Policy-Oriented Binary Classification: Improving (KD-)CART Final Splits for Subpopulation Targeting

Lei Bill Wang
The Ohio State University
Department of Economics

Zhenbang Jiao
The Ohio State University
Department of Statistics

Fangyi Wang
The Ohio State University
Department of Statistics

Abstract

Policymakers often use recursive binary split rules to partition populations based on binary outcomes and target subpopulations whose probability of this adverse binary event exceeds a threshold. We call such problems Latent Probability Classification (LPC). Practitioners typically employ Classification and Regression Trees (CART) for LPC. We prove that, in the context of LPC, classic CART and the knowledge distillation method, in which the student model is a CART (referred to as KD-CART), are suboptimal. We propose Maximizing Distance Final Split (MDFS), which generates split rules that strictly dominate CART/KD-CART under the unique intersect assumption. Under this assumption, MDFS identifies the unique best split rule. Consequently, it targets more vulnerable subpopulations than CART/KD-CART, where "more vulnerable" is defined as a higher probability of the adverse binary event. To further relax the assumption, we propose Penalized Final Split (PFS) and weighted Empirical risk Final Split (wEFS). Through extensive simulation studies, we demonstrate that the proposed methods predominantly outperform CART/KD-CART using two risk metrics. When applied to real-world datasets, MDFS generates policies that target more vulnerable subpopulations than the CART/KD-CART.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Recursive binary split rules appeal to policymakers and researchers due to their transparency and nonparametric flexibility. These rules partition a population based on a binary outcome $Y \in \{0, 1\}$ and target subpopulations with a probability of $Y = 1$ greater than 50% when implementing a policy. We call such policy targeting problems Latent Probability Classification (LPC). Practically, CART is often employed for LPC. For example, Andini et al. (2018) uses CART to find subpopulations with a higher than 50% probability of being financially constrained and recommends targeting these households with a tax credit program. In Appendix A.1, we cite 23 empirical cases of using CART for various LPC problems to demonstrate the broad applicability of the setup that this paper investigates. In these studies, researchers typically use CART to divide the samples into many nodes, estimate the probability of $Y = 1$ for each node, and target those nodes with estimated probabilities higher than a threshold of 50%, denoted as $c = 0.5$. This approach, though intuitive, is not optimal for LPC. Here, we provide a toy example in Figure 1a to illustrate the limitation of using CART for an LPC problem.

A toy example: Suppose the latent probability of a binary event $Y = 1$ is a sinusoidal function of an observable variable X, $\mathbb{P}(Y = 1|X) = \frac{\sin(2\pi X)+1}{3}$ where $X \sim \text{Unif}[0, 1]$. Figure 1a shows the function, where the green segment represents the subpopulation that should be targeted, and the orange segments represent the subpopulation that should not be targeted. In this example, we impose that the policymaker can only split the population *once* based on the value of X. CART splits at $X = 0.5$, denoted as s^{CART} in Figure 1a. The left node has $\mathbb{P}(Y = 1|X \leq 0.5) = \frac{2}{3}(0.5 + \frac{1}{\pi}) > 0.5$, whereas the right node has $\mathbb{P}(Y = 1|X > 0.5) = \frac{2}{3}(0.5 - \frac{1}{\pi}) \leq 0.5$. Consequently, we target the subpopulation by $X \leq 0.5$, i.e., left node. To demonstrate why s^{CART} is suboptimal for the LPC problem, consider an alternative split at $s^* = \frac{5}{12}$. The

left node still has $\mathbb{P}(Y = 1|X \leq \frac{5}{12}) \approx 0.571 > 0.5$, right node has $\mathbb{P}(Y = 1|X > \frac{5}{12}) \approx 0.164 < 0.5$. Only the left node ($X \leq \frac{5}{12}$) is targeted. All subgroups that are correctly targeted/not targeted by s^{CART} are also correctly targeted/not targeted by s^*. Moreover, s^* excludes the group with $\frac{5}{12} < X < 0.5$ whose $\mathbb{P}(Y = 1|X) < 0.5$ from being targeted whereas s^{CART} incorrectly targets this subgroup. We say that s^* strictly dominates s^{CART}. Section 3 formally defines strict domination.

Proposed methods: This paper proposes methods that generate improved split rules. First, we replace the CART impurity function with a weighted sum of the distances between node means and the threshold c. This method is called Maximizing Distance Final Split (MDFS). Assuming the existence of a unique intersection between $\mathbb{P}(Y = 1|X)$ and c, MDFS identifies the unique optimal split (hence, strictly dominating CART). The second method, Penalized Final Split (PFS), is a generalization of MDFS. It relaxes the unique intersect assumption and still strictly dominates CART. The third method, weighted Empirical risk Final Split (wEFS), adapts the weighted loss function from the cost-sensitive binary classification literature to our setup.

Generalizing the threshold from $c = 0.5$ to $c \in (0, 1)$: Though most of the LPC studies use 50% as the threshold ($c = 0.5$ corresponds to the classic 0-1 loss), there are more general choices. For example, Sarkar et al. (2024) uses CART to determine which forest zone has a higher than $c = 61\%$ probability of forest fire to implement early warning systems. In some other cases, a policymaker may face budget constraints and hence decides to adopt a threshold that is close to 1 (Hassanzadeh et al., 2021). In the rest of this paper, we use $c \in (0, 1)$ to denote the threshold. We assume that c is *fixed* before implementing our methods. In particular, policymakers can first tune c with CART and then implement our methods with the chosen c. In this scenario, our methods still improve over CART because our theoretical results apply broadly to *any* fixed $c \in (0, 1)$.

Extending to knowledge distillation: Knowledge distillation (KD) refers to a two-step learning algorithm. The first step trains a teacher model with a higher learning capacity, e.g., a neural network or a random forest, to learn $\mathbb{P}(Y = 1|X)$. The second step uses the learned $\mathbb{P}(Y = 1|X)$ as the response variable to train a simpler student model. The goal of the student model is to output a simple and interpretable representation of the teacher model's knowledge. In our case, the student model is a CART. We refer to the KD method with a CART student model as KD-CART. We apply the MDFS method to improve KD-CART and refer to our proposed method as KD-MDFS. This generalizes our contribution to *a larger class of advanced tree-based methods*.

Summary of contributions: Section 3 formulates the LPC problem from a wide range of empirical works and shows that the split rule generated by CART/KD-CART is strictly dominated. Section 4 proposes MDFS, which point-identifies the unique optimal split rule assuming a unique intersection between $\mathbb{P}(Y = 1|X)$ and c. In addition, we propose a consistent estimator for the MDFS split rule. Section 4.4 shows that MDFS generates policies that target *more vulnerable* subpopulations. To relax the unique intersection assumption, we further propose PFS and wEFS in Section 5. Lastly, in Section 6, we demonstrate that the proposed methods outperform their respective baselines and target more vulnerable subpopulations, using simulations with synthetic and real-world datasets.

2 RELATED LITERATURE

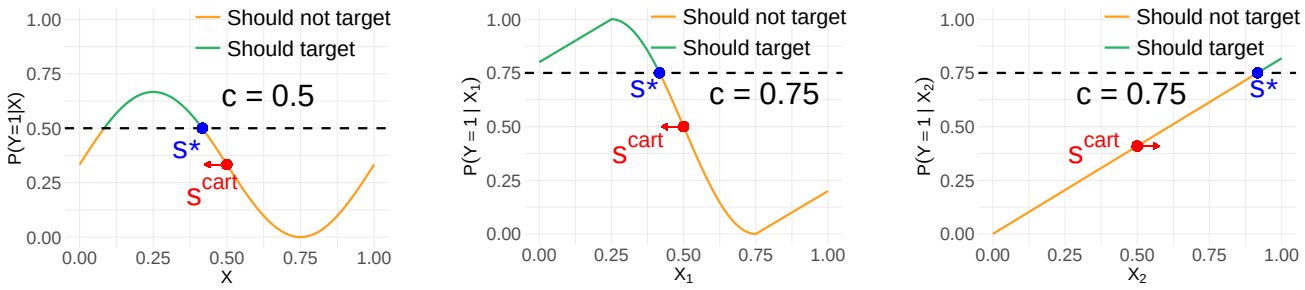
The novelty of our problem setup can be best established by comparing our work with various strands of existing frontier literature.

Nonparametric binary classification: Nonparametric binary classifiers are often not designed for transparent policymaking. Babii et al. (2024)'s theorems apply to uninterpretable deep learning. Singh and Khim (2022) outputs stochastic decision rules, which raise fairness concerns. Our output policy is nonparametric, interpretable, and deterministic.

Policy targeting: Most of the existing policy targeting literature has been developed within the causal inference framework Kitagawa and Tetenov (2018); Athey and Wager (2021); Mbakop and Tabord-Meehan (2021). Such causal inference methods are incompatible with the LPC setup, where the treatment might not have been tested in real life. We include a policy learning method from Zhou et al. (2023) as a baseline in our simulation studies to demonstrate that our proposed methods' strengths over policy learning methods for LPC problems. Two more detailed comparisons between LPC and policy learning literature are provided in Section 7 and Appendix A.2.

Tree-based methods' consistency: The tree-based method literature has a vested interest in consistency. Wager and Athey (2018); Zheng et al. (2023) show that different tree-based methods are consistent for estimating $\mathbb{P}(Y = 1|X)$. This work is interested in a different type of consistency. We show that our estimator consistently estimates the MDFS split rule.

Knowledge distillation with CART as student:



(a) Toy example (b) t_1 example (c) t_2 example

Figure 1: Illustrative examples comparing s^{CART} with s^*

Liu et al. (2018); Dao et al. (2021) use CART as the student model in knowledge distillation, termed as KD-CART in our paper. We show that the theoretical results we develop for MDFS apply to KD-CART.

3 PRELIMINARIES: (OSF)LPC

We first restrict the theoretical discussion to the following *one-split, one-feature* (OSF) LPC problem characterized as follows. Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$ be the univariate feature space and label space, respectively. Let $f : \mathcal{X} \to \mathbb{R}^+$ be a probability density function that is continuous on $\mathcal{X}$, F be its corresponding cumulative distribution function, and $\eta(x) := \mathbb{P}(Y = 1|X = x)$ be continuous. Consider a dataset comprising n i.i.d. samples, (X_i, Y_i), $i = 1, \ldots, n$, where $X_i \sim f$ and $Y_i|X_i \sim \text{Bernoulli}(\eta(X_i))$. A policymaker is allowed to split the feature space one time (i.e., one-split) using the univariate feature X (i.e., one-feature). After the split, policymakers target those node(s) whose node mean is greater than c.

3.1 CART is strictly dominated

The most common criterion function optimized by CART to determine the split s^{CART} is the weighted sum of variances of the two child nodes, i.e., $s^{CART} = \arg\min_{s\in(0,1)} \mathcal{G}^{CART}(s)$ where $\mathcal{G}^{CART}(s) = F(s)(\mu_L(s) - \mu_L^2(s)) + (1 - F(s))(\mu_R(s) - \mu_R^2(s))$, $\mu_L(s) = \int_0^s \eta(x)dF(x)/F(s)$ and $\mu_R(s) = \int_s^1 \eta(x)dF(x)/(1 - F(s))$ are the left and right node mean, respectively.

To illustrate that s^{CART} is strictly dominated, we first introduce the following definition of strict dominance for comparing two splitting rules s and s'. The definition is adapted from dominating decision rule in classic decision theory. To guide our readers through the dense notation in Definition 3.1, we provide an intuitive explanation after introducing the definition.

Definition 3.1 (Strict dominance). *If* $\exists\ \{s, s'\} \in [0,1]^2$ *such that* $\forall\ x \in [0,1], \mu_{t(x)}(s) > c \implies \mu_{t(x)}(s') > c$ *when* $\eta(x) > c$, *and* $\mu_{t(x)}(s) \le c \implies \mu_{t(x)}(s') \le c$ *when* $\eta(x) \le c$, *where* $\mu_{t(x)}(s) = \mu_L(s)$ *if* $x \le s$ *and* $\mu_{t(x)}(s) = \mu_R(s)$ *if* $x > s$, ***and*** *there exists a set* $\mathcal{A}$ *with a nonzero measure such that,* $\forall\ x \in \mathcal{A}$, *either one (or both) of the following conditions is true: (i)* $\mu_{t(x)}(s) \le c, \mu_{t(x)}(s') > c$ *when* $\eta(x) > c$; *and (ii)* $\mu_{t(x)}(s) > c, \mu_{t(x)}(s') \le c$ *when* $\eta(x) \le c$. *Then we say that splitting rule* s' ***strictly dominates*** *splitting rule* s.

Despite its dense notation, Definition 3.1 has a straightforward interpretation: split rule s' strictly dominates s if it performs no worse than s for all $x \in [0,1]$ and strictly better than s for all $x \in \mathcal{A} \subseteq [0,1]$, where $\mathcal{A}$ has nonzero measure.[1]

We show that under very general conditions, there exist some rules that strictly dominate CART's split rule in Theorem 3.2. Proofs of all lemmas and theorems are collected in Appendix B.

Theorem 3.2. *Suppose* $c \in [c_{min}, c_{max}]$ *where* $c_{min} = \min(\mu_L(s^{CART}), \mu_R(s^{CART}))$ *and* $c_{max} = \max(\mu_L(s^{CART}), \mu_R(s^{CART}))$ *and* $\eta(s^{CART}) \ne c$. *Then there exists*

$$s^* = \begin{cases} \arg\min\limits_{s\in(0,s^{CART}),\eta(s)=c} (s^{CART} - s) & \text{if } \zeta > 0, \\ \arg\min\limits_{s\in(s^{CART},1),\eta(s)=c} (s - s^{CART}) & \text{if } \zeta < 0. \end{cases}$$

where $\zeta = (\eta(s^{CART}) - c)(\mu_R(s^{CART}) - \mu_L(s^{CART}))$.

Further, all $s \in \big((s^* \wedge s^{CART}), (s^* \vee s^{CART})\big)$, ***strictly dominates*** s^{CART}.

The key challenge in understanding Theorem 3.2 is the interpretation of s^*. Figure 1a provides a graphical illustration for interpreting s^*. Given that $\eta(s^{CART}) < c = 0.5$ and $\mu_L(s^{CART}) > \mu_R(s^{CART})$, Figure 1a

[1] "No worse" means that if s targets/not target correctly at a point, then s' must also target/not target correctly at that point; "strictly better" means that at some points where s target/not target incorrectly, s' targets/not target correctly at those points.

corresponds to the first minimization problem in the definition of s^*. The minimization problem searches for s^* over the intersection of $s \in (0, s^{CART})$ and $s \in \{s : \eta(s) = c\}$, where $s^{CART} = 0.5$. In Figure 1a, there are two candidate values, $s = \frac{1}{12}$ and $s = \frac{5}{12}$, between which $s = \frac{5}{12}$ is closer to s^{CART}, and hence $s^* = \frac{5}{12}$. The graphical illustration is generalizable: s^{CART} determines which of the two minimization problems is used to determine s^*, then $\eta(s) = c$ pins down a set of candidate values of s^*, and lastly, s^* is set to be the one that is closest to s^{CART} among all candidate values.

3.2 KD-CART is strictly dominated

In the LPC setup, a teacher model learns $\eta(x), x \in [0, 1]$. Prediction based on the teacher model is denoted as $\hat{\eta}(x)$. In the LPC setup, the student model is a CART that takes in $\hat{\eta}(x)$ as the response and learns to partition the population based on $\hat{\eta}(x)$. One can show that Theorem 3.2 also applies to KD-CART, meaning there exist split rules that strictly dominate the split rule generated by KD-CART. Details are provided in Lemma B.4 in Appendix B.

4 MDFS

The suboptimality of CART for solving the OSF LPC problem motivates our proposed method MDFS, which point-identifies s^*. The optimality of s^* is discussed in Section 4.2. Since our theoretical results apply to OSF LPC, we advocate using our methods at the final splits with features identified by CART.[2] Restricting modifications to the last splits may appear trivial at first. However, note that as the tree grows deeper, the number of final splits increases exponentially, leading to non-trivial modifications to the policy designs. We substantiate this claim with empirical applications in Section 6.2.

4.1 Unique intersection assumption

Our theoretical results in Section 4 rely on the following assumption, which is quite plausible in our context. We will further relax this assumption in Section 5.

Assumption 4.1 (Unique intersection between $\eta(X)$ and c)**.** *For $X \sim$ Unif$[0, 1]$, there exists a unique s^* such that $\eta(s^*) = c$, and $\eta(X)$ is strictly monotonic and differentiable on $[s^* - \epsilon, s^* + \epsilon]$ for some $\epsilon \in (0, \min(s^*, 1 - s^*))$.*

[2]Though we focus on explaining MDFS in this section, all the theorems in this section easily extend to KD-MDFS, i.e., replacing the final splits' split criterion function of KD-CART with MDFS.

With a unique intersection between $\eta(X)$ and c, the definition of s^* is simply the X value that corresponds to the unique intersection. We argue for the plausibility of Assumption 4.1.

Uniform X. The set of quantile statistics for any continuous feature follows a uniform distribution, so we can convert a continuous variable to its quantile statistics. For a discrete X, the LPC problem is simpler. We defer the explanation to Appendix B.

The unique intersection condition is weaker than monotonicity, which is assumed by some theoretical works on CART (Blanc et al., 2020). Moreover, in many real-life applications, the monotonicity of $\eta(X)$ is a reasonable assumption. For example, in our real-world dataset application in Section 6.2, a final node is split based on blood glucose level, and the binary outcome Y is diabetic status. It is reasonable to assume that the probability of diabetes increases with blood glucose level. Nevertheless, monotonicity is *not necessary*, see Figure 1b as a non-monotonic example that satisfies Assumption 4.1.

Further, when we apply MDFS to the final nodes, the splitting feature is determined by CART. CART selects the splitting feature that gives the greatest reduction in variances (i.e., the greatest increase in purity in binary classification). Hence, when many features are available, CART is likely to pick some features that exhibit a salient trend (e.g., monotonicity), ruling out features that are more likely to violate the unique intersect assumption. Also, as we go down the tree, the domain of all nodes becomes smaller and smaller, making Assumption 4.1 easier to satisfy. For example, take Figure 1a as an example. It violates the unique intersect assumption. However, if we split the population at $X = 0.3$, then both child nodes will satisfy the unique intersection assumption.

4.2 Optimality of s^*

To characterize the optimality of s^*, we formalize the following definition.

Definition 4.2 ($\mathcal{SD}$-optimal)**.** *A split rule is said to be $\mathcal{SD}$-optimal if it strictly dominates any other rule.*

Remark 4.3. *Under Assumption 4.1, s^* is $\mathcal{SD}$-optimal.*

Proof of this Remark can be found in Appendix B.

4.3 Identification and estimation of MDFS

Theorem 4.4. *Under Assumption 4.1,* $\arg\max_s \mathcal{G}^*(s, c)$ *identifies s^*, where* $\mathcal{G}^*(s, c) = s\,|\mu_L - c| + (1 - s)\,|\mu_R - c|$.

The proof of Theorem 4.4 can be largely decomposed into two steps: first, we show that there exists an interval containing s^* in which first-order condition guarantees that s^* is the local minimum; second, we show that for all s outside of the interval, $\mathcal{G}^*(s,c) < \mathcal{G}^*(s^*,c)$.

Theorem 4.4 states that s^* maximizes the *population* objective function $\mathcal{G}^*(s,c)$. In the *finite* sample regime (denote sample size as n), we can estimate the sample version of s^* using $\{(X_i,Y_i)\}_{i=1}^n$ by maximizing the sample analogue of $\mathcal{G}^*(s,c)$. Following from Assumption 4.1, we have $s^* \in (\epsilon, 1-\epsilon)$. Define the MDFS final split estimator $\hat{s}$ as

$$\hat{s} = \underset{s\in(\epsilon,1-\epsilon)}{\arg\max}\, \widehat{\mathcal{G}}^*(s,c),$$

where $\widehat{\mathcal{G}}^*(s,c) = s\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i \le s\}}{\sum_{i=1}^n \mathbb{1}\{X_i \le s\}} - c\right| + (1-s)\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i > s\}}{\sum_{i=1}^n \mathbb{1}\{X_i > s\}} - c\right|$ is the estimator of $\mathcal{G}^*(s,c)$. The following theorem states that $\hat{s}$ is a consistent non-parametric estimator for s^*.

Theorem 4.5. *Under Assumption 4.1,* $\hat{s} \xrightarrow{p} s^*$

4.4 Policy significance of MDFS

This section relaxes the OSF LPC setup and considers the policy relevance of MDFS under many-node many-feature setup. We demonstrate two advantages of MDFS policies: (i) targeting more vulnerable subpopulations than CART by using the same amount of resources and (ii) uncovering subgroups with a higher-than-threshold probability of event $Y=1$ that CART ignores.

Suppose there are two nodes $\{t_1, t_2\}$ with the same amount of population and corresponding features $X_1, X_2 \sim \text{Unif}(0,1)$. We depict $\eta_1(X_1)$ for node t_1 in Figure 1b and $\eta_2(X_2)$ for node t_2 in Figure 1c. The analytical forms of $\eta_1(X_1)$ and $\eta_2(X_2)$ are provided in Appendix B. Here, we compare the targeted population using MDFS policy versus CART policy. Splitting nodes t_1 and t_2 individually at s^{CART} versus s^* results in different target subpopulations: s^{CART}: Target $\{X_1 < \frac{1}{2}\}$ in t_1. s^*: Target $\{X_1 < \frac{5}{12}\}$ in t_1 and $\{\frac{11}{12} < X_2 < 1\}$ in t_2. The two sets of policies target the same proportion of the population, but $\eta_1(x_1) < 0.75$ for $x_1 \in \{\frac{5}{12} < X_1 < \frac{1}{2}\}$, which is targeted by s^{CART}, whereas $\eta_2(x_2) > 0.75$ for $x_2 \in \{\frac{11}{12} < X_2 < 1\}$, which is targeted by s^*. Therefore, policies based on LPC target a **more vulnerable** subpopulation than CART/KD-CART policy.

Admittedly, the fact that LPC and CART policies target the same proportion of the population in the previous example is by construction. It is also possible that the proportion of the population targeted by the MDFS policy is bigger than that by CART or KD-CART. In this scenario, comparing the effectiveness of the two sets of policies is not straightforward. Nonetheless, LPC still has policy significance. It discovers new latent groups with a higher-than-c probability of an adversarial event happening to them (i.e., vulnerable subpopulations), e.g., $\eta_2(x_2) > 0.75$ for $\frac{11}{12} < x_2 < 1$.

We make two remarks to formalize the two advantages that the MDFS policy offers. Assuming a homogeneous targeting cost per unit of population, the targeting cost of a policy is the percentage of the subpopulation targeted. Denote all M final splitting nodes as $\{t_1, t_2, \ldots, t_M\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$. For node m, we denote the feature selected by CART as $X_{(m)}$ and the CDF of $X_{(m)}$ as F_m and let $\eta_m(x) = \mathbb{P}(Y = 1 | X_{(m)} = x)$. The costs of CART and MDFS policies are

$$C^{CART} = \sum_{m\in\mathcal{M}} \int_0^1 \mathbb{1}\{\mu_{t(x)}(s_m^{CART}) > c\} dF_m(x)$$

$$C^{*} = \sum_{m\in\mathcal{M}} \int_0^1 \mathbb{1}\{\mu_{t(x)}(s_m^{*}) > c\} dF_m(x)$$

Assume that for some $m \in \mathcal{M}$, $s_m^{CART} \neq s_m^*$.

Remark 4.6. *If $C^{CART} = C^*$, then MDFS policy targets strictly more vulnerable subpopulation than CART using the same targeting resources, where "more vulnerable" is defined as greater $\eta_m(X_{(m)})$.*

Remark 4.7. *If $C^{CART} < C^*$, then MDFS policy uncovers latent subgroups in some node m with selected feature $X_{(m)} = x$ whose $\eta_m(x) > c$ that CART ignores.*

5 PFS AND WEFS

To relax Assumption 4.1, we propose two additional methods: PFS and wEFS.

5.1 PFS

Intuitively, if μ_L is *close* to c and it is slightly higher than c, by the continuity assumption on $\eta(x)$, it's likely that there is a considerable amount of subpopulation from the left node with $\eta(x) < c$, see $0 < x < \frac{1}{12}$ and $\frac{5}{12} < x < \frac{1}{2}$ in Figure 1a as examples. This can substantially increase misclassification cases because μ_L and $\eta(x)$ are on different sides of c for all these x values. Following this intuition, we consider adding a penalty to $\mathcal{G}^{CART}$ that pushes μ_L and μ_R away from c. Let

$$\mathcal{G}^{PFS}(s,c) = \mathcal{G}^{CART} + \lambda J \tag{1}$$

where $J = F(s)W(|\mu_L - c|) + (1 - F(s))W(|\mu_R - c|)$ and $W : \mathbb{R}^+ \cup 0 \to \mathbb{R}^+ \cup 0$ is a decreasing function that penalizes small distances between the node means (i.e., μ_L and μ_R) and c, and $\lambda \geq 0$ controls the weight of the penalty term, which we denoted as J. The following theorem states that under some regularity conditions of the function W and weight λ, the split rule that minimizes $\mathcal{G}^{PFS}$, denoted as s^{PFS}, strictly dominates s^{CART}.

Theorem 5.1. *Suppose $W : \mathbb{R}^+ \to \mathbb{R}^+$ is convex, monotone decreasing, and upper bounded in second derivative. If s^{CART} is the unique minimizer for $\mathcal{G}^{CART}$ and $\eta(s^{CART}) \neq c$, then there exists a $\Lambda > 0$, such that $\forall \lambda \in (0, \Lambda)$, $s^{PFS} := \arg\min_{s \in [0,1]} \mathcal{G}^{PFS}(s, c)$ strictly dominates s^{CART}.*

We provide an intuitive explanation for Theorem 5.1 in Appendix B. Note that MDFS is a special case of PFS whose $\lambda = \infty$ and W is the identity function.

5.2 wEFS

We adapt the cost-sensitive classifier from Koyejo et al. (2014) to our setup and name it weighted Empirical Final Split (wEFS). Due to the lack of theoretical backing, we do not elaborate on wEFS. We defer the explanation for the lack of the theoretical results to Section 7. Details about wEFS can be found in Algorithm 5 in Appendix C.2.

6 EXPERIMENTS

In this section, we conduct comprehensive numerical experiments to compare the performance of our proposed methods and the classic CART under both regular and KD frameworks. Under the regular framework, we compare the classic CART with MDFS, PFS, and wEFS. Under the KD framework, we compare RF-CART with RF-MDFS. We replace KD with RF because the teacher model is a **R**andom **F**orest.

6.1 Synthetic data

Settings: We simulate synthetic datasets using 8 different data generation processes (DGP), e.g., the Friedman synthetic datasets (Friedman, 1991; Breiman, 1996). These setups are designed to capture various aspects of real-world data complexities and challenges commonly encountered when using tree-based methods, such as nonlinear relationships, feature interactions, collinearity, and noises (subtitles in Figure 2). Detailed descriptions of the DGP are provided in Appendix C.1. For each DGP, we consider a threshold of interest $c \in \{0.5, 0.6, 0.7, 0.8\}$, resulting in 32 unique tasks.

We set two stopping rules for growing the tree: (i) a max depth of $m \in \{4, 5, 6, 7\}$, (ii) a minimal leaf node size of ρn, where $\rho \in \{1\%, 2\%, 3\%\}$ and n is the sample size. We set $n = 5000$. The fitting procedures stop once either of the rules is met. These give us 12 configurations for each task. For each of the $32 \times 12 = 384$ settings, we do 50 replicates of experiments.

Evaluation metrics: We define false positives (FP) and false negatives (FN) before defining the two performance metrics that we use in the synthetic data simulations.

$$FP = \sum_{i=1}^{n} \mathbb{1}\{\hat{T}(\mathbf{X}_i) \leq c\} \mathbb{1}\{\eta(\mathbf{X}_i) > c\}$$

$$FN = \sum_{i=1}^{n} \mathbb{1}\{\hat{T}(\mathbf{X}_i) > c\} \mathbb{1}\{\eta(\mathbf{X}_i) \leq c\}$$

where $\mathbf{X}_i$ is the *multivariate* feature vector which includes *all* features in the simulation and $\widehat{T}(\mathbf{X}_i)$ is the tree-estimate of $\eta(\mathbf{X}_i)$.

We use two metrics to evaluate the performance and robustness of all methods: misclassification rate (MR) and F1 score (F1). MR and F1 are defined as

$$\text{MR} = \frac{1}{n}(FP + FN)$$

$$\text{F1} = \frac{1}{n} \frac{2(n - FP)}{2(n - FP) + FP + FN}$$

Given that our theorems state that split rules generated by (RF-)MDFS and PFS strictly dominate (RF-)CART, we expect that our proposed methods outperform their counterpart in both metrics.

Model fitting procedures: As we mentioned in Section 4, our methods and their counterparts only differ in the splits at the final level (final splits). We first implement CART until the final split is reached based on the stopping rules. Once the final split is identified, we select the splitting feature using CART, then implement MDFS, PFS and wEFS with the selected splitting feature. MDFS and PFS decide the split by optimizing $\mathcal{G}^*$ and $\mathcal{G}^{PFS}$, respectively, while wEFS selects the split by identifying s such that the minimum weighted empirical risk is obtained.

For W in $\mathcal{G}^{PFS}$ from (1), we choose $W(d) = 1 - d$. This choice of W satisfies conditions specified in Theorem 5.1. We also experiment with alternative choices, such as $W(d) = (1-d)^2$ and $W(d) = \exp(-d)$, but observed similar performance across these choices. Based on Theorem 5.1, we set $\lambda = 0.1$, a sufficiently small value that presumably satisfies the conditions of the theorem while maintaining practical effectiveness. We also propose a standard λ selection procedure inspired by cross-validation and the honest approach (Athey and Imbens, 2016) (see details in Appendix C.3).

Comparison against respective baselines: Figure 2 provides a pairwise evaluation of MR differences between CART and its three refinements—MDFS, PFS and wEFS, and between RF-CART and RF-MDFS. Each boxplot depicts the distribution of the MR difference for a single DGP across 50 replicates, with $m = 7, \rho = 2\%, c = 0.5$. Tables beneath the boxplots report one-sided p-values from paired t-tests (mean < 0) and Wilcoxon signed-rank tests (median < 0). Except for a few cases in Ball, Friedman #1, and Ring, MDFS and RF-MDFS reduce both the mean and the median MR relative to their counterparts, CART and RF-CART, at the 5% significance level. PFS and wEFS outperform CART in most cases, but have less stability than MDFS, as evidenced by their large p-values in a few cases.

The comparison in terms of F1 is similar. All proposed methods largely outperform their counterparts at the 5% significance level. We defer this result to Appendix C.4.

RF-MDFS has the strongest performance among six methods: Our methods have a clear advantage over their respective baselines, as shown by Figure 2. In addition, we compare all six methods together, instead of a pair-wise comparison. RF-MDFS performs the best out of all six methods on 26 out of the 32 tasks in terms of MR, 27 out of 32 tasks in terms of F1, as shown by Table 1 and Table 2 in Appendix C.4. These results highlight the strength of combining MDFS with KD-CART, a frontier tree-based algorithm.

Comparison with policytree: We further compare our tree-based classification algorithm with Zhou et al. (2023), referred to as policytree, implemented using R package `policytree` (Sverdrup et al., 2020). Our methods outperform policytree by a very large margin. See Section 7 for how we adapt policy learning to our framework and why our methods outperform policytree. See Appendix C.4 for the numerical results comparison between our methods and policytree.

6.2 Real-world datasets

We implement CART, MDFS, RF-CART and RF-MDFS with the Pima Indians Diabetes dataset (Smith et al., 1988) to demonstrate the policy significance of our proposed methods.[3] The Pima Indians Diabetes dataset measures health factors among Pima Indian women with the response variable being a binary indicator of diabetes status: 34.9% of the sample is diabetic. It contains 768 observations and 8 features: number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, family diabetes history index, and age. We use these eight features to search for subpopulations whose probability of having diabetes is above 60% with depth of trees fixed at $m = 3$. We pick these hyperparameter values so that one set of the empirical results matches Remark 4.6. We also experimented with other hyperparameter values. The results from these additional experiments align with Remark 4.7.

CART vs MDFS: As shown in Figure 3, CART and MDFS commonly target those with Glucose > 129.5 and BMI > 29.95, which consists of 188 observations. The two methods' targeting policies differ in two ways: CART additionally targets $127.5 <$ Glucose ≤ 129.5 and BMI > 29.95. This subgroup consists of 19 observations. MDFS additionally targets Glucose > 166.5 and BMI ≤ 29.95. This subgroup consists of 12 observations. The difference between the sizes of these two subgroups is 7, which is small relative to 188, i.e., the size of the subgroup commonly targeted by both policies. Assuming that the sample is representative of the population of all Pima Indian women, then the two sets of policies will incur a similar amount of targeting resources. The additional group targeted by CART has a 57.9% probability of having diabetes, whereas the additional group targeted by MDFS has a 66.7% probability of having diabetes. MDFS targets a subpopulation that is more prone to diabetes. This corresponds to Remark 4.6.

RF-CART vs RF-MDFS: As shown in Figure 3, both RF-CART and RF-MDFS commonly target Glucose > 157.5 and BMI > 29.95, a subgroup consisting of 92 observations. RF-MDFS additionally targets two subgroups: Glucose > 166.5 and BMI ≤ 29.95 consisted of 12 observations and $129.5 <$ Glucose ≤ 157.5 and BMI > 29.95, which consisted of 96 observations. RF-MDFS targets a much greater number of observations than RF-CART. If the sample is representative of the population, then RF-MDFS would use much more targeting resources than RF-CART. Nevertheless, RF-MDFS is useful in uncovering groups with higher than 60% probability of having diabetes that RF-CART is unable to find. For example, the first group that RF-MDFS additionally targets has a 63.6% probability of having diabetes. This aligns with Remark 4.7.

7 DISCUSSION

We discuss some additional literature that relates to this paper and then conclude.

Cost-sensitive binary classification: LPC can be related to the cost-sensitive binary classification prob-

[3]We supplement an additional forest fire empirical study in Appendix D to showcase the wide applicability of our paper.

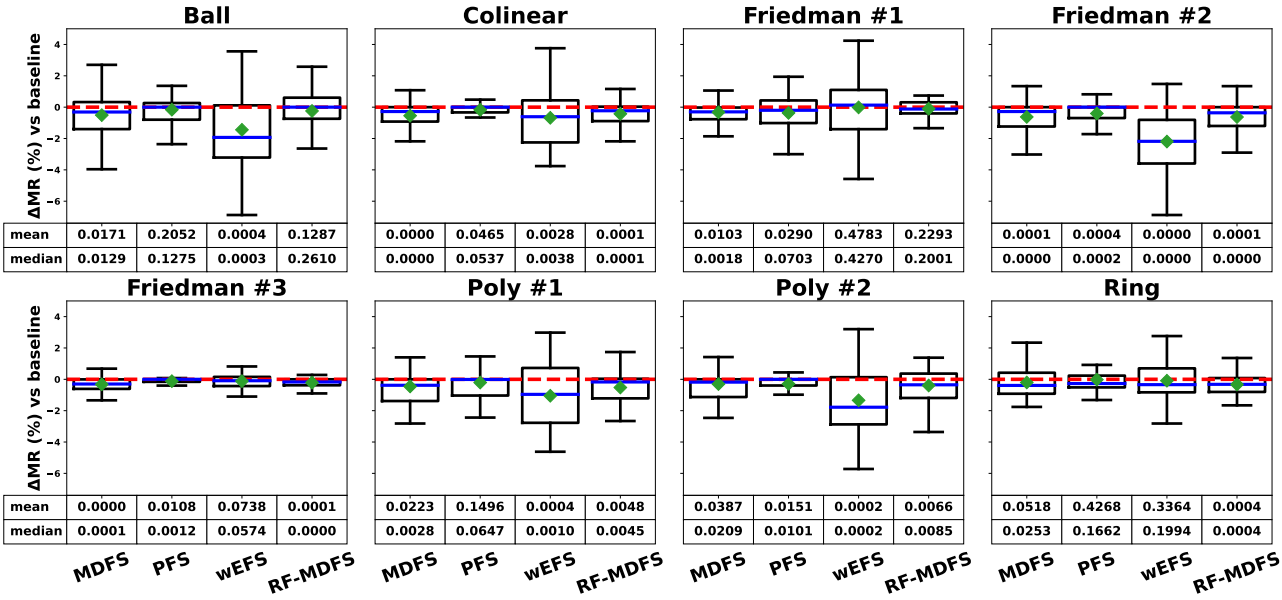


Figure 2: Boxplots of MR differences relative to baseline models. For each panel, the first three boxplots compare CART with MDFS, PFS and wEFS, and the last boxplot compares RF-CART with RF-MDFS. Negative values in the boxplots denote *improvement* in MR. Embedded tables list one-sided paired t-test and Wilcoxon signed-rank p-values for mean and median MR differences, respectively. See the boxplot for F1 scores in Appendix C.4.

lem, further highlighting the policy significance of the LPC problem. Nan et al. (2012); Menon et al. (2013); Koyejo et al. (2014) show that the optimal classifier of a cost-sensitive binary classification problem is determined by whether the latent probability is above some performance-metric-dependent threshold. Whether the theoretical results in the cost-sensitive classification literature carry over to our setup is unclear. For example, Nan et al. (2012) shows risk consistency with uniform convergence of the empirical risk. In our setup, $\{\mu_L(s), \mu_R(s)\}$ appear inside the indicator functions, and uniform convergence of empirical risk is not guaranteed even with a large sample.

Subgroup discovery: Lavrač et al. (2004); Herrera et al. (2011); Atzmueller (2015); Helal (2016) discuss a widely-used quality measure, called Weighted Relative Accuracy (WRAcc), in the subgroup discovery literature. Under the assumption that $X \sim \text{Unif}[0,1]$,

$$\text{WRAcc}_L = F(s)\left(\frac{\int_0^s \eta(x)dx}{s} - \int_0^1 \eta(x)dx\right)$$

where the subscript L indicates it is the WRAcc measure associated with the left node. $\mathcal{G}^*$ is a generalization of $\text{WRAcc}_L + \text{WRAcc}_R$. Hence, Theorem 4.4 applies to a combination of CART and WRAcc.

Remark 7.1. *Under Assumption 4.1, a CART algorithm that uses $WRAcc_L + WRAcc_R$ as the objective function is a special case of $\mathcal{G}^*$ whose $c = \ddot{s}$ where $\eta(\ddot{s}) = \int_0^1 \eta(x)dx$.*

Corollary 7.2. *Under Assumption 4.1, a CART algorithm that uses $WRAcc_L + WRAcc_R$ identifies $\ddot{s}$.*

Policy learning: We show that there is an equivalence between the policy learning problem and LPC and then explain how we can adapt a policy learning algorithm to our context. In addition, we explain why our methods should work better than adapting policy learning to LPC.

Equivalence between policy learning and LPC: Consider potential outcomes $Y_i(0) \mid X_i \sim \text{Bernoulli(c)}$ and $Y_i(1) \mid X_i \sim \text{Bernoulli}(\eta(\text{X}_\text{i}))$.

Remark 7.3. $\max_{\pi(X_i)} \mathbb{E}[Y_i(\pi(X_i))]$*, where $\pi(X_i)$ is the treatment allocation rule, is equivalent to an LPC treatment allocation rule with threshold c.*

Proof.

$$\begin{aligned}
&\mathbb{E}[Y_i(\pi(X_i))] \\
=&\mathbb{E}[\mathbb{E}[Y_i(\pi(X_i)) \mid X_i]] \\
=&\mathbb{E}\big[\mathbb{1}\{\pi(X_i)=1\}\mathbb{E}[Y_i(1) \mid X_i] \\
&\qquad + \mathbb{1}\{\pi(X_i)=0\}\mathbb{E}[Y_i(0) \mid X_i]\big] \\
=&\mathbb{E}\big[\mathbb{1}\{\pi(X_i)=1\}\eta(X_i) + \mathbb{1}\{\pi(X_i)=0\}c\big]
\end{aligned}$$

From the last line of the equation, it is not hard to see that maximizing $\mathbb{E}[Y_i(\pi(X_i))]$ means that $\pi(X_i) = 1$

CART

```
if Glucose > 127.5
    if BMI <= 29.95
        if Glucose <= 145.5
            value: 0.146, samples: 41
            value: 0.514, samples: 35
        if Glucose <= 157.5
            value: 0.609, samples: 115
            value: 0.870, samples: 92
```

MDFS

```
if Glucose > 127.5
    if BMI <= 29.95
        if Glucose <= 166.5
            value: 0.250, samples: 64
            value: 0.667, samples: 12
        if Glucose <= 129.5
            value: 0.579, samples: 19
            value: 0.739, samples: 188
```

RF-CART

```
if Glucose > 127.5
    if BMI <= 29.95
        if Glucose <= 145.5
            value: 0.193, samples: 41
            value: 0.511, samples: 35
        if Glucose <= 157.5
            value: 0.595, samples: 115
            value: 0.830, samples: 92
```

RF-MDFS

```
if Glucose > 127.5
    if BMI <= 29.95
        if Glucose <= 166.5
            value: 0.284, samples: 64
            value: 0.636, samples: 12
        if Glucose <= 129.5
            value: 0.573, samples: 19
            value: 0.713, samples: 188
```

Figure 3: The targeting policies generated by CART, MDFS, RF-CART, RF-MDFS. The red groups are the targeted subpopulations predicted to a higher than 60% probability of being diabetic. We present nodes that differ by targeting decisions due to page limit, see the full trees in Appendix D.

when $\eta(X_i) > c$, and $\pi(X_i) = 0$ otherwise. This is precisely the LPC treatment allocation rule. □

How to adapt policy learning to LPC: In our problem setup, observe that $Y_i \mid X_i \sim \text{Bernoulli}(\eta(\text{X}_\text{i}))$, which exactly corresponds to the potential outcome of treated units $Y_i(1) \mid X_i$ in the policy learning problem. Moreover, despite not observing any control unit in the data, we know that the distribution of the control unit follows Bernoulli(c). Hence, we can simply draw $Y_i(0)$ from this known distribution and input both the treated units (observed (Y_i, X_i) with treatment status $D_i = 1$) and control units (observed X_i and randomly drawn Y_i with treatment status $D_i = 0$) into any policy learning algorithm.

Why our methods should outperform policy learning: We characterize these two types of methodologies (classification vs policy learning) at a high level to demonstrate the theoretical advantage of our methods in the context of LPC.

Classification methods, including CART, MDFS, PFS, and wEFS, have a two-step procedure. First step, divide the population into nodes and learn each node's $\mathbb{P}(Y = 1 \mid X)$. Second step, compare each node's estimated conditional probability of $Y = 1$, denoted as $\hat{\mathbb{P}}(Y = 1 \mid X)$, with the threshold c to decide the targeting rule.

Policy learning algorithms, including policytree, have a three-step procedure. We take policytree as an example. First step, we randomly draw $Y_i(0)$. Second step, policytree divides the population into nodes and learn each node's $\mathbb{P}(Y(1) = 1 \mid X)$ and $\mathbb{P}(Y(0) = 1 \mid X)$. Third step, target those nodes with $\hat{\mathbb{P}}(Y(1) = 1 \mid X) > \hat{\mathbb{P}}(Y(0) = 1 \mid X)$.

These two high-level characterizations allude to why we expect that classification methods work better for LPC than policy learning. Classification methods treat the threshold c as a known quantity. In contrast, policy learning methods treat $\mathbb{P}(Y(0) = 1 \mid X)$ as an unknown quantity and estimate it for all nodes, ignoring that $\mathbb{P}(Y(0) = 1 \mid X) = c$ for all nodes.[4] The simulation results in Appendix C.4 align with this expectation.

Conclusion: Our paper points out that classic CART/KD-CART is suboptimal for LPC in each split. Based on different assumptions, we propose three alternative methods: MDFS, PFS, and wEFS. MDFS and PFS generate policy rules that strictly dominate CART. When brought to the data, our proposed methods outperform their counterparts in terms of two risk metrics with synthetic data and target more vulnerable subpopulations with real-world data.

[4] Technically, we can draw infinitely many $Y_i(0)$ to back out the information that $\mathbb{P}(Y(0) = 1 \mid X) = c$ for all nodes. Unfortunately, such an infinitely-many-draws strategy will make the policy learning algorithm computationally infeasible.

Acknowledgment

We thank our advisors, Professor Jason Blevins, Professor Yoonkyung Lee, and Professor Sebastian Kurtek, for their support throughout our PhD journey, including for this project. We would also like to thank Professor Eric Mbakop, Professor Oksana Chkrebtii, and the anonymous AISTATS reviewers for their constructive feedback.

References

Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini. Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization*, 156:86–102, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

Andrii Babii, Xi Chen, Eric Ghysels, and Rohit Kumar. Binary choice with asymmetric loss in a data-rich environment: Theory and an application to racial justice. *under revision of Quantitative Economics*, 2024. URL `chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ababii.github.io/papers/binary_choice.pdf`.

Guy Blanc, Jane Lange, and Li-Yang Tan. Provable guarantees for decision tree induction: the agnostic setting. In *International Conference on Machine Learning*, pages 941–949. PMLR, 2020.

Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

Begüm Çığşar and Deniz Ünal. Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, 2019(1):8706505, 2019.

David D Clarke, Richard Forsyth, and Richard Wright. Machine learning in road accident research: decision trees describing road accidents during cross-flow turns. *Ergonomics*, 41(7):1060–1079, 1998.

Michael Crowe, Michael O'Sullivan, Oscar Cassetti, and Aifric O'Sullivan. Weight status and dental problems in early childhood: classification tree analysis of a national cohort. *Dentistry Journal*, 5(3):25, 2017.

Aurenice da Cruz Figueira, Cira Souza Pitombo, Ana Paula Camargo Larocca, et al. Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: A study case from brazil. *Case Studies on Transport Policy*, 5(2):200–207, 2017.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. *arXiv preprint arXiv:2104.09732*, 2021.

David Feldman and Shulamith Gross. Mortgage default: classification trees analysis. *The Journal of Real Estate Finance and Economics*, 30:369–396, 2005.

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

Abdul Hannan and Jagadeesh Anmala. Classification and prediction of fecal coliform in stream waters using decision trees (DTs) for upper Green River watershed, Kentucky, USA. *Water*, 13(19):2790, 2021.

Parisa Hassanzadeh, Danial Dervovic, Samuel Assefa, Prashant Reddy, and Manuela Veloso. Tradeoffs in streaming binary classification under limited inspection resources. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9, 2021.

Sumyea Helal. Subgroup discovery algorithms: a survey and empirical evaluation. *Journal of computer science and technology*, 31(3):561–576, 2016.

Jonathan D Herman and Matteo Giuliani. Policy tree optimization for threshold-based water resources management over multiple timescales. *Environmental Modelling & Software*, 99:39–51, 2018.

Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29(3):495–525, 2011.

Ali Tavakoli Kashani and Afshin Shariat Mohaymany. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science*, 49(10):1314–1320, 2011.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27, 2014.

SVSS Lakshmi and Selvani Deepthi Kavilla. Machine learning for credit card fraud detection system. *International Journal of Applied Engineering Research*, 13(24):16819–16824, 2018.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5(Feb):153–188, 2004.

Michelle Seng Ah Lee and Luciano Floridi. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191, 2021.

Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 905–912. IEEE, 2018.

Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012042. IOP Publishing, 2021.

Mayuri Mahendran, Daniel Lizotte, and Greta R Bauer. Quantitative methods for descriptive intersectional analysis with binary health outcomes. *SSM-Population Health*, 17:101032, 2022.

J John Mann, Steven P Ellis, Christine M Waternaux, Xinhua Liu, Maria A Oquendo, Kevin M Malone, Beth S Brodsky, Gretchen L Haas, and Dianne Currier. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *Journal of Clinical Psychiatry*, 69(1):23, 2008.

Eric Mbakop and Max Tabord-Meehan. Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848, 2021.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611. PMLR, 2013.

Ye Nan, Kian Ming Chai, Wee Sun Lee, and Hai Leong Chieu. Optimizing f-measure: A tale of two approaches. *arXiv preprint arXiv:1206.4625*, 2012.

Gunda Obereigner, Pavlo Tkachenko, and Luigi del Re. Methods for traffic data classification with regard to potential safety hazards. *IFAC-PapersOnLine*, 54(7):250–255, 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

S Mehdi Saghebian, M Taghi Sattari, Rasoul Mirabbasi, and Mahesh Pal. Ground water quality classification by decision tree method in ardebil region, iran. *Arabian Journal of Geosciences*, 7:4767–4777, 2014.

Yusuf Sahin, Serol Bulkan, and Ekrem Duman. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 40(15): 5916–5923, 2013.

Mriganka Shekhar Sarkar, Bishal Kumar Majhi, Bhawna Pathak, Tridipa Biswas, Soumik Mahapatra, Devendra Kumar, Indra D Bhatt, Jagadish C Kuniyal, and Sunil Nautiyal. Ensembling machine learning models to identify forest fire-susceptible zones in northeast india. *Ecological Informatics*, 81: 102598, 2024.

Prajal Save, Pranali Tiwarekar, Ketan N Jain, and Neha Mahyavanshi. A novel idea for credit card fraud detection using decision tree. *International Journal of Computer Applications*, 161(13), 2017.

Mai Shouman, Tim Turner, and Rob Stocker. Using decision tree for diagnosing heart disease patients. *AusDM*, 11:23–30, 2011.

Shashank Singh and Justin T Khim. Optimal binary classification beyond accuracy. *Advances in Neural Information Processing Systems*, 35:18226–18240, 2022.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

Jaime L Speiser, Constantine J Karvellas, Geoffery Shumilak, Wendy I Sligl, Yazdan Mirzanejad, Dave Gurka, Aseem Kumar, and Anand Kumar. Predicting in-hospital mortality in pneumonia-associated septic shock patients using a classification and regression tree: a nested cohort study. *Journal of intensive care*, 6:1–10, 2018.

Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232, 2020.

Elena G Toth, David Gibbs, Jackie Moczygemba, and Alexander McLeod. Decision tree modeling in r soft-

ware to aid clinical decision making. *Health and Technology*, 11(3):535–545, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Hal R Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

T Waheed, RB Bonnell, Shiv O Prasher, and E Paulet. Measuring performance in precision agriculture: CART—A decision tree approach. *Agricultural Water Management*, 84(1-2):173–185, 2006.

Lei Bill Wang and Sooa Ahn. Disentangling barriers to welfare program participation with semiparametric and mixed effect approaches. *arXiv preprint arXiv:2506.03457*, 2025.

Qin-Cheng Zheng, Shen-Huan Lyu, Shao-Qun Zhang, Yuan Jiang, and Zhi-Hua Zhou. On the consistency rate of decision tree learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 7824–7848. PMLR, 2023.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

Leying Zou and Warut Khern-am nuai. AI and housing discrimination: the case of mortgage applications. *AI and Ethics*, 3(4):1271–1281, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Appendix C.5.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

We summarize the content of the appendices as follows:

A.1 Twenty-three examples of using CART for LPC problems.

A.2 Comparison between LPC and policy learning.

B Proofs for all theoretical results in the paper.

C Additional details of the simulation studies, including data generation process (C.1), full algorithms and implementation details (C.2), choice of hyperparameter λ (C.3), additional simulation results (C.4) and computational complexity analysis (C.5).

D Additional real-world data application results.

A Additional literature reivew

A.1 Extensive use of CART for LPC setups

Traffic safety: da Cruz Figueira et al. (2017) identifies potential sites of serious accidents in Brazil using different decision tree algorithms. Policymakers may choose to take more safety precautions at dangerous traffic spots whose probability of having a fatal accident is above a threshold. Other research that uses decision trees to identify dangerous traffic situations includes Clarke et al. (1998); Kashani and Mohaymany (2011); Obereigner et al. (2021).

Fraud detection: Sahin et al. (2013); Save et al. (2017); Lakshmi and Kavilla (2018) find conditions under which credit card fraudulent usage is likely to happen. Credit card companies can inform card owners when the probability of fraudulent usage is above a pre-specified threshold.

Mortgage lending: Feldman and Gross (2005); Çığşar and Ünal (2019); Madaan et al. (2021) predict different subpopulations' mortgage (and other types of loan) default rates using a decision tree. Mortgage loan lenders can use the result to determine whether to deny a loan request. Another stream of literature on mortgage lending using decision trees focuses on racial discrimination (Varian, 2014; Lee and Floridi, 2021; Zou and Khern-am nuai, 2023). Policymakers may want to intervene in situations where with high enough probability race seems to play a role in determining whether mortgage lending is denied and the probability of denial is high.

Health intervention: Mann et al. (2008); Shouman et al. (2011); Crowe et al. (2017); Speiser et al. (2018); Toth et al. (2021); Mahendran et al. (2022) classify diabetes (or other diseases) using relevant risk factors. Such classification result leads to different health interventions. Doctors may recommend various treatments based on whether the subpopulation a patient belongs to is more likely to be classified as type I or type II diabetes.

Water management: One of the objectives of Herman and Giuliani (2018) is to manage flood control and output a threshold-based water resources management policy. The authors output a set of conditions that define the subpopulation whose probability of flooding is high enough to justify policy intervention. Many other studies also use decision trees to classify whether the water quality is satisfactory, resulting in important implications to water management policies (Waheed et al., 2006; Saghebian et al., 2014; Hannan and Anmala, 2021).

A.2 Comparing LPC and policy learning/causal subgroup/individualized treatment rules

Policy targeting can be largely divided into two cases: policy targeting with treatment testing and policy targeting without treatment testing. The former is a burgeoning literature termed policy learning/causal subgroup/individualized treatment rules. The LPC framework falls under the latter, imposing no requirement on treatment testing. Here we provide two common scenarios where LPC is useful but policy learning is not possible.

What if the treatment cannot be/has not yet been tested in real life?

Our methods do not need a randomized control trial/quasi-experiment/instrumental variable for unconfoundedness, which is a key assumption for the policy learning literature. Experimental data (or observational data that satisfies unconfoundedness) is not always available.

For instance, in the case of tax credit programs (the example in our introduction), it is nearly impossible to experiment since the tax credit program is a one-time, urgent assistance package for households during financial

crisis. Policymakers are unlikely to test out the assistance package with a random set of households and then evaluate how to allocate these assistance packages at the population level. There are two reasons. One, such an experiment is ethically questionable; two, by the time such an experiment is completed, the financial crisis might have already been over.

In such cases, classic policy learning is impractical. LPC can help refine policy design by identifying and targeting the subgroups that are more likely to be financially constrained. It is much faster for the policymakers (in this case, the Italian central bank) to collect data on households' financial information (e.g., whether the household is financially constrained) than to experiment.

What if the policymakers do not have a specific treatment in mind?

Another scenario that happens often is that the policymakers may decide the treatment after finding the vulnerable subgroups. In this case, testing the treatment is not possible, since there is no treatment to begin with.

Example 1: traffic safety

da Cruz Figueira et al. (2017) uses CART to find out that for the northern part of BR-116 (a highway in Sao Paulo, Brazil), the severe traffic rate (FSI in the paper) is higher than 50% (severe means involving human injury/death, the rate is the number of severe events divided by the total number of accidents). In contrast, the southern part of BR-116 has a higher than 50% FSI during the peak hours (12 pm to 6 pm) when there is drizzle. Policymakers can analyze these two geographic regions separately and use different interventions for the northern and southern parts of BR-116.

For the northern part, there might be some geographic features (maybe curvy turns) that make the road dangerous at all times, then installing traffic mirrors/stop signs/traffic lights at curvy turns would be good solutions. For the southern part, since it is during specific times, then sending out more traffic police during those hours to check on the speed limit would work better with minimal disruption to ongoing traffic.

Example 2: welfare take-up

Another example is Wang and Ahn (2025), in their paper, they find subgroups among eligible households for WIC (a welfare program) who have a low probability of choosing to participate in WIC (they use random forest instead of CART). They then analyze which subgroup is less likely to participate for what reason: unawareness, limited usefulness, hassle or stigma. For different reasons, different interventions are suggested. For example, an information campaign is suggested for higher-educated subgroups, and choice-inducing strategies are suggested for those who are already in the programs.

In both examples, the treatment is decided after the subgroups are found. This is inherently contradictory to the premise of policy learning that treatment has to be tested before finding the targeting subgroups.

B Mathematical appendix

We state a property of the CART splitting rule in Lemma B.1. This is useful for establishing the rest of the theoretical results.

Lemma B.1. $2\eta(s^{CART}) - \mu_L(s^{CART}) - \mu_R(s^{CART}) = 0$ *and* $\mu_L(s^{CART}) \neq \mu_R(s^{CART})$,

Proof of Lemma B.1 needs the following lemma, which is proved after the proof of Lemma B.1.

Lemma B.2. *Given the problem setup, for any node t, assume that $\eta(X)$ is continuous and not a constant,* $\frac{\partial \mathcal{G}^{CART}(s)}{\partial s} = 0$ *is a necessary condition for s to be an optimal split point.*

Proof for Lemma B.1.

Proof. For a fixed $s \in [0,1]$, the impurity score after the split is given as:

$$\begin{aligned} \mathcal{G}^{CART}(s) &= (\mu_L - \mu_L^2)F(s) + (\mu_R - \mu_R^2)(1 - F(s)) \\ &= E_t - \mu_L^2 F(s) - \mu_R^2(1 - F(s)), \end{aligned} \tag{2}$$

where $E_t = \int_0^1 \eta(x)f(x)dx$ is free of s. Moreover,

$$\frac{\partial \mu_L}{\partial s} = \frac{\partial \int_0^s \eta(x)f(x)dx/F(s)}{\partial s} = \frac{f(s)}{F(s)}\left(\eta(s) - \mu_L\right) \tag{3}$$

$$\frac{\partial \mu_R}{\partial s} = \frac{\partial \int_s^1 \eta(x)f(x)dx/(1-F(s))}{\partial s} = \frac{f(s)}{1-F(s)}\left(-\eta(s) + \mu_R\right), \tag{4}$$

Using (3) and (4), we simplify the first-order derivative

$$\begin{aligned}
\frac{\partial \mathcal{G}^{CART}(s)}{\partial s} =& -2\mu_L F(s)\frac{\partial \mu_L}{\partial s} - \mu_L^2 f(s) - 2\mu_R(1-F(s))\frac{\partial \mu_R}{\partial s} + \mu_R^2 f(s) \\
=& f(s)\left(-2\mu_L\left(\eta(s) - \mu_L\right) - \mu_L^2 - 2\mu_R\left(-\eta(s) + \mu_R\right) + \mu_R^2\right) \\
=& f(s)(-2\mu_L\eta(s) + E_{t_L}^2 + 2\mu_R\eta(s) - E_{t_R}^2) \\
=& f(s)(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) \\
\propto& (2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L).
\end{aligned} \tag{5}$$

By Lemma B.2, $\frac{\partial \mathcal{G}^{CART}(s)}{\partial s} = 0$ is a necessary condition for s to be an optimal split point.

Given that f is strictly positive, we have that $\frac{\partial \mathcal{G}^{CART}(s)}{\partial s} = 0$ iff $(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) = 0$. The next step is to rule out the possibility that $\mu_R - \mu_L = 0$ outputs a local maximum.

$$\begin{aligned}
\frac{\partial^2 \mathcal{G}^{CART}(s)}{\partial s^2} =& \underbrace{(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L)}_{=0 \text{ by the first-order condition}}\frac{\partial f(s)}{\partial s} + \frac{\partial (2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L)}{\partial s} f(s) \\
=& \frac{\partial (2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L)}{\partial s} f(s) \\
\propto& \frac{\partial (2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L)}{\partial s}
\end{aligned}$$

Consider the second derivative of $\mathcal{G}(s)$ with respect to s with $\mu_L = \mu_R$, by (3) and (4):

$$\begin{aligned}
\frac{\partial^2 \mathcal{G}^{CART}(s)}{\partial s^2} \propto& \frac{\partial (2\eta(s) - \mu_L - \mu_R)}{\partial s}\underbrace{(\mu_R - \mu_L)}_{=0} + (2\eta(s) - \mu_L - \mu_R)\left(\frac{\partial \mu_R}{\partial s} - \frac{\partial \mu_L}{\partial s}\right) \\
\propto& (2\eta(s) - \mu_L - \mu_R)\left(-\frac{\eta(s)}{1-F(s)} + \frac{\mu_R}{(1-F(s))} - \frac{\eta(s)}{F(s)} + \frac{\mu_L}{F(s)}\right) \\
=& \frac{2\eta(s) - \mu_L - \mu_R}{F(s)(1-F(s))}\left(-\eta(s) + \mu_L(1-F(s)) + \mu_R F(s)\right)
\end{aligned}$$

If $\mu_L = \mu_R$,

$$\frac{\partial^2 \mathcal{G}^{CART}(s)}{\partial s^2} \propto (2\eta(s) - \mu_L - \mu_R)(-\eta(s) + \mu_L) = -2(\eta(s) - \mu_L)^2.$$

For s s.t. $2\eta(s,t) - \mu_L - \mu_R \neq 0$ and $\mu_L = \mu_R$, $\frac{\partial \mathcal{G}^{CART}(s)}{\partial s} = 0$ and $\frac{\partial^2 \mathcal{G}^{CART}(s)}{\partial s^2} < 0$: $\mathcal{G}^{CART}(s)$ reaches its local maximum, which cannot be a global minimum. Such s is not the optimal split.

Global minimum must have $(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) = 0$ and $\mu_R - \mu_L \neq 0$. Therefore, $2\eta(s,t) - \mu_L - \mu_R = 0$ holds for optimal split s in any node t and dimension p. □

Proof for Lemma B.2.

Proof. We first rule out the possibility that the optimal splitting happens at the boundary point.

Consider $s = 0$, then we write the impurity score as

$$\begin{aligned}\mathcal{G}^{CART}(0) &= (\mu_L(0) - \mu_L^2(0))F(0) + (\mu_R(0) - \mu_R^2(0))(1 - F(0)) \\ &= \mu_R(0) - \mu_R^2(0) = \bar{\eta} - \bar{\eta}^2\end{aligned}$$

where $\bar{\eta}$ denotes the mean probability of $Y = 1$ for the entire node t.

Since $\eta(X)$ is not a constant, we can find $s = s'$ such that $\mu_L(s') \neq \mu_R(s')$ and in general $\bar{\eta} = \mu_L(s')F(s') + \mu_R(s')(1 - F(s'))$, where $0 < F(s') < 1$. The impurity score for $s = s'$ is

$$\mathcal{G}^{CART}(s') = (\mu_L(s') - \mu_L^2(s'))F(s') + (\mu_R(s') - \mu_R^2(s'))(1 - F(s'))$$

Using the equality $\bar{\eta} = \mu_L(s')F(s') + \mu_R(s')(1 - F(s'))$, we can rewrite $\mathcal{G}(0)$ and show that it is strictly greater than $\mathcal{G}(s')$.

$$\begin{aligned}\mathcal{G}^{CART}(0) &= \mu_L(s')F(s') + \mu_R(s')(1 - F(s')) - (\mu_L(s')F(s') + \mu_R(s')(1 - F(s')))^2 \\ &= \mu_L(s')F(s') + \mu_R(s')(1 - F(s')) - (\mu_L(s')F(s'))^2 - ((\mu_R(s')(1 - F(s')))^2 \\ &\quad - 2(\mu_L(s')F(s'))(\mu_R(s')(1 - F(s'))) \\ &= \mu_L(s')F(s') + \mu_R(s')(1 - F(s')) - (\mu_L(s')F(s'))^2 - ((\mu_R(s')(1 - F(s')))^2 \\ &\quad - \mu_L^2(s')F(s')(1 - F(s')) - \mu_R^2(s')F(s')(1 - F(s')) \\ &\quad + \mu_L^2(s')F(s')(1 - F(s')) + \mu_R^2(s')F(s')(1 - F(s')) \\ &\quad - 2(\mu_L(s')F(s'))(\mu_R(s')(1 - F(s'))) \\ &= \mathcal{G}^{CART}(s') + \mu_L^2(s')F(s')(1 - F(s')) + \mu_R^2(s')F(s')(1 - F(s')) \\ &\quad - 2(\mu_L(s')F(s'))(\mu_R(s')(1 - F(s'))) \\ &= \mathcal{G}^{CART}(s') + (\mu_L(s') - \mu_R(s'))^2F(s'))(1 - F(s')) > \mathcal{G}(s')\end{aligned}$$

The inequality means that $s = 0$ can never be the optimal split. The proof for the case of $s = 1$ is similar. Hence, we show that the optimal split is not the boundary point.

Given that $\mathcal{G}^{CART}(s)$ is differentiable, its domain is closed and compact, and the boundary points are not the optimal splits. The first-order condition must be satisfied at all interior local optima (including the global minimum whose argument is the optimal splitting).

□

Lemma B.3. *If splitting rule s' strictly dominates s, then $R(s') < R(s)$.*

Proof for Lemma B.3

Proof. Using Definition 3.1 and contrapositive,

- when $\eta(x) > c$, $\mu_{t(x)}(s') \leq c \implies \mu_{t(x)}(s) \leq c$ and
- when $\eta(x) \leq c$, $\mu_{t(x)}(s') > c \implies \mu_{t(x)}(s) > c$.

Therefore, $\forall\ x \in [0, 1]$,

$$\begin{aligned}\mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s') \leq c\} &\leq \mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s) \leq c\} \\ \mathbb{1}\{\eta(x) \leq c\}\mathbb{1}\{\mu_{t(x)}(s') > c\} &\leq \mathbb{1}\{\eta(x) \leq c\}\mathbb{1}\{\mu_{t(x)}(s) > c\}\end{aligned}$$

Also, there exists a set $\mathcal{A} \subseteq [0, 1]$ with nonzero measure such that, $\forall\ x \in \mathcal{A}$, either (or both) of the following conditions is true

$$\begin{aligned}\mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s') \leq c\} &< \mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s) \leq c\} \\ \mathbb{1}\{\eta(x) \leq c\}\mathbb{1}\{\mu_{t(x)}(s') > c\} &< \mathbb{1}\{\eta(x) \leq c\}\mathbb{1}\{\mu_{t(x)}(s) > c\}\end{aligned}$$

$$R(s') = \int_{\dot{X}_t}^{\ddot{X}_t} \left(\mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s') \le c\} + \mathbb{1}\{\eta(x) \le c\}\mathbb{1}\{\mu_{t(x)}(s') > c\}\right) f(x)dx$$
$$< \int_{\dot{X}_t}^{\ddot{X}_t} \left(\mathbb{1}\{\eta(x) > c\}\mathbb{1}\{\mu_{t(x)}(s) \le c\} + \mathbb{1}\{\eta(x) \le c\}\mathbb{1}\{\mu_{t(x)}(s) > c\}\right) f(x)dx = R_t(s)$$

□

Proof for Theorem 3.2

Proof. By Lemma B.1, without loss of generality[5], assume

$$\mu_L(s^{CART}) < \eta(s^{CART}) = \frac{\mu_L(s^{CART}) + \mu_R(s^{CART})}{2} < \mu_R(s^{CART}).$$

We sequentially prove the two claims in the lemma: the existence of s^* and strict dominance.

Existence of s^* When $\eta(s^{CART}, t) > c$, we have $(\eta(s^{CART}) - c)(\mu_R(s^{CART}) - \mu_L(s^{CART})) > 0$, so we want to show that $\exists\ s \in (0, s^{CART})$ such that $\eta(s) = c$.

$$\mu_L = \min(\mu_L, \mu_R) \le c \implies \exists\ \tilde{s} \in [0, s^{CART}] \text{ such that } \eta(\tilde{s}) \le c$$

If $\eta(\tilde{s}) = c$, then the existence of s^* is proven. If $\eta(\tilde{s}) < c$, then by intermediate value theorem, $\exists\ s \in (\tilde{s}, s^{CART})$ such that $\eta(s) = c$, the proof for existence of s^* is complete for $\eta(s^{CART}, t) > c$. The case when $\eta(s^{CART}, t) \le c$ can be proved using the same logic.

We showed that when $\eta(s^{CART}, t) > c$, there exists s^* and the s^* is in the range $(0, s^{CART})$. By continuity of η and the definition of s^*, $\forall\ s \in (s^*, s^{CART}), \eta(s) > c$. By the same logic, we have that when $\eta(s^{CART}, t) \le c$, $\forall\ s \in (s^{CART}, s^*), \eta(s) \le c$.

Strict dominance We compare risks led by s^{CART} and s^*:
1. If $\eta(s^{CART}) \le c$, for any $s \in (s^{CART}, s^*)$, we have $\eta(s) \le c$ and

$$\mu_L(s) = \int_0^s \eta(x)f(x)dx/F(s) = \frac{\int_0^{s^{CART}} \eta(x)f(x)dx + \int_{s^{CART}}^{s^*} \eta(x)f(x)dx}{F(s)}$$
$$< \frac{\int_0^{s^{CART}} \eta(x)f(x)dx + c\int_{s^{CART}}^{s^*} f(x)dx}{F(s)} = \frac{\mu_L(s^{CART})F(s^{CART}) + c(F(s) - F(s^{CART}))}{F(s)} < c$$
$$\mu_R(s) > \frac{\mu_R(s^{CART})(1 - F(s^{CART})) - c(F(s) - F(s^{CART}))}{1 - F(s)} > c.$$

Since the order that $\mu_L(s) \le c < \mu_R(s)$ does not change and $\forall\ s \in (s^{CART}, s^*), \eta(s) \le c$, we have that when $X \in [0, s^{CART}]$, splitting rules s and s^{CART} are equivalent in classifying the latent probability; when $X \in (s^{CART}, s)$ which has a nonzero measure, splitting rule s performs strictly better than s^{CART} in classifying the latent probability; when $X \in [s, 1]$, splitting rules s and s^{CART} are equivalent in classifying the latent probability. Therefore, s strictly dominates s^{CART}.
2. If $\eta(s^{CART}) > c$, for $s \in (s^*, s^{CART})$, we have $\eta(s) > c$, s strictly dominates s^{CART}. □

Theoretical property of KD-CART

To explore the theoretical property of KD-CART, we treat it as a CART that takes true $\eta(x)$ as input and splits the population based on $\eta(x)$. KD-CART uses criterion function $\mathcal{G}^{KD} := F_X(s)\text{Var}(\eta(X)|X \le s) + (1 - F_X(s))\text{Var}(\eta(X)|X > s)$.

Lemma B.4. *Define* $s^{KD} := \arg\min_{s\in(0,1)} \mathcal{G}^{KD}(s)$. $2\eta(s^{KD}) - \mu_L(s^{KD}) - \mu_R(s^{KD}) = 0$ *and* $\mu_L(s^{KD}) \ne \mu_R(s^{KD})$.

[5]Proof for the case $\mu_R(s^{CART}) < \eta(s^{CART}) < \mu_L(s^{CART})$ is similar.

The proof of Lemma B.4 is almost identical to Lemma B.1 and is omitted here. The key take away is Theorem 3.2 also applies to s^{KD}, as proof of Theorem 3.2 only requires that $\eta(s^{CART})$ is strictly between μ_L and μ_R, Lemma B.4 shows that $\eta(s^{KD})$ satisfies this condition. As a result, KD-CART does not minimize the misclassification risk.

Proof of Remark 4.3

The proof mostly consists of proving two claims.

WLOG, $\mu_L(s^*) \leq c < \mu_R(s^*)$.

Claim 1: $\forall x \in [0,1], s^*$ targets correctly.

Proof.

By the uniqueness of intersection between $\eta(X)$ and c, we have that $\forall x \in [0, s^*), \eta(x) < c$ and that $\forall x \in (s^*, 1], \eta(x) > c$.

Consider left node,

$$\mu_L(s^*) = \frac{1}{s^*} \int_0^{s^*} \eta(x)\, dx < \frac{1}{s^*} \int_0^{s^*} c\, dx = c.$$

Hence, we do not target the left node. All $x \in [0, s^*]$ are correctly NOT targeted.

Similarly, we can show that all $x \in (s^*, 1]$ are correctly targeted. This completes the proof of Claim 1.

Claim 2: $\forall s' \neq s^*, s^*$ strictly dominates s'.

Proof.

WLOG, $s' < s^*$.

- **Case 1:** If $\mu_R(s') > c$, then we target the right node. For all $x \in (s', s^*)$, their $\eta(x) < c$, these points are incorrectly targeted. (Using the notation from Definition 3.1, $\mathcal{A} = (s', s^*)$ in Case 1)

- **Case 2:** If $\mu_R(s') \leq c$, then we do not target the right node. For all $x \in (s^*, 1]$, their $\eta(x) > c$, these points are incorrectly not targeted. (Using the notation from Definition 3.1, $\mathcal{A} = (s^*, 1]$ in Case 2)

In either case, s^* strictly dominates s'. This completes the proof for Claim 2.

Explanation for discrete X

Say $X \in \{a_1, a_2, \ldots, a_K\}$ where $a_1, a_2, \ldots, a_K$ are ordered and $a_1 < a_2 < \cdots < a_K$. Assumption 4.1 (a discrete X version) would guarantee that for one unique split rule, for all possible values of X in the left node, $\eta(X) > c$ and for all possible values of X in the right node, $\eta(X) < c$, or vice versa. We can identify this unique rule by simply comparing $\eta(a_k)$ and c for $k = 1, 2, \ldots, K$. Estimation boils down to estimating $\eta(X)$ for finitely many X, which is trivial.

We find it clearer to prove Corollary B.5 and then plug in some of the steps in the proof for Corollary B.5 into proof for Theorem 4.4. Hence, we first present the proof for Corollary B.5.

Corollary B.5. *If $X \sim \text{Unif}[0,1]$, and $\eta(X)$ is monotonic $\forall\ X \in [0,1]$ and is strictly monotonic and differentiable in a neighborhood of s^*, then* $\arg\max_s \mathcal{G}^*(s,c)$ *identifies s^*.*

Proof for Corollary B.5

Proof. WLOG, assume η is monotonically increasing with respect to x. The proof for the monotonically decreasing case is similar. We will first discuss two distinct cases: $\mu_L < \mu_R \leq c$ and $c \leq \mu_L < \mu_R$ and show that for both cases $\mathcal{G}^*$ is constant. Then, we consider the case where $\mu_L \leq c \leq \mu_R$ and show that s^* maximizes $\mathcal{G}^*$ in this case. We complete the proof by showing that $\mathcal{G}^*(s^*, c)$ is larger than the two constant $\mathcal{G}^*$ for the first two cases.

i. $\mu_L < \mu_R \leq c$, we can write $\mathcal{G}^*$ as

$$s\left(c - \frac{\int_0^s \eta(x)dx}{s}\right) + (1-s)\left(c - \frac{\int_s^1 \eta(x)dx}{1-s}\right) = c - \int_0^1 \eta(x)dx$$

Note that in this case, $\mathcal{G}^*$ is a constant.

ii. $c \leq \mu_L < \mu_R$, we can write $\mathcal{G}^*$ as

$$s\left(\frac{\int_0^s \eta(x)dx}{s} - c\right) + (1-s)\left(\frac{\int_s^1 \eta(x)dx}{1-s} - c\right) = \int_0^1 \eta(x)dx - c$$

Note that in this case, $\mathcal{G}^*$ is a constant.

iii. $\mu_L \leq c \leq \mu_R$, we can write $\max_s \mathcal{G}^*$ as

$$\max_s s\left(c - \frac{\int_0^s \eta(x)dx}{s}\right) + (1-s)\left(\frac{\int_s^1 \eta(x)dx}{1-s} - c\right)$$

The first-order condition is $c - \eta(s) - \eta(s) + c = 0$. We solve $\eta(s) = c$, hence, s^* is a local optima. Moreover, second order derivative is $-2\eta'(s^*)$ which is negative given that η is strictly increasing and differentiable in a neighborhood of s^*. Hence, s^* is a local maxima. Since $\mathcal{G}^*$ is continuous in s, we need to show that $\mathcal{G}^*(s^*)$ is larger than the boundary points for the case $\mu_L \leq c \leq \mu_R$ to claim s^* as the global maxima for case iii. There are two possibilities for the boundary points

Possibility 1: Consider s_1 and s_2 and their associated $\mathcal{G}^*(s_1)$ and $\mathcal{G}^*(s_2)$ where $\mu_L(s_1) = c$ and $\mu_R(s_2) = c$. Since s_1 is included in case ii; whereas s_2 is included in case i, once we show that $\mathcal{G}^*(s^*) > \mathcal{G}^*(s_1), \mathcal{G}^*(s_2)$, we can claim s^* to be the unique global maxima among all possible s.

$$\begin{aligned}
\mathcal{G}^*(s^*) =& s^*c - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx - (1-s^*)c \\
>& s^*c - \int_0^{s^*} \eta(x)dx + (1-s^*)c - \int_{s^*}^1 \eta(x)dx \\
=& c - \int_0^1 \eta(x)dx = \mathcal{G}^*(s_2) \\
\mathcal{G}^*(s^*) =& s^*c - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx - (1-s^*)c \\
>& \int_0^{s^*} \eta(x)dx - s^*c + (1-s^*)c - (1-s^*)c \\
=& \int_0^1 \eta(x)dx - c = \mathcal{G}^*(s_1)
\end{aligned}$$

Possibility 2: Even at the boundary point, $\mu_L \leq c \leq \mu_R$ still holds. Then, the boundary points for case iii are $s = 0$ and $s = 1$.

$$\begin{aligned}
\tilde{G}(0) =& \int_0^1 \eta(x)dx - c < \mathcal{G}^*(s^*) \\
\tilde{G}(1) =& c - \int_0^1 \eta(x)dx < \mathcal{G}^*(s^*)
\end{aligned}$$

Since under Possibility 2 case iii spans the entire domain of X, s^* is the unique global maxima. Note that Assumption 4.1 implies that $\exists\ \epsilon > 0$, such that $s^* \in [\epsilon, 1-\epsilon] \subset [0,1]$. Therefore, s^* is the global maxima among all possible s. □

Proof for Theorem 4.4

Proof. Again, assume that η is monotonically increasing in the neighborhood of s^*, the proof for the monotonically decreasing case is the same. Assumption 4.1 guarantees $\forall s < s^*, \eta(X) < c$, $\forall s > s^*, \eta(X) > c$.

Let $[s^* - \epsilon_1, s^* + \epsilon_1]$ denote the monotonically increasing η neighborhood of s^*, where $\epsilon_1 > 0$. Let $[s^* - \epsilon_2, s^* + \epsilon_2]$ denote an interval that contains s^* in which $\forall s \in [s^* - \epsilon_2, s^* + \epsilon_2], \mu_L < c < \mu_R$ and $\epsilon_2 > 0$. The existence of $[s^* - \epsilon_2, s^* + \epsilon_2]$ is guaranteed because both μ_L and μ_R are continuous in s and given Assumption 4.1, $\mu_L(s^*) < c < \mu_R(s^*)$.

Consider $s \in [s^* - \epsilon_1, s^* + \epsilon_1] \cap [s^* - \epsilon_2, s^* + \epsilon_2] = [a, b]$, where $a = \max(s^* - \epsilon_1, s^* - \epsilon_2)$ and $b = \min(s^* + \epsilon_1, s^* + \epsilon_2)$. This case is equivalent to case iii in the proof for Corollary B.5. Hence, s^* is a local maxima for the interval $[a, b]$ and $\mathcal{G}^*$ is

$$s^* c - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx - (1 - s^*)c.$$

Consider $s \leq a$. When $\mu_R \leq c$, $\mathcal{G}^*$ is $c - \int_0^1 \eta(x)dx$, the proof is identical to case i in the proof for Corollary B.5. When $\mu_R > c$,

$$\begin{aligned}
\mathcal{G}^* =& sc - \int_0^s \eta(x)dx + \int_s^1 \eta(x)dx - (1 - s)c \\
=& sc + \int_s^{s^*} \eta(x)dx - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx + \int_s^{s^*} \eta(x)dx - (1 - s)c \\
<& sc + (s^* - s)c - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx + (s^* - s)c - (1 - s)c \\
=& s^* c - \int_0^{s^*} \eta(x)dx + \int_{s^*}^1 \eta(x)dx - (1 - s^*)c
\end{aligned}$$

The proof for when $s \geq b$ is similar, those s results in higher $\mathcal{G}^*$.

Since s^* is the unique point in the interval $[a, b]$ which meets the first-order condition and the boundary points $\{a, b\}$ have lower $\mathcal{G}^*$ than s^*, s^* is the unique maxima in the interval $[a, b]$. Moreover, $s \in [0, a] \cup [b, 1]$ also have lower $\mathcal{G}^*$ than s^*, hence, s^* is the unique global maxima. □

Proof for Theorem 4.5

Proof. **Outline**: We leverage tools for studying asymptotic properties of M-estimator. To show $\hat{s} \xrightarrow{p} s^*$, we take two steps:

1. Under Assumption 4.1, we show uniform convergence of the estimator of cost function $\widehat{\mathcal{G}}^*$ to its population target $\mathcal{G}^*$, i.e., $\sup_{s \in (\epsilon, 1-\epsilon)} |\mathcal{G}^*(s, c) - \widehat{\mathcal{G}}^*(s, c)| \xrightarrow{p} 0$ as $n \to \infty$.

2. Combining with Theorem 4.4, we can apply Theorem 5.7 of Van der Vaart (2000) and get the desired result.

Step 1. For any given $c \in (0,1)$ and $\epsilon_0 > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} |\mathcal{G}^*(s,c) - \widehat{\mathcal{G}}^*(s,c)| \geq \epsilon_0 \right) \\
&= \mathbb{P}\Bigg(\sup_{s\in(\epsilon,1-\epsilon)} \Bigg| s\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\}}{\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} - c\right| - s|\mu_L - c| \\
&+ (1-s)\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i > s\}}{\sum_{i=1}^n \mathbb{1}\{X_i > s\}} - c\right| - (1-s)|\mu_R - c|\Bigg| \geq \epsilon_0 \Bigg) \\
&\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} s\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\}}{\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} - \mu_L\right| \geq \epsilon_0/2 \right) \qquad (6) \\
&+ \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} (1-s)\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i > s\}}{\sum_{i=1}^n \mathbb{1}\{X_i > s\}} - \mu_R\right| \geq \epsilon_0/2 \right) \qquad (7)
\end{aligned}
$$

We focus on showing (6) goes to 0 as $n \to \infty$, the proof for (7) is exactly the same.

For any $\delta > 0$, denote event $\mathcal{E} = \inf_{s\in(\epsilon,1-\epsilon)} n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq s\}} \geq \delta$. We have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} s\left|\frac{\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\}}{\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} - \mu_L\right| \geq \epsilon_0/2 \right) \\
&\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} s\left(\frac{\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right|}{n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} + \frac{\mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right|}{n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \right) \geq \epsilon_0/2 \right) \\
&\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \frac{s\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right|}{n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4 \right) + \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \frac{\mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right|}{n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4 \right) \\
&\leq \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} s\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4 \right) + \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} \mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4 \right) \\
&\leq \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} s\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4, \mathcal{E} \right) + \\
&\quad \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} s\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4, \mathcal{E}^c \right) + \\
&\quad \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} \mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4, \mathcal{E} \right) + \\
&\quad \mathbb{P}\left(\frac{\sup_{s\in(\epsilon,1-\epsilon)} \mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right|}{\inf_{s\in(\epsilon,1-\epsilon)} n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\}} \geq \epsilon_0/4, \mathcal{E}^c \right) \qquad (8) \\
&\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} s\left|n^{-1}\sum_{i=1}^n Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right| \geq (\epsilon_0\delta)/4 \right) + \mathbb{P}(\mathcal{E}^c) \qquad (9) \\
&+ \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \mu_L\left|n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq s\} - s\right| \geq (\epsilon_0\delta)/4 \right) + \mathbb{P}(\mathcal{E}^c), \qquad (10)
\end{aligned}
$$

where (8) is by law of total probability. For the first term in (9), let $\boldsymbol{Z}_i = (X_i, Y_i)$, $i = 1, \ldots, n$ and $f(\boldsymbol{Z}_i, s) = Y_i \mathbb{1}_{\{X_i \leq s\}}$. Notice $\mathbb{E}[f(\boldsymbol{Z}, s)] = s\mu_L$, $f(\boldsymbol{Z}, s)$ is continuous at each $s \in (0,1)$ for almost all $\boldsymbol{Z}$, since discontinuity occurs at $X_i = s$, which has measure zero for $X_i \sim \text{Unif}(0,1)$. Also, $f(\boldsymbol{Z}, s) \leq \mathbb{1}_{\{0 \leq X \leq 1\}}$ with $\mathbb{E}[\mathbb{1}_{\{0 \leq X \leq 1\}}] =$

$1 \leq \infty$. Applying uniform law of large numbers:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} s\left|n^{-1}\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right| \geq (\epsilon_0\delta)/4 \right) \\
&\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \left|n^{-1}\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right| \geq (\epsilon_0\delta)/4 \right) \\
&\leq \mathbb{P}\left(\sup_{s\in(0,1)} \left|n^{-1}\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i \leq s\} - s\mu_L\right| \geq (\epsilon_0\delta)/4 \right) \to 0, \qquad (11)
\end{aligned}$$

as $n \to \infty$. For the first term in (10), notice $F_n(s) = n^{-1}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq s\}$ is the empirical CDF of $X \sim \text{Unif}(0,1)$, with CDF $F_X(s) = s$ for $s \in (0,1)$. Applying Dvoretzky–Kiefer–Wolfowitz inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \mu_L\left|n^{-1}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq s\} - s\right| \geq (\epsilon_0\delta)/4 \right) \\
(\mu_L \in [0,1]) &\leq \mathbb{P}\left(\sup_{s\in(\epsilon,1-\epsilon)} \left|n^{-1}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq s\} - s\right| \geq (\epsilon_0\delta)/4 \right) \\
&\leq \mathbb{P}\left(\sup_{s\in(0,1)} \left|n^{-1}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq s\} - s\right| \geq (\epsilon_0\delta)/4 \right) \to 0, \qquad (12)
\end{aligned}$$

as $n \to \infty$.

For the second terms in (9) and (10), we want to show $\mathbb{P}(\mathcal{E}^c) \to 0$ as $n \to \infty$. By Glivenko–Cantelli theorem, we have $\sup_{s\in(0,1)} |F_n(s) - F_X(s)| \xrightarrow{a.s.} 0$, which implies $\sup_{s\in(\epsilon,1-\epsilon)} |F_n(s) - s| \xrightarrow{a.s.} 0$. The uniform convergence means for all $s \in (\epsilon, 1-\epsilon)$ and any $\epsilon' > 0$, there exists N such that for all $n \geq N$, we have $|F_n(s) - s| < \epsilon'$, which implies $|\inf_{s\in(\epsilon,1-\epsilon)} F_n(s) - \inf_{s\in(\epsilon,1-\epsilon)} s| = |\inf_{s\in(\epsilon,1-\epsilon)} F_n(s) - \epsilon| < \epsilon'$. This shows $\inf_{s\in(\epsilon,1-\epsilon)} F_n(s) \xrightarrow{p} \epsilon$. Finally, setting $\delta = \epsilon/2$, we have $\inf_{s\in(\epsilon,1-\epsilon)} F_n(s) - \delta \xrightarrow{p} \epsilon/2$. Hence $\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\inf_{s\in(\epsilon,1-\epsilon)} F_n(s) - \delta < 0) \to 0$ as $n \to \infty$.

Combining (9) and (10), we can show (6) goes to 0 as $n \to \infty$. Similarly, (7) goes to 0 as $n \to \infty$. This gives us $\mathbb{P}(\sup_{s\in(\epsilon,1-\epsilon)} |\mathcal{G}^*(s) - \widehat{\mathcal{G}}^*(s)| \geq \epsilon_0) \to 0$ as $n \to \infty$, which completes the proof of Step 1.

Step 2. By Theorem 4.4, we have for any $\epsilon > 0$, $\sup_{s:d(s,s^*)>\epsilon} \mathcal{G}^*(s,c) < \mathcal{G}^*(s^*,c)$. Combining with $\sup_{s\in(\epsilon,1-\epsilon)} |\mathcal{G}^*(s,c) - \widehat{\mathcal{G}}^*(s,c)| \xrightarrow{p} 0$, we apply Theorem 5.7 of Van der Vaart (2000) and conclude $\hat{s} \xrightarrow{p} s^*$.

□

Intuition for Theorem 5.1

We use Figure 4 to illustrate the intuition behind Theorem 5.1 based on the same setting as that in Figure 1a. From top to bottom, Figure 4 depicts $\mathcal{G}^{CART}$, penalty term J, $\mathcal{G}^{PFS}$ and LPC misclassification risk R as functions of feature X. The green curve shows that $\mathcal{G}^{CART}$ is minimized at s^{CART}. The green curve is also very *flat* around s^{CART}. In contrast, the blue curve shows that penalty J increases *sharply* at $X = s^{CART}$. Minimizing J would shift the split rule from s^{CART} to the left, as indicated by the arrow. Since $\mathcal{G}^{PFS} = \mathcal{G}^{CART} + \lambda J$, the gradient of $\mathcal{G}^{PFS}$ at s^{CART} is largely determined by the gradient of J, given how flat $\mathcal{G}^{CART}$ is at s^{CART} (See the black curve). Hence, minimizing $\mathcal{G}^{PFS}$ also shifts the split rule from s^{CART} to the left (i.e., towards s^*, the minimizer of the LPC misclassification

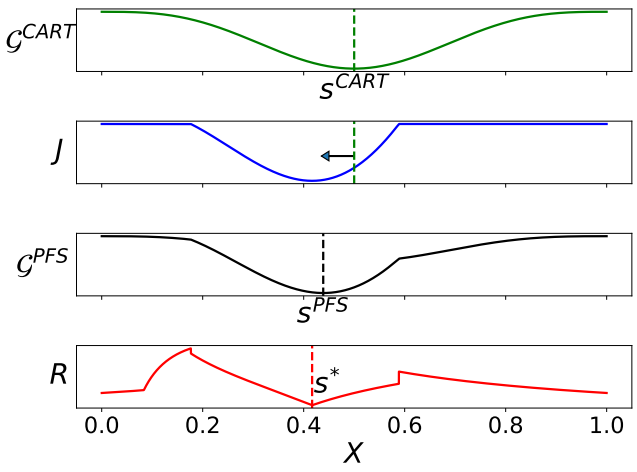


Figure 4: Visualization of the different loss components of PFS.

risk as shown by the red curve). Our formal proof in Appendix B shows that the *flatness* of $\mathcal{G}^{PFS}$ and *sharp* change in J at $X = s^{CART}$ and shifting s^{CART} towards s^* are not coincidences; they are generally true under the assumptions in Theorem 5.1.

Proof for Theorem 5.1

Proof. Consider

$$\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s} = \frac{\partial \mathcal{G}^{CART}(s,c) + \lambda W(|\mu_L - c|)F(s) + \lambda W(|\mu_R - c|)(1-F(s))}{\partial s}.$$

We focus on the new term $\frac{\partial W(|\mu_L - c|)F(s)}{\partial s}$ and $\frac{\partial W(|\mu_R - c|)(1-F(s))}{\partial s}$. By (3) and (4),

$$\begin{aligned}
\frac{\partial W(|\mu_L - c|)F(s)}{\partial s} &= f(s)W(|\mu_L - c|) + \frac{\partial W(|\mu_L - c|)}{\partial \mu_L}\frac{\partial \mu_L}{\partial s}F(s) \\
&= f(s)\left(W(|\mu_L - c|) + \frac{\partial W(|\mu_L - c|)}{\partial \mu_L}(\eta(s) - \mu_L)\right) \\
\frac{\partial W(|\mu_R - c|)(1-F(s))}{\partial s} &= -f(s)W(|\mu_R - c|) + \frac{\partial W(|\mu_R - c|)}{\partial \mu_R}\frac{\partial \mu_R}{\partial s}(1-F(s)) \\
&= f(s)\left(-W(|\mu_R - c|) + \frac{\partial W(|\mu_R - c|)}{\partial \mu_R}(\mu_R - \eta(s))\right).
\end{aligned}$$

With $\mathcal{G}^{PFS}(s,c)$'s derivative with respect to s, we evaluate how $\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}$ behaves at s^{CART}, i.e., when $2\eta(s) - \mu_L - \mu_R = 0$ and at s^*, i.e., when $\eta(s) = c$ respectively:

$$\begin{aligned}
\frac{1}{\lambda f(s)}\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^{CART}} &= W(|\mu_L - c|) - W(|\mu_R - c|) \\
&\quad + \left(\frac{\partial W(|\mu_L - c|)}{\partial \mu_L} + \frac{\partial W(|\mu_R - c|)}{\partial \mu_R}\right)(\eta(s^{CART}) - \mu_L) \qquad (13) \\
\frac{1}{f(s)}\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^*} &= (2c - \mu_L - \mu_R)(\mu_R - \mu_L) + \lambda(W(|\mu_L - c|) - W(|\mu_R - c|)) \\
&\quad + \lambda\left(\frac{\partial W(|\mu_L - c|)}{\partial \mu_L}(c - \mu_L) + \frac{\partial W(|\mu_R - c|)}{\partial \mu_R}(\mu_R - c)\right). \qquad (14)
\end{aligned}$$

Note that s^* in (14) is only guaranteed to exist when the condition of c in Theorem 3.2 is met. We will address this in the latter part of the proof. Next, consider all five possible scenarios of c's location separately. (Still, without loss of generality, assume $\mu_L(s^{CART}) < \eta(s^{CART}) < \mu_R(s^{CART})$.)

i. $\mu_L(s^{CART}) \le c < \eta(s^{CART}) < \mu_R(s^{CART})$.

By Lemma B.1,$|\mu_R(s^{CART}) - c| > |\mu_L(s^{CART}) - c|$. $W(|\mu_L - c|) > W(|\mu_R - c|)$ by the monotonicity of W and $\frac{\partial W(|\mu_L - c|)}{\partial \mu_L} + \frac{\partial W(|\mu_R - c|)}{\partial \mu_R} \ge 0$ by the convexity of W. Thus, by (13),

$$\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^{CART}} > 0. \qquad (15)$$

By Lemma B.1, and continuity of η, μ_L and μ_R,

$$(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) < 0,\ s \in (s^{sp}, s^{CART}),$$

where $s^{sp} := \max\{s : s < s^{CART}, (2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) = 0\}$. If there is no such $s^{sp} > 0$ exists, let $s^{sp} = 0$. When $s \in (\max(s^*, s^{sp}), s^{CART})$, by Theorem 3.2, $\mu_L < c$ and $\mu_R > c$.

When $s^* > s^{sp}$, define $d_L := c - \mu_L$ and $d_R := \mu_R - c$ and we have

$$0 > (2c - \mu_L(s^*) - \mu_R(s^*))(\mu_R(s^*) - \mu_L(s^*)) = (d_L - d_R)(d_L + d_R),$$

which directly suggests that $d_L < d_R$. Let $W_2(d) = d^2 + \lambda\,(W(d) - dW'(d))$. We can rewrite (14) as:

$$\begin{aligned}\frac{1}{f(s)}\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^*} &= d_L^2 + \lambda\left(W(d_L) - d_L\frac{\partial W(d_L)}{\partial d_L}\right) - d_R^2 - \lambda\left(W(d_R) - d_R\frac{\partial W(d_R)}{\partial d_R}\right)\\ &= W_2(d_L) - W_2(d_R).\end{aligned}$$

Note that $W_2^{'}(d) = d(2-\lambda W^{''}(d))$. By the bounded second derivative assumption, set $\Lambda_1 = \frac{1}{\max\{W^{''}(d_L), W^{''}(d_R)\}}$. Since $0 < d_L < d_R$, when $\lambda < \Lambda_1$, $W_2(d_L) < W_2(d_R)$,

$$\frac{1}{f(s)}\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^*} < 0.$$

When $s^* < s^{sp}$, there exists $s^{in} \in (s^{sp}, s^{CART})$ and $c > 0$ s.t.

$$(2\eta(s) - \mu_L - \mu_R)(\mu_R - \mu_L) < -c,$$

by unique global minimum assumption and continuity. With $\Lambda_1 = \frac{c}{W(0)+W'(0)}$, we have

$$\frac{1}{f(s)}\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^{in}} < 0.$$

Thus, for $0 < \lambda < \Lambda_1$, there exists a $s \in [s^*, s^{CART})$ s.t.

$$\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s} < 0. \tag{16}$$

Considering (15) and (16) jointly with the continuity of the derivative, we can conclude that there exists a $s^{**} \in (s^*, s^{CART})$ such that $\left.\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}\right|_{s=s^{**}} = 0$. s^{**} is a local minimizer of $\mathcal{G}^{PFS}(s,c)$ and is guaranteed to give a smaller risk than s^{CART} by Theorem 3.2.

Next, we argue that with some additional bound for λ, s^{**} is also the global optimal split under the penalized impurity measure. To simplify the notation in the rest of the proof, let s^{in} be s^* if $s^* < s^{sp}$. Then, we want to show $\mathcal{G}^{PFS}(s^{**}, c) < \mathcal{G}^{PFS}(s,c)$ for $s \neq s^{**}$:

(a). Consider s such that there is at least one s^{sp} in between s and s^{**}. By unique optimizer assumption for G, if there is more than one local minimum, i.e., multiple s that have $2\eta(s,t) - \mu_L - \mu_R = 0$, we assume the global minimum and the second smallest local minimum, gives by s^{sc}, different by Δ. Let $\Lambda_2 = \frac{\Delta}{W(0) - \min(W(c), W(1-c))}$. For $0 < \lambda < \Lambda_2$,

$$\begin{aligned}\mathcal{G}^{PFS}(s^{**}, c) &< \mathcal{G}^{PFS}(s^{CART}, c)\\ &= \mathcal{G}(s^{CART}, c) + \lambda\,(W(|\mu_L - c|)F(s) + W(|\mu_R - c|)(1 - F(s)))\\ &\leq \mathcal{G}(s^{CART}, c) + \lambda W(0)\\ &\leq \mathcal{G}(s^{sc}, c) + \lambda \min\{W(c), W(1-c)\} \leq \mathcal{G}^{PFS}(s,c)\end{aligned}$$

(b). Consider s such that there is no s^{sp} in between s and s^{**}.

When $s \in (s^*, s^{CART})$, $\mathcal{G}^{PFS}(s^{**}, c) < \mathcal{G}^{PFS}(s,c)$ by definition.

When $s < s^*$, by (16) and continuity of $\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s}$, there can be two possible scenarios. One, there exist a $s_0 < s^*$ such that $\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s} = 0$ and $\mathcal{G}^{PFS}(s_0, c) - \mathcal{G}^{PFS}(s^{**}, c) := \Delta_L > 0$. Then, define $\Lambda_3 = \frac{\Delta_L}{W(0) - \min(W(c), W(1-c))}$. For $0 < \lambda < \Lambda_3$, $\mathcal{G}^{PFS}(s^{**}, c) < \mathcal{G}^{PFS}(s,c)$ when $s_0 \leq s < s^*$ by definition. When $s < s_0$:

$$\begin{aligned}\mathcal{G}^{PFS}(s^{**}, c) &= \mathcal{G}^{PFS}(s_0, c) - \Delta_L\\ &= \mathcal{G}(s_0, c) + \lambda\,(W(|\mu_L - c|)F(s) + W(|\mu_R - c|)(1 - F(s))) - \Delta_L\\ &\leq \mathcal{G}(s_0, c) + \lambda W(0) - \Delta_L\\ &\leq \mathcal{G}(s_0, c) + \lambda \min\{W(c), W(1-c)\}\\ &\leq \mathcal{G}(s, c) + \lambda\,(W(|\mu_L - c|)F(s) + W(|\mu_R - c|)(1 - F(s))) = \mathcal{G}^{PFS}(s,c).\end{aligned}$$

Two, $\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s} < 0$. Then $\mathcal{G}^{PFS}(s^{**}, c) < \mathcal{G}^{PFS}(s, c)$ for $s < s^*$ follows.

When $s > s^{CART}$, we can follow the same procedure as when $s < s^*$. If there exists a $s_1 > s^*$ such that $\frac{\partial \mathcal{G}^{PFS}(s,c)}{\partial s} = 0$ and $\mathcal{G}^{PFS}(s_1, c) - \mathcal{G}^{PFS}(s^{**}, c) := \Delta_R > 0$. We can define $\Lambda_4 = \frac{\Delta_R}{W(0) - \min(W(c), W(1-c))}$ and everything else follows.

To conclude, for $0 < \lambda \leq \Lambda = \min\{\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4\}$, the theoretical optimal split chosen with respect to G^{PFS} leads to a risk smaller than s^{CART} when $\mu_L(s^{CART}) \leq c < \eta(s^{CART}) < \mu_R(s^{CART})$.

ii. $\mu_L(s^{CART}) < \eta(s^{CART}) < c \leq \mu_R(s^{CART})$. This scenario is symmetric to **i** and is also considered in Theorem 3.2. It can be proved following the same logic as in **i**.

iii. $c < \mu_L(s^{CART}) < \eta(s^{CART}) < \mu_R(s^{CART})$.

In this scenario, consider two split points located on each side of s^{CART} respectively: let s_{c1} be the largest split less than s^{CART} such that $\mu_L(s_{c1}) = c$, s_{c2} be the smallest split greater than s^{CART} such that $\mu_L(s_{c2}) = c$, s_L be the largest split less than s^{CART} such that $\eta(s_L) \geq \mu_L(s^{CART})$ and s_R be the smallest split greater than s^{CART} such that $\eta(s_R) = \mu_R(s^{CART})$. Let $s_{c1} = 0$ or $s_{c2} = 1$ if no such s_{c1} or s_{c2} exists. Both s_L and s_R's existence is guaranteed by the intermediate value theorem which we used once in Theorem 3.2.

We argue for any $s \in (\max(s_L, s_{c1}), \min(s_R, s_{c2}))$, risk is no greater than s^{CART}.

For $s \in (\max(s_L, s_{c1}), s^{CART})$, it is a piece of x that has $\eta(x) > \mu_L(s^{CART})$ being split to right.

$$\mu_R(s) > \frac{\mu_R(s^{CART})(1 - F(s^{CART})) + \mu_L(s^{CART})(F(s^{CART}) - F(s))}{1 - F(s)} > \mu_L(s^{CART}) > c$$

$$c < \mu_L(s) < \frac{\mu_L(s^{CART})F(s^{CART}) - \mu_L(s^{CART})(F(s^{CART}) - F(s))}{F(s)} = \mu_L(s^{CART}).$$

Since $c < \mu_L(s) < \mu_L(s^{CART}) < \mu_R(s)$, the expectations on both sides remain greater than c and the risk is the same as split s^{CART}. When $s \in (s^{CART}, \min(s_R, s_{c2}))$, $c < \mu_L(s) < \mu_R(s^{CART}) < \mu_R(s)$ can be obtained similarly and the risk also remains.

Note that $\mu_R(s) > \mu_L(s)$ all the way when $s \in (\max(s_L, s_{c1}), \min(s_R, s_{c2}))$. We can then define Δ the same as in scenario **i**, $\Delta_L = \mathcal{G}^{PFS}(\max(s_L, s_{c1}), c)$, and $\Delta_R = \mathcal{G}^{PFS}(\min(s_R, s_{c2}), c)$. $\Lambda_2, \Lambda_3, \Lambda_4$ are defined accordingly. For $0 < \lambda \leq \Lambda = \min\{\Lambda_2, \Lambda_3, \Lambda_4\}$, the theoretical optimal split chosen with respect to G^{PFS} leads to the same risk as s^{CART} when $c < \mu_L(s^{CART}) < \eta(s^{CART}) < \mu_R(s^{CART})$.

iv. $\mu_L(s^{CART}) < \eta(s^{CART}) < \mu_R(s^{CART}) < c$. This scenario is symmetric to **iii**. It can be proved following the same logic in **iii**.

v. $\mu_L(s^{CART}) < \eta(s^{CART}) = c < \mu_R(s^{CART})$. This scenario can be considered a special case of **i** and **ii**. By appropriate choice of Λ, it's trivial to show that the original s^{CART} remains the global optimizer and the risk stays the same.

□

Mathematical details for Figure 1b and 1c

Imagine there are two final splitting nodes $\{t_1, t_2\}$ with the same mass of population. CART selects features $\{X_1, X_2\}$ for nodes $\{t_1, t_2\}$, respectively. The pdf of X_1 and X_2 in the $\{t_1, t_2\}$ are both uniform, respectively: $f_1(x) = 1$ and $f_2(x) = 1 \ \forall \ x \in [0, 1]$.

In node t_1,

$$P(Y = 1|X_1) =: \eta_1(X_1) = \begin{cases} 1, \text{ when } X_1 \in [0, \frac{1}{4}] \\ \frac{\sin(2\pi X_1)+1}{2}, \text{ when } X_1 \in (\frac{1}{4}, \frac{3}{4}) \\ 0, \text{ when } X_1 \in [\frac{3}{4}, 1] \end{cases}$$

We depict $\eta_1(X_1)$ in Figure 1b. Since $\eta_1(X_1)$ is reflectional symmetric around $\eta_1(0.5)$, $s^{cart} = 0.5$. Splitting s^{cart}, policymakers target the left node $X_1 \leq 0.5$ only, the calculation is similar to Figure 1a in the introduction.

The unique optimal LPC solution is $s^* = \frac{5}{12}$. By shifting from s^{CART} to s^*, the policymaker excludes subpopulation $\{\frac{5}{12} < X_1 < \frac{1}{2} | t_1\}$ from the targeted group. Again, the mathematics is similar to Figure 1a in the introduction.

In node t_2,

$$P(Y = 1|X_2) =: \eta_2(X_2) = \frac{9}{11} X_2$$

The other node t_2 is depicted by Figure 1c. CART will split t_2 at $s^{CART} = 0.5$ because $\eta_2(X_2)$ is reflectional symmetric around $\eta_2(0.5)$. This would not target any subpopulations from t_2, as both the left node mean and right node mean are smaller than 0.75.

$$P(Y = 1|X_2 < 0.5) = \frac{\int_0^{0.5} \frac{9}{11} x dx}{0.5} = \frac{9}{44} < 0.75$$
$$P(Y = 1|X_2 > 0.5) = \frac{\int_{0.5}^{1} \frac{9}{11} x dx}{0.5} = \frac{27}{44} < 0.75$$

On the other hand, the optimal split for LPC is $s^* = \frac{11}{12}$. This splits t_2 into two groups, the left node is entirely below the threshold whereas the right node is entirely above, as illustrated by the orange and green segment in Figure 1c, respectively. As a result, splitting at s^* targets subpopulation $\{\frac{11}{12} < X_2 < 1 | t_2\}$.

To summarize, splitting nodes t_1 and t_2 individually at s^{CART} versus s^* results in different target subpopulations: (i) s^{CART}: Target $\{X_1 < \frac{1}{2} | t_1\}$; (ii) s^*: Target $\{X_1 < \frac{5}{12} | t_1\}$ and $\{\frac{11}{12} < X_2 < 1 | t_2\}$. Assuming that both t_1 and t_2 contain the same amount of population, the two sets of policies target the same proportion of the population, but $\eta(x, t_1) < 0.75$ for $x \in \{\frac{5}{12} < X_1 < \frac{1}{2}\}$, which is targeted by s^{CART}, whereas $\eta(x, t_2) > 0.75$ for $x \in \{\frac{11}{12} < X_2 < 1\}$, which is targeted by s^*. Therefore, policies based on LPC, i.e., s^* policy, target a **more vulnerable** subpopulation than policies based on observed Y classification, i.e., s^{CART} policy or CART/KD-CART policy. This idea appears in our diabetes empirical study, which we detail in Section 6.2.

Proof of Remark 7.1

$$\begin{aligned}
\mathcal{G}^{\text{WRacc}} =& \text{WRacc of left node} + \text{WRacc of right node} \\
=& F(s) \left| \frac{\int_0^s \eta(x) dF(x)}{F(s)} - \int_0^1 \eta(x) dF(x) \right| + (1 - F(s)) \left| \frac{\int_s^1 \eta(x) dF(x)}{1 - F(s)} - \int_0^1 \eta(x) dF(x) \right| \\
=& F(s) \left| \mu_L - \int_0^1 \eta(x) dF(x) \right| + (1 - F(s)) \left| \mu_R - \int_0^1 \eta(x) dF(x) \right| \\
=& s \left| \mu_L - \int_0^1 \eta(x) dx \right| + (1 - s) \left| \mu_R - \int_0^1 \eta(x) dx \right| \quad \text{Under Assumption 4.1}
\end{aligned}$$

C Synthetic data simulation studies

C.1 Data generation processes

1. Generate features: $X_i \sim U(0, 1), i \in \{1, 2, 3, 4, 5\}$.
2. - **Ball:** $f(X) = \sum_{i=1}^{3} X_i^2$.
   - **Friedman #1:** $f(X) = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$
   - **Friedman #2:**
     - Make transformations: $Z_1 = 100X_1, Z_2 = 40\pi + 520\pi X_2, Z_4 = 10X_4 + 1$
     - Generate responses: $f(X) = \sqrt{Z_1^2 + (Z_2 X_3 - \frac{1}{Z_2 Z_4})^2}$.
   - **Friedman #3:**
     - Make transformations: $Z_1 = 100X_1, Z_2 = 40\pi + 520\pi X_2, Z_4 = 10X_4 + 1$
     - Generate responses: $f(X) = \arctan\left((Z_2 X_3 - \frac{1}{Z_2 Z_4})/Z_1\right)$.
   - **Poly #1:** $f(X) = 4X_1 + 3X_2^2 + 2X_3^3 + X_4^4$

- **Poly #2:** $f(X) = X_1^4 + 2X_2^3 + 3X_3^2 + 4X_1$.
- **Ring:** $f(X) = |\sum_{i=1}^{3} X_i^2 - 1|$.
- **Collinear:**
  - Create correlated features: $X_{i+3} = X_i + 0.1 \cdot \mathcal{N}(0, 1)$, where $i \in \{1, 2, 3\}$.
  - Generate responses: $f(X) = \sum_{i=1}^{6} X_i$.

3. Map it to probabilities: $\eta = \text{Sigmoid}\left(f(X) - E(f(X))\right)$.
4. Generate labels: $y \sim \text{Bernoulli}(\eta)$.

C.2 Full algorithms

We provide pseudo codes for CART, PFS, MDFS, wEFS, RF-CART, and RF-MDFS. In Algorithm 1, setting 'method' to CART, PFS, MDFS, or wEFS invokes the corresponding procedure. Likewise, in Algorithm 2, choosing method = CART or MDFS produces RF-CART and RF-MDFS, respectively. The rest of the algorithms are all helper functions used in Algorithms 1 and 2.

Algorithm 1 Grow_Tree

Fit the tree-based model to $(X, \boldsymbol{y})$ when η or $\hat{\eta}$ is not provided.

Input: $X, \boldsymbol{y}, c$, current_depth, method, depth, min_samples
if current_depth = depth **or** $n(\text{unique}(\boldsymbol{y})) = 1$ **or** $n(\boldsymbol{y})$ <min_samples **then**
  Output: mean($\boldsymbol{y}$)
end if
best_feature, best_split $\leftarrow$ FindBestSplit$(X, \boldsymbol{y})$
mask $\leftarrow (\boldsymbol{x}_{\text{best feature}} < \text{best\_split})$
left_X, right_X $\leftarrow X[\text{mask}], X[!\text{mask}]$
left_y, right_y $\leftarrow \boldsymbol{y}[\text{mask}], \boldsymbol{y}[!\text{mask}]$
if current_depth = depth - 1 **or** $\min\{n(\text{left\_y}), n(\text{right\_y})|\}$ <min_samples **or** method!='CART' **then**
  if method = 'MDFS' **then**
    new_best_split $\leftarrow$ Find_PFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{y}, c, 1)$
  else if method = 'PFS' **then**
    new_best_split $\leftarrow$ Find_PFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{y}, c, 0.1)$
  else if method = 'wEFS' **then**
    new_best_split $\leftarrow$ Find_wEFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{y}, c)$
  end if
  mask $\leftarrow (\boldsymbol{x}_{\text{best\_feature}} < \text{new\_best\_split})$
  left_tree, right_tree $\leftarrow$ mean($\boldsymbol{y}$[mask]), mean($\boldsymbol{y}$[!mask])
  Output: best_feature, new_best_split, left_tree, right_tree
else
  left_tree = Grow_Tree(left_X, left_y, current_depth + 1, method, depth, min_samples)
  right_tree = Grow_Tree(right_X, right_y, current_depth + 1, method, depth, min_samples)
  Output: best_feature, best_split, left_tree, right_tree
end if

Algorithm 2 Grow_Tree_wP

Fit the tree-based model to $(X, \boldsymbol{y})$ and $\hat{\eta}$ provided by Knowledge-Distillation.

Input: $X, \boldsymbol{y}, \boldsymbol{p}, c$, current_depth, method, depth, min_samples
if current_depth = depth **or** $(\min(\boldsymbol{p}) > c$ **or** $\max(\boldsymbol{p}) < c)$ **or** $n(\boldsymbol{y})$ <min_samples **then**
  Output: mean($\boldsymbol{p}$)
end if
best_feature, best_split $\leftarrow$ FindBestSplit$(X, \boldsymbol{p})$
mask $\leftarrow (\boldsymbol{x}_{\text{best feature}} < \text{best\_split})$
left_X, right_X $\leftarrow X[\text{mask}], X[!\text{mask}]$
left_y, right_y $\leftarrow \boldsymbol{y}[\text{mask}], \boldsymbol{y}[!\text{mask}]$
left_p, right_p $\leftarrow \boldsymbol{p}[\text{mask}], \boldsymbol{p}[!\text{mask}]$
if current_depth = depth - 1 **or** $\min\{n(\text{left\_p}), n(\text{right\_p})|\}$ <min_samples **or** method!='CART' **then**
  if method = 'MDFS' **then**
    new_best_split $\leftarrow$ Find_PFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{y}, c, 1)$
  else if method = 'PFS' **then**
    new_best_split $\leftarrow$ Find_PFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{y}, c, 0.1)$
  else if method = 'wEFS' **then**
    new_best_split $\leftarrow$ Find_wEFS_BestSplit$(\boldsymbol{x}_{\text{best_feature}}, \boldsymbol{p}, c)$
  end if
  mask $\leftarrow (\boldsymbol{x}_{\text{best\_feature}} < \text{new\_best\_split})$
  left_tree, right_tree $\leftarrow$ mean($\boldsymbol{p}$[mask]), mean($\boldsymbol{p}$[!mask])
  Output: best_feature, new_best_split, left_tree, right_tree
else
  left_tree = Grow_Tree_P(left_X, left_y, left_p, current_depth + 1, method, depth, min_samples)
  right_tree = Grow_Tree_P(right_X, right_y, right_p, current_depth + 1, method, depth, min_samples)
  Output: best_feature, best_split, left_tree, right_tree

end if

Algorithm 3 Calculate_impurity

Calculate sample impurity $\hat{\mathcal{G}}^{CART}$.

Input: $\boldsymbol{x}, \boldsymbol{y}, x$
$\boldsymbol{y}_l \leftarrow \boldsymbol{y}[\boldsymbol{x} \leq x], \boldsymbol{y}_r \leftarrow \boldsymbol{y}[\boldsymbol{x} > x]$
$\mathcal{G} \leftarrow \mathrm{Var}(\boldsymbol{y}_l)\frac{|\boldsymbol{y}_l|}{n} + \mathrm{Var}(\boldsymbol{y}_r)\frac{|\boldsymbol{y}_r|}{n}$
Output: $\mathcal{G}$

Algorithm 4 Calculate_distance

Calculate the second term in (1) and $\hat{\mathcal{G}}^*$.

Input: $\boldsymbol{x}, \boldsymbol{y}, x, c$
$\boldsymbol{y}_l \leftarrow \boldsymbol{y}[\boldsymbol{x} \leq x], \boldsymbol{y}_r \leftarrow \boldsymbol{y}[\boldsymbol{x} > x]$
$\mathcal{G} \leftarrow (1 - |c - \bar{y_l}|)\frac{|\boldsymbol{y}_l|}{n} + (1 - |c - \bar{y_r}|)\frac{|\boldsymbol{y}_r|}{n}$
Output: $\mathcal{G}$

Algorithm 5 Calculate_weighted_risk

Calculate sample loss function of wEFS.

Input:$\boldsymbol{x}, \boldsymbol{y}, x, c$
$\boldsymbol{y}_l \leftarrow \boldsymbol{y}[\boldsymbol{x} \leq x], \boldsymbol{y}_r \leftarrow \boldsymbol{y}[\boldsymbol{x} > x]$
$r_l \leftarrow |\boldsymbol{y}_l[\boldsymbol{y}_l > c]|, r_r \leftarrow |\boldsymbol{y}_r[\boldsymbol{y}_r > c]|$
$\mathcal{R} \leftarrow (\mathbb{1}\{\bar{y}_l > c\}(|\boldsymbol{y}_l| - r_l) + \mathbb{1}\{\bar{y}_r > c\}(|\boldsymbol{y}_r| - r_r))c + (\mathbb{1}\{\bar{y}_l \leq c\}r_l + \mathbb{1}\{\bar{y}_r \leq c\}r_r)(1 - c)$
Output: $\mathcal{R}$

Algorithm 6 FindBestSplit

Take in data from the parent node and outputs the best feature, split point in terms of $\mathcal{G}^{CART}$.

Input:$X, \boldsymbol{y}$
best_impurity, best_split, best_feature $\leftarrow$ 0, None, None
for each feature $\boldsymbol{x}_i$ in X, i from 1 to p **do**
  sort X in terms of $\boldsymbol{x}_i$
  for x in ordered samples in $\boldsymbol{x}_i$ **do**
    impurity $\leftarrow$ Calculate_impurity($\boldsymbol{x}_i, \boldsymbol{y}, x$)
    if impurity $<$ best_impurity **then**
      best_impurity $\leftarrow$ impurity
      best_split $\leftarrow$ x
      best_feature $\leftarrow$ i
    end if
  end for
end for
Output:best_split, best_feature

Algorithm 7 Find_PFS_BestSplit

Take in data from the parent node and outputs the best feature, split point in terms of $\mathcal{G}^{PFS}(\lambda)$.

Input: $\boldsymbol{x}, \boldsymbol{y}, c, \lambda$
best_G_PFS, best_split $\leftarrow$ 0, None
sort $\boldsymbol{x}$
for x in sorted $\boldsymbol{x}$ **do**
  G_PFS $\leftarrow (1-\lambda)$ Calculate_impurity$(\boldsymbol{x}, \boldsymbol{y}, x) + \lambda$ Calculate_distance$(\boldsymbol{x}, \boldsymbol{y}, x, c)$
  if G_PFS $<$ best_G_PFS **then**
    best_G_PFS $\leftarrow$ G_PFS
    best_split $\leftarrow x$
  end if
end for
Output: best_split

Algorithm 8 Find_wEFS_BestSplit

Take in data from the parent node and outputs the best feature, split point in terms of the weighted empirical risk.

Input: $\boldsymbol{x}, \boldsymbol{y}, c$
best_risk, best_split $\leftarrow$ 0, None
sort $\boldsymbol{x}$
for x in sorted $\boldsymbol{x}$ **do**
  risk $\leftarrow$ Calculate_weighted_risk$(\boldsymbol{x}, \boldsymbol{y}, x, c)$
  if risk $<$ best_risk **then**
    best_risk $\leftarrow$ risk
    best_split $\leftarrow x$
  end if
end for
Output: best_split

How to implement KD-CART/KD-MDFS: We first use *RandomForestClassifier* from *sklearn* (Pedregosa et al., 2011) to train a random forest with the entire sample. All user-defined parameters of the random forest follow the default settings. Then, we grow an RF-CART and an RF-MDFS tree with the predicted probability output by the random forest as the response variable.

C.3 Choice of hyperparameter λ

Our simulations regarding PFS choose a small λ value of 0.1, and overall, this choice leads to lower misclassification error than CART. To identify an appropriate λ in practice, we had a standard procedure: we select the λ value using an approach that combines the idea of cross-validation and the honest approach (Athey and Imbens, 2016):

1. Divide the data into $K = 5$ folds. For a fixed candidate λ value, by leaving out fold $k \in \{1, 2, 3, 4, 5\}$, we can construct 5 trees, T^{-k}. Denote T^{-k}'s estimate for X as $T^{-k}(X)$.
2. Obtain an unbiased estimate of η for each node by feeding the left-out fold to T^{-k}. Denote this estimate for X as $\hat{\eta}^k(X)$.
3. Calculate misclassification score (MS) (i.e., *honest* approach),

$$\text{MS} = \sum_{k=1}^{5} \sum_{i \in \text{fold } k} \mathbf{1}\left\{(T^{-k}(X_i) - c)(\hat{\eta}^k(X_i) - c) < 0\right\}.$$

4. Choose λ value that minimizes MS.

We implemented this approach in simulation for PFS, the results are comparable to setting $\lambda = 0.1$.

C.4 Additional simulation results

Refer to Figure 5, Table 1, and Table 2. Figure 5 shows the F1 score comparison across all six methods under the configuration of $m = 7$, $\rho = 2\%$, and $c = 0.5$. Table 1 and Table 2 show all six methods' MR and F1 scores, respectively, across all tasks. For each task, only the best configuration is included in the two tables.

We further include a comparison of the tree-based classification algorithms and other typical baselines: logistic regression and policytree, a leading policy learning algorithm. The simulation setup is identifical to Figure 2 and Figure 5. The MR and F1 scores are summarized in Table 3 and Table 4, respectively. Our methods outperform logistic regression and policytree by a very large margin.

C.5 Computational complexity analysis

Since our proposed algorithm only operates at the last split, the tree-building processes are identical in upper levels among 4 strategies. Next, we evaluate the computational complexity of the **final split**, i.e., split at the final level.

Suppose the sample size in a node at the second to last level is n and the number of features is p. In **CART**, the dominant step is to sort the samples based on each feature. With p features to consider, the computational complexity is $O(pNlogN)$. **PFS**, **MDFS**, and **wEFS** all determine the feature to split on using **CART**. The split point decision requires another $O(N)$ of time which is still dominated by the sorting process.

D Additional empirical study details

We use publicly available data from our empirical studies. The Pima Indian Diabetes dataset can be downloaded at `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`. The Montesinho Park forest fire dataset can be downloaded at `https://archive.ics.uci.edu/dataset/162/forest+fires`.

In this section, we use forest fire dataset to illustrate an application of our methods MDFS and RF-MDFS, and compare them to CART and RF-CART respectively.

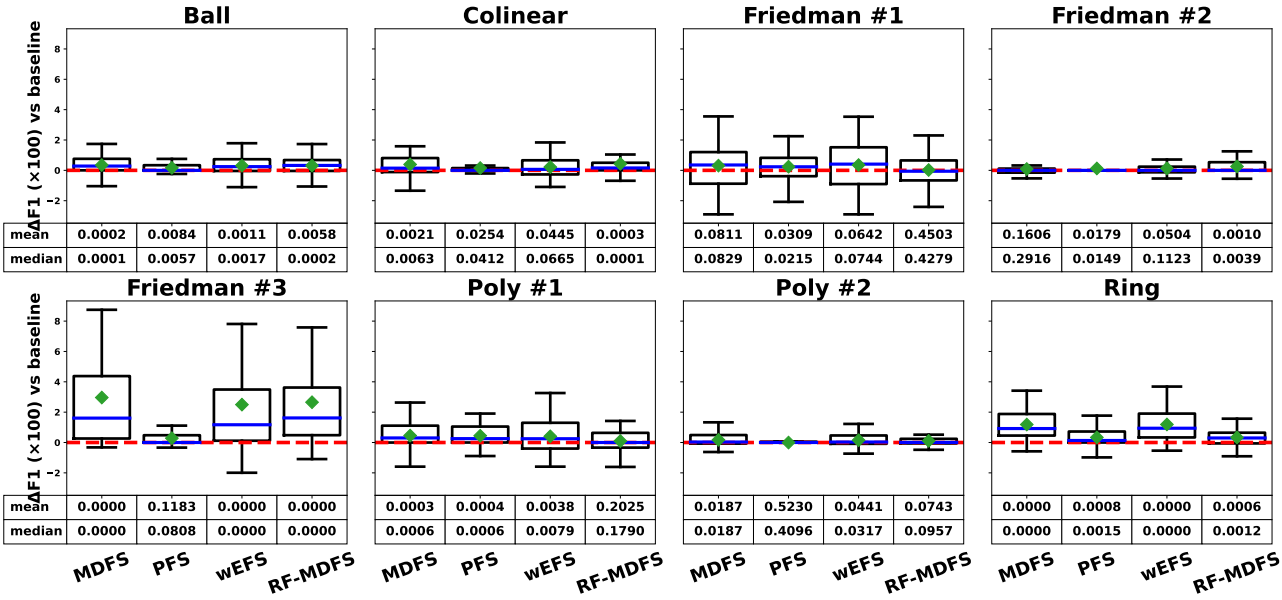


Figure 5: Boxplots of F1 differences relative to baseline models. For each panel, the first three boxplots compare CART with PFS, MDFS, and wEFS, and the last boxplot compares RF-CART with RF-MDFS. Positive values in the boxplots denote *improvement* in F1. Embedded tables list one-sided paired t-test and Wilcoxon signed-rank p-values for mean and median F1 differences, respectively.

The UCI Forest Fire dataset measures the characteristics of different forest fires in Montesinho Park with the response variable being the area burnt by a forest fire. We binarize the response by labeling those observations with a burnt area larger than 5 as 1, and 0 otherwise. 1 indicates a "big" forest fire. We search for conditions under which the probability of having a big forest fire is above 1/3. We use the following features in the dataset for this task: X and Y-axis spatial coordinates, month, FFMC, DMC, DC, ISI, temperature, RH (relative humidity), wind, and rain. All the acronyms except for RH are measures that positively correlate with the probability of a big forest fire. With a moderate sample size of 517, we set $m = 3$.

D.1 CART vs MDFS

As shown by Figure 3, the two sets of policies target many common groups except for the subgroup consisting of 43 observations defined by $FFMC > 85.85$, $temp \leq 26.0$, $DC > 766.2$. This group that MDFS additionally targets has a 37.2% probability of catching a big forest fire, higher than the threshold. This finding aligns with Remark 4.7.

D.2 RF-CART vs RF-MDFS

As shown by Figure 3, RF-MDFS also additionally targets the subgroup consisting of 43 observations defined by $FFMC > 85.85$, $temp \leq 26.0$, $DC > 766.2$. Hence, the comparison between RF-CART and RF-MDFS here also aligns with Remark 4.7.

Interestingly, both RF-CART and RF-MDFS (compared to CART and MDFS respectively) additionally target the subgroup $FFMC > 85.85$, $temp > 26.0$, $RH \leq 24.5$.

D.3 Full CART/MDFS/RF-CART/MDFS for the diabetes example

CART

```
if FFMC <= 85.85
    if temp <= 5.15
        if X <= 5.0
            value: 1.000, samples: 8
            value: 0.600, samples: 5
        if temp <= 18.55
            value: 0.172, samples: 29
            value: 0.667, samples: 9
    if temp <= 26.0
        if DC <= 673.2
            value: 0.208, samples: 216
            value: 0.310, samples: 200
        if RH <= 24.5
            value: 0.125, samples: 8
            value: 0.500, samples: 42
```

MDFS

```
if FFMC <= 85.85
    if temp <= 5.15
        if X <= 3.5
            value: 1.000, samples: 2
            value: 0.818, samples: 11
        if temp <= 18.55
            value: 0.172, samples: 29
            value: 0.667, samples: 9
    if temp <= 26.0
        if DC <= 766.2
            value: 0.244 samples: 216
            value: 0.372, samples: 43
        if RH <= 24.5
            value: 0.125, samples: 8
            value: 0.500, samples: 42
```

RF-CART

```
if FFMC <= 85.85
    if temp <= 5.15
        if wind <= 7.15
            value: 0.638, samples: 6
            value: 0.964, samples: 7
        if temp <= 18.55
            value: 0.229, samples: 29
            value: 0.594, samples: 9
    if temp <= 26.0
        if DC <= 673.2
            value: 0.220 samples: 216
            value: 0.308, samples: 200
        if FFMC <= 95.85
            value: 0.381, samples: 43
            value: 0.649, samples: 7
```

RF-MDFS

```
if FFMC <= 85.85
    if temp <= 5.15

        value: 0.814, samples: 13

        if temp <= 18.55
            value: 0.229, samples: 29
            value: 0.594, samples: 9
    if temp <= 26.0
        if DC <= 766.2
            value: 0.250 samples: 373
            value: 0.368, samples: 43
        if FFMC <= 91.95
            value: 0.578, samples: 4
            value: 0.404, samples: 46
```

Figure 6: The targeting policies generated by CART, MDFS, RF-CART, RF-MDFS. The red groups are the targeted subpopulations predicted to have a higher than 1/3 probability of catching a big forest fire.

DGP	c	CART	PFS	MDFS	WRACC	wEFS	RF-CART	RF-MDFS
Ball	0.8	9.7 (1.8)	9.2 (1.7)	9.2 (1.5)	8.9 (1.4)	8.8 (1.3)	8.8 (1.8)	**8.5** (1.2)
	0.7	13.8 (2.6)	13.7 (2.3)	13.3 (2.2)	13.8 (2.7)	14.0 (2.5)	12.8 (2.1)	**12.6** (2.0)
	0.6	14.4 (2.2)	14.0 (1.8)	13.8 (1.8)	14.2 (2.2)	13.9 (1.7)	13.6 (1.4)	**13.0** (1.4)
	0.5	11.3 (1.8)	11.0 (1.9)	10.7 (1.7)	10.7 (1.6)	10.7 (1.6)	10.5 (1.7)	**10.2** (1.6)
Friedman #1	0.8	6.7 (1.3)	6.5 (1.4)	6.4 (1.4)	6.2 (1.2)	6.0 (1.1)	**6.0** (1.0)	6.3 (1.0)
	0.7	12.5 (1.6)	12.2 (1.7)	12.2 (1.3)	12.6 (1.4)	12.6 (1.7)	11.5 (1.2)	**11.5** (1.1)
	0.6	18.5 (1.7)	18.1 (1.6)	18.0 (1.4)	18.4 (1.6)	18.5 (1.6)	17.6 (1.4)	**17.2** (1.6)
	0.5	20.4 (2.0)	20.3 (1.5)	20.2 (1.4)	20.3 (1.4)	20.1 (1.4)	**19.3** (1.3)	19.8 (1.3)
Friedman #2	0.8	8.6 (2.0)	7.8 (1.4)	**7.2** (1.3)	8.2 (1.9)	8.2 (1.9)	8.4 (1.7)	7.3 (1.4)
	0.7	8.0 (1.7)	7.6 (1.6)	7.4 (1.4)	7.8 (1.6)	7.9 (1.5)	7.8 (2.0)	**7.1** (1.6)
	0.6	6.9 (1.7)	6.6 (1.4)	6.4 (1.5)	6.8 (1.6)	6.4 (1.4)	6.5 (1.8)	**6.0** (1.5)
	0.5	5.2 (1.5)	5.0 (1.4)	5.0 (1.2)	5.0 (1.3)	5.0 (1.3)	5.0 (1.5)	**4.6** (1.0)
Friedman #3	0.8	1.9 (0.4)	2.0 (0.4)	1.8 (0.4)	2.2 (0.5)	2.1 (0.5)	1.8 (0.4)	**1.8** (0.4)
	0.7	2.6 (0.6)	2.5 (0.5)	**2.3** (0.4)	2.9 (0.7)	2.5 (0.5)	2.4 (0.5)	2.4 (0.4)
	0.6	3.2 (0.7)	3.2 (0.6)	2.9 (0.6)	3.5 (0.6)	**2.9** (0.5)	3.1 (0.5)	3.0 (0.5)
	0.5	4.0 (0.7)	3.9 (0.7)	3.6 (0.5)	3.6 (0.5)	3.6 (0.5)	3.8 (0.7)	**3.6** (0.5)
Poly #1	0.8	7.3 (1.4)	7.2 (1.3)	7.3 (1.3)	7.4 (1.3)	7.4 (1.2)	7.1 (1.5)	**6.8** (1.5)
	0.7	10.6 (2.0)	10.4 (2.0)	10.1 (1.7)	10.4 (1.9)	10.6 (2.1)	9.8 (1.9)	**9.3** (1.6)
	0.6	12.5 (1.8)	12.4 (2.8)	12.2 (2.6)	12.9 (2.5)	12.9 (2.2)	**11.8** (2.0)	11.8 (2.2)
	0.5	13.0 (1.9)	12.3 (2.0)	12.3 (1.9)	12.3 (1.7)	12.3 (1.7)	12.2 (2.1)	**12.0** (1.7)
Poly #2	0.8	10.0 (2.1)	9.7 (1.9)	9.7 (1.9)	9.9 (2.1)	9.9 (2.1)	9.3 (2.1)	**9.0** (2.0)
	0.7	9.7 (2.1)	9.5 (2.0)	9.4 (1.9)	9.7 (2.0)	9.9 (1.9)	9.5 (2.0)	**9.3** (1.8)
	0.6	8.2 (1.6)	8.1 (1.6)	7.9 (1.6)	8.1 (1.9)	8.0 (1.8)	7.8 (1.4)	**7.6** (1.3)
	0.5	6.9 (1.5)	6.9 (1.5)	6.7 (1.3)	6.6 (1.4)	6.6 (1.4)	6.3 (1.6)	**6.2** (1.4)
Ring	0.8	17.1 (2.0)	17.0 (2.0)	16.6 (2.0)	17.4 (1.8)	17.5 (1.8)	16.1 (1.7)	**15.9** (1.6)
	0.7	14.5 (1.7)	14.4 (1.4)	14.3 (1.5)	15.1 (1.4)	14.9 (1.4)	13.7 (1.3)	**13.3** (1.2)
	0.6	12.7 (1.4)	12.5 (1.4)	11.9 (1.3)	12.6 (1.4)	12.2 (1.3)	11.8 (1.2)	**11.4** (1.1)
	0.5	10.7 (1.6)	10.4 (1.5)	9.8 (0.9)	9.8 (0.9)	9.8 (0.9)	10.0 (1.2)	**9.5** (1.1)
Colinear	0.8	6.5 (1.4)	6.3 (1.3)	5.8 (1.2)	6.7 (1.3)	6.6 (1.3)	6.2 (1.2)	**5.8** (1.1)
	0.7	7.8 (1.4)	7.7 (1.4)	7.3 (1.1)	8.1 (1.4)	8.2 (1.4)	7.3 (1.6)	**7.0** (1.2)
	0.6	8.3 (1.4)	8.2 (1.3)	8.0 (1.2)	9.1 (1.6)	8.6 (1.6)	8.2 (1.5)	**7.7** (1.2)
	0.5	8.5 (1.6)	8.4 (1.5)	8.1 (1.5)	8.2 (1.4)	8.2 (1.4)	8.2 (1.4)	**7.7** (1.2)

Table 1: MR comparison of misclassification rate among all six methods. CART, PFS, MDFS, and wEFS use raw data input; CART and MDFS use random forest as a KD tool. Each entry presents the average and standard error (in parentheses) of the misclassification rate over 50 replicates. Note that the standard error is mostly from the data generation randomness, not the variability of the methods. To check for statistical significance, one should check out the t-test and rank sum test p-values in Figure 2. The complete simulation results for the simulation setups presented in Table 1 are provided in the supplemental files.

DGP	c	CART	PFS	MDFS	WRACC	wEFS	RF-CART	RF-MDFS
Ball	0.2	78.9 (4.0)	79.5 (3.4)	80.0 (3.4)	80.1 (3.3)	80.3 (3.3)	80.4 (3.4)	**80.8** (3.6)
	0.3	83.7 (2.6)	83.8 (2.6)	84.1 (2.6)	83.6 (2.9)	83.4 (2.8)	84.8 (2.6)	**85.1** (2.4)
	0.4	88.4 (1.9)	88.7 (1.7)	88.9 (1.1)	88.7 (2.0)	89.1 (1.4)	89.0 (1.3)	**89.6** (1.2)
	0.5	92.8 (1.3)	93.0 (1.3)	93.2 (1.2)	93.1 (1.1)	93.1 (1.1)	93.3 (1.2)	**93.5** (1.1)
Friedman #1	0.2	45.6 (5.7)	46.2 (5.7)	46.7 (5.4)	43.1 (6.1)	45.9 (6.0)	**49.4** (5.4)	48.4 (5.4)
	0.3	61.8 (3.7)	62.5 (4.8)	63.0 (2.8)	61.5 (3.0)	62.3 (3.0)	63.3 (3.8)	**63.8** (3.1)
	0.4	71.4 (2.5)	71.7 (2.7)	71.7 (2.4)	71.7 (2.4)	72.5 (2.2)	72.6 (2.3)	**73.0** (2.4)
	0.5	80.6 (2.1)	80.6 (1.7)	80.8 (1.7)	80.6 (1.7)	80.8 (1.7)	**81.6** (1.7)	81.2 (1.6)
Friedman #2	0.2	91.0 (2.3)	91.8 (1.5)	92.5 (1.4)	91.5 (1.9)	91.5 (1.9)	91.2 (1.7)	**92.4** (1.4)
	0.3	93.7 (1.4)	94.0 (1.3)	94.2 (1.1)	93.8 (1.3)	93.7 (1.3)	93.8 (1.7)	**94.3** (1.3)
	0.4	95.3 (1.2)	95.5 (1.0)	95.7 (1.0)	95.5 (1.1)	95.8 (0.9)	95.6 (1.3)	**96.0** (1.1)
	0.5	96.9 (0.9)	97.0 (0.9)	96.9 (0.7)	97.0 (0.8)	97.0 (0.8)	96.9 (0.9)	**97.2** (0.6)
Friedman #3	0.2	63.1 (7.7)	61.8 (13.4)	**64.7** (9.5)	50.2 (24.4)	59.2 (15.6)	63.8 (10.6)	62.1 (13.6)
	0.3	69.3 (5.0)	70.2 (5.2)	**71.6** (5.5)	66.3 (10.8)	71.4 (5.2)	70.1 (5.5)	70.7 (6.3)
	0.4	73.6 (4.4)	74.2 (4.3)	**75.9** (4.1)	73.6 (3.7)	75.7 (3.8)	74.3 (3.3)	75.1 (3.8)
	0.5	76.2 (3.3)	76.5 (3.0)	77.7 (3.0)	77.2 (3.2)	77.2 (3.2)	76.9 (3.2)	**77.7** (2.6)
Poly #1	0.2	74.7 (3.9)	75.9 (4.5)	76.3 (3.8)	75.2 (6.6)	75.5 (4.5)	75.9 (5.2)	**76.6** (5.4)
	0.3	81.9 (3.3)	82.4 (3.4)	82.7 (2.9)	82.2 (3.2)	82.0 (3.3)	83.0 (3.5)	**83.9** (2.7)
	0.4	86.7 (1.9)	86.8 (2.9)	87.1 (2.5)	86.8 (2.5)	86.9 (2.0)	87.3 (2.3)	**87.5** (2.2)
	0.5	90.0 (1.5)	90.5 (1.6)	90.5 (1.5)	90.5 (1.4)	90.5 (1.4)	90.6 (1.7)	**90.7** (1.4)
Poly #2	0.2	90.3 (2.1)	90.4 (1.9)	90.4 (1.9)	90.2 (2.2)	90.2 (2.1)	90.8 (2.0)	**91.1** (1.8)
	0.3	92.7 (1.6)	92.9 (1.5)	92.9 (1.4)	92.7 (1.4)	92.6 (1.4)	92.8 (1.5)	**93.0** (1.3)
	0.4	94.9 (1.0)	94.9 (1.0)	95.0 (1.0)	95.0 (1.2)	95.0 (1.1)	95.1 (0.9)	**95.2** (0.9)
	0.5	96.1 (0.9)	96.1 (0.9)	96.2 (0.8)	96.2 (0.8)	96.2 (0.8)	96.4 (0.9)	**96.5** (0.8)
Ring	0.2	70.2 (3.1)	71.0 (3.4)	71.4 (3.2)	69.8 (2.9)	70.2 (2.9)	71.9 (2.9)	**72.4** (2.7)
	0.3	82.5 (2.0)	82.6 (1.5)	82.8 (1.7)	82.3 (1.5)	82.6 (1.4)	83.4 (1.6)	**83.8** (1.4)
	0.4	88.1 (1.3)	88.3 (1.3)	89.0 (1.2)	88.6 (1.2)	88.8 (1.1)	89.0 (1.2)	**89.4** (1.1)
	0.5	91.7 (1.3)	92.0 (1.2)	92.5 (0.7)	92.5 (0.7)	92.5 (0.7)	92.3 (0.9)	**92.7** (0.8)
Colinear	0.2	81.3 (3.8)	81.5 (3.6)	82.9 (3.3)	80.6 (3.4)	80.8 (3.3)	81.8 (3.5)	**83.3** (3.0)
	0.3	85.9 (2.5)	86.3 (2.5)	86.8 (2.2)	85.6 (2.6)	85.5 (2.5)	86.9 (2.8)	**87.3** (2.2)
	0.4	89.2 (1.9)	89.4 (1.8)	89.7 (1.6)	88.8 (1.9)	89.3 (1.8)	89.6 (1.7)	**90.2** (1.6)
	0.5	91.5 (1.7)	91.6 (1.7)	91.9 (1.5)	91.7 (1.5)	91.7 (1.5)	91.8 (1.5)	**92.2** (1.2)

Table 2: F1 comparison of misclassification rate among all six methods. CART, PFS, MDFS, and wEFS use raw data input; CART and MDFS use random forest as a KD tool. Each entry presents the average and standard error (in parentheses) of the misclassification rate over 50 replicates. Note that the standard error is mostly from the data generation randomness, not the variability of the methods. To check for statistical significance, one should check out the t-test and rank sum test p-values in Figure 5. The complete simulation results for the simulation setups presented in Table 2 are provided in the supplemental files.

Table 3: Misclassification Rate (MR) – Lower is Better

Method	ball	collinear	friedman1	friedman2	friedman3	poly1	poly2	ring
logistics	0.307	0.274	0.372	0.221	0.136	0.334	0.221	0.279
hybrid policy tree	0.203	0.175	0.276	0.142	0.090	0.229	0.137	0.189
classic_cart	0.113	0.085	0.204	0.052	0.040	0.130	0.069	0.107
mdfs	0.107	0.081	0.202	0.050	0.036	0.123	0.067	0.098
pfs	0.110	0.084	0.203	0.050	0.039	0.123	0.069	0.104
wefs	0.107	0.082	0.201	0.050	0.036	0.123	0.066	0.098

Table 4: F1 Score – Higher is Better

Method	ball	collinear	friedman1	friedman2	friedman3	poly1	poly2	ring
logistics	0.785	0.725	0.644	0.857	0.388	0.729	0.866	0.778
hybrid policy tree	0.866	0.823	0.733	0.912	0.555	0.820	0.920	0.850
classic_cart	0.928	0.915	0.806	0.969	0.762	0.900	0.961	0.917
mdfs	0.932	0.919	0.807	0.969	0.777	0.905	0.962	0.925
pfs	0.930	0.916	0.806	0.970	0.760	0.905	0.961	0.920
wefs	0.931	0.917	0.808	0.970	0.772	0.905	0.962	0.925

CART

```
if Glucose <= 127.5
    if Age <= 28.5
        if BMI <= 30.95
            value: 0.013, samples: 151
            value: 0.175, samples: 120
        if BMI <= 26.35
            value: 0.049, samples: 41
            value: 0.399, samples: 173
    if BMI <= 29.95
        if Glucose <= 145.5
            value: 0.146, samples: 41
            value: 0.514, samples: 35
        if Glucose <= 157.5
            value: 0.609, samples: 115
            value: 0.870, samples: 92
```

MDFS

```
if Glucose <= 127.5
    if Age <= 28.5
        if BMI <= 22.25
            value: 0.000, samples: 35
            value: 0.097, samples: 236
        if BMI <= 28.25
            value: 0.222, samples: 63
            value: 0.377, samples: 151
    if BMI <= 29.95
        if Glucose <= 166.5
            value: 0.250, samples: 64
            value: 0.667, samples: 12
        if Glucose <= 129.5
            value: 0.579, samples: 19
            value: 0.739, samples: 188
```

RF-CART

```
if Glucose <= 127.5
    if Age <= 28.5
        if BMI <= 30.95
            value: 0.028, samples: 151
            value: 0.186, samples: 120
        if BMI <= 26.95
            value: 0.104, samples: 45
            value: 0.396, samples: 169
    if BMI <= 29.95
        if Glucose <= 145.5
            value: 0.193, samples: 41
            value: 0.511, samples: 35
        if Glucose <= 157.5
           value: 0.595, samples: 115
            value: 0.830, samples: 92
```

RF-MDFS

```
if Glucose <= 127.5
    if Age <= 28.5
        if BMI <= 22.25
            value: 0.010, samples: 35
            value: 0.111, samples: 236
        if BMI <= 28.25
            value: 0.225, samples: 63
            value: 0.380, samples: 151
    if BMI <= 29.95
        if Glucose <= 166.5
            value: 0.284, samples: 64
            value: 0.636, samples: 12
        if Glucose <= 129.5
           value: 0.573, samples: 19
            value: 0.713, samples: 188
```

Figure 7: The targeting policies generated by CART, MDFS, RF-CART, RF-MDFS. The red groups are the targeted subpopulations predicted to a higher than 60% probability of being diabetic.