Model Multiplicity and Predictive Arbitrariness in Recidivism Risk Assessment

Ashwin Singh[1], **Carlos Castillo**[2]

[1]TU Wien, Vienna, Austria
[2]ICREA and Universitat Pompeu Fabra, Barcelona, Spain
ashwin.singh@tuwien.ac.at, chato@icrea.cat

Abstract

Prediction tasks over individual futures, which are inherently noisy, often admit multiple similarly accurate models. When these models produce different predictions for the same individual, they raise concerns of arbitrariness in decision-making. How severe can this arbitrariness be, in theory and in practice? How can it be resolved to support high-stakes risk assessment? We address these questions through a study of a machine learning-based decision support system for recidivism risk assessment that has been in use for over 15 years. By translating complex legal rules into an algorithm for labeling post release outcomes (*recidivist* or *non-recidivist*), we first construct a dataset of thousands of inmate releases. Using this dataset, we learn interpretable models that improve predictive performance, reduce error-rate disparities between groups, and ensure that rehabilitative progress lowers risk scores. Next, we study predictive multiplicity, by first deriving a tight lower bound on the expected predictive agreement of any finite set of models over a dataset, and then by evaluating the extent to which structural diversity (e.g., different model coefficients) within this set translates to predictive multiplicity (i.e., different predictions for the same individual). Our experiments indicate that the existence of many similarly accurate models with comparable error-rate disparities does not necessarily translate into severe predictive multiplicity. Empirically, similarly performant models can exhibit substantially higher predictive agreement than worst-case theoretical guarantees suggest. We find that a simple policy that assigns each inmate the lowest risk among these models is effective for addressing predictive arbitrariness.

1 Introduction

The use of machine learning to support risk assessments of criminal recidivism is one of the most well-studied applications in algorithmic fairness research. Increased interest on this topic over the past decade can be traced back to a study of the `COMPAS` algorithm that has been a subject of considerable debate and research (Angwin et al. 2016; Rudin, Wang, and Coker 2020; Jackson and Mendoza 2020; Bao et al. 2021). Since the publication of these studies, recidivism prediction has arguably taken the role of a "model organism" for research in this area, similarly to well-studied species that are selected for intensive examination in biology research, or to platforms that have taken that role in social computing research (Tufekci 2014).

Considering the high-stakes nature of recidivism risk assessment, the European Union's AI Act (2024) categorizes it as a high-risk application of AI. To comply with the AI Act, high-risk AI systems must, among other requirements, demonstrate an appropriate level of accuracy, robustness, and resilience to inconsistencies.

In practice, these criteria are difficult to satisfy. Obtaining **reliable training data** requires determining if people released in the past recidivated, which is legally complex. For instance, due to long judicial processes and sentencing delays, individuals sometimes serve a prison sentence only to be re-incarcerated later for a crime committed before their first term. Hence, returning to prison does not always indicate recidivism. This means labeling needs to be done either manually (delaying the extraction of up-to-date training data) or by applying a complex set of rules (Karimi-Haghighi 2022). While some systems circumvent this issue by using *re-arrests* as proxy labels—such as `COMPAS` (Jackson and Mendoza 2020) or `OASYS` (Hamilton and Ugwudike 2023)—this is problematic for two reasons. First, the distinction between an *arrest* and *conviction* is best captured in the scrutiny of judicial due process. Second, arrest rates are often unequal across demographic groups. Using *re-arrest* labels can therefore reinforce bias through feedback loops, a risk the AI Act explicitly requires high-risk systems to mitigate.

Even with ground truth labels, **consistency** in recidivism risk assessment is difficult to achieve. There are often many equally accurate models that assign different predictions to the same individuals. This has been dubbed the *Rashomon Effect*:[1] when different models for the same hypothesis class perform similarly well on a prediction task (Breiman 2001). The set of all such models is called a *Rashomon Set*, and presents an opportunity to address algorithmic fairness concerns. One can, for instance, search within a *Rashomon Set* for models that satisfy desirable properties such as statistical non-discrimination criteria and monotonicity constraints (Fisher, Rudin, and Dominici 2019; Coston, Rambachan, and Chouldechova 2021; Rudin et al. 2024). Even then, predictive disagreement among such models

[1]The name "Rashomon" refers to the 1950 movie by director Akira Kurosawa, in which multiple people offer credible but mutually incompatible accounts of the same crime.

raises concerns of **arbitrariness** in the decision process. While ensembling and randomization have been proposed to resolve predictive multiplicity (Black, Raghavan, and Barocas 2022), both approaches have important limitations. Ensembling reduces interpretability, which is especially important in a high-stakes context, whereas randomization is often perceived unfavorably by stakeholders (Meyer et al. 2025). More broadly, much of the literature on model multiplicity does not engage with its implications in real-world settings (Ganesh, Taik, and Farnadi 2025).

Contributions. Our work seeks to address these concerns simultaneously. We study `RisCanvi`, a decision support system for recidivism risk assessment that has been operational across prisons in Catalonia since 2009. By translating a complex set of legal rules into an algorithm for determining post release outcomes (*recidivist* or *non-recidivist*), we construct a dataset of over 17.5K inmate releases between 2010 and 2019. Using this dataset, which is $17\times$ larger than the one `RisCanvi` is currently trained on, we build on the MILP (mixed-integer linear programming) `SLIM` formulation of Ustun and Rudin (2016) and the Rashomon Set exploration framework of Langlade et al. (2025). This allows us to learn models that are not only significantly more accurate than the one in use, but overcome important limitations. In particular, they exhibit lower error-rate disparities between subgroups, and ensure rehabilitative progress translates into lower risk scores through monotonicity constraints. Finally, we present a simple yet effective policy for addressing predictive arbitrariness in the set of all such models. Our **main contributions** are as follows:

1. We generalize the search for less discriminatory models in the Rashomon Set to $m > 2$ subgroups by adding only $O(m)$ constraints to the MILP of Langlade et al. (2025).
2. We extend the notion of self-consistency introduced by Cooper et al. (2024) to Rashomon Sets and derive a tight lower bound on expected self-consistency for any finite set of binary classifiers over a dataset.
3. We present the first (to the best of our knowledge) real-world case study of predictive multiplicity in recidivism risk assessment.
4. We propose and study a *lowest-risk policy* for addressing predictive arbitrariness, with a detailed discussion of its institutional and legal implications in recidivism risk assessment.

The rest of this paper is organized as follows. The next section provides a background on `RisCanvi`, on non-discrimination criteria for supervised learning, and on predictive multiplicity in Rashomon Sets (§ 2). Then, we describe our methodology, dataset, experimental set-up, and evaluation strategy (§ 3). Next, we presents our results (§ 4) and discussion (§ 5). The last section presents our conclusions and the limitations of this study (§ 6).

2 Background

In this section, we first provide background about the design goals and features of `RisCanvi`, and situate its use as a decision support tool for risk assessment (§ 2.1). We then introduce technical preliminaries on supervised learning and statistical non-discrimination criteria (§ 2.2). Finally, we overview related work on Rashomon Sets and the predictive multiplicity that arises from them (§ 2.3).

2.1 RisCanvi

`RisCanvi` is a decision support tool for risk assessment that has supported management of sentencing conditions and the provision of prison alternatives to inmates in Catalonia for more than 15 years (Andrés-Pueyo, Arbach-Lucioni, and Redondo 2018). It uses logistic regression with two types of features to predict recidivism risk: (i) 20 **static** features that describe immutable characteristics of the inmate (e.g., age at the onset of criminal activity), and (ii) 23 **dynamic** features that can vary over time, and in principle, should lead to risk reduction due to rehabilitative progress (e.g., level of education). An overview of features used by `RisCanvi` can be found in Appendix A. Thresholds over the predicted risk are then used to classify inmates into one of three risk levels (low, medium and high). Assigned risk levels support evaluations by the prison staff, including social workers, psychologists, and lawyers, which are considered in judicial decisions related to requests for parole, in changes to sentence conditions ("degree of imprisonment"), and in the design of rehabilitation programs.

`RisCanvi` is trained on labels that indicate whether an inmate recidivated within five years of being released. Specifically, it considers *penal recidivism*, i.e., when a person is sentenced to prison and actually enters prison to serve that prison sentence. However, reliably constructing these labels at scale is challenging. This is because penal recidivism is determined by a complex set of legal rules and applying these rules is a time-consuming process susceptible to human error. Moreover, due to limited institutional resources, this labeling process is not carried out regularly. As a result, `RisCanvi` often lacks up-to-date training data.

In principle, all prison inmates should undergo a `RisCanvi` evaluation once every six months. In reality, longitudinal evaluation records for most inmates have longer gaps. Only close to the end of their prison sentence we find that complete evaluations—having all features—are reliably found. This is partly because the law in Catalonia mandates risk assessment before an inmate is released from prison. In addition, other concerns about `RisCanvi` have been documented by external audits. Most notably, the system has a high false negative rate, and exhibits significant false positive rate disparities between nationals and foreigners (Dribia 2024).

2.2 Non-Discrimination in Supervised Learning

Let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} = \{-1, 1\}$ denote the feature space and label space respectively. In our case, $y = 1$ implies *recidivist* whereas $y = -1$ implies *non-recidivist*. A dataset consists of labeled examples $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathcal{X}$ is the

feature vector for instance i and $y_i \in \mathcal{Y}$ is the corresponding label. In supervised learning, the goal is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}$ from a specified hypothesis class $\mathcal{H}$ that minimizes a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ over $\mathcal{S}$. In binary classification, a standard choice is the 0-1 loss, $\ell(h(\mathbf{x}), y) = \mathbf{1}[h(\mathbf{x}) \neq y]$. The empirical risk of a classifier h on a dataset $\mathcal{S}$ is then $L_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(\mathbf{x}_i) \neq y_i]$.

In our work, we consider the hypothesis class $\mathcal{H}$ of linear integer scoring systems as they are widely used for recidivism risk assessments (Hoffman 1994; Hamilton et al. 2016). It consists of models that predict $h(\mathbf{x}) = 1$ if $\lambda_h^\top \mathbf{x} \geq \gamma$, and -1 otherwise. Here, $\lambda_h \in \mathbb{Z}^p$ denotes the integer coefficient vector and $\gamma \in \mathbb{R}$ denotes the decision threshold.

Let $\mathcal{G} = \{G_1, \ldots, G_m\}$ denote the partition of $\mathcal{D}$ into demographic groups induced by the intersections of sensitive attributes, and let $g : \mathcal{X} \to \mathcal{G}$ be the group membership function. A number of **statistical non-discrimination criteria** for supervised learning have been proposed in the literature; they typically equalize a group-dependent statistic (see, e.g., Barocas, Hardt, and Narayanan 2023). *Statistical Parity* (Dwork et al. 2012) equalizes prediction rates across all pairs of groups $G_i, G_j \in \mathcal{G}$:

$$\mathbb{P}(h(\mathbf{x}) = 1 \mid g(\mathbf{x}) = G_i) = \mathbb{P}(h(\mathbf{x}) = 1 \mid g(\mathbf{x}) = G_j)$$

This criterion can be insufficient when groups have unequal base rates, which is often the case with criminal recidivism. *Equalized Odds* (Hardt, Price, and Srebro 2016) addresses this limitation by conditioning a prediction $h(\mathbf{x})$ on its true outcome y. In particular, it requires the following to hold for all $y \in \mathcal{Y}$ and $G_i, G_j \in \mathcal{G}$:

$$\mathbb{P}(h(\mathbf{x}) = 1 \mid g(\mathbf{x}) = G_i, y) = \mathbb{P}(h(\mathbf{x}) = 1 \mid g(\mathbf{x}) = G_j, y).$$

2.3 Rashomon Sets and Predictive Multiplicity

The *Rashomon Effect*, i.e., the existence of multiple models with equivalent prediction accuracy, is prevalent in noisy prediction tasks such as criminal recidivism and commonly formalized through a *Rashomon Set* (Semenova et al. 2023). Let $h_{\mathcal{S}} = \arg\min_{h \in \mathcal{H}} L_{\mathcal{S}}(h)$ be the empirical risk minimizer. Then for $\epsilon \geq 0$, the ϵ-Rashomon Set consists of all models in $\mathcal{H}$ whose empirical risk lies within ϵ of $h_{\mathcal{S}}$ (Fisher, Rudin, and Dominici 2019). Formally:

$$\mathcal{R}(\mathcal{H}, \epsilon, \mathcal{S}) = \{h \in \mathcal{H} \ : \ L_{\mathcal{S}}(h) \leq L_{\mathcal{S}}(h_{\mathcal{S}}) + \epsilon\}.$$

Enumerating this set is generally intractable. Accordingly, there is a growing body of work on methods for exploring it without exhaustive enumeration across several hypothesis classes. These include linear models, generalized additive models, rule-lists, and decision diagrams (Coston, Rambachan, and Chouldechova 2021; Watson-Daniels, Parkes, and Ustun 2023; Zhong et al. 2023; Langlade et al. 2025). Two notable goals of this exploration are (i) finding less discriminatory models (LDMs) within the Rashomon Set (Gillis, Meursault, and Ustun 2024), and (ii) resolving arbitrariness in decision-making due to predictive multiplicity (Marx, Calmon, and Ustun 2020). Regarding the first goal, finding LDMs is important because a model sampled uniformly at random from the Rashomon Set may be substantially more discriminatory than the least discriminatory one (Dai et al. 2025). Regarding the second goal, some works also study arbitrariness arising from variation in the training data or learning procedure (Black and Fredrikson 2021).

Cooper et al. (2024) introduce *self-consistency* to measure how stable a prediction is across models trained with the same learning procedure on equally sized training samples. Formally, for an instance $\mathbf{x}$:

$$SC(\mathcal{A}, \mathbb{S}, \mathbf{x}) = \mathbb{E}_{h_{S_i}, h_{S_j} \sim \mu}\left[\mathbf{1}[h_{S_i}(\mathbf{x}) = h_{S_j}(\mathbf{x})]\right]$$

$$= \mathbb{P}_{h_{S_i}, h_{S_j} \sim \mu}(h_{S_i}(\mathbf{x}) = h_{S_j}(\mathbf{x}))$$

where $\mathbb{S}$ denotes the set of all equally sized subsets of $\mathcal{S}$, and μ is the distribution over models generated by applying $\mathcal{A}$ to samples $S \in \mathbb{S}$. In their framing, low self-consistency corresponds to a greater degree of arbitrariness.

Our work extends this notion to predictive multiplicity within the Rashomon Set, and derives a tight lower bound on expected self-consistency over a dataset $\mathcal{S}$ (in § 3.3).

3 Methods

In this section, we first introduce the MILP-based optimization framework with our extensions (§ 3.1), then describe our dataset, experimental setup, baselines, and evaluation metrics (§ 3.2), followed by our extension of self-consistency (Cooper et al. 2024) to study predictive multiplicity in finite model sets (§ 3.3).

3.1 Optimization Framework

We use a two-step optimization framework. First, we learn a linear integer scoring system that enforces monotonicity constraints on how static and dynamic features shape risk scores. Then, we use this scoring system to initialize a search for less discriminatory models within its Rashomon Set.

SLIM with Monotonicity Constraints. The `SLIM` (*Supersparse Linear Integer Model*) MILP, given by Ustun and Rudin (2016), is designed to learn linear scoring systems with integer coefficients. We use it to find an empirical risk-minimizer $h_{\mathcal{S}}$ for recidivism prediction, using balanced 0–1 loss as the objective. A key advantage of this approach is that modern MILP solvers either certify optimality or report an optimality gap, i.e., the difference between the incumbent feasible solution and the best bound on the objective.

Moreover, we enforce monotonicity constraints in `SLIM` by restricting feature coefficients to non-negative integers. These constraints are important from a practical standpoint because they guarantee that changes in recidivism risk align with expectations of the prison staff. Specifically, predicted risk should increase in the presence of more adverse indicators (static features) and decrease based on responsiveness to rehabilitation (dynamic features).

Finding Less Discriminatory Models. The framework proposed by Langlade et al. (2025) uses an empirical risk minimizer $h_{\mathcal{S}} \in \mathcal{H}$ to initialize the MILP that explores the corresponding ϵ-Rashomon Set. We use the solution returned by our extended `SLIM` formulation and build on their MILP formulation for scoring systems. In particular,

we modify it to minimize the maximum pairwise error-rate disparity across $m > 2$ groups. Let

$$\Delta_{\mathrm{FPR}}(h) = \max_{G_i, G_j \in \mathcal{G}} \left| \mathrm{FPR}_{G_i}(h) - \mathrm{FPR}_{G_j}(h) \right|$$

and

$$\Delta_{\mathrm{FNR}}(h) = \max_{G_i, G_j \in \mathcal{G}} \left| \mathrm{FNR}_{G_i}(h) - \mathrm{FNR}_{G_j}(h) \right|.$$

Our objective minimizes the *Equalized Odds* violation:

$$\Delta_{\mathrm{EO}}(h) = \max \left\{ \Delta_{\mathrm{FPR}}(h), \Delta_{\mathrm{FNR}}(h) \right\}.$$

Observe that the largest error-rate disparity between any two groups does not require comparing every pair explicitly. For false positive rates, the worst pair is simply the group with the highest FPR and the group with the lowest FPR; their difference equals the maximum pairwise FPR disparity. The same holds for false negative rates. We therefore introduce auxiliary variables that track the minimum and maximum FPR and FNR across groups, and minimize the larger of the two resulting gaps. This reduces the number of fairness constraints from $O(m^2)$, for all pairwise group comparisons, to $O(m)$, for m groups. Additionally, we reuse the monotonicity constraints from the `SLIM` formulation. We refer to the resulting formulation as `FairSLIM`, which is included in Appendix C.

3.2 Experimental Setup

We now describe in detail how we construct our dataset, the chronological train-test split, and the MILP configuration used to instantiate the optimization framework.

Iterative Algorithm Design. To identify recidivism reliably and at scale, we develop a labeling algorithm through an iterative process in close collaboration with researchers specializing in sentence enforcement. We begin by translating the legal rules that define penal recidivism into an initial rule-based algorithm. In each subsequent iteration, we share the labels assigned by the algorithm with the researchers. In turn, they annotate a random sample of this data and provide justifications for the incorrectly labeled instances. Then, we incorporate these justifications to refine the rule set of the algorithm. This process is repeated until labels assigned by the algorithm fully match those of the researchers. We validate the correctness of our proposed algorithm by comparing its labels against ground-truth labels for all releases between 2010 and 2015. Upon validation, we apply our algorithm to the remaining releases to construct our dataset. Appendix B provides a condensed version of the pseudo code for the labeling algorithm.

Dataset Description. The resulting dataset contains 17.5K releases from prisons in Catalonia between 2010 and 2019. Table 1 reports the number of releases and recidivism rates across groups defined by age, sex, and nationality. Age groups are split at 30, a common cutoff in criminology that reflects differences in types of crimes and social contexts between younger and older adults (Ulmer and Steffensmeier 2014). As described in § 2.1, each release is associated with 43 features from the inmate's last evaluation, typically conducted six to nine months before release.

Table 1: Distribution of recidivists and non-recidivists across groups based on sensitive attributes.

Group	# of Releases	Recidivism Rate
Age ≥ 30	14,049	0.36
Age < 30	3,493	0.50
Female	655	0.47
Male	16,887	0.38
National	11,785	0.37
Foreigner	5,757	0.42
Overall	17,542	0.39

Table 2: Overview of train-test split. Releases from train years 2015–2018 include only observed recidivists, since non-recidivists cannot be confirmed before the test year. For instance, a person released after 2015 may recidivate in 2020, which is after the test year 2019.

Split	Years	# of Releases	Recidivism Rate
Train	2010–2014	8,177	0.44
	2015–2018	1,958	1.00
Test	2019	1,881	0.29

Train-Test Split. We use a chronological train-test split. The train set includes releases from fully observed years 2010–2014 as well as recidivists from partially observed years (2015–2018) that precede the test set year, 2019. Non-recidivists for release years 2015-2018 cannot be ascertained until after the test year, and are therefore excluded from training. The test set consists of all inmates released in 2019, whose five-year observation period concluded in 2024. Table 2 summarizes this split. We did preliminary experiments using training data solely from the fully-observed years, and also probabilistically from partially observed years, but observed a consistently lower accuracy than in this set-up (results omitted for brevity).

MILP Configuration. Both `SLIM` and `FairSLIM` are implemented in Python using the Gurobi solver (Gurobi Optimization, LLC 2026). In `SLIM`, we use a finite coefficient grid that includes small incremental weights as well as larger values for highly predictive features. Since all features are processed so that larger values indicate higher risk, their coefficients λ are chosen from a set of non-negative values $\mathcal{L} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 25, 50, 100\}$ to enforce monotonicity. Intercept λ_0 is allowed to take both positive and negative values from the signed version of $\mathcal{L}$. Since all features take values in $\{0, 0.5, 1\}$ and coefficients are integers, every score lies on a grid with spacing 0.5. Thus any correctly classified point has margin at least 0.5 and any choice of $\gamma \in (0, 0.5]$ is valid. We set $\gamma = 0.1$. For robustness, we run `SLIM` for ten random seeds (four hours per seed). Although the solver does not converge within the allotted time, the majority of improvement in its objective occurs within the first 30 minutes for all seeds, with only marginal gains thereafter (details in Appendix D).

For every seed $s \in \{1, \ldots, 10\}$, we use the incumbent feasible solution h_s returned by SLIM to instantiate the FairSLIM MILP, and explore its Rashomon Set $\mathcal{R}(\mathcal{H}, \epsilon, \mathcal{S})$ over $\epsilon \in \{.01, .02, .03, .04, .05\}$. Each run is allowed 2 hours per ϵ per seed. As with SLIM, FairSLIM does not reach provable optimality within the allotted time. However, its incumbent objective stabilizes well before the time limit is reached, and exhibits much smaller optimality gaps compared to SLIM. During this search, we retain all models whose fairness objective Δ_{EO} lies within 5% of the best solution obtained by FairSLIM. We use $\mathcal{P}_s$ to denote the set (or pool) of all such models retained for seed s.

Baselines. We compare against two baselines: (i) the currently operational RisCanvi model which is based on logistic regression, and (ii) CatBoost (Prokhorenkova et al. 2018), a gradient-boosted tree method recommended by a recent third-party audit as a replacement for RisCanvi (Dribia 2024). As with SLIM, monotonicity constraints for features are also enforced in CatBoost.

Performance Metrics. We report F1 Score, Accuracy, and Balanced Accuracy for all models to assess their predictive performance. To evaluate discrimination, we report the maximum violation of error-rate parity across any two groups ($\Delta_{\text{FPR}}(h)$ and $\Delta_{\text{FNR}}(h)$). All metrics are averaged over 10 random seeds, and reported with standard deviation.

3.3 Predictive Arbitrariness and Self-Consistency

Our search for less discriminatory models using FairSLIM yields a model pool $\mathcal{P}_s \subseteq \mathcal{R}(\mathcal{H}, \epsilon, \mathcal{S})$ for every seed s. To study predictive multiplicity, we consider the aggregated pool over the ten seeded runs $\mathcal{P} = \bigcup_{s=1}^{10} \mathcal{P}_s$. We do so for two reasons. First, predictive multiplicity is a property of many competing models, and our seed-level pools $\mathcal{P}_s$ are too small to support meaningful evaluations of arbitrariness. Second, each seed-level pool $\mathcal{P}_s$ is obtained by initializing the solver with a different SLIM incumbent h_s. Aggregating across seeds therefore gives a larger and possibly more diverse set of less discriminatory models.

All models in $\mathcal{P}$ have comparable predictive performance and error-rate disparity on the train set. Therefore, any of these models could be a plausible candidate for deployment. To evaluate the resulting arbitrariness in risk assessment, we extend the notion of self-consistency (Cooper et al. 2024) to Rashomon Sets, and more generally, to any finite set of models. Formally, we define the self-consistency of a finite model set $\mathcal{P} = \{h_1, \ldots, h_K\}$ for an instance $\mathbf{x} \in \mathcal{S}$ as [2]:

$$SC_{\mathcal{P}}(\mathbf{x}) = \frac{1}{\binom{K}{2}} \sum_{i<j} \mathbf{1}[h_i(\mathbf{x}) = h_j(\mathbf{x})].$$

$SC_{\mathcal{P}}(\mathbf{x})$ can be interpreted as the probability that two models drawn uniformly at random from $\mathcal{P}$ produce the same prediction for a given instance $\mathbf{x}$.

Next, we provide a tight lower bound for expected self-consistency for any finite set of models $\mathcal{P}$ over a dataset $\mathcal{S}$.

[2]We slightly abuse notation and write $\mathbf{x} \sim \mathcal{S}$ instead of $(\mathbf{x}, y)$ to denote a feature vector sampled uniformly from a dataset $\mathcal{S}$.

Proposition 1. *Let $\mathcal{P} = \{h_1, \ldots, h_K\}$ be a finite set of binary classifiers, and let $\bar{L}_{\mathcal{S}}(\mathcal{P}) = \frac{1}{K} \sum_{h \in \mathcal{P}} L_{\mathcal{S}}(h)$ denote the average 0–1 loss of models in $\mathcal{P}$ on dataset $\mathcal{S}$.*

If $\mu = K\bar{L}_{\mathcal{S}}(P)$ and $\delta = \mu - \lfloor \mu \rfloor$, then the following tight lower bound holds for expected self-consistency:

$$\mathbb{E}_{x \sim \mathcal{S}}[SC_{\mathcal{P}}(x)] \geq 1 - \frac{1}{\binom{K}{2}} [\mu(K - \mu) - \delta(1 - \delta)]$$

The intuition behind the bound is as follows. If $r_{\mathbf{x}}$ denotes the number of models in $\mathcal{P}$ that misclassify $\mathbf{x}$, there are exactly $r_{\mathbf{x}}(K - r_{\mathbf{x}})$ pairs of models that disagree in their prediction for $\mathbf{x}$. Averaging this quantity over all instances connects pairwise disagreement to how the errors of models in $\mathcal{P}$ are distributed across $\mathcal{S}$.

Here, μ is the average number of models that misclassify an instance $\mathbf{x}$, whereas δ is simply the fractional part of this average. The bound is tight when errors are spread as evenly as possible across instances i.e., when a fraction $1 - \delta$ of instances are misclassified by $\lfloor \mu \rfloor$ models, and the remaining δ fraction of instances are misclassified by $\lfloor \mu \rfloor + 1$ models. If errors are less evenly spread, more models would misclassify the same instances, thereby increasing the average self-consistency. This result also generalizes a related disagreement bound by Black, Raghavan, and Barocas (2022) for the special case where all models in $\mathcal{P}$ have the same empirical risk. The proof of this bound is included in Appendix E.

It is worth noting that Proposition 1 does not depend on the model pool being a subset of the Rashomon Set, and therefore applies to any finite set of models. We use it to compare the predictive agreement in the model pools returned by FairSLIM against its worst-case guarantee.

Policies for Resolving Predictive Multiplicity. Finally, we evaluate three different policies for resolving predictive multiplicity: ensembling, random model selection (from $\mathcal{P}$ for every instance $\mathbf{x}$), and a *lowest-risk policy*. Since RisCanvi is used as a decision support system, we define the *lowest-risk policy* over the predicted risk. To each instance $\mathbf{x}$, it assigns the lowest possible risk score i.e., $\min_{h \in \mathcal{P}} \lambda_h^\top \mathbf{x}$. In the binary classification setting, this is equivalent to assigning the non-recidivist label to an individual if at least one model in $\mathcal{P}$ predicts that this individual will not recidivate. For the sake of completeness, we also compare it against a *highest-risk policy*, that assigns to an inmate the highest possible risk score i.e., $\max_{h \in \mathcal{P}} \lambda_h^\top \mathbf{x}$.

4 Results

In this section, we present our results, evaluating models based on their predictive utility and error-rate disparities (§ 4.1), followed by a detailed analysis of predictive multiplicity and self-consistency (§ 4.2).

4.1 Predictive Performance and Equalized Odds

Table 3 provides predictive performance and algorithmic fairness metrics for all models tested. First, we note that all models trained on post release outcomes labeled by our algorithm substantially outperform RisCanvi. While

Table 3: Predictive performance and algorithmic fairness metrics. Results for `CatBoost` and `SLIM` are averaged over ten seeds and reported as mean ± standard deviation. We highlight our proposed Lowest-Risk Policy (`FairSLIM-LRP`) which is applied to the aggregated model pool $\mathcal{P}$ with $\epsilon \leq .01$. The performance of `RisCanvi` on the train set is computed using only the fully observed years (2010–2014), otherwise its accuracy drops substantially (to 49.7%) due to errors in predicting the positive class.

Split	Model	F1 Score	Accuracy	Balanced Accuracy	FPR	FNR	$\Delta_{\text{FPR}}(h)$	$\Delta_{\text{FNR}}(h)$
Train	`RiskEval`	25.6	60.8	56.2	2.9	84.7	4.9	14.5
	`CatBoost`	80.9 ± 0.1	78.5 ± 0.1	78.1 ± 0.1	26.2 ± 0.1	17.7 ± 0.1	17.8 ± 0.7	11.5 ± 0.3
	`SLIM`	78.1 ± 0.3	76.0 ± 0.4	75.8 ± 0.6	26.1 ± 1.9	22.4 ± 1.1	30.1 ± 2.0	15.7 ± 1.9
	`FairSLIM-LRP`	71.1	72.3	73.6	14.5	38.4	14.2	8.9
Test	`RiskEval`	24.5	71.5	55.3	5.2	84.3	7.5	15.0
	`CatBoost`	58.7 ± 0.2	63.6 ± 0.2	70.7 ± 0.2	46.6 ± 0.2	12.1 ± 0.3	11.4 ± 0.4	5.4 ± 0.5
	`SLIM`	58.0 ± 0.6	64.6 ± 0.8	70.0 ± 0.6	43.0 ± 1.4	17.0 ± 0.9	2.5 ± 2.2	5.9 ± 3.2
	`FairSLIM-LRP`	58.5	71.2	70.6	27.8	31.1	0.5	5.9

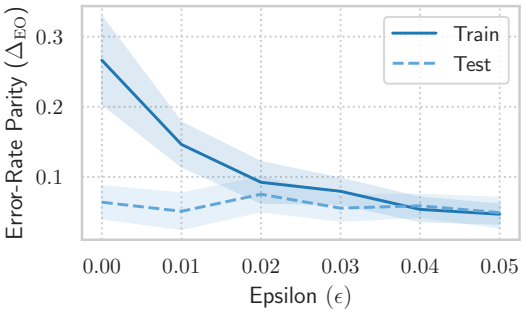


Figure 1: Error-rate parity (Δ_{EO}) achieved by the least discriminatory model learned using `FairSLIM` on the train and test sets for varying values of ϵ. A tolerance of $\epsilon \leq .01$ is sufficient to create a model pool with reduced error-rate disparities ($\leq 5\%$) on the test set.

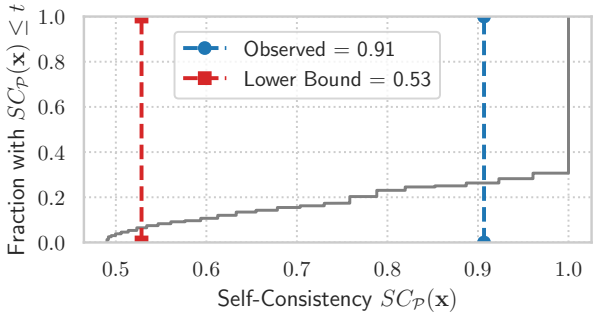


Figure 2: Empirical CDF of Self-Consistency on the test set for the aggregated model pool $\mathcal{P}$ with $|\mathcal{P}| = 78, \epsilon \leq .01$. Nearly 70% instances exhibit perfect predictive agreement across all models, resulting in high average self-consistency.

the accuracy of `RisCanvi` may seem comparable to that of alternatives, this is largely driven by it predicting the majority class. As a result, it has poor balanced accuracy and a high false negative rate. In contrast, both `SLIM` and `CatBoost` achieve balanced accuracies of approximately 70%, reducing the false negative rate by more than 65%. Second, we find that `SLIM` matches `CatBoost`'s predictive performance on the test set, while exhibiting lower error-rate disparities across subgroups. In fact, searching for less discriminatory models (LDMs) using `FairSLIM` within 1% of the best `SLIM` objective ($\epsilon \leq .01$) reduces the maximum error-rate disparity between any two subgroups to 5% on the test set, as shown in Figure 1. Together, these results undermine the case for replacing `RisCanvi` with a more complex model such as `CatBoost`, rather than with a simpler and interpretable scoring system like `SLIM`.

4.2 Predictive Multiplicity in Model Pools

We now study predictive multiplicity in the model pools returned by `FairSLIM`. In particular, we examine where and how often these models disagree on individual predictions, and evaluate a simple way of addressing the resulting predictive arbitrariness.

Self-Consistency and Structural Diversity. Despite the existence of many similarly accurate models with similar error-rate disparities across groups, model multiplicity translates into limited predictive multiplicity. Instead, the model pools returned by `FairSLIM` exhibit high predictive agreement. For instance, as shown in Figure 2, the probability that two models sampled uniformly at random from the pool assign the same prediction to a random individual from the test set is 91%. Moreover, Figure 3 shows that the average self-consistency remains substantially higher across values of ϵ than worst case analysis suggests (based on the lower bound from Proposition 1). This trend also holds at a subgroup-level (details in Appendix F), and is invariant to the size of the model pool returned by `FairSLIM` (details in Appendix G).

Disagreement in prediction is limited to instances that lie close to the decision boundary on average, as shown in Figure 4. We measure closeness in terms of the average absolute margin across $\mathcal{P}$, given by $\frac{1}{K}\sum_{h\in\mathcal{P}} |\lambda_h^\top \mathbf{x} - \gamma|$. One may think this is because the models in the pool are structurally alike, with similar feature coefficients and therefore similar risk scores. In reality, the model pool is structurally diverse, with coefficients varying significantly across both static and dynamic features (see

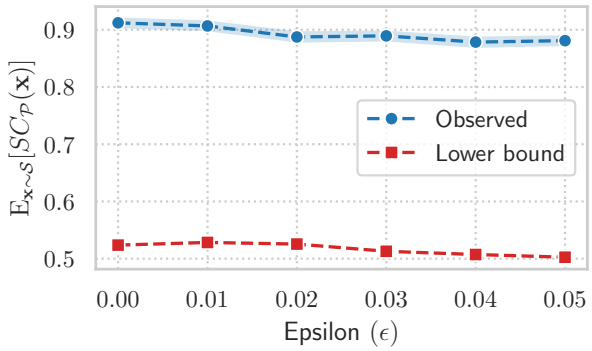


Figure 3: Observed versus worst-case expected self-consistency on the test set for varying values of ϵ. While the lower bound becomes increasingly pessimistic as ϵ grows, the observed average self-consistency remains high.

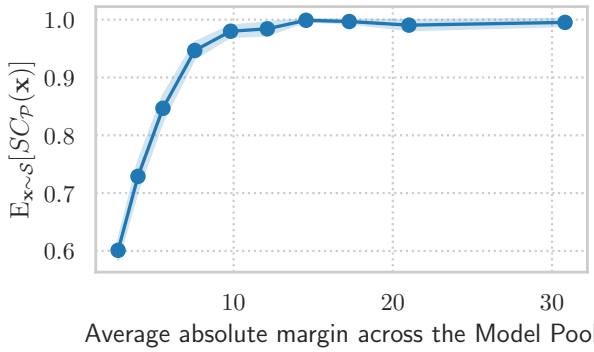


Figure 4: Expected self-consistency as a function of average absolute margin across the aggregated model pool $\mathcal{P}$ for $\epsilon \leq .01$. Predictive disagreement is concentrated among low-margin instances, while high-margin instances have almost perfectly consistent predictions.

Figure 5). Overall, we find that models largely agree in their predictions despite relying on different scoring rules, indicating that high self-consistency can coexist with structural diversity.

A Lowest-Risk Policy to Address Predictive Multiplicity. Predictive multiplicity in the model pools returned by `FairSLIM` is concentrated among approximately 30% of instances on the test set, most of which lie close to the decision boundary on average. Therefore, any mechanism for resolving the resulting arbitrariness in risk assessment will predominantly affect instances for which models cannot produce a confident prediction. In this case, we suggest applying the *favor rei* principle ("rule in favor of the defendant" [when in doubt]) and the anti-subordination principle, preferring decision-making policies that actively address systemic biases (Keswani and Celis 2024). Specifically, we propose assigning each inmate the risk computed by the model in the pool that produces the lowest risk score for that inmate. This simple policy, which we call `FairSLIM-LRP` (LRP stands for lowest risk policy) achieves higher accuracy and lower false positive rate disparity than both `CatBoost` and `SLIM` on the test set (see Table 3). In terms of relaxed equalized odds, `FairSLIM-LRP` performs just as well as the best `FairSLIM` solution, i.e., with the lowest Δ_{EO}, while achieving higher balanced accuracy, as shown in Figure 6.

The lowest-risk policy also achieves the highest balanced accuracy among the resolution mechanisms we consider, including ensembling, random model selection, and a highest-risk policy (which we test for the sake of completeness), while maintaining comparable error-rate disparities for $\epsilon \leq .03$. Figure 7 provides a more detailed comparison with various values of ϵ.

The effectiveness of `FairSLIM-LRP` can be explained by various factors, mainly that low-risk is the majority outcome and we are intervening on a relatively small number of cases. Additionally, our training data does not contain non-recidivist labels for the set of recently freed people who have been outside of prison for more than one year and less than five years—most of whom will not recidivate. This is because most recidivists reoffend within the first year of being released.

5 Discussion

Our work addresses several challenges in `RisCanvi`, a system that supports recidivism risk assessment for thousands of incarcerated people every year. We work closely with researchers in sentence enforcement from the Department of Justice of Catalonia to translate legal rules into an algorithm that systematically labels post release outcomes. This helps us obtain **reliable training data** and eliminates reliance on a time-consuming labeling process that is susceptible to human error. In practice, manual labeling of a test set would be recommended for periodic testing of the system, but not needed to generate a much larger set required for training. Using this data, we extend existing MILPs to learn interpretable models that are significantly more accurate, distribute predictive errors more evenly across groups, and ensure by design that rehabilitative progress lowers risk scores.

Our analysis of **predictive multiplicity** in the model pool highlights several important considerations. The existence of structurally diverse models does not necessarily mean vastly different predictions. Empirically, models exhibit much higher predictive agreement than what the theoretical lower bounds guarantee. Thus, it is entirely possible that the **arbitrariness** due to predictive multiplicity only affects a small proportion of instances. In our case, these instances lie close to the decision boundary on average, making their predictions ambiguous across most models.

Our results indicate that a simple policy following the *favor rei* principle that prevents arbitrariness by assigning each inmate the lowest possible risk among the models is effective. In the context of the anti-subordination principle (Keswani and Celis 2024), this policy helps address structural asymmetries created by

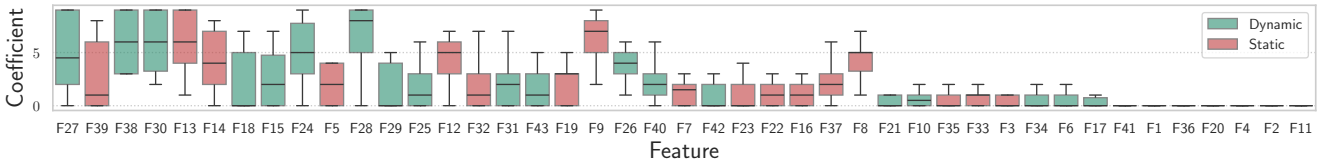
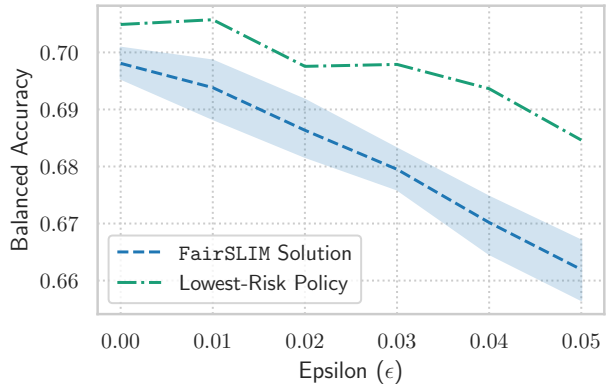
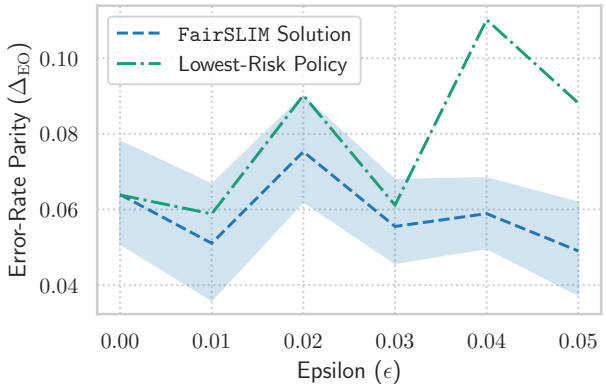


Figure 5: Variance in feature coefficients across the model pool for $\epsilon \leq .01$. Although $\mathcal{P}$ exhibits high average self-consistency, coefficients vary substantially across both dynamic and static features. Features are sorted by variance in decreasing order.



(a) Balanced accuracy of the best `FairSLIM` solution and the *lowest-risk policy* over the model pool $\mathcal{P}$ for varying values of ϵ.



(b) Error-rate parity (Δ_{EO}) of the best `FairSLIM` solution and the *lowest-risk policy* over the model pool $\mathcal{P}$ for varying values of ϵ.

Figure 6: A *lowest-risk policy* over the aggregated model pool $\mathcal{P}$ achieves higher balanced accuracy than the best `FairSLIM` solution (lowest Δ_{EO}) on the test set, while demonstrating comparable error-rate parity (Δ_{EO}) for $\epsilon \leq .03$.

predictive multiplicity. In actuarial risk assessment, the scoring rule underlying a model is typically not visible to incarcerated individuals, or the prison staff responsible for evaluating them. As a result, affected individuals lack the resources and information needed to contest assigned risk scores, even when similarly valid models could have produced less punitive outcomes. This concern is especially consequential in discrimination claims, where the burden of producing *prima facie* evidence rests with the plaintiff under EU law (Adams-Prassl, Binns, and Kelly-Lyth 2023). These considerations become even more complex once risk assessment is viewed as an ongoing institutional process rather than a one-time prediction task.

Periodic evaluations and model updates. Articles 17 and 72 of the AI Act require periodic evaluations of high-risk systems (European Union 2024), which may necessitate re-training of models on more recent data. In this setting, predictive arbitrariness becomes more complex because model pools can change over time. For instance, if an inmate is assessed by different models in consecutive evaluations, their risk assessments may fail to reflect rehabilitative progress if the updated model omits or downweights certain dynamic features. This prevents **algorithmic recourse**, i.e., the provision of information on dynamic (mutable) features whose changes are likely to reduce the computed risk (Ustun, Spangher, and Liu 2019). However, we believe that rehabilitation programs should not be driven by `RisCanvi`. Instead, they should be guided by models specifically designed to identify interventions that are most likely to reduce recidivism risk—not models that predict risk in the absence of interventions (Barabas et al. 2018). This is a separate and rather challenging task in the current institutional setting, in which, for instance, the policy of evaluating each inmate every six months is not strictly implemented, as reported in § 2.1, and in which rehabilitative progress of inmates is not guaranteed to lower their risk scores. Together, these concerns motivate our focus on models with better predictive performance, lower error-rate disparities, and explicit mechanisms for risk-reduction.

6 Conclusions

This work, first, highlights the importance of automating manual labeling procedures, as opposed to relying on proxy labels or outdated data. More specifically, the much larger set of post release outcomes labeled by our algorithm materially improves model development by increasing predictive performance, reducing error-rate disparities and aligning risk scores with domain expertise.

Second, our results show that predictive multiplicity is not solely determined by the number of sufficiently distinct and similarly performant models; what matters

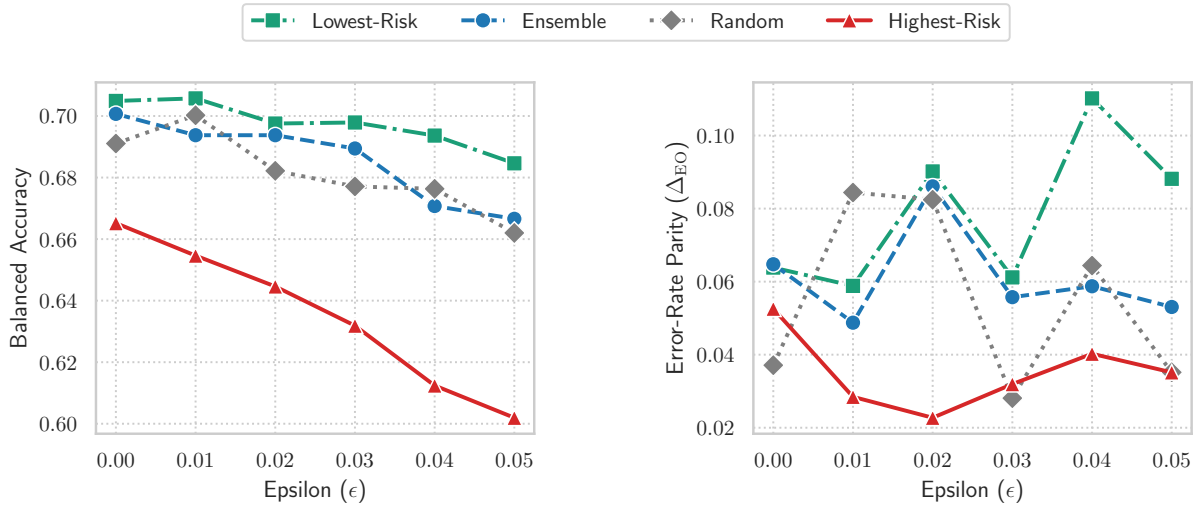


Figure 7: Comparison of policies for resolving predictive multiplicity over the aggregated model pool $\mathcal{P}$. The *lowest-risk policy* yields the highest balanced accuracy among those considered and maintains comparable error-rate parity for $\epsilon \leq .03$.

instead is the extent to which their predictions overlap. Consequently, even when many such models may exist, predictive multiplicity may be far less severe than worst-case analysis suggests. Proposition 1 serves as a useful reference for empirical estimates of predictive agreement. For instance, using procedures that independently sample from the Rashomon Set, one can estimate expected self-consistency and compare it against the worst-case guarantee. Similarly, estimating self-consistency over models that satisfy additional desiderata (statistical non-discrimination criteria, monotonicity etc.) can reveal the residual predictive arbitrariness after imposing these constraints.

A broader implication of this research is that, in the presence of model multiplicity, *where* predictive disagreement occurs can greatly inform mechanisms for addressing the resulting arbitrariness in decision-making. In particular, the geometry of disagreement (e.g., where it is concentrated relative to the decision threshold) may support policies that are grounded in existing institutional principles (e.g., resolving ambiguity in favor of subjects).

6.1 Limitations and Future Work

Our work has several limitations stemming from the institutional setting, resource availability, and the scope of our analysis. First, our analysis is based on a single operational system in Catalonia. Therefore, our empirical results are highly contextual, and may not generalize to other jurisdictions with different legal rules, risk assessment systems, and data collection practices. We cannot release the dataset, but in our camera-ready version we will provide instructions to request access to it from the corresponding authorities under a research agreement, in the same manner as we did when starting this research. Second, due to limited data, our work does not explicitly address the challenges introduced by periodic risk assessments and model updates. Third, our analysis is limited to the predictive aspect of risk assessment, whereas `RisCanvi` operates as a decision support system. An end-to-end evaluation of the effectiveness of any new model or policy is necessary, and should be conducted in a decision support setting. This means evaluating the extent to which actual users of the system understand and interpret risk scores by different models, and more importantly, whether they reach more accurate decisions in a more reliable manner using this tool. Finally, our study on predictive multiplicity is focused on the hypothesis class of integer linear models. This choice reflects the design of `RisCanvi`, and also the broader use of linear scoring systems in recidivism risk assessment, where static and dynamic features contribute additively to the final risk score. Generalizing some of the results, e.g., those related to the distribution of predictive disagreement with respect to the decision boundary, requires the exploration of a broader class of models.

Acknowledgments

This work was partially funded by contract CEJFE-2024-124 (CNR03324) with the Centre d'Estudis Jurídics i Formació Especialitzada (CEJFE). We are grateful for the support of CEJFE, especially Marian, Alba, Susana, and Abril, for sharing their expertise on sentence enforcement. We also disclose that the second author, through UPF, provided training to CEJFE in 2021 on causal analysis methods.

7 Researcher Positionality Statement

We now reflect on our position and our engagement with `RisCanvi` as an algorithmic institution (Mendonça et al. 2024). We start by acknowledging our formal training as computer scientists, with additional background in social and responsible computing. Despite our due diligence, it is possible that we overlook important considerations by researchers from other disciplines who work more closely with risk assessments in criminal justice contexts. Nevertheless, we collaborate closely with domain experts in sentence enforcement. These experts have the responsibility of generating official statistics on recidivism in Catalonia, and have years of experience in various research projects on the topic. Lastly, our lived experiences are far removed from those of incarcerated individuals who undergo routine evaluations by `RisCanvi`. This significantly limits our ability to assess the efficacy of institutional processes through which rehabilitative needs are identified and met.

8 Adverse Impacts Statement

First and foremost, our work does not engage with the abolitionist debate regarding prisons. Rather, our work is situated within the existing legal obligation in Catalonia to conduct risk assessments before the definitive release of any person. Although this position reinforces existing carceral structures, we believe it is urgent to address systemic issues in recidivism risk assessment tools because they continue to shape the lives of incarcerated people.

Second, `RisCanvi` is based on the RNR (*risk-need-responsivity*) model from criminology (Andrews, Bonta, and Hoge 1990), which requires rehabilitative interventions to (i) match their intensity to the level of recidivism **risk**, (ii) address individual criminogenic **needs**, and (iii) be tailored to individual **responsivity** to treatment. `RisCanvi` operationalizes risk and responsivity through actuarial assessments, but provides limited visibility into how rehabilitative needs are addressed in the day-to-day lives of incarcerated people. As a result, systems like `RisCanvi` are susceptible to reducing their rehabilitation to what can be measured through actuarial assessments.

Third, `RisCanvi` operates as a decision support system, in compliance with Article 22 of the GDPR that affords EU residents the right to not be subject to solely automated decisions that produce significant effects (European Union 2016). Recent work supports this operational setting, showing that recidivism risk assessments are most accurate when they follow a structured approach while trained professionals retain the agency to overrule model predictions (Portela et al. 2025). Therefore, we advocate for using such tools solely as decision support systems, operated by trained professionals under appropriate institutional incentives that promote accountability and care. Nothing on this paper should be interpreted as a justification to reduce careful scrutiny of the outputs of these systems, given that they are only as reliable as their weakest link. For instance, investigative reporting on other risk assessment tools has documented cases where record keeping errors contributed to parole denials despite substantial rehabilitative progress (Wexler 2017).

Fourth, `RisCanvi` must be used within its intended scope. However, there have been reports of multiple cases involving a high risk score computed by `RisCanvi`, which has been reduced, with justification, by human evaluation teams, but then prosecutors have cited the tool's score in parole proceedings as grounds for opposing release (Jimenez Arandia 2026). Nothing in our research should be interpreted as justifying out-of-scope uses of `RisCanvi`.

References

Adams-Prassl, J.; Binns, R.; and Kelly-Lyth, A. 2023. Directly Discriminatory Algorithms. *The Modern Law Review*, 86: 144–175.

Andrews, D. A.; Bonta, J.; and Hoge, R. D. 1990. Classification for Effective Rehabilitation: Rediscovering Psychology. *Criminal Justice and Behavior*, 17(1): 19–52.

Andrés-Pueyo, A.; Arbach-Lucioni, K.; and Redondo, S. 2018. *The RisCanvi*. John Wiley and Sons, Ltd. ISBN 9781119184256.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica*.

Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; and Venkatasubramanian, S. 2021. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*.

Barabas, C.; Virza, M.; Dinakar, K.; Ito, J.; and Zittrain, J. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 62–76. PMLR.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Black, E.; and Fredrikson, M. 2021. Leave-one-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 285–295. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 850–863. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–215.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22004–22012.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2144–2155. PMLR.

Dai, G.; Ravishankar, P.; Yuan, R.; Black, E.; and Neill, D. B. 2025. Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '25, 42–73. New York, NY, USA: Association for Computing Machinery. ISBN 9798400721403.

Dribia. 2024. Informe Tiresias: Auditoria de l'algorisme RisCanvi. Generalitat de Catalunya, Departament de Justícia, Drets i Memòria.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151.

European Union. 2016. General Data Protection Regulation. Regulation (EU) 2016/679, Article 22: Automated individual decision-making, including profiling.

European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Accessed: May 2026.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.

Ganesh, P.; Taik, A.; and Farnadi, G. 2025. Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1032–1048.

Gillis,
T. B.; Meursault, V.; and Ustun, B. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 377–387. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Gurobi Optimization, LLC. 2026. Gurobi Optimizer Reference Manual.

Hamilton, M.; and Ugwudike, P. 2023. A 'black box' AI system has been influencing criminal justice decisions for over two decades – it's time to open it up. The Conversation.

Hamilton, Z.; Kigerl, A.; Campagna, M.; Barnoski, R.; Lee, S.; van Wormer, J.; and Block, L. 2016. Designed to Fit: The Development and Validation of the STRONG-R Recidivism Risk Assessment. *Criminal Justice and Behavior*, 43(2): 230–263.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hoffman, P. B. 1994. Twenty years of operational use of a risk prediction instrument: The United States parole commission's salient factor score. *Journal of Criminal Justice*, 22(6): 477–494.

Jackson, E.; and Mendoza, C. 2020. Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not. *Harvard Data Science Review*.

Jimenez Arandia, P. 2026. Un algoritmo dificulta la reinserción de los presos vulnerables en Catalunya. Público. Updated March 11, 2026.

Karimi-Haghighi, M. 2022. *Risk Assessment in Complex Data Settings: Algorithmic Fairness and Causal Inference*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Keswani, V.; and Celis, L. E. 2024. Algorithmic fairness from the perspective of legal anti-discrimination principles. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 724–737.

Langlade, L.; Ferry, J.; Laberge, G.; and Vidal, T. 2025. Fairness and Sparsity Within Rashomon Sets: Enumeration-Free Exploration and Characterization. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1536–1547.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive Multiplicity in Classification. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6765–6774. PMLR.

Mendonça, R. F.; Almeida, V.; Filgueiras, F.; and Almeida, V. A. 2024. *Algorithmic institutionalism: the changing rules of social and political life*. Oxford University Press.

Meyer, A. P.; Kim, Y.-S.; D'Antoni, L.; and Albarghouthi, A. 2025. Perceptions of the Fairness Impacts of Multiplicity in Machine Learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Portela, M.; Castillo, C.; Tolan, S.; Karimi-Haghighi, M.; and Pueyo, A. A. 2025. A comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artificial Intelligence and Law*, 33(2): 471–517.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Rudin, C.; Wang, C.; and Coker, B. 2020. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1): 1.

Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Position: Amazing Things Come From Having Many Good Models. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 42783–42795. PMLR.

Semenova, L.; Chen, H.; Parr, R.; and Rudin, C. 2023. A Path to Simpler Models Starts With Noise. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 3362–3401. Curran Associates, Inc.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 505–514.

Ulmer, J.; and Steffensmeier, D. 2014. *The age and crime relationship: Social variation, social explanations*, 377–396. United States: SAGE Publications Inc. ISBN 9781452242255.

Ustun, B.; and Rudin, C. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3): 349–391.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 10–19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Watson-Daniels, J.; Parkes, D. C.; and Ustun, B. 2023. Predictive Multiplicity in Probabilistic Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10306–10314.

Wexler, R. 2017. Code of Silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out. Washington Monthly.

Zhong, C.; Chen, Z.; Liu, J.; Seltzer, M.; and Rudin, C. 2023. Exploring and interacting with the set of good sparse generalized additive models. In *Advances in neural information processing systems*, volume 36, 56673–56699.

A RisCanvi Features

Table 4 at the end of this document, lists all of the static and dynamic features used by `RisCanvi` for risk assessment. These were selected in 2009 from a pool of over 100 features, through a process that tested their statistical dependency with relevant outcomes (including recidivism) plus the input of domain experts (Andrés-Pueyo, Arbach-Lucioni, and Redondo 2018).

B Rule-Based Recidivism Labeling

Below, we provide pseudocode of our recidivism labeling algorithm. Note that more fine-grained details are abstracted as the algorithm is highly customized to the internal database. Put simply, it searches for crimes committed by an inmate in less than five years **after** being released for which the inmate **re-enters prison** to serve a sentence.

Algorithm 1: Rule-based labeling of penal recidivism

Input: Inmate I, release date R, release procedure P
Output: True/False

Let $t^\star \leftarrow \infty$
$D \leftarrow 5$ years
for all cases $C \in \text{Cases}(I)$ **do**
  if $\text{ProcedureID}(C) = P$ **then**
    continue {Case is related to the current release.}
  else if $\text{Verdict}(C) = \text{Conviction}$ **then**
    $M \leftarrow \{m \in \text{Crimes}(C) : \text{Date}(m) > R\}$
    $d \leftarrow \min_{m \in M} \text{Date}(m)$
  else if $\text{Verdict}(C) = \text{Preventive Sentencing}$ **then**
    $d \leftarrow \text{DetentionDate}(C)$
  else
    continue
  end if
  if $d > R$ **then**
    $t^\star \leftarrow \min\{t^\star, \text{TimeDiff}(R, d)\}$
  end if
end for
return $t^\star < D$

C FairSLIM Extension to $m > 2$ Groups

Notation. We use $[n]$ to denote the set $\{1, \ldots, n\}$, and the function $I : \mathcal{S} \to [n]$ maps instances in $\mathcal{S}$ to their indices. We define $\mathcal{S}_G = \{\mathbf{x} \in \mathcal{S} \mid g(\mathbf{x}) = G\}$. $\mathcal{L}_j$ denotes the set of values that the coefficient of feature j, i.e., λ_j can take.

The complete `FairSLIM` MILP is as follows:

$$\min_{\lambda} \quad \Delta_{\text{EO}} \tag{1}$$

$$\text{s.t.} \quad \lambda_j = \sum_{\omega \in \mathcal{L}_j} \omega \cdot u_{j\omega} \qquad \forall j \in [p] \tag{2}$$

$$\sum_{\omega \in \mathcal{L}_j} u_{j\omega} = 1 \qquad \forall j \in [p] \tag{3}$$

$$\frac{1}{n} \sum_{i=1}^{n} z_i \leq L_{\mathcal{S}}(h_{\mathcal{S}}) + \epsilon \tag{4}$$

$$M_i z_i \geq \gamma - y_i \lambda^\top \mathbf{x}_i \qquad \forall i \in [n] \tag{5}$$

$$O_i (1 - z_i) \geq y_i \lambda^\top \mathbf{x}_i - \gamma \qquad \forall i \in [n] \tag{6}$$

$$\text{FPR}_G = \frac{1}{|\mathcal{S}_G^-|} \sum_{i \in I(\mathcal{S}_G^-)} z_i \qquad \forall G \in \mathcal{G} \tag{7}$$

$$\text{FNR}_G = \frac{1}{|\mathcal{S}_G^+|} \sum_{i \in I(\mathcal{S}_G^+)} z_i \qquad \forall G \in \mathcal{G} \tag{8}$$

$$\text{FPR}_{\min} \leq \text{FPR}_G \leq \text{FPR}_{\max} \qquad \forall G \in \mathcal{G} \tag{9}$$

$$\text{FNR}_{\min} \leq \text{FNR}_G \leq \text{FNR}_{\max} \qquad \forall G \in \mathcal{G} \tag{10}$$

$$\Delta_{\text{EO}} \geq \text{FPR}_{\max} - \text{FPR}_{\min} \tag{11}$$

$$\Delta_{\text{EO}} \geq \text{FNR}_{\max} - \text{FNR}_{\min} \tag{12}$$

$$u_{j\omega} \in \{0, 1\} \qquad \forall j \in [p], \omega \in \mathcal{L}_j \tag{13}$$

$$z_i \in \{0, 1\} \qquad \forall i \in [n] \tag{14}$$

MILP Description. Equations (2) and (3) ensure that every coefficient λ_j only takes one value from $\mathcal{L}_j$ by using $u_{j\omega}$ as selector variables. Equation (4) represents the Rashomon constraint which ensures that the balanced 0–1 loss of `FairSLIM` stays within ϵ tolerance of $L_{\mathcal{S}}(h_{\mathcal{S}})$ where $h_{\mathcal{S}}$ is the incumbent feasible solution obtained from `SLIM`. Here, $z_i = 1$ if instance $\mathbf{x}_i$ has been misclassified and 0 otherwise. This is enforced by margin constraints (5) and (6). In (5), $M_i = \max_{\lambda \in \mathcal{L}}(\gamma - y_i \lambda^\top \mathbf{x}_i)$ is the maximum margin violation (during misclassification). Similarly, in (6), $O_i = \max_{\lambda \in \mathcal{L}}(y_i \lambda^\top \mathbf{x}_i - \gamma)$ is an upper bound on the margin (when points are correctly classified). Constraints (7) and (8) compute false positive and false negative rates respectively for all groups $G \in \mathcal{G}$. In (9) and (10), auxiliary variables $\text{FPR}_{\max}, \text{FPR}_{\min}, \text{FNR}_{\max}, \text{FNR}_{\min}$ track the maximum and minimum false positive and false negative rates. Constraints (11) and (12) ensure that the objective Δ_{EO} is lower bounded by the maximum error-rate disparity between any two groups $G_i, G_j \in \mathcal{G}$. Lastly, (13) and (14) simply define the domain of z_i and $u_{j\omega}$. **To summarize, our contributions are constraints (7-12)**.

D Convergence and Runtime

Figures 8 and 9 show the convergence behavior of `SLIM` and `FairSLIM` respectively. While neither MILP reaches provable optimality within the allotted runtime, the incumbent objective stabilizes rapidly, with most improvements occurring within the first 30 minutes. Overall, we observe that the optimality gap is much smaller for the `FairSLIM` runs.

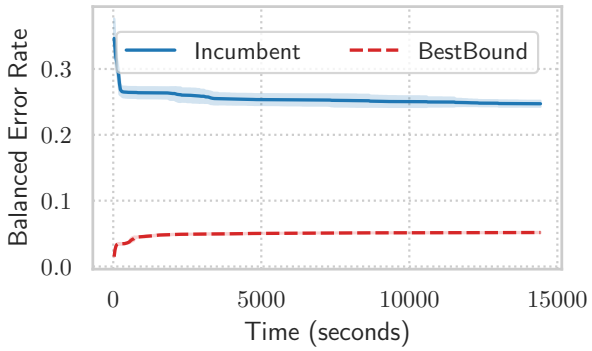


Figure 8: Convergence behavior of SLIM. Although the solver does not converge within the allotted time, the majority of improvement in the objective occurs within the first 30 minutes, with only marginal gains thereafter.

E Bound for Expected Self-Consistency

In this section, we restate our bound on expected self-consistency given in the paper, followed by its proof. For notational convenience, we encode predictions and labels as binary vectors in $\{0,1\}^{|\mathcal{S}|}$.

Proposition 2. *Let $\mathcal{P} = \{h_1, \ldots, h_K\}$ be a finite set of binary classifiers, and let $\bar{L}_{\mathcal{S}}(\mathcal{P}) = \frac{1}{K}\sum_{h\in\mathcal{P}} L_{\mathcal{S}}(h)$ denote the average 0–1 loss of models in $\mathcal{P}$ on dataset $\mathcal{S}$.*

If $\mu = K\bar{L}_{\mathcal{S}}(P)$ and $\delta = \mu - \lfloor\mu\rfloor$, then the following tight lower bound holds for expected self-consistency:

$$\mathbb{E}_{x\sim\mathcal{S}}[SC_{\mathcal{P}}(x)] \geq 1 - \frac{2}{K(K-1)}\left[\mu(K-\mu) - \delta(1-\delta)\right]$$

Proof. For each instance $(\mathbf{x}, y) \in \mathcal{S}$, let $r_{\mathbf{x}} = \sum_{h\in\mathcal{P}} \mathbf{1}[h(\mathbf{x}) \neq y]$ denote the number of models in $\mathcal{P}$ that misclassify $\mathbf{x}$. Then $K - r_{\mathbf{x}}$ models classify $\mathbf{x}$ correctly. The number of disagreeing model pairs on $\mathbf{x}$ is simply $r_{\mathbf{x}}(K - r_{\mathbf{x}})$. Hence, the self-consistency of $\mathcal{P}$ on instance $\mathbf{x}$ can be written as:

$$SC_{\mathcal{P}}(\mathbf{x}) = 1 - \frac{1}{\binom{K}{2}} r_{\mathbf{x}}(K - r_{\mathbf{x}})$$

Taking expectations over $\mathbf{x} \sim \mathcal{S}$, we get:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[SC_{\mathcal{P}}(\mathbf{x})] &= 1 - \frac{1}{\binom{K}{2}}\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}\left[r_{\mathbf{x}}(K - r_{\mathbf{x}})\right] \\ &= 1 - \frac{1}{\binom{K}{2}}\left(K\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}] - \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}^2]\right)\end{aligned} \tag{1}$$

Observe that

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}] &= \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},y)\in\mathcal{S}}\sum_{h\in\mathcal{P}}\mathbf{1}[h(\mathbf{x}) \neq y] \\ &= \sum_{h\in\mathcal{P}}\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},y)\in\mathcal{S}}\mathbf{1}[h(\mathbf{x}) \neq y] \\ &= \sum_{h\in\mathcal{P}} L_{\mathcal{S}}(h) \\ &= K\bar{L}_{\mathcal{S}}(\mathcal{P}).\end{aligned} \tag{2}$$

Substituting (2) in (1), we get

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[SC_{\mathcal{P}}(\mathbf{x})] = 1 - \frac{1}{\binom{K}{2}}\left(K^2\bar{L}_{\mathcal{S}}(\mathcal{P}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}^2]\right) \tag{3}$$

To obtain a lower bound on $\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[SC_{\mathcal{P}}(\mathbf{x})]$, we must minimize $\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}^2]$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}] = K\bar{L}_{\mathcal{S}}(\mathcal{P})$. Before that, we prove the following lemma.

Lemma 1. *Let R be a random variable which takes values in $\{0, \ldots, K\}$ with $\mathbb{E}[R] = \mu$. If $\delta = \mu - \lfloor\mu\rfloor$, then:*

$$\mathbb{E}[R^2] \geq (1-\delta)\lfloor\mu\rfloor^2 + \delta(\lfloor\mu\rfloor + 1)^2.$$

Proof. Let $\tilde{f}$ be the piecewise-linear interpolation of $f(r) = r^2$ on $\{0, \ldots, K\}$. Since f is convex, so is $\tilde{f}$. Applying Jensen's inequality to $\tilde{f}$, we obtain:

$$\mathbb{E}[R^2] = \mathbb{E}[\tilde{f}(R)] \geq \tilde{f}(\mathbb{E}[R]) = \tilde{f}(\mu) \tag{4}$$

Since $\mu \in [\lfloor\mu\rfloor, \lfloor\mu\rfloor + 1]$ and $\mu = (1-\delta)\lfloor\mu\rfloor + \delta(\lfloor\mu\rfloor + 1)$, the linearity of $\tilde{f}$ on this interval gives

$$\tilde{f}(\mu) = (1-\delta)\lfloor\mu\rfloor^2 + \delta(\lfloor\mu\rfloor + 1)^2 \tag{5}$$

Combining (4) and (5) completes the proof of this lemma. □

Now, using Lemma 1 with $R = r_{\mathbf{x}}$, $\mu = K\bar{L}_{\mathcal{S}}(\mathcal{P})$ and $\delta = \mu - \lfloor\mu\rfloor$, we get:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}^2] &\geq (1-\delta)\lfloor\mu\rfloor^2 + \delta(\lfloor\mu\rfloor + 1)^2 \\ &\geq \mu^2 + \delta(1-\delta)\end{aligned} \tag{6}$$

Combining (3) and (6), we get:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[SC_{\mathcal{P}}(\mathbf{x})] &\geq 1 - \frac{1}{\binom{K}{2}}\left(K^2\bar{L}_{\mathcal{S}}(\mathcal{P}) - \mu^2 - \delta(1-\delta)\right) \\ &\geq 1 - \frac{1}{\binom{K}{2}}\left[\mu(K-\mu) - \delta(1-\delta)\right]\end{aligned}$$

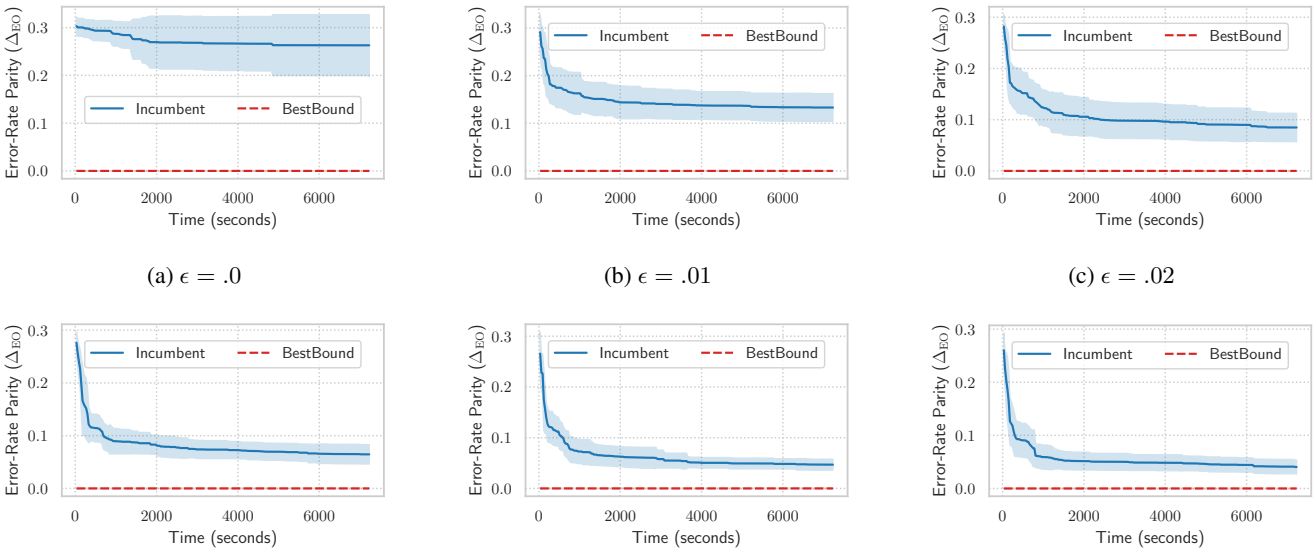
To see that the bound is tight, construct a dataset $\mathcal{S}$ in which a fraction $1-\delta$ of instances are misclassified by exactly $\lfloor\mu\rfloor$ models, and a fraction δ of instances are misclassified by exactly $\lfloor\mu\rfloor + 1$ models. Then $\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}] = (1-\delta)\lfloor\mu\rfloor + \delta(\lfloor\mu\rfloor + 1) = \mu$, so the average empirical risk of the model pool is $\bar{L}_{\mathcal{S}}(\mathcal{P}) = \mu/K$. Moreover,

$\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[r_{\mathbf{x}}^2] = (1-\delta)\lfloor\mu\rfloor^2 + \delta(\lfloor\mu\rfloor + 1)^2 = \mu^2 + \delta(1-\delta)$.

Substituting this in (3), we get:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[SC_{\mathcal{P}}(\mathbf{x})] = 1 - \frac{1}{\binom{K}{2}}\left[\mu(K-\mu) - \delta(1-\delta)\right]$$

which matches the lower bound. Hence, the bound is tight. □



(a) $\epsilon = .0$ (b) $\epsilon = .01$ (c) $\epsilon = .02$

(d) $\epsilon = .03$ (e) $\epsilon = .04$ (f) $\epsilon = .05$

Figure 9: Convergence behavior of FairSLIM across different Rashomon tolerances ϵ on the train set. Although the solver does not converge within the allotted time, the majority of improvement in the objective occurs within the first 30 minutes.

F Self-Consistency Bound for Subgroups

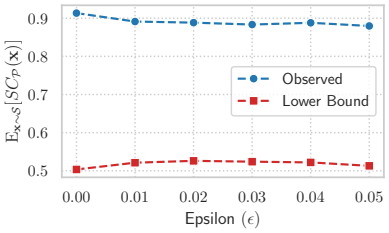
As a corollary of Proposition 2, the lower-bound for expected self-consistency of a finite model pool $\mathcal{P}$ at a subgroup-level is simply:

$$\mathbb{E}_{\mathbf{x}\sim G}\left[SC_{\mathcal{P}}(\mathbf{x})\right] \geq 1 - \frac{1}{\binom{K}{2}}\left[\mu(K-\mu) - \delta(1-\delta)\right]$$

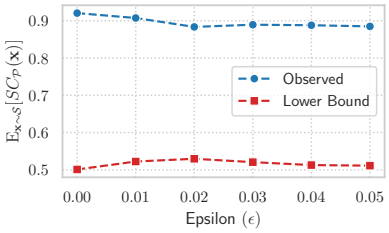
where $\mu = K\bar{L}_G(\mathcal{P})$ and $\delta = \mu - \lfloor\mu\rfloor$. In Figure 10, we compare the observed self-consistency of $\mathcal{P}$ against its worst-case guarantee for all subgroups in our data. We find that observed self-consistency is consistently higher than its theoretical lower bound across all subgroups and all values of ϵ. Thus, despite its tightness, the bound is quite conservative in practice.

G Self-Consistency and Pool Size

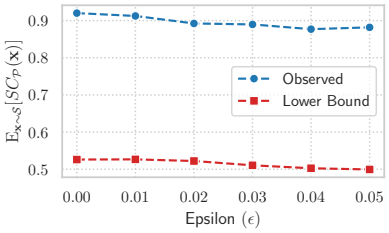
Figures 11 and 12 show the sensitivity of self-consistency of a model pool $\mathcal{P}$ with respect to its size. To produce these figures, we first sample k models uniformly at random from $\mathcal{P}$ for every combination of $k \in \{5, 10, 15, 20, 25\}$ and $\epsilon \in \{.01, .02, .03, .04, .05\}$. Then, for every sample of size k, we compute its average self-consistency on both the train and test sets. We observe that the overall trend that average self-consistency remains substantially higher across values of ϵ than worst case analysis suggests, is invariant to pool size for $k \leq 25$.



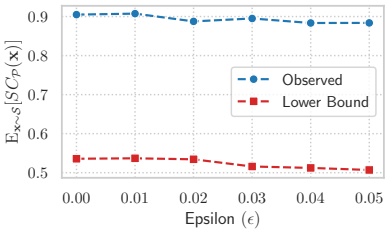
(a) Male, Age<30, National; Size=173



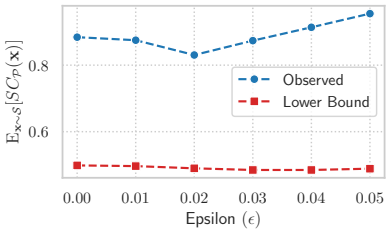
(b) Male, Age<30, Foreigner; Size=142



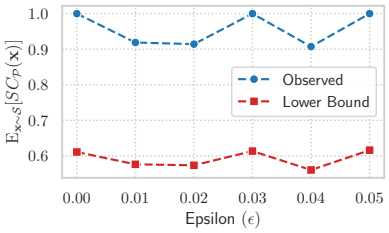
(c) Male, Age≥30, National; Size=982



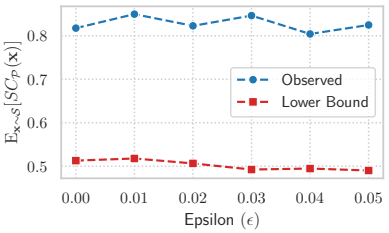
(d) Male, Age≥30, Foreigner; Size=501



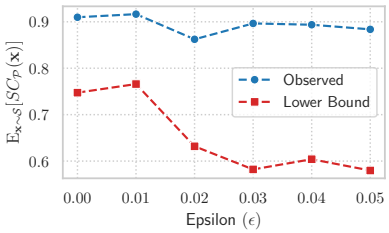
(e) Female, Age<30, National; Size=6



(f) Female, Age< 30, Foreigner; Size=4



(g) Female, Age≥30, National; Size=61



(h) Female, Age≥30, Foreigner; Size=12

Figure 10: Expected self-consistency of the model pool returned by `FairSLIM` on the **test set** at a subgroup-level versus ϵ. While both quantities vary with ϵ, a large gap between the observed values and the lower bound persists throughout. Interpretation for the smaller subgroups should be considered cautiously due to limited sample sizes.

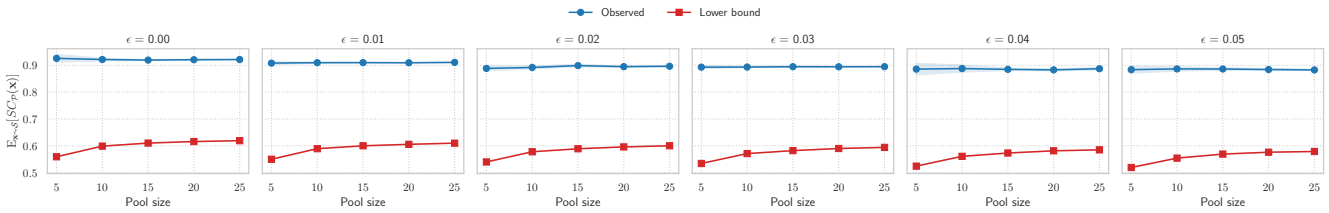


Figure 11: Expected self-consistency of the model pool returned by `FairSLIM` on the **train set** as a function of varying Pool Sizes. The large gap between the observed self-consistency and its theoretical lower-bound is persistent throughout.

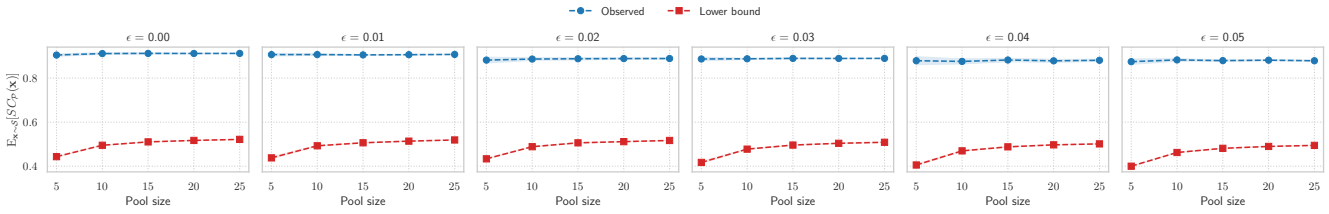


Figure 12: Expected self-consistency of the model pool returned by `FairSLIM` on the **test set** as a function of varying Pool Sizes. The large gap between the observed self-consistency and its theoretical lower-bound is persistent throughout.

Table 4: Overview of the 20 static and 23 dynamic features used in `RisCanvi` for risk assessment.

Feature	Feature Name	Feature Type
F1	Violent Base Crime	Static
F2	Age at the time of Base Crime	Static
F3	Intoxication during the Commission of Base Crime	Static
F4	Number of Victims with Injuries	Static
F5	Duration of the Penalty	Static
F6	Uninterrupted Time in Prison	Dynamic
F7	History of Violence	Static
F8	Age for Start of Criminal or Violent Activity	Static
F9	Increase in the Frequency, Severity and/or Diversity of Crimes	Static
F10	Conflicts with Other Inmates	Dynamic
F11	Non-Compliance with Criminal Measures	Dynamic
F12	Disciplinary Files	Static
F13	Evasions or Escapes	Static
F14	Regression of Degree	Static
F15	Break of Permissions	Dynamic
F16	Childhood Maladjustment	Static
F17	Distance between Residence and Penitentiary	Dynamic
F18	Level of Education	Dynamic
F19	Problems in Employment	Static
F20	Lack of Financial Resources	Dynamic
F21	Absence of Viable Future Plans	Dynamic
F22	Criminal History in the Family of Origin	Static
F23	Problematic Socialization in the Family of Origin	Static
F24	Lack of Family and Social Support	Dynamic
F25	Criminal Friendships	Dynamic
F26	Belongs to Social Risk Groups	Dynamic
F27	Prominent Criminal Role	Dynamic
F28	Victim of Gender Violence (only applicable to women)	Dynamic
F29	Current Family Burdens	Dynamic
F30	Drug Abuse or Dependence	Dynamic
F31	Alcohol Abuse or Dependence	Dynamic
F32	Severe Mental Disorder	Static
F33	Promiscuous Sexual Behavior or Paraphilia	Static
F34	Limited/No Response to Psychological and/or Psychiatric Treatment	Dynamic
F35	Personality Disorder related to Anger, Impulsivity or Violence.	Static
F36	Poor Coping with Stress	Dynamic
F37	Attempts or Behaviors of Self-Harm	Static
F38	Pro-Criminal Attitudes or Anti-Social Values	Dynamic
F39	Low Mental Capacity and Intelligence	Static
F40	Recklessness	Dynamic
F41	Impulsivity and Emotional Instability	Dynamic
F42	Hostility	Dynamic
F43	Irresponsibility	Dynamic