Machine Learning Explainability Via Microaggregation and Shallow Decision Trees

Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez and David Sánchez

Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, UNESCO Chair in Data Privacy, Av. Països Catalans 26, 43007 Tarragona, Catalonia

Abstract

Artificial intelligence (AI) is being deployed in missions that are increasingly critical for human life. To build trust in AI and avoid an algorithm-based authoritarian society, automated decisions should be explainable. This is not only a right of citizens, enshrined for example in the European General Data Protection Regulation, but a desirable goal for engineers, who want to know whether the decision algorithms are capturing the relevant features. For explainability to be scalable, it should be possible to derive explanations in a systematic way. A common approach is to use simpler, more intuitive decision algorithms to build a surrogate model of the black-box model (for example a deep learning algorithm) used to make a decision. Yet, there is a risk that the surrogate model is too large for it to be really comprehensible to humans. We focus on explaining black-box models by using decision trees of *limited depth* as a surrogate model. Specifically, we propose an approach based on microaggregation to achieve a trade-off between the *comprehensibility* and the *representativeness* of the surrogate model on the one side and the *privacy* of the subjects used for training the black-box model on the other side.

Keywords: Explainability, Machine learning, Data protection, Microaggregation, Privacy

1. Introduction

A breakthrough in artificial intelligence (AI) has occurred in the last decades with the appearance of deep learning and, specifically, deep neural networks. This forward jump has been unleashed by the increase of computing power and the availability of big data to train machine learning algorithms. Whereas deep

Email address: {alberto.blanco, josep.domingo, sergio.martinezl, david.sanchez}@urv.cat (Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez and David Sánchez)

learning algorithms can make very accurate decisions, they are black boxes in the sense that it is infeasible for humans to understand how these decisions have been reached. This is problematic for both the engineers who train the algorithms and the citizens who are affected by their decisions:

- Engineers would like to understand how the algorithm makes the decisions, in order to make sure it takes the relevant features into account. It may well occur that wrong features are used, such as in the well-known example of [25] in which an animal with a snow background was classified as a wolf, whereas an animal with a grass background was classified as a husky dog, because in the training pictures all wolves were in a snowy landscape and huskies were not.

- Citizens are affected by an increasing number of automated decisions: credit granting, insurance premiums, diagnoses, etc. While transparency measures are being implemented by public administrations worldwide, there is a risk of automated decisions becoming an omnipresent black box. This could result in formally transparent democracies operating in practice as computerized totalitarian societies. To protect citizens, explainability requirements are starting to appear in legal regulations and ethics guidelines. For example, article 22 of the EU General Data Protection Regulation (GDPR, [12]) states the right of citizens to an explanation on automated decisions. Also, the European Commission's Ethics Guidelines for Trustworthy AI [11] urge organizations making automated decisions to be ready to explain them on request of the affected citizens, whom we will call also subjects in what follows. On the other hand, IEEE has issued a report on ethically aligned design for intelligent systems [17]. Yet further, regulatory proposals for AI are ongoing at the national level in several EU countries [2].

Since it will be increasingly necessary to provide explanations, their generation must be automated. It cannot be assumed that a human auditor will be available to explain every automated decision to the affected subject. Older machine learning models, based on rules, decision trees or linear models, are good candidates to yield automated explanations; the reason is that such models are understandable by humans and are thus self-explanatory [15] as long as they are not very large (*i.e.* as long as the number of rules, the size of the decision trees or the number of explanatory attributes stay small). However, the appearance of deep learning has worsened matters: it is much easier to program an artificial neural network and train it than to understand why it yields a certain output for a certain input.

1.1. Contribution and plan of this paper

A usual strategy to generate explanations for decisions made by a black-box machine learning model, such as a deep learning model, is to build a surrogate model based on more expressive machine learning algorithms, such as the aforementioned decision rules [26, 36], decision trees [29], or linear models [32]. The

surrogate model can be trained either on the same data set as the black-box model to be explained or on new data points classified by that same model. Global surrogate models explain decisions on points on the whole domain, while local surrogate models build explanations that are relevant for a single point or a small region of the domain.

We present an approach that assumes that the party generating the explanations has unrestricted access to the black-box model and to a training data set, which can be smaller than the entire data used to train the black-box model (a representative sample suffices). We will take as surrogate models decision trees trained on disjoint subsets of the training data set.

Even though decision trees have already been employed as surrogate models [29], the originality of our proposal lies in the novel combination of microaggregation —by which we obtain *representative* trees while ensuring the *privacy* of the subjects used for training—, shallow decision trees —whose limited depth enhances *comprehensibility*— and ontologies —which we use to manage categorical attributes in a semantically consistent way. By using microaggregation to create the subsets on which surrogate models are trained, we make sure these subsets are of size at least k. This allows preserving the privacy of subjects in the training data set when the surrogate models are released. Also, the larger k, the less local and the more representative are the explanations obtained on the subsets. Yet the main emphasis is on the comprehensibility of the surrogate models, *that we measure as the inverse of the depth of the trained decision trees*. In general, the shallower a decision tree, the easier it is to comprehend it because fewer features are needed to make a decision on a subject.

Section 2 characterizes the type of explanations we seek to generate and the risks of generating them through straightforward release of surrogate models. Section 3 describes our microaggregation-based approach to generate explanations of limited size and, more precisely, trees of limited depth. Section 4 describes how to treat categorical attributes by using ontologies. The use of shallow decision trees to improve comprehensibility is discussed in Section 5. Experimental results are provided in Section 6. Finally, Section 7 gathers conclusions and future research directions.

A preliminary and short version of this work appeared in the conference paper [5]. The results on shallow trees and categorical data are new to the present article. Specifically, Section 1 has been substantially rewritten; furthermore, Section 4, Section 5, and the entire empirical Section 6 are new.

2. Explanations via surrogate models

2.1. Machine learning explanations

According to [20], an explanation for a black-box machine learning model should take into account the following properties:

- **Accuracy.** This property refers to how well an explanation predicts unseen data. Low explanation accuracy can be fine only if the black-box model to be explained is also inaccurate.

- **Fidelity.** The explanations ought to be close to the predictions of the explained model. Accuracy and fidelity are very related: if the black-box model is very accurate and the explanation has high fidelity, then the explanation has also high accuracy.
- **Consistency.** Explanations should apply equally well to any model trained on the same data set.
- **Stability.** When providing explanations to particular instances, similar instances should produce similar explanations.
- **Representativeness.** A highly representative explanation is one that can be applied to several decisions on several instances.
- **Certainty.** If the model at study provides a measure of confidence on its decisions, an explanation of this decision should reflect this.
- **Novelty.** This property refers to the capability of the explanation mechanism to cover instances far from the training domain.
- **Degree of importance.** The explanation should pinpoint the important features.
- **Comprehensibility.** Explanations should be understandable to humans. This depends on the target audience and has psychological and social implications, although short explanations generally go a long way towards comprehensibility.

Miller analyzes explainability from the social sciences perspective [19] and makes four important observations: i) people prefer *contrastive* explanations, *i.e.* why the algorithm took a certain decision does not matter as much to us as why it did not take a different decision instead; ii) people *select* only a few causes from the many causes that make up an explanation, and personal biases guide this selection; iii) referring to probabilities or statistical connections is not as effective as referring to causes; and iv) explanations are *social*, and thus should be part of a wider conversation, or an interaction between the explainer and the explainee.

In [16], the authors emphasize the importance of human field experts guiding the development of explanation mechanisms, given that current machine learning systems work on a statistical and/or model-free mode, and require context from human/scientific models to convey convincing explanations (especially for other field experts).

No single explanation model in the current literature is able to satisfy all the above properties (refer to [20, 4, 15] for extensive surveys on explainable artificial intelligence techniques). In what follows we will focus on accuracy, fidelity, stability, representativeness and comprehensibility, to which we will add privacy. See Section 2.2.2 about the privacy risks of explanations.

2.2. Risks of surrogate model release

A common strategy to provide explanations satisfying the above properties is via a surrogate model based on intrinsically interpretable algorithms. However, care must be exercised to ensure that the surrogate model does not violate trade secret, privacy and explainability.

2.2.1. Trade secret risks

A very detailed surrogate model may reveal properties of the data set that was used to train the black-box model. This may be in conflict with *trade secret*. Indeed, training data are often the result of long-term corporate experience and reflect successes and failures. It takes time to accumulate good training data. Thus, organizations owning such data regard them as a valuable asset they do not want disclosed to competitors.

At the same time, too much detail in the released surrogate model may reveal more about the black-box model to be explained than its owner is willing to disclose. Training complex models, like for example deep models, requires a costly process involving time and computing power. Hence, a well-trained black-box model is also a highly valued asset that organizations view as a trade secret.

2.2.2. Privacy risks

If the released surrogate model leaks information on the training data and these contain personally-identifiable information, then we have a conflict with privacy legislation [12].

2.2.3. Comprehensibility risks

A complex surrogate model, even if based on intrinsically interpretable algorithms, may fail to be comprehensible to humans. We illustrate this risk in Figures 1 and 2.

Figure 1 shows a simple data set with two continuous attributes, represented by the two dimensions of the graph, and a binary class attribute, represented by the color of points in the graph. Thus, points represent the records in the data set. Figure 2 shows a surrogate model consisting of a decision tree trained on the example data set. With 303 nodes and depth 19, this model is not very useful as an explanation to humans: it is very hard to comprehend it.

3. Microaggregation-based surrogate models

To avert the risks identified in Section 2.2 while achieving as many of the properties listed in Section 2.1 as possible, we need a method to construct surrogate models that keep at bay leakage and complexity. To that end, we propose to provide subjects with partial or local explanations, that cover an area of the original training data set close to the subject (that is, attribute values similar to the subject's). Algorithm 1 describes a procedure for the owner of the training data and the black-box model to generate cluster-based explanations.

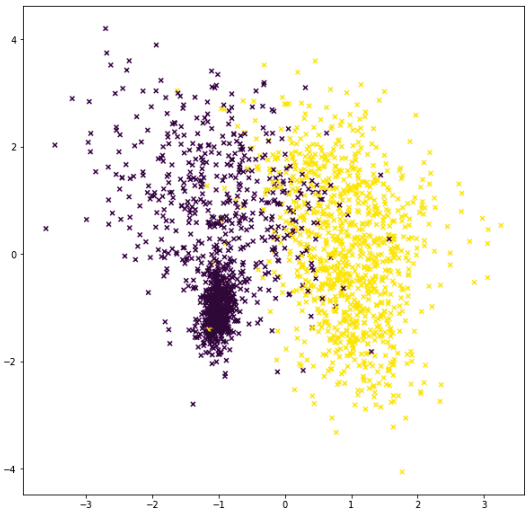


Figure 1: Example data set

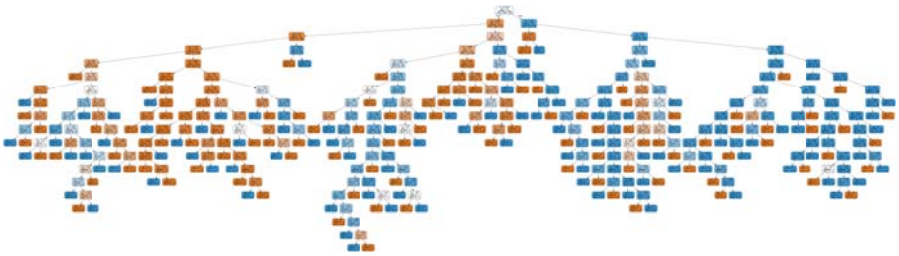


Figure 2: Decision tree trained on the example data set

Then, Protocol 1 shows how a subject obtains an explanation close to her. The fact that explanations are cluster-based favors *stability*: all instances in the cluster are similar and they are explained by the same interpretable model, so that a subject will obtain explanations that were created for subjects similar to her.

Algorithm 1: Generation of cluster-based explanations

Input: Training data set $\mathbf{X}$

Compute a clustering $C(\mathbf{X})$ for $\mathbf{X}$ based on all attributes except the class attribute

for *each cluster* $C_i \in C(\mathbf{X})$ **do**

  Compute a representative, *e.g.* the centroid or average record $\tilde{c}_i$

end

for *each cluster* $C_i \in C(\mathbf{X})$ **do**

  Train an interpretable model, such as a decision tree DT_i

end

Protocol 1 (Provision of explanations).

1. *A subject submits a query* $\hat{x}$ *to the black-box model.*
2. *The black-box model returns to the subject:*
   (a) *A decision* $d = f(\hat{x})$*;*
   (b) *The closest representative* $\tilde{c}_x = \arg\min_{\tilde{c}_i} \mathrm{dist}(\tilde{c}_i, \hat{x})$ *for some distance* dist*;*
   (c) *The interpretable model* DT_x *corresponding to the cluster represented by* $\tilde{c}_x$*.*

A shortcoming of Protocol 1 is that the decision output by the interpretable model DT_x on input $\hat{x}$ may not match the decision $d = f(\hat{x})$ made by the black-box model. This is bad for fidelity, but can be fixed by returning the closest representative to $\hat{x}$ *whose decision tree yields* d. In this way, the explanation provision is *guided* by the black-box model. The search for a valid representative is restricted by a parameter N: if none of the decision trees associated with the N representatives closest to $\hat{x}$ matches the decision of the black-box model, the closest representative is returned. Whereas this may hurt the fidelity of the explanations, returning an arbitrarily distant cluster representative would be of little explanatory power. The guided provision is formalized in Protocol 2.

Protocol 2 (Guided provision of explanations).

1. *A subject submits a query* $\hat{x}$ *to the black-box model.*
2. *The black-box model owner does:*
   (a) *Compute the decision* $d = f(\hat{x})$ *using the black-box model;*
   (b) *Sort the list of cluster representatives by their distance to* $\hat{x}$*, so that* $\tilde{c}_i$ *is the* i*-th closest representative;*
   (c) **let** $i = 1$*;*
   (d) **let** $found = 0$*;*

(e) **repeat**
   i. **let** DT_i *be the interpretable model corresponding to the cluster represented by* $\tilde{c}_i$*;*
   ii. **if** $DT_i(\hat{x}) = d$ **then** $found = 1$ **else** $i = i + 1$*;*
   until $found = 1$ *or* $i > N$*;*
(f) **if** $found = 1$ **then return** d*,* $\tilde{c}_i$ *and* DT_i
   else return d*,* $\tilde{c}_1$ *and* DT_1*.*

We choose microaggregation [7, 10, 21] as the type of clustering in Algorithm 1, because it allows enforcing that clusters consist of at least a minimum number k of records. This minimum cardinality allows trading off *privacy* and *representativeness* for *comprehensibility* of explanations:

- Parameter k ensures that returning the representative $\tilde{c}_x$ in Protocol 1 is compatible with k-anonymity [27, 10] for the subjects in the training data set. Indeed, the representative equally represents k subjects in the training data set. In this respect, the larger k, the more privacy.
- Additionally, large values of k result in clusters that contain larger parts of the domain, thus yielding explanations with higher representativeness.
- While choosing large values for k has a positive effect on privacy and representativeness, it does so at the expense of comprehensibility. A small k results in very local explanations, which in principle should imply simpler and thus more comprehensible surrogate models.

Specifically, we compute microaggregation clusters using MDAV (Mean Distance to Average Vector), a well-known microaggregation heuristic [10]. We recall it in Algorithm 2.

Algorithm 2: MDAV

Input: $\mathbf{X}$, k
Output: $\mathbf{C}$: set of clusters

$\mathbf{C} \leftarrow \emptyset$
while $|\mathbf{X}| \geq 3k$ **do**
  $x_c \leftarrow$ **mean_record**$(\mathbf{X})$
  $x_r \leftarrow \text{argmax}_{x_i}$ **distance**(x_i, x_c)
  $x_s \leftarrow \text{argmax}_{x_i}$ **distance**(x_i, x_r)
  $C_r \leftarrow$ **cluster**$(x_r, k, \mathbf{X})$ `// Algorithm 3`
  $C_s \leftarrow$ **cluster**$(x_s, k, \mathbf{X})$
  $\mathbf{C} \leftarrow \mathbf{C} \cup \{C_r, C_s\}$
  $\mathbf{X} \leftarrow \mathbf{X} \setminus C_r \setminus C_s$
end
if $2k \leq |\mathbf{X}| < 3k$ **then**
  $x_c \leftarrow$ **mean_record**$(\mathbf{X})$
  $x_r \leftarrow \text{argmax}_{x_i}$ **distance**(x_i, x_c)
  $C_r \leftarrow$ **cluster**$(x_r, k, \mathbf{X})$
  $\mathbf{C} \leftarrow \mathbf{C} \cup \{C_r\}$
  $\mathbf{X} \leftarrow \mathbf{X} \setminus C_r$
else
  $\mathbf{C} \leftarrow \mathbf{C} \cup \{\mathbf{X}\}$
end
return $\mathbf{C}$

Algorithm 3: cluster

Input: x, k, $\mathbf{X}$
Output: C: cluster

$C \leftarrow \{x\}$
while $|C| < k$ **do**
  $x_i \leftarrow \text{argmin}_{x_i}$ **distance**(x_i, x)
  $C \leftarrow C \cup \{x_i\}$
  $\mathbf{X} \leftarrow \mathbf{X} \setminus \{x_i\}$
end
return C

Figure 3 depicts the representatives (centroids) of clusters computed by MDAV with $k = 200$ on the example data set of Figure 1. The figure also shows the decision trees that are obtained as explanations for three of the clusters.

The computational complexity of Algorithm 1 is dominated by the cost of microaggregation. Although optimal microaggregation (that is, yielding maximally homogeneous clusters) is NP-hard [23], most microaggregation heuristics take time $O(n^2)$ in the number n of records of the data set [10, 9] and a recent heuristic takes linear time $O(n)$ in most instances [30]. Furthermore, blocking strategies splitting the data set into smaller blocks to be microaggregated

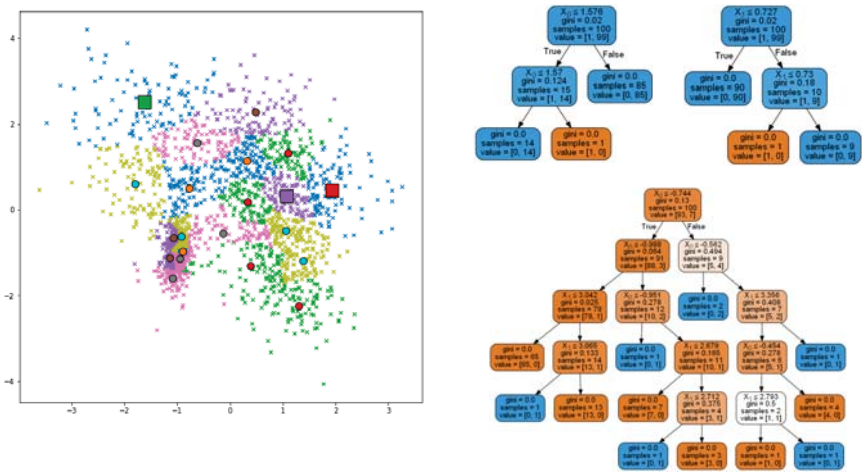


Figure 3: Left, clusters produced by MDAV with $k = 200$ for the data set of Figure 1; for each cluster, points are in a different color and the centroid is depicted. Right, decision tree-based explanations generated for the three clusters whose centroids have been marked with □ symbols on the left figure.

separately can be used to make $O(n^2)$ practical for large n. Aside from microaggregation, computing cluster representatives and training a decision tree as interpretable model within each cluster take time linear in the number of clusters, that is $O(n/k)$; for each cluster training the model takes time $O(k + m)$, where k is the size of the cluster, m the number of attributes and both k and m are typically much smaller than n. Thus the overall complexity of Algorithm 1 is $O(n^2 + n/k)$ and with the heuristic [30] it can be as low as $O(n)$ for most instances.

Regarding Protocol 1, it has cost $O(n/k)$ linear in the number of clusters (finding the representative closest to the subject), because the cost of computing the black-box decision can be assumed constant. Finally, Protocol 2 is dominated by the cost of sorting cluster representatives by their distance to the subject, which takes time quasi-linear in the number of clusters, that is $O((n/k)\log(n/k))$.

Note that, whereas Algorithm 1 is more costly than Protocols 1 and 2, it needs to be run only once and it can be precomputed before any explanations are interactively provided by the protocols.

4. Ontology-based semantic treatment of categorical attributes

Accurate data analyses, including machine learning processes such as data clustering or neural network modeling, require that attribute values be accurately compared and aggregated. Whereas numerical attributes can be treated

through standard arithmetical operators (for example, the arithmetical difference to compare values or the mean to aggregate values), very often attributes contained in data sets are categorical. This is the case of most demographic attributes, such as occupation, marital status or education. Unfortunately, finding analogs of arithmetical operators for categorical attributes is not straightforward.

As stated in [34], retaining the semantics of categorical attributes plays an important role to enable useful and accurate data analyses. A standard way to capture, model and exploit the semantics underlying categorical data is by means of *ontologies*. An ontology is a machine-readable, structured knowledge source that explicitly and consensually represents the concepts in a domain and their semantic interrelations [14]. Ontologies have been successfully applied in a wide range of fields, including medicine, energy, transportation and military information, as they allow comparing concepts by measuring the resemblance between their meanings, rather than just their lexicographic or distributional similarities.

Semantic data analyses require a suitable ontology for each attribute in the data set. In this way, the values of the attributes can be matched to concepts in the ontology, and the semantic relationships modeled in the ontology can be exploited to compare, aggregate and transform categorical values in a semantically consistent manner.

Ontologies suitable for attribute domains can be either extracted from existing knowledge sources or created from scratch. In the former case, large knowledge bases such as WordNet [13] (a structured thesaurus containing more than 150,000 general concepts) or SNOMED-CT [31] (the largest medical knowledge base currently available) can be employed. In the latter case, ontologies can be created by generalizing and taxonomically structuring the concepts in the attribute domain. See in Figure 4 an example of an ontology created for an attribute *Education* that specifies the education level of a subject.

In the sequel we detail how ontologies can be used to measure the semantic distance between attribute values in order to compare and aggregate them.

4.1. Ontology-based semantic distance

Since no arithmetic functions can be applied to categorical values, alternatives are needed to calculate the distances between them. The straightforward option of using Boolean equality/inequality operators [35] presents some limitations that may degrade utility. Under this approach, two concepts are said to be at distance 0 if they are equal and at distance 1 otherwise, thereby placing closely related (but not equal) concepts at the same distance as radically different concepts. Other approaches, such as measuring the lexicographic resemblance between categorical values or their distribution poorly capture the semantics underlying categorical data.

A better option to retain the utility of concepts for data analytics is to calculate the *semantic distance* between them, which is a function $distance : c_1 \times c_2 \rightarrow \mathbb{R}$ mapping a pair of concepts (corresponding to categorical attribute

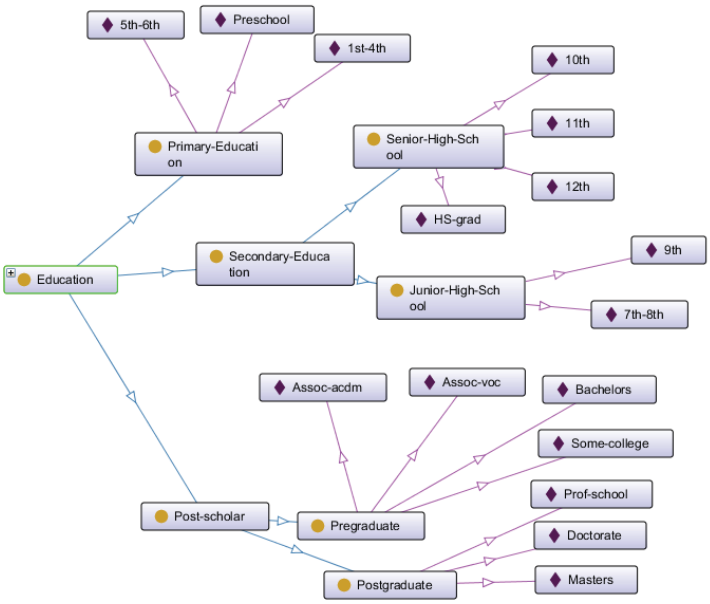


Figure 4: Example of an ontology describing the taxonomic structure of the domain of the categorical attribute *Education*

values) to a real number that quantifies the distance between their meanings according to the semantic evidences gathered from an ontology.

Among the variety of ontology-based semantic distance measures proposed in the literature [3], the measure proposed in [28] stands out as achieving state-of-the-art accuracy and being efficient to calculate. In addition, this measure fulfills the non-negativity, reflexivity, symmetry and triangle inequality properties, thereby defining a coherent total order within the categories of an attribute. Due to these reasons, we have chosen it to measure the semantic distance of attribute values in data analyses.

The authors of [28] compute the distance $X \times X \to \mathbb{R}$ between two values v_1 and v_2 of a categorical attribute X, whose domain is modeled in the taxonomy $\tau(X)$, as a logarithmic function of their number of non-common taxonomic ancestors normalized by their total number of ancestors:

$$\begin{aligned} &distance(v_1, v_2) \\ &= \log_2 \left(1 + \frac{|\phi(v_1) \cup \phi(v_2)| - |\phi(v_1) \cap \phi(v_2)|}{|\phi(v_1) \cup \phi(v_2)|} \right) \end{aligned} \tag{1}$$

where $\phi(v_i)$ is the set of taxonomic ancestors of v_i in $\tau(X)$, including itself.

We use the above-described distance in Algorithm 2 (MDAV), lines 4, 5 and 13, and Algorithm 3 (clustering), line 3, in order to calculate the distance

between pairs of categorical values. Furthermore, we use it in Protocol 1 (provision of explanations), step 2b, and Protocol 2 (guided provision of explanations), step 2b, in order to calculate cluster representatives and sort them.

4.2. Calculation of cluster representatives

During the clustering process, cluster representatives are calculated by aggregating the attribute values in the cluster. Whereas numerical values can be aggregated arithmetically by using the mean, categorical values cannot. A straightforward alternative is to select the most common value in the cluster (the mode) as the representative [1]. A better approach is to take into account the semantics of concepts in the cluster. To do this, we use the notion of *marginality* [8]. Based on on a semantic distance (like the above-described one), marginality measures how far each value of a categorical attribute in the data set lies from the "center" of the taxonomy that describes the attribute's domain. The marginality $m(\cdot,\cdot)$ of each value v_j of a categorical attribute X with respect to its domain of values $Dom(X)$ is computed as follows:

$$m(Dom(X), v_j) = \sum_{v_l \in Dom(X) - \{v_j\}} distance(v_l, v_j) \tag{2}$$

where $distance(\cdot,\cdot)$ is a semantic distance between two values (see Section 4.1). The smaller $m(Dom(X), v_j)$, the less marginal (that is, the more central) v_j is with regard to $Dom(X)$.

Based on marginality, the representative of a cluster formed by categorical values can be selected as the value belonging to the domain of the categorical attribute that minimizes the sum of distances with respect to *all other values in the data set* [8], that is, the least marginal value in the minimum taxonomy including all values that appear in the data set. Formally, given a sample $S(X)$ of a categorical attribute X in a certain cluster, the marginality-based representative of that cluster is defined as:

$$Representative(S(X)) = \arg \min_{v_i \in \tau(S(X))} m(S(X), v_i) \tag{3}$$

where $\tau(S(X))$ is the minimum taxonomy extracted from $\tau(X)$ that includes all values in $S(X)$. Note that the search space of the cluster representative is extended to all concepts in $\tau(S(X))$, and not only to the values in the cluster. In this manner, the centroid can be selected in a more accurate way from a broader set of candidates.

Specifically, we use the above-described representative calculation in: i) the algorithm for generating cluster-based explanations (Algorithm 1), line 3, to calculate the cluster representative of a cluster of categorical values, ii) the MDAV algorithm (Algorithm 2), lines 3 and 12, to obtain the mean record of the set X by computing the marginality-based representative of the categorical attributes in the data set.

5. On the depth of decision trees

In principle, the fewer the records of the cluster on which a decision tree is trained, the fewer the nodes and the higher the understandability of the tree. Thus, the more local explanations are, the more understandable they can be expected to be.

However, the size of a tree depends on its breadth and its depth, and we argue that depth is more relevant for comprehensibility than breadth:

- The depth of the tree measures how many feature-based decisions must be chained in order to reach the overall decision. Even if decisions are binary (two alternatives), a sequence of, say, 19 decisions like in the tree of Figure 2 is hard to follow by a human. As mentioned above, it is observed in [19] that people focus only on a few explanatory causes.
- The breadth of the tree depends on the number of alternatives per decision. A large number of alternatives may often be understandable if the number of decisions is small. For example, a decision tree containing a single decision that assigns subjects an income tax rate depending on their income interval is easy to understand even if, say, 30 different income intervals are defined.

If we want a decision tree to stay shallow (that is, with low depth) no matter the size k of the cluster, we must enforce a maximum tree depth. Otherwise, larger clusters are likely to result in deeper trees, as illustrated in the experiments of Section 6.

The price paid for limiting the maximum depth of explanatory decision trees is likely to be a loss of accuracy, *unless* the black-box decisions turn out to depend on only a few input features. In that case, shallow trees may perform even better than deep trees. In the experiments of Section 6.6, we show that this is the case for the Adult data set.

6. Empirical work

We have applied the above-described methodology to three data sets: a synthetic numerical data set, a real numerical data set and a real data set with a mix of categorical and numerical attributes. The code to generate the synthetic data set and the ontologies for categorical attributes, as well as to conduct all experiments reported in this section is available in Github[1].

6.1. Experiments on synthetic numerical data with unbounded trees

We generated a data set consisting of 1,000,000 records, each with 10 numeric continuous attributes and a single binary class labeled using the `make_classification` method from Scikit-learn[2]. Out of the 1,000,000 records, we reserved 2/3 to train the models, and the remaining 1/3 to validate them.

[1] Download address: `https://github.com/ablancoj/explainability`

[2] `https://scikit-learn.org/stable/index.html`

We took as a black-box model a neural network denoted by ANN with three hidden layers of 100 neurons each, which achieved 96.6% classification accuracy. We also trained a decision tree on the whole data set, to take its size, depth and accuracy as references. This global decision tree had size 71,617 nodes, depth 56 and its classification accuracy was 91.88%. We expected our local decision trees (trained on a single cluster) to achieve a similar accuracy on average, although it could happen that clusters containing points from a single class would produce more accurate classifiers.

Then, we tested our cluster-based mechanism for different values of k. As stated in Section 3, smaller values of k could be expected to yield *simpler* classifiers. Instead of directly choosing arbitrary values for k, we chose several percentages of the 670,000 records of the training data set that we wanted the clusters to contain, ranging from 0.1% to 30%; this translated to k values ranging from 670 to 201,000.

The experiment was as follows. For each value of k, we used MDAV to obtain a clustering of the training data set. Then we computed the centroid representatives of clusters, and we trained a decision tree for each cluster. After that, we measured the classification accuracy and the fidelity of the explanations. Classification accuracy was computed in the usual manner, with the ground truth being the labels in the evaluation data set (1/3 of the original data set, that is, the 330,000 records not used for training). Fidelity was computed as the classification accuracy with respect to the decisions made by the black-box model.

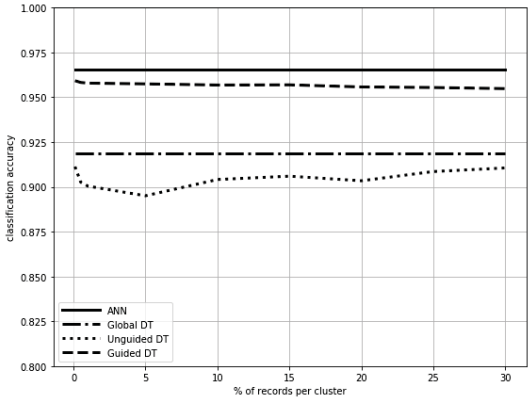
Figure 5a shows the **accuracy** of our local explanations, which for all values of k is lower than the accuracy of the black-box model ANN by around 5% to 6% in the unguided approach (Protocol 1) and by only around 1% in the guided approach (Protocol 2, with $N = 3$). The accuracy of the unguided approach of Protocol 1 is within 2% of the global decision tree mentioned above, with the guided approach of Protocol 2 clearly outperforming both.

In fact, black-box models have largely replaced intrinsically explainable models based on decision trees or rules because the former are more accurate. Even if in Figure 5 the global decision tree is not *very* inferior to the ANN, it is too big and deep a tree to be comprehensible, and thus is nearly as bad as the ANN from the explainability point of view. To obtain shallow and hence comprehensible decision trees whose accuracy comes close to that of the ANN, they must be guided by the ANN.

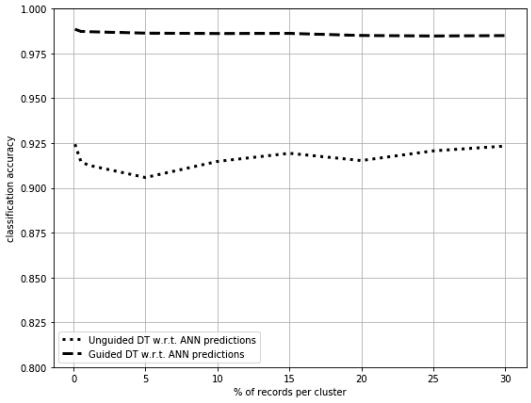
Moreover, it is important to note that accuracy is not very affected by the value of k. This same behavior has been observed for several different generated data sets, hence it cannot be attributed to randomness.

Figure 5b, on the other hand, depicts the **fidelity** of our explanations with respect to the black-box model. For Protocol 1 (unguided approach) the explanations coincide with the black-box model for over 90% of the decisions. When using Protocol 2 (guided by the ANN with $N = 3$), these results improve to up to 98% coincidence, which demonstrates that our method achieves a high accuracy and fidelity with respect to the black-box model.

Figure 6 deals with the **comprehensibility** of the explanations, by depicting



(a) Accuracy



(b) Fidelity

Figure 5: Synthetic numerical data. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$.

the number of nodes of the decision trees trained on the clusters, as a function of the cluster size k: since there is one decision tree per cluster, the box plot represents for each k the median and the upper and lower quartiles of the number of nodes per decision tree. We can see that for the smallest k (representing 0.1% of the training set, in our case $k = 670$), the number of nodes for most decision trees is below 100.

We argue that decision trees with 100 or more nodes are not very useful as explanations of a decision. Since according to Figure 5a k does not significantly affect accuracy, *one should take the smallest k that is deemed sufficient for* **privacy** (explanations are best understood if trees have no more than 10 or 20 nodes).

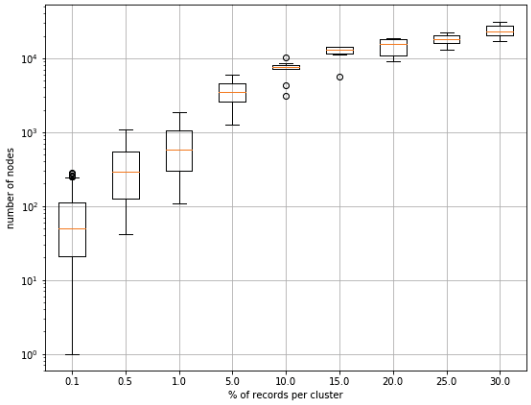


Figure 6: Synthetic numerical data. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees.

6.2. Experiments on synthetic numerical data with shallow trees

We also conducted experiments where the depth of explanatory decision trees was bounded to 4 for the sake of comprehensibility.

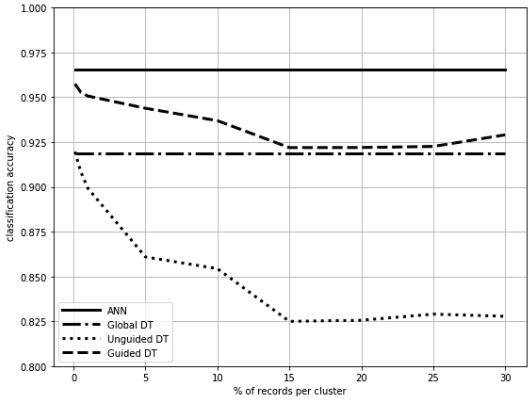
Figure 7a displays the accuracy of local explanations with such shallow trees as a function of the cluster size k. Unlike in Figure 5 (trees with unbounded depth), we can see that the accuracy decreases as the cluster size increases, for both the guided and unguided approaches. The reason is that a shallow tree can explain better a small set of decisions than a large set.

Figure 7b depicts the fidelity of local explanations with trees of depth at most 4 with respect to the black-box model. It can be seen that the fidelity also decreases with the cluster size, for both the guided and unguided approaches. Whereas the black-box model does not depend on the cluster size, shallow trees work best in small clusters. Hence, for the sake of both accuracy and fidelity, one should choose the smallest cluster size k that is viewed as sufficient for the privacy of the subjects in the training data. As in the previous section, we show the accuracy of the global decision tree as a reference, even if this tree is too big and deep from the comprehensibility point of view.

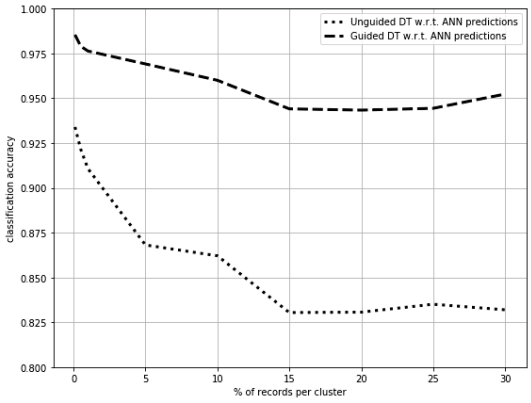
Figure 8 reports the number of nodes of the resulting shallow trees. Unlike for trees with unbounded depth (Figure 6), the number of nodes for shallow trees does not grow beyond 35. Thus, limiting the tree depth also results in a smaller number of nodes, which doubly improves comprehensibility.

6.3. Experiments on the Activity data set with unbounded trees

To experiment with large real numerical data we used the "PAMAP2 Physical Activity Monitoring" data set from the UCI Machine Learning Reposi-



(a) Accuracy



(b) Fidelity

Figure 7: Synthetic numerical data. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$. Decision trees are limited to depth 4.

tory[3]. This data set contains continuous measurements of 3 inertial body sensors (placed on the arm, chest and ankle) and a heart-rate monitor worn by 9 subjects who performed 18 different activities such as walking, cycling, watching TV, etc. First, as recommended by the releasers of the Activity data set [24], we created a data set discarding the transient activity (*e.g.* going to the next activity location). Second, for simplicity, we mapped the various types of activity into two categories indicating whether the activity involved displacement or not (*e.g.* walking and cycling were mapped to "displacement" and watching TV to "not displacement"). As result, we obtained a data set containing 1,942,872 records of which 1,136,540 records were labeled as "displacement" and 806,332

[3]`https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring`

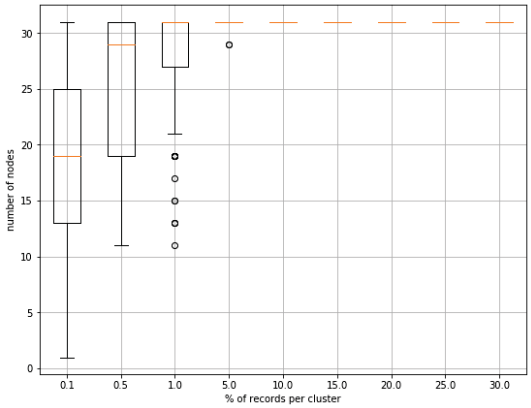


Figure 8: Synthetic numerical data. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees. Decision trees are limited to depth 4.

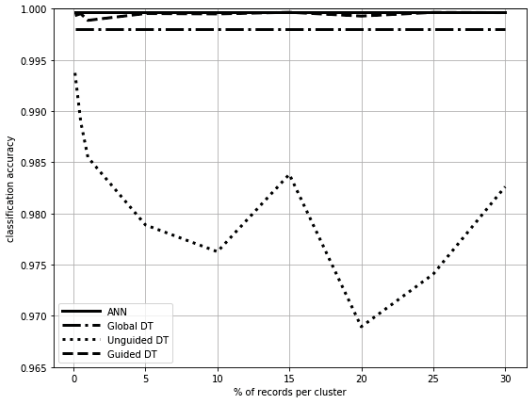
as "not displacement".

Each record contained 54 numerical attributes corresponding to timestamp, label, heart rate and 17 sensor data feeds for each of the 3 inertial sensors. Given an unlabeled record, the classification task in the experiment consisted in deciding whether the subject was performing at that instant an activity involving physical displacement.

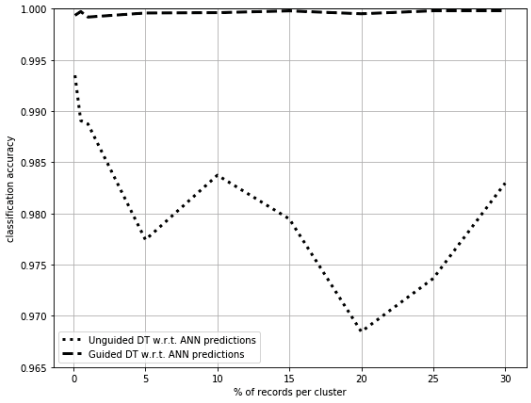
The setup for the experiment was the same as for the synthetic numerical data set, namely training a black-box model and a decision tree on the whole data set. Then, we computed clusters of records via microaggregation for different values of k and trained a decision tree on each cluster. The black-box model obtained over 99.9% accuracy, whereas the decision tree trained on the whole data set achieved 99.7% accuracy and had size 5,335 nodes. Figure 9 shows the accuracy and fidelity for the Activity data set, with the guided approach achieving an accuracy close to that of the black-box model and the unguided approach ranging between 97% and 99.4% accuracy. Figure 10 reports the number of nodes of the decision trees trained on clusters, as a function of the cluster size k.

6.4. Experiments on the Activity data set with shallow trees

The trees obtained on the Activity data set with unbounded depth were not as big as those for the synthetic data set; as an example, the decision tree trained on the whole Activity data set had only 5,335 nodes, compared to 71,617 nodes for the synthetic numerical data set. Nonetheless, unbounded trees were still too big and deep for comprehensibility. Thus, we also tested our method with shallow trees of maximum depth 4. As shown in Figure 11, and similarly to the synthetic numerical data set, the accuracy and fidelity of the obtained trees dropped as the cluster size k increased: larger clusters required deeper trees to



(a) Accuracy



(b) Fidelity

Figure 9: Activity data set. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$.

maintain the same accuracy and fidelity, but we capped depth to 4. The reward for depth capping is shown in Figure 12, where it can be seen that the resulting trees were substantially smaller than those in Figure 10.

6.5. Experiments on categorical data with unbounded trees

To experiment on a real data set containing categorical data we used Adult, a standard data set also hosted in the UCI Machine Learning Repository[4]. Adult contains over 48,842 records of census income information and has 14 attributes reporting both numerical and categorical values. For each categorical attribute we created an ontology describing the concepts in the attribute domain

[4] `https://archive.ics.uci.edu/ml/datasets/adult`

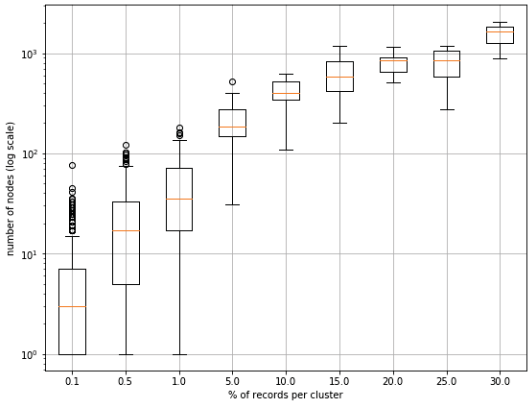


Figure 10: Activity data set. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees.

in a taxonomic way (see Section 4). Ontologies for attributes *Occupation* and *Native-country* were extracted from WordNet. For the rest of the categorical attributes (*i.e.*, *Age*, *Workclass*, *Education*, *Marital-status*, *Relationship*, *Race* and *Sex*) ontologies were manually created.

Figure 13a displays the accuracy of the local explanations. The guided approach was as accurate as the black-box model. The unguided approach did not perform so well: for small clusters the unguided approach did not even match the global decision tree, which it nearly managed to do for larger clusters. As in previous sections, we report the accuracy of the global decision tree as a reference, even if this tree is too big and deep to be comprehensible.

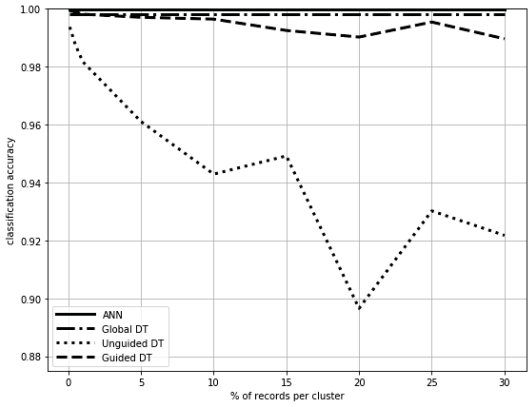
Figure 13b depicts the fidelity with respect to the black-box model. The guided approach achieved fidelities around 0.97 for any cluster size. In contrast, the unguided approach achieved fidelities between 0.76 and 0.83, with the highest figures occurring for larger clusters. Thus, for larger clusters, guiding the explanation provision by the black-box model was somewhat less necessary.

Figure 14 represents the number of nodes of decision trees as the cluster size increases. The fact that larger clusters resulted in larger trees may explain why guidance was less important for larger clusters: as decision trees grew bigger, their accuracy and fidelity improved anyway.

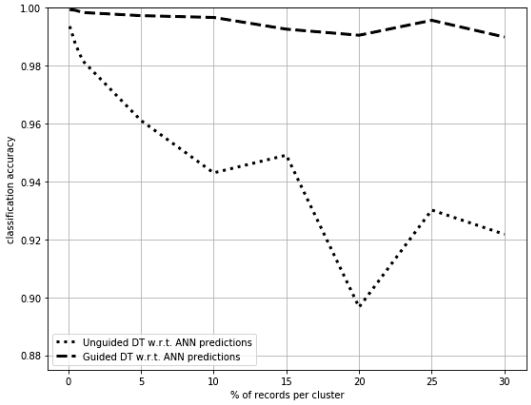
6.6. Experiments on categorical data with shallow trees

When we capped the depth of the decision trees to 4, we obtained the accuracy and fidelity results in Figure 15. As in previous sections, the accuracy of the (non-comprehensible) global decision tree is given as a reference.

Here we observe significant differences with what happened when using shallow trees with the synthetic and the Activity data sets (Figures 7 and 11, respectively). Whereas for the two numerical data sets the accuracy and fidelity



(a) Accuracy



(b) Fidelity

Figure 11: Activity data set. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$. Decision trees are limited to depth 4.

of shallow trees worsened as the cluster size grew, for the Adult data set they improved with the cluster size. This behavior is counterintuitive, because one would expect shallow trees to have more difficulties to explain larger clusters than smaller clusters. In fact, the reason must be sought in the specificity of the Adult data set: in [22] it is shown that decision trees with depths between 2 and 13 achieve practically the same accuracy for the Adult data set, with deeper trees performing slightly worse. This may indicate that the class attribute in Adult depends on a few feature attributes and hence decisions can be reproduced more accurately by shallow decision trees than by deep trees —the latter take too many features into account and actually introduce noise.

Finally, Figure 16 depicts the number of nodes of the shallow decision trees used for the various cluster sizes. Similarly to what occurred for the two numerical data sets, shallow trees for Adult had much fewer nodes than trees with

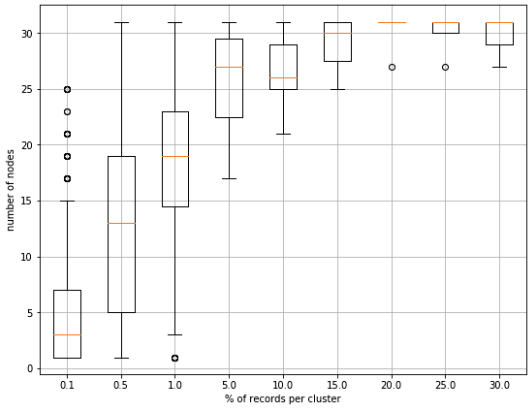


Figure 12: Activity data set. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees. Decision trees are limited to depth 4.

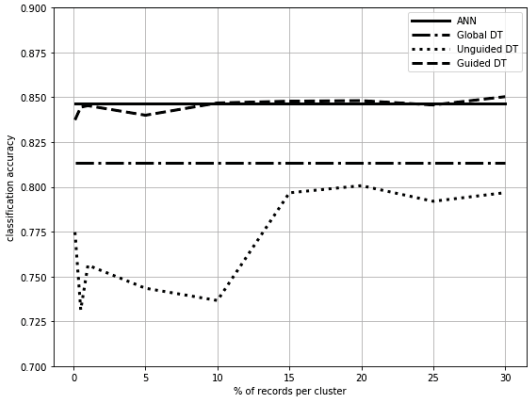
unbounded depth, which definitely improved their comprehensibility.

6.7. Examples of decisions for the Adult data set

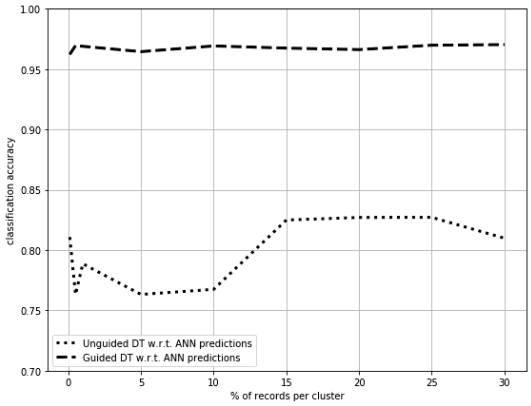
To illustrate the results that can be obtained when querying a decision system based on our mechanism, we sampled 5 individuals from the Adult data set. They are depicted in Table 1a.

Note that individuals 6096 and 13891 belong to the class *over 50K*, while the rest belong to the class *below 50K*. All individuals are white Americans, which reflects the imbalance of the data set w.r.t. race and country of origin. Two of the individuals are Female, while the rest are Male. We configured our mechanism for a representativeness of 10%, which in the Adult training data set means taking $k = 2,020$. We used the guided approach with $N = 3$ and shallow trees of maximum depth 4. The overall results for this configuration have been shown in Section 6.6 above.

Figures 17 to 21 report the obtained decision trees, with the decision paths highlighted. A general inspection of the five trees shows that the attribute *Sex*, which is known to be a highly informative attribute in the Adult data set, does not appear in any of these trees. The reason for this is that this attribute is highly relevant during the clustering procedure, meaning that females are grouped with females and males with males, with very little overlapping; hence, the attribute is irrelevant within each cluster. Consistently with this, Table 1b shows that female instances were associated with female centroids and male instances were associated with male centroids. Another relevant attribute that explains the decisions in the trees is *Marital-status*: being divorced or, in general, not married, increases the chances of belonging to class *below 50K*, especially for males. *Education* and *Workclass* are also relevant features in the trees, as well as *Capital-gain*, *Capital-loss* and *Age*. In contrast, *Native-country* is not very



(a) Accuracy



(b) Fidelity

Figure 13: Adult data set. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$.

relevant, which is in line with the instances of Table 1 being assigned centroids with different *Native-country*.

6.8. Execution times

The execution time of our mechanism is heavily dominated by the microaggregation procedure in Algorithm 2. Fortunately, microaggregation needs to be run *only once* and it can be precomputed before starting to provide explanations. As discussed in Section 2, the MDAV clustering algorithm has complexity quadratic in the size of the data set. The complexity of our implementation is $O(\frac{6n^2}{2k})$. The denominator k is usually omitted, since for typical uses of MDAV, namely enforcing k-anonymity, $k \ll n$ (otherwise, the utility of the anonymized data set is much too degraded). However, we choose k polynomially in n as $k = rn$, being r the *representativeness* of the clusters, that is, the minimum

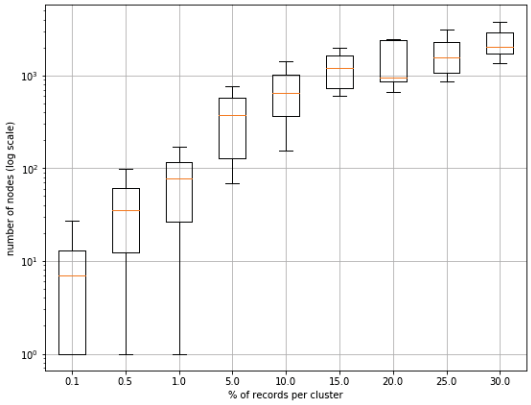


Figure 14: Adult data set. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees.

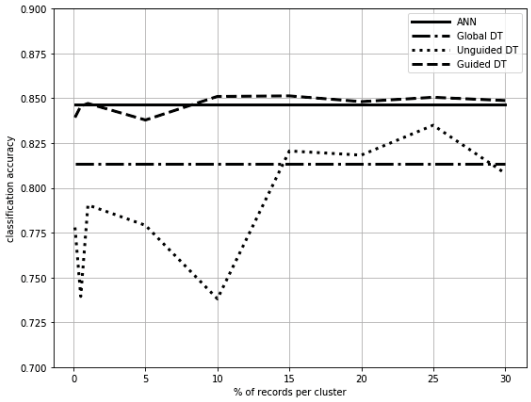
fraction of records of the data set per cluster. The minimum value r can take is $1/n$, resulting in $k = 1$, but in this case no clustering needs to be performed, since clusters are of a single record. Therefore, the minimum value of r is $2/n$ for $k = 2$. The maximum value would be $r = 1$, in which case $k = n$; again, in this case no clustering needs to be performed, since the whole data set would be included in a single cluster. Moreover, for values of r larger than $1/3$, only one cluster is computed (which is a linear procedure in the size of the data set) and, since the remaining records are fewer than $2k$, they are all included in a single cluster. Thus, the values of r that require substantial computation are those in the range $[2/n, 1/3]$. For $r = 2/n$, the MDAV algorithm is quadratic in the size of the data set ($O((3/2)n^2)$). On the other hand, for $r = 1/3$, the complexity is linear in the size of the data set ($O(9n)$).

Figure 22 shows the execution times of the clustering procedure for the three considered data sets and for the relevant range of r.

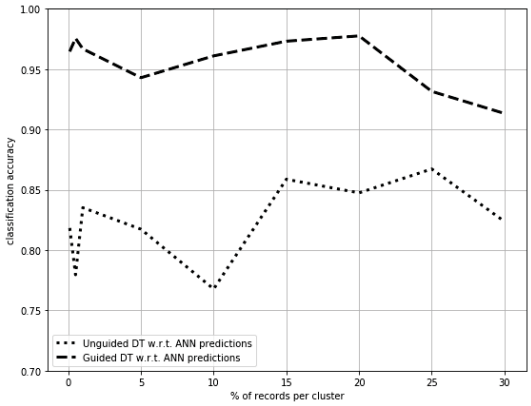
7. Conclusions and future research

We have presented an approach based on microaggregation that allows deriving explanations of machine learning decisions while controlling their accuracy, fidelity, representativeness, comprehensibility and privacy preservation. In addition, being cluster-based, our explanations offer stability by design.

Future research will involve trying different distances in Protocols 1 and 2 and also in the microaggregation algorithm, in order to improve the trade-off between the above properties. Options to be explored include various semantic distances. Another interesting line is to study how the proposed approach could be adapted to explain decisions on high-dimensional data [6], unstructured text or images [18, 33].



(a) Accuracy



(b) Fidelity

Figure 15: Adult data set. Accuracy and fidelity of the decision trees for each cluster size k. For the guided approach we use $N = 3$. Decision trees are limited to depth 4.

On the other hand, in this paper we have assumed that explanations are generated by the owner of the black-box model and the training data set. It is worth investigating the case in which a third party or even the subjects themselves generate the explanations. In this situation the black-box model owner may limit access to her model to protect her trade secrets. We plan to explore ways to generate microaggregation-based explanations that are compatible with such access restrictions. Possible strategies include cooperation between subjects and/or smart contracts between the generator of explanations and the owner of the black-box model.

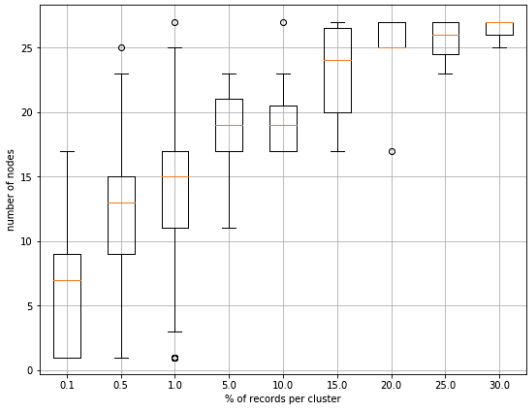


Figure 16: Adult data set. Comprehensibility of explanations: the box plot represents for each cluster size k the median and lower and upper quartiles of the number of nodes of within-cluster decision trees. Decision trees are limited to depth 4.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: European Commission (project H2020-871042 "SoBigData++"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705) and Spanish Government (projects RTI2018-095094-B-C21 "CONSENT" and TIN2016-80250-R "Sec-MCloud"). The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO or any of the funders.

References

[1] Abril, D., Navarro-Arribas, G. and Torra, V, 2010. Towards semantic microaggregation of categorical data for confidential documents. In 7th International Conference on Modeling Decisions for Artificial Intelligence – MDAI 2010, pp. 266-276. Springer.

[2] Access Now, 2018. Mapping Regulatory Proposals for Artificial Intelligence in Europe. `https://www.accessnow.org/cms/assets/uploads/2018/11/mapping_regulatory_proposals_for_AI_in_EU.pdf`

[3] Batet, M. and Sánchez, D., 2015. A review on semantic similarity. In Encyclopedia of Information Science and Technology, Third Edition, pp. 7575-7583. IGI Global.

[4] Biran, O. and Cotton, C., 2017. Explanation and justification in machine learning: a survey. In IJCAI-17 Workshop on Explainable AI – XAI (Vol. 8).

[5] Blanco-Justicia, A. and Domingo-Ferrer, J., 2019. Machine learning explainability through comprehensible decision trees. In IFIP Cross Domain Conference for Machine Learning and Knowledge Extraction – CD-MAKE 2019, pp. 15-26. Springer.

[6] Deng, T., Ye, D., Ma, R., Fujita, H. and Xiong, L., 2020. Low-rank local tangent space embedding for subspace clustering. Information Sciences 508:1-21.

[7] Domingo-Ferrer, J., and Mateo-Sanz, J.M., 2002. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Enginering 14(1):189-201.

[8] Domingo-Ferrer, J., Sánchez and Rufian-Torrell, G., 2013. Anonymization of nominal data based on semantic marginality. Information Sciences 242:35-48.

[9] Domingo-Ferrer, J., Sebé, F., and Solanas, A., 2008. A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications 55(4):714–732.

[10] Domingo-Ferrer, J. and Torra, V., 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2):195-212.

[11] European Comission's High-Level Expert Group on Artificial Intelligence, 2018. Draft Ethics Guidelines for Trustworthy AI.

[12] European Union, 2016. General Data Protection Regulation. Regulation (EU) 2016/679. `https://gdpr-info.eu`

[13] Fellbaum, C. (ed.), 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

[14] Guarino, N. (ed.), 1998. Proc. of the 1st International Conference on Formal Ontology and Information Systems – FOIS 1998. IOS Press.

[15] Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F., 2018. A survey of methods for explaining black box models. ACM Computing Surveys 51(5):art. 93.

[16] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H., 2019. Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. e1312.

[17] The Institute for Electrical and Electronics Engineers, 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. `https://ethicsinaction.ieee.org/#read`

[18] Laiadi, O., Ouamane, A., Benakcha, A., Taleb-Ahmed, A. and Hadid, A., 2019. Learning multi-view deep and shallow features through new discriminative subspace for bi-subject and tri-subject kinship verification. Applied Intelligence 49(11):3894-3098.

[19] Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267:1-38.

[20] Molnar, C., 2018. Interpretable machine learning: A guide for making black box models explainable. Leanpub. Available at `https://christophm.github.io/interpretable-ml-book/`

[21] Mortazavi, R., Jalili, S. and Gohargazi, H., 2013. Multivariate microaggregation by iterative optimization. Applied Intelligence, 39(3):529-544.

[22] Nguyen, D., 2019. Learning data science — Predict Adult income with decision tree. ITNEXT-LINKIT, Mar. 17. `https://itnext.io/learning-data-science-predict-adult-income-with-decision-tree-ae8dd57a76cc`

[23] Oganian, A. and Domingo-Ferrer, J., 2001. On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe 18(4):345–354.

[24] Reiss, A., and Stricker, D., 2012 Introducing a New Benchmarked Dataset for Activity Monitoring. The 16th IEEE International Symposium on Wearable Computers (ISWC), IEEE.

[25] Ribeiro, M. T., Singh, S., and Guestrin, C., 2016. Why should I trust you? Explaining the predictions of any classifier. In 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD'16, pp. 1135-1144. ACM.

[26] Ribeiro, M. T., Singh, S., and Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations. In 32nd AAAI Conf. on Artificial Intelligence–AAAI'18, pp. 1527-1535. AAAI.

[27] Samarati, P., and Sweeney, L., 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report.

[28] Sánchez, D., Batet, M., Isern, D. and Valls, A., 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39(9):7718-7728.

[29] Singh, S., Ribeiro, M. T., and Guestrin, C., 2016. Programs as black-box explanations. arXiv preprint arXiv:1611.07579.

[30] Soria-Comas, J., Domingo-Ferrer, J. and Mulero, R., 2019. Efficient near-optimal variable-size microaggregation. In Modeling Decisions for Artificial Intelligence – MDAI 2019, pp. 333-345. Springer.

[31] Spackman, K.A., Campbell, K.E. and Cote, R.A., 1997. SNOMED RT: a reference terminology for health care. In Proceedings of AMIA Annual Fall Symposium, pp. 640-644.

[32] Strumbelj, E., and Kononenko, I., 2010. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research 11:1–18.

[33] Terhörst, P., Damer, N., Kirchbuchner, F. and Kuijper A., 2019. Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. Applied Intelligence 49(8):3043-3060.

[34] Torra, V., 2010. Towards knowledge intensive data privacy. In Data Privacy Management and Autonomous Spontaneous Security, pp. 1-17. Springer.

[35] Torra, V., 2004. Microaggregation for categorical variables: amedian based approach. Privacy in Statistical Databases – PSD 2004, pp. 162-174. Springer.

[36] Turner, R., 2016. A model explanation system. In IEEE Intl. Workshop on Machine Learning for Signal Processing–MLSP'16, pp. 1-6. IEEE.

Table 1: Five individuals in the Adult data set and their closest representatives

(a) 5 randomly chosen individuals from the Adult data set.

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
6096	0.767123	Self-emp-inc	Masters	0.933333	Divorced	Exec-managerial	Not-in-family	White	Male	0.278283	0.0	0.612245	United-States	1
368	0.616438	Private	Some-college	0.666667	Divorced	Handlers-cleaners	Unmarried	White	Female	0.000000	0.0	0.408163	United-States	0
28817	0.315068	Private	12th	0.533333	Never-married	Craft-repair	Not-in-family	White	Male	0.000000	0.0	0.510204	United-States	0
13891	0.589041	Private	Masters	0.933333	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.076881	0.0	0.540816	United-States	1
29901	1.232877	Private	HS-grad	0.600000	Widowed	Adm-clerical	Unmarried	White	Female	0.000000	0.0	0.255102	United-States	0

(b) Centroids closest to the selected individuals.

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
6096	0.6202	State-gov	Bachelors	0.8556	Never-married	Adm-clerical	Husband	White	Male	0.0591	0.0201	0.4932	Philippines
368	0.4077	Local-gov	Some-college	0.5923	Separated	Adm-clerical	Wife	White	Female	0.0032	0.0056	0.3753	Cuba
28817	0.4454	Local-gov	Assoc-acdm	0.6526	Separated	Adm-clerical	Husband	White	Male	0.0065	0.0131	0.3756	Philippines
13891	0.5269	Local-gov	Some-college	0.7058	Separated	Adm-clerical	Husband	White	Male	0.0063	0.0229	0.4448	Cuba
29901	0.6074	Federal-gov	Bachelors	0.6174	Separated	Adm-clerical	Husband	White	Female	0.0032	0.0322	0.3838	Dominican-Republic

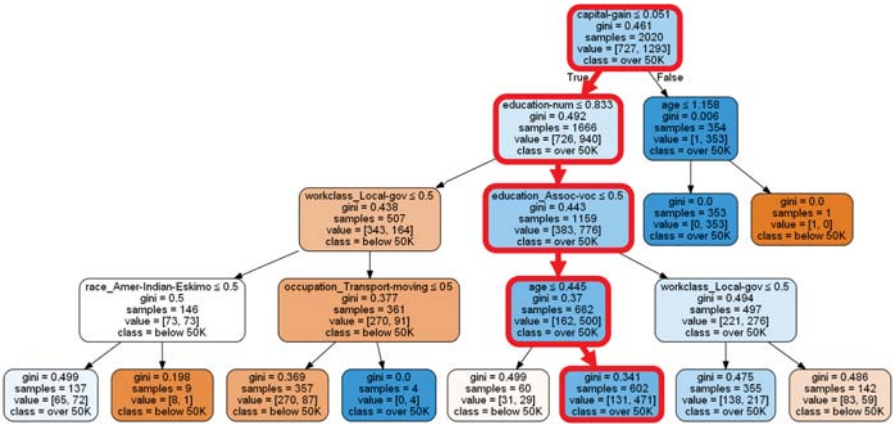


Figure 17: Returned decision trees for query on user 6096: class above 50K. Representativeness set to 10%. Guided approach with $N = 3$ and trees limited to depth 4.

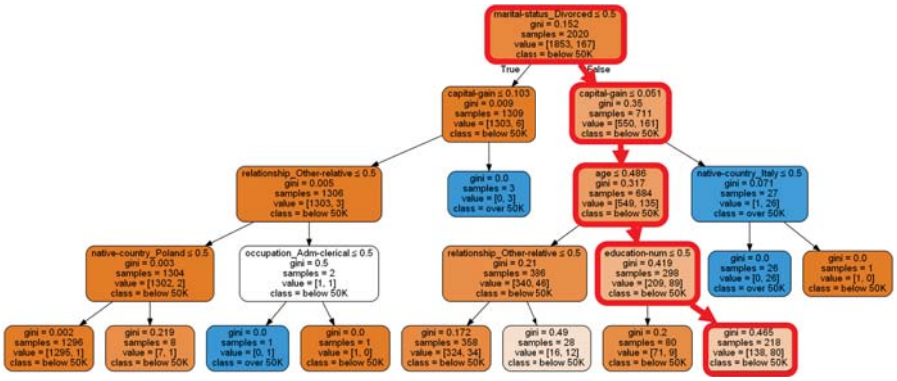


Figure 18: Returned decision trees for query on user 368: class below 50K. Representativeness set to 10%. Guided approach with $N = 3$ and trees limited to depth 4.

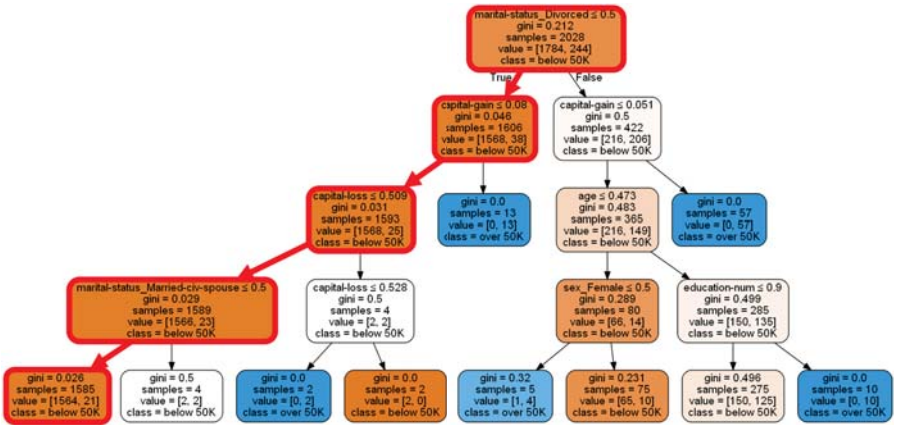


Figure 19: Returned decision trees for query on user 28817: class below 50K. Representativeness set to 10%. Guided approach with $N = 3$ and trees limited to depth 4.

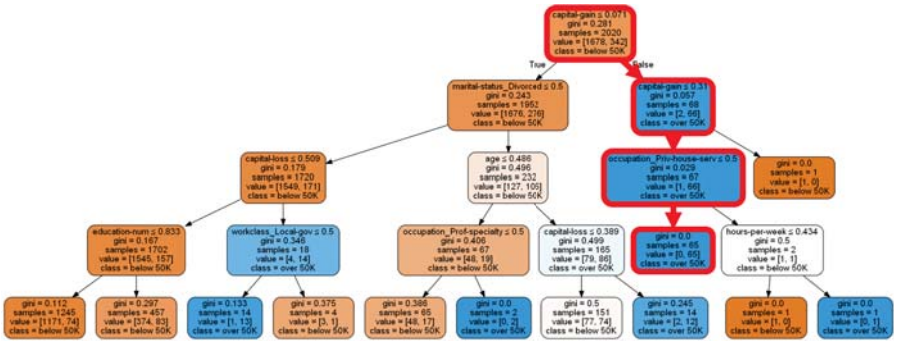


Figure 20: Returned decision trees for query on user 13891: class over 50K. Representativeness set to 10%. Guided approach with $N = 3$ and trees limited to depth 4.

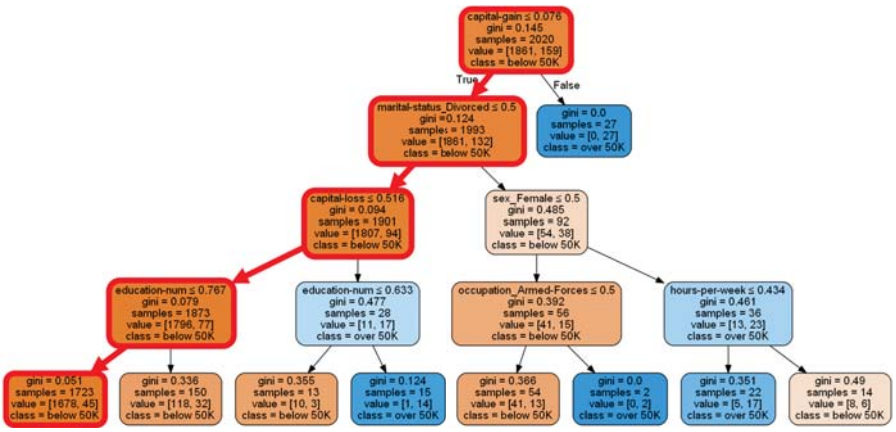


Figure 21: Returned decision trees for query on user 29901: lass below 50K. Representativeness set to 10%. Guided approach with $N = 3$ and trees limited to depth 4.

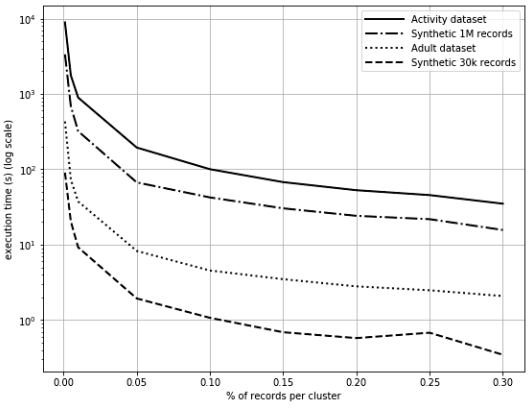


Figure 22: Execution times of the clustering procedure for the three considered data sets, as a function of the cluster representativeness r (size of clusters relative to size of the data sets)