Application of the Representative Measure Approach to Assess the Reliability of Decision Trees in Dealing with Unseen Vehicle Collision Data

Application of the representative measure approach to assess the reliability of decision trees in dealing with unseen vehicle collision data

Javier Perera-Lago[1][0009−0009−4536−4020],
Victor Toscano-Duran[1][0009−0006−1316−9026],
Eduardo Paluzo-Hidalgo[1,2][0000−0002−4280−5945],
Sara Narteni[3][0000−0002−0579−647X], and
Matteo Rucco[4][0000−0003−2561−3340]

[1] Department of Applied Mathematics I, University of Sevilla, Sevilla, Spain
{jperera,vtoscano}@us.es
[2] Department of Quantitative Methods, Universidad Loyola Andalucía, Campus Sevilla, Dos Hermanas, Seville, Spain
epaluzo@{uloyola,us}.es
[3] Cnr-Istituto di Elettronica, Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), Genoa, Italy
sara.narteni@ieiit.cnr.it
[4] Data Science Department, Biocentis, Milan, Italy
matteo.rucco@biocentis.com

Abstract. Machine learning algorithms are fundamental components of novel data-informed Artificial Intelligence architecture. In this domain, the imperative role of representative datasets is a cornerstone in shaping the trajectory of artificial intelligence (AI) development. Representative datasets are needed to train machine learning components properly. Proper training has multiple impacts: it reduces the final model's complexity, power, and uncertainties. In this paper, we investigate the reliability of the ε-representativeness method to assess the dataset similarity from a theoretical perspective for decision trees. We decided to focus on the family of decision trees because it includes a wide variety of models known to be explainable. Thus, in this paper, we provide a result guaranteeing that if two datasets are related by ε-representativeness, i.e., both of them have points closer than ε, then the predictions by the classic decision tree are similar. Experimentally, we have also tested that ε-representativeness presents a significant correlation with the ordering of the feature importance. Moreover, we extend the results experimentally in the context of unseen vehicle collision data for XGboost, a machine-learning component widely adopted for dealing with tabular data.

Keywords: Decision trees · XGboost · Representativeness · Feature importance

1 Introduction

In the contemporary landscape of technological evolution, the imperative role of representative datasets stands as a cornerstone in shaping the trajectory of artificial intelligence (AI) development. As AI algorithms increasingly influence decision-making processes, the need for robust, unbiased, and comprehensive datasets becomes ever more critical [21]. The frameworks that provide the guidelines and the technical requirements for engineering trustworthy data-driven AI systems provide at least two possible interpretations of representativeness [1, 5, 10, 11, 17, 22, 31]: 1) attribute coverage and completeness: how much the samples in a dataset describe the phenomena under observations. The question mark can be interpreted both in terms of the number of features that have been recorded in the dataset and the number and sparsity of samples concerning the dynamics of the system. 2) similarity among datasets: given two or more datasets, how much are they different, and what is their impact on a data-driven model derived from the datasets. Attribute coverage and completeness can be measured with different approaches. Methods such as Population Parity and Disparate Impact quantify and identify disparities in the distribution of different demographic groups within a dataset [26, 33]. Statistical techniques like Feature Divergence and Kernel Density Estimation can be used to assess the similarity between feature distributions in the dataset and the real-world population [29, 30]. Fairness Indicators and Counterfactual Fairness detect and mitigate biases in AI models trained on biased datasets [15, 34]. Outlier Analysis and Clustering Techniques can be used to identify the points that might compromise dataset representativeness [6, 20]. Temporal Drift Detection and Cohort Analysis are explored, shedding light on techniques that ensure the dataset's representativeness remains intact as conditions evolve over time [27, 32]. Transfer Learning and Adversarial Training are discussed in adapting models to diverse domains, mitigating biases associated with domain-specific features [12, 25]. Crowdsourcing and Expert Evaluation are presented as methods to incorporate diverse perspectives, providing insights into potential biases and limitations in the dataset from end-users and domain experts [18, 19]. Intersectional Approaches and Subgroup Analysis are considered, emphasizing the importance of examining the interaction between multiple demographic attributes for a more comprehensive assessment of representativeness [9].

On the other hand, assessing the similarity of datasets is crucial for ensuring that machine learning models trained on these datasets generalize well to real-world scenarios. Various methods have been developed to measure the likeness between datasets, to identify potential discrepancies, and to ensure that the data used for model training accurately represents the target domain. Metrics such as Wasserstein distance, Jensen-Shannon divergence, and Bhattacharyya distance offer quantitative measures of dissimilarity between probability distributions, providing insights into the overall similarity of datasets [3, 28]. Additionally, domain adaptation techniques, such as Maximum Mean Discrepancy (MMD) and Kernelized Discrepancy (KMM), focus on aligning feature distributions between source and target domains, ensuring a seamless transition from one dataset to

another [14, 16]. Understanding and mitigating dissimilarities between datasets are paramount to building robust and generalizable machine learning models, as models trained on dissimilar datasets may exhibit poor performance when deployed in real-world applications. Therefore, a comprehensive assessment of dataset similarity is indispensable for fostering reliable and effective AI systems. In this paper, we use the measure proposed in [13] to quantify the similarity between datasets and how their difference has an impact on a machine learning component, decision trees and, in particular, for the eXtreme Gradient Boosting (aka XGBoost). XGBoost stands out as a powerful and versatile machine learning algorithm with significant implications for trustworthy AI. Its importance lies in its ability to enhance model performance across various tasks, such as classification, regression, and ranking. XGBoost excels in handling complex datasets, mitigating overfitting, and providing robust predictions. The algorithm's interpretability features, such as the ability to generate feature importance scores and decision trees, contribute to the transparency of AI models—a crucial aspect for ensuring trustworthiness. Moreover, XGBoost's regularized learning objectives and advanced optimization techniques foster model generalization and prevent overfitting, reinforcing the reliability of AI systems. The algorithm's widespread adoption in both research and industry attests to its effectiveness and underscores its role in building trustworthy AI models that prioritize accuracy, interpretability, and generalization [7].

This paper is organized as follows. In Section 2, the main concepts about decision trees are provided. Then, in Section 3, we discuss the relationship between the ε-representativeness measure and the predictions of decision trees proving theoretical results. In Section 4, we explore experimentally the correlation between ε-representativeness and the similarity between the explanations provided by decision trees and XGBoost by the ordering of the feature importance. Finally, conclusions and future work are discussed in Section 5.

2 Classification with decision trees and XGBoost

In this section, we introduce all the needed concepts about classification and the model family of decision trees. For a further understanding of Machine Learning, we refer to [23], and for geometric results for point clouds we refer to [4].

Let (X, λ_X) be a dataset for classification with $X \subseteq \mathbb{R}^n$ the set of data with size N and $\lambda : X \to \{1, \ldots, c\}$ the class labelling. For each $x \in X$, the d-th coordinate coordinate of x will be denoted by $x_d \in \mathbb{R}$ and will be called the d-th feature of x. The objective of classification is to find a function $f : \mathbb{R}^n \to \{1, \ldots, c\}$ that approximates λ_X. In the literature, there exist different families of functions f, such as artificial neural networks and support vector machines, which are called models. We will focus on the family of decision trees $\mathcal{T}$ which are supervised learning models with desirable properties such as that they are interpretable by definition. The most basic model of the family is decision trees (DT) (see [23, Section 8.3.3]).

A (binary) DT is a rooted tree representing a partition of the feature space. It contains a set of ordered nodes $\{n_1, \cdots, n_r\}$ where n_1 will be called the root node. A node is said to be internal if it is connected to two children nodes, opening new branches in the tree. Conversely, terminal nodes, also known as leaves, do not have children and represent the endpoints of a branch in the tree. Let $I \subset \{1, \cdots, r\}$ be the subset of indices corresponding to the internal nodes, and L be the subset of indices corresponding to the terminal nodes. The root node n_1 is considered an internal node, so $1 \in I$.

Given a data $x \in X$, x traverses the internal nodes of the DT from the root n_1 to one of the terminal nodes. Each internal node n_i is associated with one of the features $d_i \in \{1, \cdots, n\}$, and with one condition in terms of an inequality bounded by a threshold value $c_i \in \mathbb{R}$ that will send the data to one of the two children nodes based on the inequality. This inequality is called a decision rule. When x reaches an internal node n_i, it is sent to its left child if $c_i - x_{d_i} > 0$ and to its right child otherwise. The margin $\mu_i > 0$ of an internal node n_i is the minimum value $|c_i - x_{d_i}|$ for all the examples $x \in X$ reaching n_i. Each terminal node n_j is associated with an integer $k_j \in \{1, \cdots, c\}$ representing a class label.

There exist different training algorithms for DT. In our case, let us consider first a DT with just one node that splits the dataset into two subsets based on one of the features. The feature and the splitting condition are chosen based on maximizing the purity of the nodes. A node is considered pure when all the data reaching that node has the same label. Then, the process is iterated recursively until a desired depth is reached or the purity can not be improved. There exist different purity measures. The most common ones are the entropy and the Gini index.

Let us denote by N_i the number of data from X reaching the node n_i. The number of examples of class k reaching n_i will be denoted by $N_{i,k}$. Then, let us use the following notation: $p_i = N_i/N$, $p_{i,k} = N_{i,k}/N_i$. The entropy of a node n_i is $E(n_i) = -\sum_{k=1}^{c} p_{i,k} \log p_{i,k}$, and the Gini index of n_i is $G(n_i) = \sum_{k=1}^{c} p_{i,k}(1 - p_{i,k})$. Assume that we fix a purity measure and we denote it as I. The information gain for an internal node n_i whose two children nodes are n_{i_1} and n_{i_2} is:

$$IG(n_i) = I(n_i) - \frac{N_{i_1}}{N_i} I(n_{i_1}) - \frac{N_{i_2}}{N_i} I(n_{i_2})$$

Feature importance (FI) quantifies the impact of a particular feature $d \in \{1, \cdots, n\}$ in increasing the purity of the decision tree. It is calculated as:

$$FI(d) = \sum_{\substack{i \in I \\ d_i = d}} N_i \cdot IG(n_i)$$

The goodness of the classification given by $T \in \mathcal{T}$ for the dataset (X, λ_X) is measured by the accuracy, with formula:

$$\mathrm{Acc}(T, (X, \lambda_X)) = \sum_{j \in L} p_j \cdot p_{j,k_j}$$

A more complex model belonging to $\mathcal{T}$ is called XGBoost [8]. A thorough description of XGBoost is out of the scope of this paper but, roughly speaking, it is an ensemble of decision trees, i.e., it combines the predictions of several DTs that are sequentially built, each correcting errors of the previous one. Then, during the training a gradient descent optimization algorithm is used.

3 Representativeness and decision trees

In this section, we introduce a metric to compare different training datasets and we show a theoretical result proving the relation between this metric and the final performance of a DT.

Given a dataset (X, λ_X), the ability of a DT to predict new unseen data will depend on the completeness and quality of the information learned during training. It becomes then a vital task to find measures to compare datasets expected to induce similar predictions. In [13], a measure based on computational topology was proposed. This measure is called the ε-representativeness of a dataset. Given another dataset (Y, λ_Y) with the cardinal of Y smaller than the one of X, we say that $y \in Y$ is an ε-representative of $x \in X$ if $||y - x||_\infty \leq \varepsilon$ and $\lambda_X(x) = \lambda_Y(y)$, and we say that (Y, λ_Y) is an ε-representative dataset of (X, λ_X) if for all $x \in X$ there exists $y \in Y$ that is an ε-representative of x. In general, we will consider Y to be a subset of X but it is not a necessary condition. We will add a superscript $*$ (for example N^*) when referring to the dataset (Y, λ_Y). A dataset (Y, λ_Y) that is representative of (X, λ_X) is said to be γ-balanced if each $y \in Y$ is representative of exactly γ data examples of X and each $x \in X$ is represented by a single example $y \in Y$.

Knowing the concept of representativeness, we present the following result:

Theorem 1. *Let be $T \in \mathcal{T}$ a binary DT, (X, λ_X) a dataset and (Y, λ_Y) a γ-balanced ε-representative dataset of (X, λ_X). If $\varepsilon < M = \min_{i \in I} \mu_i$, then*

$$\mathrm{Acc}(T, (X, \lambda_X)) = \mathrm{Acc}(T, (Y, \lambda_Y))$$

Proof. Let be $x = (x_1, \cdots, x_n)^T \in X$ and $y = (y_1, \cdots, y_n)^T \in Y$ an ε-representative of x. That means that $|y_i - x_i| \leq \varepsilon \ \forall i \in \{1, \cdots, n\}$. Assume that y reaches the tree through the root node n_1 (root node) and is sent to its left child. That means that $0 < c_1 - y_{d_1}$. By the definition of margins, we can improve the inequality by applying that $\mu_1 \leq c_1 - y_{d_1}$. Since y is ε-representative of x, we have that $x_{d_1} \leq y_{d_1} + \varepsilon$. Combining both inequalities we have that $\mu_1 - \varepsilon \leq c_1 - x_{d_1}$. Since $\varepsilon < M \leq \mu_1$, then $0 < c_1 - x_{d_1}$, meaning that x is also sent to the left child. Analogously, if we assume that y is sent to the right child of n_1, we can show that x is also sent to the right child. If we apply this same reasoning to all the internal nodes that y passes through, we see that x follows the same path through the tree and, consequently, reaches the same terminal node n_j and is classified with the same label k_j.

By the definition of γ-balance, for each $y \in Y$ of class k there are exactly γ examples from X of class k ε-represented by y, and all of them reach the

same terminal node of T. It also follows from the definition of γ-balance that $N = \gamma \cdot N^*$, $N_j = \gamma \cdot N_j^*$ and $N_{j,k} = \gamma \cdot N_{j,k}^*$. Consequently, $p_j = N_j/N = (\gamma \cdot N_j^*)/(\gamma \cdot N^*) = N_j^*/N^* = p_j^*$ and $p_{j,k} = N_{j,k}/N_j = (\gamma \cdot N_{j,k}^*)/(\gamma \cdot N_j^*) = N_{j,k}^*/N_j^* = p_{j,k}^*$. Consequently:

$$\mathrm{Acc}(T, (X, \lambda_X)) = \sum_{j \in L} p_j \cdot p_{j,k_j} = \sum_{j \in L} p_j^* \cdot p_{j,k_j}^* = \mathrm{Acc}(T, (Y, \lambda_Y))$$

□

4 Experiments

The experiments developed are organized as follows. Firstly, a $2D$ synthetic dataset was used as an illustration of the proposed methodology. Then, the experimentation was extended to the Vehicle Platooning (also called Collision) dataset [24]. Given a dataset, subsets of the set with different values of ε-representativeness will be considered and the ordering of the features based on the importance compared. The results suggest that the lower values of ε will produce similar explanations of the input data and similar decision boundaries. The code for the experiments is available in a GitHub repository [5].

4.1 Synthetic $2D$ dataset

In this experiment, we used a $2D$ dataset generated using the Python Scikit-learn package[6] as an example. The dataset comprises 200 data distributed in two noisy concentric circles representing distinct classes as shown in Figure 1.

The methodology followed in the experiments is summarized in the following steps. The dataset was split into a training set composed of the 75% of the data and a test set with the remaining 25%. Then, two random subsets containing the 40% of the training set were considered, from now on Subset 1 and Subset 2, and their ε-representativeness was computed, obtaining $\varepsilon = 0.756$ for the Subset 1 and $\varepsilon = 0.497$ for the Subset 2. A DT was trained with the training set and the subsets using the Gini index and a maximum depth of 4 obtaining the DTs of Figure 2. As we can see, the resultant decision rules after training the DT with the training set are similar with a lower value of ε. The accuracy values for the DTs on the test set were: 0.84 for the train dataset, 0.94 for Subset 1, and 0.82 for Subset 2. To determine the similarities between the DTs, we will consider the ordering of the features importance. The first feature ranked the most important for both the training set and Subset 2 while the ordering was the opposite for Subset 1 (See Table 1).

[5] `https://github.com/Cimagroup/Application_Representative_Measure_Reliability_DT`

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

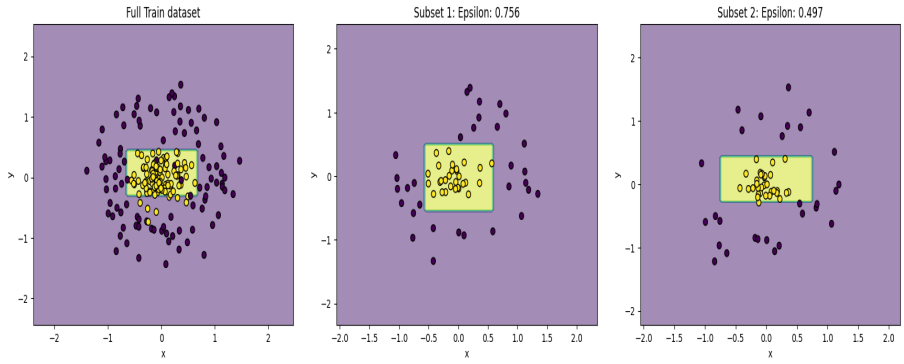

Fig. 1. Full synthetic dataset generated using Scikit-learn and the two random subsets. From left to right: (1) the training set; (2) a subset composed of a 40% of the training set and $\varepsilon = 0.756$; (3) a subset composed of 40% of the training set and $\varepsilon = 0.497$. We can also see the decision boundaries of the DT trained using each set of data.

Table 1. Feature importance percentage for the training set, Subset 1, and Subset 2. We can see that the most important feature for the training set and Subset 2 is the same.

Data	x_1	x_2
Training set	40.4	**59.6**
Subset 1	**50.37**	49.62
Subset 2	18.4	**81.6**

4.2 The collision dataset

In this experiment, we used the binary classification Collision Dataset [24] which consists of predicting whether a platoon of vehicles will collide based on features such as the number of cars or their speed. It is composed of 107,210 data with 23 numerical features. Each class is composed of 69,348 examples and 37,862 examples, respectively.

We followed a similar methodology to the one proposed for the experiment in Section 4.1. The dataset was split into a training set composed of the 75% of the data and a test set with the remaining 25%. Then, two random subsets containing the 10% of the training data were considered, from now on Subset 1 and Subset 2, and their ε-representativeness was computed, obtaining $\varepsilon = 0.539$ for Subset 1 and $\varepsilon = 0.655$ for Subset 2. A DT was trained with the training set and the subsets using the Gini index and a maximum depth of 10. The following accuracy values were obtained on the test set: 0.874 for the training dataset, 0.841 for Subset 1, and 0.84 for Subset 2. To determine the similarities between the DTs, we will consider the ordering of the feature importance. For comparing the ordering of the feature importance, we used a metric developed in [2, Section 4.2]. Let x and y be two ordered sets whose elements are the features of the dataset

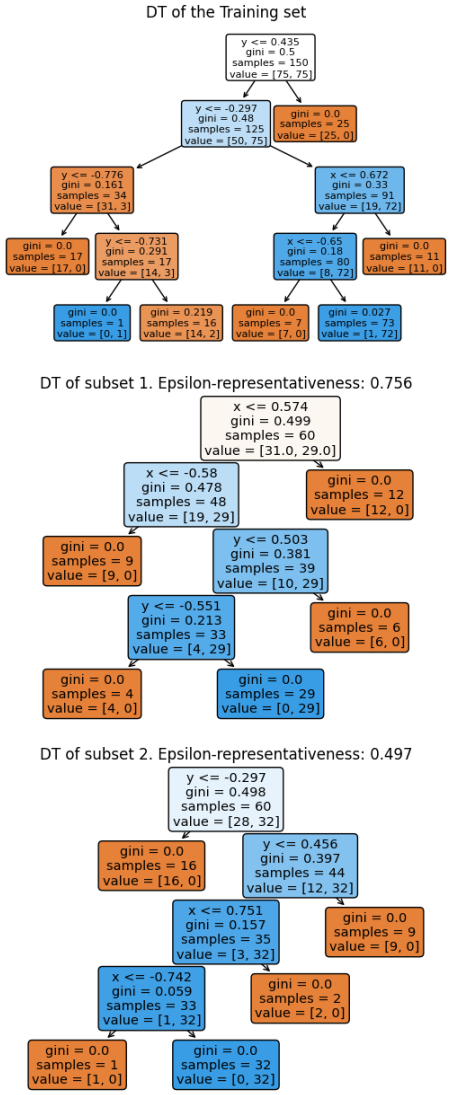

Fig. 2. DT for the sets in Figure 1. From top to bottom: (1) the training set; (2) a subset composed of a 40% of the training set and $\varepsilon = 0.756$; (3) a subset composed of 40% of the training set and $\varepsilon = 0.497$.

ordered by their importance. Then, for each feature, we compute the absolute value of the difference between the position of the feature in x and y. Finally, the mean of these differences is computed. For Subset 1, we obtained a value of 1.74 for this metric, and for Subset 2 the value was 1.83, where we can highlight the higher similarity of variable importance for the decision tree generated by

the subset with lower epsilon, that is the Subset 1. The ordering of the feature importance for the three DTs is displayed in Table 2. Finally, the experiment was repeated for 100 subsets and the Spearman's correlation (Sp) between the ε-representativeness and the metric of the ordering of the feature importance was computed obtaining significant correlation ($Sp = 0.51$, p-value$= 5.2 \times 10^{-8}$).

Additionally, we trained Gradient Boosting Classifier(XGBoost) with the training set and the subsets using the Friedman Mean Squared Error as a criterion and a maximum depth of 10. The number of boosting stages to perform was set to 25. The following accuracy values were obtained: 0.912 for the training dataset, 0.882 for Subset 1, and 0.876 for Subset 2. In this case, Subset 1 has a similarity of 0.696 for the feature importance, and Subset 2 has a similarity value of 1.823, reaching better similarity with lower ε. The ordering of the feature importance for the three XGBoost trained models is displayed in Table 2. Finally, the experiment was repeated for 20 subsets and the Spearman's correlation (Sp) between the ε-representativeness and the metric of the ordering of the feature importance was computed obtaining significant correlation ($Sp = 0.673$, p-value$= 1.79 \times 10^{-14}$).

5 Conclusions and future work

The results of this paper are two-fold. Firstly, we proved that similar accuracy is obtained when certain conditions about representativeness are satisfied when using DTs. Secondly, by experimentation, we have compared the feature importance ordering for different subsets of the training data with different values of representativeness reaching a significant correlation between them. According to our results, representative sets produce similar explanations of the dataset. Finally, we extend our experiments to a more complex decision tree model, XGBoost. In the future, we would like to provide theoretical guarantees regarding the ordering of the importance of the features and distance-based comparison between the decision rules of DT.

Acknowledgements

We want to thank Maurizio Mongelli and Miguel A. Gutierrez-Naranjo for the insightful discussions and ideas. The work was supported in part by the European Union HORIZON-CL4-2021-HUMAN-01-01 under grant agreement 101070028 (REXASI-PRO) and by TED2021-129438B-I00 / AEI/10.13039/501100011033 / Unión Europea NextGenerationEU/PRTR.

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. In: 2019 IEEE international conference on systems, man and cybernetics (SMC). pp. 4184–4191. IEEE (2020)

Table 2. Ordering of the feature importance for DT and XGBoost trained with the training set and the random subsets. The number indicates the position in the importance order. For example, the most important feature for DT trained on the training set is d_ms.

	DT			**XGBoost**		
	Training set	**Subset 1**	**Subset 2**	**Training set**	**Subset 1**	**Subset 2**
F0	2	4	2	2	3	2
d_ms	10	10	10	10	10	10
d0	6	8	9	5	7	9
v0	23	23	20	23	23	23
prob	5	5	8	6	5	7
Int_dv1	15	14	12	13	13	12
Int_dv2	12	15	19	16	15	21
Int_dv3	14	17	14	14	14	13
Int_dv4	7	9	7	9	9	8
Int_dv5	20	12	21	18	16	19
Int_dv6	16	16	23	15	17	20
Int_dv7	1	1	1	1	1	1
Int_dd2	18	22	16	19	19	14
Int_dd3	19	18	15	20	20	16
Int_dd4	3	3	3	4	4	4
Int_dd5	13	13	13	12	12	15
Int_dd6	17	21	17	17	18	18
Int_dd7	11	11	11	11	11	11
duration	22	20	22	21	22	22
freq	4	2	4	3	2	3
ampl	8	7	6	7	8	6
Fresponse	21	19	18	22	21	17
KInt	9	6	5	8	6	5

2. Barrera-Vicent, A., Paluzo-Hidalgo, E., Gutiérrez-Naranjo, M.A.: The metric-aware kernel-width choice for lime. In: Longo, L. (ed.) Joint Proceedings of the xAI-2023 Late-breaking Work, Demos and Doctoral Consortium co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), Lisbon, Portugal, July 26-28, 2023. CEUR Workshop Proceedings, vol. 3554, pp. 117–122. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3554/paper21.pdf
3. Bhattacharyya, A.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Biometrika **34**(3/4), 291–302 (1943)
4. Boissonnat, J.D., Chazal, F., Yvinec, M.: Geometric and Topological Inference. Cambridge Texts in Applied Mathematics, Cambridge University Press (2018)
5. Brundage, e.a.: Toward trustworthy ai development: Mechanisms for supporting verifiable claims. arXiv preprint arXiv:2110.05282 (2021)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) **41**(3), 1–58 (2009)
7. Chen, T., Guestrin, C.: Xgboost: A scalable and accurate implementation of gradient boosting. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 785–794 (2016)
8. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785
9. Crenshaw, K.: Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum pp. 139–167 (1989)
10. Diakopoulos, N.: Accountability in algorithmic decision making: A visual analysis framework. Data Society Research Institute (2016)
11. Floridi, L., Taddeo, M.: Soft ethics, the governance of the digital and the general data protection regulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2133), 20180081 (2018)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
13. Gonzalez-Diaz, R., Gutiérrez-Naranjo, M.A., Paluzo-Hidalgo, E.: Topology-based representative datasets to reduce neural network training resources. Neural Computing and Applications (2022)
14. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. In: Advances in neural information processing systems. pp. 513–520 (2007)
15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
16. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. Advances in neural information processing systems **19**, 601 (2007)
17. Jobin, A., Ienca, M., Vayena, E.: Global data justice. Ethics and Information Technology **21**(2), 87–96 (2019)
18. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 453–456 (2013)

19. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work. pp. 37–46 (2008)
20. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Outlier detection in axis-parallel subspaces of high-dimensional data. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 555–564 (2009)
21. Liang, W., Tadesse, G.A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., Zou, J.: Advances, challenges and opportunities in creating data for trustworthy ai. Nature Machine Intelligence **4**(8), 669–677 (2022)
22. Liu, Y., Zhao, Z., Guan, Y., Song, S., Cao, L., Chen, H.: Towards trustworthy ai: A cross-disciplinary survey. arXiv preprint arXiv:1907.02527 (2019)
23. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2 edn. (2018)
24. Mongelli, M., Ferrari, E., Muselli, M., Fermi, A.: Performance validation of vehicle platooning through intelligible analytics. IET Cyber-Physical Systems: Theory & Applications **4**(2), 120–127 (2019). https://doi.org/https://doi.org/10.1049/iet-cps.2018.5055, `https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cps.2018.5055`
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
26. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568 (2008)
27. Rahman, S.M.M.M., Davis, D.: Detecting and mitigating concept drift in an adaptive learning system for personalized news recommendation. Knowledge-Based Systems **166**, 132–145 (2019)
28. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**(2), 99–121 (2000)
29. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons (2015)
30. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. Annals of the Institute of Statistical Mathematics **64**(5), 1009–1044 (2012)
31. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. In: 2017 IEEE international conference on data science and advanced analytics (DSAA). pp. 520–529. IEEE (2017)
32. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine learning **23**(1), 69–101 (1996)
33. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180 (2017)
34. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning (Vol. 28, No. 3) (2013)