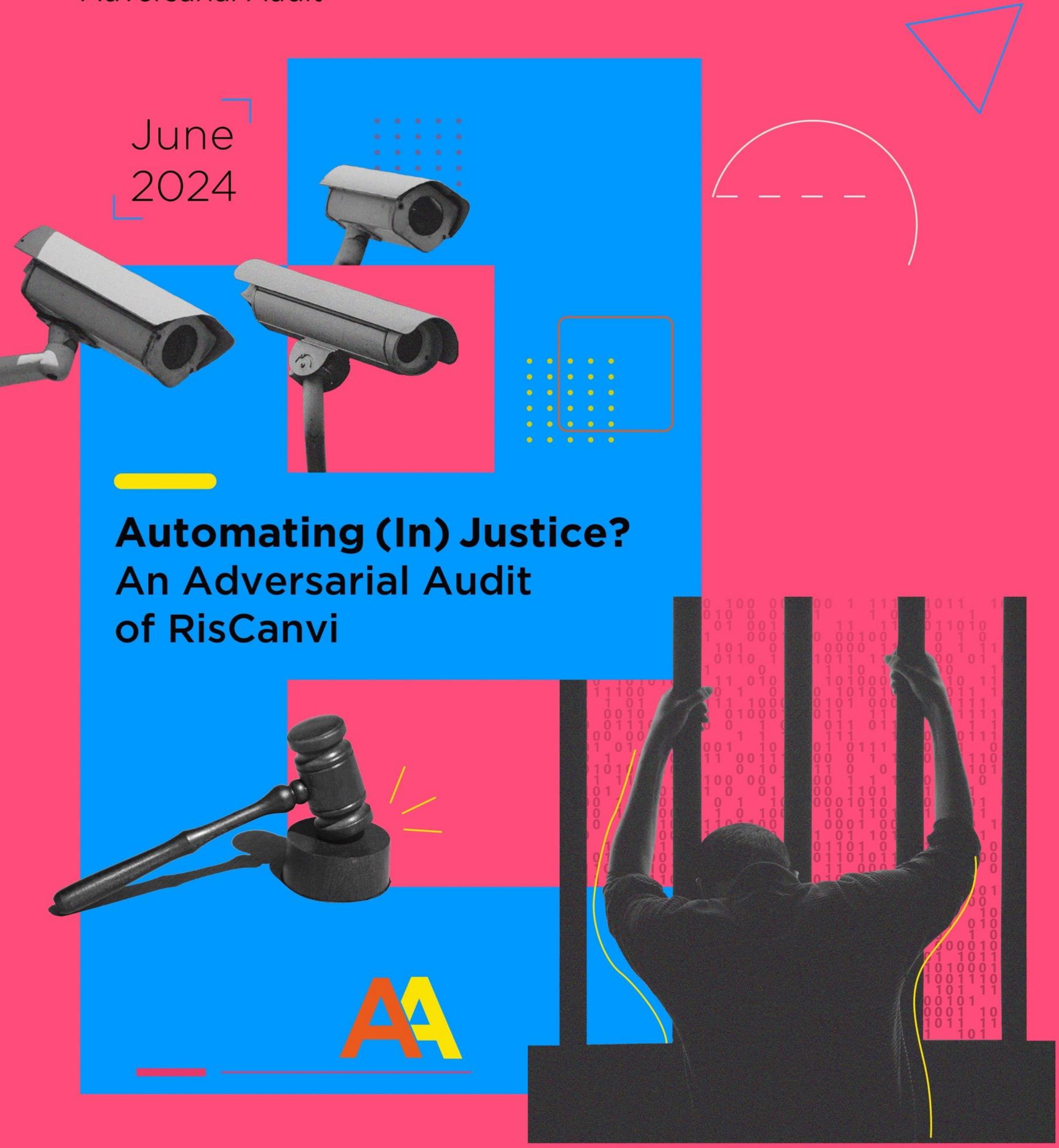
Adversarial Audit

June 2024



Automating (In) Justice? An Adversarial Audit of RisCanvi

eticas Foundation

Irīdia

DIVERSIFAIR
Diversify with Intersectionally Fairer AI

Funded by
the European Union

TABLE OF CONTENTS

TABLE OF FIGURES

To Julia Angwin and her team at ProPublica, who were the first to shed light on the use of algorithms in the criminal justice system.

Executive Summary

Criminal justice has undergone a profound transformation with the integration of predictive algorithms, as exemplified by systems like COMPAS or PredPol, both used in the United States. This paradigm shift has sparked intense debates surrounding **transparency, bias, and the reliability of algorithmic decision-making**.

At Eticas we wanted to contribute to the debate by conducting the **first ever adversarial audit of an AI criminal justice system in Europe**: the RisCanvi tool. In use since 2009 in Catalonia, Spain, and specifically designed for assessing recidivism and violence risk among inmates. RisCanvi wields influence over parole and sentencing decisions, and while some authors have raised concerns about its fairness, accuracy and transparency, most people, inmates, lawyers and court actors are unaware of its existence or inner workings.

In our quest to understand RisCanvi, the Eticas team conducted an Adversarial Audit using a dual-method approach. This involved an Ethnographic Audit, which included interviews with inmates and staff both in and outside the criminal justice system, and a Comparative Output Audit, based on public data on the inmate population and recidivism, and comparing it with RisCanvi's risk factors and risk behaviors.

What we have found with **RisCanvi is a system that is not known by those whom it impacts the most, inmates; that is not trusted by many of those who work with it, who are also not trained on its functioning and weights; that is opaque and has failed to adhere to current regulation** on the use of automated decision-making systems in Spain, where AI audits are required since 2016. Above all, however, our data shows that RisCanvi may not be fair nor reliable, and that it has failed to do what AI should do best: standardize outcomes and limit discretion. Consistent with earlier studies, **we do not find RisCanvi to be reliable, as this would require a clear relationship between risk factors, risk behaviors and risk scores**.

Based on the available data, we conclude that **RisCanvi does not work, and is not currently able to provide the necessary guarantees to inmates, lawyers, judges and criminal justice authorities**.

As with any adversarial audit, our findings are not final. We were not able to access system data, therefore we could not confirm our conclusions. But there seems to be enough data on the table to grant **further scrutiny of the system**. The way things stand, whenever a low-risk inmate engages in recidivism, it is impossible to know if the failure to categorize them correctly is the result of an unavoidable error rate or a feature in an unreliable system. Likewise, when an inmate is denied access to increased levels of freedom due to high-risk, it is currently unclear whether this is a fair decision.

As established in the recently approved **AI Act, using AI in sensitive settings such as the criminal justice system requires an increased level of transparency and scrutiny**, both internal and external, and consistent efforts to inspect and monitor system performance and impact. We do not find this to be the case in the deployment and use of RisCanvi, adding **legal compliance** to the many challenges of this Ai risk tool.

***

This report is our fourth in the Eticas Adversarial Auditing Program. We encourage you to access earlier reports on AI systems covering topics such as gender bias risk assessment, AI-induced inequalities and labor violations in ride-hailing apps, and the impact of facial recognition on people with disabilities. Stay tuned for Eticas future work on blindness and emotion recognition, social media and mental health and the use of AI in housing, banking and social services.

This audit was conducted as part of the DIVERSIFAIR project. DIVERSIFAIR has received funding from the European Education and Culture Executive Agency (EACEA) in the framework of Erasmus+, EU solidarity Corps A.2 - Skills and Innovation under grant agreement 101107969.

Funded by the European Union

DIVERSIFAIR
Diversify with Intersectionally Fairer AI

Key Findings

RisCanvi is a predictive algorithm used to assess recidivism and violence risk and determine access to parole in the Catalan (Spain) prison system. Established in 2009, as of 2022 it impacts 7,713 individuals in the Spanish prison system, according to the latest available data from the Ministry of Justice. The system has **never been audited**, even though Spanish regulations require that all automated systems impacting on individual rights be audited since 2016.

Awareness and **understanding** of the RisCanvi system **vary significantly** among stakeholders. While some professionals are highly familiar with it, others admit to having little to no awareness of it. This is particularly concerning as most of those with no awareness of the existing or functioning of RisCanvi are inmates.

The **inner working** of RisCanvi remains **opaque**, even to frontline staff like psychologists. The variables and weights behind risk ratings are not known to those using the system.

Professionals can influence RisCanvi scores but **rarely do so**, with alterations occurring in less than 5% of cases. This means that in fact RisCanvi is a fully automated (AI) system without meaningful human intervention ("human in the loop").

Inmates **lack legal support** and awareness of the system and their own risk classifications, preventing their meaningful participation or contestation of RisCanvi outcomes.

There is no robust relationship between RisCanvi risk factors, risk behaviors and risk scores. The results of our reverse-engineering efforts show **random correlations between the relevant factors**. This points to a **lack of standardization in the assignation of risk to inmates**.

Static unchangeable **factors** like childhood problems play a **crucial role** in behavior evaluations, leading to bias against certain demographics and inmates that once were children born in difficult environments.

The current RisCanvi structure **does not provide the necessary guarantees of fairness, explainability and robustness** to inmates, lawyers, judges and criminal justice authorities.

Acknowledgments

We express our sincere gratitude to Iridia - Center of Defense of Human Rights for their collaborative role in conducting the interviews, which proved pivotal in comprehending the intricacies of the RisCanvi algorithms. Our deep appreciation extends to the professionals and advocates in the criminal justice landscape who generously shared their invaluable insights during interviews conducted from July to October 2023. The collective contributions of former inmates, social educators, psychologists, a jurist, RisCanvi validator, lawyers, an activist, and representatives supporting incarcerated individuals and their families have significantly enriched the content of this report.

Project team: Adversarial Audits
Research Director: Gemma Galdon-Clavell, Founder and CEO of Eticas
Research Lead: Mat Mastracci, Team Lead Researcher at Eticas
Research Team:

- Luis Rodrigo González Vizuet, Researcher at Eticas
- Miguel Azores, Junior Researcher at Eticas
- Arancha Cienfuegos, Chief of Operations Officer at Eticas

Recommended citation:

Galdon Clavell, G; Mastracci, M., Vizuet, L., Azores, M. (2024). Automating (in)justice? An adversarial audit of RisCanvi", Eticas Foundation.

This audit was conducted as part of the DIVERSIFAIR project. DIVERSIFAIR has received funding from the European Education and Culture Executive Agency (EACEA) in the framework of Erasmus+, EU solidarity Corps A.2 - Skills and Innovation under grant agreement 101107969. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Culture Executive Agency (EACEA). Neither the European Union nor the EACEA can be held responsible for them.

Automating (in)justice?: An adversarial audit of RisCanvi

1. Introduction: AI in the criminal justice system

In recent years, the criminal justice system has witnessed a significant transformation in its approach to addressing risk and recidivism. This shift has been driven by a growing emphasis on data-driven decision-making, with predictive algorithms emerging as a powerful tool for estimating recidivism rates among individuals involved in the justice system. Governments are increasingly turning to algorithms for various purposes in the context of safety and security, including predicting criminal behavior or assessing risk both at the individual and collective level for entire populations and places (Schulberg, 2021).

These tools rely on **statistical calculations** to determine risk (Schulberg, 2021). They use many indicators, such as an individual's criminal history, demographic details, and behavioral patterns, to gauge their likelihood of reoffending.

While the adoption of such algorithms has generated both optimism and controversy, it reflects a compelling effort to enhance the justice system's efficiency and fairness. These algorithms aim to identify effective rehabilitation and intervention strategies for individuals in need. However, the implementation of these algorithms also brings to the fore a host of ethical and legal concerns, demanding meticulous consideration and oversight.

The utilization of actuarial recidivism risk prediction instruments to gauge an offender's dangerousness and the consequent severity of their punishment or the concession of parole by the judicial system has sparked discussions on discrimination against individuals belonging to **socially salient groups**. **Members of these groups** are statistically more likely to reoffend, making the application of these risk prediction instruments potentially discriminatory, unfair, and, in the absence of compelling reasons, morally impermissible (Angwin et al., 2016; New and Castro, 2018; Hannah-Moffat and Montford, 2019; Starr, 2014). Additionally, it is crucial to consider that risk assessment tools can serve as valuable aids in gauging sentencing and recidivism potential when adequately validated, with well-trained staff and third-party accountability mechanisms. The neutrality of these tools is paramount; they should ideally be unbiased, impartial "voices of reason" in legal proceedings. However, the effectiveness of these tools is ultimately **contingent on the impartiality and objectivity of their creators**. If developers harbor preconceived notions about the relationship between gender, race, and criminal behavior, it can perpetuate bias within the tool itself, undermining its intended fairness and accuracy (Center for Digital Ethics & Policy, 2018).

One significant challenge related to these risk algorithms is their tendency to **conflate correlation with causation**. While it is crucial to recognize that a risk score merely represents a correlation, denoting the likelihood of reoffending, the interpretation by legal professionals often deviates from this understanding. The scores are at times erroneously treated as indicators of inherent dangerousness, introducing a critical distortion in how courts perceive an offender's risk. This misinterpretation, as documented by scholars such as Cole and Angus (2002), Bonta and Andrews (2007), and Hannah-Moffat and Maurutto (2010), has tangible repercussions on sentencing decisions. Courts, influenced by this misperception, might inadvertently shape the type, duration, and conditions of sentencing, revealing a nuanced challenge in disentangling the correlation-causation conundrum within the criminal justice algorithmic landscape.

A noteworthy limitation in the realm of risk scores lies in their inherent **inability to discern nuances between various types of recidivism**. The algorithms, by design, do not evaluate the specificity of an individual's likelihood to breach a probation order versus engaging in a violent offense. This lack of granularity in their assessment renders them less effective when it comes to making critical distinctions on the nature and risks of reoffending. Additionally, their predictive accuracy is notably diminished when it comes to forecasting violent recidivism, a vital facet that demands heightened precision within the criminal justice system. As illuminated by scholars such as Campbell and Gendreay (2008), these algorithms tend to exhibit superior performance in predicting low-level criminal behavior while faltering in their capacity to differentiate and accurately predict outcomes associated with more severe, violent offenses. This drawback underscores the complex challenge of fine-tuning algorithmic tools to encompass a broader spectrum of criminal behaviors with varying levels of severity.

An additional criticism of predictive recidivism algorithms in the criminal justice system brings to light a complex interplay between technological advancements, ethical considerations, and the fundamental goals of **rehabilitation**. These algorithms, intended to enhance decision-making processes, can significantly impact the lives of individuals within the criminal justice system. However, a critical examination of their design reveals potential pitfalls that may challenge the core principles of rehabilitation. In the pursuit of fairness and unbiased decision-making. There has been a growing discourse around the ethical implications of including or excluding certain factors in algorithmic risk assessments. One such factor that demands attention is gender. The study conducted by Gavazzi et al. (2005) sheds light on the undeniable fact that females exhibit distinct risks and needs compared to males across various life domains. Consider a scenario where an algorithm, in its attempt to treat all individuals equally, overlooks the unique challenges faced by female inmates. This oversight could result in a failure to address their specific rehabilitation needs adequately. The study emphasizes that neglecting gender-specific considerations in prison settings not only diminishes rehabilitation efforts but may also contribute to an increase in recidivism rates.

To **navigate the challenges** posed by algorithmic recidivism prediction, there is a need for a cross-disciplinary dialogue involving developers, data scientists, analysts, criminologists, and others. This discourse should focus on the legal, socio-political, and discriminatory ramifications of algorithm use. However, even with these efforts, it remains challenging to insulate predictive recidivism models from bias. Predicting recidivism probabilities will always carry an inherent bias due to the data and variables used by these models. The potential solution may lie in designing algorithms that prioritize consistency by adhering to strictly defined legal criteria, albeit with limitations in predicting outcomes like recidivism. Such an approach shifts the focus from concerns about bias, equity, accuracy, and fairness to a more consistent application of the law or policy. However, this approach also has its pitfalls, as it may not adequately consider extenuating circumstances (Fernando Ávila, Kelly Hannah-Moffat & Maurutto, 2010).

The deployment of algorithms for predicting recidivism transcends the realm of academic discourse; it profoundly impacts real-world scenarios, exerting a tangible influence on legal decisions and, consequently, the lives of individuals. This underscores the critical need for an unwavering commitment to principles of fairness, ethics, and the equitable application of the law within the intricate landscape of the criminal justice system. As these predictive tools play a pivotal role in shaping the fate of individuals within the legal framework, it becomes imperative to navigate the delicate balance between technological advancements and the ethical considerations essential for upholding justice. The stakes are high, as decisions guided by algorithmic predictions carry tangible consequences, making it essential to uphold a steadfast commitment to ensuring fairness and justice in the application of these tools.

2. Auditing criminal justice AI

This report is inspired by the groundbreaking work of ProPublica (Angwin et. al. 2016) on one of the first recidivism tools ever implemented and scrutinized, COMPAS. This pioneering work evaluated the algorithm's accuracy particularly across different racial groups in the US. Through an extensive analysis of COMPAS scores, criminal records, and subsequent recidivism data for a vast cohort of over 10,000 individuals who had been arrested in Broward County, Florida, between 2013 and 2014, and methodically comparing COMPAS's predicted recidivism risk categories for each defendant with the actual recidivism rates observed over a two-year span, the ProPublica team found that **Black inmates** were 77% more likely to be erroneously classified as **higher risk individuals**. Conversely, white defendants were more likely to be underestimated as low-risk individuals. The research showed that while mistakes occurred at similar rates for both black and white inmates, the types of errors varied depending on race.

Since then, other AI systems deployed in the context of criminal justice have been reported on and analyzed.[1] However, interest over these systems seemed to peak in 2016-2019, and for the last 5 years scrutiny over these systems seems to have dwindled. Moreover, all known scrutiny and reporting has covered systems deployed in the US, with the rest of the world embarking in similar initiatives without these being reported on or studied. This report hopes to bring new life an attention to the use of AI tools in the criminal justice system, and prompt a global debate on the possibilities, risks and challenges of automating and predicting risk and making life-altering decisions based on algorithmic outcomes.

As algorithms in the criminal justice systems have proliferated for some time, there is abundant literature on their risks. Kehl et. al, (2017) has classified most approaches to AI risk in this context in 3 main groups: **opacity**, **bias/unreliability** and **inaccuracy**. These are not mutually exclusive: in the case of a predictive policing algorithm used to allocate law enforcement resources in a city, the algorithm's opaqueness may make it challenging for both citizens and law enforcement agencies to comprehend how it selects high-crime areas. This lack of transparency can result in public distrust and, in turn, hinder the algorithm's effectiveness. Additionally, if the algorithm is not carefully designed and tested for bias, it may unfairly target marginalized communities, compounding issues of bias and unreliability. Debates around fairness may also emerge when determining the algorithm's criteria for identifying 'high-crime' areas, as different stakeholders may have diverse opinions on what constitutes a fair allocation of resources.

Lack of **transparency** has been highlighted by many authors like one of the main challenges faced by these systems (Pasquale, 2015; Blacklaws, 2018; Abiteboul & Dowek, 2020; Diakopoulous, 2020), even though some argue that the main question is not whether machine learning algorithms are opaque, as they inherently are, but whether they are more or less

[1] See Annex I for an overview of COMPAS, Predpol, CORELS, ORAS and others

opaque than human decision-making in sentencing and criminal justice (Chiao, 2022),. Others (Ryberg & Petersen, 2022) point to their impact on **due process** and whether the use of opaque AI tools may obscure the very principles upon which justice systems are built – fairness and justice.

As for **bias**, the COMPAS case described above is a good example of how bias may be captured and amplified by AI systems (Larson et al., 2016).. These advanced algorithms, for all their intricacies, often find themselves entangled in the web of inherent biases inherited from conventional risk assessments. Their reliance on extensive datasets, often reflecting the inequalities woven into society's fabric, coupled with the intricate process of data extraction, can paradoxically give birth to new biases, unintentionally perpetuating pre-existing disparities (Ávila, Hannah-Moffat, & Maurutto, 2021). This might well extend to **racial discrimination** (Huq, 2019)

Finally, AI systems may also suffer from **technical inaccuracies**. The fluidity and subjectivity (Završnik, 2018) surrounding the interpretation of metrics and fairness indicators make this a very clear challenge. The lack of universally agreed-upon definitions and indicators for many of the metrics used in risk and fairness assessments exacerbates the problem., This inherent variability sparks ongoing debates about the fundamental principles that should govern algorithmic decision-making in diverse contexts (Plesničar and Šugman Stubbs, 2019). In the meantime, people's access to freedom continues to be determined by an increasing number of AI systems.

An additional issue of concern for some authors is the growing role of **private companies** in providing public services, including AI tools, to the criminal justice system. The move towards privatization enables these companies to shield their algorithms under the banner of trade secrets and intellectual property protections (Wexler, 2018), thereby rendering inmates and those representing them unable to ensure the accuracy of risk score results (Carlson, 2023). Also, algorithmic secrecy is sometimes justified by the notion of "public interest" (Ryan, 2020). The argument goes that in the pursuit of public safety, certain details about the functioning of these algorithms are better left undisclosed. This should be at least controversial, as a fundamental pillar of the criminal justice system is the transparency of the codes that determine what constitutes a crime and what are the elements that contribute to sentencing. Claiming that potential criminals could game the law/algorithm if they knew how it works goes against the very basis of modern law in democratic societies.

3. Auditing RisCanvi

RisCanvi, which stands for "risk change" in Catalan, is a multi-level risk assessment protocol developed in Catalonia to assess and manage the risk of violence and recidivism among inmates. The genesis of RisCanvi can be traced back to 2009 when the Catalan Department of Justice, concerned about the rise in violent recidivism among released offenders, created an expert group that recommended the creation of a risk assessment protocol for managing recidivism, particularly with dangerous offenders. Responding to this call, a collaborative effort led by a university professor and the Group of Advanced Studies in Violence resulted in the development of RisCanvi (Digital Future Society, 2023).

Initially RisCanvi was primarily aimed at estimating the risk of re-offense among specific categories of offenders, such as murderers and sex offenders, as they approached the end of their sentences (Andrés-Pueyo et al., 2017). However, the evolving needs of the Catalan prison system led to a significant expansion in RisCanvi's scope. Over the years, it transformed into a multi-level risk assessment protocol encompassing not only violent crimes but also behaviors related to self-directed violence, intra-institutional violence, general reoffending, and potential breaches of prison furlough or parole (Jiménez Arandia, 2023). Indeed, the RisCanvi protocol now encompasses five key behavioral aspects:

- Self-Directed Violence (VA). This category includes assessing the risk of self-injury, suicide attempts, and self-harm among inmates.
- Intra-Institutional Violence (VI): RisCanvi evaluates the potential for violent behavior or aggression within the prison, directed towards fellow inmates or staff members.
- General Recidivism (GR): In addition to violent crimes, RisCanvi assesses the likelihood of inmates committing any type of offense upon release, whether violent or non-violent.
- Violent Recidivism (VR): That is the penitentiary re-entry for a violent crime committed in the community, which may have occurred after completing the sentence, during a release permit or in any other situation of the intern who is unable to obtain release.
- Breach of sentence (RC): This category involves predicting the probability of inmates failing to comply with the conditions of their sentences, such as not returning from authorized leaves.

3.1 Adversarial auditing methodology

In auditing RisCanvi, we've used a socio-technical approach informed by our Adversarial Auditing Guide[2]. This means that we have combined the methods and results of an ethnographic (qualitative) audit and a comparative output (quantitative) audit to better understand the system, its logics and impact.

We have also collaborated closely with a civil society organization, Iridia, that defends the civil and political rights of inmates.[3] Whenever possible, we rely on local, experienced organizations to ensure that our work with vulnerable communities follows the highest standards of respect, empathy and true collaboration, and also as a means to train and empower these organizations to have a voice in the AI debate. Iridia has been crucial to the development of the methodology, identification of key actors, trust-building and conducting the interviews.

In July-October of 2023, a series of 18 **interviews** were conducted with a diverse array of professionals and advocates within the criminal justice landscape[4]. Six former inmates shared their perspectives on July 6, 18, and four times on September 30, 2023. Two social educators in prison were interviewed on July 6 and September 28, 2023, respectively. Three

[2] See, Eticas (2023). Adversarial Algorithmic Auditing Guide. Association Eticas Research and Innovation. https://eticas.ai/case-study/adversarial-algorithmic-auditing-guide/

[3] https://iridia.cat/en/

[4] The full list of interview questions can be found in Annex 3: Interview Questions.

psychologists specializing in various aspects of the penitentiary system participated in interviews held on July 12, 14, and 17, 2023. Further insights were gleaned from one jurist and validator of RisCanvi on July 18, 2023. Four lawyers, including one previously employed in a prison, were interviewed on September 23, 21, and October 1, 2013, along with a lawyer engaged in doctoral research on September 12, 2023. Finally, one activist advocating for the rights of incarcerated individuals and one representative from an entity supporting the relatives of incarcerated individuals in Catalonia provided their perspectives on September 12 and 30, 2023, respectively

The qualitative data collected through observation, interviews, and surveys served as a crucial lens into the tangible repercussions of RisCanvi, unveiling insights into awareness, functionality, disparities, decision influences, and transparency challenges. The qualitative work enabled us to identify clear gaps in comprehension among inmates, sparking concerns about the algorithm's precision, impartiality, and practical applicability.

Simultaneously, we analyzed publicly available data on 3,600 individuals released from Catalan prisons in 2015. As the data and variables of RisCanvi are not transparent and the system has never been audited, we had to rely on public, related datasets.

Investigating relationships between RisCanvi's 43 risk factors and recidivism outcomes, we challenge fundamental assumptions about risk differentiation and weighting. Initial cluster analysis reveals the difficulties of RisCanvi in categorizing the population into distinct risk groups, while deeper exploration exposes inconsistencies between risk classifications and expected distributions. Through rigorous multivariate regression modeling, we gauge the predictive validity and scrutinize potential biases.

3.2 Data overview and limitations

Of the very few publicly available datasets related to RisCanvi factors and evaluations, only two contain recent enough information to be relevant to our study. Both are published by the Center of Legal Studies and Specialized Training (CEJFE).[5] One of these datasets, Catalan Prison Recidivism Rate in Parole 2020 (CPRR), does not include all the RisCanvi factors[6] so we had to discard it. The other dataset is the Catalan Prison Recidivism Rate (CPRR)[7] focused on individuals from the inmate population who were released via permanent release, conditional release, or suspension of sentence in 2015 and were then tracked until December 31, 2019, constituting a follow-up period ranging between four to five years.[8]

The CPRR dataset includes 379 variables with 3,651 observations, incorporating RisCanvi's factors. However, only 2,726 observations contain at least one RisCanvi variable (risk factor), and so the number of useful observations is significantly reduced once the data with missing values is discarded, as shown in Figure 1.

[5] See https://cejfe.gencat.cat/ca/recerca/opendata/presons/taxa-reincidencia-2020/index.html
[6] See https://cejfe.gencat.cat/ca/recerca/opendata/presons/reincidencia-llibertat-condicional
[7] See https://cejfe.gencat.cat/ca/recerca/opendata/presons/taxa-reincidencia-2020/index.html
[8] See https://cejfe.gencat.cat/ca/recerca/opendata/presons/reincidencia-llibertat-condicional

Figure 1: Dataset observations

RisCanvi Evaluations	Observations with at least one valid variable [9]	Observations with information in all variables
First Complete [10]	1,921	308
First Screening	805	791
Second Complete [11]	805	801
Second Screening[12]	1,921	365

In order to complete the sample, we used alternative variables within the same database[13]. This approach allowed us to go from the figures shown in Figure 1 to a total of 1,889 useful observations to be used to **reverse engineer the black box of RisCanvi and find out which of its factors have higher or lower impacts on the automated risk classification**. This adversarial auditing process, conducted through statistical regression analysis, allows us to openly and transparently question whether the ways in which RisCanvi classifies risk are fair and acceptable in the context of a prison system designed to foster the rehabilitation of offenders.

3.3 Existing studies

The existence of RisCanvi is not widely known or discussed in the region. The research team that developed the system has published works (Karimi-Haghighi & Castillo, 2021, 2022) describing a good predictive accuracy of the system, with Area Under the Curve (AUC) values ranging from 0.79 to 0.87, and also highlighting the role of RisCanvi in standardizing risk assessment, facilitating individualized management and improving information sharing between relevant staff. Indeed, some interviewees highlighted how a strength of RisCanvi "*is that in some way, it [established] some positive criteria that are based on many studies and meta-analyses.*" (Psychologist, human rights organization) Others pointed to greater consistency across staff: "*A very important strength is that all professionals in the organization focus attention on the same criteria... I think this makes us fairer in decision-making processes.*" (Prison administrator). We did find that most interviewees struggle to find strengths in the system, with one psychologist stating: "*There's a moral strength, which is at least they tried.*" This points to the crucial need to inspect RisCanvi not only from a technical perspective, but also organizational.

Most of the literature on RisCanvi points to its shortcomings. Jiménez Aranda (2023), for instance, has addressed the **lack of independent studies** appraising the effectiveness and impact of RisCanvi on inmates. Alemán Aróstegui (2023), in his assessment of the integration of RisCanvi into Penitentiary Law Practices has highlighted a worrisome trend towards an over-reliance on algorithmic outcomes, which undermines the principles of rehabilitation and social reintegration in favor of risk minimization. He argues that this approach risks reducing incarcerated individuals to mere objects of assessment and control, neglecting their rights

[9] At least of the variables contains some information.
[10] Composed of 43 variables identified with the IDs 207-249.
[11] Composed of 10 variables identified with the IDs 303-312.
[12] Composed of 10 variables identified with the IDs 260-302.
[13] Annex 5: Alternative variables for RisCanvi factors" shows the correspondence between each RisCanvi Complete factor (43 factors) and its respective alternative variable.

and perpetuating a punitive rather than rehabilitative approach to penitentiary decision-making. The author also argued that despite its purported intent to usher in a risk-focused paradigm, RisCanvi fails to consistently address disparities, as the influence of risk levels significantly fluctuates based on the nature of the committed crime. This uneven impact perpetuates an unjust system wherein individuals convicted of different criminal typologies face divergent consequences. Moreover, the protocol's sway extends to shaping the legal status of those in custody, casting them more as objects of risk management rather than recognizing their status as subjects of law. The deficiency in transparency, coupled with these legal criticisms, raises ethical concerns, and underscores the imperative for a more open approach to align with legal standards.

The issue of the system's **predictive accuracy** has also been highlighted. The Center for Legal Studies and Specialized Training (CEJFE) report on recidivism from 2014[14], established the algorithm's sensitivity at a high 77% (meaning that out of 100 inmates considered medium or high risk, 77 did reoffend). and the specificity at 57% (meaning that out of 100 inmates who did not reoffend, 57 were labeled low-risk). However, Lucía Martínez Garay (2016) found that the CEJFE study does not clearly differentiate sensitivity (percent of recidivists correctly predicted as high/medium risk) from the positive predictive value (percent recidivated among those predicted high/medium risk). The **positive predictive value of RisCanvi, based on Martinez Garay's data, is only 17.9%.**

Another author, Gimeno Beviá (2023), has highlighted the risks associated with **false negatives** in the RisCanvi algorithm, warning against its sole reliance for parole decisions. The author stressed concerns about transparency, particularly regarding the algorithm's functioning and criteria, which may hinder prisoners' legal representatives from effectively challenging its outcomes.

More recent studies such as Karimi-Haghighi and Castillo (2021, 2022) have corroborated the system's high predictive accuracy, but also hinted to some possible challenges. Their 2022 findings uncovered a tendency for the tool to overstate the risk of violent recidivism, and also potential bias against specific demographic and background factors such as sex, nationality, age, birthplace, age of first recorded criminal activity, mental disorders/substance abuse, and socioeconomic status. In their 2021 analysis the authors juxtaposed RisCanvi with a machine learning model, which marginally outperformed RisCanvi in accuracy. Nonetheless, both methodologies exhibited biases, highlighting the imperative of meticulous calibration and fairness considerations in the deployment of such tools within criminal justice frameworks.

3.4 How RisCanvi works

The RisCanvi risk assessment protocol is informed through human input collected by various professionals, including psychologists, legal criminologists, social workers, and social educators. Inmates undergo biannual assessments based on interviews, and the collected data is inputted into a computer program featuring a deterministic statistical model. A team of one hundred individuals act as "validators" by checking if the information entered into

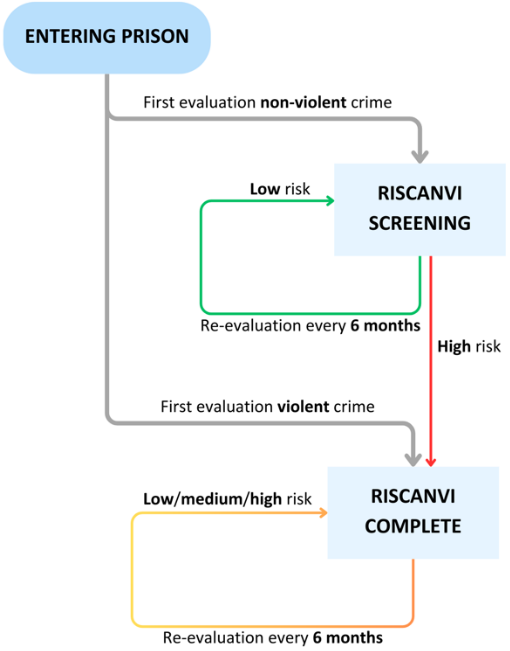
[14] See https://cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2015/taxa_reincidencia_2014/tasa_reincidencia_2014_cast.pdf

RisCanvi by other professionals matches the risk level determined by the algorithm. The outcome of the risk assessment is represented through a color-coded system: red (indicating high risk), yellow (medium risk), and green (low risk). This dynamic assessment is updated biannually to accommodate changes in an individual's circumstances or behavior.

We have found that the exact calculations behind the final "low, medium or high" risk ratings are **opaque**. An educator we interviewed admitted that "*We were not explained [how it works], and we do not know it.*" A psychologist agreed: "No one knows exactly, except the tool designers. Not even management." The risk ratings are not accompanied by any kind of explanatory report, as highlighted by a lawyer: "You just know the headline, if it's high, medium or low, nothing more." Professionals can manually override the rating if they disagree but must provide justification (which is not needed if the algorithmic outcome is accepted). So while RisCanvi is seen as "essential" in parole and sentencing decisions, frontline staff and prisoners themselves lack transparency into the inner calculations and processes behind RisCanvi's final risk determinations.

RisCanvi determines a risk score following a dual assessment approach, incorporating both screening and complete evaluations, with distinct sets of risk factors tailored to enhance its precision. The initial screening, RisCanvi-S, applies 10 key risk factors upon an inmate's entry a penitentiary center, and classifies them as either "low risk" or "high risk." For individuals designated as "high risk" during the RisCanvi-S evaluation, a more detailed examination ensues through a more comprehensive assessment, RisCanvi-C, which includes an analysis of 43 risk factors. The following figure summarizes the workflow of the RisCanvi's protocol with all the relevant details:

Figure 2: RisCanvi Protocol Workflow


ENTERING PRISON
First evaluation non-violent crime
Low risk
RISCANVI SCREENING
Re-evaluation every 6 months
High risk
First evaluation violent crime
Low/medium/high risk
RISCANVI COMPLETE
Re-evaluation every 6 months

RisCanvi-C distinguishes between 16 static and 27 dynamic elements,[15] these factors collectively form the backbone of RisCanvi's dual assessment methodology, as illustrated in Figure 2, and are divided into four macro areas representing a distinct domain of influence on an individual's risk of recidivism. The criminal/penitentiary domain incorporates factors related to an individual's criminal history, behavior within the penal system, specific characteristics of the committed offenses and includes variables such as the nature of the violent index offense, age at the time of the offense, history of violence, and disciplinary records. The biographical domain focuses on personal background and life circumstances, examining elements that can have long-lasting effects on behavior, including variables such as poor childhood adjustment, educational level, employment-related issues, and the absence of viable future plans. The family/social domain explores the impact of familial and social connections on an individual's risk profile with variables like the criminal history of family members, difficulties in socialization or development within the family, lack of family or social support, and association with criminal or antisocial friends. Finally, the clinical domain addresses mental health and clinical factors that may contribute to an individual's risk of reoffending covering issues like substance abuse, mental disorders, problematic sexual behavior, and the response to psychological or psychiatric treatments.

Dynamic risk factors, unlike static elements, are characterized by their capacity to evolve over an individual's life course. Examples include unemployment and peer group influences, with the pace of change varying between stable dynamics (e.g., personality traits) and acute dynamics (e.g., drug use) that may fluctuate daily (Coid et al., 2016). RisCanvi, akin to OASys (see Annex 1), integrates a comprehensive assessment model that incorporates both static and dynamic risk factors. In contrast to systems like LSI-R and Static-99R, which predominantly rely on static elements. In RisCanvi, 62.8% of the factors evaluated are dynamic and 37.2% are static.

Within 4 weeks after the initial completion of RisCanvi-S (R-S) once an inmate enters a correctional facility or prison, a multidisciplinary team assigned to the inmate convenes a meeting to review the case, supplementing evidence for each R-S factor and rendering judgments on their presence or absence. This process is conducted under the supervision of a validator, ensuring the integrity of the assessment. If the screening indicates high risk or presents a potential special case, the comprehensive RisCanvi Complete (R-C) protocol is activated.

The R-C protocol encompasses a thorough examination of 43 risk factors spanning criminal, personal, social, and clinical domains. Each domain is scrutinized by professionals with expertise in the relevant field, such as criminologists assessing criminal history and psychologists evaluating clinical items. Information is collated through interviews, file reviews, and collaboration with other services. In subsequent team meetings, all members engage in extensive discussions regarding the evidence, collectively determining the presence, probability, or absence of each factor through consensus. The validator ensures that ratings align with scale criteria before utilizing the e-RisCanvi software, integrated with the inmate management system, to generate automated risk level results (low, medium, high) for the five assessed criteria. Teams retain the authority to override the rating, accompanied by a justification, when deemed necessary.

[15] A detailed list of factors can be found in Annex 4: Risk Factors of RisCanvi Screening (RisCanvi-S) and Complete Assessment (RisCanvi-C) Versions

Subsequent re-evaluations occur every 6 months. In cases where critical events, such as self-harm or aggression, emerge, more frequent assessments may be conducted. The e-RisCanvi software serves as a repository of evidence, rating justifications, and results. Throughout the process, the validator assumes a pivotal role, overseeing the entire process from evidence gathering to final risk rating. The validator's final validation officially concludes the assessment.[16]

Figure 3: Target population

Within the confines of Catalonia's prison system[17], a substantial and diverse population of **7,713** individuals, as of 2022, the latest available data from the Ministry of Justice[18], represents a complex mosaic of **Spanish** (**3,949**) and **foreign** nationalities (**3,764**). RisCanvi is used on pre-trial detainees, convicted prisoners, those in closed and open regimes, and even individuals on temporary parole or furloughs. The only exception are juveniles under the age of 18, who are governed by a separate youth justice system.

The inmate population has been steadily decreasing in the region since 2020.

3.5 RisCanvi's impact on prison decisions

RisCanvi plays a crucial role in determining outcomes such as granting or removing parole and conditional day-time release in the prison system. According to an educator we interviewed, "*For open prison regimes, a person with medium RisCanvi risk will find it challenging. A person with very low risk will have it easier to achieve an initial third grade*." This sentiment underscores the real-world impact of RisCanvi on the possibilities afforded to individuals based on their automated risk assessments and **refutes the idea that RisCanvi is "just a tool"**. As revealed by a prison psychologist "For any issue implying exits, benefits like third grade, conditional release, furloughs, it is obligatory and binding to conduct the RisCanvi [assessment]. If the RisCanvi [assessment] is not conducted, the decision for furlough, conditional release, or third grade does not happen." A researcher we interviewed reinforced this idea, stating that "*In theory, it depends on the people making that decision, but in practice, it conditions it.*" The researcher further illustrated the tangible impact, stating, "*For example, a low risk will not hinder granting a 48-hour furlough, but faced with a high risk, it is very difficult they will grant 48 hours.*"

In our interviews, we aimed to explore further how the interaction between the technical system and human input interrelate, both formally and informally. An educator pointed out the potential to "*soften*" scores slightly by highlighting positive aspects when adding data to the system, but also underscored the impossibility of omitting negative background factors, the static elements of RisCanvi, which could not be changed. Echoing this sentiment, a

[16] The details regarding the RisCanvi risk assessment process and protocols are based on information provided in the 'Manual d'aplicació del protocol de valoració RisCanvi' (RisCanvi assessment protocol application manual) available at https://justicia.gencat.cat/web/.content/home/ambits/reinsercio_i_serveis_peni/manual-aplicacio-protocol-avaluacio-riscanvi.pdf

[17] In Catalonia there are currently 11 prison facilities, see http://www.prisonobservatory.org/upload/PrisonconditionsinSpain.pdf

[18] See https://www.idescat.cat/indicadors/?id=aec&n=15859&lang=en

psychologist from a human rights organization supported the notion that risk scores likely remain unchanged, pointing to the importance of providing context through additional reporting. However, a prison subdirector stated that professionals have a distinct and specific role, limited to inputting evidence, while it is external validators that assess risk factors.

This interviewee added that even though teams can adjust the final rating if they disagree with it, such alterations require consensus among the treatment team and must be thoroughly justified. According to a prison psychologist, **human changes in risk levels are rare**, estimated to occur in "*less than 5% of cases,*" even though apparently, they are encouraged. A researcher interviewed also stated that while protocols to change algorithmic outputs exist, these are hardly ever used, and a lawyer expressed skepticism at the actual possibility to alter risk levels. In essence, although protocols theoretically allow for score adjustments, real-world instances suggest such changes are uncommon. Indeed, in 2021 the local press reported that "*RisCanvi's final score has only been contradicted 3.2% of the time*," (La Vanguardia, 2021).

This data also points to a more final and automated role of RisCanvi than the authorities would like to admit, which contradicts the idea of the algorithmic system being an advisor or contributor to human-ed decisions. Alemán Aróstegui's (2023) extensive research with key actors involved in the decision-making processes affected by RisCanvi confirms the overwhelming influence of RisCanvi's algorithmic outcomes on final decisions and uncovers a complex interplay between the algorithm's perceived objectivity and the practical challenges encountered by decision-makers when contesting its outcomes.

Altogether, our research and other studies show how without proper procedures to define and implement meaningful human supervision, **the requirement for a human in the loop withers away in practice through a combination of lack of awareness, training, enforcement and procedures that make contradicting AI systems more onerous than abiding by their outputs**.

This is particularly relevant in the context of increasing AI regulation, as the role of the human in the loop can sometimes establish the difference between fully automated systems, subject to high-risk precautions, and advisory systems that may not have the same transparency, accountability and explainability requirements.

3.6 Who Knows RisCanvi?

A common theme in the adversarial audits conducted by our team is the lack of awareness of the mere existence of automated or algorithmic systems in decision-making by those impacted by them.[19] We found that while former inmates had varying degrees of awareness of the existence of RisCanvi, none of them were aware of its existence while they were in prison. One former inmate admitted, "*I didn't know about it during my time in prison,*" highlighting the system's limited reach or visibility among incarcerated individuals. Another former prisoner mentioned, "*I don't know what RisCanvi is,*" reflecting the widespread unawareness among this demographic. Another former inmate shared that: "*You don't know*

[19] See, for instance, Eticas. (2022). The External Audit of the VioGén System. Association Eticas Research and Innovation. https://eticasfoundation.org/gender/the-external-audit-of-the-viogen-system/

the questions asked will populate a database translated into parameters. No one tells you anything… you answer innocently without realizing." An activist working on inmate rights conceded "*I absolutely don't know what RisCanvi is,"* while a lawyer admitted, "*I first learned about it in a specialized course.*"

The experience of the inmates was confirmed by other professionals. A prison psychologist corroborated that **most inmates** are **unaware** of the tool's usage and implications at all: "*Very few inmates know they are evaluated for RisCanvi… they do not know that this algorithm decides on the quality and circumstances of whether they will be granted more furloughs.*" Similarly, a legal expert validated that "*they never at any point have access to what is being asked*" in the system's background calculations. An academic researcher concurred "*most of the inmates do not even know what RisCanvi is.*" as "*They do not explicitly inform the inmate this test is for this purpose.*"

We found the experience to be different for professionals. According to an educator from a penitentiary, "*It's something routine for any professional in Catalonia*." A psychologist in a penitentiary also acknowledged that "*professionals are obliged to use this tool,*" while another psychologist working in a human rights organization mentioned "*I mostly study it rather than apply it.*"

These contrasting experiences indicate differing levels of understanding and familiarity with the system among professionals, advocates and those impacted by the prison context. They point to a dynamic we have also identified in other adversarial audits: **those directly impacted by automated decision-making systems are the least aware of their existence and in a weak position to uphold their rights** in AI contexts. The activists and lawyers that defend those impacted by such systems are also highly unaware of their existence and inner workings, pointing to **information and power asymmetries** made worse by the introduction of automated systems.

We delved into this issue by asking different stakeholders about access to legal support for inmates during the RisCanvi assessment. A psychologist from a human rights organization stated definitively that to her awareness legal support on RisCanvi risk levels did not exist, "*no, or not that I know of.*"

These implementation choices have direct impacts on issues related to explainability, trust and recourse, and therefore on the legitimacy of AI systems used in the criminal justice system. In the case of RisCanvi, inmates do not even know what their risk level is. As confirmed by an educator, "*Do they tell you*: *'you have a medium risk'? Do they tell you that as an inmate? No, no, they don't.*" Without transparency around the tool determining their futures, inmates feel powerless and unable to meaningfully participate. As one former inmate described: "*At no time are you informed. If you're not interested in the topic, you don't even find out [about] what it is… It's a question that's cooked internally for their own interests.*" As one former inmate stated: "*In general, inmates never trust anything that comes from the prison administration, because in the end it's what confines you, crushes you."* Another commented: "I don't trust it… [RisCanvi] does not tell you anything until you see the person on the outside."

Interestingly, in the case of RisCanvi the inability to build trust on the system has also impacted on its legitimacy among corrections staff and legal experts. A prison psychologist called it: "*An institutional scam… because RisCanvi's own creators as well as those responsible for forcing us to use it know perfectly well that it does not work.*" The psychologist also added that "*specific* *concerns include reliance on outdated risk factors, the high rate of false positives upwards of 95% in some studies, and predictive validity as low as 13%. Without staff confidence, the system lacks credibility*". An attorney commented: *"I don't believe a tool, an algorithm, can predict human behavior... although it collects a lot of information and the result it gives apparently seems objective, we are talking about people*."

It seems remarkable that a system designed to improve the prison system and operational for 15 years would generate such distrust among all involved stakeholders, pointing to a **lack of evaluation studies, participation mechanisms, and robust implementation plans**. As frequently identified in AI policy implementation, the inability to incorporate socio-technical aspects and monitoring of impacts often leads to the continued use of very problematic technical solutions.

3.7 RisCanvi's fairness and reliability

While the interviews conducted for this study allowed us to explore issues related to the functioning, awareness and legitimacy of the system, the data analysis is a door to RisCanvi's fairness and reliability as a prediction tool in the context of the criminal justice system and the rights of the inmate population.

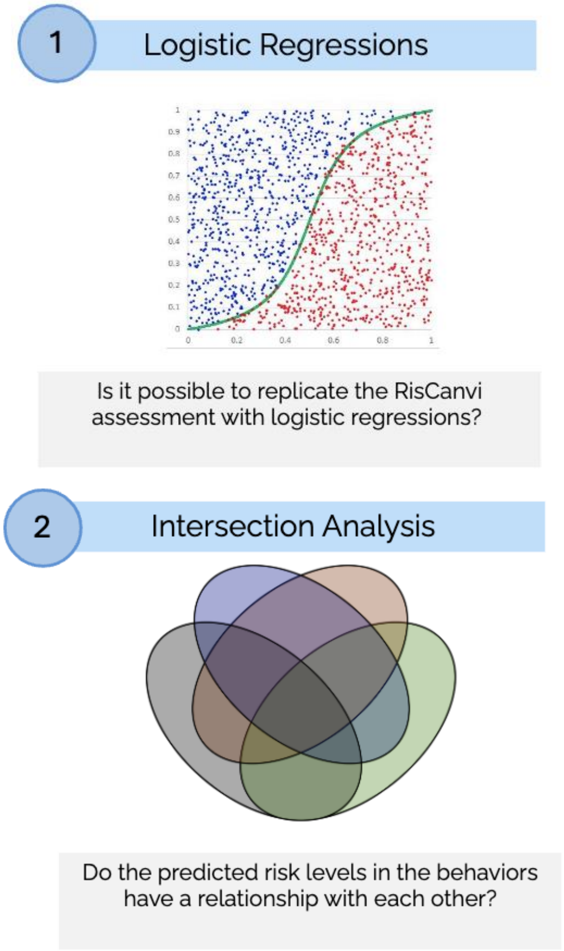
Indeed, many of our interviewees mentioned issues related to fairness in RisCanvi. Several believed that "**foreigners**", "**young inmates**", and **those "who committed crimes at a young age"** tend to receive **higher RisCanvi risk scores**. As one psychologist explained, these groups often lack social stability factors so "*static factors*" like "*your age at first offense*" weigh them down despite good behavior. An educator agreed that those "*with **substance abuse** have a much higher elevated risk level.*" She added that "*it does not differentiate by type of crime, but rather social structures*" - those lacking family ties or education perform worse regardless of offense. However, one prison subdirector asserted firmly that "*there are no biases based on population*" since RisCanvi simply estimates risk factors which can be present across groups. A prison psychologist critiqued that the tool was "*created for violent crimes*" which are a minority - it often underestimates recidivism risk for "*sexual abusers or scammers*", labeling them low-risk simply because they lack the antisocial traits of concern for violent offenders. In the end, while the official stance is that the algorithm itself does not introduce demographic biases, frontline staff widely report systematic disadvantages faced by young, foreign, socially unstable inmates based on the risk factors prioritized within the system. The suitability of applying the same tool across diverse offense types was also questioned.

Additional concerns included contributing to tunnel vision, limited reliability and predictive validity, and inadequate implementation conditions like time pressures. An attorney summarized: "*You cannot use people as simple statistical numbers.*" By engaging in a statistical analysis of RisCanvi data, we aim to find out whether the perceptions expressed by the stakeholders interviewed hold true.

When conducting an adversarial audit, we leverage data to reverse engineer a system. As we have access to limited data points, most of which refer to a system's impacts, the starting point is always to see what outcomes can tell us about inputs. **Reverse engineering, which is the heart of an adversarial audit, is like being able to taste a cake and having to find ways to write the list of ingredients and cooking steps**. In the case of RisCanvi, the challenge was to determine the relationship between risk factors with risk levels, and the impact of risk factors on how risk-level is predicted for each of the different RisCanvi behaviors. Thanks to the existing literature, we had access to a list of ingredients. In this case, the key was to establish which of those ingredients (behaviors) played a role when determining an inmate's likelihood to re-engage in criminal activity.

We conducted several statistical experiments to reverse engineer RisCanvi, which are summarized and sequenced in Figure 4 below.

Figure 4: Analysis developed as part of the adversarial audit.



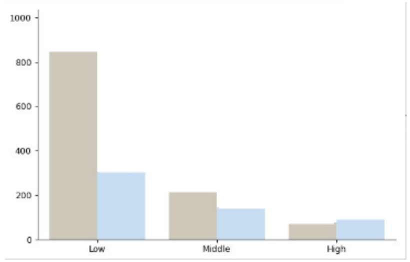
1 Logistic Regressions

Is it possible to replicate the RisCanvi assessment with logistic regressions?

2 Intersection Analysis

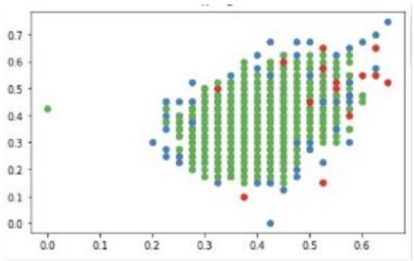
Do the predicted risk levels in the behaviors have a relationship with each other?

3 RisCanvi's Factors Prevalence



Are there factors with higher or less impact on the risk classification and RisCanvi assessments?

4 Spectral Clustering

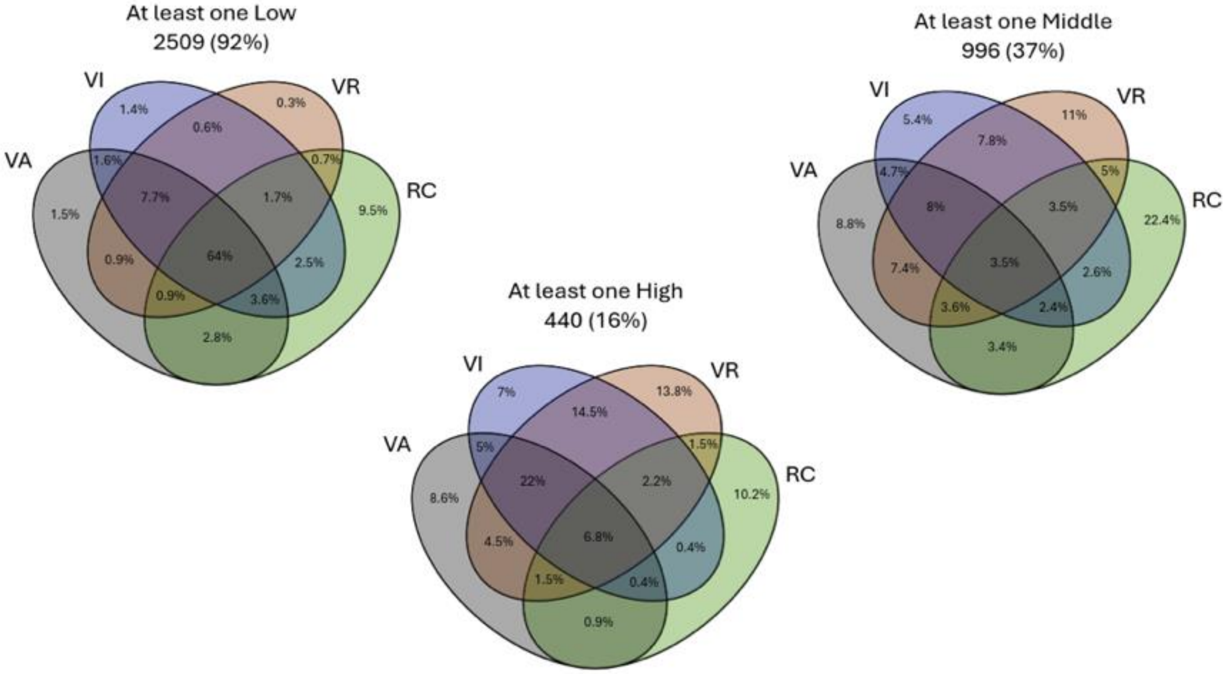


Does the existing data on the RisCanvi factors statistically determine the different behaviors assessed by the protocol?

The first step was to conduct logistic and multinomial regressions to replicate the functioning of the RisCanvi tool (Annexes 6 and 7), as the use of these methods had been confirmed during the interviews we conducted. We used these methods on our dataset but found high standard errors for all RisCanvi factors, general fitting problems and errors in predicting risk levels for each behavior. Overall, **we found no significant relationship between RisCanvi factors and behaviors**, which was concerning. If different risk behaviors have different weights, the dataset should reveal some strong relationships pointing to the weights used. As in, if the cake is very sweet, we would assume that sugar is one of the main ingredients used. In this case.

As the regression models did not shed light on RisCanvi's algorithm, we developed a more detailed analysis to find out which factors have higher and lower impact on the risk classification of inmates in the RisCanvi system. In this instance we did an intersection analysis using Venn diagrams (Figure 5). We developed a Venn diagram for four risk levels out of the 5 used by RisCanvi, where each ellipsis represents a relevant behavior: **violent recidivism (VR), breach of sentence (RC), intra-institutional violence (VI) and self-directed violence (VA)**. The intersections of the ellipses correspond to the union between these behaviors in the context of the assigned level of risk.

Figure 5: Intersection analysis



The diagram corresponding to the low-risk level (upper left) shows that 92% percent of the people assessed by RisCanvi in the CPRR database have at least a low-risk level score in one of the four behaviors. Of this 92% we know that 12.7% have a low probability of displaying one of RisCanvi's risk behaviors. Therefore, 87.3% of the low-risk population is classified as having a low probability of presenting more than one behavior.

The second Venn diagram corresponds to the medium risk level (upper right). In this case, a medium risk of breach of sentence (RC) is present in the aggregated 46.4% of the middle-risk population, showing how only this factor explains why inmates may be assigned a medium risk. In the low-risk scenario, we found a more stable distribution of low risks, where most inmates showed low risk in all relevant behaviors, and not just one.

For individuals with at least one high risk behavior, we found that a high risk of breach of sentence or violent recidivism, or a combination of at least two behaviors (VA, VI and/or VR) concentrates most of the individuals with a high-risk classification (46%). This points to a different weighting for violent and non-compliant behaviors, where violent behaviors would be assigned higher weights when assigning risk to an inmate.

To further scrutinize the system, we shifted our focus from risk behaviors to risk factors, which several interviewees flagged as a matter of concern, particularly due to the prevalence of static factors such as the "age at the time of the index offense" which may be difficult to justify in an AI system deployed in a legal context where the goal of the prison system is to foster rehabilitation and inmate reformation.

At the heart of this concern lies the fear that the weight assigned to static factors could disproportionately skew the final risk assessment, potentially undermining the efficacy of human intervention and rehabilitation efforts. By prioritizing elements that remain unalterable over time, RisCanvi might run the risk of diminishing the potential impact of proactive interventions aimed at fostering positive behavioral changes among inmates.

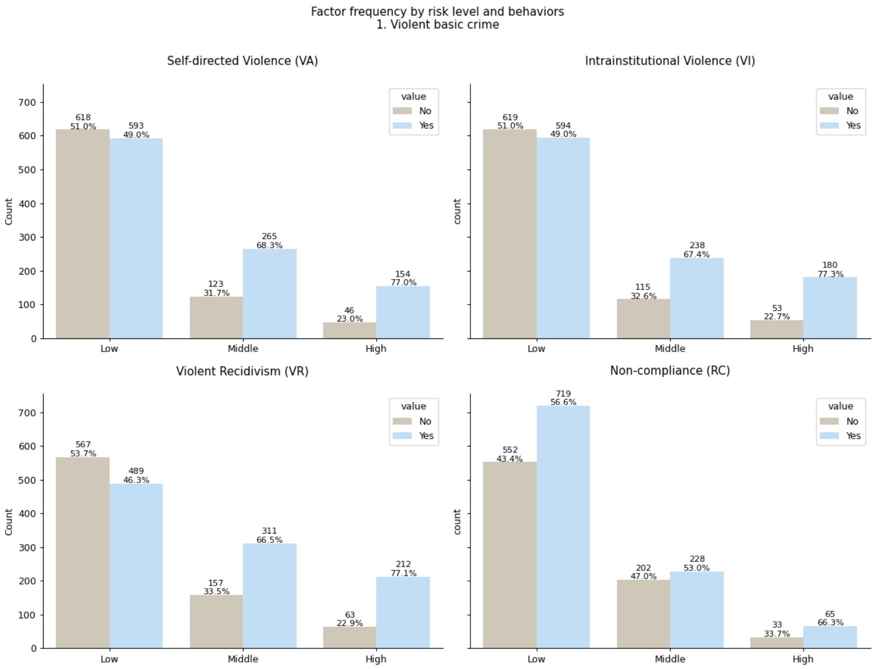
We explored RisCanvi's predictive evaluations for each behavior in a panel composed of four bar charts or histograms[20]. The histograms show both the number of observations (yes or no) and the percentage of active vs the total for each risk-level.[21] We zoomed in on 4 risk factors that we found were particularly relevant: violent base crime, severity/diversity of crimes, belonging to social groups and poor childhood adjustment.

Using the same example as in Figure 6, the percentage of *Yes* at the low level is 49%, at the medium level 68% and, finally, 77% at the high-risk level. Therefore, in this case, there is an increase in the percentage through the risk levels. If we analyze Figure 6 corresponding to the violent base crime factor, we observe that in general there is a decrease in the number of people in the medium and high risk for all behaviors. On the other hand, in the same panel, we observe an increase in the percentage of Yes in the medium and high levels of predicted risk. Using Figure 6 (row one, column three) as a reference, we can conclude that violent base crime is a factor of relevance in the high-risk classifications, especially in VR. In VR behavior we observe that there is a greater number of people classified as high risk who have the characteristic of engaging in violent base crime and also, that their percentage of presence in the risk classifications increased from 44% (low risk) to 77% (high risk). The number and percentage of VR, VI or VA behavior differ from the values obtained for RC behavior. Consequently, we can infer that the violent base crime factor has a greater weight or presence in assigned risk of violent behaviors.

[20] The results are mapped in Annex 8: Table to interpret Factor Analysis.

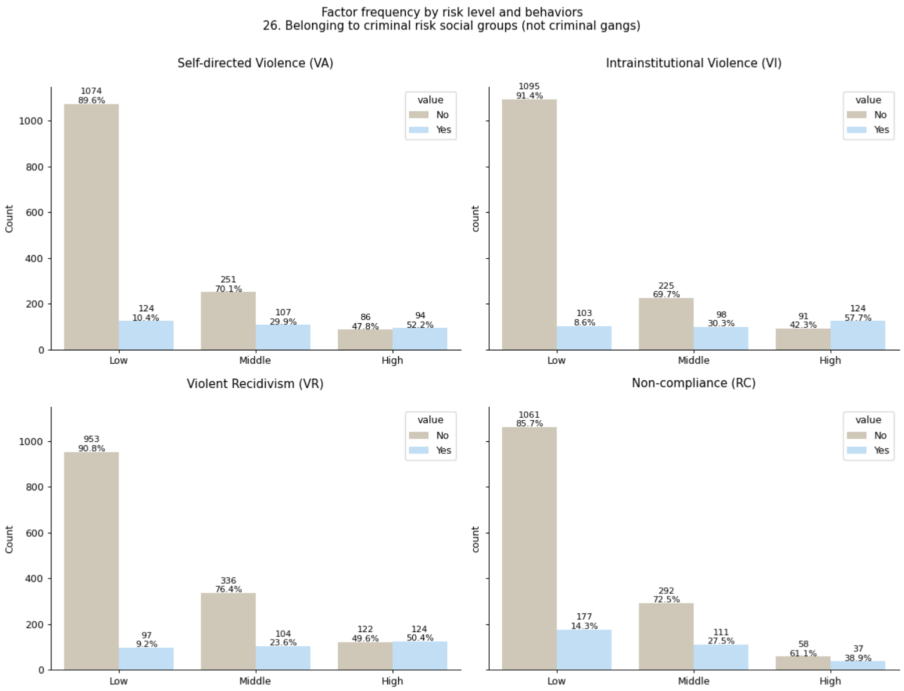
[21] We can perform an example of the above described. In **Error! Reference source not found.**, in the first pair of columns of the upper left graph, we observe the values 618 and 593. The value of 618 corresponds to the number of persons deprived of liberty who do not have a violent crime but who are classified as low risk for self-directed violent behavior. On the other hand, the value 593 corresponds to those persons who did have a violent crime and who are also considered low risk. Therefore, the percentage represents the proportion of *No* and *Yes* in the risk level they are at ($\frac{618}{618+593}$ and $\frac{6593}{618+593}$)**.**

Figure 6: Violent basic crime factor analysis



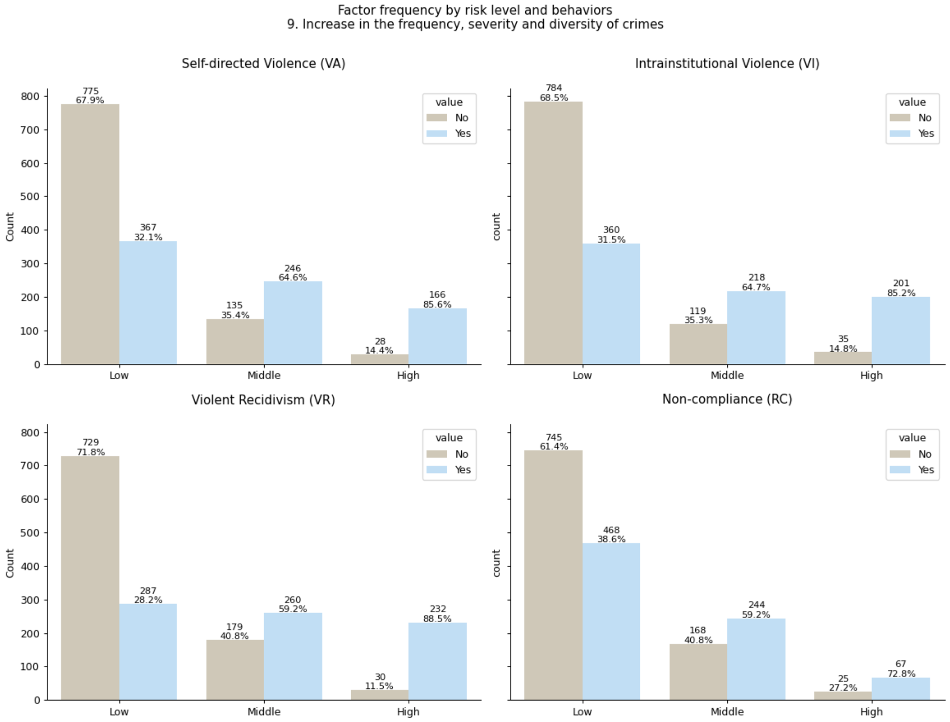
Another relevant risk factor assessed is belonging to social groups at risk of crime. In this case, we found that for behaviors VA and VI, the number of inmates with this factor remains flat through the different risk scores. For VR, the risk increases and for RC it decreases. In all behaviors, the presence of the factor in high predicted risk scores increases, but does not seem to be a determining factor (Figure 7).

Figure 7: Belonging to criminal risk social groups factor analysis



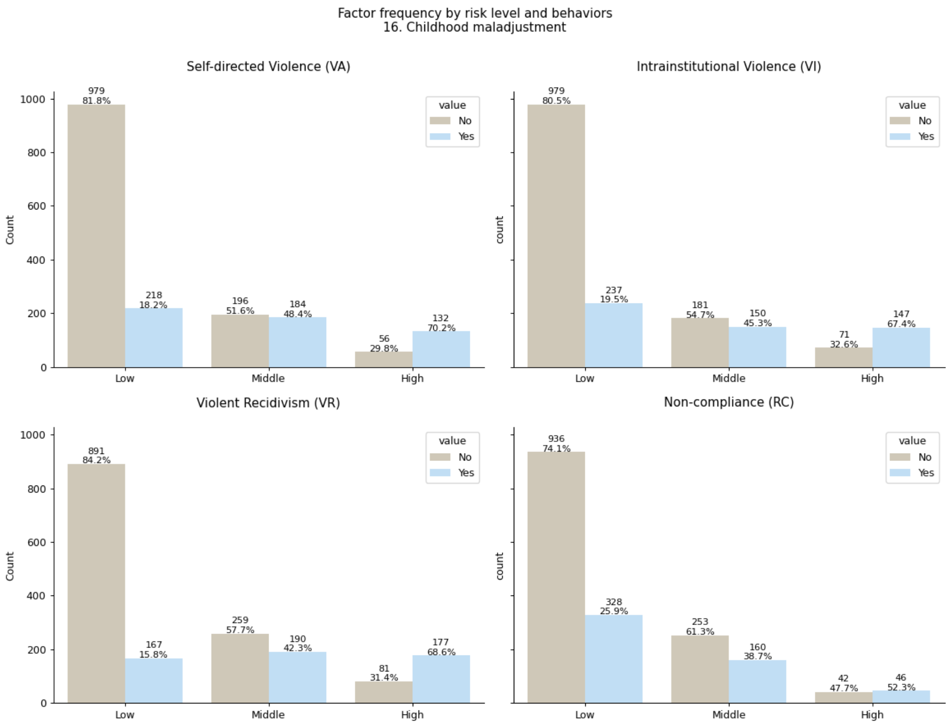
We expected to find a strong correlation between those inmates with a “Yes” for the risk factor “increase in the frequency, severity and diversity of crimes,” as RisCanvi is a tool to assess recidivism risk. However, we found different patterns. While people with this factor represent more than 70% of all high-risk cases, for inmates with a non-compliant (risk of breach of sentence) behavior this factor is very low. If RisCanvi was a transparent system, we could have an open debate on why this combination of factor and behavior is relevant when determining a recidivism risk score, and there may well be a justification for it, but we have not been able to find it during our research.

Figure 8: Increase in the frequency, severity and diversity of crimes analysis.



Finally, RisCanvi includes both dynamic and static factors. These static factors cannot change over time, like an assessment of "Childhood mismatch" or "poor childhood adjustment." For this risk factor, our analysis shows a significant increase in the frequency percentage for inmates deemed high-risk. This increase is especially significant in VA, VI, and VR behaviors, which is problematic. Several of our interviewees expressed concern over the reliance on static risk factors: "*It contains what are known as static risk factors that are not modifiable: .a person cannot modify that item.*" (Prison psychologist) Another psychologist explained: "*You generate powerlessness in the person because there are things that will not change.*"

Figure 9: Poor Childhood adjustment analysis



Overall, our RisCanvi factor prevalence and relationship analysis found that:

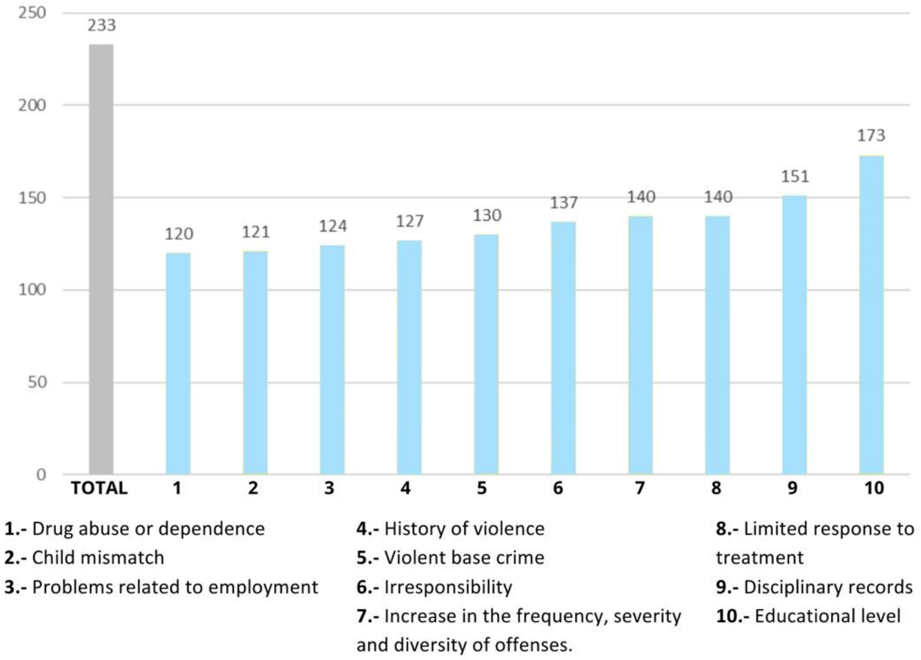
- The RisCanvi assessment weighs factors differently in predicting behavioral risk.
- Fixed factors such as poor childhood adjustment (which cannot be changed over time) prevail in the prediction of high-risk in RisCanvi behaviors even though it is a factor present across the board in the prison population.
- Factors that would be expected to have a greater presence and impact on the risk assignation of the four RisCanvi behaviors, such as increased frequency, severity and diversity of offences, are not prevalent in the behavior of breach of sentence (RC).
- The RisCanvi factor analysis suggests that the algorithm assigns weight to each factor when predicting risk, but no clear or predictable pattern is observed.

So far, all our reverse engineering efforts proved unsuccessful in finding a robust internal logic for RisCanvi. When we found strong correlations, these appeared random and did not seem to fit the leads we found when interviewing relevant stakeholders, the relevant literature on recidivism or our common sense. Therefore, we turned our attention to one specific risk score: **high-risk.** We examined the 10 most frequent risk factors when assessing risk for the behavior "violent recidivism", due to its relevance for the ultimate goal of RisCanvi: to prevent recidivism.

We created Figure 10 with data from the 233 people in our dataset classified as having a high risk of violent recidivism.[22]

[22] To interpret Figure 10: when the factor is a dichotomous variable, we consider it as active when the answer is Yes. When the variable is categorical, we take the response with the value furthest away from

Figure 10: Most present risk factors in prisoners classified as “high risk” of violent recidivism



Total: Total number of prisoners classified as high risk for violent recidivism.

The figure above shows that most inmates considered at high risk of violent recidivism have a low level of education. In RisCanvi education is classified into three levels of academic achievement: low level for primary education, medium level for secondary education or professional training and high level for higher education or university studies. The data at hand undeniably reveals a glaring **overrepresentation of individuals with a low level of education** within the Catalan prison population. This stands in stark contrast to the general population, as 76% of the prison population have only primary studies, while in society that is true for less than 52% of adults. Low educational level is a factor present in 74% of prisoners classified as high risk of violent recidivism. What we do not know is whether inmates can change their educational level by studying while in prison. The opacity of RisCanvi means we do not know whether this is a stable or dynamic factor or how the relevant staff would take into account efforts to increase one’s educational level while in prison.

After educational level, some of the most present risk factors among those deemed high risk are dynamic factors such as disciplinary records, response to treatment or irresponsibility. They are changing factors that allow the assessment of behavior to evolve as individuals receive specialized assistance and take part in rehabilitation programs. Our data (Figure 11) show that these dynamic factors are apparently more relevant in classifying risk, which is consistent with a system that values and accounts for a positive evolution in the areas where

the socially desirable (e.g., opting for 'low' when considering educational levels categorized as low, medium, or high).

an inmate can change and therefore alter their chances and improved their possibilities of an early release.

Figure 11: Comparative based on specific factors

Risk Factor	% of Inmates with active factor	% of Inmates with high risk for violent recidivism	Difference
Disciplinary records	40%	76%	36 p.p.
Limited response to treatment	57%	78%	21 p.p.
Increase in frequency, severity and diversity of offenses	45%	74%	29p.p.
Drug abuse or dependence	31%	67%	36p.p.
Problems related to employment	46%	76%	30 p.p.
Irresponsibility	47%	76%	29 p.p.

However, we are also surprised to see static factors among the most present in predicted high risk individuals. Riscanvi factors 2 and 3, for instance, relate to childhood and employment history prior to being incarcerated. For "childhood mismatch", RisCanvi's files define it as behavioral problems or pattern of misbehavior common during childhood, including low school performance or truancy. It is difficult to understand how inmates can be encouraged to engage in rehabilitation efforts when their childhood follows them in ways they cannot escape or outdo. Other variables used in RisCanvi such as Criminal history in the family of origin or Problematic socialization or upbringing in family of origin are variables that behave in similar, deterministic ways; they are linked to past circumstances that the individual cannot change. These variables may tie the individual to their assessment of risk and discriminate against certain demographic groups without considering their ability to change or rehabilitate during their time in prison.

Up until this point, the data found in our reverse-engineering exercise has led to partial findings and a feeling among the team that RisCanvi tends to make random decisions, as its risk scores do not seem to be anchored in strong factors or behaviors. In this sense, our assessment would validate the findings of Martínez Garay (2016) questioning CEJFE's data.

Our final test consisted on spectral clustering to see whether we could identify groups of individuals with similar characteristics based on the RisCanvi factors.

As the logistic and multinomial regressions did not show statistically significant relationships[23] between factors, behaviors and risk levels in the CPRR dataset, we chose to use Artificial

[23] See Annex 7: Confusion matrix and accuracy table on multinomial model.

Intelligence to audit RisCanvi. We chose an unsupervised learning method[24] to identify patterns in the data and find distinctive features among a number of groups. Our hypothesis was that if we could identify statistically diverse groups, we could then infer that the RisCanvi factors were driving the classification and therefore find a decision-making logic in the system. This would allow us to measure its dynamics and outputs, and establish that RisCanvi's risk scores were not being assigned in ways that we could only describe as random or almost random.

To determine whether the RisCanvi factors were groupable, we used the available observations of the **alternative variables** of the RisCanvi factors (explained in the previous sub-section) and calculated a **Dice similarity matrix**. We then applied a first **hierarchical cluster** model based on the similarity matrix, then applied another clustering model called **spectral clustering** to reinforce the hierarchical clustering findings.

As described earlier on, our analysis is based on a dataset of 1,889 observations with 43 variables (coded like the 43 factors), most of them dichotomous or categorical, which makes the finding of patterns or classifications more complex (as the statistical distribution is not a typical distribution). To facilitate the classification task, we calculated a Sorensen-Dice similarity matrix. This matrix helps in the interpretation of similarity between factors and in the implementation of clusters. The Sorensen-Dice index[25] is constructed as follows:

The Dice coefficient between *u* and *v*, is

$$\frac{c_{TF} + \ c_{FT}}{2c_{TT} + c_{Ft} + c_{TF}}$$

Where c_{ij} is the number of occurrences of $u[k] = i$ and $v[k] = j$ for $k < n$**.**

Thanks to the Sorensen-Dice coefficient, the similarity of all observations can be interpreted with a value ranging from 0 to 1. One corresponds to two equal observations and zero corresponds to two opposite observations. Consequently, the similarity matrix allows us to interpret more easily how similar the observations are taking into account the 43 variables equal to the RisCanvi factors.

The matrix also helps to reduce the complexity of the model. With the Dice matrix, the classification algorithm must focus on finding groups within the matrix results and not on the 43 factors. For example, the algorithm could find those groups of observations that are similar to each other (values close to one), or those observations that share average similarities (values around 0.5).

Once the problem of comparing multiple dichotomous and categorical variables with the similarity matrix has been solved, we applied the classification algorithms. As mentioned

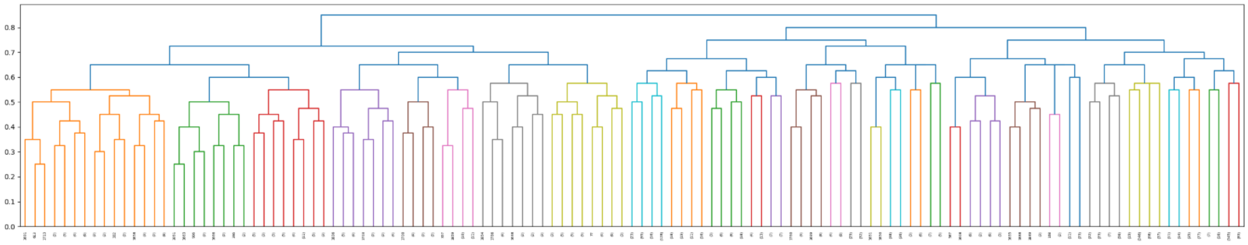
[24] Unsupervised learning in artificial intelligence is a type of machine learning that learns from data without human supervision. Unlike supervised learning, unsupervised machine learning models are given unlabeled data and allowed to discover patterns and insights without any explicit guidance or instruction

[25] Additionally, the same clustering exercise was realized with the Jaccard, Anderberg and Ochiai metrics. Considering these metrics, we do not find any relevant difference in significance, number of clusters or dissimilarity between groups.

above, what is relevant in applying these algorithms is finding indications that the data can be grouped and separated from each other using the RisCanvi factors. For this, two types of clusters were used: hierarchical[26] and spectral.[27]

We thus created a dendrogram[28] (Figure 12) to visualize how the different clusters are structured based on the RisCanvi data. **If inmates are being categorized along a risk scale based on a particular set of characteristics (RisCanvi factors), the dendrogram would show clear groupings**. However, we found multiple ramifications, which means that the observations are not easily grouped. This is consistent with our previous findings with this dataset, but not the best outcome in terms of the robustness, reliability and fairness of the RisCanvi algorithm. All our tests (using different linkage methods, including nearest point, farthest point, and average WPGMA linkage) returned similar results.

Figure 12: Hierarchical clustering dendrogram



We even performed a series of Silhouette tests[29] using an additional similarity matrix[30] as a reference, to find out how statistically different the clusters are from each other. Figure 13 shows the Silhouette score compared to a different Euclidean matrix as a reference. We find two relevant results: one, that the values of the Silhouette index are close to zero. This means that there is overlap between clusters; and so that among the clusters there are elements that could belong to multiple clusters. Second, that the Silhouette score does not increase at the same rate as the clusters (x-axis). This means that even with the significant increase in clusters

[26] The methods used were nearest point, farthest point, and average (WPGMA) linkage methods.

[27] Hierarchical clustering is a methodology capable of working with an indeterminate number of clusters. Unlike other models, it is not necessary to specify an initial number of clusters. The model itself iterates over the observations and the similarity matrix. Such iteration allows finding a split that maximizes the differences between clusters. Additionally, it is a technique that is not very sensitive to outliers, since the assignment of outliers is performed after the classification of most of the observations.

[28] A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities.

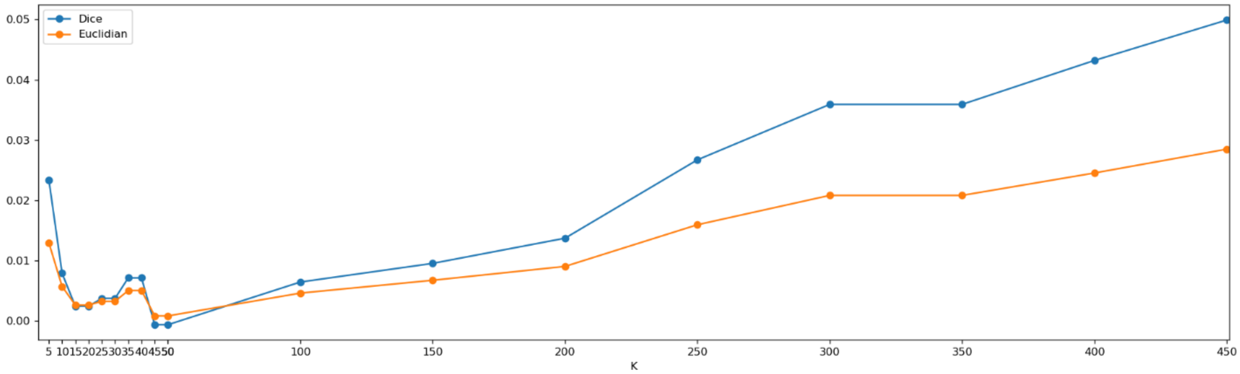
[29] The Silhouette Score is a metric used to measure the goodness of a clustering techniques. Its value has a range between -1 to 1. One, means that the clusters are clearly distinguishable between them. Zero, refers to an overlapping of clusters. Minus one, means an incorrect assignation of groups. The next formula shows its construction. b_i represents the minimum average distance from i to all clusters to which i does not belong. a_i refers to the average distance between i and all the other data points in the cluster to which i belongs.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

[30] The Euclidian distance matrix is commonly used on this type of algorithms, we add this metric to show a baseline to compare the present results with the Sorensen-Dice metric.

or subdivisions (x-axis of the graph) has no impact on the separation of clusters. This second finding, means that the observations and information is highly concentrated among themselves. In this sense, the RisCanvi algorithm, like any other classification algorithm, would also present problems in identifying clusters, even more so if it is required to classify individuals into three categories.

Figure 13: Silhouette score of hierarchical clustering

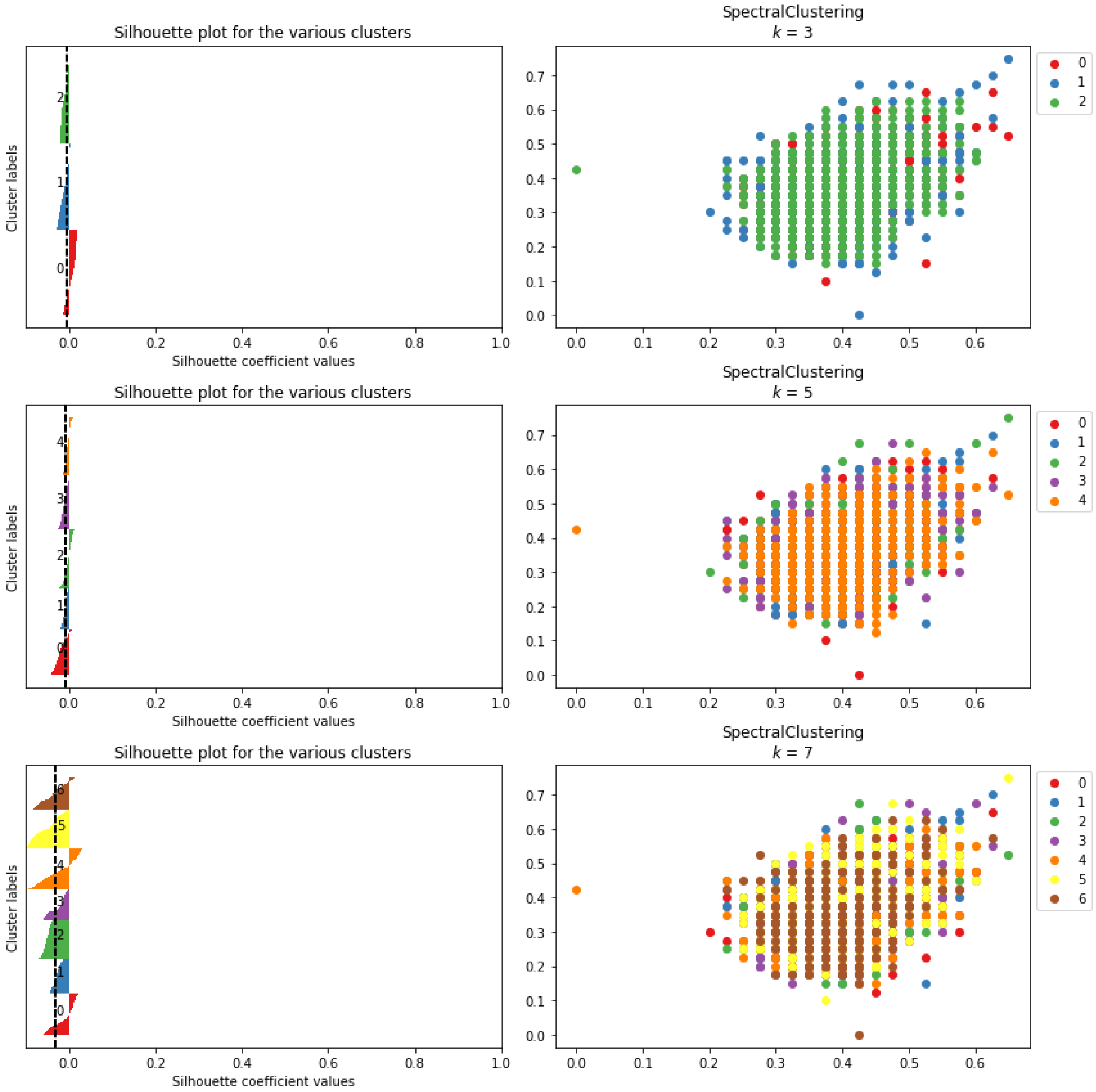


Finally, we used **spectral clustering**, a technique that usually performs well in identifying groups in complex point densities. One of the strengths of this approach is its ability to find patterns, spot anomalies, check assumptions and test hypothesis in overlapping point clouds or more complex features where other algorithms such as K-Means may fail to identify statistical significance (Wang, Qian & Davidson; 2014). Spectral clusters have shown good performance with categorical variables and statistical distributions of complex information (Mbuga & Tortora, 2022), which is the case of RisCanvi.

We used two parameters: the similarity matrix previously calculated following Sorensen-Dice, and the Radial Basis Function (RBF) kernel. We also used different approximations using the Silhouette Score for clusters three, five and seven. The results are shown in Figure 14. The graphs on the left side of the panel correspond to the Silhouette test represented in bar charts. The right side of the panel shows a two-dimensional representation of the alternative variables and their similarity matrix.

As can be seen in the left column of Figure 14, there is no cluster or average that is close to the value "1". This means that the **persons classified as high-risk and therefore deprived of their freedom have similar characteristics in all the clusters**. The right column of the same graph confirms the condition of similarity among the observations by visualizing a circle-like condensation of points.

Figure 14: Spectral clustering analysis results



This latter experiment confirms our findings throughout this adversarial audit: that **RisCanvi risk categories and risk behaviors do not assign risk levels (low, medium, high) in consistent ways**. When reverse-engineering a system like RisCanvi, the expectation was that we would find that the three main risk scores correspond to different sets and combinations of categories and behaviors.

4. Conclusion and recommendations

Our adversarial audit of RisCanvi points to some serious concerns. If we were reverse-engineering bakery products by combining lists and ingredients with the taste and appearance of different products, we would find similarities indicating the choice in ingredients and preparation techniques, and so we would be able to classify the relevant products into bread, pastries and desserts, for instance.

One of the contributions of AI, and the use of data and algorithms, to make decisions, is the possibility to apply a set of indicators and rules to standardize outcomes. **The promise of AI rests, precisely, in its ability to normalize outcomes, remove discretion and ensure consistency in decision-making processes**: faced with the same data and characteristics, an AI system will make consistent decisions. Humans, on the other hand, may find themselves making different decisions for similar cases, or using discretion in ways that are difficult to justify. The system we have audited seems to behave in random ways, in the sense that similar combinations of factors and behaviors can lead to very different risk levels. Likewise, the inmates with high risk do not seem to have consistent attributes in common. But **if factors and behaviors are not determining who get access to increased levels of freedom, who does?**

One of the worst outcomes when incorporating data and AI systems in decision-making processes is to end up with worse human procedures and guarantees but also worse data-driven decisions. AI policies fail when staff is discouraged from actively intervening or having agency in decision-making, as the hope is that the data knows best and will fix everything, and when human decisions are replaces by data-driven systems with opaque indicators and unaccountable outputs.

What we have found with **RisCanvi is a system that is not known by those whom it impacts the most, inmates; that is not trusted by many of those who work with it, who are also not trained on its functioning and weights; that is opaque and has failed to adhere to current regulation** on the use of automated decision-making systems in Spain, where AI audits are required since 2016. Above all, however, our data shows that RisCanvi may not be fair nor reliable, and that it has failed to do what AI does best: standardize outcomes and limit discretion. Consistent with earlier studies, **we do not find RisCanvi to be reliable, as this would require a clear relationship between risk factors, risk behaviors and risk scores**.

As with any adversarial audit, our findings are not final. We were not able to access system data, and so could not confirm our conclusions. But there seems to be enough data on the table to grant **further scrutiny of the system**. The way things stand, whenever a low-risk inmate engages in recidivism, it is impossible to know if the failure to categorize them correctly is the result of an unavoidable error rate or a feature in an unreliable system. Likewise, when an inmate is denied access to increased levels of freedom due to high-risk, it is currently unclear whether this is a fair decision.

Using AI in sensitive settings such as the criminal justice system should require an increased level of transparency and scrutiny, both internal and external, and consistent efforts to inspect and monitor system performance and impact. We do not find this to be the case in the deployment and use of RisCanvi.

Our conclusions are in line with some of the things mentioned by interviewees. Several emphasized the need for enhanced communication, both to corrections staff regarding proper implementation and limitations of RisCanvi, and to inmates regarding scoring and consequences. As one attorney and researcher stated, in a desired future "*There would be transparency in relation to its use and, above all, so that inmates would be informed of its use, its application and its consequences.*" Some interviewees suggested changes, like making the tool optional rather than compulsory. Others believed the current RisCanvi system is fundamentally flawed beyond repair and should be eliminated or replaced by a better system. In the words of one prison psychologist: "*RisCanvi is like a house that has so many structural defects, so many facade defects, so many partition defects that it will not be worth rehabilitating, it has to be demolished and rebuilt.*" Likewise, a former inmate recommended elimination given doubts about the tool's validity: "*I would really eliminate it... if the algorithm in the end turns out to be not reliable, what purpose does it serve?*

Based on the available data, we can only conclude that RisCanvi does not work, and is not currently able to provide the necessary guarantees to inmates, lawyers, judges and criminal justice authorities.

As we unravel the layers of RisCanvi, drawing insights from both human experiences and data-driven analyses, a set of imperative recommendations surface. These stem from the algorithm's intricacies and from the varied perspectives of stakeholders within and beyond the criminal justice system, but also, and most crucially, from the obligations generated by the recent passing of the Ai Act in Europe, which identifies the criminal justice system as a high-risk scenario and requires and increased level of guarantees and scrutiny.

Our call to action is grounded on the pursuit of a justice system that transcends mere algorithmic efficiency, incorporating fairness, equity, and human-centric values and ensuring that hard-won, established rights are not eroded or lost in the black box of AI processes. The following recommendations serve as guiding principles for the successful incorporation of AI solutions in high-risk contexts and when fundamental rights are at stake.

- Take proactive steps to **improve the transparency** of the RisCanvi algorithm by sharing indicators, weights and ratings in a clear and accessible manner with professionals in the prison system, lawyers and judges

- Establish robust protocols to guarantee that **inmates have access to legal support and clear mechanisms to contest and appeal RisCanvi risk scores** if they find them incorrect or want to assess how their risk score was calculated.

- Conduct **external, independent and recurrent audits** of RisCanvi, with a specific focus on its predicted and recorded outputs, assessed over time. Such audits should be end to end and socio-technical, to ensure adherence to the requirements laid out in the AI Act. A version of the audits should be made public.

Going beyond mere algorithmic analysis, this report aims to contribute to a fairer and more equitable world by showing the possibilities of reverse engineering opaque AI systems. By coupling personal narratives with data-driven insights, we hope to have shown that there is no AI black box. All those affected by AI systems can and should scrutinize their functioning and impact, and ensure their rights are protected.

We would like to close this report by extending an invitation to all readers and actors, both within and beyond the criminal justice system, to read, share and use this report to build a movement that pushes for better AI for all. If this piece contributes to a future where AI fosters and promotes individual and collective rights, the time and effort devoted to it will have been worthwhile.

References

Abiteboul, S., & Dowek, G. (2020). Fairness, Transparency, and Diversity. In *The Age of Algorithms* (pp. 115-124). Cambridge: Cambridge University Press. doi:10.1017/9781108614139.014

Alemán Aróstegui, L. (2023). El uso de RISCANVI en la toma de decisiones penitenciarias (the use of RISCANVI in prison decision making). *Estudios Penales y Criminológicos*. https://doi.org/10.15304/epc.44.8884

Andrés-Pueyo, A., Arbach-Lucioni, K., & Redondo, S. (2017). The RisCanvi: A New Tool for Assessing Risk for Violence in Prison and Recidivism. In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton (Eds.), *Handbook of Recidivism Risk/Needs Assessment Tools* (1st ed., Chapter 13). https://doi.org/10.1002/9781119184256.ch13

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, 18, 1-78. https://doi.org/10.48550/arXiv.1704.01701

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ávila, F., Hannah-Moffat, K., & Maurutto, P. (2021). The seductiveness of fairness: Is machine learning the answer? – Algorithmic fairness in criminal justice systems. In *The Algorithmic Society: Technology, Power, and Knowledge* (pp. 87). London: Routledge.

Bagaric, M., Svilar, J., Bull, M., Hunter, D., & Stobbs, N. (2022). The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence. *American Criminal Law Review*, 59(1). https://ssrn.com/abstract=3795911

Baudin, C., Nilsson, T., Sturup, J., Wallinius, M., & Andiné, P. (2021). A Static-99R Validation Study on Individuals With Mental Disorders: 5 to 20 Years of Fixed Follow-Up After Sexual Offenses. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.625996

Bellio, N. (2021, May 25). In Catalonia, the RisCanvi algorithm helps decide whether inmates are paroled. *AlgorithmWatch*. https://algorithmwatch.org/en/riscanvi/

Blacklaws, C. (2018). Algorithms: Transparency and Accountability. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170351. https://doi.org/10.1098/rsta.2017.0351

Bonta, J., & Andrews, D. A. (2007). Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation. *Public Safety Canada*.

Carlson, A. M. (2017). The Need for Transparency in the Age of Predictive Sentencing Algorithms. *Iowa Law Review*, 103(1), 303-329.

Center for Digital Ethics & Policy. (2018, May 7). Sentence by Numbers: The Scary Truth Behind Risk Assessment Algorithms. *Loyola University*. https://www.luc.edu/digitalethics/researchinitiatives/essays/archive/2018/sentencebynumbersthescarytruthbehindriskassessmentalgorithms/

CEJFE. (2015). *Tasa de reincidencia penitenciaria 2014: Investigación propia*. Manel Capdevila Capdevila (Coord.), Berta Framis Ferrer, Carles Soler Iglesias, Paula Ribas Plano, Alba Hostench Fontàs, Sandra Màrquez Postigo, Laura Ruiz Sarrión, Aroa Arrufat Pijuan, Ruth Díez Lerma, Antonio Andrés Pueyo, & Marta Blanch Serentill. Área de Investigación y Formación en Ejecución Penal. https://cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2015/taxa_reincidencia_2014/tasa_reincidencia_2014_cast.pdf

CEJFE. (2023). *Tasa de reincidencia penitenciaria 2020: Investigación propia*. Manel Capdevila Capdevila (Coord.), Berta Framis Ferrer, Carles Soler Iglesias, Paula Ribas Plano, Alba Hostench Fontàs, Sandra Màrquez Postigo, Laura Ruiz Sarrión, Aroa Arrufat Pijuan, Ruth Díez Lerma, Antonio Andrés Pueyo, & Marta Blanch Serentill. Área de Investigación y Formación en Ejecución Penal. https://cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2023/taxa-reincidencia-penitenciaria/Taxa_reincidencia_penitenciaria_2020_RESUMEN_EJECUTIVO_ESP.pdf

Chiao, V. (2022). Sentencing and the right to reasons. In J. Ryberg & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence*. Oxford University Press.

Coid, J. W., Ullrich, S., Kallis, C., et al. (2016). Improving risk management for violence in mental health services: a multimethods approach. In NIHR Journals Library (Ed.), *Programme Grants for Applied Research* (No. 4.16.), Chapter 18, Development of a dynamic risk assessment for violence. Southampton, UK. Available from: https://www.ncbi.nlm.nih.gov/books/NBK396458/

Cole, D., & Angus, G. (2003). Using pre-sentence reports to evaluate and respond to risk. *Criminal Law Quarterly*, 47(3), 302-264.

Diakopoulos, N. (2020). Transparency. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780190067397.013.11

Digital Future Society (January 2023). Algorithms in the public sector: four case studies of ADMS in Spain. https://digitalfuturesociety.com/report/algorithms-in-the-public-sector-four-case-studies-of-adms-in-spain/

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Eticas (2023). Adversarial Algorithmic Auditing Guide. Association Eticas Research and Innovation. https://eticas.ai/case-study/adversarial-algorithmic-auditing-guide/

Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the Compas: Validation Data on Two Risk-Needs Tools. *Criminal Justice and Behavior*, 35(9), 1095-1108. https://doi.org/10.1177/0093854808320497

Fitzgibbon, D. (2008). Fit for purpose? OASys assessments and parole decisions. *Probation Journal*, 55(1), 55-69. https://doi.org/10.1177/0264550507085677

Fitzgibbon, W., & Green, R. (2006). Mentally Disordered Offenders: Challenges in using the OASys risk assessment tool. *British Journal of Community Justice*, 4.

Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice*, 34(5), 523-529. https://doi.org/10.1016/j.jcrimjus.2006.09.007

Gavazzi, S. M., Yarcheck, C. M., & Lim, J.-Y. (2005). Ethnicity, gender, and global risk indicators in the lives of status offenders coming to the attention of the juvenile court. *International Journal of Offender Therapy and Comparative Criminology*, 49(6), 696–710. https://doi.org/10.1177/0306624X05276467

Gimeno Beviá, J. (2023). Predictive Policing and Predictive Justice in the Spanish Legal System: Current Situation and Lege Ferenda Ideas Before Future Applications. *International Association of Penal Law*. https://www.penal.org/sites/default/files/files/Spanish%20report%20-English%20Version-.pdf

Karimi-Haghighi, M., & Castillo, C. (2021). Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (pp. 210–214). https://doi.org/10.1145/3462757.3466150

Karimi-Haghighi, M., & Castillo, C. (2022). Quantitative analysis of disparate effects of RisCanvi for estimating the risk of violent recidivism. Technical Report, Web Science and Social Computing Research Group, Universitat Pompeu Fabra. https://chato.cl/papers/karimi_haghighi_2022_quantitative_analysis_disparate_effects_riscanvi_revi.pdf

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, *Harvard Law School*. https://dash.harvard.edu/handle/1/33746041

Hannah-Moffat, K., & Maurutto, P. (2010). Re-contextualizing Pre-sentence Reports: Risk and Race. *Punishment and Society*, 12(3), 262-286. NCJ Number 231688. https://doi.org/10.1177/1462474510377592

Hannah-Moffat, K., & Struthers Montford, K. (2019). Unpacking Sentencing Algorithms: Risk, Racial Accountability, and Data Harms. In J. W. de Keijser, J. V. Roberts, & J. Ryberg (Eds.), *Predictive Sentencing: Normative and Empirical Perspectives*. Hart Publishing.

Howard, P. (2011). Hazards of different types of reoffending. Ministry of Justice Research Series 3/11.

Howard, P. D., & Dixon, L. (2012). The Construction and Validation of the OASys Violence Predictor: Advancing Violence Risk Assessment in the English and Welsh Correctional Services. *Criminal Justice and Behavior*, 39(3), 287-307. https://doi.org/10.1177/0093854811431239

Hsu, C.-I., Caputi, P., & Byrne, M. K. (2009). The Level of Service Inventory—Revised (LSI-R): A Useful Risk Assessment Measure for Australian Offenders? *Criminal Justice and Behavior*, 36(7), 728-740. https://doi.org/10.1177/0093854809335409

Huq, A. Z. (2019). Racial Equity in Algorithmic Criminal Justice. *Duke Law Journal*, 68, 1043-1134. https://scholarship.law.duke.edu/dlj/vol68/iss6/1

Jiménez Arandia, P. (2023). *Algorithmic transparency in the public sector*. (First edition). Govern obert ; 9.

La Vanguardia. (2021, December 6). Un algoritmo impreciso condiciona la libertad de los presos. https://www.lavanguardia.com/vida/20211206/7888727/algoritmo-sirve-denegar-permisos-presos-pese-fallos.html

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Latessa, E. J., Lemke, R., Makarios, M., Smith, P., & Lowenkamp, C. T. (2010). The creation and validation of the Ohio Risk Assessment System (ORAS). *Federal Probation*, 74(1), 16–22. https://psycnet.apa.org/record/2010-16555-002

Lovins, B. K., Latessa, E. J., May, T., & Lux, J. (2018). Validating the Ohio Risk Assessment System Community Supervision Tool with a Diverse Sample from Texas. *Corrections*, 3(3), 186-202. https://doi.org/10.1080/23774657.2017.1361798

Lowder, E. M., Morrison, M. M., Kroner, D. G., & Desmarais, S. L. (2019). Racial Bias and LSI-R Assessments in Probation Sentencing and Outcomes. *Criminal Justice and Behavior*, 46(2), 210-233. https://doi.org/10.1177/0093854818789977

Lowenkamp, C. T., & Bechtel, K. (2007). Predictive Validity of the LSI-R on a Sample of Offenders Drawn From the Records of the Iowa Department of Corrections Data Management System. *Federal Probation*, 71(3), 25-29. https://www.uscourts.gov/sites/default/files/71_3_4_0.pdf

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13, 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Martínez Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia. *Revista Española De Investigación Criminológica*, 14, 1–31. https://doi.org/10.46381/reic.v14i0.97

Mbuga, F., & Tortora, C. (2022). Spectral Clustering of Mixed-Type Data. *Stats*, 1-11. https://doi.org/10.3390/stats5010001

Moore, R., & Howard, P. D. (Eds.). (2015). A compendium of research and analysis on the Offender Assessment System (OASys) 2009–2013. National Offender Management Service, Ministry of Justice Analytical Series.

Morton, S. (2009). Can OASys deliver consistent assessments of offenders? Results from the inter-rater reliability study. *Ministry of Justice Research Summary* 1/09. http://webarchive.nationalarchives.gov.uk/20110201125714/http://www.justice.gov.uk/publications/docs/oasys-research-summary-01-09.pdf

New, J., & Castro, D. (2018). How policymakers can foster algorithmic accountability. Washington, D.C.: Center for Data Innovation. www2.datainnovation.org/2018-algorithmic-accountability.pdf

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Phenix, A., & Epperson, D. L. (2016). Overview of the Development, Reliability, Validity, Scoring, and Uses of the Static-99, Static-99R, Static-2002, and Static-2002R. In A. Phenix & H. Hoberman (Eds.), *Sexual Offending. Springer*, New York, NY. https://doi.org/10.1007/978-1-4939-2416-5_19

Ryan, M. J. (2020). Secret Algorithms, IP Rights, and the Public Interest. *Nevada Law Journal*, 21(1), 61. https://scholars.law.unlv.edu/nlj/vol21/iss1/3

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.6ed64b30

Ryberg, J., & Petersen, T. S. (2022). Sentencing and the Conflict Between Algorithmic Accuracy and Transparency. In J. Ryberg, & J. V. Roberts (Eds.), *Sentencing and Artificial Intelligence.* Oxford University Press.

Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review*, 66(4), 803-872.

Šugman Stubbs, K., & Plesničar, M. M. (2018). Subjectivity, Algorithms, and the Courtroom. In A. Završnik (Ed.), *Big Data, Crime and Social Control* (1st ed.). Routledge.

Van Eijk, G. (2021). Algorithmic reasoning: The production of subjectivity through data. In *The Algorithmic Society: Technology, Power, and Knowledge* (pp. 119). London: Routledge.

Wang, X., Qian, B. & Davidson, I. On constrained spectral clustering and its applications. Data Min Knowl Disc 28, 1–30 (2014). https://doi.org/10.1007/s10618-012-0291-9

Wexler, R. (2018). Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. 70 *Stanford Law Review*, 1343. https://www.stanfordlawreview.org/print/article/life-liberty-and-trade-secrets/

Wooldridge, J. (2012). Introductory econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning.

Završnik, A. (Ed.). (2018). *Big Data, Crime and Social Control* (1st ed.). Routledge.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18(5), 623-642. https://doi.org/10.1177/1477370819876762

La educación en las prisiones catalanas, U. R. del D. a. (s/f). Escuela y prisión en Cataluña. Gencat.cat. página 25 (Link)

Annexes

Annex 1: Algorithms in the criminal justice system: an overview

1.1 COMPAS

Used in some U.S. states, including New York, Wisconsin, California, Florida's Broward County, and others
Years of operation: 1990s – Ongoing

Racial bias within criminal justice systems has garnered increased attention as predictive algorithms play a central role in influencing critical decisions such as bail, sentencing, and parole. Among these algorithms, **COMPAS**, developed by Northpointe Inc., stands as one of the most prominent commercial risk assessment tools. In 2016 (Angwin, J. et al.), ProPublica embarked on a rigorous investigation to scrutinize the potential presence of racial bias within COMPAS's recidivism algorithm. Their overarching objective was to evaluate the algorithm's accuracy, particularly across different racial groups, with a focus on assessing whether certain groups were more prone to being erroneously classified as either higher or lower risk individuals.

To achieve this, ProPublica undertook an extensive analysis of COMPAS scores, criminal records, and subsequent recidivism data for a vast cohort of over 10,000 individuals who had been arrested in Broward County, Florida, between 2013 and 2014. In their investigation, they methodically compared COMPAS's predicted recidivism risk categories for each defendant with the actual recidivism rates observed over a two-year span. Statistical tests were employed to isolate the influence of race from other variables, including criminal history, age, and gender. The results of the analysis revealed several pivotal findings. While the COMPAS algorithm exhibited moderate overall accuracy in predicting recidivism risk, achieving correct predictions in 61% of cases, the study also showed concerning trends. **Black defendants** were notably 77% more likely than their white counterparts to be erroneously classified as **higher risk individuals**. Conversely, white defendants were more inclined to be underestimated as low-risk individuals. The research showed that while mistakes occurred at similar rates for both black and white inmates, the types of errors varied depending on race. Even after accounting for other influential factors, race remained a significant predictor of being inaccurately categorized as higher risk.

Another notable study (Dressel & Farid, 2018) shed light on the **accuracy and fairness** of recidivism prediction algorithms compared to human judgment. This study compared the accuracy and fairness of COMPAS with predictions made by individuals with limited criminal justice expertise responding to an online survey. Key findings revealed that COMPAS had an accuracy rate of around 65%, a figure comparable to predictions by human participants and simple machine learning models. This suggests that there may be an inherent accuracy limit in recidivism prediction. Although COMPAS employs 137 features in its predictions, a basic logistic regression model using only age and total prior convictions achieved similar results.

This implies that COMPAS's predictions may not be significantly more advanced than a basic linear model.

1.2 LSI-R

Used in the U.S. State of Washington
Years of operation: 1999 – Ongoing

The Level of Service Inventory-Revised (LSI-R) is a risk assessment tool utilized by the Washington State Department of Corrections to evaluate an **offender's potential for reoffending**. Developed in Canada during the 1980s, the LSI-R consists of **54 questions** organized into **ten domains** that encompass various aspects of an offender's life, including criminal history, education, finances, family, and personal problems. Offenders' responses are **scored** to generate an LSI-R score, which **can range from 1 to 54**. A higher score indicates a greater likelihood of reoffending, while a lower score suggests a reduced probability. The LSI-R's background is rooted in its reputation as a valid measure for predicting reoffending, leading to its adoption by the Washington State DOC in 1999. This tool has been designed to examine a broad spectrum of an offender's life circumstances and behavior, providing valuable insights into their risk profile. By analyzing an offender's responses across multiple domains, the LSI-R aims to assist in making evidence-based decisions within the criminal justice system, ultimately contributing to more effective offender management and reducing recidivism rates.

Van Eijk (2021) underscores the LSI-R's extensive influence, serving as a benchmark for other risk assessment tools such as **OASys** in England and Wales, **RISc** (Risico Inschattings Schalen) in the Netherland and **RITA** (Riski-ja tarvearvio) in Finland and categorizing it as a 'third-generation' instrument due to its inclusion of **dynamic risk factors**. Lowenkamp & Bechtel (2007) conducted a study on the LSI-R's predictive validity for probationers and parolees, concluding that it is a valuable predictor of recidivism. Their findings, based on a sample of 1,145 individuals, underline the tool's reliability and effectiveness in predicting recidivism. Hsu et al. (2009) conducted a study with over 78,000 Australian offenders, and while not entirely positive, they report significant insights. They found **no gender differences** in LSI-R total scores and observed that the tool effectively distinguished risk levels based on sentence order. This suggests equitable application of the LSI-R between genders and its ability to differentiate between risk levels across offender groups. Flores et al. (2006) also provides support for the LSI-R, emphasizing the importance of staff training and agency experience in maintaining its predictive validity. Their study underscores the LSI-R's effectiveness in informing correctional decisions and contributing to better outcomes for offenders.

However, criticism arises from Lowder et al. (2019), which suggests **potential racial disparities** in sentencing decisions based on LSI-R results. The study analyzed 11,792 probationers found that at low-risk levels, White probationers received longer sentences than Black probationers, indicating potential racial disparities in sentencing. However, there was no racial difference in sentencing at higher risk levels. The study also showed that LSI-R assessments had similar predictive validity for probation outcomes, regardless of race. When an alternative risk classification system was used, only minor variations were observed.

1.3 Static-99R

Used in Canada and the United States
Years of operation: 1999 – Ongoing

The STATIC-99R is an actuarial risk assessment tool used to estimate the likelihood of sexual and violent reconviction for adult males with prior sexual offense convictions, including first-time offenders, that has been commonly used in Canada and the United States (Phenix & Epperson, 2016). Unlike LSI-R, Static-99R as the name suggests lack of the so-called dynamic factors.
Static-99R comprises **ten items**, including age at release from the index sex offense (Item 1), cohabitation history (Item 2), prior convictions for non-sexual violence (Items 3 and 4), prior sex offense charges or convictions (Item 5), prior sentencing dates (Item 6), convictions for non-contact sex offenses (Item 7), unrelated victims (Item 8), stranger victims (Item 9), and male victims (Item 10). These items are scored to determine an individual's risk level (Baudin et al., 2021).

1.4 Offender Assessment System (OASys)

Used in England and Wales
Years of operation: 2002 - Ongoing

The Offender Assessment System (OASys) is a validated general risk assessment tool used by the prison and probation services in England and Wales. Developed by the Home Office in 2002, OASys serves as an actuarial risk and needs assessment, generating a summary risk score to evaluate the likelihood of reoffending and the risk of harm to self and others. The tool underwent significant updates in August 2009, introducing the OASys General reoffending Predictor (OGP) and the OASys Violence Predictor (OVP). These additions replaced the old OASys score, offering enhanced predictive capabilities (Howard, 2011).

OASys comprises 14 subsections, addressing various aspects of an individual's life, and the August 2009 update introduced 'layered OASys,' providing Basic, Standard, and Full assessments. These assessments share a similar structure but differ in length. The OGP and OVP play a crucial role, predicting the likelihood of nonviolent and violent proven reoffending, respectively. They combine information on identified static and dynamic risk factors. OASys also includes an electronic version (eOASys) introduced in 2005 (Howard, 2011). OASys assessments must be conducted by prison or probation staff with the necessary knowledge of offender behaviors, and ongoing refresher training is recommended. The tool has several strengths, including a dedicated section for assessing intervention suitability and a self-assessment component allowing individuals to record their views on their risk/needs. Empirically grounded in the 'what works' evidence base and risk-need-responsivity principles, OASys drew from effective practice guidelines and empirical grounding of the LSI-R and the Assessment Case management and Evaluation (ACE) (Moore and Howard, 2015).

There is currently no international research available on this aspect. In terms of predictive accuracy, OASys has demonstrated moderate to high accuracy in various studies. Improvements were observed when used in conjunction with the OGP and OVP. The tool contributes to risk practice by creating awareness of static and dynamic risk factors, informing

Pre-Sentence Reports, and identifying targets for treatment/change (Howard and Dixon, 2012). While the tool has strengths, concerns have been raised about its accuracy in predicting recidivism in sub-groups, such as those with mental disorders and ethnic minorities (Fitzgibbon and Green, 2006; Fitzgibbon, 2008). Additionally, some subsections exhibit limited inter-rater reliability (Morton, 2009). Ongoing research and updates, including the development of OGP2, OVP2, and the OASys Sexual reoffending Predictor (OSP), are underway. Howard's recommendations emphasize recognizing the importance of positive factors during assessments and monitoring their recording (Moore and Howard, 2015).

1.5 PredPol

Used in some U.S. states, including California
Years of operation: 2010 – Ongoing

Another notable predictive algorithm utilized within the criminal justice system is **PredPol** (Ryan, 2020). PredPol is specialized software designed for police departments, enabling law enforcement to strategically allocate their limited resources to areas statistically deemed most likely to witness future criminal activity. While this innovative approach appears to offer valuable insights into crime prevention, concerns surrounding **potential biases** have also surfaced. Indeed, Lum & Isaac (2016) revealed how predictive policing algorithms, like PredPol, amplify bias and discrimination in law enforcement. These systems, designed to identify likely targets for police intervention and prevent crimes through **statistical predictions**, have raised legitimate concerns regarding their inherent biases and the implications for marginalized communities. To illustrate these issues, the study delves into the case of Robert McDaniel, a 22-year-old black man living in Chicago's South Side, who found himself on the Chicago Police Department's "heat list." These individuals were identified as potentially involved in violent crimes based on an analysis of geographical location and arrest data. This system exemplifies the shift towards predictive policing, attempting to prevent crimes before they occur.

The pivotal concern raised in the study is that police-recorded data sets are inherently flawed due to **systematic bias**. Police databases do not represent a complete census of all criminal offenses, nor do they offer a representative random sample. Empirical evidence suggests that the police's implicit or explicit considerations of race and ethnicity in their determinations of whom to detain, search, and which neighborhoods to patrol introduce bias into these datasets. Predictive policing algorithms, like PredPol, learn and replicate patterns in the data provided to them. When this data is biased, the algorithms inadvertently perpetuate these biases. As a result, if police data over-focuses on certain ethnic groups and neighborhoods, predictive policing models will amplify the perceived prevalence of crimes in these areas, irrespective of the actual crime rates. This cyclical process results in a feedback loop that reinforces existing biases. Furthermore, the study highlights that community-driven factors, including levels of community trust in the police and the amount of local policing, contribute to bias in police records. The study emphasizes the unequal crime reporting rates across precincts due to these factors. Importantly, any bias present in police-recorded data is compounded when used to inform predictive policing models. The PredPol algorithm, for example, tends to reinforce biases in police data rather than correcting for them. Thus, this approach exacerbates unequal policing by over-representing certain communities, particularly those with high proportions of non-white and low-income residents. Such disparities raise concerns about discriminatory policing and its associated consequences, including the exacerbation of social and economic inequalities.

1.6 Ohio Risk Assessment System (ORAS)

Used in the U.S. State of Missouri
Years of operation: 2019 - Ongoing

The Verified Risk Assessment Tool, grounded in the Ohio Risk Assessment System (ORAS), plays a crucial role in the Missouri Department of Corrections' approach to evaluating and managing the risks posed by individuals involved with the criminal justice system. Functioning as a dynamic tool, it assesses various factors influencing an individual's likelihood of re-offending. The primary objective is to provide personalized interventions and programs tailored to the specific needs of each individual throughout their engagement with the department aimed at reinserting individuals into society. Originating in Ohio, the ORAS was developed to enhance consistency and communication among criminal justice agencies by assessing the risk and needs of offenders at different stages within the system. The tool was strategically designed to predict recidivism and classify offenders based on their risk levels. Additionally, it identifies criminogenic needs and potential barriers to programming, aligning with the principles of effective classification.

Latessa's work in 2010 highlights the goals of ORAS, emphasizing its predictive capabilities and its role in identifying criminogenic needs and barriers to programming. This comprehensive approach enhances the efficiency of resource allocation and decision-making, contributing to more effective interventions. Lovins' research in 2017 builds on these foundations by emphasizing the importance of local validation and norming of risk assessment tools. The study specifically evaluated the applicability of the Ohio Risk Assessment System-Community Supervision Tool to a Texas population. While the Ohio version remained predictive, adjustments made for Texas-specific legal factors and sociopolitical differences significantly strengthened the instrument. The study also provided insights into gender, race, and ethnicity differences, emphasizing the need for context-specific considerations.

1.7 CORELS

Used in: unknown
Years of operation: unknown

CORELS, an acronym for "Certifiable Optimal RulE ListS," is a specialized algorithm tailored for generating rule lists within the realm of supervised machine learning. Rule lists serve as predictive models relying on a set of features. In the specific context of CORELS, it is purpose-built for crafting optimal rule lists optimized for categorical feature spaces.

According to some studies (Angelino et al., 2018; Rudin, 2019), when juxtaposed with the widely adopted COMPAS algorithm for recidivism prediction, CORELS emerges as a **compelling alternative**. Ryberg and Petersen (2022) highlights CORELS as an intuitive machine learning model that excels in identifying if-then patterns, predominantly based on age and criminal history. For instance, this model discerns that if an offender possesses either a history of more than three prior crimes, falls within the age range of 18–20 and is male, or falls within the age range of 21–23 with two or three prior crimes, they are forecasted to recidivate within a two-year period; otherwise, not. A distinctive attribute of CORELS is its **accessibility and transparency**, a stark contrast to the complex proprietary nature of COMPAS.

In empirical studies conducted by Angelino et al. (2018), CORELS shines, generating concise rule lists often comprising merely 3-4 rules, yet achieving a level of predictive accuracy akin to COMPAS, especially concerning recidivism prediction using the ProPublica dataset[31]. Notably, CORELS rule lists hinge exclusively on age, prior criminal records, and gender, deliberately excluding explicit consideration of race. This stands in stark contrast to the concerns raised about COMPAS' potential racial bias (Larson et al., 2016). CORELS' transparency and simplicity foster constructive discussions around fairness and bias, in stark contrast to COMPAS's complex and proprietary approach, which conceals the rationale behind its predictions. CORELS underlines the importance of a **few key variables, such as age and criminal history**, in determining recidivism risk, as underscored by Dressel and Farid (2018). Furthermore, the efficiency with which Angelino et al. demonstrates CORELS' exploration of rule lists, problem-solving capabilities, and certification of optimality within a reasonable timeframe is a breakthrough compared to traditional methodologies. This aligns perfectly with Rudin's argument (2019) that CORELS' transparent and straightforward nature could potentially supplant the opaque black box algorithms like COMPAS for recidivism prediction while delivering similar levels of accuracy.

CORELS' success in generating interpretable and optimal rule lists that rival the performance of intricate black box models like COMPAS reinforces the idea that there's potential to replace proprietary algorithms with more transparent and accountable alternatives within the domain of recidivism prediction. This holds the promise of a more equitable and fair approach to predicting recidivism.

[8] See https://github.com/propublica/compas-analysis

Annex 2: Description of Risk factors in RisCanvi C

Criminal/Penitentiary	
Violent index offense (1)	Age at the time on index offense (2)
Intoxication during the perpetration of the index offense (3)	Victims with injuries (4)
Length of criminal convictions (5)	Time served in prison (6)
History of violence (7)	Start of the criminal or violent activity (8)
Increasing of the frequency, seriousness, and diversity of the offenses (9)	Conflict with other inmates (10)
Failure to accomplishment of penal measures (11)	Disciplinary reports (12)
Escapes or absconding (13)	Grade regression (14)
Breaching prison permission (15)	

Biographical	
Poor childhood adjustment (16)	Distance from residence to prison (17)
Educational level (18)	Problems related with employment (19)
Lack of financial resources (20)	Lack of viable plans for the future (21)

Family/Social	
Criminal history of family or parents (22)	Difficulties in the socialization or development in the origins family (23)
Lack of family or social support (24)	Criminal or antisocial friends (25)
Member of social vulnerable groups (26)	Relevant criminal role (27)
Gender violence victim (only women) (28)	Dependent family charges (29)

Clinical	
Drug abuse or dependence (30)	Alcohol abuse or dependence (31)
Severe mental disorder (32)	Sexual promiscuity and/or paraphilia (33)
Limited response to psychological and/or psychiatric treatments (34)	Personality disorder related to anger, impulsivity, or violence (35)
Poor stress coping (36)	Self-injury attempts/behavior (37)

Attitudes/ Personality	
Pro-criminal or antisocial attitudes (38)	Low mental ability (39)
Recklessness (40)	Impulsiveness and emotional instability (41)
Hostility (42)	Irresponsibility (43)

Annex 3: Interview Questions

Opening Questions

Do you have knowledge of the RisCanvi system?

This question seeks to determine whether the interviewee is familiar with the RisCanvi system. It serves as an opening question to gauge their awareness of the system. The response helps establish a baseline understanding and their general knowledge about RisCanvi. Their level of awareness is valuable, even if they may not be experts in the field.

What is your experience with RisCanvi?

This question delves deeper into the interviewee's practical experience with the RisCanvi system. It aims to explore their interactions, usage, or involvement with the system in their professional capacity. Experience could include the use of RisCanvi in their decision-making processes and its impact on their work.

Can you briefly explain how the RisCanvi system works?

This inquiry seeks to assess the interviewee's ability to provide a concise overview of how the RisCanvi system operates. It encourages them to outline the system's functioning, including elements such as the type of data it processes, its internal mechanisms, and the outcomes it produces. This question helps ensure that interviewees can articulate their understanding of the system.

Based on your professional experience, do you believe the RisCanvi system treats certain groups differently than others? For example, do some groups tend to receive lower RisCanvi scores than others?

This question investigates whether the interviewee perceives any disparities or biases in the application of the RisCanvi system. It focuses on their professional experience and judgment regarding whether RisCanvi may affect different groups, such as inmates, in distinct ways. Specifically, it inquiries about potential variations in RisCanvi scores between different demographic or behavioral groups and seeks to understand the interviewee's reasoning for such observations.

Professional & Legal Involvement

What is the role of professionals administering the questionnaire? Can they change the score? How often do they make changes?

This question seeks to understand the responsibilities of the professionals involved in the administration of the RisCanvi questionnaire. It also addresses whether these professionals have the authority to alter the risk score and the frequency at which such changes occur. This inquiry aims to uncover the extent of human intervention in the assessment process.

Can inmates receive legal support before or during their RisCanvi assessment?

This question aims to determine if inmates can seek legal guidance or representation during their RisCanvi assessments. Understanding the availability of legal support is critical in ensuring that inmates' rights are protected throughout the evaluation process.

Who decides to grant third-degree progression and/or parole? What role does the RisCanvi score play in this decision?

This set of questions delves into the decision-making process related to third-degree progression and parole. It seeks to identify the key decision-makers and assess the influence of the RisCanvi score in these determinations. Understanding the decision hierarchy and the role of the RisCanvi score provides insights into the system's overall structure and the factors considered during critical decisions affecting inmates.

Confidence and Perceptions

Do you believe that inmates are adequately informed about how the RisCanvi system works? Are they informed about their RisCanvi risk score?

These questions aim to assess the level of information provided to inmates regarding the RisCanvi system and whether they are aware of their individual risk scores. It explores the transparency and communication of the system's functioning to those it directly impacts.

Do you trust the RisCanvi system? Why? Please provide a brief explanation.

This question seeks to understand the level of trust professionals have in the RisCanvi system. Respondents are encouraged to provide a concise explanation for their trust or lack thereof. Trust is a crucial factor in evaluating the system's credibility.

Based on your professional experience, do inmates trust the RisCanvi system? Why? Please explain briefly.

Here, the focus is on the trust inmates place in the RisCanvi system. The question aims to explore whether inmates have confidence in the system and the reasons behind their trust or skepticism. Inmates' perceptions are valuable indicators of the system's effectiveness.

What are the key strengths of the RisCanvi system?

This question encourages respondents to identify and highlight the strengths and advantages of the RisCanvi system. It provides an opportunity to recognize the aspects of the system that are functioning well and contributing positively to decision-making.

What are the main weaknesses of the RisCanvi system, including issues like "false positives"?

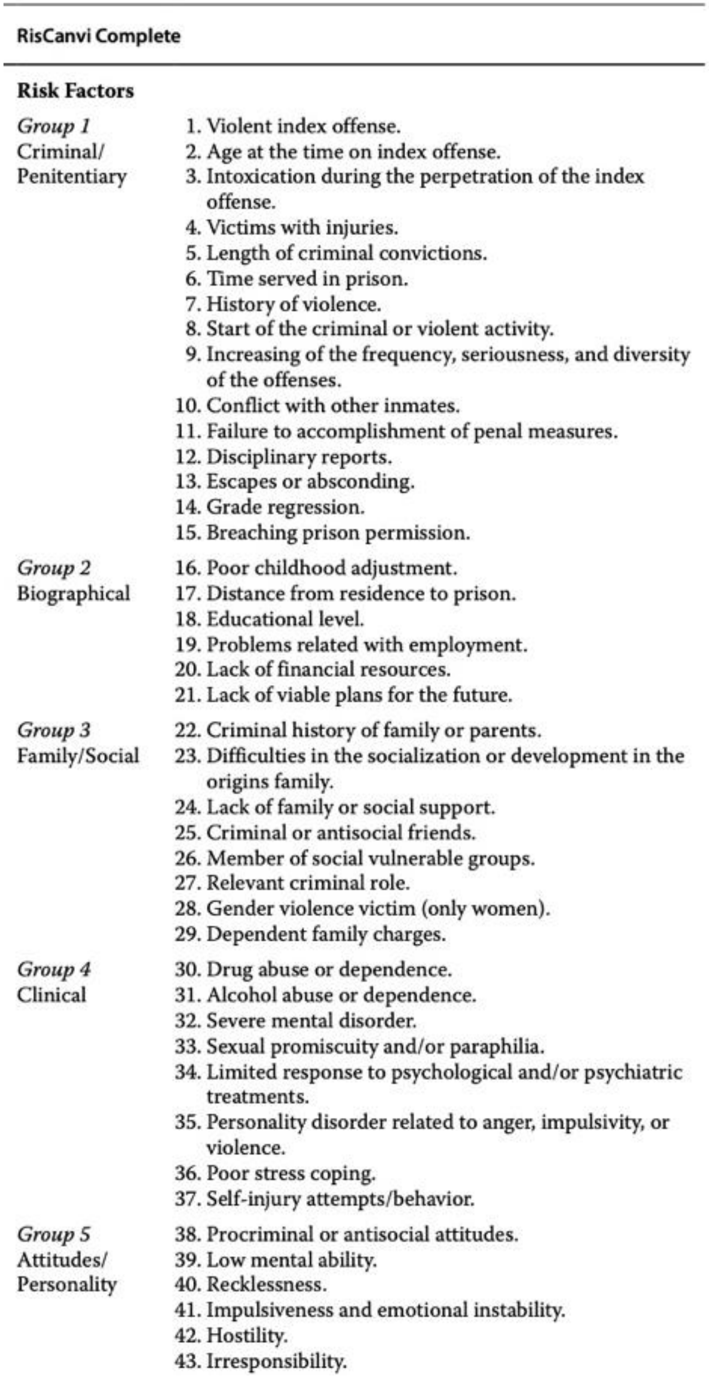
The query is directed at uncovering the weaknesses and shortcomings of the RisCanvi system. It specifically mentions "false positives," which refers to cases where the system incorrectly identifies someone as high risk. Addressing these issues is essential for improving the system's accuracy.

What changes would you propose to improve the RisCanvi system? If you don't see any issues, please explain.

This open-ended question invites respondents to suggest modifications or improvements to enhance the RisCanvi system. Professionals are encouraged to provide constructive feedback or, if they believe the system is functioning well, explain why they find it satisfactory. It can provide valuable insights for system refinement.

Annex 4: Risk Factors of RisCanvi Screening (RisCanvi-S) and Complete Assessment (RisCanvi-C) Versions

Source: Andrés-Pueyo et al., 2017: 260



RisCanvi Complete	
Risk Factors	
Group 1 Criminal/ Penitentiary	1. Violent index offense.
	2. Age at the time on index offense.
	3. Intoxication during the perpetration of the index offense.
	4. Victims with injuries.
	5. Length of criminal convictions.
	6. Time served in prison.
	7. History of violence.
	8. Start of the criminal or violent activity.
	9. Increasing of the frequency, seriousness, and diversity of the offenses.
	10. Conflict with other inmates.
	11. Failure to accomplishment of penal measures.
	12. Disciplinary reports.
	13. Escapes or absconding.
	14. Grade regression.
	15. Breaching prison permission.
Group 2 Biographical	16. Poor childhood adjustment.
	17. Distance from residence to prison.
	18. Educational level.
	19. Problems related with employment.
	20. Lack of financial resources.
	21. Lack of viable plans for the future.
Group 3 Family/Social	22. Criminal history of family or parents.
	23. Difficulties in the socialization or development in the origins family.
	24. Lack of family or social support.
	25. Criminal or antisocial friends.
	26. Member of social vulnerable groups.
	27. Relevant criminal role.
	28. Gender violence victim (only women).
	29. Dependent family charges.
Group 4 Clinical	30. Drug abuse or dependence.
	31. Alcohol abuse or dependence.
	32. Severe mental disorder.
	33. Sexual promiscuity and/or paraphilia.
	34. Limited response to psychological and/or psychiatric treatments.
	35. Personality disorder related to anger, impulsivity, or violence.
	36. Poor stress coping.
	37. Self-injury attempts/behavior.
Group 5 Attitudes/ Personality	38. Procriminal or antisocial attitudes.
	39. Low mental ability.
	40. Recklessness.
	41. Impulsiveness and emotional instability.
	42. Hostility.
	43. Irresponsibility.

Annex 5: Alternative variables for RisCanvi factors

These variables are a set of fields within the database that have a greater number of observations and contain the same information as the factors. The differences between the alternative variables and RisCanvi factors are as follows:

1. Type of coding: most of the factors are dichotomous and some alternative variables are categorical. For example, the RisCanvi factor for violent recidivism has a value of 1 for *yes* and 0 for *no*. The alternative variable has three possible responses: 1 for violent crime, 2 for non-violent crime, and 3 for violent crime. In contrast, the alternative variable has three possible responses: 1 for violent crime, 2 for nonviolent crime, and 9 for no response. Therefore, the alternative variable was harmonized to have the same values and the same responses as the RisCanvi factor.
2. Type of variable: some factors are represented as categorical variables when the alternative variables are continuous. For example, the age factor is coded into three groups: one for those under 22 years of age, two for those between 22 and 28 years of age, and three for those over 28 years of age. In contrast, the alternative variable is the age of the person deprived of liberty. As in the previous example, these variables were also transformed to represent the same values as the factors.

RisCanvi Factor	Description	Alternative variable	1st Evaluation	2nd Evaluation
Violent Index Offense	Refers to the use of physical violence, coercion or threats at the time of carrying out the base offense.	v72	v207	v260
Age at the time on index offense	Age of the subject at the time of committing the crime.	v139	v208	v261
Intoxication during the perpetration of the index offense	The subject was intoxicated by alcohol or psychotropic substances at the time of the commission of the base offense.	v76	v209	v262
Victims with injuries	Number of victims with physical or psychological injuries of moderate or severe severity, i.e. requiring professional care.	v77	v210	v263
Length of criminal convictions	Refers to the length of the effective sentence served by the inmate.	v86	v211	v264
Time served in prison	Time spent in prison in days since last admission for release, voluntary admission	Not possible to substitute	v212	v265

	or return from furlough/release up to the time of evaluation.			
History of violence	History of violent behavior prior to the base offense	v78	v213	v266
Start of the criminal or violent activity	Age of the subject at the time of committing the first violent incident or the first crime	v79	v214	v267
Increasing of the frequency, seriousness, and diversity of the offenses	The commission of different types of crimes and/or an increase in the seriousness or quantity of crimes	v80	v215	v268
Conflict with other inmates	It refers to whether the inmate generates arguments or fights; or if he also receives pressure from other subjects (victim of harassment or extortion) or if, on the contrary, he exerts harassment or extortion on fellow inmates.	v104	v216	v269
Failure to accomplishment of penal measures	This risk factor refers to the non-compliance with the penal measures imposed on the company.	v105	v217	v270
Disciplinary reports	Only the commission by the inmate of serious disciplinary offenses "art. 108RP" and very serious "art. 109RP" while serving the current or previous custodial sentence.	v106	v218	v271
Escapes or absconding	Evasions or escapes from a correctional facility or from previous incarcerations	v107	v219	v272
Grade regression	Negative evolution in relation to penitentiary treatment (previous or current incarcerations).	v108	v220	v273
Breaching prison permission	Refers to the non-return of furloughs (only the last incarceration is considered).	v109	v221	v274
Poor childhood adjustment	Behavioral problems or pattern of misbehavior common to childhood. Low	v29	v222	v275

	school performance or truancy is also considered.			
Distance from residence to prison	Refers to the residence where the inmate will go to live, e.g., on the occasion of a leave of absence	Not possible to substitute	v223	v276
Educational level	Level of education completed	v30	v224	v277
Problems related with employment	Chronic unemployment, job instability, frequent change of jobs, etc.	v31	v225	v278
Lack of financial resources	Economic level of the subject in the last year or before his admission to a penitentiary center in the event that it has been more than 12 months.	v32	v226	v279
Lack of viable plans for the future	Inability or unwillingness to consider viable medium- and long-term plans for the future.	v33	v227	v280
Criminal history of family or parents	First- or second-degree relatives who have committed criminal conduct are taken into account.	v34	v228	v281
Difficulties in the socialization or development in the origins family	The subject has been a victim or witness of violent behavior, abuse or neglect in the family environment (biological family, adoptive family, foster care, etc.).	v35	v229	v282
Lack of family or social support	This factor refers to the lack of support for regular contact with family and friends. Frequency of visits, letters and telephone contacts in the last 12 months are taken into account.	v36	v230	v283
Criminal or antisocial friends	The inmate is part of an organized criminal activity or has links to criminal networks.	v37	v231	v284
Member of social vulnerable groups	The inmate belongs to social groups, other than criminal gangs, at risk of committing criminal acts such as being a drug dealer, prostitution-related activities, etc.	v38	v232	v285

Relevant criminal role	The subject is considered and respected within the criminal subculture attentive to his criminal record.	v39	v233	v286
Gender violence victim (only women)	The inmate is a victim of interpersonal violence (physical, sexual or psychological) by her partner or ex-partner (only the last 12 months are considered).	v40	v234	v287
Dependent family charges	The inmate is responsible for minor children, parents, etc.	v41	v235	v288
Drug abuse or dependence	Subjects whose lives are negatively affected by the use of drugs (legal or illegal).	v42	v236	v289
Alcohol abuse or dependence	Alcohol consumption interferes negatively with the subject's family, work or social life.	v43	v237	v290
Severe mental disorder	Subject diagnosed with severe mental disorder second DSM-V or ICD-10.	v44	v238	v291
Sexual promiscuity and/or paraphilia	It refers to whether the subject presents risky sexual promiscuity, irresponsible hypersexuality, violent sexual behavior, sexual deviance, sexual perversion or disorder of sexual preference.	v45	v239	v292
Limited response to psychological and/or psychiatric treatments	The inmate shows no adherence (or poor results) to psychological, psychiatric or pharmacological treatment.	v46	v240	v293
Personality disorder related to anger, impulsivity, or violence	The subject must meet the criteria established in the DSM-V or ICD-10 for a cluster B personality disorder. The possibility that the subject suffers from habitual anger, impulsivity or violent behavior is also considered.	v47	v241	v294
Poor stress coping	The inmate has a lack of cognitive and behavioral resources in the face of a given stressful situation.	v48	v242	v295

Self-injury attempts/behavior	It refers to a history of self-inflicted violence behaviors understood as suicide attempts and/or self-injury.	v49	v243	v296
Pro-criminal or antisocial attitudes	Subject who displays manifest attitudes of an antisocial nature characteristic of criminal subcultures with the justification of the use of violence, as well as criminal behavior. criminal behavior.	v50	v244	v297
Low mental ability	The inmate shows poor language proficiency (knows the language), as well as inadequate performance on new tasks and poor reading comprehension. The evaluation is performed by means of a diagnostic instrument	v51	v245	v298
Recklessness	Engagement in risky activities and preference for new experiences rather than routines	v52	v246	v299
Impulsiveness and emotional instability	Propensity to react unexpectedly and explosively both behaviorally and emotionally. Shows dramatic fluctuations in mood or behavior.	v53	v247	v300
Hostility	Refers to aggressive behavior (either verbal or physical) toward others.	v54	v248	v301
Irresponsibility	The subject does not fulfill the obligations or commitments acquired with others (has little sense of responsibility).	v55	v249	v302

Annex 6: Logit and multinomial regression

As a first approach to the RisCanvi quantitative analysis, a set of logistic regressions and a subsequent multinomial logistic regression were performed. The logistic models were performed for two reasons. First, the type of behavioral variables corresponds to categorical variables, therefore, logistic models fit the statistical distribution of RisCanvi's variables (Wooldridge, 2012). Second, the same type of regressions has been used by authors such as Andrés-Pueyo, Arbach-Lucioni, and Redondo (2018) to predict the risk level for specific behaviors.[18] We compiled three binomial logistic regression tables corresponding to the three risk levels for each behavior found.

The multiple regressions models were developed by Eticas.ai using the dataset CPRR. The next tables are the regression results for each model.

Each table is related to a risk level and contains four models corresponding to the four RisCanvi's behaviors contained in the CPRR dataset. The columns represent the causal relationship of the factors and the risk prediction for each behavior. The table shows the **betas (values without parentheses)** that represent the expected change in log odds of having the outcome per unit change of each RisCanvi's factor. **The value inside parentheses represents the standard errors** determining the statistical significance of the beta. Statistical significance helps to determine if the relationship between variables exists or not in a statistical perspective.

Additionally, in the table we have included a **McFadden's R-squared to explain the variance of the models**. In the case of regression models, if the R's value is equal to "1", it means that the models are perfectly fitted, however a "perfect" scenario is very rare and it is frequent that a R-squared value equal to 1 shows biased estimations, overfitting, high correlation of variables, distribution assumptions violation, low data, etc. All tables show a R-squared value equal to 1 in this case.

Figure 15: Low risk prediction logistic regression model

	Model I	Model II	Model III	Model IV
	Self-directed Violence	Intra-institutional Violence	Violent recidivism	Non-compliance
Constant	45.99	21.31	49.72	37.94
	(-1133.27)	(-164.68)	(nan)	(-176.3)
Violent Index Offense	-1.27	-1.25	-2.03	-0.45
	(-522)	(-109.32)	(nan)	(-99.4)
Intoxication during the perpetration of the index offense	-6.73	-2.06	-9.42	-4.58
	(-416.94)	(-97.2)	(nan)	(-105.17)
History of violence	-18.91	-5.64	-34.96	-17.23
	(-478.88)	(-78.06)	(nan)	(-68.92)
Increasing of the frequency, seriousness, and	-11.89	-8.3	-25.72	5.95

diversity of the offenses				
	(-417.77)	(-72.66)	(nan)	(-58.32)
Conflict with other inmates	2.3	-7.29	-18.9	6.76
	(-325.41)	(-123.16)	(nan)	(-168.18)
Failure to accomplishment of penal measures	-2.98	-2	23.36	-31.78
	(-198.63)	(-102.14)	(nan)	(-76.84)
Disciplinary reports	-1.94	-8.58	-4.84	-13.74
	(-207.59)	(-72.12)	(nan)	(-78.49)
Escapes or absconding	-27.31	4.11	-27.28	-37.42
	(-494.86)	(-140.77)	(nan)	(-262.03)
Grade regression	-11.39	-2.88	6.07	-39.06
	(-732.53)	(-113.93)	(nan)	(-123.58)
Breaching prison permission	8.09	2.93	0.34	-0.58
	(-311.7)	(-119.97)	(nan)	(-100.2)
Poor childhood adjustment	-7.24	3.97	-6.29	2.2
	(-379.35)	(-84.23)	(nan)	(-61.87)
Problems related with employment	9.41	-5.13	-0.77	1.78
	(-517.04)	(-93.46)	(nan)	(-46.34)
Lack of financial resources	1.65	0.93	-10.87	-42.95
	-381.9	-138.09	(nan)	-88.86
Lack of viable plans for the future	-7.83	-8.02	-11.26	2.22
	-295.61	-94.49	(nan)	-281.45
Criminal history of family or parents	-8.05	-0.13	-9.54	26.53
	-408.8	-137.49	(nan)	-109.82
Difficulties in the socialization or development in the origins family	8.6	-3.25	4.91	-36.77
	-435.55	-73.45	(nan)	-97.83
Lack of family or social support	6	-0.91	-3.52	-39.41
	-234.74	-141.68	(nan)	-130.39
Criminal or antisocial friends	-0.24	-2.73	16.58	8.28
	-609.91	-95.29	(nan)	-76.97

Member of social vulnerable groups	-0.58	-2.66	5.09	1.95
	-196.7	-131.51	(nan)	-153.79
Relevant criminal role	-8.19	-1.47	1.98	3.18
	-957.14	-203.61	(nan)	-269.64
Gender violence victim (only women)	17.07	1	13.61	-5.29
	-250.16	-87.79	(nan)	-115.4
Dependent family charges	-27.95	-6.03	-28.43	10.26
	-428.44	-89.43	(nan)	-89.87
Drug abuse or dependence	-5.67	2.21	-14.28	0.71
	-480.8	-71.18	(nan)	-190.77
Alcohol abuse or dependence	-8.95	0.51	-4.82	17.72
	-1873.42	-545.38	(nan)	-6915.69
Severe mental disorder	-0.85	0.44	10.17	-16.78
	-26596.17	-3962.63	(nan)	-357.55
Sexual promiscuity and/or paraphilia	-6.32	0.58	-19.23	-4.29
	-180.16	-97.57	(nan)	-168.12
Limited response to psychological and/or psychiatric treatments	-10.55	-5.85	-5.87	13.48
	-634.97	-94.49	(nan)	-290.56
Personality disorder related to anger, impulsivity, or violence	-18.32	2.41	2.85	8.08
	-135.32	-64.27	(nan)	-298.69
Poor stress coping	-10.86	-3.05	18.52	1.24
	-443.75	-83.86	(nan)	-168.24
Self-injury attempts/behavior	-11.41	-12.03	-17.86	-4.22
	-307.72	-209.33	(nan)	-164.54
Pro-criminal or antisocial attitudes	0.79	-4.98	-0.47	-22.7
	-663.49	-144.52	(nan)	-166.25
Recklessness	-3.85	-6.14	-14.49	13.01
	-165.12	-60.19	(nan)	-158.77
Impulsiveness and emotional instability	4.44	-4.43	-15.47	12.77

	-623.33	-213.23	(nan)	-260.21
Hostility	1.09	0.64	-16.92	-4.49
	-398.02	-92.05	(nan)	-154.48
Irresponsibility	14.7	3.11	3.75	-4.23
	-593.45	-143.57	(nan)	-254.59
Age at the time on index offense (Between 22 and 28 years old)	17.51	-0.04	13.09	-14.31
	-626.63	-80.89	(nan)	-137.02
Age at the time on index offense (Over 28 years old)	-10.33	-6.47	-5.95	-0.37
	-472.87	-90.03	(nan)	-112.23
Victims with injuries (1 victim)	-17.95	-3.66	-45.44	-1.96
	-543.01	-128.19	(nan)	-152.38
Victims with injuries (Over 1 victim)	-3.13	-1.42	-17.31	-2.41
	-522.34	-87.03	(nan)	-103.79
Length of criminal convictions (Between 2 and 6 years)	-6.43	-4.19	-16.1	14.94
	-388.15	-175.42	(nan)	-130.69
Length of criminal convictions (Over 6 years)	-12.56	-6.77	-14.56	-2.91
	-761.14	-113.66	(nan)	-114.36
Time served in prison (Between 1 and 3 years)	-3.78	-1.92	-0.73	-32.16
	-796.56	-152.6	(nan)	-223.41
Time served in prison (Over 3 years)	7.95	5.84	-5.37	11.33
	-572.45	-103.19	(nan)	-86.3
Start of the criminal or violent activity (Between 16 and 30 years old)	-6.52	11.26	-5.36	18.58
	-635.07	-121.85	(nan)	-80.44
Start of the criminal or violent activity (Over 30 years ol)	12.25	-3.25	12.95	19.6
	-873.2	-306.07	(nan)	-6883.8

Distance from residence to prison (Between 100 and 300 km)	-3.75	-0.85	0.14	21
	-528.45	-217.48	(nan)	-178.66
Distance from residence to prison (Over 300 km)	0.2	0.15	0.19	2.07
	-1.17E+11	-3566795.93	(nan)	-469.18
Educational level (High Graduate)	-5.48	0.67	-2.2	-3.97
	-239.6	-82.7	(nan)	-62.91
Educational level (Middle Undergraduate)	9.39	11.86	50.55	23.08
	-662.2	-127.03	(nan)	-186.81
Number of observations	308	308	308	308
R-squared	1	1	1	1

Probability under the assumption of no effect

*: p<.1	**: p<.05	***: p<.01

Figure 16: Middle risk logistic regression model

	Model I	Model II	Model III	Model IV
	Self-directed Violence	Intra-institutional Violence	Violent recidivism	Non-compliance
Constant	-95.29	-13.78***	-4.27**	-5.23**
	(-281.89)	(-5.26)	(-2.02)	(-2.16)
Violent Index Offense	16.67	-3.92**	-0.93	-2.24*
	(-312.13)	(-1.98)	(-0.81)	(-1.17)
Intoxication during the perpetration of the index offense	14.02	0.77	1.24*	-1.4
	(-175.82)	(-1.62)	(-0.74)	(-1.08)
History of violence	27.73	-0.6	1.67**	2.97**
	(-304.52)	(-1.51)	(-0.73)	(-1.37)
Increasing of the frequency, seriousness, and diversity of the offenses	10.91	3.09**	1.53**	-0.35
	(-122.87)	(-1.54)	(-0.76)	(-0.93)
Conflict with other inmates	-19.56	0.03	-0.02	3.17*
	(-406.91)	(-1.7)	(-0.97)	(-1.65)

Failure to accomplishment of penal measures	21.39	5.91**	-0.55	6.49***
	(-356.41)	(-2.34)	(-0.69)	(-1.88)
Disciplinary reports	-2.34	5.34**	2.90***	1.22
	(-369.98)	(-2.42)	(-0.97)	(-1.13)
Escapes or absconding	47.93	0.16	4.14**	0.27
	(-461.28)	(-1.81)	(-1.67)	(-1.93)
Grade regression	2.38	-3.17	-2.76**	6.03***
	(-345.89)	(-1.96)	(-1.21)	(-1.66)
Breaching prison permission	1.91	-1.88	-1.15	2.51
	(-188.97)	(-2.82)	(-1.4)	(-1.86)
Poor childhood adjustment	9.07	-1.73	0.23	-1.6
	(-362.45)	(-1.22)	(-0.66)	(-1.1)
Problems related with employment	-10.87	0.68	-1.71**	-0.1
	(-188.31)	(-1.12)	(-0.81)	(-0.81)
Lack of financial resources	-22.48	2.96**	1.98**	7.23***
	(-252.41)	(-1.51)	(-0.82)	(-1.89)
Lack of viable plans for the future	57.78	5.11**	-1.19	1.7
	(-284.5)	(-2.36)	(-0.99)	(-1.37)
Criminal history of family or parents	13.83	0.32	0.77	-4.75***
	(-276.41)	(-1.63)	(-0.77)	(-1.63)
Difficulties in the socialization or development in the origins family	-44.4	0.53	-1.06	6.73***
	(-327.98)	(-1.25)	(-0.84)	(-1.89)
Lack of family or social support	-37.63	-4.09	0.47	5.33***
	(-177.66)	(-3)	(-1.09)	(-1.81)
Criminal or antisocial friends	14.79	1.02	-2.40**	-0.36
	(-242.48)	(-1.25)	(-1.21)	(-1.41)
Member of social vulnerable groups	-0.95	1.1	-3.13**	-5.80**
	(-163.96)	(-1.85)	(-1.25)	(-2.29)
Relevant criminal role	47.04	4.69	-9.36	3.44
	(-481.92)	(-3.11)	(-18.37)	(-2.86)

Gender violence victim (only women)	-29.97	-0.74	-1.86**	0.11
	(-152.25)	(-1.33)	(-0.74)	(-0.81)
Dependent family charges	49.35	1.17	1.16	0.34
	(-432.39)	(-1.29)	(-0.83)	(-1.13)
Drug abuse or dependence	17.87	-0.23	0.89	1.21
	(-139.09)	(-1.29)	(-0.7)	(-1.26)
Alcohol abuse or dependence	-81.39	-6.19	-1.47	2.12
	(-539.1)	(-4.32)	(-2.32)	(-2.23)
Severe mental disorder	0.07	-1.08	0.18	3.58*
	(-1570.66)	(-39.96)	(-1.36)	(-1.86)
Sexual promiscuity and/or paraphilia	10.02	-1.42	0.18	-3.58**
	(-214.26)	(-1.36)	(-0.71)	(-1.73)
Limited response to psychological and/or psychiatric treatments	31.29	4.31*	-0.77	-3.01*
	(-359.94)	(-2.21)	(-1.12)	(-1.81)
Personality disorder related to anger, impulsivity, or violence	43.25	2.33	1	0.27
	(-128.44)	(-1.86)	(-0.74)	(-1.2)
Poor stress coping	-5.38	-1.38	1.03	0.69
	(-307.83)	(-1.76)	(-0.8)	(-1.15)
Self-injury attempts/behavior	94.3	6.87**	1.79	5.94**
	(-369.83)	(-2.73)	(-1.33)	(-2.55)
Pro-criminal or antisocial attitudes	-69.62	-7.21**	-0.7	3.00*
	(-403.77)	(-3.4)	(-1.08)	(-1.7)
Recklessness	9.13	3.77*	0.36	-3.17**
	(-203.18)	(-1.98)	(-0.83)	(-1.37)
Impulsiveness and emotional instability	-91.01	-4.78*	-2.92*	-9.27***
	(-387.64)	(-2.53)	(-1.6)	(-3.14)
Hostility	-27.64	-2.25	1.75**	0.92
	(-421.2)	(-1.55)	(-0.84)	(-1.06)
Irresponsibility	-26.99	-2.96*	0.84	1.35

	(-324.41)	(-1.53)	(-0.89)	(-1.35)
Age at the time on index offense (Between 22 and 28 years old)	-45.51	-2.63*	-0.41	0.26
	(-530.35)	(-1.5)	(-1.04)	(-1.21)
Age at the time on index offense (Over 28 years old)	10.26	5.36**	0.59	1.17
	(-213.62)	(-2.16)	(-0.91)	(-1.23)
Victims with injuries (1 victim)	29.81	-1.41	3.80***	1.7
	(-348.06)	(-3.51)	(-1.23)	(-1.78)
Victims with injuries (Over 1 victim)	14.84	0.39	0.1	1.11
	(-236.93)	(-1.43)	(-0.97)	(-1.15)
Length of criminal convictions (Between 2 and 6 years)	15.18	2.2	1.83	-5.17***
	(-212.51)	(-2.06)	(-1.19)	(-1.87)
Length of criminal convictions (Over 6 years)	-5.75	4.15*	-0.43	-1.19
	(-446.07)	(-2.13)	(-1.1)	(-1.87)
Time served in prison (Between 1 and 3 years)	16.21	2.2	-1.42	6.73***
	(-428.99)	(-2.49)	(-1.58)	(-2.43)
Time served in prison (Over 3 years)	-5.86	1.39	1.57	0.12
	(-257.55)	(-1.72)	(-1.24)	(-1.3)
Start of the criminal or violent activity (Between 16 and 30 years old)	21.39	0.79	1.44	2.14
	(-459.57)	(-2.26)	(-1.48)	(-1.63)
Start of the criminal or violent activity (Over 30 years ol)	44.87	-0.52	0.05	-14.75
	(-322.86)	(-2.32)	(-1.54)	(-19.05)
Distance from residence to prison (Between 100 and 300 km)	-8.84	-4.24	-10.02	-1.29
	(-6510.12)	(-7.72)	(-18.39)	(-1.87)
Distance from residence to prison (Over 300 km)	-0.66	-0.41	-1.95	-4.47

	(-23783528776.05)	(-70.31)	(-20.54)	(-20.77)
Educational level (High Graduate)	-10	-2.73	-0.32	1.78*
	(-118.3)	(-1.71)	(-0.78)	(-1)
Educational level (Middle Undergraduate)	15.94	1.12	-2.93**	-7.27***
	(-170.75)	(-2.26)	(-1.2)	(-2.45)
Number of observations	308	308	308	308
R-squared	1	1	1	1

Probability under the assumption of no effect

*: p<.1	**: p<.05	***: p<.01

Figure 17: High risk logistic regression model

	Model I	Model II	Model III	Model IV
	Self-directed Violence	Intra-institutional Violence	Violent recidivism	Non-compliance
Constant	-28.75	-16.8	-30.02	-17.11
	(-410.97)	(-315.74)	(-3918.75)	(-274.09)
Violent Index Offense	2.22	5.55	5.68	2.57
	(-214.68)	(-163.8)	(-1611.19)	(-296.28)
Intoxication during the perpetration of the index offense	1.54	4.72	9.24	0.24
	(-128.91)	(-117.57)	(-2221.95)	(-167.49)
History of violence	2.85	3.93	3.77	-0.47
	(-150.14)	(-126.21)	(-1830.77)	(-247.79)
Increasing of the frequency, seriousness, and diversity of the offenses	6.46	0.07	9.65	1.76
	(-153.06)	(-112.48)	(-952.41)	(-186.17)
Conflict with other inmates	0.87	8.76	-2.95	-7.38
	(-141.72)	(-117.16)	(-3184.08)	(-374.7)
Failure to accomplishment of penal measures	-6.99	-3.53	1.67	2.48
	(-172.18)	(-144.43)	(-2039.06)	(-183.99)
Disciplinary reports	10.07	-0.73	7.03	-0.67
	(-377.76)	(-147.37)	(-1365.27)	(-293.77)
Escapes or absconding	2.97	0.13	3.62	10.1
	(-202.64)	(-128.98)	(-1868.59)	(-276.36)
Grade regression	8.14	8.44	-0.16	11.34
	(-148.63)	(-115.78)	(-2350.85)	(-286.32)

Breaching prison permission	-5.94	0.14	1.57	7.57
	(-402.01)	(-247.18)	(-2648.89)	(-227.43)
Poor childhood adjustment	3.13	1.32	0.82	-3.39
	(-136.96)	(-122.98)	(-1321.68)	(-216.28)
Problems related with employment	-4.03	0.36	-1.34	-2.22
	(-310.63)	(-123.18)	(-2363.01)	(-223.59)
Lack of financial resources	7.58	-5.07	1.37	5.58
	(-159.76)	(-150.2)	(-940)	(-191.7)
Lack of viable plans for the future	0.66	-0.32	9.01	-1.79
	(-231.29)	(-148.34)	(-1850.5)	(-256.92)
Criminal history of family or parents	-3.1	2.57	3.8	5.17
	(-187.43)	(-160.37)	(-2056.82)	(-223.16)
Difficulties in the socialization or development in the origins family	2.66	0.83	1.67	-0.45
	(-188.37)	(-134.54)	(-1996.39)	(-206.34)
Lack of family or social support	4.11	-1.39	7.56	8.3
	(-211.58)	(-243.01)	(-4141.64)	(-220.75)
Criminal or antisocial friends	0.52	1.11	0.04	-4.16
	(-255.36)	(-176.37)	(-2713.92)	(-316.1)
Member of social vulnerable groups	-8.66	-2.27	1.43	0.65
	(-175.84)	(-224.02)	(-1752.13)	(-333.74)
Relevant criminal role	-6.58	-2.16	6.05	-0.98
	(-365.83)	(-264.07)	(-2472.34)	(-1189.55)
Gender violence victim (only women)	-4.79	1.17	1.28	-2.72
	(-172.22)	(-96.74)	(-1393.68)	(-193.56)
Dependent family charges	4.96	3.38	12.03	-3.86
	(-256.26)	(-115.76)	(-2653.8)	(-293.33)
Drug abuse or dependence	2.86	2.42	-0.18	0.95
	(-147.11)	(-156.51)	(-825.55)	(-238.88)
Alcohol abuse or dependence	1.15	-0.89	1.83	-4.37
	(-476.07)	(-382.84)	(-3205.91)	(-802.38)
Severe mental disorder	-4.51	-2.97	-7.06	-2.75
	(-645.21)	(-325.86)	(-4971.78)	(-423.11)
Sexual promiscuity and/or paraphilia	2.81	-3.98	-2.69	3.06
	(-200.25)	(-103.75)	(-1645.78)	(-239.47)

Limited response to psychological and/or psychiatric treatments	-3.68	-2.67	1.72	2.89
	(-295.22)	(-210.39)	(-1521.24)	(-246.7)
Personality disorder related to anger, impulsivity, or violence	1.62	-2.29	-6.54	-1.28
	(-164.12)	(-129.63)	(-1517.28)	(-240.35)
Poor stress coping	8.04	2.03	-9.09	-2.19
	(-141.88)	(-109.03)	(-1431.06)	(-258.78)
Self-injury attempts/behavior	-6.7	2.5	8.57	-1.71
	(-205.65)	(-180.65)	(-2554.86)	(-394.26)
Pro-criminal or antisocial attitudes	7.94	7.15	8.39	3.4
	(-229.8)	(-107.29)	(-2328.84)	(-451.75)
Recklessness	3.58	3.17	8.43	0.06
	(-113.59)	(-119.7)	(-1311.59)	(-274.56)
Impulsiveness and emotional instability	11.21	8.85	6.52	7.36
	(-214.92)	(-176.28)	(-2046.34)	(-275.52)
Hostility	2.05	3.89	1.14	-2.71
	(-187.45)	(-137.9)	(-1239.4)	(-235.42)
Irresponsibility	1.38	3.15	-10.42	-0.91
	(-217.58)	(-130.07)	(-1580.75)	(-191.33)
Age at the time on index offense (Between 22 and 28 years old)	9.18	1.09	-9.22	-1.16
	(-203.63)	(-169.23)	(-2140.41)	(-218.89)
Age at the time on index offense (Over 28 years old)	4	0.27	0.87	2.96
	(-289.48)	(-158.83)	(-3375.82)	(-256.03)
Victims with injuries (1 victim)	11.24	2.44	2.1	3.77
	(-242.2)	(-253.58)	(-4260.36)	(-292.51)
Victims with injuries (Over 1 victim)	-5.55	2.46	1.56	-0.87
	(-238.76)	(-117.17)	(-1138.67)	(-166.3)
Length of criminal convictions (Between 2 and 6 years)	2.83	-0.83	1.59	7.7
	(-229.68)	(-151.32)	(-1276.83)	(-253.18)
Length of criminal convictions (Over 6 years)	-0.81	-0.12	2.08	1.58
	(-340.89)	(-184.91)	(-2075.75)	(-228.69)
Time served in prison (Between 1 and 3 years)	2.69	-1.63	0.35	-0.41
	(-283.11)	(-298.23)	(-4746.22)	(-278.65)
Time served in prison (Over 3 years)	-9.05	-5.99	-9.62	-6.63

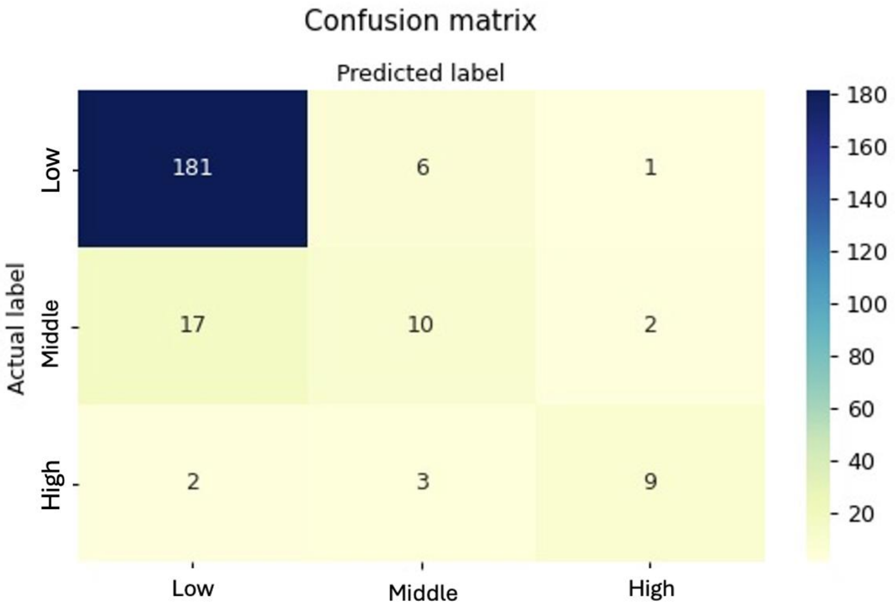
	(-296.08)	(-119.98)	(-2814.15)	(-188.24)
Start of the criminal or violent activity (Between 16 and 30 years old)	-2.24	-7.46	-3.86	-6.1
	(-301.44)	(-254.75)	(-4136.17)	(-260.77)
Start of the criminal or violent activity (Over 30 years old)	2.29	2.26	-5.73	-1.58
	(-328.55)	(-267.21)	(-3701.66)	(-1182.6)
Distance from residence to prison (Between 100 and 300 km)	0.12	0.68	10.02	0.01
	(-955.84)	(-360.75)	(-3359.35)	(-278.26)
Distance from residence to prison (Over 300 km)	-0.18	-0.11	-0.1	-0.2
	(-7645330.74)	(-324506.76)	(-180403255.3)	(-13644.67)
Educational level (High Graduate)	5.28	-0.51	4.02	-1.69
	(-203.76)	(-246.97)	(-1904.42)	(-186.55)
Educational level (Middle Undergraduate)	-13.2	-12.29	-12.7	-2.36
	(-333.35)	(-258.95)	(-2626.19)	(-316.27)
Number of observations	308	308	308	308
R-squared	1	1	1	1

Probability under the assumption of no effect

*: p<.1	**: p<.05	***: p<.01

Annex 7: Confusion matrix and accuracy table on multinomial model.

Figure 18: Confusion matrix of multinomial logistic regression.



	Low	Middle	High
Low	181	6	1
Middle	17	10	2
High	2	3	9

Figure 19: Fitting metrics of multinomial regression model

	Precision	Recall	F1-Score	Support
Low	0.91	0.96	0.93	188
Middle	0.53	0.34	0.42	29
High	0.75	0.64	0.69	14
Accuracy			0.87	231
Macro Avg	0.73	0.65	0.68	231
Weighted Avg	0.85	0.87	0.85	231

Annex 8: Table to interpret Factor Analysis.

This table is designed to facilitate the interpretation of the factor analysis panels found in the report, In the rows, we find three possible scenarios: The number of people presenting the factor (*Yes*) can decrease, stay the same (flat), or increase. The presence of a factor may have more, the same, or fewer people at the low, medium, and high levels. For instance, in Figure 6, the number of people with *Yes* in self-directed violence behavior is decreasing, going from 593 to 265, and finally, to 154 people with high-risk levels.

Figure 20: Panel interpretation tool

Possible scenarios		Yes %		
		Decreasing	Flat	Increasing
Count (Absolute value)	Decreasing	It is not a major factor in the risk assessment.	It is not a major factor in the risk assessment.	It is a factor that prevails in medium and high-risk levels. It is not a factor characteristic of people at high risk levels.
	Flat	No cases were found in the analysis.	It is a general condition of inmates. It does not play a fundamental role in risk assessment.	It is a factor that prevails in high risk levels. This factor is present in the other levels, being a common characteristic of the population.
	Increasing	No cases were found in the analysis.	It is a particular characteristic of the risk levels but does not represent a differentiator in the evaluation.	The weight of this factor is significant. This factor is characteristic of the middle and high-level population.

The columns in the table represent the trend of observations expressed as a percentage of the total. Using the same example as in Figure 6 (see main report), the percentage of *Yes* at the low level is 49%, at the medium level 68% and, finally, 77% at the high-risk level. Therefore, in this case, there is an increase in the percentage through the risk levels.

Adversarial Audit