IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2021, accepted April 11, 2021, date of publication April 21, 2021, date of current version May 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074676

Explainable Machine Learning for Default Privacy Setting Prediction

SASCHA LÖBNER[1], WELDERUFAEL B. TESFAY[1], TORU NAKAMURA[2], AND SEBASTIAN PAPE[1]

[1]Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt, 60323 Frankfurt am Main, Germany
[2]Information Security Laboratory, KDDI Research, Inc., Fujimino 356-8502, Japan

Corresponding author: Sascha Löbner (sascha.loebner@m-chair.de)

This work was supported by the European Union's Horizon 2020 Research and Innovation Program through the Project CyberSec4Europe under Agreement 830929.

ABSTRACT When requesting a web-based service, users often fail in setting the website's privacy settings according to their self privacy preferences. Being overwhelmed by the choice of preferences, a lack of knowledge of related technologies or unawareness of the own privacy preferences are just some reasons why users tend to struggle. To address all these problems, privacy setting prediction tools are particularly well-suited. Such tools aim to lower the burden to set privacy preferences according to owners' privacy preferences. To be in line with the increased demand for explainability and interpretability by regulatory obligations – such as the General Data Protection Regulation (GDPR) in Europe – in this paper an explainable model for default privacy setting prediction is introduced. Compared to the previous work we present an improved feature selection, increased interpretability of each step in model design and enhanced evaluation metrics to better identify weaknesses in the model's design before it goes into production. As a result, we aim to provide an explainable and transparent tool for default privacy setting prediction which users easily understand and are therefore more likely to use.

INDEX TERMS Privacy preference, privacy setting, machine learning, explainability, interpretability.

I. INTRODUCTION

Nowadays, many internet service providers are interested in retrieving personal data when a user requests online service access. The General Data Protection Regulation (GDPR)[1] requires to give the users a choice on that, but since many companies still want to get users' data for their business model, they have a strong incentive to work with bad default settings and dark patterns to lure the users into consent. Moreover, not accepting the settings recommended by the service provider almost always ends in the denial of the service. Besides this, most users are overstrained or even not aware of their self-privacy preferences [30]. Solove [44] claims that the principle of privacy self-management that includes the rights to notice, access and consent to the collection, use, and disclosure of personal data is beyond the users' capabilities and almost always ends in insufficient control over a user's private data. Therefore, a user-specific and tailored evaluation of anonymization and usability is mandatory.

To support users in acting according to their preferences and to prevent them from providing privacy-related data unintentionally, a support tool that can decrease the complexity of privacy settings and helps the users to better understand their privacy preferences is highly preferable. But privacy-friendly providers do not only need an adequate tool which the users can operate, they also need to demonstrate that their model is considering the users' interests. Therefore, if we want to support users with a Machine Learning (ML) approach, we also need to explain its results.

A tool fulfilling these requirements besides providing explainability was already proposed by Nakamura *et al.* [30]–[32]. Their two models can automatically predict 75 unknown privacy features from 5 features received from the user with an accuracy of 82 % and 85 % and are trained and tested based on a survey with 10,000 Japanese participants. Their models utilize the ML methods of K-means and the Support Vector Machine

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao Liu.

[1]Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L119/1.

(SVM). An SVM-based model by Nakamura *et al.* [33] was tested in a user evaluation and therefore underpins the need and applicability of default privacy setting prediction in a real-world scenario.

Although the overall performance of these models is still good and the contribution was verified in a user evaluation [33], the question of how a user can understand the privacy setting prediction of such a ML approach has not yet been fully investigated. With the increasing demand for trust and explainability, also supported by many new regulatory obligations to ML models such as stated by the GDPR, a user right to demand transparency has been strengthened. To be more precise, if decision making or profiling is involved, information must be provided in an understandable way to the user so that fairness and transparent processing of the model's logic can be assessed.

In this paper, we focus on increased interpretability, visualization, and validation of the achieved results. Our aim is not to improve the accuracy but to add interpretability to allow the user a better understanding of the results while preserving accuracy as best as possible. To achieve better insights from the beginning and to have a better feature selection approach compared to the previous models, we introduce a new explainable model based on a K-means clustering and an extended Iterative Dichotomiser 3 (ID3) [23] classification model. Each step building the model is evaluated, explained, and discussed in detail.

The remainder of the paper is structured as follows. Section II-D provides an overview of the scores used for the evaluation. In Section II we discuss related work in privacy setting prediction and have a closer look at explainability and interpretability. In Section III we explain the methodology of the paper. In Section IV, we compare the previous work of Nakamura *et al.* [30]–[32], and in Section III-C we elicit the requirements investigated in this paper. In Section V we present our approach including the requirements of Section III-C. In Section VI we present our results that are then further discussed in Section VII and the results are evaluated against the requirements. Finally in Section VIII we summarize the main conclusions and point out fields for further investigation.

II. RELATED WORK

This section provides an overview of related literature in the field of privacy setting prediction and introduces studies that deal with the challenges of explainability.

A. PRIVACY SETTING

The right to notice, access, and consent to the collection, use and disclosure of privacy-related personal data can be summarized as privacy self-management [44]. Although this right provides persons with comprehensive possibilities to manage and control their privacy, this concept exceeds the capacity of the average user [44]. Finally, this results in insufficient management of personal data.

Apart from this, Consolvo *et al.* [10] show that the majority of people do not read data policies although they are essential for a personalized privacy setting. According to their study, a lack of knowledge of privacy-related technologies raises the hurdle to assess the own privacy concern as well [10]. This mismatch between intended and real privacy settings is also shown by Madejski *et al.* [28], who investigate privacy settings in an online social network service.

Besides the lack of understanding privacy-related technologies an experimental study by Acquisti and Grossklags [1] identified a knowledge gap in legal forms when privacy policies are accepted. Their findings are supported by Pollach [37] who also confirm the existence of a knowledge gap in privacy-related, technical, and legal terms.

One approach to close this knowledge gap on the users' side is to provide privacy setting support systems as introduced by Fang and LeFevre [15] through their privacy wizard that addresses the privacy settings for social networks. This privacy wizard automatically keeps the interaction of a user as low as possible by predicting the privacy preferences of a user from a set of observed examples. Besides this, there are also approaches that try to develop a language that describes privacy policies [3], [11], [12]. Backes *et al.* [2] utilize abstract syntax and semantics to compare the privacy policies of enterprises. Similar to this, Tondel *et al.* [45] present a metric that facilitates the comparability of machine-readable policies. According to the opinion of Sadeh *et al.* [41], ML approaches will play a significant role in the future because their predictions of preferences are likely to exhibit a better fit to the real preferences than user-selected preferences. For instance, Tondel *et al.* [45] suggest a ML approach that generates preferences in the context of privacy agents. Their solution also has the advantage to disconnect the privacy preference self-assessment from a specific situation where the achievement of a service might exhibit a higher importance and the privacy preferences might be neglected [45]. An open issue of these studies is the justification why a user should trust the privacy setting predictions of such approaches. This can be only achieved by adding transparency and explainability in every step of the model design, as well as explaining the overall prediction in an easy understandable manner. To the best of our knowledge, this is the first paper addressing explainable machine learning for default privacy setting prediction.

B. EXPLAINABILITY AND INTERPRETABILITY

In the field of ML, a clear differentiation of explainability and interpretability is blurred within the literature. According to Lipton [24], who analyze the different approaches in detail, a large group of researchers who utilize interpretability to reinforce trust exists. They claim that other groups who combine interpretable models and models that uncover causality in the data exist as well. Lipton *et al.* also identified researchers who relate interpretability to understandability or intelligibility, aiming to understand in detail how a model works. Also, post-hoc interpretation is identified

that compared to other definitions neglects to explain how a method works in detail.

Besides this, Gilpin *et al.* [17] define interpretability alone as insufficient for humans to trust black box models. They claim that explainable models are interpretable by default but up to them, this does not hold vice versa.

In contrast to this, Biran and Cotton [5] state out that explainability and interpretability are closely related. They propose that a system is interpretable if the behavior of that system can be understood by a human. This can happen by introspection or a provided explanation.

This paper will follow Miller [29] and treat interpretability and explainabililty as interchangeable. As shown in Figure 1, we treat interpretability as an additional channel besides the evaluation metric that only provides a performance measure of the model but cannot answer the question why a certain decision in the model was made.

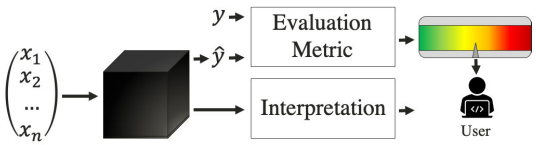


FIGURE 1. Demand for interpretability [24].

C. CHALLENGES OF EXPLAINABILITY

With the introduction of regulations that strengthen the right of a user to understand and interpret ML-based predictions as covered by the General Data Protection Regulation (GDPR) in Europe, approaches that increase the interpretability of ML models have to be taken into account [18]. Related challenges are described in the following studies.

Hall and Gill [20] highlight that particularly strongly regulated industries, e.g. banking, insurance, and healthcare are nowadays restricted from using simple and linear models for predictions instead of more complex machine and deep learning approaches. Moreover, they line out the existence of a trade-off between interpretability and more complex models with higher accuracy.

Another major problem is the absence of a best ML model. Different ML approaches tend to produce different predictions for the same instance but exhibit the same accuracy score for a given evaluation metric [20]. To overcome these problems, Hall and Gill present several model-agnostic and model-specific approaches to increase the explainability.

Others such as Guidotti *et al.* [19] claim that the decisions of ML models are interlinked with the digital trace of people. Thereby, especially daily provided data such as location, purchases and comments can contain human biases that are adopted by ML models and therefore can result in discriminatory and wrong results. Moreover, Guidotti *et al.* claim that the term explainability is misleading because the degree of expertise needed is not clearly defined.

To define requirements of an explainability approach with the aim to increase interpretability and trust, Petkovic *et al.* [36] experimented with random forest-based classifications trees. Thereby, they suggest 6 questions a report for experts and non-experts should answer. These questions take into account the balance of the training data, the importance of features, direction of features, feature interaction, and understandability of the report.

In much more detail, the problems and information a report should provide are considered and summarized by Sokol and Flach [43]. They propose a framework for the systematic assessment of explainable approaches. This framework evaluates explanatory systems against the five requirements (functional, operational, usability, safety, and validation).

D. EVALUATION METRICS

The most common way to calculate the accuracy of a classification model is to compare the correct and false predicted instances. While this score is easy to understand, low accuracy scores for small classes are very likely to be hidden. The common accuracy is presented in Equation 1.

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i), \tag{1}$$

with actual y, predicted $\hat{y}$ and n instances.

Because the class size of our data is imbalanced, we require a score that takes the accuracy per class into account. While this can also be computed with the arithmetic mean, the harmonic mean is often preferred because it is more sensitive to outliers. Therefore, systematically underperforming classes are more likely to be detected. This can significantly improve the accuracy and dissolve discrimination of minorities in the data. The average classaccuracy$_{HM}$ is presented in Equation 2.

$$averageclass\ accuracy_{HM} = \frac{1}{\frac{1}{K} \sum\limits_{k \in K} \frac{1}{recall_k}}, \tag{2}$$

with $k \in \{1, 2, \ldots, K\}$ and K numbers of classes.

Another common method for evaluating a classification result is the confusion matrix and related evaluation metrics such as precision, recall and F1-score. In this study the classes will represent groups of users with a certain level of privacy awareness. In the following, we will explain how the scores are calculated in our approach.

From a confusion matrix the precision can be defined as the ratio of instances that were predicted to have the privacy concern of class 2 that exhibit actually the privacy concern of class 2 and instances that are wrongly assigned to class 2 [38].

In contrast to this, the recall describes the ratio of instances that actually exhibit the privacy awareness level of class 2 and are correctly assigned to class 2 [38]. Therefore, the recall is also known by the term True Positive Rate (TPR).

Taking into account recall and precision, the F1-score is defined as the ratio of actual class 2 privacy concern predictions to the arithmetic mean of instances assigned to the

TABLE 1. Queried data type.

No.	Categories
1	Address and telephone number
2	Email address
3	ID for online services
4	Purchase record
5	Bank account
6	IP address (unique network id 192.168.xxx.xxx)
7	Browsing history
8	Logs on a search engine
9	Personal info (age, gender, income)
10	Contents of email, blog, twitter etc.
11	Cookie (Cookies in Your Internet Web Browser)
12	Social Info. (Membership, Religion, volunteer records)
13	Medical Info
14	Hobby
15	Location Info
16	Official ID

TABLE 2. Intended purpose of data.

No.	Data purpose
A	Providing the service
B	System administration
C	Marketing
D	Behavior analysis
E	Recommendation

privacy concern of class 2 and instances that actually exhibit a privacy concern of class 2.

Because the F1-score combines the previous scores, it is a commonly used and good measure for model performance [23]. Again making use of the harmonic mean, here the harmonic mean of precision and recall, imbalanced classifications are likely to be detected.

III. METHODOLOGY

An overview of data collection, data structure, and data preparation are provided in the following paragraph.

A. DATA COLLECTION

The web-based survey in this study is designed to gain more insights into the users' privacy concerns in relation to their willingness to share personal data for a given purpose. In the survey, the answers of 10,000 Japanese participants were collected. As shown in Table 1, 16 data categories are defined. For each intended purpose, given in Table 2 the survey participants were asked to fill out their willingness to share each of these 16 data categories on a Likert scale. The Likert scale is designed within a range from 1-6 mapping to the Likert types from ''strongly disagree'' to ''strongly agree''.[2] The categories and data purposes are closely related to the ones from the Platform for Privacy Preferences Project (P3P) [48]. Moreover, some demographic and device data were added to achieve more insights into the users' behavior.

The resulting questionnaire finally exhibits 80 $(5 * 16)$ combinations of categories and purposes that in the following will be treated as feature values. By handing the survey via a web-based system to 10,000 participants, the final data set consists of 800,000 data points. In this paper, we also call the 80 questions features and the value of the feature is termed feature value. Talking about the 10,000 participants in the models, the terminology instances is used.

[2]A Likert scale consists of a minimum of four Likert types which can be mapped to a composite score [7].

TABLE 3. Distribution of age.

Gender	Age	ratio (%)
Male	20s	10.0
Male	30s	10.0
Male	40s	10.0
Male	50s	10.0
Male	Over 60	10.0
Female	20s	10.0
Female	30s	10.0
Female	40s	10.0
Female	50s	10.0
Female	Over 60	10.0

TABLE 4. Distribution of device.

Mobile phone	ratio (%)
iPhone	23.5
Android	30.0
Others	1.7
Not smart phone	44.9

B. DATA COMPOSITION

As shown in Table 3 the participants of the survey are uniformly distributed with regard to gender and age. Investigations on the digital nativity of users as introduced by [39] did not have a significant impact [30].

Table 4 informs about the used devices the online survey was accessed from. 53.5 % of participants used a smartphone to access the survey, whereby the share of Android users is 6.5 % higher compared to iPhone users.

The distribution of feature values that are directly mapped to the values of the Likert scale among all instances is important for a better understanding of the data. It is particularly noticeable that the number of instances is highest for the Likert values 1 (39.69 %), and 2 (29.85 %) and decreases significantly over 3 (18.24 %), 4 (8.45 %) and 5 (3.07 %) up to 6 (0.69 %).

C. REQUIREMENT ELICITATION

The aim of this chapter is to line out open research questions of the preceding papers of Nakamura *et al.* [30]–[32] that form the starting point of the current paper.

Analyzing the Model 1 (SVM) and Model 2 (Combined Scheme), three requirements have been derived that will be tackled in this paper.

- **R1 Appropriate Feature Selection**
  The first requirement is to find a method to optimize the selection of the best combination of n out of the 80 features for questioning the user. Table 5 for Model 1 and Table 6 for Model 2 show that the selection is done randomly and not user-specific in the previous approaches [30]–[32].

- **R2 Provide Interpretability**
  Current data protection regulations such as the GDPR formulate the need of explainability, especially for the affected user of an ML-based application. Therefore, the second requirement aims to make the provided ML model more transparent.
- **R3 Enhance Evaluation Metrics**
  Basic evaluation metrics exhibit the risk of systematic discrimination of small classes or user groups in the data set. Therefore, this requirement aims to focus on a more detailed evaluation of the model, to detect such misclassifications of underrepresented or small classes in the data set.

D. TECHNICAL IMPLEMENTATION

All tests were implemented in the programming language Python 3.7.4 and run on a late 2013 MacBook Pro with a 2.4 GHz Intel Core i5, 8 GB (1600MHz) memory and an Intel Iris 1536 MB graphic card. In the implementation the ML library `scikit-learn` [35] and the fundamental package for scientific computing `numpy` [21] were used.

IV. PREVIOUS APPROACHES

In this section we provide a brief overview of the four previous papers. In general, all papers aim to provide a solution to reduce the burden of privacy setting for web-based services according to the user's privacy preferences. Therefore, the first papers [30], [31] propose a machine learning-based approach for default privacy setting prediction that is further improved in the third paper [32]. In the fourth paper [33] the model is tested and evaluated in an user experiment.

A. TECHNICAL DESCRIPTION

The first three papers present two different approaches, while both are based on the same data. The first approach utilizes a Support Vector Machine (SVM) while the second approach uses a combination of K-means and SVM and is therefore called Combined Scheme. The following sections give an overview of the models with the final parameter setting but do not explain the process of parameter setting from the beginning.

1) MODEL 1: SUPPORT VECTOR MACHINE

Model 1 consists of 75 sub-models (Model 1.1 to Model 1.75) that are based on a classification SVM approach [30], [31]. In the learning phase, 5 of the 80 features are chosen randomly and used for training and prediction. Nakamura *et al.* [32] determine the number of selected questions $n = 5$ with $1 \leq n \leq 80$ by a performance indicator with several rounds of testing with randomly chosen features. As shown in Figure 2, each user has to answer 5 preselected questions and the 75 remaining features are predicted by the models. Therefore, each sub-model is trained with the 5 preselected features to predict one of the 75 features. Thereby, the feature values have been reduced from 1-6 to 1-3, mapping 1 maps to [1, 2]; 2 : [3, 4]; and 3 : [5, 6]

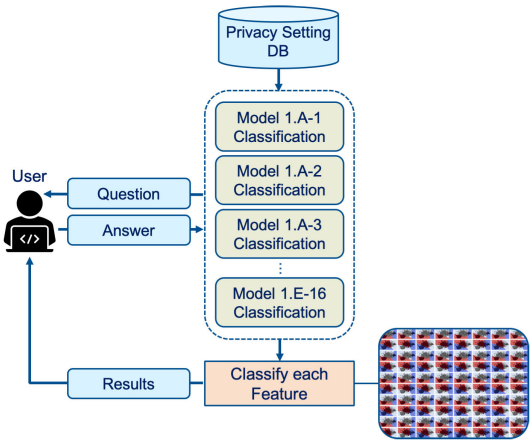


FIGURE 2. Flow chart Model 1.

TABLE 5. Results of SVM-scheme with optimization (#Training data = 9,000, #Test data = 1,000).

Combination					Accuracy TRD	Accuracy TED
A-8	B-12	C-16	D-14	E-11	0.8589	0.8566
B-7	C-12	D-6	D-14	D-15	0.8540	0.8519
B-12	B-15	D-5	D-8	E-6	0.8510	0.8470
B-7	C-16	D-11	D-14	E-11	0.8540	0.8518
B-4	B-15	D-14	E-6	E-11	0.8522	0.8491
B-8	C-16	D-14	E-10	E-11	0.8547	0.8525
A-8	B-12	D-6	D-14	E-11	0.8545	0.8531
B-4	B-15	D-6	D-14	E-11	0.8528	0.8510
A-3	A-16	C-12	D-11	E-3	0.8504	0.8480
B-7	B-12	D-14	D-15	E-6	0.8531	0.8503
B-7	C-14	D-10	D-16	E-11	0.8524	0.8499
B-7	C-12	D-10	D-16	E-11	0.8515	0.8486
A-2	B-7	D-14	D-16	E-11	0.8547	0.8532
A-12	B-7	C-14	D-6	D-15	0.8537	0.8518
A-12	B-8	C-16	E-10	E-11	0.8526	0.8500

and used as acceptance levels for the SVM classification models. For the evaluation metric of the model, the percentage of correctly guessed values is calculated by comparing the 75 predicted values with the original values, as given in Equation 1 with $\hat{y}_i :=$ predicted feature value, $y_i :=$ actual feature value.

In the guessing phase, five questions were chosen randomly as input for the trained SVM models. Then a prediction is made and the accuracy is calculated by the percentage of correctly guessed values.

Finally, an average model accuracy is calculated by repeating the random choice of input questions 15 times. The results are shown in Table 5.

2) MODEL 2: COMBINED SCHEME

This model, first introduced in Nakamura *et al.* [30], [31], consists of two sub-models and combines Model 2.A (K-means) and Model 2.B (SVM) as shown in Figure 3. In a first step, unsupervised learning with K-means is utilized to generate K clusters from the training data set. Each instance is assigned to one cluster k with $1 \leq k \leq K$. Then the feature values of the gravity point are set as target values for the instances in the referring cluster. Sub-model two is

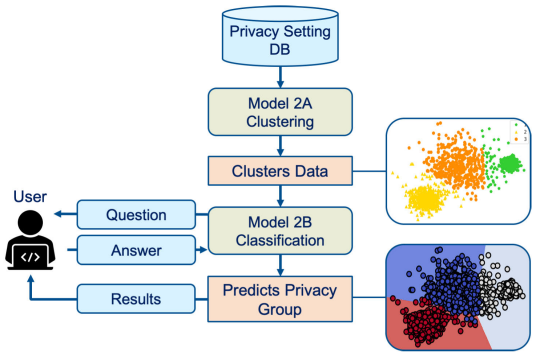


FIGURE 3. Flow chart Model 2.

TABLE 6. Accuracy of Combination Scheme (#Training data = 9,000, #Test data = 1,000).

Combination					Cluster Accuracy TRD	Accuracy TRD	Accuracy TED
A-11	A-15	B-4	C-2	D-6	0.7314	0.8169	0.8174
A-12	B-7	B-8	D-11	E-9	0.7490	0.8211	0.8217
B-6	B-7	D-7	E-10	E-11	0.7247	0.8224	0.8233
A-10	B-4	D-4	E-6	E-8	0.7441	0.8205	0.8206
A-10	B-4	D-6	D-9	E-6	0.7460	0.8194	0.8200
A-10	B-4	D-6	D-9	E-7	0.7638	0.8234	0.8250
A-10	B-4	D-7	D-9	E-6	0.7594	0.8223	0.8230
A-10	B-4	D-9	E-4	E-6	0.7510	0.8192	0.8195
A-11	B-4	B-8	D-10	E-6	0.7432	0.8206	0.8207
A-11	B-4	D-10	E-6	E-13	0.7559	0.8212	0.8214
A-13	B-4	D-11	E-6	E-11	0.7430	0.8211	0.8224
A-16	B-6	B-10	D-8	E-6	0.7577	0.8231	0.8235
B-4	B-10	D-4	D-13	E-7	0.7456	0.8230	0.8239
B-4	D-4	D-6	D-13	E-12	0.7495	0.8237	0.8244
B-4	D-6	D-9	E-4	E-7	0.7408	0.8232	0.8243

based on a supervised learning approach using an SVM. Similar to Model 1, the input of the SVM are the selected questions $n = 5$ with $1 \leq n \leq 80$ of all features. The SVM model then uses these selected 5 features to assess the instance to one of k clusters, whereby the classes of the SVM are mapped on the cluster of the K-means model. Finally, the accuracy is calculated as in Equation 1 with $\hat{y}_i :=$ centroid value for instance i in the predicted cluster, $y_i :=$ actual feature value.

The results of this scheme for 9,000 instances in the training and 1,000 instances in the testing set are shown in Table 6.

B. USER EVALUATION

Nakamura *et al.* [33] did a user experiment for investigating users' impressions of default privacy setting prediction system and the impact of nudge effect, that is, observing the results of four groups given different suggestions from; (1) original model, (2) privacy-biased model, (3) open-biased model, and (4) random suggestion model. The experiment was conducted from March 26 to April 2, in 2018, and the number of participants was 552. The participants were divided into four groups, each of which is assigned the previous four models. They first answered to five predictor questions and got predicted answers for remaining 75 questions from assigned model. If the predicted answers were different from their opinions, they changed these answers following to their opinions. In the study the acceptance rate was based on the ratio of all 75 non-predictor questions where the participants did not change the predicted answers. After engaging with the prediction system experiment, participants gave their impression of the system and experiment via a survey, for example, "This system would be convenient to help control the disclosure and protection of my personal data" [33].

The main two results of the experiment and survey are (1) majority of the participants is positive for such a prediction system, and (2) neither the privacy-biased model or the open-biased model produced a statistically significant difference in the proportion of accepted predictions. (1) means there is a potential need for such a prediction system. (2) suggests such a prediction system can easily push users towards openness or sharing if the system has evil intention. This result leads to the requirement of explainability for a default privacy setting prediction system.

V. APPROACH

In this section a new Model 3 (Interpretable Scheme) is introduced as an extension of the Combined Scheme that is based on two successive sub-models. The Combined Scheme was preferred because it exhibits only two sub-models that have to be explained, compared to 75 sub-models in the SVM-based scheme.

Moreover, to tackle *R1* (Appropriate Feature Selection), Model 3 is based on K-means and an extended Iterative Dichotomiser 3 (ID3) algorithm that uses methods of the C4.5 such as pre- and post-pruning. The model tackles this requirement with an impurity metric-based approach, utilizing the Information Gain (IG). The idea behind this is to derive a decision tree that enables the user to answer personalized questions iteratively that were learned from a training data set. Although building the tree can be computationally time-intensive, the classification of an unseen user is very quick after the learning phase because no further computation is necessary.

To tackle *R2* (Provide Interpretability), model-specific methods are used to provide further insights into how the model was built and how the model makes decisions. The interpretability is increased by a visualization that supports the user in understanding how Model 3 makes decisions. The visualization is prepared in a simplified way for experts and non-experts. The aim of the provided visualizations is on the one hand to increase trust and understanding but also on the other hand to show possible constellations where the algorithm might not provide reliable results.

To tackle *R3* (Enhance Evaluation Metrics), each sub-model is measured against enhanced evaluation metrics and the results of the metrics are explained to increase also the explainability. Besides the basic calculation of accuracy Equation 1, we prefer the *averageclass accuracy*$_{HM}$ (cf. Equation 2) that is useful for an unbalanced data set because it takes the deviation of small classes more into account and therefore lowers the burden to identify

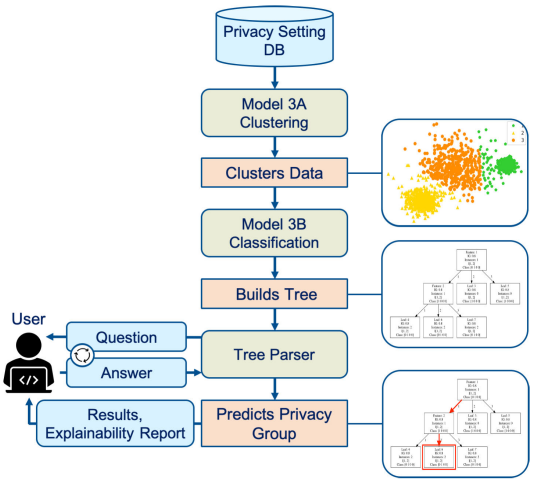


FIGURE 4. General flow chart of Model 3.

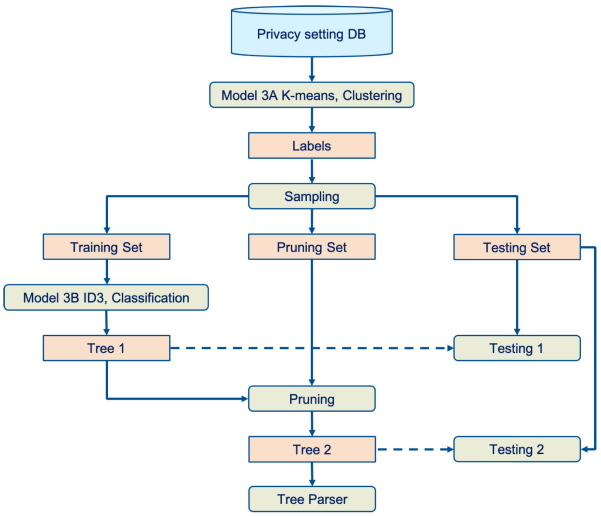


FIGURE 5. Technical flow chart of Model 3.

classification problems [23]. Overall, more complex evaluation metrics based on the confusion matrix such as Precision, Recall, and F1-score will be introduced to achieve a better understanding of the False-Positives and False-Negatives.

Figure 4 shows a general view flow chart of the explainable scheme. Model 3 is divided into an unsupervised classification (Model 3.A) and a supervised clustering model (Model 3.B) similar to the combined scheme. The big difference is that Model 3.B creates a classification tree. Starting with the same question at the beginning, the user is asked up to 4 individual questions in relation to the previous answer. Using this more user-oriented approach, we aim to achieve a better performance compared to a model where every user is asked the same question. The required amount of questions can also be reduced to a minimum because the algorithm stops earlier, if a user can be clearly classified based on less than 5 questions.

Figure 5 provides more technical insights into the Model 3 approach. In the first step, the privacy setting database is classified with an unsupervised learning approach. To group the privacy concern, we aim to map the clusters on the Westin/Harris Privacy Segmentation model [10]. This model defines people with the highest privacy concern as fundamentalist, people with a medium privacy concern and a balanced privacy attitude as pragmatist and people with no privacy concern as unconcerned.

To provide a better understanding how the classes differ from each other, we visualize the overall IG and the IG in between classes. After that, instances are labeled, the data set is sampled and split into a training, pruning, and testing set. The training set is used to train the classification model with the class label from the previous step as Y. The performance of the resulting classification $\hat{Y}$ is evaluated against the testing set and then further improved with the pruning set. After improving the classification tree with the pruning set the resulting classification tree is evaluated again with the testing set. The pruning is repeated until all post-pruning conditions are met. In our approach, we have used error pruning, cutting the leaf if the missclassificationrate in the parent node is higher compared to the missclassificationrate in the leaf. Although it is also possible to cut leaves that do not exhibit a minimum number of instances, this method could not increase the results significantly, which is the reason why the input variable `min_samp` was in all tests set to 0. The resulting classification tree can now be handed to the graphic user interface that we call tree parser where the user can parse the tree. Finally, we generate a brief report summarizing the main information of the tree path and results from the evaluation of the model.

VI. RESULTS

In this chapter, we will start by presenting the results of the different steps from Figure 5 and then present the explainability report.

A. CLUSTERING

Nakamura *et al.* [30], [31] have already tested K-means [27], Ward's method [47], and DB-Scan [13] whereby K-means performed best. In this paper, we introduce the Gaussian Mixture Model (GMM) [6] and Spectral Clustering [34], [46] and test it against a Mini Batch K-means algorithm [42] that finds a better starting point and relies on randomly sampled subsets of the original data set. Each model is run 10 times and the best result is chosen. To evaluate the performance of the clustering results we use the Silhouette Score. We have chosen GMM and Spectral Clustering because they are known to deal with high dimensions well. To understand the performance of the algorithms better, we compare the visualization of the data set that we achieve by reducing the dimensions from 80 features into a two-dimensional plot. To do this, we utilize the t-Stochastic Neighborhood Embedding (t-SNE) [26] and the Principal Component Analysis (PCA) [9], [25]. Figure 6 shows the resulting plots. To underpin the visualization results, we use the Silhouette Score in Table 7 to evaluate the performance of the overall clustering result, for the cluster

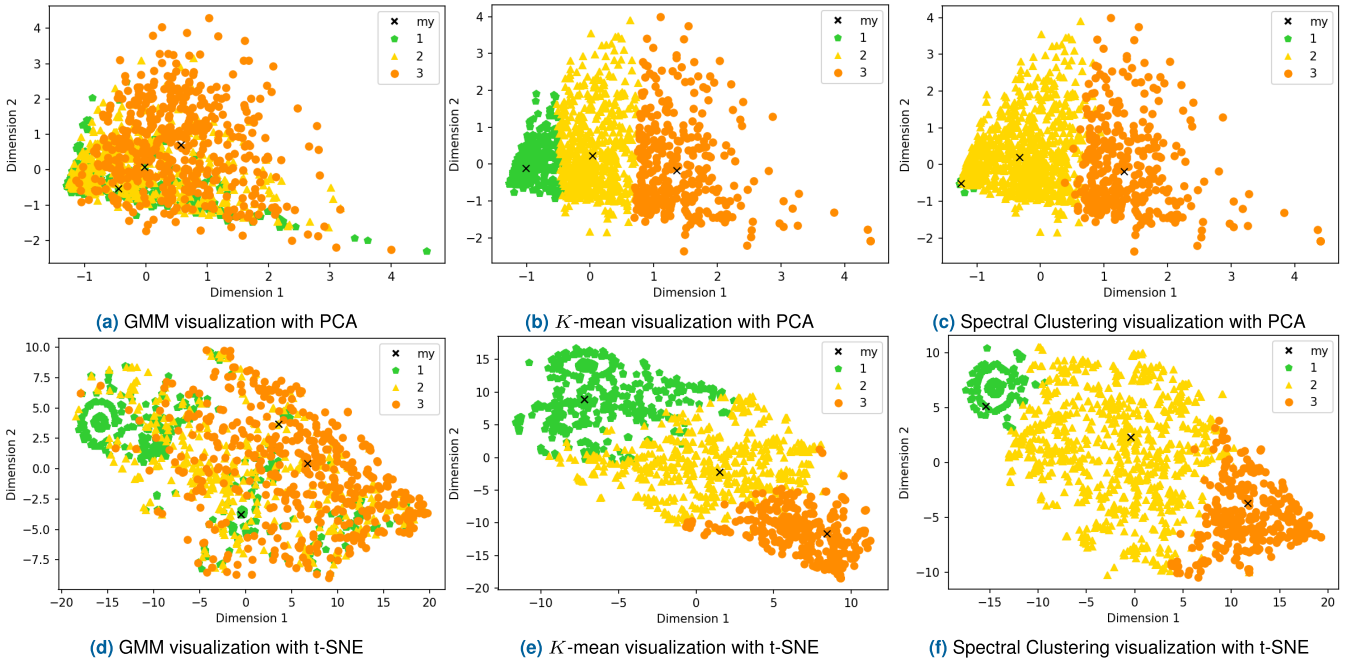


FIGURE 6. Clustering results for 3 clusters, using GMM, *K*-means, and Spectral Clustering algorithms and utilizing PCA, and t-SNE for visualization.

TABLE 7. Silhouette Score for different clustering algorithms GMM, *K*-means and Spectral Clustering with $K = 3$.

$K = 3$	Silhouette Score
GMM	0.0095
K-means	0.2574
Spectral Clustering	0.1209

size $K = 3$. The visualization of the GMM shows completely overlap clusters as visualized with the PCA (see Figure 6a) and t-SNE (see Figure 6d). The Silhouette Score is very close to zero, this means that most of the points could also be placed in another cluster as well. For K-means, both visualizations with PCA and t-SNE (see Figure 6b, 6e) show a very well separation of the clusters. While there exists an overlap in the border region, no points are placed close to the centroid of another cluster. This very clear visualization is also underpinned by the highest Silhouette Score of 0.2574 compared to the other methods. For the Spectral Clustering, the visualizations (see Figure 6c, 6f) also show clear borders between the clusters. Nevertheless, the Silhouette Score is much lower compared to K-means. Therefore, K-means was chosen for further investigation.

In a second step, we assess the optimal cluster size of K-means with $K \in 4, 5$. Again, we have chosen the best out of 10 runs per algorithm. The criteria of visualization used above (see Figure 7) and Silhouette Score (see Table 8), the Westing/Harris Privacy Segmentation model [10] was taken into account.

While the visualization for $K = 4$ has clear borders and well-placed centroids as shown for PCA (see Figure 7a) and for t-SNE (see Figure 7c). For $K = 5$ Figure 7b shows the PCA visualization that looks not that clear. Especially for

TABLE 8. Silhouette score for *K*-means $K \in 2, 3, 4, 5$.

	Silhouette Score						Mean Answers				
K	1	2	3	4	5	All	1	2	3	4	5
2	0.46	0.27				0.38	1.57	3.03			
3	0.45	0.11	0.19			0.26	1.25	2.20	3.35		
4	0.46	0.11	0.17	0.32		0.25	1.23	2.13	3.17	4.77	
5	0.47	0.14	-0.04	0.20	0.16	0.24	1.18	1.97	2.54	3.04	4.20

the t-SNE visualization (see Figure 7d) shows an overlapping area between cluster 2, 3 and 4 can be identified. This result is supported by a lower Silhouette Score for these clusters. While the overall score between the 3 and 4 cluster solution is very close together, the 4 cluster solution has the lowest score and is therefore rejected. Based on the Westing/Harris Privacy Segmentation Model we expect a cluster that can be mapped to an unconcerned behavior, the mean answers show that this is only possible with the 4 cluster solution.

Finally, we map the 4 cluster result on the Westin/Harris Privacy Segmentation Model in Table 9. Taking the mean and median into account, the mapping of the first cluster to the privacy concern fundamentalist fits very well. The second cluster is somewhere between fundamentalist and pragmatist and therefore labeled pragmatist (low). The third cluster is labeled as pragmatist. Because there seem to be fewer instances in the data that are completely unconcerned, we label the 4$^{\text{th}}$ cluster unconcerned (low).

B. CLASSIFICATION

At the beginning, the data is sampled by stratified sampling [23] and split into a training, pruning and testing set. To achieve an equal distribution of each cluster in the training set and to compensate the low size of cluster 4 as shown

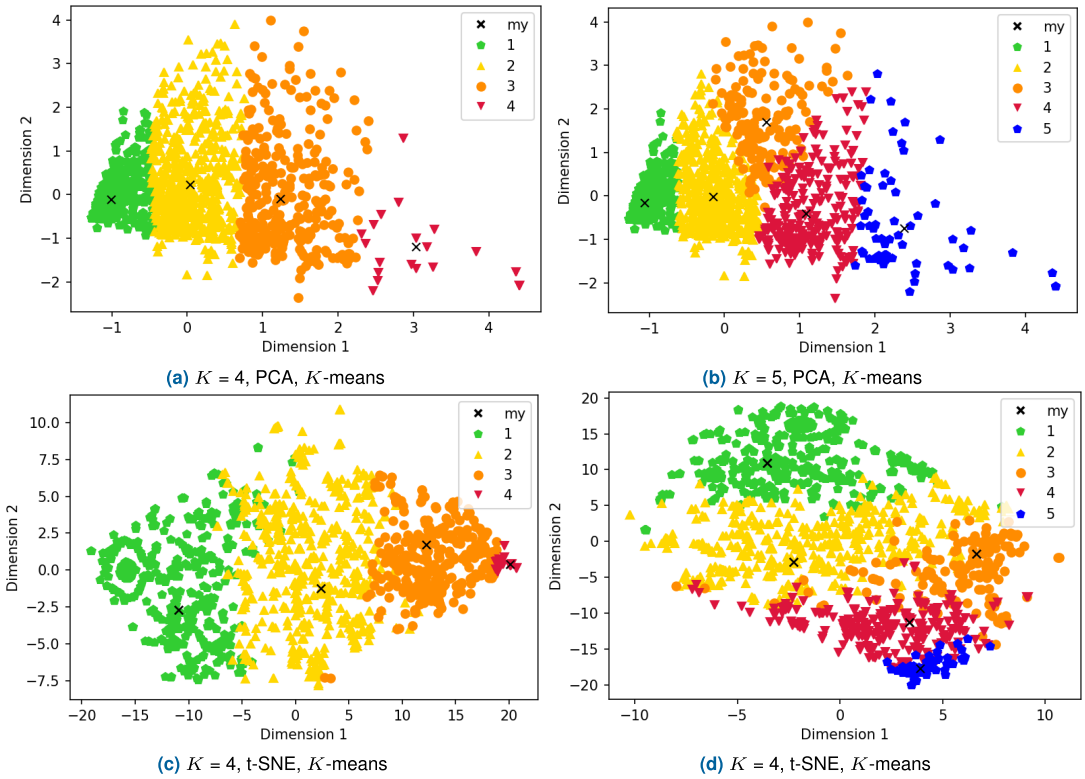


FIGURE 7. *K*-means with $K = 4$, and $K = 5$ and visualization with PCA, and t-SNE.

TABLE 9. Mapping on the Westin/Harris Privacy Segmentation Model.

Cluster ID	Privacy Concern	Mean Answer	Median Answer
1	Fundamentalist	1.2274	1
2	Pragmatist (low)	2.127	2
3	Pragmatist	3.1673	3
4	Unconcerned (low)	4.7703	5

TABLE 10. Cluster sizes before and after using SMOTE.

Cluster	1	2	3	4
Before	1811	1993	1120	76
After	1993	1993	1993	1993

in Table 10, the data is oversampled with the Synthetic Minority Over-sampling Technique (SMOTE) [8]. Figure 8 shows clearly visible the increase of cluster 4.

To determine the best size of the training data, different sizes were tested, while the remaining data was split equally into the testing and pruning set. Figure 9 shows the results of running the algorithm several times. To better identify the trend of the *averageclass accuracy*$_{HM}$ a linear regression with 5 basis functions is utilized. From this figure, we can assume that the best size for the training data is between 5000 and 7000. Results below 5000 can be rejected because the model picks up too much noise and has therefore a low *averageclass accuracy*$_{HM}$ for the testing set [40]. Results higher than 7000 should be neglected because the algorithm begins to overfit, indicated by the decreasing *averageclass accuracy*$_{HM}$ for the training and testing data. We have chosen 5000 instances for the training data because this most likely to prevents overfitting and the data can perfectly be split into 4 folds for cross validation.

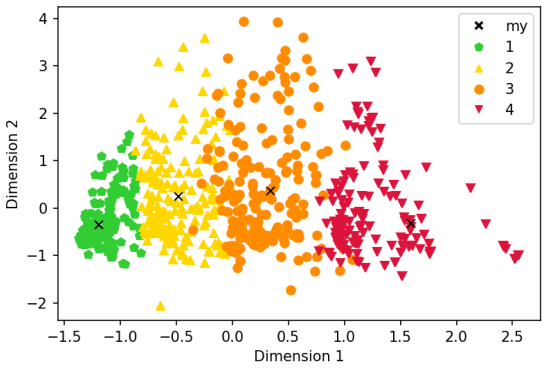
TABLE 11. Confusion matrix of basic ID3 algorithm, maximum depth = 5, imputity metric = IG.

75.83*	Predicted				Instances
Actual	**94.59**	5.41	0	0	905
	9.63	**83.35**	7.02	0	997
	0.18	19.29	**77.5**	3.04	560
	0	0	42.11	**57.89**	38
Instances	931	921	582	66	2500

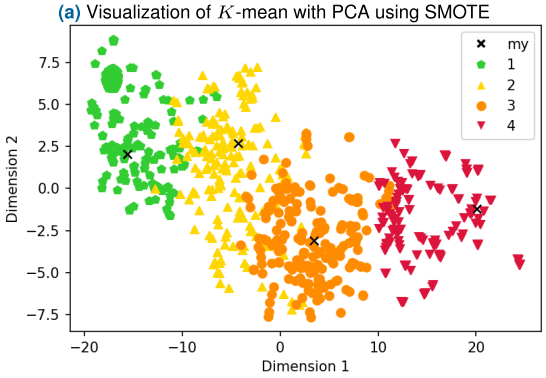
*$averageclass\ accuracy_{HM}$

In the following, we describe the steps during the optimization phase of the model. First we use a data set of 5000 instances without SMOTE and a testing set of 2500 instances. Table 11 shows the relative confusion matrix with the recall at the principal diagonal. The model-based on an ID3 algorithm with a maximum depth of 5 that uses the impurity metric information gain has an *averageclass accuracy*$_{HM}$ of 75.83 % with a significantly low recall in cluster 4 and a high rate of people who were assigned to class 3 and belong actually in class 4.

Table 12 shows a data set of 5000 instances with SMOTE and a testing set of 2500 instances. The *averageclass accuracy*$_{HM}$ has significantly increased to 81.72 %. The recall in class 4 has significantly increased to 73.68 % and there was also an increase in class 3, nevertheless the recall in class 1 and 2 has slightly decreased.



(a) Visualization of K-mean with PCA using SMOTE



(b) Visualization of K-mean with t-SNE using SMOTE

FIGURE 8. Visualization of SMOTE using t-SNE, and PCA utilizing *K*-mean with $K = 4$.

TABLE 12. Confusion matrix of ID3 algorithm, maximum depth = 5, imputity metric = IG, SMOTE.

81.72*	Predicted				Instances
Actual	**93.48**	6.41	0.11	0	905
	8.43	**80.14**	11.23	0.20	997
	0.18	11.43	**81.96**	6.43	560
	0	0	26.32	**73.68**	38
Instances	931	921	582	66	2500

$^*averageclass\ accuracy_{HM}$

Next, the standard ID3 algorithm was extended in several rounds of parameter testing and evaluation. First, we introduced reduced error pruning with the pruning set with 2500 instances for reduced error pruning. Second, all leaves with less than 4 instances in the training and pruning set together were cut in a simple post-pruning approach. Third, we have analyzed the tree and inserted some extra leaves in the following two cases where the ID3 algorithm would return the probability of the parent node. First, if the class label of the parent node for a higher feature value compared to an existing leaf with a smaller feature value is smaller, the class label of the leaf with the highest feature value is returned. Second, if the class label of the parent node for a lower feature value compared to an existing leaf with a smaller feature value is higher, the class label of the leaf with the lowest feature value is returned.

The final result of these extensions is given in Table 13. The *averageclass accuracy*$_{HM}$ has increased to 86.64 % and the recall of class 4 to 86.83 % while there was no decrease of the recall in any other class.

TABLE 13. Confusion matrix of ID3 algorithm, maximum depth = 5, imputity metric = IG, SMOTE, post pruning, min samples = 4.

86.64*	Predicted				Instances
Actual	**93.48**	6.3	0.22	0	905
	7.72	**81.84**	10.23	0.2	997
	0.17	9.46	**85.18**	5.18	560
	0	0	13.16	**86.83**	38
Instances	924	926	586	64	2500

$^*averageclass\ accuracy_{HM}$

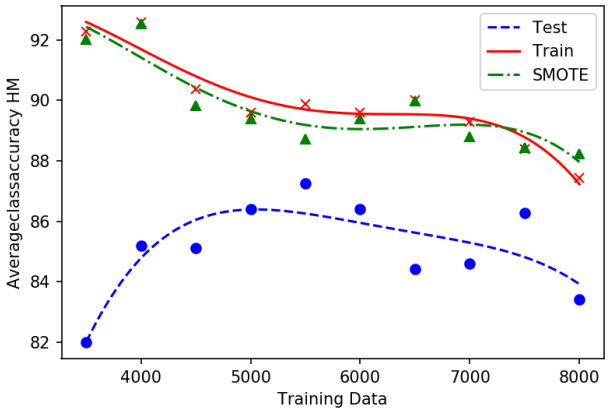


FIGURE 9. Determine the size of the training set.

TABLE 14. Cross validation of the best result.

Data set	Fold 1	Fold 2	Fold 3	Fold 4
1	TEST	PRUNE	TRAIN	TRAIN
2	TEST	TRAIN	PRUNE	TRAIN
3	TEST	TRAIN	TRAIN	PRUNE
4	PRUNE	TEST	TRAIN	TRAIN
5	TRAIN	TEST	PRUNE	TRAIN
6	TRAIN	TEST	TRAIN	PRUNE
7	PRUNE	TRAIN	TEST	TRAIN
8	TRAIN	PRUNE	TEST	TRAIN
9	TRAIN	TRAIN	TEST	PRUNE
10	PRUNE	TRAIN	TRAIN	TEST
11	TRAIN	PRUNE	TRAIN	TEST
12	TRAIN	TRAIN	PRUNE	TEST

To validate this result as resilient and to avoid a "lucky split" we have used a 4-fold cross validation. In a standard setup with only training and testing set, the first fold is used as the test set and the remaining 3 sets are used for testing. Then the second fold becomes the training set and again the remaining sets are used for testing. This procedure is repeated until every fold has been used for training [23]. Therefore with a standard 4-fold cross validation only 4 results can be achieved. In our approach we have a testing, a pruning and a training set. As shown in Table 14 each fold has a size of 2500 instances. We start with fold 1 for testing, fold 2 for pruning and combine fold 3 and 4 because we require 5000 instances for the training. Then we swap the folds until every combination was tested. This leads to 12 runs and produces a more reliable cross validation result, also avoiding a lucky split in the pruning set.

Again, our approach is based on 5000 instances in the training, 2500 instances for the pruning, and 2500 instances for the test data. Using the harmonic mean, the 12 resulting

TABLE 15. Mean of all 12 cross validation results.

	Final Result			Cross Validation		
Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	91.56	93.48	92.51	90.69	92.39	91.53
2	88.12	81.85	84.87	85.59	80.84	83.15
3	81.4	85.18	83.25	80.59	82.54	81.56
4	51.56	86.84	64.7	51.00	84.21	63.52
HM	74.14	86.64	79.90	73.09	84.77	78.50

TABLE 16. Overall accuracy based on harmonic and arithmetic mean and per class accuracy using the cluster centroid (a vector of $\mathbb{R}^{80}$) to predict the feature values in the scales from 1-6 and 1-3.

Likert Types	Class Accuracy 1	2	3	4	Avarage-classaccuracy$_{HM}$	Overall Accuracy
1-6	83.08	62.99	57.64	51.66	61.90	68.87
1-3	95.02	79.36	73.11	65.63	76.86	83.33

confusion matrices were summarized. From this, the precision, recall, and F1-Score were calculated, which was also done for the final model. These results are shown in Table 15.

So far we have evaluated the performance of the clustering and the performance of the classification based on the clustering results. In our last step, we will use the centroid of each cluster from the clustering result with $K = 4$ to make a privacy prediction for each unknown feature value, using the derived tree of the final classification result. For example, if the first instance is classified in class 2, the feature values of the centroid are set for all unknown feature values. Thereby, we round each value in the centroid that is a vector of $\mathbb{R}^{80}$ to integer values. Table 16 shows for the test data, the result with the true Likert types from 1-6 and the result with the linear merged Likert types as in the previous papers from 1-3. Again, the possible feature values are mapped on the Likert types. The accuracy for class $k \in \{1, 2, 3, 4\}$ is calculated by dividing the number of correctly predicted feature values in the respective class ($\hat{y}_{i,k} = y_{i,k}$) by the sum of all feature values in that class (n_k) similar to Equation 1. The sum of all feature values can be computed as instances in the respective class times features. To make the results comparable to the previous papers we calculate the overall accuracy similar to Equation 1. To get a better understanding how the accuracy in the classes differs we use Equation 2 to calulate the averageclassaccuracy$_{HM}$ based on the harmonic mean. In this case we replace the recall$_k$ with the accuracy$_k$.

For both scales, the accuracy for class 3 and 4 is much lower compared to class 1 and 2. For example, 51.66 % of the feature values in the range from 1-6 were predicted correctly. In comparison to the chance of 16.67 % predicting 51.66 % of the feature values of all test instances that were classified in class 4 correct is still a good result. Taking into account that guessing a feature value correctly on a scale from 1-6 is much harder than guessing on a scale from 1-3, we assess the results for the scale from 1-6 as the better result. No information is lost by merging feature values and the privacy prediction can be made more granular. Nevertheless, the result for class 1 is much better compared to the other classes. One reason for this could be that the median answer in cluster 1 is 1 and the mean answer 1.2274 (see 9). This in combination with the highest occurrence of feature value 1 is likely to leverage the accuracy when predicting the feature values of the fundamentalist. Taking another look at the overall accuracy and the overall accuracy$_{HM}$, the number of possible feature values have to be taken into account. 68.87 % of the feature values from 1-6 were guessed correctly which is much better than the chance of 16.67 %. Nevertheless, the accuracy in class 4 is far behind the accuracy in the other classes. While the overall accuracy is still high, the overall accuracy$_{HM}$ is slightly lower, indicating correctly that the accuracy is not homogeneous over the 4 classes. Finally, in all classes the prediction of the feature values is significantly higher than the chance, so the overall model can provide benefits to the users when making privacy setting predictions.

C. EXPLAINABILITY

Until now, the visualization of the clustering results were presented and the choice for the size of training, pruning and testing sets was justified. We have seen the evaluation metrics precision, recall and F1-score and the final result was validated with cross validation. In this section we will add some more explainability with the aim to provide more insights into how decisions are made by the model.

1) CLUSTERING

To get better insights into the clustering results, we take a look at the features that separate the data best, using the Information Gain (IG). Guidotti *et al.* [19] mention feature importance as an effective solution to provide global or local explanations to a black box model. To calculate the feature importance, we use the IG as introduced in [23]. We have chosen the IG as score for feature importance because the IG will later be used as impurity metric in the classification. First, we calculate the IG for each of the 80 feature values using the 10,000 instances and the labels of the clustering result. Figure 10 shows the features with the highest IG on the right that separate the data best to the features with the lowest IG on the left that separate the data worst. The overall entropy of the data is 1.6349 bit. The closer the IG to the entropy, the better a feature separates the data. Especially the data purpose *Behavior Analysis (D)* and the category *Cookies (Web Browser) (11)* are often represented in the top ranking. In the left corner, the data purpose *Providing the service (A)* is present.

While the IG for the whole data provides a good first impression which feature is most important to separate the data, in a second step we take a closer look at each cluster to better understand the differences in the clusters. Therefore, Figure 11 presents the IG, mean and median value for the instances that are assigned to cluster 1 and cluster 2 with an entropy of 0.9984 bit. Thereby, the figure shows only the top 10 and worst features. In the top ranking on the right, especially the data purpose *Behavior Analysis (D)* and *Recommendation (E)* are present as well as the category *Cookies (Web Browser) (11)*. A good separation of these features is also indicated by the mean and median that are placed for

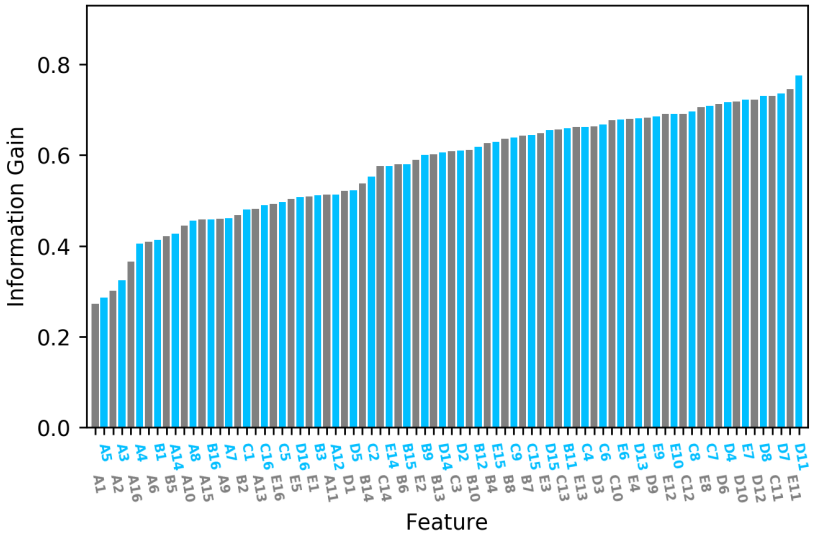


FIGURE 10. Information Gain of all clusters.

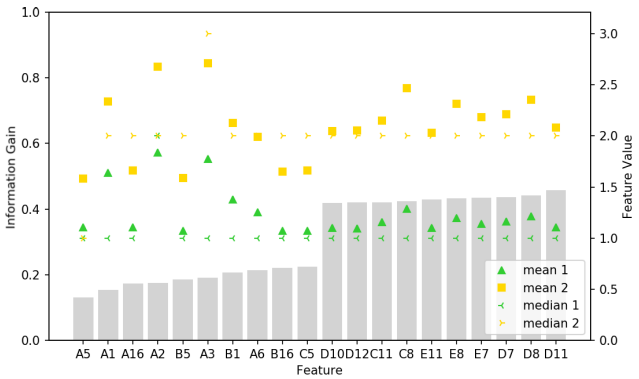


FIGURE 11. Information Gain, mean and median of cluster 1 and 2.

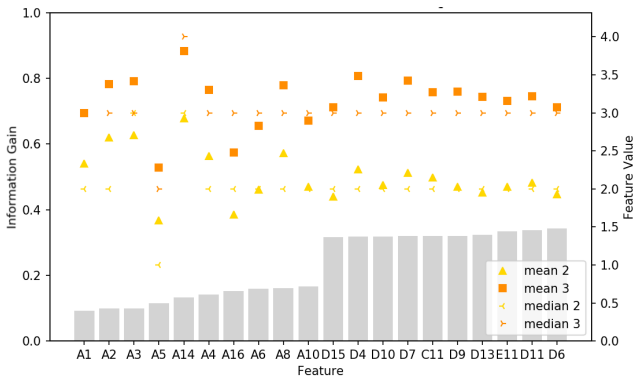


FIGURE 12. Information Gain, mean and median of cluster 2 and 3.

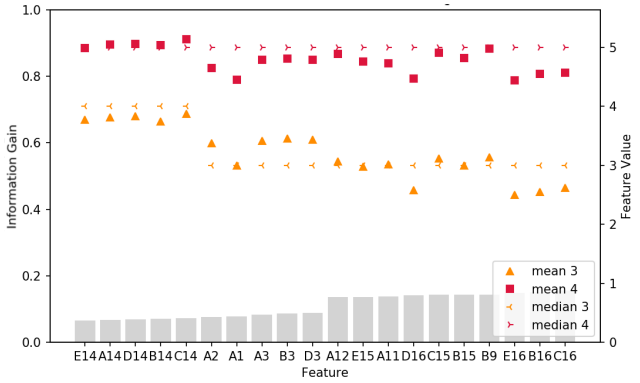


FIGURE 13. Information Gain, mean and median of cluster 3 and 4.

cluster 1 around the feature value 1 and for cluster 2 around the feature value 2. While the IG of the features at the left is significantly lower, the mean and median of the features cannot be separated as well as compared to the features with a higher IG. Especially the data purpose *Providing the service (A)* is omnipresent.

Comparing the instances of cluster 2 and cluster 3 (see Figure 12) with an entropy of 0.9425 bit, the result is quite similar to cluster 2 and cluster 3. For the features that separate the data the best, mean and median are centered for cluster 2 close to the feature value 2 and for cluster 3 close to feature value 3. For the features with a lower IG, this does not hold.

When it comes to the comparison of cluster 3 and cluster 4 (see Figure 13) with an entropy of 0.3414 bit, the results look different. All features are well-separated, although the features with a top-ranking are further apart. In the best and worst ranked features, no favorites can be identified.

2) CLASSIFICATION

To increase the explainability of the classification, we focus in this paper on the visualization of the decision tree. While others like Huysmans *et al.* [22] show that decision tables for small classification trees are more user friendly, our approach for a decision tree with 159 nodes builds on a reduced view of the tree. Moreover, we do not show and count nodes that return the probability of the parent node. While a tree with 159 nodes cannot be presented very well, the tree can be pruned more restrictively. Figure 14 shows every leaf that did

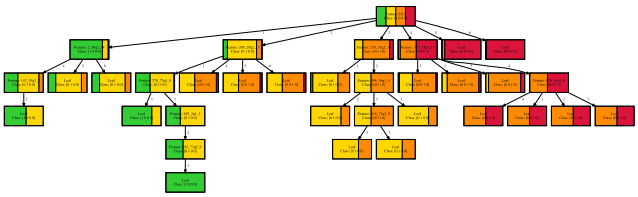


FIGURE 14. Final result of the classification, min_samp = 50, colored class probabilities.

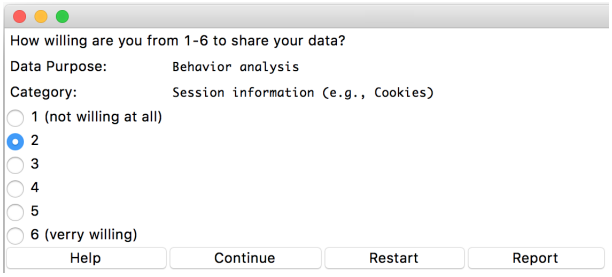


FIGURE 15. Prototype of GUI to interact with the tree parser.

not have a minimum number of 50 instances after the pruning phase.

As a result of more restrictive pruning, the tree has only 33 nodes excluding nodes that would return the probability of the parent node and a F1-score of 77.62 compared to 79.9 in the final result. The nodes are colored by the share of each remaining class. This can be interpreted as the probability to be classified to a specific class. The classification in the root node is equal for every class because when using SMOTE, all classes have the same probability. Instead of reducing the tree size to a minimum, the Figures 16 and 17 show only a summarized extract of the whole tree for 2 different paths.

Each node displays the feature acronym, the mapped class, and the probability of each remaining class. The red errors show the answers of the user and indicate the individual path of a user through a tree. Besides the nodes that are on the red path, all other possible answers for these nodes are displayed with their probability. By this, users can see how a different answer in one of the nodes would have influenced the classification result. Generally, a different answer closer to the root node has a bigger impact on the classification compared to the features closer to the leaves. The closer a feature to the root node, the better a feature separates the whole data into 4 classes and the higher the impact of an answer. Going down the branch, e.g. answering feature B15 in Figure 16 with a 6 has the same result as answering a 3. This shows that the impact of a strongly deviating answer is lower.

D. EVALUATION OF DIFFERENT CLUSTER SIZES K ON MODEL PERFORMANCE

In Section VI-A Clustering we chose the cluster size $K = 4$ because of the explainability and the mapping to the Westin/Harris Privacy Segmentation model. We have already shown that this choice is reasonable because of a positive Silhouette Score for all clusters. In this section we are comparing

TABLE 17. Precision, Recall and F1-Score for $K = 3$ and $K = 5$, ID3 algorithm: maximum depth = 5, imputity metric = IG, SMOTE, post pruning, min samples = 4.

	$K = 3$			$K = 5$		
Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score
1	90.44	92.76	91.59	91.02	91.87	91.44
2	86.35	82.43	84.34	83.17	74.60	78.65
3	85.13	88.46	86.76	52.59	69.58	59.90
4				77.14	70.26	73.54
5				57.35	78.64	66.33
HM	87.25	87.68	87.46	76.21	69.03	72.44

TABLE 18. Overall accuracy based on harmonic and arithmetic mean and per-class accuracy using the cluster centroid (a vector of $\mathbb{R}^{80}$) to predict the feature values in the scales from 1-6.

K	Class Accuracy					Avarageclass	Overall
	1	2	3	4	5	$\text{accuracy}_{\text{HM}}$	Accuracy
3	83.04	63.12	59.11			66.96	69.24
4	83.08	62.99	57.64	51.66		61.90	68.87
5	85.47	70.23	41.60	62.57	58.07	60.11	68.84

the model performance for clusters with $K = 3$ and $K = 5$, also to get a better idea of the choice $K = 4$.

To rerun the experiment, we utilize a simplified cross validation using only the splits 1, 5, 9, and 10 from Table 14. This approach provides more resilient testing results compared to a single run because every set was used once as a test set. The parameter settings used for the extended ID3 are the same as for $K = 4$ (see Table 17). Comparing the score for $K = 3$ with $K = 4$ (see Table 15) and $K = 5$ the Precision, Recall and F1-score are the highest over all classes and in the harmonic mean. In comparison to that, the F1-score for $K = 5$ is low, especially for class 3. Taking again a look at the Silhouette Score (see Table 7) the negative score of -0.04 has already indicated that there exist more instances that could be placed in another class than instances that are best placed in class 3. This is now also reflected in the classification result.

Table 18 shows the results for the prediction of the feature values on a likert scale 1-6, based on the clusters' centroids. Although the overall accuracy only slightly differs from each other, the $\text{averageclassaccuracy}_{\text{HM}}$ for K = 3 is a little bit higher.

For the reduced likert scale with the feature values 1-3, again the overall accuracy only slightly differs. Nevertheless, the $\text{averageclassaccuracy}_{\text{HM}}$ for K = 4 is somewhat lower.

Although the comparison of different Ks has shown that $K = 3$ performs best for both likert scales, the model performance seems to be roughly at the same level, so we conclude that it is fine to use $K = 4$. Another reason for this choice is the fourth class that maps much better to the privacy class of unconcerned from the Westin/Harris Privacy Segmentation model than class 3 for $K = 3$ does.

E. PERFORMANCE EVALUATION

In this section we give an impression of the performance of the tree parser in a Graphic User Interface (GUI) prototype that is built with the Python package tkinter.

Figure 15 shows the GUI prototype that is used to access the tree parser. The user can see the data purpose (A-E)

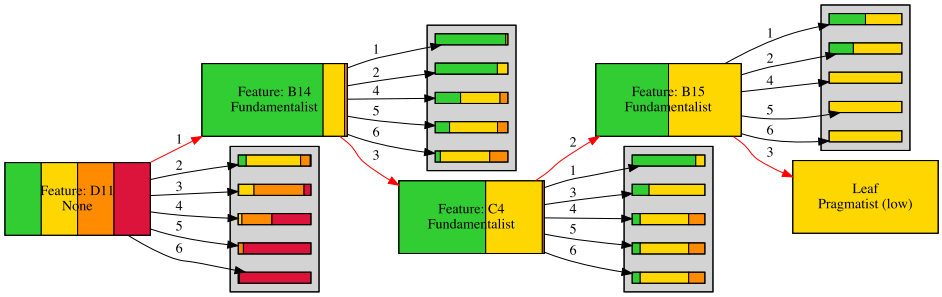


FIGURE 16. Path of answering 1, 3, 2, 3, with colored class probabilities and short branches.

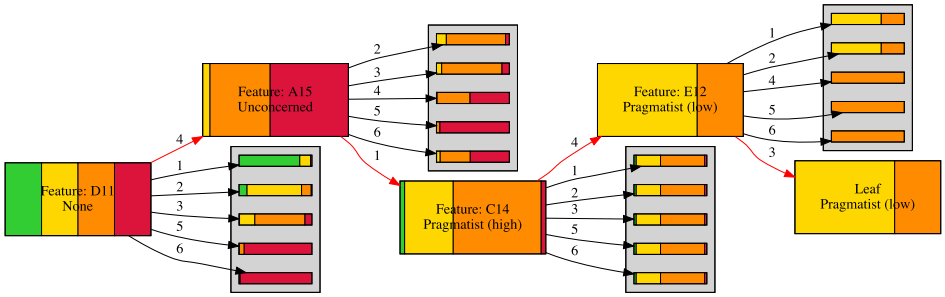


FIGURE 17. Path of answering 4, 1, 4, 3, with colored class probabilities and short branches.

TABLE 19. Overall accuracy based on harmonic and arithmetic mean and per-class accuracy using the cluster centroid (a vector of $\mathbb{R}^{80}$) to predict the feature values in the scales from 1-3.

K	Class Accuracy					Avarageclass	Overall
	1	2	3	4	5	$\text{accuracy}_{\text{HM}}$	Accuracy
3	94.88	79.69	75.05			82.38	83.96
4	95.02	79.36	73.11	65.63		76.86	83.33
5	96.08	86.06	63.55	75.19	77.17	78.10	83.50

and the Category (1-16). The first feature a user is asked about is always the root node D11 with the data purpose *Behavior Analysis* and the category *Session information (e.g., Cookies)*. The user can use the radio buttons to choose the own privacy preference from 1-6 for the given feature. The input is confirmed with the continue button. The tree parser then follows the given feature value in the decision tree and the GUI asks to provide the privacy preference (feature value) for the feature of the new child. The times were measured 5 times and the mean is provided. For testing we have used the feature values as shown in Figure 16. Starting the GUI takes around 1 seconds. When the user chooses a radio button and clicks continue it takes less than 0.01 seconds to display the next node based on the provided answer. Generating the classification into a cluster takes less than 0.2 seconds. To generate a path such as in Figure 16 and Figure 17 it takes around 1 second. We have not implemented the prediction of the single feature values yet but we estimate the generation of the missing feature values to take less than 0.02 seconds. The estimation will not be time-intensive because the rounded feature values of the 4-cluster centroids are already known and the correct centroid has just to be returned based on the classification result.

Generally, the critical performance bottleneck in future will be the GUI and the users connection. Nevertheless, the provided impression of the GUI performance shows that the model can be used without any negative delay in a real world example.

VII. DISCUSSION

In this chapter, we will discuss and evaluate how well the requirements are met by the presented model and provide a more detailed interpretation of the current results.

A. FEATURE SELECTION

R 1 aimed to find an appropriate feature selection. This requirement is met perfectly by using a decision tree because users can follow their path based on their provided answers. Thereby each question contributes to classify the users into one of the 4 classes. As shown in Section VI-C1 there is a huge discrepancy of most relevant features between cluster 4 and the other clusters. This underpins our approach to have an appropriate feature selection. We have also restricted the tree to a depth of 5 so that maximum 5 feature values can be set by the user. A deeper tree has also been tested but the gain in accuracy is negligible. Taking a closer look at the decision tree, it stands out that in the special case of answering feature D11 with a feature value of 6, the tree immediately classifies the user to the lowest privacy concern. One could argue that this behavior can cause errors if the user makes a mistake because the decision is made solely based on one answer. Moreover, the user might mistrust the model if a decision is made too quickly. Another point of discussion is

what happens if the user does not give an answer at all. Using SMOTE, the probability to belong in one of the four classes is equal in the root node because each cluster that was used to train the classification tree has the same size. Nevertheless, this does not reflect the real ratio that would map most of the users as a fundamentalist in class 1. While on the one hand, an error message can be raised and the process is canceled, on the other hand, setting the most restrictive privacy setting that maps to a fundamentalist is also reasonable. Overall, the feature selection with the tree parser worked quite well and can ask user-specific questions by following the branches through the tree.

B. INTERPRETABILITY

R2 is about increasing the explainability of the presented model. In this approach we have taken a closer look at the clustering of submodel 3.1 using a visualization of the results and the IG. Alternative results that are not part of our final model have been analyzed to create a better understanding of the trade-off between accuracy and interpretability.

Especially the IG can provide deep insights into how the model makes decisions because also our extended classification algorithm ID3 is based on this impurity metric. One drawback of using pure IG is its dependency on the entropy. We have seen that if we take only the data from cluster 3 and cluster 4 the entropy is only 0.3414 bit. This indicates that there are difficulties in separating this data. This could be one reason why Figure 13 shows a different separation of the features that split the data the best, compared to the other clusters.

In the classification result, we provide simplified and summarized tree paths that we assess to be easily understandable at a glance by experts and non-experts. Especially displaying the probability of each class per node aims to give the user a feeling how her answers have influenced her classification result. This has, on the one hand, the advantage that the user can see quickly what would have happened if a different choice was made, on the other hand, the whole tree is not disclosed. Therefore, misuse such as model inversion attacks that aim to reveal sensitive features or recover training data [16] can be complicated. In case the user interrupts the algorithm in one of the child nodes, the algorithm could easily return the probability of the current node but as Figure 16 shows really well, this can easily result in a misclassification. E. g. for the first and second child, the classification would map the user as a fundamentalist, for the third child, the decision is very close to chance and only the last question brings confidence. In contrast to this, Figure 17 elucidates that the user in this scenario is not 100 % fundamentalist as classified. This is a very important information for the user because it indicates the reliability of her personalized classification. While most user groups have a high probability to be assigned to one specific class, the prediction is less reliable for a small group of users. The main reason for this is data shortage in a specific path. Although noisy nodes are compensated by error pruning, we assess user feedback as essential to further improve the classification tree and to identify noisy subgroups in the dataset that were not revealed in the post-pruning. Very restrictive pruning (see Figure 14) has been withdrawn because balancing accuracy and explainability is very difficult and the loss of accuracy for this example could not be justified well.

C. ENHANCE EVALUATION METRICS

R3 is about introducing more advanced evaluation metrics. To evaluate the clustering we have presented the Silhouette Score in Table 8. To have the overall Silhouette Score of a two-cluster result is very interesting because a first impression of the maximum possible score is achieved. Although a score close to 1 indicates a perfect separation of the data, this appears to be unrealistic for the presented data. While the first cluster is always well-separated from the other clusters with a score of around 0.45 for each K, we face a lot of difficulties in the other clusters. One reason is a high ratio of instances with a feature value of 1. These instances cannot be assigned to another cluster which increases the silhouette score of cluster 1. In contrast to this, the feature values for the other clusters are much more diverse which increases the probability for an instance to be placed in another cluster. Surprisingly, there are nearly no instances with continuous feature values of 6, which causes a lower score for the higher clusters. This effect is also shown in the mean of cluster 4 for $K = 4$ where the mean answers are with 4.7703 further away from 6 than the mean answer of cluster 1 with 1.2274 from 1. This is also reflected in the uneven distribution of instances over the clusters.

In the classification result we have focused on the confusion matrix and related scores. In Tables 11, 12 and 13 the increase of the recall by expanding the ID3 algorithm is clearly illustrated. Taking a closer look at Table 15 the low precision of class 4 can be easily detected. Here, the precision describes how many of the instances that have been assigned to class 4 are actually in class 4. As shown in Table 13 there are actually 38 instances in class 4 what is quite low compared to the size of the other classes. Therefore, a small ratio of people that are predicted in class 4 and are actually in another class has a disproportional effect on the precision of class 4. Generally, for such an imbalanced class, information from the precision is biased and weak. Moreover, the noise cannot be eliminated by using SMOTE because new instances are created based on the existing ones. Reasoned by the fact that the F1-score is calculated based on the harmonic mean of precision and recall, the result is biased by the class 4 precision score. Nevertheless, at least we learn that the size of class 4 has a negative effect on the model performance. Comparing the final result and the cross validation, we learn that a slightly lower performance of the classification can be expected in the real world.

Besides the performance of each individual submodel, also the performance of the overall model has to be taken into account, which is presented in Table 16. Under equal conditions that include only three different feature values and the same evaluation metric for accuracy (cf. Equation 1),

the performance of model 3 with an overall score of 83.33 % is in between model 1 with an average overall score of 85 % and model 2 with an average score of 82 %. Nevertheless, the low score in class 3 and 4, especially for 6 feature values cannot be neglected. Taking into account a study of [10] who find that ''users do not do what they say, and they do not know what they claim to know'' users face serious problems especially when answering questions about privacy-related technologies they claimed to know. This statement implies that the data is biased and does not reflect the real behavior of the user.

D. IMPACT

As shown in the user evaluation, users evaluate tools that provide support in privacy setting as very helpful. Nevertheless, one point of criticism was the non-transparent decision making of the model. With the model 3 approach, the decision and reliability for every user should be much clearer and the understanding of the model structure is facilitated.

Regarding the dissemination of such a tool in Europe, several advantages can be emphasized. First, the burden of setting privacy preferences is leveraged because the required time to find personalized preferences is reduced to a minimum while at the same time the trust in the model is increased. Second, our approach enables a wide range of different features that can be personalized with our tool. With such a solution, companies do not need to group the features and can receive much more precise settings while the privacy of the user is maximized. Third, by using such a tool, general privacy awareness is increased because the user gets insights into her own privacy behavior. With our tool of mapping and visualization it is even possible to receive feedback about one's own privacy sensitivity. Implementing tools for privacy setting prediction can also be seen as a chance for companies to stand out with a data-friendly service. By increasing privacy awareness also the discussion about fair handling of private data can be pushed forward and vulnerabilities in the data protection regulations which a few market participants take advantage of become more likely to be fixed.

E. LIMITATIONS

The first limitation of the presented model is the struggle of users to answer privacy-related questions as described by Consolvo *et al.* [10]. If the user cannot assess her privacy concern if asked directly, the whole model is somehow biased. Regarding the presented data, users seem to overestimate their own privacy concerns. In general, this is not a major issue because the privacy settings will be more restrictive and a user might think twice to disclose private data if a service is denied based on the predicted settings.

Second, there could also be biases in the sampling process. For example, maybe certain groups of people are not interested in participating in studies about privacy because they are not concerned about privacy at all and cannot see any benefit from it. The other way around, people that are concerned about privacy so much that they do not want to participate in any study at all might also exist.

Third, the fourth cluster was very small and not well-separated from the other clusters. To use this model in a real-world scenario, more data for this cluster in particular is likely to significantly increase the accuracy.

Fourth, the presented model 3 consists of 2 submodels that can be described as whitebox model. While whitebox models are known to be easily interpretable, more complex blackbox models such as deep neuronal networks might achieve a higher accuracy.

Fifth, for interpretable ML further approaches such as adversarial training to increase adversarial robustness, and influential samples measuring how the model's output is influenced if a certain data point is removed [4] might be interesting to test in future. Moreover, especially for blackbox models the usage of local interpretable model-agnostic explanations (LIME) could be used [20].

VIII. CONCLUSION

In this paper, we have presented and evaluated an explainable machine learning model for default privacy setting prediction based on a K-means clustering and extended ID3 classification tree. Finally, we mapped the users based on their classification to one of these clusters that were built upon the privacy preferences of a training group and used the centroid of the respective group for privacy setting prediction. Compared to the previous studies, we have elicited and enhanced the requirements of building an appropriate feature selection with the tree-based approach, introducing the tree parser that asks the user individual questions based on the classification tree and follows the respective path of the tree until a classification can be made. We provide enhanced interpretability by visualizing the classification and clustering results and investigating the information gain of the most and less important features in each cluster and the probability to be classified to a certain class in each node in the classification tree. Moreover, we give insights into what would have happened if a different answer were given. We enhance the evaluation metrics by introducing the silhouette score to evaluate the clustering and the confusion matrix based metrics recall, precision and F1-score. Besides this, we use the $\text{averageclassaccuracy}_{\text{HM}}$ to detect minority classes and biases in the classification result. Finally, we compared the scores of the presented explainable model and the two previous models. Although the accuracy is not distributed equally among all clusters, we conclude that with a slight loss of overall accuracy from 85 % to 83.33 % the interpretability and reliability of the model have been significantly increased. In future studies, a user evaluation of the presented model and especially the methods to increase interpretability are of particular interest. For the future, another user evaluation using the new explainable Model 3 is of high interest, to evaluate the usability of the explainability and the trustworthiness of our privacy setting prediction approach. Such an evaluation could also test different ways of displaying information and provide more insights into the

required depths for increasing transparency and explainability of our presented approach. A further question to answer in future is how a direct user feedback on the user-specific privacy setting prediction, e.g. the correction of some of the predicted feature values by the user, can be used to improve the ML algorithm. Such improvements can include keeping the algorithm up to date or extending the decision tree for groups of instances that were underrepresented in the training data.

HUMAN RESEARCH DISCLOSURE

This research project will not be subject to the Ethics Committee's obligation to assess pursuant to section 4, subsection 1, sentence 2 of the joint Ethics Committee's rules of procedure [14]. The project has been classified as ethically acceptable.

REFERENCES

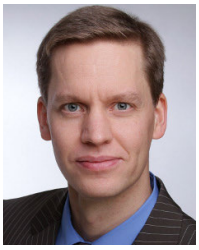
[1] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Secur. Privacy Mag.*, vol. 3, no. 1, pp. 26–33, Jan. 2005.
[2] M. Backes, G. Karjoth, W. Bagga, and M. Schunter, "Efficient comparison of enterprise privacy policies," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2004, pp. 375–382.
[3] K. Bekara, Y. B. Mustapha, and M. Laurent, "XPACML extensible privacy access control markup Langua," in *Proc. 2nd Int. Conf. Commun. Netw.*, Nov. 2010, pp. 1–5.
[4] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 648–657.
[5] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. IJCAI Workshop Explainable AI (XAI)*, vol. 8, 2017, pp. 1–6.
[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[7] H. N. Boone and D. A. Boone, "Analyzing Likert data," *J. Extension*, vol. 50, no. 2, pp. 1–5, 2012.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[9] J. Choo, S. Bohn, and H. Park, "Two-stage framework for visualization of clustered high dimensional data," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, Oct. 2009, pp. 67–74.
[10] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge, "Location disclosure to social relations: Why, when, & what people want to share," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2005, pp. 81–90.
[11] L. F. Cranor, "P3P: Making privacy policies more useful," *IEEE Secur. Privacy*, vol. 1, no. 6, pp. 50–55, Nov. 2003.
[12] A. Dehghantanha, N. I. Udzir, and R. Mahmod, "Towards a pervasive formal privacy language," in *Proc. IEEE 24th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Apr. 2010, pp. 1085–1091.
[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[14] *Ethics Committee of the Faculty of Economics and Business of Goethe University Frankfurt (GU) and the Gutenberg School of Management & Economics of the Faculty of Law*. Accessed: Mar. 18, 2021. [Online]. Available: https://www.wiwi.uni-frankfurt.de/en/research/ethics-committee.html
[15] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 20105, pp. 351–360.
[16] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.
[17] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 80–89.
[18] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Oct. 2017.
[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Jan. 2019.
[20] P. Hall and N. Gill, *An Introduction to Machine Learning Interpretability*. Sebastopol, CA, USA: O'Reilly Media, 2018.
[21] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[22] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decis. Support Syst.*, vol. 51, no. 1, pp. 141–154, Apr. 2011.
[23] J. D. Kelleher, B. M. Namee, and A. D'arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA, USA: MIT Press, 2015.
[24] Z. C. Lipton, "The mythos of model interpretability," 2016, *arXiv:1606.03490*. [Online]. Available: http://arxiv.org/abs/1606.03490
[25] M. Lovric, *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer-Verlag, 2011.
[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[27] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[28] M. Madejski, M. Johnson, and S. M. Bellovin, "A study of privacy settings errors in an online social network," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops*, Mar. 2012, pp. 340–345.
[29] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.
[30] T. Nakamura, S. Kiyomoto, W. B. Tesfay, and J. Serna, "Easing the burden of setting privacy preferences: A machine learning approach," in *Proc. Int. Conf. Inf. Syst. Secur. Privacy*. Cham, Switzerland: Springer, 2016, pp. 44–63.
[31] T. Nakamura, S. Kiyomoto, W. B. Tesfay, and J. Serna, "Personalised privacy by default preferences-experiment and analysis," in *Proc. ICISSP*, 2016, pp. 53–62.
[32] T. Nakamura, W. B. Tesfay, S. Kiyomoto, and J. Serna, "Default privacy setting prediction by grouping user's attributes and settings preferences," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Cham, Switzerland: Springer, 2017, pp. 107–123.
[33] T. Nakamura, A. A. Adams, K. Murata, S. Kiyomoto, and N. Suzuki, "The effects of nudging a privacy setting suggestion algorithm's outputs on user acceptability," *J. Inf. Process.*, vol. 27, pp. 787–801, 2019.
[34] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[36] D. Petkovic, R. Altman, M. Wong, and A. Vigil, "Improving the explainability of random forest classifier–user centered approach," in *Proc. Pacific Symp. Biocomput. Pacific Symp. Biocomput.*, vol. 23. Singapore: World Scientific, 2018, pp. 204–215.
[37] I. Pollach, "What's wrong with online privacy policies?" *Commun. ACM*, vol. 50, no. 9, pp. 103–108, Sep. 2007.
[38] D. Powers, "Evaluation: From precision, recall and F-factor to roc, informedness, markedness & correlation," *Mach. Learn. Technol.*, vol. 2, pp. 2229–3981, Jan. 2008.
[39] M. Prensky, "Digital natives, digital immigrants Part 2: Do they really think differently?" *Horizon*, vol. 9, no. 5, pp. 1–6, 2001.
[40] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*. [Online]. Available: http://arxiv.org/abs/1811.12808
[41] N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao, "Understanding and capturing people's privacy policies in a mobile social networking application," *Pers. Ubiquitous Comput.*, vol. 13, no. 6, pp. 401–412, Aug. 2009.
[42] D. Sculley, "Web-scale k-means clustering," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 1177–1178.

[43] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 56–67.
[44] D. J. Solove, "Introduction: Privacy self-management and the consent dilemma," *Harvard Law Rev.*, vol. 126, no. 7, p. 1880, 2012.
[45] I. A. Tondel, Å. A. Nyre, and K. Bernsmed, "Learning privacy preferences," in *Proc. 6th Int. Conf. Availability, Rel. Secur.*, Aug. 2011, pp. 621–626.
[46] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.
[47] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, Mar. 1963.
[48] *The Platform for Privacy Preferences 1.0 (P3P1. 0) Specification*, World Wide Web Consortium, Cambridge, MA, USA, 2002.

SASCHA LÖBNER received the B.Sc. degree in economics and business administration and the M.Sc. degree in business informatics from Goethe University Frankfurt. During his master's degree, he specialized in machine learning, distributed systems, and high performance computer applications. He is currently a Research and Teaching Assistant with the Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt. He is working in the field of privacy preserving machine learning and especially federated learning.

WELDERUFAEL B. TESFAY received the B.Sc. degree in computer science and engineering from the Mekelle Institute of Technology (MIT), Ethiopia, the M.Sc. degree in computer science and engineering with specialization in mobile systems from the Luleå University of Technology, and the Ph.D. degree in computer science *(summa cum laude)* from Goethe University Frankfurt. He is currently a Senior Researcher with the Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt. His research interests include information privacy, data protection regulations, and applied machine learning. He was a recipient of the Best Demo Award of The Web Conference 2018.

TORU NAKAMURA received the B.E., M.E., and Ph.D. degrees from Kyushu University, in 2006, 2008, and 2011, respectively. In 2011, he joined KDDI, and he moved to KDDI R&D Laboratories, Inc., (currently renamed KDDI Research, Inc.), in the same year. In 2018, he moved to the Advanced Telecommunications Research Institute International (ATR). Since 2020, he has been a Researcher at KDDI Research, Inc., where he is currently a Research Engineer with the Information Security Laboratory. His current research interests include security, privacy, and trust, especially privacy enhanced technology and analysis of privacy attitudes. He is a member of IEICE and IPSJ. He received the CSS2016 SPT Best Paper Award.



SEBASTIAN PAPE received the Diploma degrees in mathematics (Dipl.-Math.) and computer science (Dipl.-Inform.) from the Darmstadt University of Technology and the Dr. rer. nat. degree from the University of Kassel. From 2005 to 2011, he worked as a Research and Teaching Assistant with the Database Group (lead by Prof. Dr. Lutz Wegner), University of Kassel. From 2011 to 2015, he was a Senior Researcher and Teaching Assistant at the Software Engineering for Critical Systems Group (lead by Prof. Dr. Jan Jürjens), TU Dortmund University. From October 2014 to January 2015, he was a Visiting Researcher (of Prof. Dr. Fabio Massacci) at the Security Group, University of Trento. From October 2018 to August 2019, he was standing in as a Professor for business informatics at the University of Regensburg. He is currently a Senior Researcher at the Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt.