Measuring Feature Importance Through Counterfactual Distributions

Abstract

Understanding the importance of individual features in machine learning models is critical for interpretability, especially in a field where no definitive ground-truth exists. This paper provides an alternative perspective by proposing a novel local feature importance method that can be applied to any model. The general idea is to obtain two sets of positive and negative counterfactuals, estimate their underlying distributions with Kernel Density Estimations (KDE), and rank the features where positive and negative counterfactuals differ the most. We anchor our approach within a solid mathematical framework, demonstrating that it satisfies key properties to serve as a measure of dissimilarity and distance.

We demonstrate the effectiveness of our method by illustrating and comparing the results with traditional and well-established local feature importance scores. We incorporate faithfulness metrics as an additional point of comparison, complementing our analysis and providing a more comprehensive evaluation of the differences and similarities between metrics. The results obtained on both datasets show that the metric we introduced is generally better compared to the other both on comprehensiveness and sufficiency. Consequently, this reinforces the usefulness of complementary introduction of our metric, not only because it represents a different point of view, capturing distributional aspects of the data, but also because of the results obtained.

1 INTRODUCTION

Explainable AI (XAI) addresses one of the most pressing challenges in modern machine learning: understanding and interpreting how a model produces its outputs. With the extensive adoption of black-box models in sensitive and high-stakes domains—such as medical diagnosis, credit scoring, and criminal justice—the need for interpretable methods has become not merely desirable but indispensable. In these contexts, the consequences of opaque decision-making can be severe, ranging from ethical concerns to life-altering outcomes (see [Solon Barocas, 2023], [Ingram, 2020], [Balasubramaniam et al., 2023]). One way to tackle this problem is to detect which features are more involved in generating the output. To do so, the scientific community mostly developed two classes of methods: intrinsic and post-hoc methods. Intrinsic methods leverage the internals of a given classifier to measure the degree to which each feature contributes to the model's predictions, whereas post-hoc methods operate independently of the classifier by analyzing its outputs. The choice of method type is subject of scientific debate, with conflicting opinions and arguments ([Rajbahadur et al., 2022]). Among the various post-hoc methods proposed, counterfactual (CF) explanations, introduced by Wachter et al. ([Wachter et al., 2017]), have emerged as a common apporach for model interpretability. A CF explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred". Counterfactuals are human-friendly explanations, because they are contrastive to the current instance and because they are selective, meaning they usually focus on a small number of feature changes.

In mathematical terms, let us consider a black-box model $M\colon \mathbb{R}^d \to \{0,1\}$ and a certain input $x \in \mathbb{R}^d$ for which the output $y = M(x)$ is observed. Counterfactuals are solutions to

$$x_{cf} = \arg\min_{x'} \mathcal{L}(M(x'), y_{target}) + \lambda d(x, x')$$

where $\mathcal{L}$ is a loss function, for instance $(M(x') - y_{target})^2$, λ is a regularization parameter, y_{target} is the desired outcome and $d(\cdot,\cdot)$ is a distance function function ensuring that x and x' remain close. Building on this foundation, [Dandl et al., 2020] proposed an alternative formulation that

Submitted to the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025). **To be used for reviewing only**.

minimizes a four-objective loss function:

$$L(x, x', y', X^{obs}) =$$
$$= (o_1(\hat{f}(x'), y'), o_2(x, x'), o_3(x, x'), o_4(x', X^{obs}))$$

where o_i are proper objective function that encodes desirable properties on the counterfactual: o_1 reflects that the prediction of our counterfactual x' should be close to the desired outcome, o_2 encodes the similarity between x' and x, o_3 sparsity and o_4 that counterfactuals have likely feature values/combinations. From an operational point of view, when generating CFs, we provide a list of x_{cf} based on different notions as proximity, diversity, sparsity, and actionability (see papers [Mothilal et al., 2020], [S., 2023] and [Karimi et al., 2019]). The generation of CFs often relies on techniques such as trial and error or optimization algorithms like NSGA-II ([Deb et al., 2002]). The advantages and limitations of CFs have been widely studied in recent years. Please refer to [Verma et al., 2020] for a wider discussion.

In this work, we propose a novel approach to rank dimensions where positive and negative CFs differ the most, enhancing interpretability and enabling better decision-making in high-stakes applications. Section 3 starts with the intuition and definition of formula (2) exploring its properties and geometrical meaning. We conclude the section with two important results, Theorems 3.1 and 3.2, showing that our equation defines a measure of dissimilarity and a distance. Section 4 is devoted to the experiments. Here, we compare formula (2) with DiCE local feature importance scores as well as other metrics on various datasets. In general, the contributions of this work can be summarized as follows:

1. We propose an innovative local feature importance method that uses positive and negative counterfactuals to capture the relevance of features in generating a specific output. Our formulation takes into account value support and distributional density, thus capturing nuances in the data
2. We effectively prove that our proposed measure satisfies the mathematical properties required to qualify as a measure of dissimilarity, and for $k = 1$, it adheres to the axioms of a metric. This formal grounding ensures that the measure is not only conceptually valid but also mathematically rigorous.
3. We perform extensive analysis on a two datasets for several variables comparing our metric with other standard feature importance scores, underlying the value of our metrics.
4. We show that by comparing local importance scores with the faithfulness metrics Completeness and sufficiency, our framework generally produces better results.

2 RELATED WORK

Local interpretation methods are used to explain individual predictions in machine learning models where the relationship between features and outputs is complex or opaque. These methods attempt to uncover the contributions of individual features to the model's decision process for a particular prediction, in contrast to global interpretation methods that explain overall model behavior. Broadly speaking, most local methods focus on calculating the importance of each feature for the specific instance under analysis ([Molnar, 2022]). Several methods have been proposed in recent years to tackle this challenge. For instance, Individual Conditional Expectation (ICE), first introduced by Goldstein et al. [2015], is a technique that visualizes how the model's prediction changes when a feature varies for a single instance. ICE plots display one line per instance, illustrating how the prediction responds to changes in that instance's features. This method allows for an instance-level understanding of feature impact, but it can be difficult to interpret when dealing with high-dimensional data. LIME (Local Interpretable Model-agnostic Explanations) ([Ribeiro et al., 2016]) provides another approach, where it approximates the complex model locally with an interpretable surrogate model. The idea is to perturb the input data around the instance of interest and fit a simpler model to these perturbed instances, which can then be interpreted in a straightforward manner. While LIME has become quite popular for providing insights into black-box models, its reliance on local linearity makes it sensitive to the choice of proximity measures and perturbation methods ([Thomas Altmann, 2019]).

CFs ([Wachter et al., 2017]) offer an alternative by identifying the minimal change in features required to alter the model's prediction. CFs focus on generating "what if" scenarios, offering users a counterfactual instance that illustrates what would need to happen for a different outcome. The main limitation of counterfactual explanations is the fact that they are instance-specific, i.e. no general information about the model reasoning as a whole is extracted as showed in [Guidotti, 2022] and [Setzu et al., 2021].

Anchors, introduced by [Ribeiro et al., 2018], describe a prediction as being anchored by certain feature values, which lock the model's prediction in place. The idea behind Anchors is to find a local subset of features that sufficiently explain a prediction by ensuring that the prediction would not change when those features remain fixed.

Shapley Values ([Lundberg and Lee, 2017]) come from cooperative game theory and represent the average marginal contribution of a feature across all possible coalitions. Shapley values assign a score to each feature based on how much it contributes to the prediction in comparison to all other possible subsets of features. SHAP has become one of the most widely used methods for feature attribution,

as it provides a theoretically grounded explanation that satisfies several desirable properties, such as fairness and consistency, even tough it may suffer from correlation among features ([Marcilio and Eler, 2020].
Although several techniques exist, in the literature LIME, SHAP and CF are among the most commonly referred approaches (Mishra et al. [2021], [Collaris et al., 2022]) and have been widely adopted in various applications. However, it is important to note that different feature importance methods can yield different rankings of features, even for the same dataset and model. As highlighted by [Rajbahadur et al., 2022], papers in this field rarely provide a clear justification for choosing one particular method over others. This lack of transparency makes it challenging to evaluate or select the best method for any given problem. Additionally, since there is no universally agreed-upon "ground truth" in model interpretability ([Chakraborty et al., 2017], [Haufe et al., 2024], Harel et al. [2023]), the validity of these methods often relies on empirical validation rather than rigorous theoretical proofs.

In this context, our approach, which to the best of our knowledge is the only one with this specific formulation, seeks to provide a mathematically sound alternative by offering a method that is based on distributional discrepancies between local neighborhoods of the data. Our approach directly measures the distance between distributions to assess feature importance, providing an interpretable framework for understanding model behaviour. In addition, we compare the different metrics from different perspectives: in particular, we use the feature agreement and as a final term of comparison we use two well-established metrics of faithfulness.

3 RANKING DIMENSIONS WITH POSITIVE AND NEGATIVE CFS

Let us assume the generation of two sets of CFs: positive CFs C^+ and negative CFs C^- each with cardinality m. Positive CFs C^+ are those that successfully flip the output y, while negative CFs C^- fail to achieve the desired outcome. Our goal is to rank the dimensions in the input space $\mathbb{R}^d$ (or in the activation space of a neural network layer, for neural models) where C^+ and C^- differ the most. This ranking provides insights into which dimensions are most critical in determining the model's output.

A straightforward approach is to compute the variability of each dimension between C^+ and C^-. Specifically, for each dimension i we calculate:

$$Var_i = \frac{1}{m}\sum_{j=1}^{m}(x_{i,j}^{+} - x_{i,j}^{-})^2.$$

where $x_{i,j}^{+}$ and $x_{i,j}^{-}$ represent the i-th entry of the j-th counterfactual in C^+ and C^- respectively.

While variability provides a simple and effective heuristic, it does not account for the underlying data distribution. To address this, we propose an alternative method based on Kernel Density Estimation (KDE). For continuous or ordinal variables, KDE allows us to estimate the probability distributions of the dimensions in C^+ and C^-. The estimated distributions are respectively given by

$$P(x)_i = \frac{1}{mh_1}\sum_{j=1}^{m} K\Big(\frac{x - x_{i,j}^{+}}{h_1}\Big),$$
$$Q(x)_i = \frac{1}{mh_2}\sum_{j=1}^{m} K\Big(\frac{x - x_{i,j}^{-}}{h_2}\Big)$$

where $K(\cdot)$ is the kernel function (e.g., Gaussian or Epanechnikov kernel; see [Weglarczyk, 2018]), h_1 and h_2 are bandwidth parameters for C^+ and C^-. By comparing the above distributions $P(x)_i, Q(x)_i$ for each dimension we can capture more nuanced differences in how positive and negative CFs behave. This method is particularly useful in cases where feature variability alone may not fully capture the importance of a dimension.
In order to ground this intuition within a mathematical framework we introduce a notion that captures how the two distribution differs.

3.1 THE NOTION OF OVERLAP OF FUNCTIONS

To formalize the intuition about how two distributions p and q differ, we introduce a measure of overlap $o(p, q)$. Before presenting the formal definition, we outline some desirable properties of such a measure:

1. If $p = \mathbf{1}_{[0,1]}$ and $q = \mathbf{1}_{[1,2]}$, the overlap measure should yield a small value, as the support of p and q are disjoint.
2. On the other hand, if $p = q$ then $d(p, q)$ should attain its maximum value, indicating high overlap.
3. Beyond support overlap, the measure should account for the densities of p and q. For instance, if $\text{supp}(p) = [0, 1]$ and $\text{supp}(q) = [0, 100]$, but
$$\int_1^{100} q(x)\,dx \approx \epsilon$$
where ϵ is very small, the overlap measure $o(p, q)$ should remain high, indicating a limited contribution of q outside $\text{supp}(p)$.

Given these considerations, we define the notion of overlap as

Definition 1. Let p, q real positive integrable functions, the notion of overlap is defined as

$$o(p, q) = \frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p(x), q(x))dx}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p(x), q(x))dx} \qquad (1)$$

and given $k \in \mathbb{R}$ such that $k \geq 1$, we define the measure $d_k(p,q)$ as

$$d_k(p,q) = k - o(p,q). \tag{2}$$

In figure 1 we provide a graphical example of the computation of the above formula for several variables from *Diabetes* dataset. In concrete, to compute integrals we use np.trapz from *Numpy* which applies the composite trapezoidal rule for approximating the value.

3.2 MATHEMATICAL AND GEOMETRICAL PROPERTIES OF $d_1(p,q)$

Remark 3.1. We now explore a set of properties and results that will help clarify the meaning of the overlap measure and provide the foundation for the main results of this paper, Theorems 3.2. and 3.1. The discussion will be done for $k = 1$ as it is the most relevant, all the results can be immediately adapted to $d_k(p,q)$.
First of all we note that since $p(x), q(x) \geq 0$ and

$$\text{supp}(p) \cap \text{supp}(q) \subset \text{supp}(p) \cup \text{supp}(q)$$

we can conclude that

$$0 \leq o(p,q) \leq 1$$

which implies that

$$0 \leq d_1(p,q) \leq 1.$$

If p=q then $d_1(p,q) = 0$ and if the two density estimations share a 0 measure support, as $\mu(\text{supp}(p) \cap \text{supp}(q)) = 0$, we obtain that $d_1(p,q) = 1$. Furthermore, we have a monotonicity property that takes into account densities: if $p(x), g(x), q(x)$ are three density estimations sharing the same support with the property that, almost everywhere,

$$p(x) \leq g(x) \leq q(x) \implies d_1(p,q) \leq d_1(g,q).$$

It is important to underline since we are working with integrals, we have to consider statements as $p < q$ or $p \neq q$ almost everywhere for a given measure μ. More properly, (2) is defined on

$$L^1(\mathbb{R}) \times L^1(\mathbb{R}) = \frac{\mathcal{L}^1(\mathbb{R}) \times \mathcal{L}^1(\mathbb{R})}{\sim_\mu}$$

which is the quotient space of $\mathcal{L}^1(\mathbb{R}) \times \mathcal{L}^1(\mathbb{R})$ with the equivalence relation induced by μ.

Finally, we are left to show that the notion (2) introduced, satisfies property 3 in Section 3.1.

Proposition 3.1. *Let p, q be two real probability distributions, supp$(p) = [a,b]$, supp$(q) = [a,c]$ such that $b < c$. Moreover, assume that*

$$|p(x) - q(x)| < \delta \quad \text{for } x \in [a,b], \qquad \int_b^c q(x) = \epsilon.$$

Then $o(p,q)$ attains values close to 1, and so $d_1(p,q)$ to 0.

Proof. See Appendix. □

There is an important corollary of previous Proposition. Several density function are defined on $\mathbb{R}$, even though they are concentrated on a small region as in the case of $N(\sigma, \mu^2)$ and $\Gamma(\alpha, \beta)$ (a list of possible Kernel functions can be found in paper [Weglarczyk, 2018]). Assume now that $p(x), q(x)$ are two density estimation defined on $\mathbb{R}$, then we can concentrate our overlap measure on a finite proper interval D without any relevant information loss. This is a result of the continuity of $d_1(p,q)$, Proposition 3.1 and the following trivial remark

Remark 3.2. Let $f \in L^1(\mathbb{R})$ then $\exists a \in \mathbb{R}_{\geq 0}$ such that $\forall \epsilon > 0$

$$\int_a^\infty |f(x)| < \epsilon, \quad \int_{-\infty}^{-a} |f(x)| < \epsilon.$$

Example 3.1. Assume the following density functions

$$p(x) = \begin{cases} \frac{1}{2} & \text{if } x \in [0,2] \\ 0 & \text{otherwise} \end{cases}, \quad q(x) = \begin{cases} \frac{19}{40} & \text{if } x \in [0,2] \\ \frac{1}{60} & \text{if } x \in (2,5] \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the overlap value is

$$d_1(p,q) = 1 - \frac{19/20}{1 + 1/20} = \frac{2}{21} \approx 0.095,$$

indicating that the two densities share highly the same domain.

In the appendix we prove the following important Proposition.

Proposition 3.2. *Given p, q real positive integrable functions such that $p \neq q$, then $d_k(p,q) \geq k - 1$.*

For $k \geq 2$ in the additional material we show how to use Proposition 3.2 along with some computation to derive the following

Theorem 3.1. *The measure $d_k(p,q)$ is a metric dissimilarity measure on Ω for $k \geq 2$, where*

$$\Omega = \{f \in L^1(\mathbb{R}) \mid \textit{supp}(f) \neq \emptyset \textit{ and } f(x) \geq 0\}$$

Our focus, as we stated at the beginning of the Section, is $k = 1$, for which we can obtain a stronger result for $d_1(p,q)$. In short, invoking previous results, using the fact that Jaccard distance satisfy triangle inequality and by dominated convergence, we can obtain the following central Theorem

Theorem 3.2. *$d_1(p,q)$ is a metric on Ω. Moreover, $d_k(p,q)$ is a dissimilariy measure on Ω for $k \in [1,2)$.*

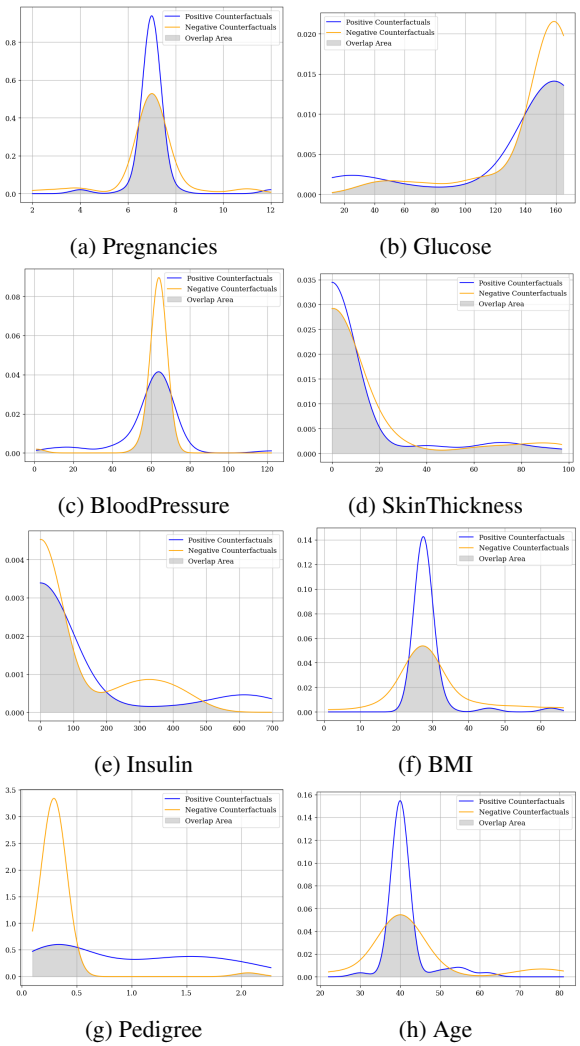
As a consequence, Theorems 3.1 and 3.2 convey that formula defined in (2) actually represents a way of calculating in mathematical terms the dissimilarity between two probability distributions. There is more, Theorem 3.2 establishes that $d_1(p, q)$ satisfies the axioms of a distance: non-negativity, symmetry, the triangle inequality, and that $d_1(p, q) = 0$ if and only if $p = q$. This result is significant because it implies that d_1 induces a metric topology on the space Ω. This topology provides a rigorous framework for comparing probability distributions. From a topological perspective, the metric d_1 ensures that the space Ω is metrizable. For instance, d_1 can be used to define neighborhoods of probability distributions, facilitating the study of their stability, convergence, or variation under perturbations.

4 EXPERIMENTAL RESULTS

To demonstrate the applicability of the proposed framework, we conduct experiments on numerical features on classical datasets: *Diabetes* and *Heart Disease*. To compute CFs we use the Diverse Counterfactual Explanations (DiCE) for ML library which is based on [Mothilal et al., 2020] that generates CF explanations for any ML model. This library also allow us to generate positive and negative CFs. We take inspiration from Paper [Wiegreffe and Pinter, 2019] and compare the results using $d_k(p, q)$ with local feature importance scores using DiCE. In order to provide a complete picture, for the analyzed variables we also compute local SHAP values ([Lundberg and Lee, 2017]) and local LIME values ([Ribeiro et al., 2016]). The code is avaialbe at `https://github.com/EddieConti/Local_Importance`.

4.1 DISSIMILARITY ANALYSIS ON DATASETS

In the datasets we performed iteratively the same analysis. In particular, we generate with DiCE library 50 elements for each set C^+ and C^-. Since the dissimilarity measure $d_1(p, q)$ is computed from the distributional discrepances between C^+ and C^-, it is sensitive to the generation of these two sets, which is a random process. In order to reduce the randomness, not only we computed a consistent number of element in each set, but we performed the analysis 10 times and took the mean value to obtain our dissimilarity measure $d_1(p, q)$ and the DiCE local feature importance. Below in Figure 2 we report the an example of the distribution of feature importance for the first instance of the test set for which we performed the analysis 100 times. Clearly, as we aggregated the results for the whole test set for each dataset to compare the different metrics (d_1, DiCE, LIME and SHAP), we reduced the number of iteration per instance to 10. In the case of *Diabetes* we use LogisticRegression, while for *Heart Disease* we use RandomForestClassifier. It is important to emphasise that the same analyses could be conducted with other types of models, as all measures to capture feature importance are model-agnostic.
First of all, let us visualize an instance of $d_1(p, q)$. We recall that $d_1(p, q)$ conveys how different are the two distributions: the more dissimilar are the two distributions, the higher the value $d_1(p, q)$ would be reflecting the fact that different values must be used for that variable to alter the output result.



(a) Pregnancies (b) Glucose

(c) BloodPressure (d) SkinThickness

(e) Insulin (f) BMI

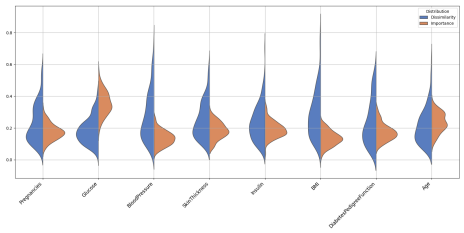
(g) Pedigree (h) Age

Figure 1: An instance of one iteration of the dissimilarity analysis for the *Diabetes* dataset based on the various entries of C^+ and C^- for the first element of the test dataset. In grey it is highlighted the overlap area.

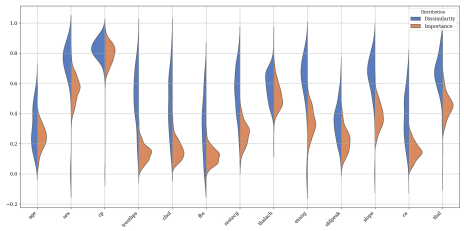
We are now ready to show the comparison between the dissimilarity measure and DiCE local scores. The results are summarized in Figure 2.

4.2 COMPARISON WITH OTHER MEASURES

In order to validate our framework and formulation, we computed for both dataset for the same model and instance SHAP values and LIME values. in the supplementary mate-



(a) Analysis on the *Diabetes* dataset.



(b) Analysis on the *Heart Disease* dataset.

Figure 2: The distribution of the feature importance scores for a specific instance over the iterations for $d_1(p, q)$ and DiCE scores.

rial, in Table 3 and Figure 4, we show an example for the first instance of the dataset of the various calculated values of feature importance using the 4 metrics.

Even tough, as we already mentioned, there is no ground-truth in feature importance scores and approaches ([Rajbahadur et al., 2022]), we use *FeatureAgreement* introduced in [Krishna et al., 2024] with $k = 4$, to evaluate the agreement/disagreement on the feature importance scores across the various metrics we used. Given two explanations (in our case a vector consisting of feature importance values) E_a and E_b, feature agreement is formulated as

$$\frac{|TF(E_a, k) \cap TF(E_b, k)|}{k} \tag{3}$$

where $TF(E, k)$ returns the set of top-k features of explanation E based on the magnitude of the feature importance values. Formula (3) it computes the fraction of common features between the sets of top-k features of two explanations. The results in Tables 4 and 5 in the supplementary material are the aggregation of (3) for the whole test set and therefore we report the mean with the confidence interval, computed as $2\sigma/\sqrt{n}$.

4.3 ANALYSIS OF RESULTS

The variability observed Figure 2 reflects both the use of a random counterfactual generator and the inherent nature of feature importance, which can be volatile due to the complexity of the relationships between inputs and outputs. To address this, we mitigated randomness by running multiple experiments and reporting the average values.

Interestingly, the variability appears to be influenced by the dataset's complexity. For instance, in the simpler Diabetes dataset in Figure 2a, the variability is relatively low compared to the Heart Disease dataset in Figure 2b, which involves more features. This suggests that the complexity of the dataset—and the underlying relationships we aim to infer—plays a crucial role in explainability.

The metric also leverages Kernel Density Estimation (KDE) to identify the distribution of the values analyzed, providing an alternative insight into the interplay between features and their impact. This aspect becomes particularly important in more complex datasets, where a single value might not fully capture the nuances of feature importance.

It is important to point out that the literature shows a general lack of absolute truth in the topic of attributing importance to features as an explanation of a model ([Harel et al., 2023], [Rajbahadur et al., 2022], [Jacovi and Goldberg, 2020]). In our experiments, we used two different datasets and calculated the feature agreement. In Table 4 for the *Diabetes dataset*, the d_1 metric demonstrates moderate alignment with the other metrics (DiCE, SHAP, and LIME). The values of agreement between d_1 and the other metrics range from 0.46 to 0.57. This suggests that in general d_1 identifies half the features of the other metrics. In contrast, DiCE, SHAP, and LIME exhibit a much higher degree of alignment with each other, with SHAP and LIME having the strongest agreement of 0.8393. This strong alignment indicates that these three metrics tend to highlight the same features with high consistency.

In Table 5 for the *Heart* dataset, we observe a similar trend, but with d_1 showing even more pronounced divergence. The agreement value between d_1 and the other metrics drops to around 0.33, which reinforces the idea that d_1 captures a different aspect of feature importance. This lower alignment suggests that d_1 is not simply mirroring the perspectives of the other metrics, but rather, it is focusing on different features or dimensions of the data.

The fact that d_1 captures distinct aspects of feature importance means it might be revealing information that the more aligned metrics miss. This makes d_1 a valuable tool for gaining additional perspectives compared to more conventional methods like SHAP, LIME, or DiCE.

In the next section, we will explore how d_1's distinctive perspective can be validated in terms of faithfulness and whether it provides a more meaningful or nuanced explanation of the data, offering a stronger case for its inclusion as a key metric for feature explanation.

4.4 EVALUATION OF FAITHFULNESS METRICS

The results obtained by using the Feature Agreement measure (3) are used to understand whether there are differences in stating which features are more important. However, in order to provide a true assessment of how the metrics perform, and in particular, the one we have introduced we estimate the *correctness* of an explanation. In particular, we follow the frameworks proposed in [Zhou and Shah, 2023] and [Chan et al., 2022], which assess the faithfulness of feature attribution evaluations. The key intuition behind these metrics is that altering an important feature should significantly impact the model's prediction and the magnitude of this impact reflects the quality of the explanation.
Among the various methods to estimate faithfulness, we adopt **comprehensiveness** and **sufficiency** introduced in [DeYoung et al., 2020]. Given an input $x = (x_1, \ldots, x_L)$ and an explanation $e = (e_1, \ldots, e_L)$, if we denote $\tilde{x}_e^{(l)}$ the input with l most important features removed according to e, comprehensiveness is defined as

$$k(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\tilde{x}_e^{(l)}) \tag{4}$$

where f is the function we want to explain. In simple terms, comprehensiveness measures how much the model prediction deviates from its original value when important features are removed sequentially (larger values indicate better explanations). If we denote $\hat{x}_e^{(l)}$ the input with the l most important features present, sufficiency is defined as

$$\sigma(x, e) = \frac{1}{L+1} \sum_{l=0}^{L} f(x) - f(\hat{x}_e^{(l)}) \tag{5}$$

and it measures the gap to the original model prediction that remains when features are successively inserted from the most important to the least. Therefore, a smaller value is desirable.

In our experiments, we compute (4) and (5) using the *predict-proba* function as f, again reporting the mean value and the confidence interval as in Section 4.2. The explanations are generated using d_1, DiCE, SHAP and LIME. Since we are working with local feature importance estimations, in order to obtain $\tilde{x}_e^{(l)}$ we mask the top relevant features, according to the explanation, using the mean ([Covert et al., 2021]). In addition, we plot the trend of *predict-proba* when gradually removing top 4 relevant features.

In the case of Comprehensiveness shown in Table 1 , for the *Diabetes* dataset, d_1 shows higher variability but is statistically superior when considering the standard errors of the other three metrics. While d_1 has a larger spread, it outperforms the other methods in terms of the overall mean, making it the preferable choice. For the Heart dataset, the results align with LIME showing better results compared to the other metrics.
Regarding Sufficiency in Table 2, d_1 generally produces better results with less variability overall. For *Diabetes dataset*, the high variability of DiCE, SHAP, and LIME makes it difficult to claim that our results are statistically significantly better, but the reduced mean value and variability highlight d_1 as a stronger performer. For Heart, both d_1 and LIME provide better results compared to the others.
In general, d_1 stands out as the more stable and reliable method across both datasets. However, for certain cases, as in the *Heart* dataset, LIME also demonstrates competitive performance.

Method	Diabetes	Heart
d_1	$\mathbf{0.5405 \pm 0.2559}$	$\mathbf{1.8017 \pm 0.0965}$
DiCE	0.0487 ± 0.1250	1.2355 ± 0.0849
SHAP	0.1300 ± 0.1423	1.3689 ± 0.0716
LIME	0.1277 ± 0.1328	$\mathbf{1.8163 \pm 0.0815}$

Table 1: Comprehensiveness results for Diabetes and Heart datasets. Higher values are preferable

Method	Diabetes	Heart
d_1	$\mathbf{-0.1288 \pm 0.1444}$	$\mathbf{2.2129 \pm 0.1357}$
DiCE	0.3942 ± 0.2758	2.7227 ± 0.1166
SHAP	0.3748 ± 0.2797	2.7044 ± 0.1307
LIME	0.3699 ± 0.2799	$\mathbf{2.2662 \pm 0.1346}$

Table 2: Sufficiency values with standard errors for the different methods. Smaller values are preferable

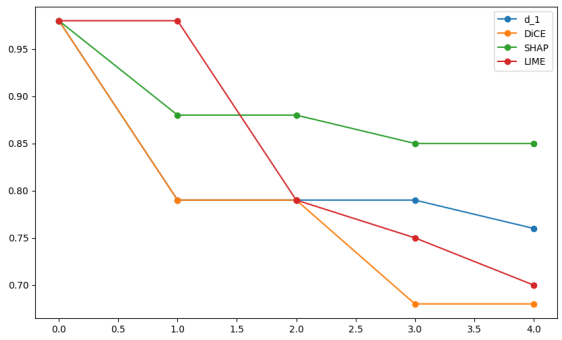


Figure 3: The trend of $k(x, e)$ when gradually removing the relevant features according the different explanations in *Heart Disease* dataset for the first entry of the test set. A downward trend indicates that features are being removed that actually played a relevant role in determining the class for the instance under consideration.

5 LIMITATIONS AND FUTURE PROSPECTS

The formula proposed in Definition 2 provides a method to quantify the dissimilarity between two data distributions, offering a novel perspective on how features contribute to a given output. Importantly, this approach can naturally extend beyond binary classifiers. For instance, in the case of a continuous output space $\mathbb{R}$, it is sufficient to define a subset $A \in \mathbb{R}$ where C^+ represents the data points for which $f(x) \in A$, and C^- corresponds to those with $f(x) \in A^c$.
In our analysis of feature importance using this framework, we treated feature importance independently, a common simplification in mathematical modeling (see [Mothilal et al., 2020], [Molnar, 2022], [Schölkopf et al., 2021]).
This work represents an initial exploration of this dissimilarity framework, with promising results across diverse datasets. However, several open questions remain. Future studies could investigate the impact of dependencies among features, the possibility of extending this framework to a multivariate setting using multivariate kernel density estimation (KDE), or addressing the challenges of extending the formula to categorical variables.
Moreover, our method is dependent on the CFs generation mechanism of DiCE and the KDE estimations. This represents a key starting point for future work, as our framework will benefit from improved CF generation methods and more accurate distribution estimations.

6 CONCLUSIONS

We would like to emphasize once again that, in the area of model explainability, there is no "absolute truth." Papers such as [Mishra and Sadia, 2023] and [Rajbahadur et al., 2022] highlight the usefulness of employing a plurality of viewpoints when explaining a model, as each method has its own strengths. In the case of the d_1 metric introduced in (2), Section 4.2 demonstrates that it captures different aspects compared to the more aligned metrics. Furthermore, when evaluating faithfulness, a critical measure of explanation correctness, Section 4.4 shows that the d_1 metric generally outperforms the other three metrics considered in terms of comprehensiveness and sufficiency.
Consequently, we believe that the method we have introduced represents a promising starting point for the study of local feature importance. It is well-founded in mathematical terms and offers valuable insights. Further research, involving additional datasets and exploring the role of the kernel in KDE or improvements in CF generation, will help emphasize the added value of the d_1 metric in greater depth

References

Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. Transparency and explainability of ai systems: From ethical guidelines to requirements. *Information and Software Technology*, 2023.

S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, and et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2017.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

Dennis Collaris, Hilde J.P. Weerts, Daphne Miedema, Jarke J. van Wijk, and Mykola Pechenizkiy. Characterizing data scientists' mental models of local feature importance. In *Nordic Human-Computer Interaction Conference*. Association for Computing Machinery, 2022.

Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res.*, 2021.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael Emmerich, and Heike Trautmann, editors, *Parallel Problem Solving from Nature – PPSN XVI*, pages 448–469. Springer International Publishing, 2020.

Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Trans. Evol. Comput.*, pages 182–197, 2002.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expecta-

tion. *Journal of Computational and Graphical Statistics*, 2015.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 2022.

Nimrod Harel, Uri Obolski, and Ran Gilad-Bachrach. Inherent inconsistencies of feature importance, 2023.

Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Danny Panknin, and Ahcène Boubekki. Explainable ai needs formal notions of explanation correctness, 2024.

Katrina Ingram. Ai and ethics: Shedding light on the black box. *The International Review of Information Ethics*, 2020.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *ArXiv*, abs/1905.11190, 2019.

Sven Kosub. A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, pages 36–38, 2019.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024.

Royden H. L. and Fitzpatrick P. M. *Real Analysis*. Pearson Modern Classic, fourth edition edition, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777. Curran Associates Inc., 2017.

Wilson Estecio Marcilio and Danilo Medeiros Eler. From explanations to feature selection: assessing shap values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020.

Alok Mishra and Halima Sadia. A comprehensive analysis of fake news detection models: A systematic literature review and current challenges. *Engineering Proceedings*, 2023.

Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *ArXiv*, abs/2111.00358, 2021. URL `https://api.semanticscholar.org/CorpusID:240354648`.

Christoph Molnar. *Interpretable Machine Learning*. Independently published, 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pages 607–617, 2020.

Gopi Krishnan Rajbahadur, Shaowei Wang, Gustavo A. Oliva, Yasutaka Kamei, and Ahmed E. Hassan. The impact of feature importance methods on the interpretation of defect classifiers. *IEEE Transactions on Software Engineering*, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Baron S. Explainable ai and causal understanding: Counterfactual approaches considered. *Minds and Machines*, pages 347–377, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021.

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 2021.

Arvind Narayanan Solon Barocas, Moritz Hardt. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

et al. Thomas Altmann. *Seminar on Limitations of Interpretable Machine Learning Methods*. Independently published, 2019. URL `https://slds-lmu.github.io/iml_methods_limitations/`.

Sahil Verma, John P. Dickerson, and Keegan E. Hines. Counterfactual explanations for machine learning: A review. *ArXiv*, abs/2010.10596, 2020.

Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*, 2017.

Stanislaw Weglarczyk. Kernel density estimation and its application. *ITM Web of Conferences, ISSN , e-ISSN 2271-2097, Irregular*, 2018.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20. Association for Computational Linguistics, 2019.

Yilun Zhou and Julie Shah. The solvability of interpretability evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, 2023.

Measuring Feature Importance Through Counterfactual Distributions (Supplementary Material)

A APPENDIX

Proof of Proposition 3.1

Proof. To show this, suppose the following scenario: $\text{supp}(p) = [a, b]$, $\text{supp}(q) = [a, c]$ such that $b < c$. Moreover, assume that

$$|p(x) - q(x)| < \delta \quad \text{for } x \in [a, b], \qquad \int_b^c q(x) = \epsilon.$$

Under this assumptions,

$$\begin{aligned} d_k(p, q) = k - \frac{\int_a^b \min(p(x), q(x))dx}{\int_a^c \max(p(x), q(x))dx} &= k - \frac{\int_a^b \min(p(x), q(x))dx}{\int_a^b \max(p(x), q(x))dx + \int_b^c q(x)dx} \\ &= k - \frac{\int_a^b \min(p(x), q(x))dx}{\int_a^b \max(p(x), q(x))dx + \epsilon} \\ &= k - \frac{\int_a^b \min(p(x), q(x))dx}{\int_a^b \max(p(x), q(x)) + \frac{\epsilon}{b-a}dx} \\ &= k - \frac{\int_a^b \min(p(x), q(x))dx}{\int_a^b \max(p(x) + \frac{\epsilon}{b-a}, q(x) + \frac{\epsilon}{b-a})dx} \end{aligned}$$

Now, from the last equation, we can obtain

$$k - \frac{1 - \epsilon}{1 - \delta(b - a)} \leq d(p, q) \leq k - \frac{1 - \epsilon - \delta(b - a)}{1 + \delta(b - a)}$$

since

$$\int_a^b q(x) = 1 - \epsilon.$$

The inequality can be obtained by observing that $|p(x) - q(x)| < \delta$ and so $q(x) - \delta < p(x) < q(x) + \delta$ which implies

$$\min(p(x), q(x)) \leq q(x), \quad \max(p(x), q(x)) \geq q(x) - \delta$$

and

$$\min(p(x), q(x)) \geq q(x) - \delta, \quad \max(p(x), q(x)) \leq q(x) + \delta.$$

Now, as desired, if ϵ and δ are relatively small, $d(p, q)$ takes on a value close to $k - 1$, i.e., indicating an high overlap. □

Derivation of Theorem 3.1

Proposition 6.1. *Given p, q real positive integrable functions such that $p \neq q$, then $d_k(p,q) > k-1$.*

Proof. Consider two functions p, q such that $p \neq q$. It is straightforward that if $\text{supp}(p) \neq \text{supp}(q)$ then $d(p,q) > k-1$ by how it is defined the measure of overlap. Indeed, in the worst case scenario $p = q$ in $\text{supp}(p) \cap \text{supp}(q)$, but as $\text{supp}(p) \neq \text{supp}(q)$, then

$$\text{supp}(p) \cap \text{supp}(q) \subsetneq \text{supp}(p) \cup \text{supp}(q)$$

which implies

$$\frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p(x), q(x))dx}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p(x), q(x))dx} < 1.$$

As a consequence, let us assume $\text{supp}(p) = \text{supp}(q) = D$. Since $p \neq q$, at least one among

$$A = \{x \in D \mid p(x) < q(x)\}, \quad B = \{x \in D \mid q(x) < p(x)\}$$

has a positive measure. Assume $\mu(A) > 0$, therefore

$$\frac{\int_D \min(p(x), q(x))dx}{\int_D \max(p(x), q(x))dx} = \frac{\int_A p(x) + \int_{D\setminus A} q(x)}{\int_A q(x) + \int_{D\setminus A} max(p(x), q(x))} < 1$$

because

$$\int_{D\setminus A} q(x) \leq \int_{D\setminus A} max(p(x), q(x)),$$
$$\int_A p(x) < \int_A q(x).$$

The case $\mu(B) > 0$ is analogous. □

Now let us recall the following definition.

Definition 2. A dissimilarity measure (DM) d on X is a function $d\colon X \times X \to \mathbb{R}$ such that

$$\begin{aligned}
&\exists\, d_0 \text{ s.t. } -\infty < d_0 \leq d(x,y) < \infty \quad \forall x, y \in X,\\
&d(x,x) = d_0 \quad \forall x \in X,\\
&d(x,y) = d(y,x) \quad \forall x, y \in X.
\end{aligned}$$

If in addition

$$\begin{aligned}
&d(x,y) = d_0 \iff x = y,\\
&d(x,z) \leq d(x,y) + d(y,z) \quad \forall\, x, y, z \in X,
\end{aligned}$$

d is called a metric DM on X.

Let us investigate the triangular inequality

$$d_k(p,q) \leq d_k(p,g) + d_k(g,q)$$

which in our case becomes

$$k - \frac{\int_{\text{supp}(p)\cap\text{supp}(r)} \min(p,r)}{\int_{\text{supp}(p)\cup\text{supp}(r)} \max(p,r)} \leq\ k - \frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p,q)}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p,q)} + k - \frac{\int_{\text{supp}(q)\cap\text{supp}(r)} \min(q,r)}{\int_{\text{supp}(q)\cup\text{supp}(r)} \max(q,r)}.$$

and simplified it becomes

$$\frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p,q)}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p,q)} + \frac{\int_{\text{supp}(q)\cap\text{supp}(r)} \min(q,r)}{\int_{\text{supp}(q)\cup\text{supp}(r)} \max(q,r)} - \frac{\int_{\text{supp}(p)\cap\text{supp}(r)} \min(p,r)}{\int_{\text{supp}(p)\cup\text{supp}(r)} \max(p,r)} \leq k$$

which is always satisfied if $k \geq 2$ because every term on the left is positive and bounded by 1. As a consequence of this result combined with Proposition 3.2, noticing that the symmetry is guaranteed and that d_0 in Definition 2 in our case is $k-1$, we have the following

Theorem 6.1. *The notion of overlap $d_k(p,q)$ is a metric dissimilarity measure on $\Omega \subset L^1(\mathbb{R})$ s.t. supp$(f) \neq \emptyset$ and $f(x) \geq 0$ for all $f \in \Omega$, for $k \geq 2$.*

Derivation of Theorem 3.2 According to our definition, we are left to analyze the case for $k \in [1,2)$ in (2). First of all we note that the Jaccard distance defined fro two sets A, B

$$d_J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

satisfies the triangle inequality (cfr. Paper Kosub [2019]). As a consequence, if we consider the functions $p(x) = \mathbf{1}_A, q(x) = \mathbf{1}_B, r(x) = \mathbf{1}_C$

$$d_1(p,q) \leq d_1(p,g) + d_1(g,q)$$

and of course if we add on the lhs $k-1$ and on the rhs $2(k-1)$ the inequality still holds, hence

$$d_k(p,q) \leq d_k(p,g) + d_k(g,q).$$

Now, the idea is to prove if the inequality is valid for general characteristic functions, therefore if

$$d_k(\alpha\mathbf{1}_A, \gamma\mathbf{1}_C) \leq d_k(\alpha\mathbf{1}_A, \beta\mathbf{1}_B) + d_k(\beta\mathbf{1}_B, \gamma\mathbf{1}_C) \quad \alpha,\beta,\gamma \in \mathbb{R}_{\geq 0}.$$

Without loss of generality we can assume that $\alpha \leq \beta \leq \gamma$, so we have to prove that

$$1 - \frac{\alpha|A \cap C|}{\gamma|A \cup C|} \leq 1 - \frac{\alpha|A \cap B|}{\beta|A \cup B|} + 1 - \frac{\beta|B \cap C|}{\gamma|B \cup C|}$$

which can be rewritten as

$$\frac{\beta|B \cap C|}{\gamma|B \cup C|} + \frac{\alpha|A \cap B|}{\beta|A \cup B|} - \frac{\alpha|A \cap C|}{\gamma|A \cup C|} \leq 1$$

but this is true since we know that

$$\frac{|B \cap C|}{|B \cup C|} + \frac{|A \cap B|}{|A \cup B|} - \frac{|A \cap C|}{|A \cup C|} \leq 1,$$

$$0 \leq \frac{\alpha}{\beta}, \frac{\alpha}{\gamma}, \frac{\beta}{\gamma} \leq 1.$$

We observe that it does not affect the order of α, β, γ since we have a minimum over a maximum and so the ratio is always less or equal than 1. Now, let us assume that $p(x), q(x), r(x)$ are simple functions, i.e.

$$p(x) = \sum_{i=1}^{k} a_i \mathbf{1}_{A_i}, \quad q(x) = \sum_{i=1}^{k} b_i \mathbf{1}_{B_i}, \quad r(x) = \sum_{i=1}^{k} c_i \mathbf{1}_{C_i}.$$

where $k \in \mathbb{N}$, $a_i, b_i, c_i \geq 0$ and A_i, B_i, C_i disjoint. It is trivial now, since the inequality holds for general characteristic functions, that

$$d_k(p,q) \leq d_k(p,g) + d_k(g,q).$$

Any function in $L^1(\mathbb{R})$ can be approximated with a sequence of simple functions (cfr. L. and M. [2017]). By dominated convergence theorem we can exchange the limit with the integral and so, combining it with the results for simple functions, we conclude that

$$d_k(p,q) \leq d_k(p,g) + d_k(g,q) \quad p,q,r \in L^1(\mathbb{R}).$$

As a conclusion, we can extend Theorem 3.1 to the case $k \geq 1$, and so we obtain Theorem 3.2.

ADDITIONAL TABLES AND FIGURES

In this section we provide the table for an example of the feature importance scores for the first entry of the *Diabetes* set along with the barplot visualization and the tables for the feature agreement computed for both *Diabetes* and *Heart* datasets.

Feature	d_1	DiCE	LIME	SHAP
Pregnancies	0.303746	0.118	0.124801	0.070353
Glucose	0.144349	0.476	0.424472	0.283522
BloodPressure	0.235353	0.138	-0.000319	0.012571
SkinThickness	0.419093	0.078	-0.023098	-0.020959
Insulin	0.186120	0.340	0.036109	0.021043
BMI	0.205907	0.232	-0.205435	-0.077856
PedigreeFunction	0.465278	0.076	-0.020605	-0.027462
Age	0.281309	0.098	0.010362	0.018021

Table 3: An instance of feature importance scores accross the various methods for the first entry of the *Diabetes* dataset. We can see, in this case, more moderate variability for d_1.

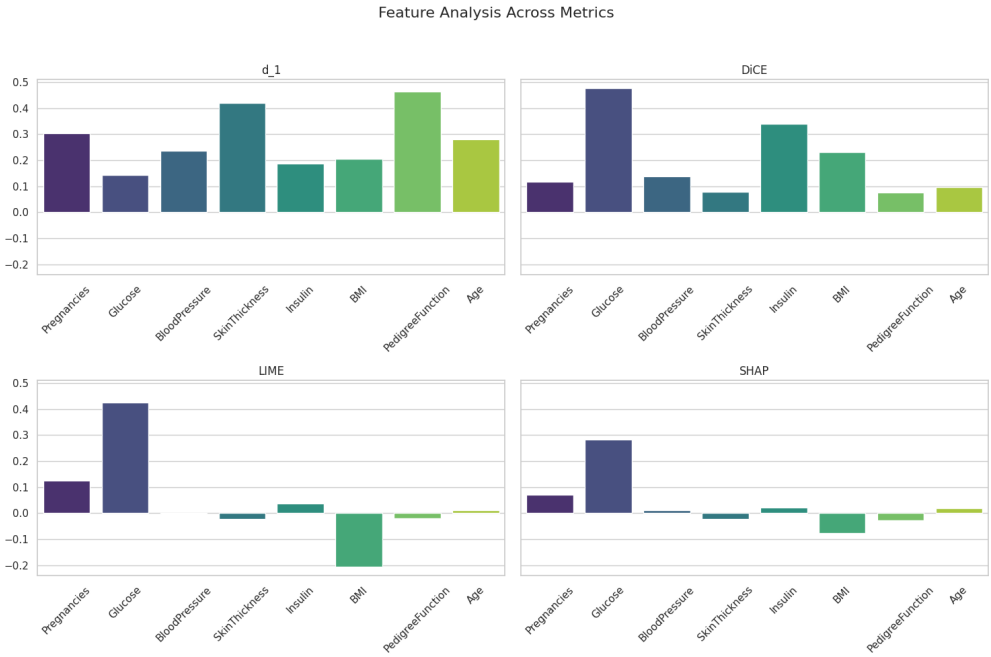


Figure 4: The visualization of Table 3 with barplots.

	d_1	DiCE	SHAP	LIME
d_1	1.0000 ±0	0.4610 ±0.0382	0.5714 ±0.0327	0.5341 ±0.0315
DiCE	0.4610 ±0.0382	1.0000 ±0	0.6721 ±0.0289	0.7110 ±0.0276
SHAP	0.5714 ±0.0327	0.6721 ±0.0289	1.0000 ±0	0.8393 ±0.0242
LIME	0.5341 ±0.0315	0.7110 ±0.0276	0.8393 ±0.0242	1.0000 ±0

Table 4: Feature Agreement Matrix for the *Diabetes* dataset.

	d_1	DiCE	SHAP	LIME
d_1	1.0000 ± 0	0.3451 ± 0.0454	0.3341 ± 0.0329	0.3549 ± 0.0381
DiCE	0.3451 ± 0.0454	1.0000 ± 0	0.5317 ± 0.0268	0.7134 ± 0.0218
SHAP	0.3341 ± 0.0329	0.5317 ± 0.0268	1.0000 ± 0	0.6890 ± 0.0274
LIME	0.3549 ± 0.0381	0.7134 ± 0.0218	0.6890 ± 0.0274	1.0000 ± 0

Table 5: Feature Agreement Matrix for the *Heart* dataset.