arXiv:2608.02697v1 [cs.LG] 3 Aug 2026

Measuring Explainer Stability via Attribution Separability

Eddie Conti[1] (✉), Álvaro Parafita[1], Axel Brando[1]

Barcelona Supercomputing Center, Barcelona, Spain econti@bsc.es
parafita.alvaro@gmail.com axelbrando@gmail.com

Abstract. Attribution methods (AMs) assign an importance score to each feature and are widely adopted to explain black-box models. However, most methods can produce variable attribution scores due to stochastic components in their definition. In this paper, we propose a distribution-based framework to capture the stability of attribution scores. In particular, our approach allows to understand the degree of separability in the ranked attribution vector and obtain the largest index for which a feature ranking remains reliable. We further extend this framework to compare AMs based on the robustness of their rankings across a dataset. Through experiments, we demonstrate how to apply our method to evaluate explainer stability. Overall, our approach provides a complementary criterion for evaluating the stability of AMs.

Keywords: Explainability · Attribution · Interpretability · Stability · Evaluation

1 Introduction

Explainable AI (XAI) aims to make the decision-making of models understandable to humans. In recent years, the adoption of complex architectures in high risk scenarios has raised concerns among researchers for its potential ethical and social implications, such as unfair treatment, bias amplification or discrimination [35], [18]. For these reasons, it is essential to understand how these models reason and generate predictions.

AMs or explainers [13] aim to detect which features are most relevant in generating the model output (feature attribution). In formal terms, we consider a model that maps input data to an output variable, $f : \mathcal{X} \to Y$, where $f(x)$ is the prediction for a given instance $x \in \mathcal{X} \subseteq \mathbb{R}^n$. In general, Y may represent either discrete classes or a continuous range of values; in this work, we focus on the common binary classification setting, where $Y = \{0, 1\}$. Then, an AM is a function $\alpha_f \colon \mathbb{R}^n \to \mathbb{R}^n$, denoted α for simplicity, where for $\alpha(x) = (\alpha_1, \ldots, \alpha_n)$, α_i represents the importance of feature x_i for the model's decision process. In the literature we refer to them as local explanations, because they analyze a specific instance of the dataset.

Over the past decade, the XAI community has developed a wide range of AMs [26,37], alongside several proposals for their evaluation [21,3]. Despite this

methodological richness, explanations suffer from various critical issues ranging from variability, inconsistency, and untruthfulness [22,20].

In this work we address a fundamental challenge in XAI: many AMs incorporate **stochastic components**—such as Monte Carlo sampling in SHAP [25] and LIME [30], or random perturbations in DiCE [27]—which can **yield explanations that vary significantly across runs**. This instability, acknowledged by several studies [5,34,12], raises concerns about reproducibility and reliability. As highlighted by Pawlicki [29], stability is a prerequisite for user trust and for the generalizability of insights derived from explanations.

In this work, we focus on explainers with stochastic components and introduce a novel distribution-based strategy to **quantify the stability of feature rankings**. In particular, we propose a metric that captures the maximum k for which the top-k ordering remains robust across multiple executions of an AM. From this proposal, we can measure a specific aspect of explanation stability, namely the degree of separability between feature importance values in the attribution vector. In particular, our contributions are:

- We introduce a novel approach to quantify the stability of AM rankings based on distributional analysis. Our approach, mathematically grounded, allows to assess to what extent the ranking is significant.
- We formally show—and validate empirically—how feature importance separation influences the overall stability of AMs.
- We conducted experiments demonstrating the application of our metric to analyze and compare AMs across different models and datasets.

2 Related Work

The issue of explanation stability has been investigated deeply in the literature. A first set of approaches aligns with the taxonomy identified by Carvalho et al. [9] and Gawantka et al. [14], which define stability respectively as "how similar the explanations are for similar instances" (an idea also supported by Alvarez-Melis [4]) or "high stability of an explanation is observed when the explanation undergoes minimal changes in response to minor variations in unimportant features of the data instance." Consistent with this definition, metrics have been developed that model perturbations at the input level. Alvarez-Melis et al. [5] formalize the first stability metric for local explanation methods, arguing that explanations should be robust to local perturbations of the input. Agarwal et al. [2] deem that the approach of Alvarez-Melis et al. [5] does not leverage potentially meaningful information—such as the model's internal representations—for evaluating stability, and implicitly assumes that f behaves similarly on inputs x and x' that are close. As a consequence, authors propose several *relative stability* metrics: Relative Input Stability, which measures the relative distance between explanations w.r.t the distance between inputs; Relative Representation Stability, which instead uses internal representations of x and x'; and Relative Output Stability, which computes the relative distance replacing x and x' with their logits outputs. Similarly, Butt et al. [8] focus on consistency in feature importance

values across perturbed inputs of a given instance x: they generate a set of perturbations and aggregate the resulting variations in feature importance across these perturbations.

From a different perspective but still at input-level, Pawlicki [29] study the stability of SHAP under three types of input perturbations: randomly shuffling values, randomly inserting the median, and adding Gaussian noise, to investigate how attribution values are affected. The author concludes that shuffling features has a more significant impact on SHAP stability, and that the stability of explanations varies across datasets, likely due to differences in complexity and characteristics—as also observed by Butt et al. [8].

A different approach [38] proposes the concept of explanation stability accounting for perturbations directly on α. Summarizing their idea, they consider $\alpha_i \in \{0, 1\}$ (i.e., features are either relevant or not relevant), and stability means that the prediction does not change even if more explanatory features are added to $\alpha(x)$. To account for small perturbations, the authors consider to alter few entries of α. Hence the robustness of explanations is studied by analyzing whether small modifications in feature selection affect the model's prediction.

In this scenario, to the best of our knowledge, we introduce a novel approach that does not rely on perturbations or modifications of the explainer. Instead, we analyze the stability of an AM with a **distributional analysis**. Specifically, we estimate the distribution of each feature's importance scores across runs and employ a **metric to quantify the overlap between these distributions**. A lower overlap indicates that the feature rankings are more distinct and, consequently, more reliable. This allows us to assess whether the resulting ranking is significant and to what extent, while also enabling comparison across AMs.

3 Problem Definition

We now formalize our notion of stability. Specifically, we seek to answer the following:

***Research Question**: given an ordered vector of feature importances produced by an AM, to what extent can we trust the resulting ranking? More specifically, how confident can we be that the first feature is truly more important than the second, the second more important than the third, and so on?*

To address this issue, we model each feature attribution as a distribution rather than a single deterministic value. Given the attribution vector $\alpha(x) = (\alpha_1(x), \ldots, \alpha_n(x))$ and its associated dimension-wise distributions $p_1, \ldots, p_n$, we define the induced ranking $r(x)$ as the permutation of features sorted according to their expected attribution values:

$$\tilde{\alpha}_i = \mathbb{E}_{\alpha_i \sim p_i}[\alpha_i] \quad \Rightarrow \quad r(x) = \text{argsort}(\tilde{\alpha}(x)). \tag{1}$$

The key question then becomes whether the ordering induced by $r(x)$ is reliable. In particular, we seek to quantify how distinguishable consecutive features

in the ranking are and, consequently, how much confidence can be placed in their relative positions. For this purpose, we perform pairwise comparisons between the distributions p_i and p_j associated with each pair of features in $r(x)$. For every pair, we compute a distance $d(p_i, p_j)$, where d is a metric taking values in a bounded interval (e.g., $[0, 1]$) and measuring the degree of separation between the two distributions.

Our notion of attribution separability provides a flexible way to analyze a key component of stability from multiple perspectives. Within this general framework, we focus on two main aspects: (i) the **evolution of separability** across the attribution vector, and (ii) the notion of ***k*-stability**, which identifies the largest prefix of top-ranked features whose ordering can be considered reliable. In Appendix A.1, we show how the framework naturally **extends to binary explanations or group-based explanations** (i.e., $\alpha \in [0, 1]^n$).

Finally, we emphasize that the framework uses the **assumption** of local independence, that is, the features used by a model can be considered independent within a neighborhood around the sample to explain. This assumption is common in the AM literature: see [25,30,36,31,16,23].

3.1 The choice of distance metric

Although the framework can use any metric, in this paper we use the following measure to quantify the distance between two distributions, introduced by Conti et al. [10]:

$$d(p, q) = 1 - \frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p(x), q(x))dx}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p(x), q(x))dx}. \tag{2}$$

This measure is the continuous generalization of the Jaccard distance, a metric widely adopted to measure the overlap between sets. We employ this d since it is proved to be a metric, with values in $[0, 1]$, and particularly suitable to capture distributional discrepancies.

To validate our framework we prove the following.

Proposition 1 (Well-posedness in the ideal case). *Let $x \in \mathbb{R}^n$ be a point to be explained with ground-truth importances $f_1 > \cdots > f_n$. Consider an AM α with a stochastic component that produces, at each run $s = 1, \ldots, m$, an attribution vector $\alpha^{(s)} = \{\alpha_1^{(s)}, \ldots, \alpha_n^{(s)}\}$. Denote by $p_{i,m}$ the empirical probability density function associated with the values $\{\alpha_i^{(s)}\}_{s=1}^m$.*

Assume that, for each i, $p_{i,m}$ converges in distribution to the Dirac measure δ_{f_i} as $m \to \infty$. Then, for the distance metric $d(\cdot, \cdot)$ defined in (3), we have

$$d(p_{i,m}, p_{i+1,m}) \to 1 \quad \text{as } m \to \infty, \quad \forall i \in \{1, \ldots, n-1\}.$$

This result, proved in Appendix A.4, confirms that the metric is well-posed in the sense that, if an AM were perfectly faithful and its stochasticity vanished asymptotically, then it would achieve the maximum possible stability.

3.2 *k*-Stability

While the evolution of separability can be directly analyzed by tracking how the values of $d(p_i, p_j)$ change across the attribution vector, the notion of k-stability requires a formal definition. In many applications, an AM is used to explain a prediction by highlighting the most relevant features; thus, we are particularly interested in assessing how reliably the top portion of the ranking can be trusted. This motivates the following definition:

Definition 1. *Given a threshold $l \in [0, 1]$ and an instance x, we say that the attribution method $\alpha(x)$ is* k-stable*, where*

$$k := \arg\max_{s \in [n]} \text{ such that } d(p_i, p_j) \geq l \quad \forall\, i, j \in \mathcal{A}_s(r(x)),$$

and $\mathcal{A}_s(r(x))$ denotes the set of the top s indices in the ranked list $r(x)$.

In simple terms, k represents the largest prefix of the ranking $r(x)$ for which all consecutively-ranked pairs (p_i, p_j) have a distance $d(p_i, p_j)$ exceeding the threshold l. This formulation allows us to assess the degree of separability among the top-ranked features in the attribution ranking. While it is possible that features lower in the ranking (i.e., the "tail") may also be well-separated in terms of d, the focus is to ensure that the top positions are genuinely distinguishable because we expect those features to be more relevant.

The notion of *k-stability* induces a total order over the set of AMs for the instance x, $\mathbb{A}_x := \{\alpha(x) : \alpha \text{ is an AM}\}$.

Definition 2. *Let $\alpha_1(x), \alpha_2(x) \in \mathbb{A}_x$ be two attribution maps for the same instance x, with corresponding stability levels k_1 and k_2. Given a fixed threshold $l \in [0, 1]$, we say that $\alpha_1(x)$ is more stable that $\alpha_2(x)$ if $k_1 \geq k_2$.*

Combining Definition 1 and Definition 2, two AMs are compared by the proportion of instances in $\mathcal{X}$ for which one exhibits higher k-stability than the other; the AM that prevails on the majority of instances is considered more stable on average.

4 Validation of the proposed stability metric

Before turning to the experiments, we need to clarify that the adopted strategy is directly aligned with the research question we aim to address. In particular, we want to show that the proposed metric—and therefore the resulting k-stability values—are sensitive to the degree of separability between feature importance values, which in turn reflects the reliability of the induced ranking.

To support this intuition, for this demonstration, in Appendix A.4 we consider a simplified setting where attribution scores are assumed to follow Gaussian distributions centered at their ground-truth values. Under this assumption, the distance between two distributions can be expressed as:

$$d(p, q) = 1 - \frac{2(1 - \Phi(z))}{2\Phi(z)} = \frac{2\Phi(z) - 1}{\Phi(z)},$$

where Φ is the cumulative distribution function of the standard Gaussian and $z = \frac{\Delta}{2\sigma_i}$, with Δ denoting the difference between the ground-truth attribution values of two features and σ_i their shared standard deviation. From this expression, it follows that when $\Delta \to 0$, we obtain $d(p, q) \to 0$, while in the limit $\Delta \to \infty$, we have $d(p, q) \to 1$.

This shows that the proposed distance is intrinsically sensitive to the separation between feature importances: the more distinguishable the ground-truth attributions are, the higher the induced distance. This supports the interpretation that attribution separability is a fundamental component of stability, and justifies its role in the definition of k-stability.

This theoretical result is supported by empirical evidence. As detailed in the supplementary material (A.5), we construct synthetic datasets with known ground-truth feature importances to evaluate the behavior of the metric in a controlled setting. Figure 1 shows that, when keeping the range of importance values fixed ($[0, 0.4]$) and increasing the number of features, feature importances naturally begin to overlap, increasing their stability, with our metric capturing this phenomenon: we observe a clear downward trend in k-stability, indicating that the metric is directly affected by the reduced separability.

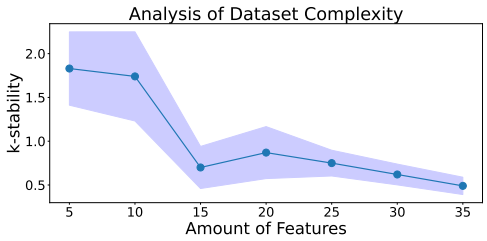


Fig. 1: When feature importance values fall in a certain range (here $[0, 0.4]$), and we increase the number of features, the k-stability value naturally decreases.

To further confirm this behavior, we also decrease separability by sampling ground-truth importances from progressively narrower intervals ($[0.4, 1]$, $[0.6, 1]$, and $[0.8, 1]$) for 5 and 10 features. As expected, smaller intervals lead to lower separability, which in turn decreases k-stability. See the supplementary material for these results. Overall, these controlled experiments provide evidence that the proposed metric responds coherently to changes in feature importance separability, supporting its adequacy for the research question.

5 Experiments

In this section, we empirically evaluate the proposed framework introduced in Section 3 and demonstrate the applicability of the stability metric for analyzing AMs. Experimental code will be made publicly available upon acceptance.

Dataset. We use four benchmark datasets: Diabetes [33], Heart Disease [19], Mobile [17], and Churn [32].

Methods. We consider four explanation methods: **LIME**, **SHAP**, **DiCE**, and a random baseline (**RAND**) used for calibration. Given the connection between LIME and SHAP [26], we ensure consistency by using the same background dataset and the same `num_samples` parameter. Across all experiments, we evaluate 50 instances from each test set and run each explainer 50 times per data point.

Stability Estimation. Our approach relies on estimating the distributions $p_1, \ldots, p_n$ of attribution scores. In practice, we approximate each distribution from the previous 50 explainer runs using Gaussian Kernel Density Estimation (KDE) with Silverman's rule for bandwidth selection. Based on these estimates, we assess whether the ranking induced by $\tilde{\alpha}(x)$ (see Equation 1) reflects meaningful distributional differences leveraging the metric d.

We design the following experiments to assess the applicability of our metric:

- Section 5.1: we compute the stability of attribution vectors for the four explainer methods on a Random Forest (RF) model.
- Section 5.2: we use k-stability to compare the robustness of different explainers across datasets and models (RF, Logistic Regression (LG), Decision Tree (DT), and Support Vector Machine (SVM)).
- Section 5.3: we study the effect of the number of runs on stability estimation.
- Section 5.4: we report additional experiments left to the appendices.

5.1 Analysis of Separability in the Attribution Vector

In this section, we analyze how separability evolves across the attribution vector. The first experiment computes the values of d_1 for consecutive feature pairs in the ranked attribution vector, across all four datasets. More precisely, given the attribution vector sorted in decreasing order of importance $(a_1, a_2, \ldots, a_m)$, we compute $d_1(a_i, a_{i+1})$ for all $i = 1, \ldots, m-1$. The resulting trends, shown in Figure 2, summarize how well the attribution distributions are separated at different positions in the ranking.

For the first two datasets, SHAP produces a constant value of 1, reflecting the fact that its attributions are near-deterministic (meaning, due to the large sample size, the attribution values are identical across runs, up to decimal precision) and therefore maximally separable. In the more complex Heart and Mobile datasets, SHAP still exhibits higher separability overall, but its curve peaks toward the tail, suggesting increased overlap among the less relevant features. As expected, all attribution methods outperform the RAND baseline. These results show how the framework allows a detailed examination of separability, highlighting differences across AMs and datasets.

We extend the separability analysis by computing the pairwise distance between all features and averaging the resulting heatmaps over 50 instances. Overall, Figure 3 highlights that separability tends to increase with the positional distance between features, and the tail shows less separability than the more

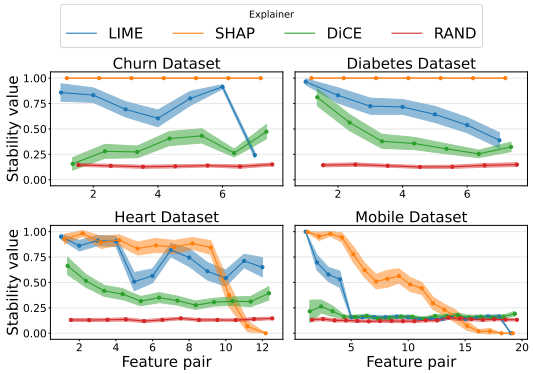


Fig. 2: The stability values for the analyzed datasets calculated on the ordered feature pairs with the confidence interval $\pm 2\sigma/\sqrt{n}$, where $n = 50$ is the number of instances. Here the x-axis refers to the feature pair (α_i, α_{i+1}). In the case of Churn and Diabetes, SHAP is achieves constantly the value of 1.

important variables. This graph provides an overview of the separability across all features, including those that are far apart in the ranking, thereby offering a more detailed view of the structure of the attribution vector.

5.2 Comparing k-stability values of AMs

Following Definition 2, we compare the four AMs across all datasets. In Table 1 and Table 2 we report the results for the RF and SVM models (LG and DT left to the appendix). Both tables are consistent: first, we note that the last column being consistently equal to 1 confirms that the RAND explainer serves as a reliable baseline, thereby validating the calibration of our analysis. Second, SHAP generally outperforms the other methods in the datasets under analysis, achieving higher proportions of larger k-stability values than LIME, DiCE, and RAND. LIME also performs better than DiCE and RAND. In contrast, DiCE consistently shows lower stability, reaching values close to the random baseline, particularly in the Mobile dataset.

Overall, our findings are consistent with previous studies reporting that SHAP tends to be more stable than LIME [8,28]. However, according to the No Free Lunch Theorem for explainability, no AM can be expected to consistently outperform all others across datasets and models. For this reason, our goal is not to establish the superiority of a particular method, but rather to exemplify the application of a well-defined framework to quantitatively assess a specific property: the reliability of the feature ranking induced by its attribution scores.

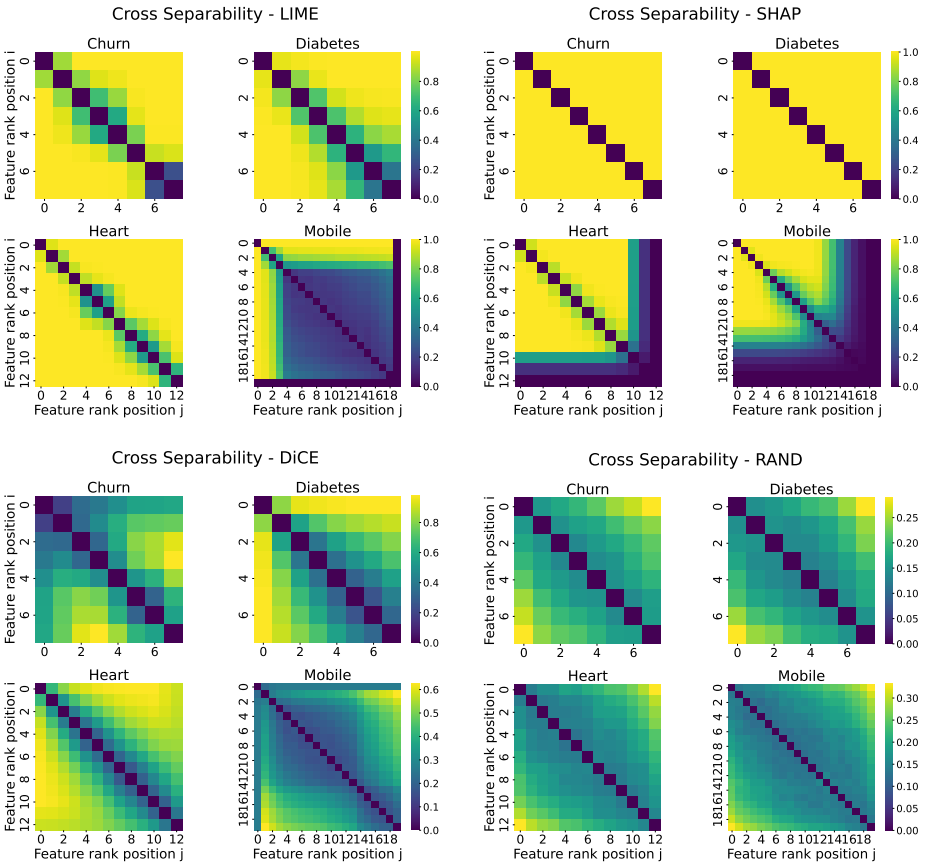


Fig. 3: Average heatmap for the 50 points analyzed with different explanation methods. Each entry represents the symmetric distance $d(p_i, p_j)$. From left to right: LIME, SHAP; DiCE and RAND.

Table 1: Pairwise comparison of k-stability values across datasets for the RF model. For each dataset we report the proportion of instances for which the row method Lime (L), SHAP (S), DiCE (D) and RAND (R) achieves a k-stability greater than or equal to the column method.

	Churn				**Diabetes**				**Heart**				**Mobile**			
	L	S	D	R	L	S	D	R	L	S	D	R	L	S	D	R
L	1.00	0.00	0.98	1.00	1.00	0.00	0.96	1.00	1.00	0.36	0.88	1.00	1.00	0.16	1.00	1.00
S	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.74**	**1.00**	**0.98**	**1.00**	**0.96**	**1.00**	**1.00**	**1.00**
D	0.20	0.00	1.00	1.00	0.30	0.00	1.00	1.00	0.32	0.16	1.00	1.00	0.18	0.04	1.00	1.00
R	0.18	0.00	0.96	1.00	0.10	0.00	0.34	1.00	0.20	0.16	0.64	1.00	0.00	0.02	0.82	1.00

Table 2: Pairwise comparison of k-stability values across datasets for the SVM model.

	Churn				**Diabetes**				**Heart**				**Mobile**			
	L	S	D	R	L	S	D	R	L	S	D	R	L	S	D	R
L	1.00	0.00	0.97	1.00	1.00	0.00	1.00	1.00	1.00	0.30	0.83	1.00	1.00	0.17	1.00	1.00
S	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.80**	**1.00**	**0.97**	**1.00**	**0.97**	**1.00**	**1.00**	**1.00**
D	0.17	0.00	1.00	1.00	0.23	0.00	1.00	1.00	0.37	0.13	1.00	1.00	0.17	0.03	1.00	1.00
R	0.17	0.00	0.60	1.00	0.13	0.00	0.63	1.00	0.23	0.07	0.63	1.00	0.00	0.02	0.73	1.00

5.3 The impact of the number of runs

Finally, we analyze how the distribution of d_1 for the ordered feature pairs evolves when varying the number of runs of the method. In Figure 4, we report the stability values along with confidence intervals $\pm 2\sigma/\sqrt{n}$. The results indicate that the proposed metric exhibits a **limited sensitivity to the number of runs** for LIME, SHAP and DiCE, suggesting that reliable estimates can be obtained without excessive computational cost. A mild deviation is observed in the Mobile dataset when using only 5 runs, where the distances between feature distributions appear overestimated. Only in the case of RAND more runs are necessary, but this is due to the random behavior of the explainer

5.4 Other Experiments

We have decided to keep the main body of the paper focused on the key experiments and insights. However, for the sake of completeness, we have conducted numerous additional experiments left to the appendix, which we summarize briefly here:

- **Group-based stability** (Appendix A.1): in many cases, the assignment vector is binary, i.e. $\alpha(x) \in [0,1]^n$. In this scenario, we show how our strategy can be used to identify a set of features that can be distinguished from the rest.
- **Hyperparameter discussion** (Appendix A.3): we analyze the impact of our framework's parameters; specifically, the threshold l and the KDE kernel. We show how reducing l increases the k-stability values (as we require less separability between distributions), as expected, whilst highlighting Silverman's rule as a robust strategy for inferring the distributions.
- **Synthetic validation of mathematical results** (Appendix A.5): we complement Section 4 by performing the analysis of k-stability in synthetic settings with 5 and 10 features. In general, the results confirm that greater separability of features leads to higher values of k-stability.

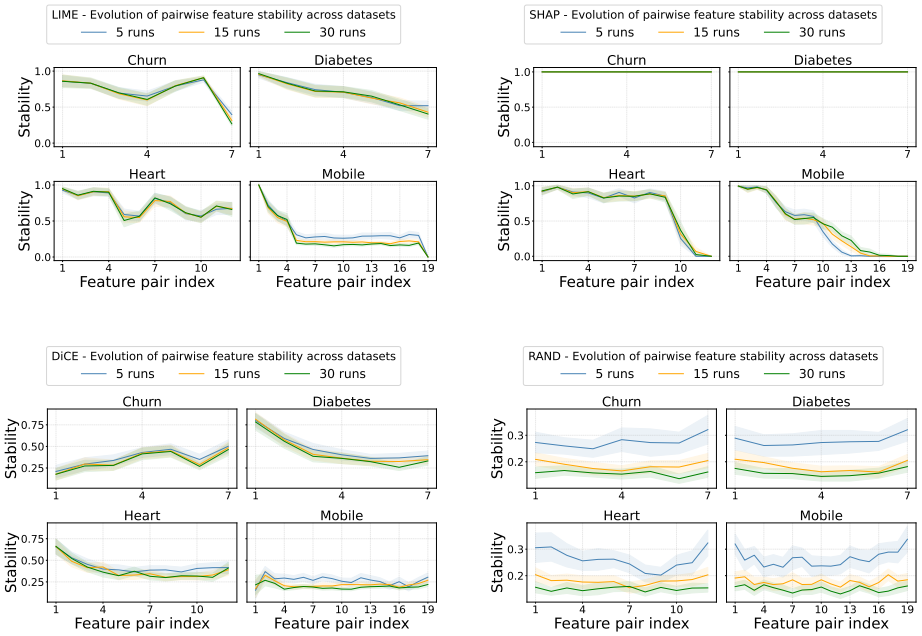


Fig. 4: Stability values across different numbers of runs for LIME (top-left), SHAP (top-right), DiCE (bottom-left), and RAND (bottom-right) on the considered datasets. For Churn and Diabetes, SHAP yields perfect stability since the attribution vectors were quasi-deterministic.

6 Limitations and Further Details

In this work we proposed an alternative method to quantify the stability of AMs, which in turn provides a meaningful way to compare explainers, as shown in Section 5.2. Our contribution is not to establish which AM is superior, but rather to introduce a comparative metric that enables more nuanced evaluations on a specific aspect of the instability. It is important to stress that no AM can be considered universally more appropriate than others [24,26]. Instead, the community has emphasized the need for multi-level and in-depth comparisons [6,7,3], and our work contributes to this perspective.

Naturally, our approach comes with some limitations. First, our empirical evaluation focuses on controlled settings, namely classification tasks on standard tabular benchmark datasets. While this choice allows us to isolate and study the properties of the proposed metric, future work should investigate its applicability to other data modalities, such as images, text, and time series.

Second, the framework requires selecting both a density estimation procedure and the threshold parameter l. Although we discuss these design choices and

their impact in the appendix, their selection may influence the resulting stability estimates.

Finally, as discussed in Appendix A.4, low k-stability values do not necessarily imply that an attribution method is unfaithful. Moreover, even when an AM faithfully reflects the model's decision process, stochastic components in the explanation procedure may induce variability across runs, leading to overlapping attribution distributions and consequently lower stability scores. Therefore, k-stability alone cannot capture the overall quality of an explanation, but rather should be interpreted as a measure of ranking reliability and eventually combined with other metrics.

7 Conclusions

In this work, we have formalized a novel method for measuring the stability of AM rankings, providing an additional dimension for explanation quality. Our experiments show that our method achieves consistent results even with a low number of explainer executions, and highlight SHAP as the most k-stable explainer for the datasets tested.

The strategy we have defined is not limited to measuring ranking stability. It can also be employed to study feature-pair separability within an attribution vector and to identify groups of features that are significantly more prominent than the rest (group-based explanations).

This flexibility allows to measure explainer stability from a variety of perspectives, which is fundamental for determining the right explainer for any given task. In line with the XAI community, which has highlighted the importance of a multi-dimensional evaluation of explanation quality, our approach provides a principled methodology to measure ranking reliability, thereby complementing explainer assessment based on other considerations, such as faithfulness or simplicity.

Acknowledgements

Eddie Conti has been partially supported by the predoctoral grant FI-STEP (2025 STEP 00108) from the Research and University Department of the Generalitat de Catalunya and cofunded by the European Social Fund Plus. Álvaro Parafita acknowledges his AI4Science fellowship within the "Generacion D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR. Axel Brando received funding from the Horizon Europe Programme under the AI4DEBUNK Project (https://www.ai4debunk.eu), grant agreement num. 101135757.

References

1. Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence **298**, 103502 (2021)
2. Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., Lakkaraju, H.: Rethinking stability for attribution-based explanations. ArXiv (2022)
3. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: Openxai: Towards a transparent evaluation of model explanations. Advances in neural information processing systems **35**, 15784–15799 (2022)
4. Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems **31** (2018)
5. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods (2018)
6. Belaid, M.K., Bornemann, R., Rabus, M., Krestel, R., Hüllermeier, E.: Compare-xai: Toward unifying functional testing methods for post-hoc XAI algorithms into a multi-dimensional benchmark. In: World Conference on Explainable Artificial Intelligence. pp. 88–109. Springer (2023)
7. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. Data Mining and Knowledge Discovery **37**(5), 1719–1778 (2023)
8. Butt, T., Iqbal, M.: Towards reliable explainable ai: A novel stability metric for trustworthy interpretations (04 2025). `https://doi.org/10.13140/RG.2.2.28294.69448`
9. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
10. Conti, E., Parafita, Á., Brando, A.: Cid: Measuring feature importance through counterfactual distributions. In: Northern Lights Deep Learning Conference 2026 (2025)
11. Covert, I., Lee, S.I.: Improving kernelshap: Practical shapley value estimation using linear regression. In: International conference on artificial intelligence and statistics. pp. 3457–3465. PMLR (2021)
12. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems **32** (2019)
13. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM (2019)
14. Gawantka, F., Just, F., Savelyeva, M., Wappler, M., LÃ¤ssig, J.: A novel metric for evaluating the stability of xai explanations. Advances in Science, Technology and Engineering Systems Journal **9**, 133–142 (2024). `https://doi.org/10.25046/aj090113`
15. Goldwasser, J., Hooker, G.: Stabilizing estimates of shapley values with control variates. In: World Conference on Explainable Artificial Intelligence. pp. 416–439. Springer (2024)
16. Goldwasser, J., Hooker, G.: Statistical significance of feature importance rankings. In: The 41st Conference on Uncertainty in Artificial Intelligence (2024)
17. Hu, N.: Classification of mobile phone price dataset using machine learning algorithms. In: 2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML). pp. 438–443. IEEE (2022)

18. Ingram, K.: AI and ethics: Shedding light on the black box. The International Review of Information Ethics (2020)
19. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart Disease. UCI Machine Learning Repository (1989)
20. Ju, Y., Zhang, Y., Yang, Z., Jiang, Z., Liu, K., Zhao, J.: Logic traps in evaluating attribution scores. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5911–5922 (2022)
21. Kadir, M.A., Mosavi, A., Sonntag, D.: Evaluation metrics for xai: A review, taxonomy, and practical applications. In: 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES). pp. 000111–000124. IEEE (2023)
22. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. Transactions on Machine Learning Research (2024)
23. Leemann, T., Kirchhof, M., Rong, Y., Kasneci, E., Kasneci, G.: When are post-hoc conceptual explanations identifiable? In: Uncertainty in Artificial Intelligence. pp. 1207–1218. PMLR (2023)
24. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4768–4777. Curran Associates Inc. (2017)
26. Molnar, C.: Interpretable Machine Learning. Independently published, 2 edn. (2022), `https://christophm.github.io/interpretable-ml-book`
27. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20) (2020)
28. Nayebi, A., Tipirneni, S., Foreman, B., Reddy, C.K., Subbian, V.: An empirical comparison of explainable artificial intelligence methods for clinical data: a case study on traumatic brain injury. In: AMIA annual symposium proceedings. vol. 2022, p. 815 (2023)
29. Pawlicki, M.: Towards quality measures for xai algorithms: Explanation stability. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2023)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: "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. pp. 1135–1144. Association for Computing Machinery (2016)
31. Samiei, S., Baratalipour, N., Yadav, P., Roy, A., He, D.: Addressing stability in classifier explanations. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 1920–1927. IEEE (2021)
32. Sharma, N., Awasthi, A., Parmar, D.S., Chouhan, T., Singh, A., Rawat, D.: Predictive analytics for customer retention: A catboost model for churn detection. In: 2025 International Conference on Networks and Cryptology (NETCRYPT). pp. 1024–1029. IEEE (2025)
33. Sigillito, V.: Diabetes dataset. `https://www.kaggle.com/datasets/` (1990), research Center, RMI Group Leader, Applied Physics Laboratory, The Johns Hopkins University, Laurel, MD, USA
34. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2019)

35. Solon Barocas, Moritz Hardt, A.N.: Fairness and Machine Learning: Limitations and Opportunities. MIT Press (2023)
36. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems **41**(3), 647–665 (2014)
37. Wang, Y., Zhang, T., Guo, X., Shen, Z.: Gradient based feature attribution in explainable AI: A technical review. arXiv preprint arXiv:2403.10415 (2024)
38. Xue, A., Alur, R., Wong, E.: Stability guarantees for feature attributions with multiplicative smoothing. ArXiv (2023)
39. Zhou, Z., Hooker, G., Wang, F.: S-lime: Stabilized-lime for model explanation. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 2429–2438 (2021)

A Supplementary Material

A.1 Group-based Stability

In this section we extend the notion of separability of attribution scores to *binary explanations*, where an AM identifies a set of relevant features rather than producing only a ranking. Let $G \subseteq [n]$ denote the selected features and $\bar{G}$ its complement. We define the *group-based stability* of G as:

$$\text{GroupStab}(G) = f\big(\{d(p_i, p_j) : i \in G, ; j \in \bar{G}\}\big),$$

where f is an aggregation function such as the mean, median, or minimum. This formulation captures different notions of separability between relevant and non-relevant features.

The framework naturally supports several explanation strategies, including top-k features, threshold-based selection, and cumulative attribution criteria. In the special case where G contains the top-k ranked features, we recover a condition analogous to k-stability:

$$d(p_i, p_j) \geq l \qquad \forall i \in G, ; \forall j \in \bar{G},$$

for a threshold $l \in [0, 1]$. This allows identifying the largest value of k for which the selected group remains sufficiently separable.

In Figure 5 we apply this framework to LIME explanations in the case of the RF model. For each instance and each value of k, we verify whether the top-k features satisfy the separability condition with $l = 0.9$. The result is binary: a value of 1 is assigned if the condition holds and 0 otherwise. The resulting curves report, across 50 instances, the proportion of explanations certified as separable.

The analysis highlights several trends. First, evaluating separability as a function of k provides a principled way to determine the size of the explanation according to the desired level of stability. Second, datasets such as Churn and Heart exhibit non-monotonic certification curves, suggesting that while the precise ordering of important features may be unstable, groups of features can still remain jointly separable (e.g., in Churn with 6 features we have a certification rate $\approx 60\%$, while with 4 of $\approx 30\%$). Finally, the Mobile dataset shows consistently low certification rates, likely due to the intrinsic complexity of the dataset and the difficulty of the explainer in identifying clearly separable features.

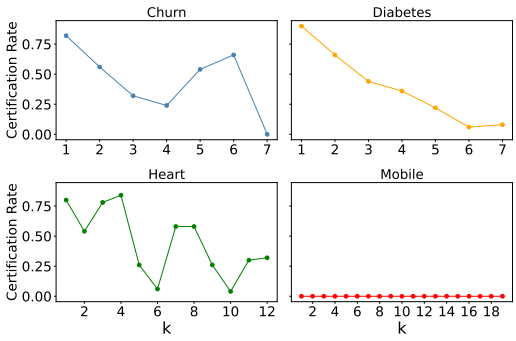


Fig. 5: Proportion of instances in which the LIME top-k features form a separable group according to the group-stability condition.

A.2 Additional comparison experiments

In this section, we supplement the experiments comparing attribution methods. In particular, we find that SHAP produces more separable rankings for DT and LG as well (Table 3, Table 4), achieving higher values in our stability metric

Table 3: Pairwise comparison of k-stability values across datasets for the DT model.

	Churn				**Diabetes**				**Heart**				**Mobile**			
	L	S	D	R	L	S	D	R	L	S	D	R	L	S	D	R
L	1.00	0.00	0.93	1.00	1.00	0.00	1.00	1.00	1.00	0.23	1.00	1.00	1.00	0.23	1.00	1.00
S	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.90**	**1.00**	**0.97**	**1.00**	**0.93**	**1.00**	**1.00**	**1.00**
D	0.50	0.00	1.00	1.00	0.20	0.00	1.00	1.00	0.27	0.03	1.00	1.00	0.07	0.00	1.00	1.00
R	0.30	0.00	0.60	1.00	0.17	0.00	0.63	1.00	0.20	0.03	0.90	1.00	0.00	0.00	0.80	1.00

Table 4: Pairwise comparison of k-stability values across datasets for the LR model.

	Churn				**Diabetes**				**Heart**				**Mobile**			
	L	S	D	R	L	S	D	R	L	S	D	R	L	S	D	R
L	1.00	0.00	0.97	1.00	1.00	0.00	1.00	1.00	1.00	0.13	0.90	1.00	1.00	0.20	1.00	1.00
S	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**	**1.00**	**0.97**	**1.00**	**0.90**	**1.00**	**1.00**	**1.00**
D	0.20	0.00	1.00	1.00	0.30	0.00	1.00	1.00	0.60	0.10	1.00	1.00	0.07	0.03	1.00	1.00
R	0.13	0.00	0.60	1.00	0.20	0.00	0.63	1.00	0.60	0.16	0.90	1.00	0.00	0.00	0.80	1.00

A.3 Hyperparameter discussion

This section briefly analyzes the sensitivity of the proposed k-stability measure to two key hyperparameters: the density estimator and the threshold l, using the Diabetes dataset and LIME as an illustrative case.

Regarding the density estimator, the use of Gaussian kernels with Silverman's rule is considered a robust and standard choice. Other bandwidth values are tested in Figure 6. Although different bandwidths produce slightly different inferred distributions of attribution values, the resulting k-stability remains very similar across settings, except for $bandwidth = 1$, which yields lower stability. Together with previous findings showing that fewer samples are sufficient to separate attribution values, this suggests that k-stability is not strongly dependent on the modeling capacity of the density estimator, supporting Silverman's rule as a safe default.

The threshold l has a more pronounced effect. It determines how large the distributional distance between feature attributions must be to ensure a stable ordering. Lower values of l allow more overlap between feature importance distributions, potentially leading to less reliable rankings, while higher values impose stricter separability requirements. There is no universally optimal choice of l, as it depends on the application context (e.g., more conservative settings in high-stakes domains such as medical diagnosis, versus more relaxed exploratory analyses). In this work, $l = 0.9$ is used. In Figure 6, different threshold values are compared, showing that lower l values lead to higher k-stability, since the stability condition is more easily satisfied.

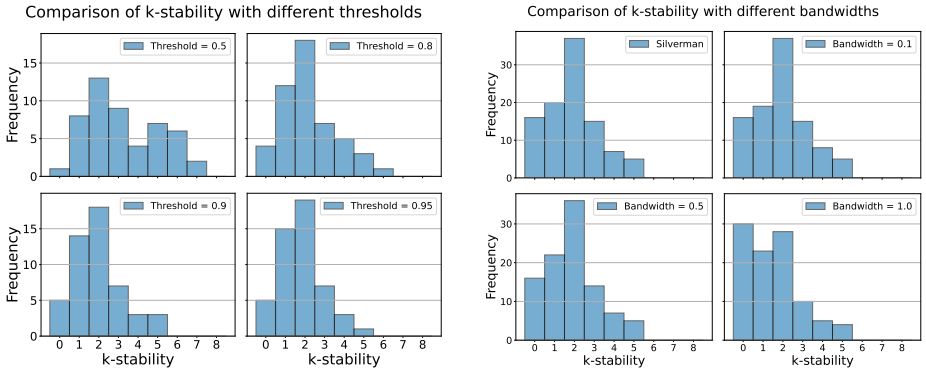


Fig. 6: Distributional comparison of k-stability values under different parameter settings. On the left, different threshold values l (0.5, 0.8, 0.9, and 0.95) are considered. On the right, different bandwidth values are compared: Silverman's rule, 0.1, 0.5, and 1.

A.4 Theoretical Guarantees and Limitations

We begin by establishing a preliminary result that guarantees the well-posedness of our stability metric in the ideal scenario where AMs converge to deterministic feature importances. Throughout the following discussion, we will assume that the distribution of attribution scores is approximately Gaussian. This hypothesis not only enables us to derive analytical results for the proposed metric, but it is also consistent with assumptions commonly adopted in the related literature [11,39,15,16,3,1]. Moreover, in the following discussion, we evaluate explainers with their hyperparameters fixed; this is essential to model the variance of the attribution scores, since increasing the sampling parameters (e.g., in SHAP, by raising the number of coalitions or background points) would eventually lead to quasi-deterministic attributions, thus removing the variability that our analysis aims to characterize.

We recall the definition of the metric d:

$$d(p,q) = 1 - \frac{\int_{\text{supp}(p)\cap\text{supp}(q)} \min(p(x), q(x))dx}{\int_{\text{supp}(p)\cup\text{supp}(q)} \max(p(x), q(x))dx}, \tag{3}$$

Proposition 2 (Well-posedness in the ideal case). *Let $x \in \mathbb{R}^n$ be a point to be explained with ground-truth importances $f_1 > \cdots > f_n$. Consider an AM α with a stochastic component that produces, at each run $s = 1, \ldots, m$, an attribution vector $\alpha^{(s)} = \{\alpha_1^{(s)}, \ldots, \alpha_n^{(s)}\}$. Denote by $p_{i,m}$ the empirical probability density function associated with the values $\{\alpha_i^{(s)}\}_{s=1}^m$.*

Assume that, for each i, $p_{i,m}$ converges in distribution to the Dirac measure δ_{f_i} as $m \to \infty$. Then, for the distance metric $d(\cdot,\cdot)$ defined in (3)*, we have*

$$d(p_{i,m}, p_{i+1,m}) \to 1 \quad \text{as } m \to \infty, \quad \forall i \in \{1, \ldots, n-1\}.$$

Proof. Since $p_{i,m} \xrightarrow{d} \delta_{f_i}$ and $p_{i+1,m} \xrightarrow{d} \delta_{f_{i+1}}$ as $m \to \infty$, with $f_i \neq f_{i+1}$, the two limiting measures are mutually singular. By definition of the distance $d(\cdot,\cdot)$, since they are separated, this implies

$$d(\delta_{f_i}, \delta_{f_{i+1}}) = 1.$$

By the continuity of d, we obtain

$$d(p_{i,m}, p_{i+1,m}) \to d(\delta_{f_i}, \delta_{f_{i+1}}) = 1,$$

which proves the claim.

This result confirms that our metric is well-posed in the sense that, if an AM were perfectly faithful and its stochasticity vanished asymptotically, then it would achieve the maximum possible stability under the d metric. We will focus later on the case of having a couple of feature importances f_i, f_j to be equal and analyze how the metric d behaves in this situation.

Remark 1 (Limitations of k-stability). The proposed metric has two complementary limitations that clarify its interpretation.

First, k-stability may attain its maximum value even for completely unfaithful attribution methods. In particular, even when the learned attributions converge to arbitrary but distinct constants (i.e., degenerate distributions), the resulting pairwise separability can be maximal, leading to perfect stability. This shows that k-stability alone cannot distinguish faithful from unfaithful explanations, and should not be interpreted as a direct measure of correctness.

Second, even in the case of faithful and unbiased attribution methods, stability is fundamentally constrained by stochasticity and by the intrinsic closeness of the true feature importances. When the separation between ground-truth values is small relative to the estimator variance, distributional overlap can prevent the stability condition from being satisfied, even asymptotically.

Let us now formalize the previous remark in a simplified setting. Consider an unbiased and faithful AM $\alpha : \mathbb{R}^n \to [0,1]^n$ (which requires to rescale in the case of DiCE, or taking the absolute value in LIME and SHAP). We focus on two features f_i, f_{i+1}, with $f_i > f_{i+1}$ (the case $f_i = f_{i+1}$ is studied in A.6), and assume equal variances[1] $\sigma_i^2 = \sigma_{i+1}^2$. As a first step, we observe that if the variances coincide, the two Gaussian distributions intersect at exactly one point. Let $p(x) = \mathcal{N}(f_i, \sigma_i^2)$ and $q(x) = \mathcal{N}(f_{i+1}, \sigma_{i+1}^2)$. Solving $p(x) = q(x)$ is equivalent to requiring $\log(p(x)) - \log(q(x)) = 0$. From the expression of the Gaussian density, this leads to

$$\log(p(x)) = \log\Big(\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-f_i)^2}{2\sigma_i^2}}\Big) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2}\frac{(x-f_i)^2}{\sigma_i^2},$$

then

$$\log(p(x)) - \log(q(x)) = \frac{1}{2}\Big(-\log(\sigma_i^2) + \log(\sigma_{i+1}^2) - \frac{(x-f_i)^2}{\sigma_i^2} + \frac{(x-f_{i+1})^2}{\sigma_{i+1}^2}\Big). \quad (4)$$

Since $\sigma_i^2 = \sigma_{i+1}^2$, (4) reduces to

$$\begin{aligned} -\frac{1}{\sigma_i^2}(x^2 + f_i^2 - 2xf_i) + \frac{1}{\sigma_i^2}(x^2 + f_{i+1}^2 - 2xf_{i+1}) &= 0 \\ \frac{1}{\sigma_i^2}(2xf_i - f_i^2 - 2xf_{i+1} + f_{i+1}^2) &= 0 \\ x &= \frac{f_i + f_{i+1}}{2}. \end{aligned}$$

Returning to our main discussion, we now consider the distance d (3) applied to p and q. Since the support of a Gaussian distribution is $\mathbb{R}$, the integrals extend

[1] Assuming equal variances simplifies the exposition without altering the substance of the argument. In the general case, different variances would lead to additional intersection points and require a case-by-case analysis. However, the core intuition concerning the separability between the two distributions remains unchanged.

over the whole real line (we omit this notation for brevity). In this case, the metric reduces to

$$d(p,q) = 1 - \frac{\int \min(p(x), q(x))}{\int \max(p(x), q(x))}.$$

In this setting there exists only one intersection $\tilde{x} = (f_i + f_{i+1})/2$, and since $f_i > f_{i+1}$

$$d(p,q) = 1 - \frac{\int_{-\infty}^{\tilde{x}} p(x) + \int_{\tilde{x}}^{\infty} q(x)}{\int_{-\infty}^{\tilde{x}} q(x) + \int_{\tilde{x}}^{\infty} p(x)} = 1 - \frac{(1 - \int_{\tilde{x}}^{\infty} p(x)) + (1 - \int_{-\infty}^{\tilde{x}} q(x))}{\int_{-\infty}^{\tilde{x}} q(x) + \int_{\tilde{x}}^{\infty} p(x)} \tag{5}$$

Now, let us set

$$\Delta := f_i - f_{i+1}, \qquad z := \frac{\Delta}{2\sigma_i} = \frac{f_i - f_{i+1}}{2\sigma_i}.$$

Using the Gaussian cumulative distribution function Φ, the relevant integrals are

$$\int_{-\infty}^{\tilde{x}} p(x) = \Phi\big((\tilde{x} - f_i)/\sigma\big) = \Phi(-z) = 1 - \Phi(z),$$
$$\int_{\tilde{x}}^{\infty} p(x) = 1 - \Phi(-z) = \Phi(z)$$

To compute the integrals concerning $q(x)$, we can observe that

$$q(x) = \mathcal{N}(f_{i+1}, \sigma_i^2) = N(f_i - \Delta, \sigma_i^2) = p(x + \Delta).$$

As a consequence,

$$\int_{-\infty}^{\tilde{x}} q(x)\,dx = \int_{-\infty}^{\tilde{x}} p(x + \Delta)\,dx = \int_{-\infty}^{\tilde{x}+\Delta} p(t)\,dt = \Phi\Big(\frac{\tilde{x} + \Delta - f_i}{\sigma_i}\Big) = \Phi(z)$$

since

$$\tilde{x} + \Delta - f_i = \frac{f_i + f_{i+1}}{2} + (f_i - f_{i+1}) - f_i = \frac{f_i - f_{i+1}}{2} = \frac{\Delta}{2}.$$

Now, $\int_{\tilde{x}}^{\infty} q(x) = 1 - \Phi(z)$ and so (5) becomes:

$$d(p,q) = 1 - \frac{2(1 - \Phi(z))}{2\Phi(z)} = \frac{2\Phi(z) - 1}{\Phi(z)}.$$

As immediate observations we have:

- For $\Delta \to 0^+$ (namely $z \to 0^+$), we have $\Phi(z) \to 1/2$ and so
$$d(p,q) \to \frac{2 \cdot \frac{1}{2} - 1}{\frac{1}{2}} = 0.$$
- For $\Delta \to \infty$ (namely $z \to \infty$), we have $\Phi(z) \to 1$ and so
$$d(p,q) \to \frac{2 \cdot 1 - 1}{\frac{1}{2}} = 1.$$

In conclusion, we have formally shown in a simple setting how the separation affects the metric d and so the stability of the ranking of the AM.

A.5 Experiments on synthetic datasets

To validate the theoretical analysis discussed above, we designed a synthetic experiment. The dataset consists of $n = 10$ features with ground-truth coefficients evenly spaced in the interval $[r_{min}, 1]$, where r_{min} is a parameter (indicating the minimum of the range) controlling the closeness of the features:

$$\mathbf{f} = (f_1, \ldots, f_n), \quad f_i = r_{min} + (1 - r_{min})\frac{n-i}{n-1}, \quad i = 1, \ldots, n.$$

The input values $X \in \mathbb{R}^{N \times n}$ for each feature are sampled independently from a standard normal distribution:

$$X_{ij} \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, N, \; j = 1, \ldots, n,$$

where $N = 2000$ is the total number of samples. The continuous target $y \in \mathbb{R}^N$ is generated as a linear combination of the features with added Gaussian noise:

$$y_i = \sum_{j=1}^{n} f_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, N,$$

where we set $\sigma = 0.5$. The binary target is then obtained by thresholding y at 0: $y_i^{\text{bin}} = \mathbf{1}\{y_i > 0\}$ for $i = 1, \ldots, N$.

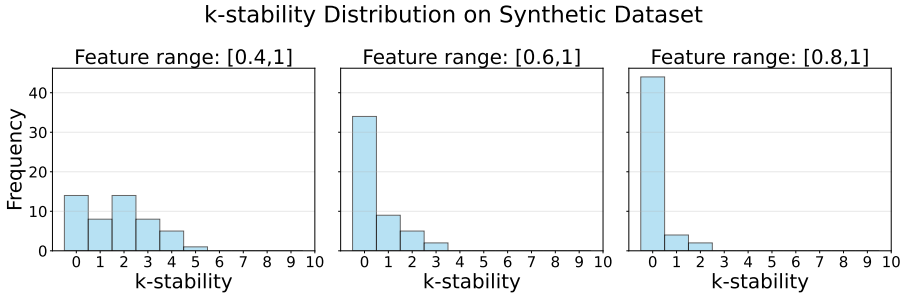
For the analysis, a RF classifier was trained on the synthetic datasets, and LIME explanations were computed 50 times for each of the first 50 test instances. The results are reported in Figure 7. We consider two settings with 10 and 5 features, respectively, while progressively reducing the separation between the ground-truth feature importances by sampling the coefficients from the intervals $[0.4, 1]$, $[0.6, 1]$, and $[0.8, 1]$.

In both cases, k-stability decreases as the feature importances become closer, confirming that the proposed metric is sensitive to attribution separability. Moreover, the 5-feature setting generally yields higher stability values than the 10-feature setting, suggesting that increasing the number of features further reduces the separability of the induced rankings. Overall, these controlled experiments support the theoretical analysis and validate the ability of k-stability to capture changes in feature importance separation.

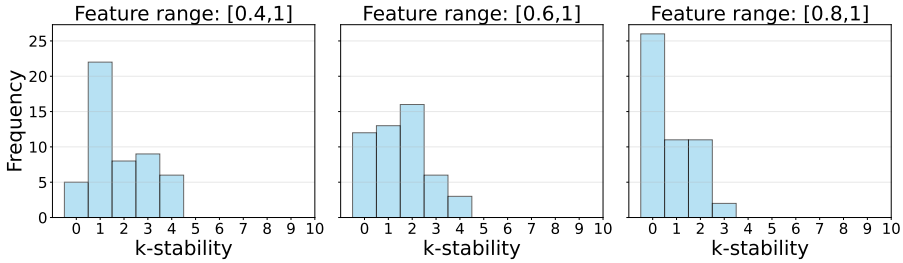
A.6 The case $f_i = f_{i+1}$

To complement the previous analysis, we now focus our attention to the case of an explainer converging in mean to the same value for two attributions and what insights we can derive from the metric d. Let $p(x) = \mathcal{N}(\mu, \sigma_1^2)$, $q(x) = \mathcal{N}(\mu, \sigma_2^2)$ with $\sigma_1^2 \neq \sigma_2^2$. Following the same strategy, the points of intersection are solutions to

$$-\frac{1}{2}\log\sigma_1^2 - \frac{(x-\mu)^2}{2\sigma_1^2} = -\frac{1}{2}\log\sigma_2^2 - \frac{(x-\mu)^2}{2\sigma_2^2},$$



(a) Synthetic dataset with 10 features.



(b) Synthetic dataset with 5 features.

Fig. 7: Validation of k-stability on synthetic datasets. Ground-truth feature importances are sampled from progressively narrower intervals, reducing their separation. In both settings, k-stability decreases as feature importances become closer, while the 5-feature dataset exhibits generally higher stability than the 10-feature dataset.

which can be written as

$$(x-\mu)^2\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)=\log\frac{\sigma_2^2}{\sigma_1^2} \implies (x-\mu)^2=\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}.$$

Now, since the right term is always positive, the solutions are:

$$x=\mu\pm\sqrt{\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}}.$$

We are now ready to compute the metric d in this case. Let us assume without loss of generality $\sigma_1<\sigma_2$ (if $\sigma_1=\sigma_2$ then the two distributions are identical and therefore their feature importances indistinguishable). Let us denote the two intersections as

$$x_{1,2}=\mu\mp R, \qquad R=\sqrt{\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\frac{\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2}}.$$

Since, the narrower Gaussian p prevails near μ and the wider one q prevails at the tails, we have the natural partition

$$\int \min(p,q) = \int_{-\infty}^{x_1} p(x) + \int_{x_1}^{x_2} q(x) + \int_{x_2}^{\infty} p(x),$$
$$\int \max(p,q) = \int_{-\infty}^{x_1} q(x) + \int_{x_1}^{x_2} p(x) + \int_{x_2}^{\infty} q(x).$$

By using the normal cumulative distribution function Φ and exploiting the symmetry of $x_{1,2} = \mu \mp R$ w.r.t, μ, we have, letting $r_1 = R/\sigma_1$ and $r_2 = R/\sigma_2$,

$$\int_{-\infty}^{x_1} p(x)\,dx = \Phi(-r_1) = 1 - \Phi(r_1), \qquad \int_{-\infty}^{x_1} q(x)\,dx = \Phi(-r_2) = 1 - \Phi(r_2);$$
$$\int_{x_1}^{x_2} q(x)\,dx = 2\Phi(r_2) - 1, \qquad \int_{x_1}^{x_2} p(x)\,dx = \Phi(r_1) - \Phi(-r_1) = 2\Phi(r_1) - 1;$$
$$\int_{x_2}^{\infty} p(x)\,dx = 1 - \Phi(r_1), \qquad \int_{x_2}^{\infty} q(x)\,dx = 1 - \Phi(r_2).$$

Now adding up the terms:

$$\int \min(p,q) = \big(1 - \Phi(r_1)\big) + \big(2\Phi(r_2) - 1\big) + \big(1 - \Phi(r_1)\big) = 1 + 2\big(\Phi(r_2) - \Phi(r_1)\big),$$

$$\int \max(p,q) = \big(1 - \Phi(r_2)\big) + \big(2\Phi(r_1) - 1\big) + \big(1 - \Phi(r_2)\big) = 1 + 2\big(\Phi(r_1) - \Phi(r_2)\big).$$

As a consequence, setting $A = \Phi(r_1) - \Phi(r_2)$,

$$d(p,q) = 1 - \frac{\int \min(p,q)}{\int \max(p,q)} = 1 - \frac{1-2A}{1+2A} = \frac{4A}{1+2A}.$$

As immediate observations we have:

- If $\sigma_1 \to \sigma_2$ then $R \to 0$, therefore $r_1, r_2 \to 0$ and $A \to 0$. As a consequence $d(p,q) \to 0$: the densities are almost equivalent and the stability metric yields small values.
- If $\sigma_2/\sigma_1 \to +\infty$ (or symmetrically $\sigma_1/\sigma_2 \to 0^+$), then we have

  $$r_1 \to +\infty, \qquad r_2 \to 0,$$

  hence $A = \Phi(r_1) - \Phi(r_2) \to 1 - \frac{1}{2} = \frac{1}{2}$ and so

  $$d(p,q) \to \frac{4 \cdot (1/2)}{1 + 2 \cdot (1/2)} = 1.$$

  The intuition is that a distribution sharply concentrated around its mean reflects a confident feature importance assignment, whereas a distribution with larger variance indicates greater uncertainty. The metric d correctly captures this difference by assigning a large distance between the two distributions. However, when the corresponding features have the same importance, their relative order remains indistinguishable. Therefore, while the metric is well-posed, this example highlights a limitation of k-stability.