Lab Course "Ethical AI": Explaining Discrimination with Data from the European Social Survey

Leo Holzhauer and **Annalena Kofler** and **Christopher Sendlinger**
firstname.lastname@tum.de
Technical University Munich

Abstract

Discrimination and its versatile forms have sparked a significant public interest in the recent years. As various social factors influence the subjective feeling of being discriminated against, the question arises whether discrimination can be predicted using Machine Learning approaches. Our work for this project is based on the data set of the European Social survey conducted in 2018-2020 and is consisting of 49,000 participants. After establishing a data preparation pipeline, we utilized unsupervised methods to find general patterns in the data which turned out to be not related to discrimination. To predict discrimination directly, we compared different supervised models like decision trees, random forests with boosting, Support Vector Machines, logistic regression, and neural networks. Multiple approaches for feature sub-selection and adjusting the imbalanced classes were considered. XGBoost with all features and weighted classes performed best on the F1 score. It was possible to find and explain contributing variables to discrimination and understand false predictions with SHAP and cosine similarity.

1 Introduction

With the recent rapid advancements in applications for Artificial Intelligence (AI), concerns for the ethical use of these applications are growing simultaneously. Security features to ensure privacy and fairness of AI applications as well as AI applications in a technical field with direct ethical considerations, such as fake image detection, are being researched and developed. However, little emphasis has been put on explaining predictions and on the application area of social sciences (Grimmer et al., 2021).

In the social sciences, data is primarily used to model and understand social structures (Grimmer et al., 2021). This way, non-beneficial conditions, and processes become visible, which have the potential to cause social grievances or distress and suffering for the individuals involved. An improved understanding of the influential factors enables the development of efficient intervention responses and therefore offers a promising use case for a way to apply AI in an ethical relevant and significant way. An example for a non-beneficial condition is the social exclusion based on unemployment, which was modeled with various Machine Learning (ML) methods from (Serrano et al., 2019). With the developed application social workers in Northern Spain were supported in an efficient use of their limited temporal capacities to search for socially excluded individuals.

In this work, a similar approach was used to model the social structures involved in another condition which implies individual suffering: the self-reported experienced discrimination of individuals. After an initial data investigation and preparation, unsupervised as well as supervised ML methods were used to find factors most influential for the perceived feeling of being discriminated against.

However, it has to be considered that feeling discriminated against is a subjective opinion that depends on a multitude of different social factors. Firstly, the participant has to be informed and educated about the meaning of discrimination and its different forms. Secondly, the person has to be able to understand and conclude whether they are discriminated against. Thirdly, the participant has to be willing to share this sensitive and personal information in a face-to-face conversation with the interviewer. As a result, the possibility of answers not reflecting the true feeling of the respondent to this question has to be taken into account during evaluation. Therefore, an analysis to better explain model performance deviations and similar data points with different target values was done.

2 Data Investigation

2.1 Data Set

This project is based on the data collected in round 9 of the European Social Survey (ESS) which is an academically-driven project conducted every two years in 38 countries up to date. The general goal of the survey is the monitoring and evaluation of public attitudes and values. The chosen data set (ESS9-2018 (European Social Survey European Research Infrastructure , ESSERIC)) was acquired in the years 2018-2020 and contains data from at that time 25 European and six non-European countries. The full list of countries can be found in Appendix A.1. The participants were selected via random stratified sampling methods. To

ensure a high response rate, the participants were called or visited up to eight times and participation benefits such as vouchers were offered in some countries. The conducted face-to-face interview had a duration of approximately one hour and contained 512 questions in different categories. The names of the categories and illustrative examples for questions can be found in Table 1 in the appendix (, ESSERIC).

Since a high dimensional feature space with a limited number of data points can propose challenges for a ML algorithm, we decided to manually reduce the number of questions from 529 to 113 with a focus on potential indicators for discrimination. A clear choice for a feature to include was for example "Belong to minority ethnic group", while a clear choice to leave it out was "Interview length in minutes". If a question received more invalid than valid answers, it was not included in further analysis. The answers to the survey questions are based on different scales such as categorical or numerical, dependent on the type of the question. For example questions like "How religious are you?" could be answered with an integer value from 0 ("Not religious at all") to 10 ("Very religious"), while a question asking for the annual net pay could be answered with any numerical value. For each question there were the options "Refusal", "Don't know", and "No answer" that were encoded as outlier values in the data set. As an example, outliers were encoded with 77, 88, 99 for the question "How religious are you?".

The target variable we tried to learn and predict is formulated as "Would you describe yourself as being a member of a group that is discriminated against in this country?". If this question was answered with "Yes", it was possible to select the specific form of discrimination. Since the subcategories contained less than 1000 data points each (see Appendix A.4), we decided to focus on the binary variable of discrimination. A total of 3,593 participants referred to themselves as being discriminated against, while 43,062 people do not feel discriminated. An illustration of this distribution can be found in Figure 10. The fact that the data distribution is imbalanced has to be taken into account in the data preparation for the ML approaches. When the number of discriminated people is normalized by the number of participants per country, it can be observed that some countries have a high percentage of people feeling discriminated (c.f. Figure 1). These countries are Iceland, the United Kingdom, France, and Montenegro which can be explained with the dominating subcategories in the specific country (see Appendix A.4).

2.2 Correlation of Variables

A variant to determine linear dependencies between feature x and target y is the Pearson Correlation coefficient (PCC). A value of $+1$ corresponds to a strong linear proportionality between x and y, whereas a value of -1 corresponds to a strong inverse linear proportionality. With a value of 0 it can be said with certainty that there

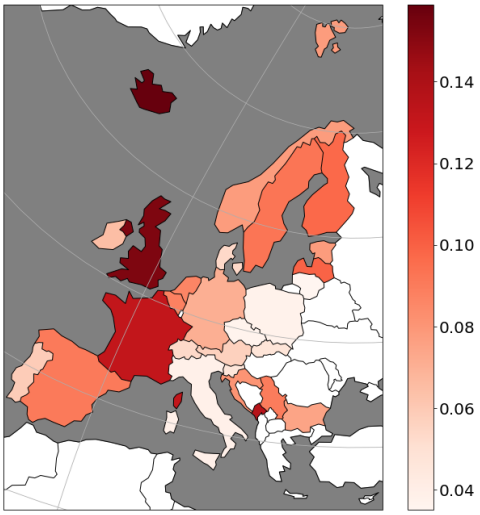


Figure 1: Percentage of participants per country that describe themselves as being discriminated against.

is no linear relation between x and y.

We calculated the PCC for all preselected features. The overall tendency is that the linear correlations within the data set are low, which is described in detail in Appendix A.3. In particular, the correlations with all other features and the target variable "Member of group discriminated against in this country" are below an absolute value of 0.25, which can be considered a weak linear correlation. Nevertheless, when sorted by magnitude, the resulting order of features leaves room for reasonable conclusions, as can be observed in Figure 2.

Figure 2: Pearson Correlation coefficients for "Member of group discriminated against in this country" (strongest 15, sorted by magnitude).

Put into perspective, the relatively high correlations of the features "Trust in the police", "Hampered in daily activities by illness/disability/infirmity/mental problem" as well as "Belong to minority ethnic group in country" w.r.t. the target are likely to mirror an existing correlation observed by individuals experiencing discrimina-

tion because of their ethnicity or disability. However, this mere quantitative analysis alone is not enough but rather intends to serve as a guideline for setting priorities for a further qualitative investigation based on expert knowledge from the field of the social sciences. In this work, we used the results of the correlation analysis as a selection criterion to reduce the number of inputs for the supervised ML models. The resulting subset was part of an additional evaluation dimension, further described in Subsection 3.2.

3 Data Preparation

A general overview of the necessary steps for the preparation of the data is highlighted with blue color in Figure 3. This process contains preprocessing steps followed by different options for the selection of features. These options were evaluated separately to determine whether a subset of the questions resulted in an improved performance of the ML algorithm. For supervised learning, the data was split into a training and test set. On the training set, we experimented with different sampling techniques with the goal of increasing the performance.

3.1 Preprocessing

In order to prepare the data for the different ML procedures, it was necessary to encode outliers, exclude NaN values, and represent the unordered, categorical variable "country" using one-hot encoding. This way, the number of data points was reduced from 49519 to 46655 and the number of features increased from 113 to 139. For more detailed information, the interested reader may refer to Appendix A.5.

3.2 Feature Selection

We integrated feature selection into the data preparation pipeline to evaluate the influence of different feature sets on the performance of the ML models, this way aiming on drawing further conclusions on the interdependencies within the data set. A better model performance based on a specific feature set may indicate higher correlations with the target of the features within this set, therefore enabling a deeper understanding of the importance of factors influencing the perception of discrimination. From the already manually selected and preprocessed 139 features, we extracted three additional subsets of these features. Firstly, the 30 features having the strongest linear correlation with the target according to the PCC values. Secondly, a subset based on iterative queries to the GPT-3 davinci model of OpenAI (Brown et al., 2020). To generate this subset, we stated a prompt to the model containing all of the feature definitions and the question to extract the ten most influential ones w.r.t. the target. This procedure was repeated ten times, the results were listed and sorted in descending order according to their number of appearances in the answers. This way, another subset with 30 features was generated. This subset had also the purpose to test the assumption of whether or not there is already implicit knowledge about the influential factors on discrimination in the training data of GPT-3. Thirdly, all of the features within the ESS category "Human values" were used to form the last subset.

3.3 Splitting

The data set was split into a training and test set in a ratio of 90/10. From the training data set, a validation data set was separated for hyperparameter optimization when needed. The remaining training data was used to learn the model parameters.

3.4 Sampling

As imbalanced distributions of the target variable are a well-known problem in the ML community (Chawla et al., 2002; He et al., 2008), we expected challenges in the supervised learning approaches. To mitigate such problems, we investigated the performance of different sampling methods. As a baseline, we trained the different ML methods using the original data set which will be called "no sampling" in the following. Additionally, over- and undersampling can be utilized to engineer a data set with balanced class labels. In undersampling, the number of data points in the majority class is reduced randomly until it is equal to the minority class. In oversampling, new data instances of the minority class are generated artificially via discrete interpolation between minority data points until the classes are balanced. Here, we used the method SMOTEN (Chawla et al., 2002). Furthermore, it is possible to assign weights to each data point that inform the model of the importance of a specific data point. The weights were chosen as the ratio of elements in the minority over the majority class. In the following, we refer to this method as "weighting".

4 Unsupervised Learning

4.1 Principal Component Analysis (PCA)

Including the one-hot encoded country information, the data is located in a 139-dimensional feature space. To reduce the number of features, one can use methods like Principal Component Analysis (PCA). For visualization purposes, a target dimension of $k = 2$ is common. Only PCC as feature selection gave visual distinct structures, pictured in Figure 4. The other selections formed uninformative big clouds of all data samples. Principle component analysis leads to the fact that the y-component is mostly defined by three features (mother/father/self born in the country), which can be interpreted as the heritage of a person. Red data points correspond to participants where neither they, nor their parents are born in this country, while dark blue values represent participants where they and their parents are born in this country. Unfortunately, the dimensional reduction does not carry any information about our target variable which can be seen in Figure 4. Nevertheless, this shows that there are underlying structures in our data that can be used for our classification task.

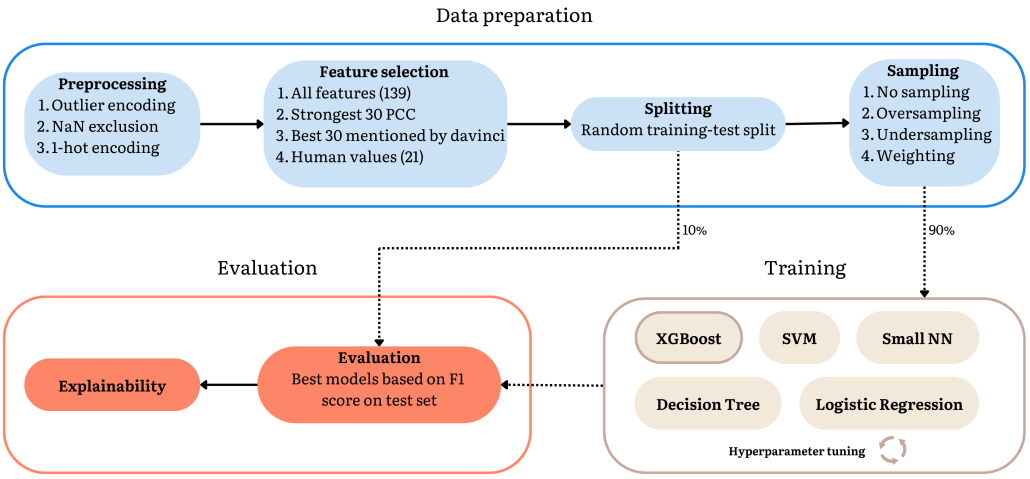


Figure 3: Data pipeline for supervised approaches.

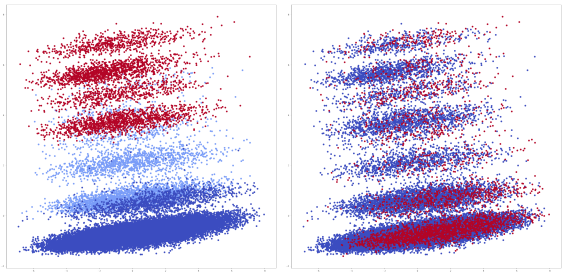


Figure 4: PCA (k=2), coloring based on persons heritage (left) / target variable (right)

4.2 K-Means Clustering

Building upon the results of dimensionality reduction, the next step was to identify clusters in the data. The simplest method to use here is K-Means, which builds K groups of similar data points. Our goal was again to differentiate between the participants based on the different answers they gave to the target question. The results on the reduced data using PCA ($k = 2$) however were not promising, which can be seen in Figure 4. No clear distinctions between the different answers to the target question can be made.

4.3 Hierarchical Clustering

Hierarchical clustering builds up clusters based on the similarity between different objects in the data set. Unlike K-means, it can provide insight into the composition of clusters, which may be helpful for our task. Also, this method can not find clusters that separate our target, reasons for that are described in the following subsection.

4.4 Conclusion

The tested unsupervised methods are not promising for our data and the question we want to answer, which has multiple reasons. The first problem was that the high dimensionality lies within the nature of our data set. To address this, standard approaches like PCA were used to reduce the dimensions with low variance and give the clustering methods a better chance to find structures within. On our data, this led to point clouds where both target classes are indistinguishable and mixed up. Arising from this is the second problem, the used clustering methods can not identify meaningful clusters for the given question. This problem is again rooted in the data since the difference between a discriminated and a not discriminated participant often lies in very small details and not fundamentally different beliefs. This fine contrast can not be picked up by the used methods.

5 Supervised Methods

Although no general structure was found in the data, supervised learning techniques can be employed to predict the feeling of discrimination of individual participants. Different supervised algorithms have proven successful in the context of predicting environmental attitudes and beliefs based on ESS8 data (Yektansani and Azizi, 2021). These models are: random forest, logistic regression, and neural networks. In addition to these algorithms, we experimented with decision trees and Support Vector Machines (SVM). For the random forest with boosting, we specifically used the XGBoost classifier.

The F1 score and the weighted F1 score were investigated as evaluation metrics for the model performance. The weighted F1 score calculates the mean of the F1 scores of the individual classes while considering the actual occurrences in the data set. Therefore, it is specifically designed for imbalanced data sets and shows good overall values for all investigated models (c.f. Figure 6a). However, we observed an undesirably large value of False Positive instances in the confusion matrices. The reason for this is that the weighted F1 score assigns greater contribution to the class with more examples in the data set. In our case, we observed a larger number of discriminated people being labeled as not discriminated. Since our focus lied on the accurate prediction of truly discriminated people, we decided to use the normal F1

score to prevent a large influence of not discriminated participants. The unweighted F1 score has the best trade-off between correctly identifying the most discriminated people and keeping the wrong identifications of not discriminated participants moderately low. The best hyperparameters for each model were optimized using the F1 score.

5.1 Decision Tree

Providing the least complex ML architecture from the mentioned methods, decision trees are the most promising when it comes to explainability and inferring knowledge about the dependencies of model input and output. An example to exploit the explaining qualities of the decision trees can be found in Appendix A.6. The used hyperparameters are: min_samples_leaf:0.001, max_depth: $\{3, 99\}$. The best performing model evaluated uses feature set "dvc" and method "wgh".

5.2 XGBClassifier

XGBoost (Chen and Guestrin, 2016) implements ML algorithms using gradient boosting. The XGBClassifier combines the predictive power of decision trees with gradient boosting, which usually outperforms random forests. The used hyperparameters are: Booster: gbtree, max_depth:10, learning_rate:0.1, n_estimators:125. The best performing model evaluated uses feature set "all" and method "wgh".

5.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) try to find a decision boundary between the different classes of the data set. The used hyperparameters are: loss: hinge, penalty: elasticnet. The best performing model evaluated uses feature set "all" and method "wgh".

5.4 Logistic Regression

While SVMs try to maximize the margin among class variables, logistic regression aims at maximizing the conditional likelihood of the data. Although the approaches are similar, one might outperform the other dependent on the data. The optimized hyperparameters for logistic regression are the solver (newton-cg, lbfgs, sag, saga), the regularization type (none, L_1, L_2) with the inverse regularization strength C, as well as the tolerance for stopping. The best performing model used the solver saga, L_2 regularization with $C = 0.1$, and a tolerance of 0.001.

5.5 Neural Network (NN)

Investigations regarding the application of deep neural networks (NNs) to social science data have shown that small NNs with a few hidden layers and a limited number of neurons per layer perform best (Serrano and Bajo, 2019). Similarly to this research, we chose small architectures for hyperparameter tuning which included networks with one hidden layer and 20, 50, or 100 neurons as well as two or three hidden layers where the number of neurons depended on the number of inputs $(N_{\text{in}}, 0.5 \cdot N_{\text{in}}), (N_{\text{in}}, 0.5 \cdot N_{\text{in}}, 0.2 \cdot N_{\text{in}})$. Furthermore, we included the activation functions ReLU, tanh, and Sigmoid in the hyperparameter search. The following hyperparameters were fixed due to computational limitations: optimizer Adam, a constant learning rate of 0.001, batch size of 200, and a stopping tolerance of 0.001. The best performing model has ReLU as the activation function and a single hidden layer with 20 neurons. Since the MLPClassifier class of sklearn does not support weighting, those values are set to zero in Figures 6b and 6a.

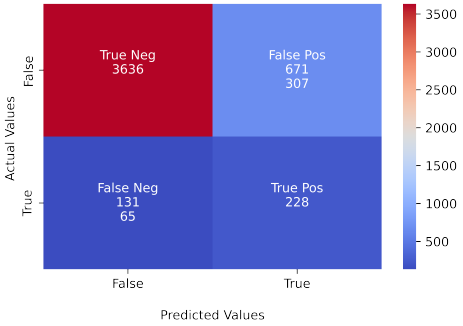


Figure 5: Confusion matrix for the best model. Total numbers for True/False Positive/Negative and number of explainable False Positive/Negative in the test set

5.6 Best Model

The overall best model regarding the unweighted F1 score, is the XGBClassifier using the feature set "all" and weighting of the classes (c.f. Figure 5) . The unweighted F1 score of 0.362 and a precision of 0.254 make it the most predictive model that was trained. Considering the highly unbalanced data set and the inherently difficult target, this result is good.

6 Explainability

To understand the predictive power and decision-making process of the model, different analyzing methods are used and combined.

6.1 SHAP

Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) is a game theoretic approach to explaining the output of ML models. It can give insight into the impact of different input features on the output classification of the model. Figure 7 shows a summary plot for our best XGBClassifier; positive SHAP values are contributing towards classifying a person as discriminated while negative have the opposite effect. The coloring indicates what values of the feature generate the given SHAP value. The feature "Belong to minority group" (blgetmg) with the answer "No" (Value 2, Red) counteracts discrimination, while answer "Yes" (Value 1, Blue) is a strong indicator for it. Every feature can be ana-

	Decision Tree				XGBoost				SVM				Logistic Regression				Small NN			
	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn
non	0.892	0.892	0.893	0.886	0.904	0.897	0.898	0.886	0.887	0.886	0.886	0.886	0.903	0.893	0.895	0.886	0.906	0.888	0.896	0.886
ove	0.797	0.748	0.731	0.670	0.898	0.846	0.874	0.774	0.891	0.772	0.834	0.647	0.897	0.801	0.838	0.703	0.899	0.825	0.860	0.744
und	0.829	0.829	0.863	0.733	0.816	0.792	0.792	0.690	0.764	0.789	0.824	0.709	0.817	0.814	0.798	0.693	0.764	0.804	0.789	0.700
wgh	0.781	0.781	0.854	0.733	0.859	0.833	0.834	0.768	0.817	0.821	0.801	0.556	0.811	0.814	0.801	0.696	0.000	0.000	0.000	0.000

(a) Weighted F1 scores

	Decision Tree				XGBoost				SVM				Logistic Regression				Small NN			
	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn	all	pcc	dvc	hmn
non	0.098	0.098	0.103	0.000	0.209	0.139	0.156	0.000	0.011	0.000	0.000	0.000	0.202	0.078	0.114	0.000	0.288	0.022	0.124	0.000
ove	0.175	0.210	0.190	0.145	0.271	0.262	0.247	0.120	0.225	0.246	0.221	0.154	0.279	0.261	0.238	0.156	0.264	0.210	0.227	0.127
und	0.278	0.278	0.279	0.160	0.317	0.280	0.269	0.160	0.272	0.251	0.272	0.153	0.323	0.264	0.275	0.159	0.257	0.261	0.266	0.157
wgh	0.257	0.257	0.282	0.160	0.362	0.302	0.300	0.145	0.308	0.267	0.257	0.155	0.292	0.266	0.279	0.161	0.000	0.000	0.000	0.000

(b) F1 scores

Figure 6: Overview of metrices for all supervised methods (decision tree, XGBoost, SVM, logistic regression, small NN), dependent on the chosen feature selection method (columns) and sampling method (rows). The abbreviations can be found in Table 2.

lyzed in this way giving a good overview of high-risk characteristics.

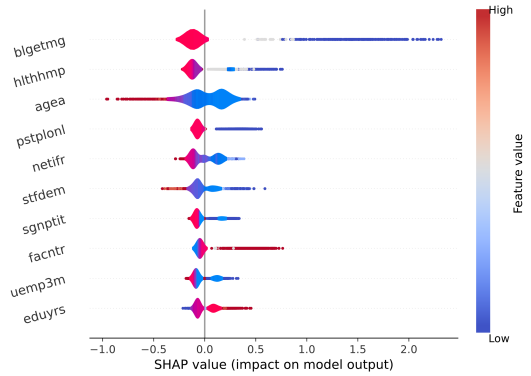


Figure 7: SHAP summary plot: Positive SHAP values are contributing towards classifying a person as discriminated while negative ones have the opposite effect. The coloring indicates what values of the feature generate the given SHAP value. For abbreviations see Table 2

6.2 Cosine Similarity

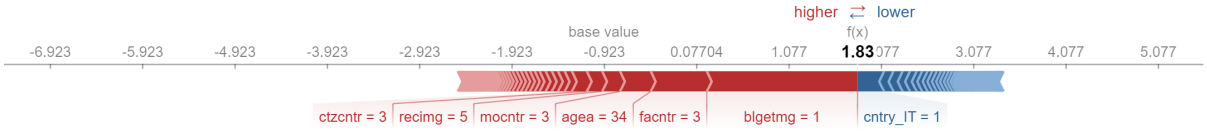
Cosine similarity allows us to calculate a value in the range $[-1, 1]$ that describes how similar two vectors are, with the maximum being identical to one another and the minimum being the exact opposite of each other. Analyzing the predictive power of our model using this method yields an explanation of why the model falsely classifies some samples. If we can find True Positives/Negatives for our wrongly classified False Positives/Negatives with a high similarity score, we can infer that our model made that mistake based on the subjectivity of the target variable. Using a threshold of 0.9 for the described similarity score, 65 of 131 participants, which answered that they feel discriminated against but our model predicted that they are not discriminated against (False Negative), can be explained by other similar persons that are True Negatives. 307 of 671 False Positives can be explained by an equivalent in the True Positives. (Figure 5) Again, this shows the underlying problem for the target variable, discrimination, is a very subjective topic. Persons that are very similar regarding the survey answers can have very different feelings and understandings of discrimination.

6.3 Selected Example

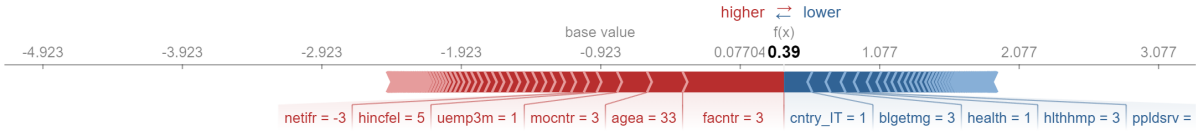
Using the explained similarity method and the SHAP values False Positives/Negatives become explainable. To illustrate this, a False Positive is compared with its corresponding similar True Positive (>0.999). Figure 8a shows person #2843 (TP) and Figure 8b shows person #1830 (FP). Most descriptive similarities: Both are living in Italy; neither they nor their parents were born there; both are in their early thirties. Their only crucial difference is that person #2843 belongs to a minority group and the other has not answered this question. Because of these similarities in the most important features regarding discrimination, the model misclassified the second person.

7 Conclusion and Outlook

The explainability results allow us to understand why the unsupervised methods failed on our classification task. Two very similar persons can answer differently on whether or not they feel discriminated against. The differences lie within small deviations and not fundamentally different lifestyles or circumstances, which could be identified using the survey data.



(a) Person #2843 classified as True Positive.



(b) Person #1830 classified as False Positive.

Figure 8: Comparison of the most influential SHAP values and their impact on the prediction of a True Positive and a False Positive.

Overall, no relevant clusters were found in unsupervised methods based on discrimination. However, statistical analysis and explainability of supervised methods yielded promising and interpretable results. Based on the findings, certain features are very likely to be good indicators for discrimination, e.g. "Belong to minority ethnic group" or "Hampered in daily activities by illness/disability/infirmity/mental problem".

Limitations were encountered during the elaboration of this work. On the one hand, discrimination is overall challenging to predict due to similar answers of participants, the imbalanced data set, and a used average over various heterogeneous European countries. On the other hand, discrimination is inherently a difficult target. Since it is a self-reported value, knowledge about discrimination is required to access it as well as a sensitization. Furthermore, it is influenced by many factors, e.g. potentially different definitions of discrimination dependent on the country or the individuals social environment.

Further qualitative investigation incorporating expert knowledge from the social sciences is required to evaluate the validity of the correlations and similarities found. Additionally, a differentiation w.r.t. the multitude of reasons for discrimination might be promising for improved insights into the model dependencies. However, these improvements are currently limited by the small amount of positive target values and only feasible if the amount of data points per ESS cohort will improve significantly in the future.

Acknowledgments

We would like to thank Tobias Eder for his supervision, advice, and ideas regarding our project. The bi-weekly discussion sessions were very helpful, interesting, and a lot of fun.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

N. V. Chawla, K.W. Bowyer, L.O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

European Social Survey European Research Infrastructure (ESSERIC). 2021. Ess9 data documentation. *Sikt - Norwegian Agency for Shared Services in Education and Research*.

. European Social Survey European Research Infrastructure (ESSERIC). 2021. Ess9 - integrated file, edition 3.1 [data set]. *Sikt - Norwegian Agency for Shared Services in Education and Research*.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2021. Machine learning for social science: An agnostic approach. *Annu. Rev. Polit. Sci.*, 24(1):395–419.

H. He, Y. Bai, and E.A. Garcia. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of International Joint Conference on Neural Networks*, pages 1322–1328.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Emilio Serrano and Javier Bajo. 2019. Deep neural network architectures for social services diagnosis in smart cities. *Future Generation Computer Systems*, 100:122–131.

Emilio Serrano, Mari Carmen Suárez-Figueroa, Jacinto González-Pachón, and Asunción Gómez-Pérez. 2019. Toward proactive social inclusion powered by machine learning. *Knowledge and Information Systems*, 58(3):651–667.

Kiana Yektansani and Seyed Soroosh Azizi. 2021. Using machine learning to predict consumers' environmental attitudes and beliefs. techreport 313902, Agricultural and Applied Economics Association.

A Appendix

A.1 Participating Countries

The following at that time European countries participated in the ESS9-2018 survey: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, United Kingdom. Additionally, the survey was conducted in six non-European countries: Albania, Iceland, Montenegro, Norway, Serbia, Switzerland.

A.2 Categories of The Data Set

Table 1 lists the categories chosen for the ESS9 survey with explicit example questions for each category.

A.3 Pearson Correlation Coefficients

In general, the overall correlations within the data are very low, which can be observed in the heat map shown in Figure 9.

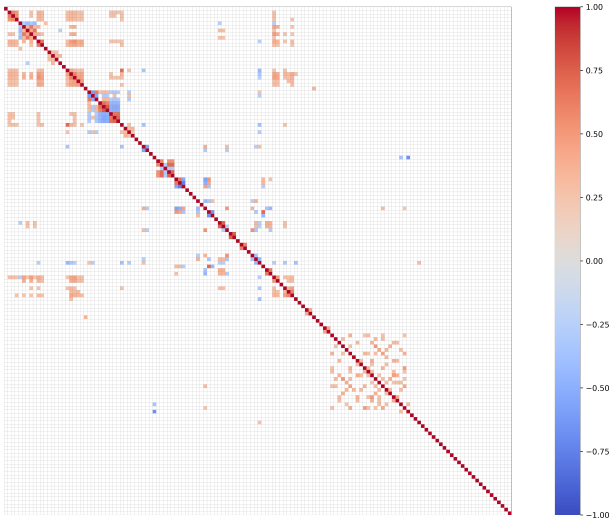


Figure 9: Pearson Correlation coefficients for all preselected features based on the a whole data set (values within [0.25, -0.25] are set to 0 for visibility reasons).

One possible explanation for this is the method of data collection. A social survey most likely will be designed to cover various aspects efficiently, therefore not having many questions with correlations in their answers. However, from the correlation analysis, some clusters with dependencies could be determined. An example is the light red cluster on the bottom right, representing questions from the section "Human values".

A.4 Target Variable

The distribution of people (not) feeling discriminated against can be found in Figure 10.

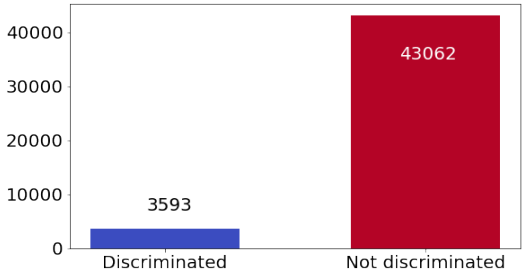


Figure 10: Distribution of participants feeling discriminated against in their country.

If the question of feeling discriminated against was answered with "Yes", it was possible to select the specific form of discrimination from the following set of potential answers: color/race, nationality, religion, language, ethnic group, age, sexuality, disability, or other. The overview of the chosen subcategories can be found

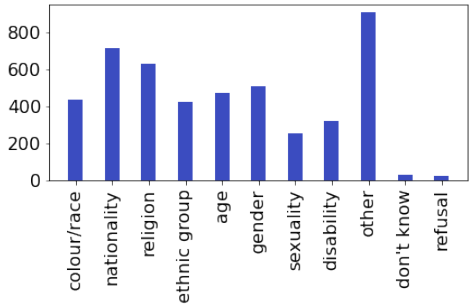


Figure 11: Distribution of participants feeling discriminated against dependent on the reason.

in Figure 11. Since the subcategories of discrimination only contain less than 1000 data points each, it was not feasible to learn the type of discrimination specifically. Therefore, we focused only on the binary variable. In order to understand the high numbers of people feeling discriminated against in specific countries (c.f. Figure 1), we investigated the subcategories of discrimination for these countries: The UK and France have a high percentage of people being discriminated based on color/race. In Iceland, discrimination based on age and gender are dominating while Montenegro leads the categories of discrimination based nationality and religion.

Category	Example questions
Media and social trust	Time in minutes spent on consuming news per day
Politics	Signed petition in last 12 months
Subjective well-being, social exclusion, health, religion, national and ethnic identity	Subjective happiness, Parents born in this country
Timing of life	Ever given birth to/fathered a child
Socio-demographics	Level of education
Justice and fairness	Decisions in country politics are transparent
Human values	Important to care for others

Table 1: Categories and exemplary questions of ESS9 data set (, ESSERIC).

Abbreviation	Description
agea	feature "Age"
AI	Artificial Intelligence
all	subset with all 139 preselected features
blgetmg	feature "Belong to minority group"
dvc	subset with best 30 features mentioned by davinci
eduyrs	feature "Years in education"
ESS	European Social Survey
facntr	feature "Father born in country"
hlthhml	feature "Hampered in daily activities"
hmn	subset with all features from ESS category "Human values"
ML	Machine Learning
netifr	feature "Net income fair"
NN	neural network
non	no sampling applied to training data
ove	oversampling applied to training data
pcc	subset with best 30 features based on PCC calculations
PCA	Principal Component Analysis
PCC	Pearson Correlation coefficient
pstplonl	feature "Posted about politics online"
sgnptit	feature "Signed petition"
SHAP	Shapley Additive Explanations
stfdem	feature "Satisfied with democracy"
SVM	Support Vector Machine
uemp3m	feature "Unemployed more than 3 months"
und	undersampling applied to training data
wgh	weighting applied to training data

Table 2: Abbreviations and their corresponding description

A.5 Details of Data Preprocessing

As described in Section 2.1, the options "Refusal", "Don't know", and "No answer" were included as outlier values, e.g. 7, 8, 9 for a question with the answers "Yes" (1) and "No"(0) in the data set. The gap between true answers and outliers increased significantly when large numerical values could be given as answers. This can be observed for the question "annual net pay" where the outliers are encoded with $7 \cdot 10^9, 8 \cdot 10^9, 9 \cdot 10^9$. While some ML algorithms can deal with such distorted distributions, approaches like neural networks require benefit from a standardized and normalized input. Therefore, we decided to encode outliers dependent on the individual question. For a categorical scale with a center value, the outliers were mapped to this central value. If the scale had no such value, a center value was created artificially via shifting the upper part of the scale and assigning the outliers to the newly created center value. For numerical answers, the outliers were mapped to the median value.

Furthermore, all data points containing NaN values were removed from the data set. This step reduced the number of samples from 49519 to 46655 and filtered out all data points corresponding to the countries Cyprus and Hungary.

The variable indicating the country of a participant was encoded using one-hot encoding. Integer encoding is not suitable for this variable since putting the

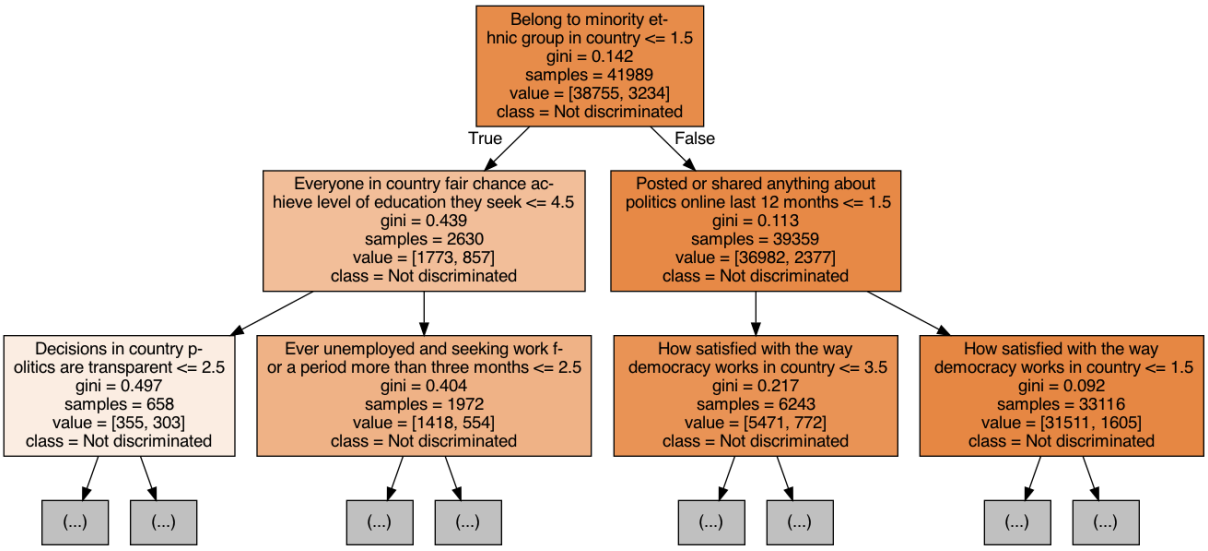


Figure 12: Decision tree of model trained with feature set "all" and sampling method "non", displayed with a max_depth = 2. For interpretation of the decision boundaries, please refer to the ESS documentation (European Social Survey European Research Infrastructure , ESSERIC).

countries in a specific order has no intrinsic meaning and might only mislead the learning algorithm. Including the one-hot encoded country information, the total number of features is increased from 113 to 139.

A.6 Decision Tree

The visualized output of the decision tree models is useful to infer the importance of the features w.r.t. the target. Together with the results of the PCC calculations and the methods for explainability, expert knowledge from the field of the social sciences can be used to test or create assumptions on the dependencies of the features and the target. An example with a reduced depth to enable visualization is to be found in Figure 12. Additionally, a brief analysis of the paths resulting in the leaf nodes with respondents feeling discriminated against representing the majority class was performed. For the stated example, one of the paths resulting in a distribution of 35 non-discriminated and 73 discriminated training examples is associated with the following ordered decision rules:

1. "Belong to minority ethnic group in country" <= 1.5,
2. "Everyone in country fair chance achieve level of education they seek" <= 4.5,
3. "Decisions in country politics are transparent" <= 2.5,
4. "Confident in own ability to participate in politics" <= 3.5,
5. "Mother's highest level of education" <= 162.5,
6. "Most people try to take advantage of you, or try to be fair" <= 4.5 and
7. "Take part in social activities compared to others of same age" <= 2.5.