Where Does Hardness Come From? An Instance Hardness Analysis of Substance-Use Prediction in Brazilian Adolescents

Thalles Mota[1], Leonardo Arruda Vilela Garcia[1], Claudia Martins[1]

Instituto de Computação – Programa de Pós-graduação em Computação Aplicada
Universidade Federal de Mato Grosso (UFMT) – Cuiabá – MT – Brazil
thallesleandromota@gmail.com

Abstract. The intrinsic difficulty of a classification dataset is routinely measured, yet its origin is rarely examined: is an instance hard because of the features that describe it, or because of the target to be predicted? We study this through Instance Hardness (IH) in a real, highly imbalanced problem — predicting frequent substance use among 165,838 Brazilian adolescents from the 2019 PeNSE survey, a setting that is hard by construction: the data are self-reported and the positive class has a prevalence of only about 4%. Hardness is quantified per instance in two independent ways — empirically, as the misclassification propensity across a committee of four learning architectures, and geometrically, from local class overlap (kDN), which uses no model. Three findings emerge across ten data partitions. First, IH is largely but not uniformly model-invariant (Spearman $\rho > 0.94$ among the four committee members; ≈ 0.79–0.97 against a fifth, neural architecture), which justifies treating hardness as a property of the data rather than of any single learner. Second, local overlap explains only part of the difficulty: kDN predicts true errors with an AUC of ≈ 0.73, against ≈ 0.95 for empirical IH, leaving a substantial residual attributable to non-local structure such as label noise or feature interactions. Third, through a controlled ablation, hardness responds to the target, not the features — removing an entire block of strong features leaves the hardness distribution essentially unchanged, whereas enlarging the target raises mean hardness roughly 2.5×. In this domain, difficulty is governed by what is predicted, not by which variables are used to predict it; past a reasonable feature set, the performance ceiling is more productively addressed by reconsidering the target than by further feature engineering.

CCS Concepts: • **Computing methodologies** → **Supervised learning by classification**.

Keywords: class overlap, data complexity, imbalanced classification, instance hardness

1. INTRODUCTION

In supervised classification, the difficulty of a problem is usually summarized by aggregate performance: an accuracy, an F-score, an AUC. Yet a growing body of data-centric work argues that such aggregate views hide where the difficulty actually lies, and that examining the data at the level of individual instances is more informative [Smith et al. 2014; Torquette et al. 2024]. Instance Hardness (IH), introduced by Smith et al. [2014], formalizes this idea: it quantifies, for each example, its propensity to be misclassified by a diverse set of learners, and a companion family of hardness measures attempts to explain *why* particular instances are hard [Lorena et al. 2019; Torquette et al. 2024]. Across many datasets, class overlap has been identified as a principal source of this difficulty [Smith et al. 2014].

What this literature rarely asks, however, is where the hardness of a *given* problem comes from in the first place. Hardness is treated as a measured property of a fixed dataset, but a dataset is not given by nature: it is the result of choosing which variables to record and which target to predict. An instance may be hard because the available *feature space* fails to separate it, or because the *target* itself conflates heterogeneous behaviors into one label. The distinction is practically decisive. If hardness

stems from the features, collecting more informative variables can reduce it; if it stems from the target, no amount of feature engineering will reduce it, and the performance ceiling is a property of the prediction task rather than of the model. This motivates our research question: *to what extent is the hardness of a prediction problem attributable to the feature space, and to what extent to the definition of the target?*

We investigate this question in a real, imbalanced problem — predicting frequent substance use among Brazilian adolescents from the 2019 PeNSE survey [IBGE 2021] — chosen precisely because it is difficult: the data are self-reported, behavioral, and the positive class is rare. Our contributions are threefold. First, we show that hardness in this domain is largely *model-invariant*: IH estimates from four architectures agree strongly and remain substantially concordant with a fifth architecture from a distinct model family, which justifies treating hardness as a property of the data. Second, we *decompose* hardness by contrasting empirical IH with a fully model-blind measure of local class overlap, and show that local overlap accounts for only part of the difficulty. Third, through a controlled ablation that separately varies the feature space and the target, we show that hardness responds to the target, not the features. To our knowledge, this is the first study to isolate these two sources of hardness experimentally in an applied setting.

2. RELATED WORK

Instance hardness and complexity measures. The study of data difficulty has shifted from aggregate dataset complexity toward instance-level analysis. Smith et al. [2014] classified over 190,000 instances with a diverse set of learners to estimate IH empirically and proposed a set of hardness measures — among them the k-Disagreeing Neighbors (kDN), which captures local class overlap — to explain it. Broader catalogs of dataset-level complexity measures [Lorena et al. 2019] and recent surveys dedicated to instance hardness [Torquette et al. 2024] consolidate these tools and their use in data-centric analysis. A recurring finding, which our results revisit in a real imbalanced setting, is that class overlap is a dominant but not exclusive contributor to hardness [Smith et al. 2014].

Uses of hardness. Beyond diagnosis, hardness has been put to work. In dynamic classifier and ensemble selection, local hardness indicates which base learners are competent in confusing regions of the space [Cruz et al. 2018]. Hardness has also been used to order training examples in curriculum learning, presenting easier instances first [Nunes et al. 2021], and to support selective prediction and data-quality auditing [Paiva et al. 2022]. These applications rely on the *identity* of the hard instances being stable — a point our ablation bears on directly.

This work. Unlike studies that take a dataset as given and measure or exploit its hardness, we ask where that hardness originates. By holding the instances fixed and varying, in turn, the feature space and the target definition, we separate two sources of difficulty that are usually entangled, and we do so in an applied public-health problem rather than on synthetic benchmarks.

3. METHODOLOGY

This study is a secondary, exploratory analysis that reuses a set of previously trained and calibrated classifiers to investigate the *origin* of instance-level hardness. Rather than proposing a new classifier, we treat hardness itself as the object of study and contrast three variants of the same prediction problem under two independent hardness measures. The design isolates two candidate sources of difficulty — the feature space and the target definition — so that their effects can be compared directly. Fig. 1 gives an overview of the pipeline. Throughout, we use *hardness* to mean the propensity of an individual instance to be misclassified, estimated independently of any single learner: a property of the instance within a given problem formulation, rather than of the classifier applied to it.

Data and target definition. We use microdata from the 2019 National Survey of School Health

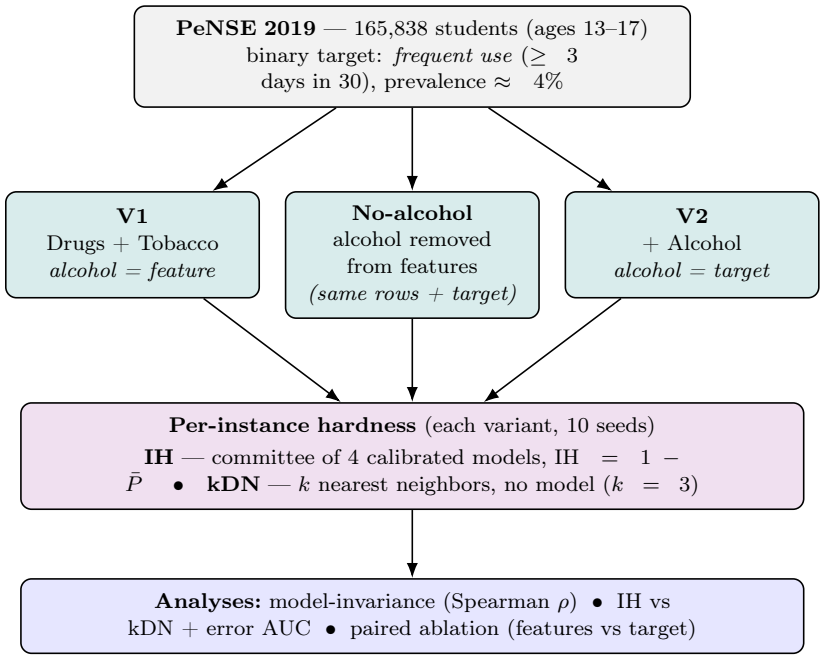

Fig. 1. Methodology overview. A binary frequent-use target is split into three variants: V1 (baseline), *no-alcohol* (alcohol removed from the features, same rows and target as V1, which isolates the feature effect), and V2 (alcohol folded into the target). For each variant, hardness is measured two independent ways — empirically (IH) and geometrically (kDN) — and the per-instance values feed three analyses.

(PeNSE), a nationally representative survey of Brazilian adolescents. The analysis unit is the student, restricted to ages 13–17, yielding 165,838 records described originally by 306 variables. After preprocessing — removal of attributes with more than 10% invalid responses, systematic treatment of non-response codes, one-hot encoding of nominal attributes, and ordinal-to-numeric mapping — we engineer a binary target indicating *frequent* substance use, defined as use on three or more days in the previous 30. This frequency threshold yields a markedly imbalanced problem, with a positive prevalence of roughly 4% in the base variant, which is itself part of what makes the problem hard.

Three problem variants. The core of the design is a controlled contrast between three variants of the problem that differ in a single, well-defined way each (Figure 1). Variant V1 (Drugs+Tobacco) defines the target from illicit-drug and tobacco use, while alcohol-related variables remain available as predictive features. The no-alcohol variant is generated by copying V1 row for row and removing the entire block of alcohol features; by construction it shares the exact same instances, the same target, and the same partitions as V1, differing only in the absence of the alcohol features. The contrast between V1 and the no-alcohol variant therefore isolates the effect of the *feature space* on hardness, holding the target fixed. Variant V2 (Drugs+Tobacco+Alcohol) instead folds alcohol into the target itself. Writing d, t and a for the reported number of days of illicit-drug, tobacco and alcohol use in the previous 30, the two targets are $y_{V1} = \mathbb{1}[d \geq 3 \vee t \geq 3]$ and $y_{V2} = \mathbb{1}[d \geq 3 \vee t \geq 3 \vee a \geq 3]$, so that frequent alcohol use also counts as a positive case.

We note explicitly that the V1–V2 contrast alters both the target and the feature pool at once, and is therefore interpreted only as the effect of *enlarging the target*, not as the effect of alcohol as a target in isolation. The clean, single-factor ablation is the V1–no-alcohol contrast.

Two independent hardness measures. The empirical measure uses four architectures — XGBoost [Chen and Guestrin 2016], CatBoost [Prokhorenkova et al. 2018], LightGBM [Ke et al. 2017], and Logistic Regression — each previously optimized (150 trials with a tree-structured Parzen esti-

mator [Akiba et al. 2019]) and isotonically calibrated within the broader study; we reuse these models without retraining. For each instance we collect the probability assigned to its true class by every model, and define $IH = 1 - \bar{P}$, where $\bar{P}$ is the mean true-class probability across the committee. By construction IH is partially tied to the models, a point we keep in mind throughout the interpretation.

To test whether the agreement observed among these four reflects genuine diversity of inductive bias rather than correlated failure modes among gradient boosting variants, we add a fifth architecture from a distinct family: a multi-layer perceptron (scikit-learn defaults, early stopping, no hyperparameter search). It draws on the same feature ranking as the logistic regression, so that architecture is the only factor that varies, and is isotonically calibrated on the same split under the protocol described above. Because it is not tuned, it is expected to perform below the optimized architectures; the invariance test rests on rank agreement rather than absolute performance, so this does not compromise it. The MLP serves as an invariance probe only: the committee that defines IH remains the four tuned architectures. Its concordance with the committee, and a robustness check contrasting the four-member IH against a hypothetical five-member one, are reported with the results.

The geometric measure is the k-Disagreeing Neighbors (kDN), the fraction of an instance's k nearest neighbors that belong to a different class, computed directly from the standardized feature space with no model whatsoever. Because the positive class is rare, we anchor the analysis at a small neighborhood, $k = 3$, and report $k \in \{1, 3, 5\}$ as a sensitivity check. The independence between IH (derived from models) and kDN (derived from geometry) is what makes their agreement — and especially their disagreement — informative rather than tautological.

Protocol and analyses. To separate genuine structure from partition noise, the entire procedure is repeated under ten random seeds, and all reported quantities are summarized as mean and standard deviation across seeds. Hardness is computed on a held-out test set that was never used during model optimization or calibration. The partition design is inherited from the broader study: within each of the five Brazilian macro-regions, one federal unit is drawn at random for training and the remaining units form the holdout, which is then split into calibration and test halves. Training and evaluation therefore occur in different states, and the ten seeds correspond to ten distinct draws of which states compose the training set. We treat this only as a source of run-to-run variability (see Section 6).

On the resulting per-instance hardness values we run three analyses, one per objective. First, to test whether hardness is a property of the data or of the model, we measure the rank agreement (Spearman's ρ) between the per-instance IH of every pair of architectures. Second, to decompose hardness, we compare empirical IH against geometric kDN, both through their rank correlation and through the area under the ROC curve (AUC) of each measure as a predictor of true classification error — where error is defined independently, from the predicted class versus the true label, rather than from IH. Third, to locate the origin of hardness, we perform a paired comparison of IH between the no-alcohol variant and V1 on the same instances, assessed by a paired Wilcoxon test together with effect magnitude and a practical-equivalence margin; the same logic, applied to the V1–V2 contrast, measures the effect of changing the target.

4. RESULTS

We report results across the ten data partitions, summarizing every quantity as mean and standard deviation over seeds. The three analyses map to the three objectives: hardness as a property of the data, the decomposition of hardness into local overlap and a residual, and the origin of hardness in the feature space versus the target.

Hardness is largely, but not uniformly, model-invariant. Table I reports the rank agreement (Spearman's ρ) between the per-instance IH of every pair of architectures. Among the four committee members, agreement is uniformly high: across all pairs and all variants, ρ never falls below 0.94, with standard deviations under 0.02. Adding a fifth architecture from a different family lowers this floor:

Table I. Inter-architecture agreement of per-instance IH (Spearman's ρ, mean ± sd over ten seeds). Agreement is high among the four committee members and lower, but still substantial, against the fifth architecture used as an invariance probe.

Variant	Architecture pair	Spearman ρ
V1 (Drugs+Tobacco)	XGBoost–CatBoost	0.965 ± 0.016
	XGBoost–LightGBM	0.967 ± 0.019
	XGBoost–LogReg	0.947 ± 0.016
	CatBoost–LightGBM	0.965 ± 0.018
	CatBoost–LogReg	0.947 ± 0.015
	LightGBM–LogReg	0.948 ± 0.015
No-alcohol	XGBoost–CatBoost	0.965 ± 0.012
	lowest pair (LightGBM–LogReg)	0.960 ± 0.016
V2 (+ Alcohol)	XGBoost–LightGBM	0.991 ± 0.003
	lowest pair (LightGBM–LogReg)	0.978 ± 0.003
Fifth architecture (MLP), mean over its four pairs with the committee		
V1 (Drugs+Tobacco)	MLP–committee	0.786 ± 0.084
No-alcohol	MLP–committee	0.851 ± 0.068
V2 (+ Alcohol)	MLP–committee	0.972 ± 0.011

cross-family agreement averages 0.79 in V1 (ranging from 0.64 to 0.92 across seeds), 0.85 in the no-alcohol variant, and 0.97 in V2. Agreement therefore rises sharply with the prevalence of the positive class, extending a pattern already visible within the committee, where V2 also shows the highest concordance ($\rho > 0.97$ throughout). Less expected is that the advantage of V2 is concentrated in the easy instances: restricted to the hardest 5% of cases, cross-family agreement converges to 0.82–0.85 in all three variants. Hardness is thus a robust property of the data at the population level, but the identity of the hardest instances is more architecture-dependent than aggregate agreement suggests. Part of the gap is attributable to the MLP being a weaker learner (AUROC ≈ 0.90 against ≈ 0.94) rather than to inductive bias alone, and the two cannot be fully separated here.

The choice of committee is not load-bearing. The per-instance IH of a hypothetical five-member committee correlates with the reported four-member IH at $\rho = 0.97$–0.998, and mean IH differs by less than 0.001 in every variant. Every result reported below is unchanged under either committee: the V2/V1 ratio is 2.45× against 2.43×, and the median paired difference of the feature ablation is -0.0004 under both.

Local overlap explains only part of the difficulty. Table II compares the two hardness measures. The rank correlation between empirical IH and geometric kDN is only moderate ($\rho \approx 0.49$ in V1 and the no-alcohol variant), indicating that the two notions of hardness are related but far from interchangeable. The discrimination analysis sharpens this: as a predictor of true classification error, kDN alone reaches an AUC of only ≈ 0.73, well below the ≈ 0.95 of empirical IH. Since kDN never sees the models, its 0.73 is an honest measure of how much of the difficulty is local class overlap — a substantial but minority share. The gap up to IH's 0.95 is the fraction of hardness that local geometry does not capture, and that must instead arise from non-local structure such as label noise or non-linear feature interactions. We note that empirical IH is partially tied to the models by construction, so its high AUC is expected; the informative quantity is the persistent gap between the two honest signals. The sensitivity check on the neighborhood size shows a monotone trend rather than a plateau: kDN's discriminative power rises with k (AUC 0.63, 0.74 and 0.80 for $k = 1, 3, 5$ in V1, with the same ordering in the other two variants), so the anchor at $k = 3$ is conservative and a larger neighborhood would capture somewhat more of the difficulty. The conclusion is unaffected: even at $k = 5$ the gap to empirical IH remains substantial, and local overlap still accounts for a minority of the hardness.

Hardness comes from the target, not the features. The central result concerns the origin of hardness, shown in Fig. 2. Removing the entire block of alcohol features (V1 → no-alcohol), on the

Table II. Decomposition of hardness. Rank correlation between empirical IH and geometric kDN, and the AUC of each measure as a predictor of true classification error (mean ± sd over ten seeds). kDN's honest AUC of ≈ 0.73 shows local overlap is a minority of the difficulty; the gap to IH's ≈ 0.95 is the unexplained, non-local share.

Variant	IH–kDN ρ	$\mathrm{AUC}_{\mathrm{kDN}}$	$\mathrm{AUC}_{\mathrm{IH}}$
V1 (Drugs+Tobacco)	0.486 ± 0.006	0.738 ± 0.017	0.971 ± 0.003
No-alcohol	0.499 ± 0.004	0.737 ± 0.022	0.970 ± 0.003
V2 (+ Alcohol)	0.663 ± 0.008	0.732 ± 0.025	0.917 ± 0.011

same instances and with the same target, barely moves hardness: mean IH shifts from 0.057 to 0.063, the median paired difference is −0.0004, and 79% of instances change by less than 0.02 in absolute IH. This near-equivalence is stable across all ten seeds. By contrast, redefining the target to include frequent alcohol use (V1 → V2) raises mean IH roughly 2.5×, from 0.057 to 0.142; the same shift appears in geometric kDN (from 0.055 to 0.148) and in the committee error rate (from 9% to 18%). The three independent indicators move together, which rules out an artifact of any single measure.

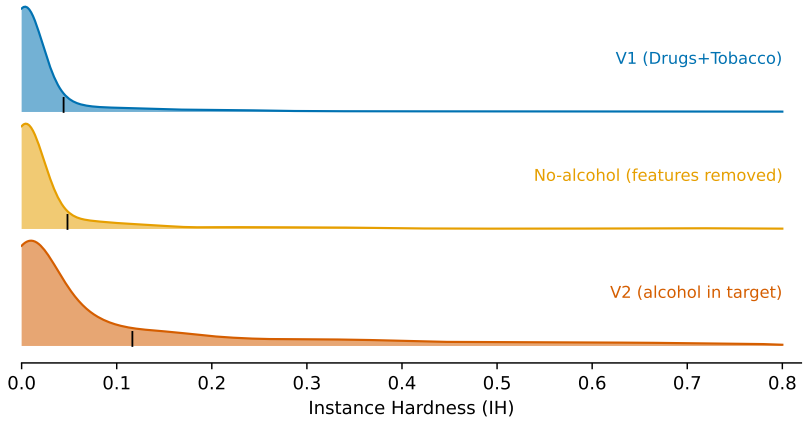

Fig. 2. Origin of hardness. Per-instance IH density for the three variants (all ten seeds pooled; tick marks show the means). Removing the entire alcohol feature block (V1 → no-alcohol), on the same instances and target, leaves the distribution essentially unchanged — the two upper ridges coincide in shape and mean. Redefining the target to include alcohol (V2) shifts mass toward higher hardness, raising mean IH roughly 2.5×. Difficulty tracks the target, not the features.

We stress the asymmetry of evidence. The V1–no-alcohol contrast is a clean, single-factor ablation: it changes only the features and shows that the *structure* of hardness — its population level and distribution — is preserved, with only a modest reshuffling of individual instances (mean absolute change ≈ 0.022 against a near-zero mean change, i.e. instance-level changes that cancel out). The V1–V2 contrast changes the target and, with it, every indicator of difficulty. Taken together, the results indicate that in this domain the difficulty of the problem is governed by what is being predicted, not by which variables are available to predict it.

5. DISCUSSION

Three results, taken together, point to a single conclusion about where the difficulty of this problem resides. First, the strong agreement between architectures (Table I) shows that hardness is not an idiosyncrasy of any one learner: models spanning gradient boosting, a linear classifier and a neural network concentrate their difficulty on the same instances. This is what licenses speaking of *the* hardness of an instance, rather than the hardness relative to a given classifier. Second, the gap

between the honest discriminative power of local overlap (kDN, AUC ≈ 0.73) and that of empirical hardness (IH, AUC ≈ 0.95) shows that local class overlap, while real, accounts for only part of the difficulty. A substantial residual remains that geometry in the immediate neighborhood does not see, and that is most plausibly attributable to label noise — expected in self-reported adolescent behavior — or to non-linear interactions among features that no single neighborhood captures. Because the three variants share the same instances, measurement noise is constant across them and cannot explain the contrast between variants.

The third result is the central one. When the feature space is stripped of an entire block of strong predictors while the target is held fixed, the distribution of hardness is essentially unchanged (Fig. 2); when the target is enlarged while drawing on a comparable feature pool, hardness, local overlap, and error rate all rise together. The difficulty of predicting adolescent substance use is thus governed far more by *what* is being predicted than by *which* variables are used to predict it. This has a concrete methodological implication. Faced with a hard problem of this kind, the common reflex is to collect more or better features; our results suggest that, past a reasonable feature set, such effort yields little, because the difficulty is encoded in the definition of the positive class. The performance ceiling is then a property of the prediction task, and is more productively addressed by reconsidering the target — its granularity, its threshold, the behaviors it conflates — than by further feature engineering.

A subtle but important point concerns the instance-level reshuffling we observe under the feature ablation: although the hardness distribution is preserved in aggregate, individual instances do change in difficulty, with changes that cancel out across the population. Hardness is therefore stable as a property of the *problem* without being stable as a property of every *instance*. This distinction matters for any application that relies on per-instance hardness — for curriculum ordering, for selective abstention, or for auditing — since the identity of the hard cases is more fragile than their overall prevalence.

6. LIMITATIONS

Several limitations bound the scope of these conclusions. First, the contrast that changes the target (V1 vs. V2) necessarily alters both the target and the feature pool at once; it should therefore be read as the effect of *enlarging the target*, not as an isolated effect of alcohol as a target. The clean, single-factor evidence comes from the V1–no-alcohol ablation, which varies only the features. Second, empirical IH is by construction tied to the committee of models used to compute it; we mitigate the resulting circularity by validating against a fully model-blind measure (kDN) and by defining classification error independently of IH, but the absolute level of IH-based discrimination should not be read as a model-free quantity. Third, the data partitions are inherited from a broader study whose design draws training and test sets in a way we do not vary here beyond the ten seeds; we treat this only as a source of run-to-run variability, and we do not claim the hardness estimates generalize to partitioning schemes unlike the one used. Fourth, the analysis rests on a single survey of a single country and year, with a rare positive class; while this is precisely the kind of difficult, real-world setting in which the question is interesting, the quantitative values — and possibly the balance between local and non-local hardness — may differ in other domains. Fifth, the additional architecture used to probe model invariance was not tuned and performs below the committee members; the resulting drop in cross-family rank agreement conflates two effects — distinct inductive bias and lower model quality — that this design cannot separate.

7. CONCLUSION

We set out to identify where the difficulty of a hard, imbalanced prediction problem originates, taking frequent substance use among Brazilian adolescents as a case study. Three findings converge. Per-instance hardness is largely model-invariant (Spearman $\rho > 0.94$ across four architectures, and ≈ 0.79–

0.97 against a fifth from a distinct family), which licenses treating it as a property of the data rather than of any single learner. A model-blind measure of local class overlap captures only part of that hardness (AUC ≈ 0.73 against ≈ 0.95 for empirical IH), leaving a substantial residual attributable to non-local structure such as label noise or feature interactions. And a clean single-factor ablation shows that removing an entire block of strong features barely moves the hardness distribution, while enlarging the target raises mean hardness roughly $2.5\times$. In this domain, difficulty is governed by what is predicted, not by which variables are used to predict it.

The practical lesson is that, past a reasonable feature set, the performance ceiling of such tasks is more productively addressed by reconsidering the target — its granularity, its threshold, and the behaviors it conflates — than by collecting further features. The same caveat bounds downstream uses of per-instance hardness, such as curriculum ordering or selective abstention. Because our strongest evidence rests on a single ablation and a single survey, future work should replicate the analysis across additional feature blocks, alternative target definitions, and other domains and partitioning schemes, to establish how far the dominance of the target over the features generalizes.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, USA, pp. 2623–2631, 2019.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, pp. 785–794, 2016.

Cruz, R. M. O., Sabourin, R., and Cavalcanti, G. D. C. Prototype selection for dynamic classifier and ensemble selection. *Neural Computing and Applications* 29 (2): 447–457, 2018.

IBGE. *Pesquisa Nacional de Saúde do Escolar: 2019*. IBGE, Coordenação de População e Indicadores Sociais, Rio de Janeiro, 2021.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the International Conference on Neural Information Processing Systems*. Long Beach, USA, pp. 3149–3157, 2017.

Lorena, A. C., Garcia, L. P. F., Lehmann, J., de Souto, M. C. P., and Ho, T. K. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys* 52 (5): 1–34, 2019.

Nunes, G. H., Martins, G. O., Forster, C. H. Q., and Lorena, A. C. Using instance hardness measures in curriculum learning. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. Online, pp. 177–188, 2021.

Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., and Lorena, A. C. Relating instance hardness to classification performance in a dataset: a visual approach. *Machine Learning* 111 (8): 3085–3123, 2022.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the International Conference on Neural Information Processing Systems*. Montréal, Canada, pp. 6639–6649, 2018.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. An instance level analysis of data complexity. *Machine Learning* 95 (2): 225–256, 2014.

Torquette, G. P., Nunes, V. S., Paiva, P. Y. A., and Lorena, A. C. Instance hardness measures for classification and regression problems. *Journal of Information and Data Management* 15 (1): 152–174, 2024.

This study was supported by the Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) through a Technological Product scholarship (Edital FAPEMAT no. 012/2024 — Mestrado com Produto Tecnológico).