HAL open science

Machine Learning Explainability Through Comprehensible Decision Trees

Alberto Blanco-Justicia, Josep Domingo-Ferrer

To cite this version:

Alberto Blanco-Justicia, Josep Domingo-Ferrer. Machine Learning Explainability Through Comprehensible Decision Trees. 3rd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Aug 2019, Canterbury, United Kingdom. pp.15-26, 10.1007/978-3-030-29726-8_2 . hal-02520062

HAL Id: hal-02520062

https://inria.hal.science/hal-02520062v1

Submitted on 26 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Machine Learning Explainability Through Comprehensible Decision Trees

Alberto Blanco-Justicia and Josep Domingo-Ferrer

Universitat Rovira i Virgili
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
{alberto.blanco, josep.domingo}@urv.cat

Abstract. The role of decisions made by machine learning algorithms in our lives is ever increasing. In reaction to this phenomenon, the European General Data Protection Regulation establishes that citizens have the right to receive an explanation on automated decisions affecting them. For explainability to be scalable, it should be possible to derive explanations in an automated 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 size* as a surrogate model. Specifically, we propose an approach based on microaggregation to achieve a trade-off between *comprehensibility* and *representativeness* of the surrogate model on the one side and *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

Since the turn of the century, big data are a reality. One of the main uses of this wealth of data is to train machine learning algorithms. Once trained, these algorithms make decisions, and a good number of decisions affect people: credit granting, insurance premiums, diagnosis, 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, [6]) states the right of citizens to an explanation on automated decisions. Also, the European Commission's Ethics Guidelines for Trustworthy AI [5] 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.

To be scalable, explanations must be automatically generated: even if a human auditor was able to produce a compelling explanation, one cannot assume that such an 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 understandable by humans and are thus self-explanatory, 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.

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 [10, 14], decision trees [12, 1], or linear models [13]. The surrogate model is trained 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 in 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 the training data set. We will take as surrogate models decision trees trained on disjoint subsets of the training data set. We focus on the comprehensibility of the models which we measure as the inverse of the number of nodes of the trained decision trees. In general, the fewer the nodes of a decision tree, the easier it is to comprehend it.

Section 2 characterizes the type of explanations we seek to generate and the risks of generating them through straighforward release of surrogate models. Section 3 describes our microaggregation-based approach to generate explanations of limited size. Experimental results are provided in Section 4. Finally, Section 5 gathers conclusions and future research directions.

2 Explanations via surrogate models

2.1 Machine learning explanations

According to [9], 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 [8] 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 did it 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 [7], 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 [9, 2] 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 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.

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. Traning 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.

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 [6].

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, 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. 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 explanations can be expected to be similar.

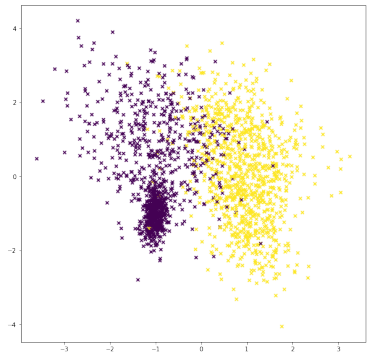


Fig. 1. Example data set

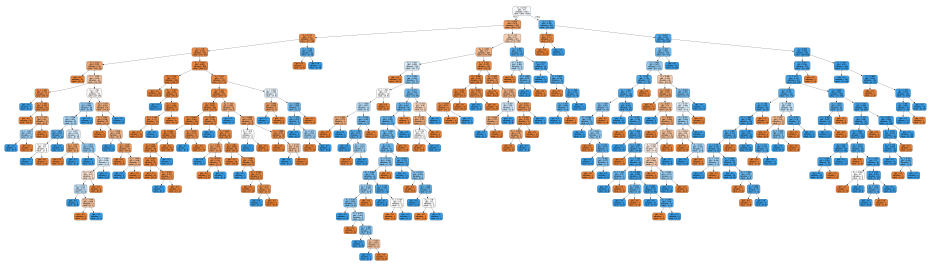


Fig. 2. Decision tree trained on the example data set

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 and 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 closest representatives to $\hat{x}$ matches the decision of the black-box model, the decision tree corresponding to the closest representative is returned. While this may hurt the fidelity of the explanations, returning the tree of 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) **let** *U be the list of cluster representatives $\tilde{c}_i$ ordered by their distance to $\hat{x}$, being $\tilde{c}_1$ the 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 [3, 4] 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 [11, 4] 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, that have the advantage of consisting of simpler and thus more comprehensible surrogate models.

Specifically, we compute microaggregation clusters using MDAV (Mean Distance to Average Vector), a well-known microaggregation heuristic [4]. 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 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.

Algorithm 3: cluster

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

```
C ← {x}
while |C| < k do
    x_i ← argmin_{x_i} distance(x_i, x)
    C ← C ∪ {x_i}
    X ← X \ {x_i}
end
return C
```

4 Empirical work

We generated a data set consisting of 30,000 records, each with 10 numeric continuous attributes and a single binary class label using the `make_classification` method from Scikit-learn[1]. Out of the 30,000 records, we reserved 2/3 to train the models, and the remaining 1/3 to validate them. The code to generate the data set and conduct all the experiments reported in this section is available as a Jupyter notebook[2].

We took as a black-box model a neural network denoted by ANN with three hidden layers of 100 neurons each, which achieves 94.22% classification accuracy. We also trained a decision tree on the whole data set, to check its accuracy and its number of nodes. The classification accuracy of this global decision tree is 88.37%, and it has 2,935 nodes. 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 produce *simpler* classifiers. Instead of directly choosing arbitrary values for k, we chose several percentages of the 20,000 records of the training data set that we wanted the clusters to contain, ranging from 0.1% to 50%; this translated to k values ranging from 20 to 10,000.

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 10,000 records not used for training). Fidelity was computed as the classification accuracy with respect to the decisions made by the black-box model.

[1] https://scikit-learn.org/stable/index.html
[2] Download address: https://www.dropbox.com/s/ex46twifl780fj4/MDAV-DT-Explainability.ipynb?dl=0

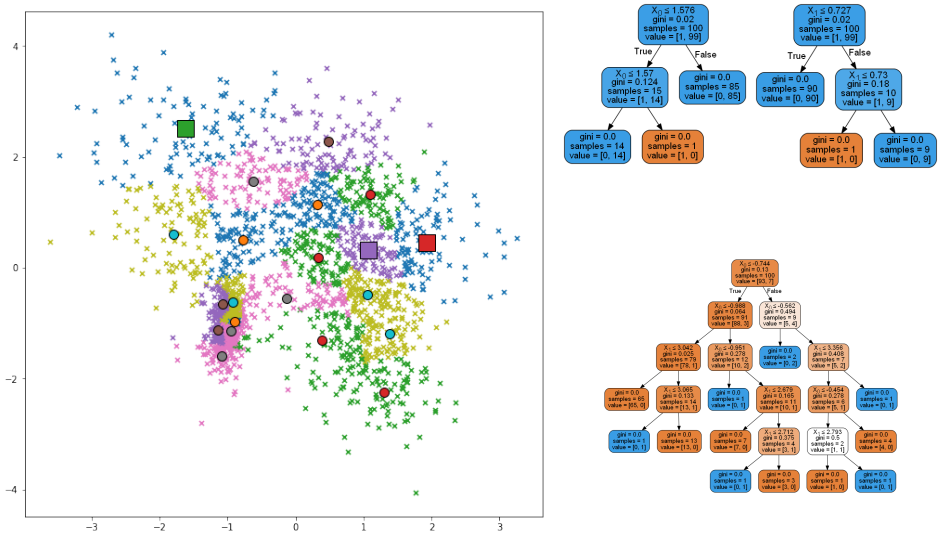
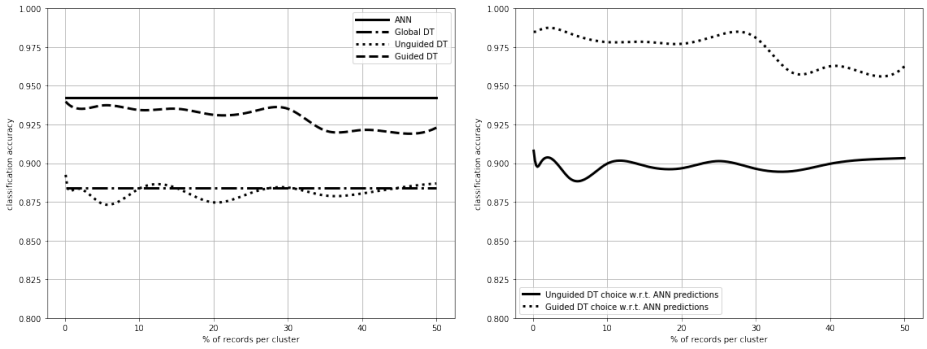


Fig. 3. Left, clusters produced by MDAV with $k = 200$ for the data set of Figure 1; for each cluster, points are in a different color and the centroid is depicted. Right, decision tree-based explanations generated for the three clusters whose centroids have been marked with □ symbols on the left figure.

Figure 4a 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% in the unguided approach (Protocol 1) and by only around 2% in the guided approach (Protocol 2, with $N = 3$). On the other hand, the accuracy of the unguided approach of Protocol 1 was basically the same as that of the global decision tree mentioned above, with the guided approach of Protocol 2 clearly outperforming both.

Moreover, it is important to note that accuracy is not very affected by the value of k, although very small values of k seem to produce slightly better results. This same behavior has been observed for several different generated data sets, so it cannot be attributed to randomness. In fact, even our unguided approach outperforms the global decision tree by around 5% in classification accuracy when trained on very small clusters (0.1% to 1%, or $k = 20$ to $k = 200$ for our data set size). The most plausible reason for this phenomenon is that for these small values of k, a substantial number of clusters are such that all records in the cluster have the same class attribute value. For these clusters, the decision tree is trivial. This hypothesis is further supported by Figure 5, discussed in more detail below, where for small values of k we find decision trees containing 0 nodes: these must correspond to clusters whose records all belong to the same class. Whether this is beneficial from the point of view of explainability is to be further explored.



(a) Accuracy (b) Fidelity

Fig. 4. Accuracy and fidelity of the decision trees for each value of k. For the guided approach $N = 3$ was used.

Figure 4b, 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 90% of the decisions. When using Protocol 2 (guided by the ANN with $N = 3$), these results improve to up to 97% coincidence, which demonstrates that our method achieves a high accuracy and fidelity with respect to the black-box model.

Figure 5 deals with the **comprehensibility** of the explanations, by depicting the number of nodes of the decision trees trained for each choice of 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 small k (up to 1% of the training set, in our case $k = 200$), the number of nodes per decision tree is well 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 4a 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).

5 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 based on clusters 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.

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

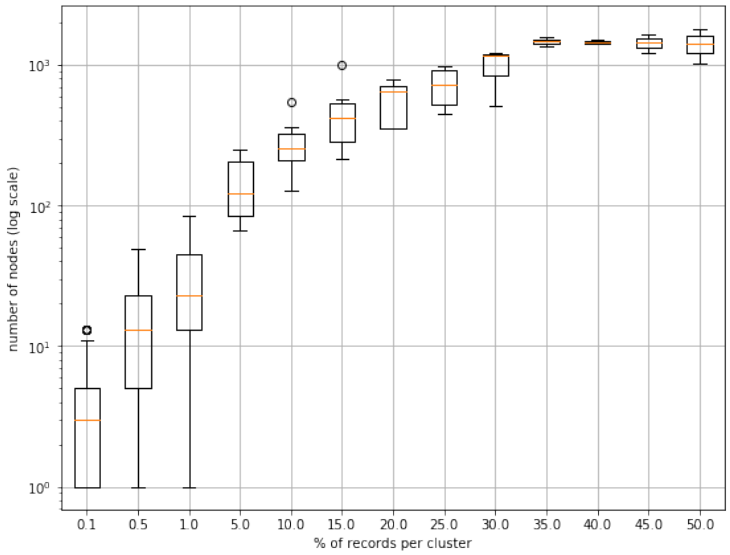


Fig. 5. Comprehensibility of explanations: the box plot represents for each k the median and lower and upper quartiles of the number of nodes per decision tree.

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 his model to protect his trade secrets. We will 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.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: European Commission (project H2020-700540 "CANVAS"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705) and Spanish Government (project RTI2018-095094-B-C21 "CONSENT"). 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. Alonso, J.M., Ramos-Soto, A., Castiello, C. and Mencar, C., 2018. Hybrid data-expert explainable AI beer style classifier. In IJCAI-18 Workshop on Explainable Artificial Intelligence (XAI 2018).

2. Biran, O. and Cotton, C., 2017. Explanation and justification in machine learning: a survey. In IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI 2017).
3. 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.
4. Domingo-Ferrer, J. and Torra, V., 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 11(2), pp.195-212.
5. European Comission's High-Level Expert Group on Artificial Intelligence, 2018. Draft Ethics Guidelines for Trustworthy AI.
6. European Union, 2016. General Data Protection Regulation. Regulation (EU) 2016/679. https://gdpr-info.eu
7. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Mller, H., 2019. Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p.e1312.
8. Miller, T., 2018. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence.
9. Molnar, C., 2018. Interpretable machine learning: A guide for making black box models explainable. Leanpub. Available at https://christophm.github.io/interpretable-ml-book/
10. 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.
11. Samarati, P., and Sweeney, L., 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report.
12. Singh, S., Ribeiro, M. T., and Guestrin, C., 2016. Programs as black-box explanations. arXiv preprint arXiv:1611.07579.
13. Strumbelj, E., and Kononenko, I., 2010. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research, 11:1–18.
14. Turner, R., 2016. A model explanation system. In IEEE Intl. Workshop on Machine Learning for Signal Processing–MLSP'16. IEEE.