.

With the data at our hands, there is no feasible way around this problem: the inclusion of every additional country would mean a multiplication of the data wrangling process for the nationally organized data sources of the respective country regarding population and abortion statistics; an inclusion of abortions numbers from prior years is not possible since the Abortion Atlas was created for the first time

in 2021. This makes the proposed *future* expansion to additional countries and subsequent years even more crucial: Including more entity-time combinations will add new value combinations (at the very least through changing Abortion Index Scores) that then increase the explanatory power of the models. As we show below, the first preliminary results of our models support the intuition of a relevant effect size especially of the explanatory geographic variable and the respective abortion score, subsequent research seems justified based on these findings.

[15]

Software :**GitHub**[7] was used as the source code management software, used to host git repositories, store differente copies and version, and allow collaborative work between team members. **Google Colab** [8] was used as the main python coding cloud app. It gave us access to runtimes of up to 12 hours of Google Compute Engines's GPU and TPU in order to run our python code.

Evaluation method: For the different machine learning models, we are using *Accuracy score*, *Confusion Matrix*, *ROC Curves*, and *F1 Score* to evaluate our results[10]. **Accuracy** is used in order to compute how many times the classifier was right (when predicted and actual value are the same) and divide it by the number of samples in the training set.

$$Accuracy = \frac{(TruePositive + TrueNegative)}{TotalSamples}$$

The **Confusion Matrix** is used in order to visualize the results of the prediction model. Here we can observe the True Positives, False Positives, False negatives, and True Negatives our model has produced with its current configuration.

The *receiver operating characteristic* (**ROC**) curve plots the true positive rate against the false positive rate (ratio of negative instances that are falsely classified as positive instances)[9]. This metric can help select a threshold for a classifier, looking to maximize true positives as well as minimize false positives.

Lastly, the **F1** score is used as a metric to evaluate the precision and recall of a prediction model. A good F1 score would mean that a model has low false positives and low false negatives.

$$F1Score = \frac{TruePositive}{TotalPositive + \frac{FalseNegative + FalsePositive}{2}}$$

With these metrics we will try to determine which of the machine learning models evaluated (Linear Regression Model Weighted or Unweighted / Random Forrest Weighted or Unweighted or Undersampled / Support Vector Machine) and their parameters is the most useful for our class prediction.

Experimental details: Due to the great data unbalance present in our dataset, initially we had to optmize the class weights based on the distribution of our Y variable, so that the model would give more importance to the least common class. The optimized values of these weights are calculated in the internal "class_weight" function in all of the evaluated models. Using parameter tunning, the optimal number for max_iter of 1500, which is the maximum number of iterations needed for the Logistic Regression Model to converge, was found between the values of (500, 1000, 1500, 2000). This value allows us to run the model maintaining the most efficient computational cost.

Since our dataset contains numeric variable values that are different in scale within our X values we decided to perform a standardization to have a common scale while building our machine learning model. Due to the fact that the most popular techniques for scaling numerical data prior to modeling is the StandardScaler we decided to start with that for our base line model. Because machine learning models learn from combining input variables to an output variable, differences in the scales across input variables may increase the difficulty of the problem being modeled.

After experiencing some issues with parameter and hyperparameter tuning for our Logistic Regression we decided to focuse on the Random Forest Classifier in order to further understand exactly what is being predicted and what it means in context. We once again used the Scikit-Learn library, which include for example the train-test-split function so that we could fit and evaluate the model on separate chunks of the dataset. From sklearn.metrics we used accuracy_score, confusion_matrix, roc_auc_score, and classification_report. Since our first unweighted Random Forest Model was only predicting class 0, and hence forth completely ignoring the minority class in favor of the majority class, we decided to introduce a class wight to our model. To then further improve our AUROC value we then used a Balanced Random Forest with Random Undersampling. However, after evaluating the Feature Importance based on feature permutation for our three Random Forest Models we realised that not only F1 scores were concerning but the features that were found to be most important did not bring any explanatory value in a real life scenario.

In order to better compare our results from Logistic Regression Model and the Random Forest Classifier, we decided to run a Support Vector Machine algorithm as well. Since SVMs are considered one of the most robust prediction methods for balanced classification and they can be easily tweeked with skewed class distributions to handle imbalanced datasets. These algorithms find a hyperplane

decision margin that best splits the measurments into two classes. We reported the performance of this model by calculating the mean area under the ROC curve, by gathering three repeats of 10-fold cross-validations. Furthermore for comparison we calculated the accuracy scores weighted and unweighted, as well as F1 scores and a confusion matrix.

Results: After running the different model versions, we obtained the following results:

Classification Model	Accuracy		AUROC	Compute Time (sec)
	Balance	Unbalance		
Logistic Regression (unweighted)	0.500	0.994	0.902	9.418
Logistic Regression (weighted)	0.845	0.861	0.902	21.41
Random Forrest (unweighted)	0.994		0.902	6.11
Random Forrest (weighted)	0.861		0.845	5.683
Random Forrest with Random Undersampling	0.861		0.904	2.116
Support Vector Machine	0.882		0.858	411.716

Figure 3. Results of Evaluated Models

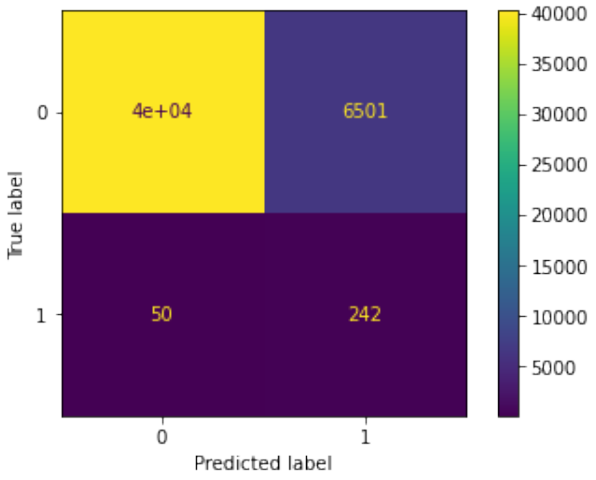

Figure 4. Confusion Matrix of Random Forrest Classifier (weighted)

All of the confusion matrixes look like the weighted Random Forrest Classifier model 4. Models that include computanionally generated weights assign both labels, and with more than 40.000 abortions correctly predicted as performed on Germans, more than 80% of the labels per class were attached correctly. Thanks to the weights attached to the classes, in a similar manner more than 80% of the almost 300 abortions performed on foreigners were correctly predicted. Given that all of our variable opreate exclusively on the state-level, we consider this explanatory capacity of 80% per class of our models a great success, especially compared to their unweighted counter parts.

Comment on your quantitative results. We were fully expecting that our unweighted models would not be able to predict the minority class, since we knew that our data was heavily skewed and therefore, unbalanced. At a first glance, however, all weighted and balanced models yielded better results than we expected since the accuracy score for all of our weighted models were in the high 80s, which were promising high explanatory values. However since accuracy is not a great measure of classifier performance when the classes are imbalanced, we also generated the Area under the Receiver Operating Characteristic (ROC) Curve, measuring the overall summary of diagnostic accuracy. Since the AUC equals 0.5 when the ROC curve corresponds to random chance and 1.0 for perfect accuracy, we were quite content at first with our results in the high 80s and low 90s. However, as soon as we look at our F1 Scores, as well as the feature importance of our Model, we realised that these values did measure the performance of our Models accurately.

The poor performance of our model on these metrics shows that the low number of German states involved in combination with abortion indices for only a single year lays too heavily on the explanatory power of our models. We anticipated that this problem, which is inherent in our available data, might constitute a challenge for our project. Our models provide a satisfying answer to the class imbalance; however, the problems in the data structure itself could neither be avoided nor solved by the current means at our hand. Given that we have had only state-level data from the 16 German states available, and that we could therefore only combine a total of 32 value combinations.

We, therefore, consider the current state of our project a partial success nevertheless, given that we detected the hidden performance issues, and were able to identify and explain the issue as outlined in the analysis below. This in turn allowed us to identify clear strategies for solutions and further expansions in more resourceful project settings in the conclusive remarks of our project paper.

5. Analysis

The particularities of our data discussed in earlier chapters are reflected in the results our models produce. Given the extreme imbalance of our data, unbalanced models behave essentially like dummy models: Aiming for high accuracy scores, they label every abortion as domestic without exception. This results in accuracy scores that are logically equal to the class distribution: less than 1% of our data represents abortions performed on foreigners, thus labelling every abortion as domestic is a successful strategy in more than 99% of the cases. In order to produce meaningful results, however, we have to focus on foreign abortions: Mislabelled German (domestically performed) abortions are less problematic than mislabeled foreign abortions.

Our balanced models have more explanatory power because they can provide precisely this added value in comparison to the unbalanced classifiers. Following this focus on labelling foreign abortions, our preliminary factor analysis becomes relevant: While keeping the flaws of our models in mind, the most important factors among the interaction terms are pointing in a clear direction: 13 of the 15 most important independent variables are distance-based, if one considers predictive power for abortions on foreigners. Even though these insights must be taken with caution, they are in line with the initial intuition: Proximity to neighboring countries in combination with the level of restrictiveness within these countries´ abortion regimes is a strong predictor for a high number of foreign abortions.

Going back to the paper´s starting assumption, this comes as no surprise: Berlin and Brandenburg, the closest states to Poland (a country with highly restrictive policies), have the highest number of foreign abortion seekers within Germany. States with big communities of Polish foreigners or Germans with Polish migrant backgrounds, however, such as North Rhine Westphalia, do not show higher numbers for abortions performed on foreigners, both in absolute and relative numbers. These two independent variables, people with migrant backgrounds and foreigners, seem to play only a minor role in explaining the cases of abortions performed on foreigners. While the data included needs to be expanded and diversified, there is a clear trend regarding what matters when it comes to explaining abortions performed on foreigners: proximity to neighboring countries and the degree of restrictiveness of their respective abortion regimes.

6. Conclusions

The project aimed to provide a classifier model that predicts labels for abortions in a binary fashion, as either being performed on German or non-German women. With models that include computationally generated weights, more than 80% of the labels per class were attached correctly. Given the exclusively state-level nature of our data, this explanatory capacity is a remarkable success. Preliminary factor importance analysis furthermore suggests that of the three considered explanatory variables, geographical proximity combined with abortion restrictiveness scores is the decisive factor for these predictions, and migrant background share and foreign population are of little relevance. The poor performance indicated for instance by the F1 scores results from the flaws of our data structure. Variables that exclusively operate on a state level capture only 32 value combinations constitut a clear limitation to the models´ success. However, this clear identification of the source of problems - the low number of distinct value combinations - allows for three concrete suggestions for refined future research projects: A first option would be to include data that operate on a lower level, ideally on the level of the individual abortion. The binary classifier logic we developed for this project would allow an easy incorporation of such information since the dataframe already lists one observation per abortion. However, such individual data on abortions is not publicly available, the only way to gain access to this type of information would therefore be through non-public channels.

A second option would be to include updated data on abortion numbers and abortion scores in upcoming years. Including abortion numbers from past years currently already available would not be sufficient for this: In order to generate new value combinations, the independent variables would need to change. Unfortunately, only the abortion numbers - so the binary dependent variable - are updated on an annual basis. Only subsequent editions of the first abortion atlas could provide new value combinations for our current independent variables populated by the German state.

The third and most promising approach would therefore be to expand the developed approach beyond the German states and include data from further European countries. Including state-level data from further countries would multiply the amount of value combination for each independent variable and improve our model drastically. Unfortunately, accessing and processing the required census and abortion data from additional European countries would certainly not have been feasible within this project, given the nationally specific data structure and the extremely time-consuming wrangling process. However, the first promising results of the present project might justify a subsequent project in order to further pursue the developed pioneering approach to quantify abortion migration across European borders by using classifier models.

7. Contributions

Since we had to build most of the dataset from scratch, each team member was in charge of wrangling and standardising a portion of the data. With the help of a pipeline used for the transformation a standardization of the data developed by Justus and Niklas, The Abortion Score index (Niklas), the foreign population per state (Justus), the geographical distance of states to European countries (Alvaro), and migrant population state (Niklas and Justus) datasets were built and eventually joined into a master dataset used for the production of the different prediction models.

Once that was done, Niklas wrote the Basic Model and a first Dummy Model, as well as the first Logistic Regression Model. Afterwards, Justus used that to implement a weighted Logistic Regression Model and calculate balanced and unbalanced accuracy. Together the two worked on the confusion Matrix, AUROC, Precision-Recall Curves, and plotting the Receiver Operating Characteristic (ROC)

Curve. Justus then created three Random Forest Classifiers, as well as the SVM Model, and evaluated those. Niklas and Justus then analyzed together what they had computed. They generated a Random Forest Classification Feature Importance, plotted the top 15 most important features and calculated feature importance for each class separately. The team wrote each paper together. The final written result was then formatted by Alvaro.

References

[1] Europe abortion access project, 2020.

[2] N. Arya. Logistic regression for classification, 2022.

[3] P. Ayoub and D. Paternotte. *LGBT activism and the making of Europe: A rainbow Europe?* Springer, 2014.

[4] P. M. Ayoub. Cooperative transnationalism in contemporary europe: Europeanization and political opportunities for lgbt mobilization in the european union. *European Political Science Review*, 5(2):279–310, 2013.

[5] I. C. Education. Random forrest, 2020.

[6] E.-M. Euchner and I. Engeli. Conflict over values in the european multilevel space: The case of morality issues. In *European Values*, pages 65–79. Routledge, 2018.

[7] GitHub, Inc. Github.

[8] Google Research. Colaboratory.

[9] A. Géron. *Hands-on Machine Learning with Scikit-Learn, Keras Tensor Flow*, chapter 14. O'Reilly Media, Inc., 2019.

[10] J. Jordan. Evaluating a machine learning model., July 2017.

[11] C. Z. Mooney. The politics of morality policy: Symposium editor's introduction. *Policy Studies Journal*, 27(4):675–680, 1999.

[12] J. M. Permoser. What are morality policies? the politics of values in a post-secular world. *Political Studies Review*, 17(3):310–325, 2019.

[13] C. Robinson and B. Dilkina. A machine learning approach to modeling human migration, June 2018.

[14] M. A. Schwartz and R. Tatalovich. The rise and fall of moral conflicts in the united states and canada, 2019.

[15] K. Singh. How to improve class imbalance using class weights in machine learning, October 2020.

[16] D. T. Studlar and G. J. Burns. Toward the permissive society? morality policy agendas and policy directions in western democracies. *Policy Sciences*, 48(3):273–291, 2015.

Notes

[1] https://www-genesis.destatis.de/genesis//online?operation=table&code=23311-0006&bypass=true&levelindex=1&levelid=1664961110565#abreadcrumb

[2] https://www.epfweb.org/node/857

[3] https://www-genesis.destatis.de/genesis/online#astructure

[4] https://www.bamf.de/DE/Themen/Forschung/Veroeffentlichungen/Migrationsbericht2020/PersonenMigrationshintergrund/personenmigrationshintergrund-node.html

[5] https://simplemaps.com/data/world-cities