Risk Assessment in Complex Data Settings: Algorithmic Fairness and Causal Inference

Marzieh Karimi-Haghighi

TESI DOCTORAL UPF / year of the thesis: 2022

THESIS SUPERVISOR
Carlos Castillo
Department of Information and Communication Technologies

"A man should not be judged by his fame, power, or money, but rather by how much love he gives to others."
Sandranil Biswas

Acknowledgments

Many thanks to my supervisor, Carlos Castillo, for his great guidance and support, kindness and high sense of responsibility during my Ph.D. He was always available and very fast in sending feedback. My Ph.D. under his supervision was also a good opportunity to meet his lovely wife and their lovely cats!

Thanks to other people and collaborators who helped me and I learned from them during my Ph.D. journey: Songul Tolan, Davinia Hernandez, Marius Miron, Kristian Lum, Antonio Andres Pueyo, Aroa Arrufat, Emilia Gomez, Manuel Capdevila, and Veronica Moreno.

Special thanks to my lovely husband Morteza for his emotional support and to my dear mom and dad (rest in peace) and siblings for their remote support.

Also, thanks to the UPF administration staffs who are always at hand to speed up paper work.

This thesis was supported by UPF Ph.D. fellowship, HUMAINT programme (Human Behaviour and Machine Intelligence), Centre for Advanced Studies, Joint Research Centre, European Commission, EU-funded ”SoBigData++” project, ICREA Academia programme, and the National Research Agency of the Spanish Ministry of Science.

Abstract

Structured risk assessment tools are sometimes appropriate alternatives to traditional prediction methods due to their higher accuracy and scalability. However, there are still challenges with regards to these tools such as limited predictive performance, different validity with respect to some demographics, and effectiveness. In this thesis, we try to address these issues in the two application areas of recidivism risk in criminal justice and dropout risk in higher education domain.

We suggest a scenario to efficiently save time, expenses and staff in a data-driven assessment of violent recidivism risk. Using Machine Learning (ML) methods, we model risk change with an AUC of 0.74-0.78 and select only a fraction of inmates with the highest probability of risk change for the next evaluation. We include a cost-benefit analysis which leads to fewer evaluations in exchange for some small number of missed/undetected changes. Importantly, by adjusting decision boundaries, we mitigate the model's disparate impact in the rate of evaluation across some demographics.

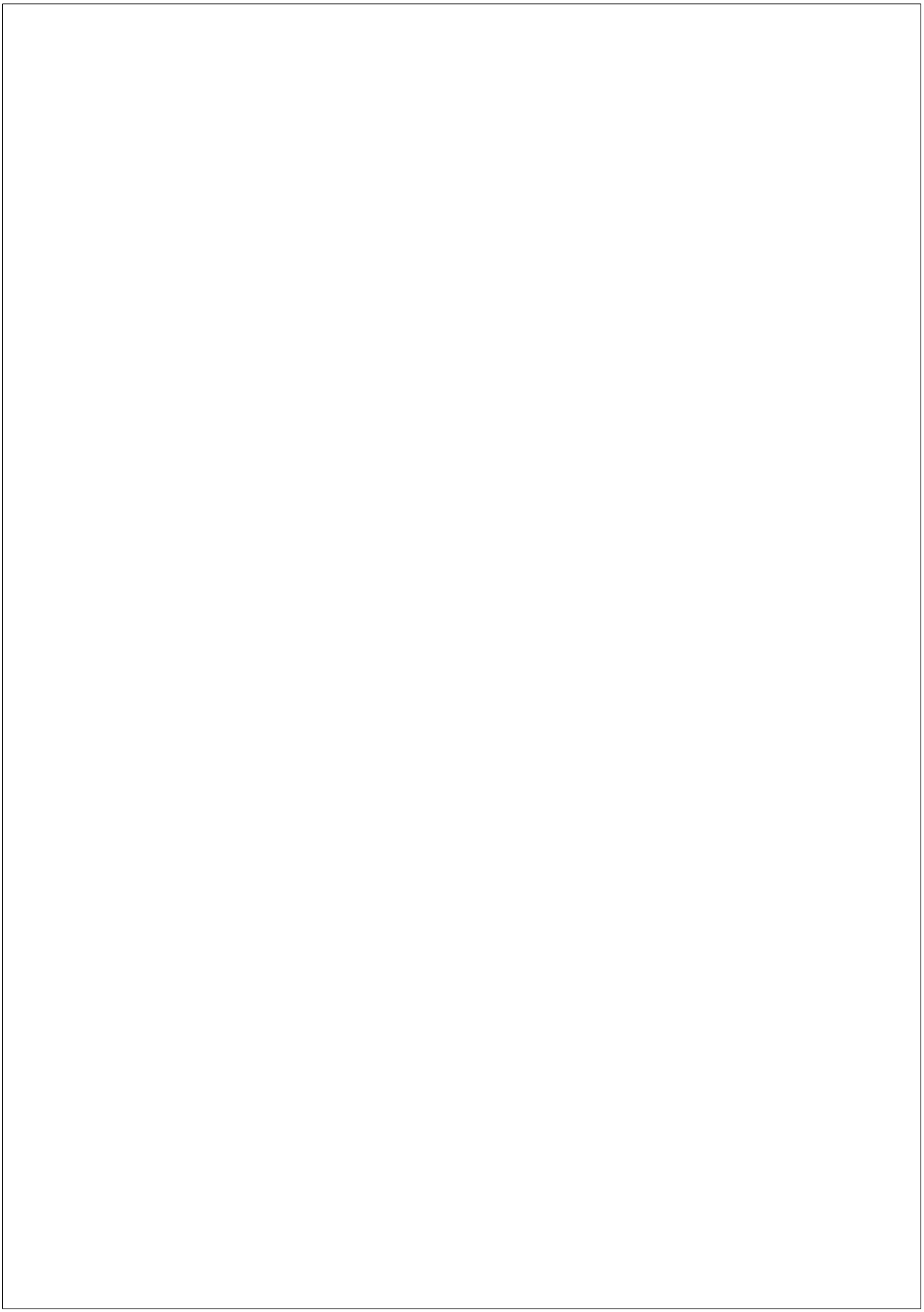
Using ML methods, we try to assess risks in a more accurate manner and with algorithmic fairness guarantees. We obtain ML-based prediction models with AUC of 0.76 and 0.73 in predicting violent and general recidivism respectively, which are a little more accurate than the manually-created formula used in RisCanvi. We also create ML prediction models for dropout and underperformance risks in undergraduate students with AUC of 0.77-0.78 based on the data available at the enrollment time, which is consistent with the AUC values in similar previous studies. To improve algorithmic fairness of risk prediction models across some sensitive groups, we minimize the disparities in generalized false positive rate through a mitigation process while maintaining calibration across groups.

We determine the effect of a treatment on an outcome risk using statistical causal inference methods. In several scenarios, we show that a reduction in university workload (first year credits) reduces dropout risk. We also indicate that conditional release (C.R.) can reduce general and violent recidivism, and seems effective at promoting a safe and supervised

return to the community while protecting public safety. As well, in contrast with the policy of assigning C.R. to people estimated as low-risk by the risk assessment tools, our results show that granting C.R. to cases with medium estimated risk may lead to potentially larger reductions in recidivism rates.

Preface

Contents

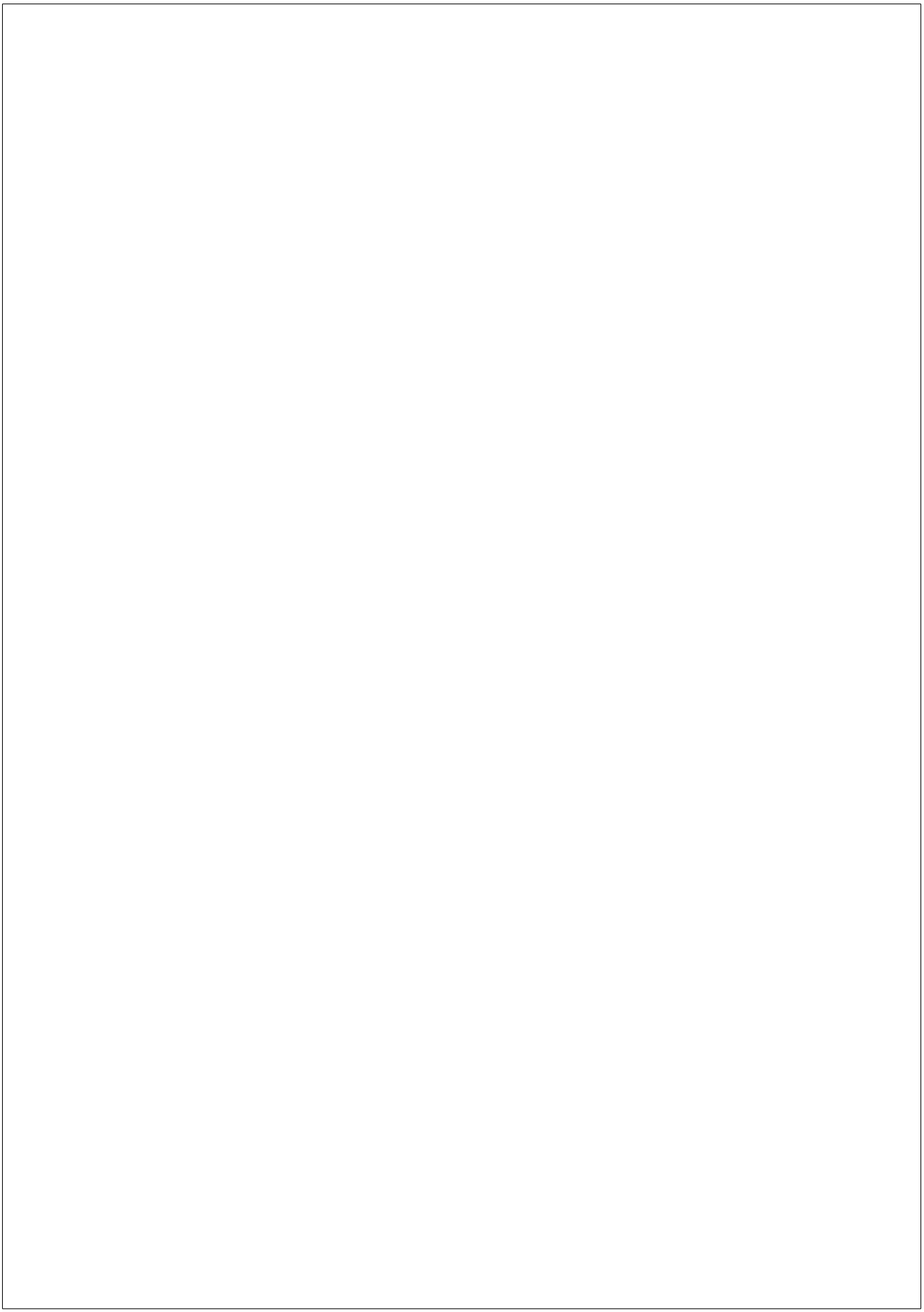


List of Figures

List of Tables

Part I

Introduction and Background



Chapter 1

INTRODUCTION

1.1 Introduction

Risk assessment is the systematic process of identifying and evaluating potential risks and their consequences that may affect individuals and/or society or environment [Rausand, 2013]. This process is necessary in highly consequential decisions such as:

- estimating the risk of human exposure to chemicals in order to maintain **public health** [Asante-Duah, 2002],
- assessing the risk to **information security** to protect business information assets [Shameli-Sendi et al., 2016],
- **auditing** risk assessment to prevent various business risks such as fraud risk [Allen et al., 2006],
- estimating multiple risks in **criminal justice** system such as violence, and recidivism risks [Kemshall, 2003] to keep the community safety.

The first generation of risk assessments were professional judgments made by clinical or correctional staff who relied on their personal training and experience. These unstructured clinical judgments were subject to human error and cognitive biases [Bell and Mellor, 2009, Lopez, 1989].

Also, there is no clear evidence that decisions were consistent for all professionals, because their judgements were based on individual case analysis or professional experience, without considering relevant risk factors, method for combining them, or applicable theory.

The second generation were actuarial assessments of risk that emerged in the 1970s [Hoffman and Beck, 1974, Nuffield, 1982, Bonta et al., 1998, Hanson and Bussiere, 1998]. They were based on numeric predictions derived from analyses of static risk factors. This generation was more accurate and reliable than professional judgments, when making predictions of human behaviour, because of incorporation of actuarial, objective, and evidence-based criteria for assessing risk [Dawes et al., 1989, Grove et al., 2000, Ægisdóttir et al., 2006, Andrews et al., 2006, Bonta and Andrews, 2007]. One of the limitations of the second generation was that its predictions were unable to handle individual patterns well because each individual was associated with the findings obtained based on a group of people and their behaviour [Hart et al., 2007, Bickley and Beech, 2001]. These measures could not address the reasons for the behaviours they tried to predict, they indicated associations rather than explaining the causation [Quayle and Taylor, 2004]. Another drawback of the second generation was its inability to capture dynamic changes in individuals' behaviors and needs over time.

In the late 1970s and early 1980s, a third generation of risk assessments came out [Bonta and Andrews, 2007]. It was an extended version of actuarial risk assessment that incorporated both static and dynamic factors [Bonta and Wormith, 2007, McDermott et al., 2008, Clarke et al., 2017]. Since dynamic factors are changeable and may be related to risk, their incorporation into risk assessments helped practitioners target and monitor risk reduction efforts such as rehabilitation programs [Beggs and Grace, 2010, Cording et al., 2016, Bonta, 2002].

While third generation instruments helped practitioners allocate supervision and intervention resources, fourth generation includes structured risk assessment instruments that integrate systematic intervention and structured monitoring of individuals over time to maximize treatment and supervision benefits [Bonta and Andrews, 2007, Andrews et al., 2006, An-

drews et al., 2000]. Fourth generation instruments focus on responsivity considerations that can help practitioners efficiently integrate case planning and risk management efforts [Hart and Boer, 2011].

The adoption of structured risk assessment tools constituted major progress during the past 40 years. In comparison to the past generations such as traditional clinical judgments and unstructured risk assessment instruments, these structured tools show higher accuracy and better performance, although they are still far from perfect [Grove et al., 2000, Hanson, 2005]. In addition, there has been an increase in the accuracy and accordingly acceptability of these tools due to developments in statistics and computer science, large databases availability, and inexpensive computing power [Berk, 2012]. Also, these improvements have expanded the applicability of the tools based on Machine Learning (ML) in different areas [Raz and Michael, 2001, Alberts and Dorofee, 2003, Allen et al., 2006, Anenberg et al., 2016, Berk and Hyatt, 2015, Berk et al., 2016, Berk, 2017]. ML-based methods can accurately discover patterns in historical data and efficiently find associations between input variables and the predicted output [Langley and Simon, 1995].

In this thesis, we focus on two applications of risk assessment tools in criminal justice and education domains which are explained in the following sections.

1.1.1 Criminal Justice

The criminal justice system has applied a range of risk assessment tools to identify the risk level of harm, sexual, criminal, and violent offending, as well consider treatment and rehabilitation programs for offenders since the 1920s [Kehl and Kessler, 2017]. As community safety has been one of the fundamental goals of intervention with offenders, the need for accurate risk assessments in this domain has intensified in recent decades. These tools have been used by police, officers, and psychologists in different decision making areas such as pre-trial risk assessment, sentencing, probation, and parole [Kehl and Kessler, 2017, Lowenkamp, 2009, Monahan and Skeem, 2016, Wright et al., 1984, Funk, 1999, Meredith et al., 2007].

Several semi-structured risk assessment tools have been created and used in different countries to estimate potential criminal risks. Among the most widely used risk assessment tools in the U.S., we introduce some of them as follows:

- **COMPAS:** Correctional Offender Management Profile for Alternative Sanctions is used to assess the risks of general and violent recidivism, and failure to appear in court (FTA) [Brennan et al., 2009].
- **ORAS:** Ohio Risk Assessment System is applied to estimate recidivism risk [Latessa et al., 2009].
- **PCRA:** Post Conviction Risk Assessment is a tool for estimating the post-conviction reoffense under supervision [Johnson et al., 2011].
- **SAVRY:** Structured Assessment of Violence Risk in Youth is used for assessing violence risk in adolescents, between the approximate ages of 12 and 18 [Borum et al., 2020].

Some of the risk assessment tools created and used in Canada are listed below:

- **LSI-R:** Level of Service Inventory-Revised is a tool for estimating recidivism risk [Andrews et al., 2000].
- **SAQ:** Self-Appraisal Questionnaire is used for assessing the recidivism risk [Loza, 2018].
- **SARA:** Spousal Assault Risk Assessment is applied in estimating domestic violence [Kropp and Hart, 2000].
- **SVR-20:** Sexual Violence Risk-20 is a tool to assess sexual violence [Hart and Boer, 2011].
- **PCL-R:** Psychopathy Checklist-Revised is used to estimate the risk of violent recidivism [Hare, 2003].

- **VRAG:** Violence Risk Appraisal Guide is applied to determine the probability of recidivism by mentally ill offenders [Harris et al., 1993].

- **HCR-20:** Historical, Clinical, and Risk Management is a tool to assess the risk of violence [Douglas and Webster, 1999].

Several risk assessment tools have been developed in European countries, they include:

- **OASys VP (OVP):** Offender Assessment System Violence Predictor is used in England and Wales to estimate violent recidivism risk [Howard and Dixon, 2012].

- **SVG[1]-10:** A screening instrument which is developed in Austria for predicting the risk of violent recidivism [Rettenberger et al., 2010b].

- **CBR:** Crime Scene Behavior Risk measure is used in Germany to estimate the risk of sexual recidivism [Dahle et al., 2014].

- **RisCanvi:** A multi-scale risk assessment tool that is developed in Spain to estimate the recidivism risk [Andrés-Pueyo et al., 2018].

1.1.2 Education

Among the challenges and risks that threaten the educational communities, students dropout and underperformance are significant problems which can have a negative impact on students, their families and society [Pascarella and Terenzini, 2005, Tinto, 2017]. These problems are more serious in higher education [Bukralia et al., 2015]. Several predictive analysis techniques have been used to estimate these academic risks using different kinds of student-related data [Liz-Domínguez et al., 2019]. Such techniques are involved in the definition of learning analytics (LA), which is the measurement, collection, and analysis of data about learners and their

[1] Screeninginstrument zur Vorhersage des Gewaltrisikos

environments for the purpose of understanding and improving learning outcomes [Ferguson, 2012].

The predictive analyses are the basis of tools such as Early Warning Systems (EWS) that can help in early identification of at-risk students. EWS are used to predict future risks, such as the likelihood of students failing or dropping out, and alert of such risks so that corrective measures can be taken [Liz-Domínguez et al., 2019]. Machine Learning (ML) algorithms have been accurate and effective methods at this predictive task [Plagge, 2013, Kemper et al., 2020, Aulck et al., 2016, Nagy and Molontay, 2018, Del Bonifro et al., 2020, Albreiki et al., 2021]. Some of EWS that have been applied in the education domain are listed below:

- **CS:** Course Signals system is used to predict students' performance in their courses. [Arnold and Pistilli, 2012].
- **DC:** Degree Compass is a course recommendation system that suggests the best patterns of courses that a higher education student should take to maximize his/her probability of success [Denley, 2013].
- **SE:** Student Explorer system is used for the purposes of identifying students in need of academic support [Krumm et al., 2014].
- **LADA:** Learning Analytics Dashboard for Advisers is a tool to support the decision-making process of academic advisers through comparative and predictive analysis [Gutiérrez et al., 2020].

In addition, assuming underperformance and dropout as a continuous process of student disengagement with the course, teachers, and institution, different screening instruments have been used in higher education for early identification of students at risk of dropout or failure in their studies [Casanova et al., 2021, Goad et al., 2021, Dyrbye et al., 2011, Ganschow and Sparks, 1991].

On Table 1.1, the median predictive performance of some of risk assessment tools (RAT) and ML algorithms used in the two applications of criminal justice [Haarsma et al., 2020, Grann et al., 1999, Karimi-Haghighi

and Castillo, 2021b] and higher education [Aulck et al., 2016, Nagy and Molontay, 2018, Huang et al., 2020] are shown in terms of AUC-ROC.

Table 1.1: AUC-ROC in risk assessment tools. See text for references.

Application	RAT	AUC	Predicted target
Criminal Justice	COMPAS	0.67	General & violent recidivism, pretrial misconduct
	ORAS	0.66	General recidivism
	LSI-R	0.64	General recidivism
	SARA	0.70	Domestic violence
	SAVRY	0.71	Violent risk in youth
	SVR-20	0.78	Sexual violence
	PCL-R	0.72	Violent recidivism
	RisCanvi	0.72 0.70	Violent recidivism General recidivism
Higher Education	LR[1]	0.73 0.76 0.49-0.75	Dropout Early dropout Student performance (low/high)
	RF[2]	0.69 0.74 0.51-0.68	Dropout Early dropout Student performance (low/high)
	K-NN[3]	0.66 0.76	Dropout Early dropout
	DT[4]	0.62 0.50-0.69	Early dropout Student performance (low/high)
	DL[5]	0.81	Early dropout

The tools show similar AUC ranges for general recidivism (between 0.64 and 0.70), violent recidivism (between 0.67 and 0.72) and domestic violence (with AUC of 0.70) prediction. In sexual violence, we can observe high predictive performance of these tools (with AUC value of 0.78).

The predictive performance of ML-based models in ”Early Dropout”, which is dropout risk detection at enrollment time, show AUC values between 0.62 and 0.81. The AUC values for ”Dropout” prediction, that is detecting dropout risk during first academic year, are between 0.66 and 0.73. Student performance level is predicted with AUC values between 0.49 to 0.75.

In general, we can see that even if they belong to different domains, these instruments are similar in the sense that they address a difficult predictive task. This is evident from the fact that they exhibit a wide range of predictive accuracy, and that their accuracy is sometimes arguably acceptable, but rarely high.

1.2 Motivation

In this thesis, we focus on two high-stakes risks in the criminal justice and higher education; recidivism risk of incarcerated persons after release from prison and early dropout risk of undergraduate students in higher education. In the two following sections, we introduce the two risks, their social and economic costs, and assessment of these risk using structured risk assessment tools. There are also other significant risks that can be managed in a similar manner and have many challenges in common with the ones we analyze such as sexual violence risk, risk of a mental disorder, sepsis risk in patients, and risk of fraud in a transaction.

[1]Logistic Regression
[2]Random Forests
[3]K-Nearest Neighbors
[4]Decision Tree
[5]Deep Learning

1.2.1 Recidivism Risk

There are various definitions for the recidivism label among countries. In some countries, reimprisonment counts as recidivism. Others apply this label earlier in the process, such as when a person is reconvicted or even just re-arrested. There are also another difference among countries in whether or not they count certain low-level offenses such as misdemeanors, fines, or traffic violations. Additionally, follow-up times, which is the period after release from incarceration, are often inconsistent between and even within jurisdictions. If one country measures recidivism using a one year follow-up period, another uses two years, and a third uses five years, the data cannot be accurately compared.

According to a study on recidivism rates in 11 countries within a 2-year follow-up period, reported rearrest rates were between 26% and 60%, reconviction rates ranged from 20% to 63%, and reimprisonment rates varied from 14% to 45% [Yukhnenko et al., 2019a]. In Catalonia, the reimprisonment rate within 5.5 years is reported to be 30% for people released from prison in 2010 [Capdevila et al., 2015]. The recidivism rates show that nearly a third of the incarcerated people return to prison. This issue prompted two types of studies with the aim of reducing recidivism rate:

- With the emergence of structured risk assessment tools, efforts have been made in different countries to develop these tools in criminal justice system for accurately assessing the risk of recidivism [Desmarais and Singh, 2013, Howard and Dixon, 2012, Kröner et al., 2007, Rettenberger et al., 2010b, Dahle et al., 2014]. Applying Machine Learning (ML) algorithms in these tools has increased their efficiency and accuracy compared to unstructured professional judgement used in the 1980s [Hanson, 2005, Berk and Hyatt, 2015, Berk et al., 2016, Berk, 2017]. However, they may cause unfairness for some sensitive groups and hence need improvements [Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Tolan et al., 2019].
- In Europe, the prison population reported in January 2021 is 0.1% of the total European population [Marcelo F. Aebi, 2022]. Although

this rate is not as high as the incarceration rate in the United States (which is 1% [Travis et al., 2014, Loeffler and Nagin, 2022]), the rise of "mass incarceration" in all countries during half a century has caused an increasing attention to assessing the effects on crime rates as well as their social and economic costs [Raphael and Stoll, 2009, Durlauf and Nagin, 2011, Spelman, 2020, Loeffler and Nagin, 2022]. In this regard, several studies have focused on the effect of incarceration and its alternatives (programs providing an alternative to prison) on recidivism [Loeffler and Nagin, 2022, Vass, 1990, Dynia and Sung, 2000, Cid, 2009].

1.2.2 Education Dropout Risk

Looking at dropout rates in higher education we observe that more than half of the Brazilian students (52%), more than one-third of the European Union students (36%) and US students (39%) discontinue their studies before graduation [Vossensteyn et al., 2015, Shapiro et al., 2017, OECD, 2012]. Such high dropout rates are alarming, as lead to professional, social and financial losses impacting students, institutions, and society [Bukralia et al., 2015, Pascarella and Terenzini, 2005, Tinto, 2017]. Especially, dropout rate among first-year university students is the highest and academic failure in this year is a strong predictor of dropout [Tinto, 2010, Herbaut, 2021]. Therefore, early recognition of vulnerable students who are at risk of failing the courses or dropping out is essential to prevent them from quitting their studies [Márquez-Vera et al., 2016].

As ML models are often used in the detection of dropout risk and these models may cause disparities, algorithmic fairness considerations should be taken into account in the performance of these models [Gardner et al., 2019, Hutt et al., 2019, Kizilcec and Lee, 2020, Karimi-Haghighi et al., 2021].

Also, when developing ML models, different descriptive statistics and base rates observed in demographic or social sub-groups of students need to be considered. For example, in the US, ethnic minority university students have lower graduation rates compared to White students [Shapiro

et al., 2017]. In the UK, elder students at point of entry are more likely to drop out after the first year compared to younger students who enter university directly from high school [Larrabee Sønderlund et al., 2019]. In addition, the interaction between social origin and academic performance of the students should be taken into consideration. In this regard, there is evidence that students from advantaged backgrounds are much less likely to drop out after academic failure than disadvantaged students [Herbaut, 2021]. Other factors such as having a scholarship or being employed can have an influence on student's performance and dropout [Modena et al., 2020, Choi, 2018, Olaya et al., 2020, Masserini and Bini, 2021].

1.3 Thesis Contribution

I this thesis, we try to assess risks in efficient methods in terms of accuracy and algorithmic fairness, as well as, study the effects of some treatments on the risks using causal inference methods. Our contribution is shown on a diagram in Figure 1.1. In this diagram, we divide our contribution in several parts which are explained as follows:

Application: We apply our goals on assessing significant risks of general and violent recidivism in criminal justice and dropout and underperformance in higher education field. These are among the risks that are of great concern in these two application areas. We believe our contribution can be adapted and applied to other risks in different application fields.

Features and output risk (ground truth): In the two applications, protected groups, treatment variable, and ground truth are not explicitly provided in a tabular form. To determine the output risks, we need to follow sophisticated and complex rules. These rules are not formally encoded, but uncovered with domain experts. Not all relevant groups are explicitly given, but are constructed from the data. Sensitive or protected groups are found based on the application, risk base rates and distribution of important features in these groups. The treatment variable is obtained from the features based on the differences observed in the risk base rate

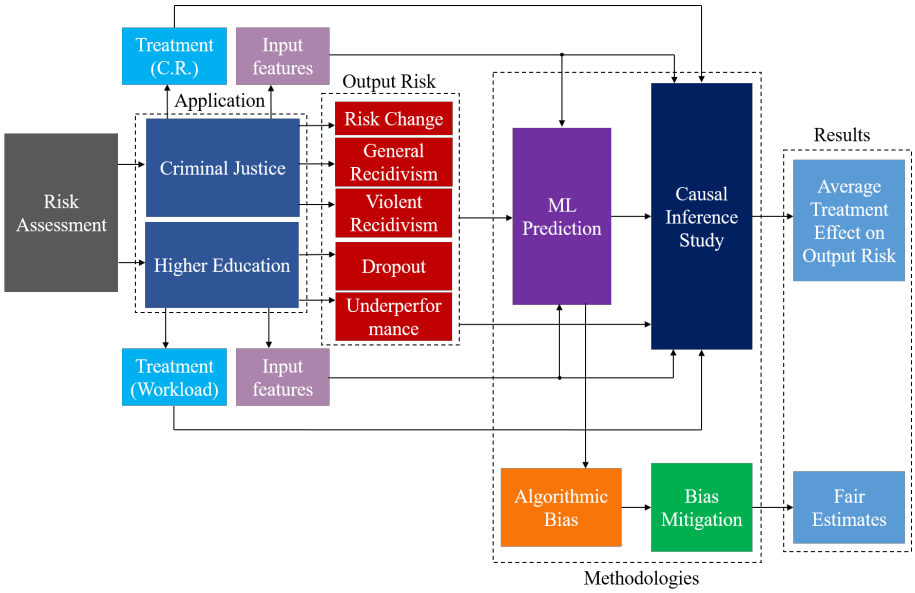


Figure 1.1: Thesis contribution diagram

and descriptive statistics of other features on this variable.

Methodologies: We apply two sets of methodologies, in each we need the risk prediction model that is obtained by applying the related input features and output risk to a Machine Learning (ML) method. In one methodology set, we try to reach algorithmic fairness in the estimates obtained from the risk prediction model. For this purpose, we check for the algorithmic bias in the predictions of sensitive groups and then try to mitigate the obtained bias. In the other methodology set, we aim to obtain the effect of a treatment on the output risk using causal inference methods. The treatment variables are conditional release (C.R.) in criminal justice and number of first year credits (workload) in higher education.

Results: Two sets of results are obtained with respect to each of the applied methodologies. Fair estimates of the risk predictions towards protected groups and the average treatment effect on the output risk.

1.3.1 Thesis Structure

The structure of the thesis is summarized below and also highlighted in Figure 1.2. The thesis consists of four parts and each part contains chapters.

In Part I, we present the introduction (Chapter 1), and background and related work (Chapter 2).

Part II includes our following studies on algorithmic fairness:

- **Chapter 3: Efficiency and Fairness in Recurring Data-Driven Risk Assessments of Violent Recidivism** (**Published in the Proceedings of ACM SAC 2021 [Karimi-Haghighi and Castillo, 2021a]**)

  In this chapter, we study a scenario in criminal justice domain. In this scenario, the inter-evaluation period of a state-of-the-art risk assessment instrument, RisCanvi, depends on the characteristics of each inmate. In the scenario, only a fraction of the inmates, those with the highest probability of having changed risk, are selected for the next evaluation. Our work is based on a cost-benefit analysis which leads to fewer evaluations in exchange for some missed/undetected risk changes. Importantly, we analyze if this method leads to discriminatory outcomes across some characteristics, including disparate impact in the evaluation rates along nationality and age. By adjusting decision boundaries we are able to mitigate the disparate impact and ensure equality in the rate of evaluation.

- **Chapter 4: Enhancing a Recidivism Prediction Tool With Machine Learning:Effectiveness and Algorithmic Fairness** (**Published in the Proceedings of ICAIL 2021 [Karimi-Haghighi and Castillo, 2021b]**)

  We address a key application of Machine Learning (ML) in the legal domain in this chapter. We investigate how ML may be used to increase the effectiveness of RisCanvi risk assessment tool, without introducing undue biases. The two key dimensions of this analysis are predictive accuracy and algorithmic fairness. It is described how

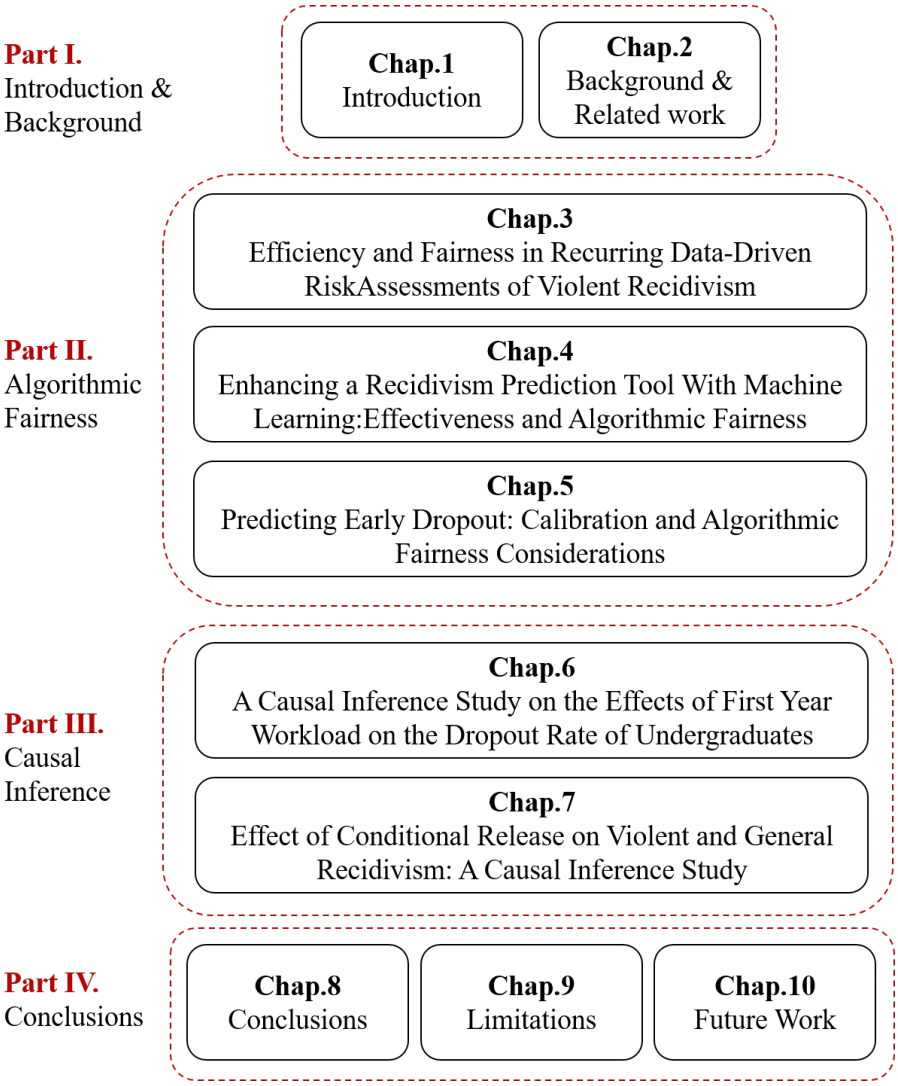


Figure 1.2: Thesis structure

effectiveness and algorithmic fairness objectives can be balanced, applying a method in which a single error disparity in terms of generalized false positive rate is minimized, while calibration is

maintained across groups.

- **Chapter 5: Predicting Early Dropout: Calibration and Algorithmic Fairness Considerations** (**Published in the Proceedings of LAK 2021 [Karimi-Haghighi et al., 2021]**)

  In this chapter, the problem of predicting dropout risk in undergraduate studies is addressed from a perspective of algorithmic fairness. We develop a machine learning method to predict the risks of university dropout and underperformance. The objective is to understand if such a system can identify students at risk while avoiding potential discriminatory biases. We analyze the discriminatory outcomes for some sensitive groups in terms of prediction accuracy (AUC) and error rates (Generalized False Positive Rate, GFPR, or Generalized False Negative Rate, GFNR). Then, we address the disparities through a mitigation process that does not affect the calibration of the ML model.

Part III includes our studies on causal inference as follows:

- **Chapter 6: A Causal Inference Study on the Effects of First Year Workload on the Dropout Rate of Undergraduates** (**Published and nominated for the best paper award in the Proceedings of AIED 2022 [Karimi-Haghighi et al., 2022]**)

  In this chapter, we evaluate the risk of early dropout in undergraduate studies using causal inference methods, and focusing on groups of students who have a relatively higher dropout risk. Among important drivers of dropout over which the first-year students have some control, we find that first year workload (i.e., the number of credits taken) is a key one, and we mainly focus on it. We determine the effect of taking a relatively lighter workload in the first year on dropout risk using causal inference methods: Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), Augmented Inverse Propensity Weighted (AIPW), and Doubly Robust Orthogonal Random Forest (DROrthoForest).

- **Chapter 7: Effect of Conditional Release on Violent and General Recidivism: A Causal Inference Study** (**Submitted to Journal of Quantitative Criminology**)

  We study the effect of conditional release (C.R.), which is similar to ”parole” in the US, on general and violent recidivism in different prison centers in Catalonia . We study men and women separately because our observations show substantial differences in their profiles. We apply several causal inference methods: Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse Propensity Weighting (AIPW) and determine the Average Treatment Effect (ATE) of C.R. on recidivism within 2-5 years of release. We are also interested to compare this effect in different recidivism risk levels estimated by RisCanvi risk assessment tool.

In Part IV, we present the thesis’ conclusions (Chapter 8), limitations (Chapter 9), and future work (Chapter 10).

Chapter 2

BACKGROUND AND RELATED WORK

2.1 Challenges of Risk Assessment

Risk assessment tools (RATs) are widely used in several decision making processes. Some of application areas of RATs are shown on Table 2.1 with their predicted risk outputs.

Although structured RATs have been appropriate alternatives to traditional prediction methods especially in terms of their accuracy [Grove and Meehl, 1996, Grove et al., 2000, Kirton and Kravitz, 2011], there are still some challenges with regards to the performance of these tools which are described in the following sections.

2.1.1 Predictive Accuracy

The moderate or limited predictive performance of some RATs prevent using them in some decision making that requires a very high level of accuracy, for example in criminal justice domain using some RATs (such as PCL-R, PCL:SV, HCR-20, VRAG, OGRS, RM2000V, LSI/LSI-R, GSIR,

[1]Failure To Appear

Table 2.1: Risk assessment tools in different application areas

Application	Predicted risk	Reference
Health system	Behavioral & psycho-social risks	[Phillips et al., 2014]
	Atherosclerotic cardiovascular disease	[Lloyd-Jones et al., 2019]
Ecosystem protection	Wildfire risk	[Calkin et al., 2011]
	Air pollution health risk	[Anenberg et al., 2016]
Information security	Risk of information disclosure or disruption	[Alberts and Dorofee, 2003]
Auditing	Auditor risk	[Allen et al., 2006]
Project management	Project failure risk	[Raz and Michael, 2001] [Zeng et al., 2007]
Education retention	Dropout risk	[Cohen, 2017] [Casanova et al., 2021]
Criminal justice	Parolee risk	[Meredith et al., 2007]
	Risk of FTA [1] in court (flight risk)	[Lowenkamp, 2009]
	Risk of recidivism	[Latessa et al., 2010] [Monahan and Skeem, 2015]
	Domestic violence risk	[Hilton et al., 2010]

and VRS) in preventive detention decisions [Yang et al., 2010] or in clinical settings applying RATs in predicting patient falls [Myers and Nikoletti, 2003]. Some RATs have weak performances, for instance one named DASH (Domestic Abuse, Stalking and Honour Based Violence), that helps

police officers in predicting revictimization[1], shows predictions which are little better than random [Turner et al., 2019]. Also, evaluating the performance of some risk scales (Manchester Self-Harm Rule, ReACT Self-Harm Rule, SAD PERSONS scale, Modified SAD PERSONS scale, Barratt Impulsiveness Scale), which identify patients who repeat self-harm within 6 months of presentation, shows that most scales performed significantly no better than clinician and patient estimates of risk and even some performed considerably worse [Quinlivan et al., 2017]. Therefore, using risk scales to determine patient management or predict self-harm is not recommended by these authors.

Furthermore, there is some inconsistency in the predictive power of some RATs such as Sexual Violence Risk-20 (SVR-20) in which the predictive performance depends on the recidivism criterion and offender subgroup [Rettenberger et al., 2011]. Such limitations of variable predictive accuracy for different sexual offender subgroups and reoffence categories have also been indicated in some other RATs for sexual offenders (Static-99, Rapid Risk Assessment for Sexual Offense Recidivism, Sex Offender Risk Appraisal Guide, and Sexual Violence Risk) [Rettenberger et al., 2010a]. Also, a study on the performance of recidivism RATs used in US correctional settings shows that the variability of predictive validity of such RATs due to offender characteristics, settings, and recidivism may impact the feasibility of implementing these tools [Desmarais et al., 2016].

2.1.2 Algorithmic Bias

There are substantial variations in the predictive accuracy and heterogeneities in the validity of some commonly used violence RATs (HCR-20, LSI-R, PCL-R, SARA, SAVRY, SORAG, Static-99, SVR-20, and VRAG) with respect to different demographics [Singh et al., 2011]. Their predictive validity is higher when the demographic characteristics of the tested sample are closer to the original validation sample of the tool. Also, the

[1]If the primary victim of the index incident was a primary or secondary victim at the subsequent incident (regardless of whether the incident was flagged as domestic abuse in the police systems), this is defined as revictimization [Turner et al., 2019].

tools designed for more specific populations show more accuracy at detecting the risk than tools intended to be used for broader populations. Such a disparate impact of predictions with regards to a protected class (ethnicity, gender) of demographics has been detected in widely-used RATs such as Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). This tool has been found to have biases across race based on a study by ProPublica [Angwin et al., 2016, Larson et al., 2016] and gender [Hamilton, 2019]. These studies showed that the COMPAS is biased against African American defendants and over-predict risk for women. However, the COMPAS developer (Northpointe) rejected the findings of the ProPublica by claiming that their algorithm is fair because it is well calibrated [Dieterich et al., 2016]. Another study focused on SAVRY [Tolan et al., 2019, Miron et al., 2020], shows that although ML models could be more accurate than the simple summation used to compute SAVRY scores, they would introduce discrimination against some groups of defendants compared to the current method.

In addition, there is empirical evidence that the fairness of some algorithmic pretrial RATs such as Public Safety Assessment (PSA), Virginia Pretrial Risk Assessment Instrument (VPRAI), and Federal Pretrial Risk Assessment (PTRA) may be affected by the classification indicators (such as accuracy and error rates indicators) and their thresholds [Zottola et al., 2022]. With regards to error rates indicators, people of color were more often misclassified as false positive and had higher false positive rates than white people, particularly at the low threshold. However, using the high threshold, the differences in the error rates of people of color and white people were much less pronounced, and accordingly provided less evidence regarding bias.

There are also biases with respect to socioeconomic factors such as education, employment, income and housing. In this regard, it has been critically shown that these factors marginality contribute to a higher risk score in some widely used risk assessment tools such as the COMPAS (US), the Level of Service Inventory-Revised (LSI-R) (Canada, US), the Offender Assessment System (OASys) (UK) and the Recidive InschattingsSchalen (RISc) (the Netherlands). These RATs result in a higher

likelihood of a (longer) custodial sentence for underprivileged offenders compared to their more privileged counterparts, which in turn leads to produce sentencing disparities as well as to reproduce social inequalities [Van Eijk, 2017].

In the education field, disparities may be found in algorithmic predictions with respect to race/ethnicity, gender, and nationality groups, as well as sociodemographic, disability, and military-connected status [Baker and Hawn, 2021, Coleman, 2019]. In the performance of models predicting the risk of university course failing, worse results are obtained for African-American compared to non-African-American students, as well as for male than female students, but the results are inconsistent across university courses [Hu and Rangwala, 2020]. In a study on predicting college graduation, algorithms show higher false positive rates for White students than others, higher false negative rates for Latino students compared to others, and higher false negative rates for male than non-male students [Anderson et al., 2019]. In predictive models of college dropout, research shows lower true negative rates and better recall for students who are not White or Asian compared to others, as well as lower true negative rates for male compared to female students [Yu et al., 2021]. In a study on MOOC dropout prediction, it is found that several algorithms performed worse for female students compared to male students [Gardner et al., 2019].

ML can be used in some cases to mitigate biases, however, there are limitations to this approach due to the human bias in the data and ML algorithms trained with such data will reproduce, not eliminate, the bias [Lum, 2017].

2.1.3 Effectiveness of Risk Assessment

A basic consideration in using risk assessment tools is their effectiveness in the desired area. According to a study on mental health care, there is no evidence that risk assessment is effective in relation to self-harm or suicide reduction [Wand, 2011]. Similarly, it has been shown that the majority of mental disorder patients categorized as being at high risk will not commit any harmful acts [Ryan et al., 2010]. Such clinical decisions made on the

basis of risk assessment also divert resources away from patients classified as low risk, even though a significant proportion do go on to commit a harmful act. In the criminal justice domain, there is a tremendous pressure to focus resources on defendants who are assessed as low-risk, for instance alternative programs to prison such as parole or conditional release are assigned to lower risk offenders [Bonta and Andrews, 2007, Andrés-Pueyo et al., 2018].

However, this question arises that how can we promote the reduction of risk and not merely its assessment? [Monahan and Skeem, 2015]. Before making decisions inline with the risk estimated by the risk assessment tools, we need to firstly check whether the risk assessment is appropriate for the desired application. If so, it is important to investigate whether resource allocation to the treatment of a particular group is the best choice for both selected and non-selected groups and the society.

Notably, based on the risk-need-responsivity (RNR) principle for offender risk assessment, rehabilitation efforts will be effective when risk instruments are evidence-based and the level of rehabilitation services matches the level of risk, type of criminogenic need, and learning style and motivations (responsivity) of the individual being treated. In other words, people at high-risk to reoffend with many potentially changeable criminogenic needs should be targeted for the strongest rehabilitation efforts. By contrast, minimal efforts should be reserved for those at low-risk and those with few criminogenic needs [Bonta and Andrews, 2007]. So, inappropriate matching of treatment intensity with offender risk level can waste treatment resources and in some situations actually make matters worse.

2.2 Addressing the Challenges

With regards to the challenges of risk assessment introduced in Section 2.1, there are several studies which try to provide solutions to such problems. In this part, we describe some of them focusing mostly on risk assessment applications in criminal justice and education which are the related work

to this thesis.

2.2.1 Increasing Predictive Accuracy

There are some studies suggesting frameworks to enhance risk assessment tools (RATs) in several applications such as business continuity management systems [Torabi et al., 2016], healthcare [Dueñas-Espín et al., 2016, Brigell et al., 2020], ecological risk assessment process [Dale et al., 2008], and recidivism risk assessment [Helmus et al., 2012, Knight and Thornton, 2007, Labrecque et al., 2014].

In criminal justice, recidivism estimates of RATs are generally developed on samples of offenders with average age below 50 years and since criminal behavior of all types declines with age [Hirschi and Gottfredson, 1983, Sampson and Laub, 2017], as a result, actuarial scales tend to overestimate recidivism for older offenders. Hence, in a study, a revised scoring system for two risk assessment tools (Static-99 and Static-2002) is developed using new age weights, which describes recidivism risk of older offenders more accurately [Helmus et al., 2012]. In a study on sexual offenders, some procedures are suggested for improving the decision-making algorithms used in the risk assessment of sexual recidivism [Knight and Thornton, 2007]. The study of the dynamic validity is essential, according to a survey, focused on the LSI-R RAT, which shows that reassessing offenders for recidivism risk is important [Labrecque et al., 2014]. By relying on the fact that changes in the risk score relate to recidivism, it is indicated that the use of risk score change can add incremental validity to the utility of this RAT.

Machine Learning (ML) algorithms have been developed for predicting dropout risk in higher education. The potential of ML is highlighted in student retention and success and it is shown that dropout can be accurately predicted even when ML predictions are based on a single term of academic transcript data [Aulck et al., 2016]. Using students' demographics and academic transcripts, the strongest predictions of students dropout are obtained by regularized logistic regression with AUC value of 0.73. In another study, early university dropout is predicted based on only available

data at the time of enrollment using several ML models [Nagy and Molontay, 2018]. Their best results are obtained from Gradient Boosted Trees and Deep Learning algorithms showing AUC values of 0.808 and 0.811 respectively. Similarly, several ML methods have been used to predict the course dropout among first-year undergraduate students before the student starts the course or during the first year [Del Bonifro et al., 2020]. Another reference develops a model to predict real-time dropout risk for each student during an online course using a combination of variables from the Student Information Systems and Course Management System [Bukralia et al., 2015]. Evaluating the predictive accuracy and performance of various data mining techniques, the study results show that the boosted C5.0 decision tree model achieves 90.97% overall predictive accuracy in predicting student dropout in online courses. Student academic performance is estimated in two levels of low or high in a research using big data (seven datasets of learning logs within three universities) and different ML algorithms [Huang et al., 2020]. Their results show the best AUC value of 0.75 using logistic regression algorithm in predicting student performance.

2.2.2 Improving Algorithmic Fairness

According to several studies, Machine Learning (ML) algorithms may cause discriminatory outcomes with respect to some sensitive groups [Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Barocas et al., 2017, Mehrabi et al., 2021, Zou and Schiebinger, 2018]. Establishing algorithmic fairness has been performed according to many different definitions, some of them incompatible with one another [Narayanan, 21, Narayanan, 2018]. Satisfying all of the fairness definitions simultaneously and even maximizing fairness and accuracy at the same time is impossible and there are necessary trade-offs between different metrics [Berk, 2019, Berk et al., 2018, Chouldechova, 2017, Kleinberg et al., 2016]. On Table 2.2, some of fairness definitions considered in previous work are summarized. The definitions can be based on predicted outcome, predicted and actual outcomes, or predicted probabilities and actual outcome.

Table 2.2: Literature review of some fairness definitions

Definition base	Fairness	Unfairness	Reference
Predicted & actual outcomes	Predictive equality	Disparate error rate	[Verma and Rubin, 2018] [Corbett-Davies et al., 2017]
	Equalized odds	Disparate mistreatment	[Hardt et al., 2016] [Zafar et al., 2017] [Woodworth et al., 2017]
Predicted outcome	Statistical parity (group fairness)	Disparate impact	[Zemel et al., 2013] [Kamiran and Calders, 2009] [Kamishima et al., 2011] [Dwork et al., 2012] [Johndrow et al., 2019] [Feldman et al., 2015]
Predicted probabilities & actual outcome	Well-calibration	Uncalibrated scores	[Chouldechova, 2017] [Kleinberg et al., 2016] [Berk et al., 2021]

One of the algorithmic fairness definitions is predictive equality which is based on the ”Separation” criterion and means that the classifier shows equal false positive rate (FPR) for both protected and unprotected groups [Verma and Rubin, 2018, Corbett-Davies et al., 2017]. In some studies, potential algorithmic discrimination is mitigated by satisfying equalized odds which means avoiding disparate mistreatment along different sensitive groups [Hardt et al., 2016, Zafar et al., 2017, Woodworth et al., 2017]. In other words, a classifier satisfies equalized odds if protected and unprotected groups have equal false positive rate and equal true positive rate, which is another variation of the ”Separation” criterion.

Other studies try to overcome disparate impact (the difference in probabilities of a positive outcome across two groups) and approach statis-

tical parity (group fairness) in which the same probability of receiving a positive-class prediction is obtained for the cases in protected and unprotected groups. In this regard, a classification scheme for learning unbiased models on modified biased training data is introduced [Kamiran and Calders, 2009]. Also a regularization approach which is applicable to any prediction algorithm with probabilistic discriminative models is suggested [Kamishima et al., 2011]. By formulating fairness as an optimization problem, a learning algorithm is introduced which can lead to both group fairness and individual fairness [Zemel et al., 2013]. Another method is to remove all information regarding protected variables from the data to which the models will be trained [Johndrow et al., 2019].

Another important fairness definition is calibration [Berk et al., 2021]. Calibration means that the output of the classifier is not merely a score, but an estimate of the probability of the positive class. A score is said to be well-calibrated if it reflects the same likelihood of being classified in the positive class irrespective of the individuals' group membership [Chouldechova, 2017]. In this regard, some work has tried to minimize error disparity across groups while maintaining calibrated probability estimates [Pleiss et al., 2017].

There are various studies on algorithmic fairness of Risk Assessment Tools (RATs) applied in the criminal justice domain. In studying algorithmic fairness of the US widely-used program of Correctional Offender Management Profiles for Alternative Sanctions (COMPAS), the ProPublica findings on disparate impact of this algorithm on black defendants [Angwin et al., 2016, Larson et al., 2016] were refuted by COMPAS developer (Northpointe), proving that their algorithm is fair because it is sufficiently calibrated and satisfies predictive parity [Dieterich et al., 2016]. In contrast to the COMPAS, studies on other risk assessment tools such as the Post Conviction Risk Assessment (PCRA), the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI) show no significant racial bias in the recidivism prediction [Skeem and Lowenkamp, 2016, Perrault et al., 2017].

There are comparatively less works looking at algorithmic fairness of university dropout risk assessment. A study [Gardner et al., 2019]

considers algorithmic fairness of predictive models of students dropout in MOOCs in terms of accuracy equity using the Absolute Between-ROC Area (ABROCA) metric. The method they apply to improve algorithmic fairness is slicing analysis, which is also used in another study to analyze fairness across sociodemographic groups in a predictive ML modeling of on-time college graduation [Hutt et al., 2019]. In another research, with the aim of satisfying some fairness definitions, post-hoc adjustments are applied to a predictive model of students success in higher education. The fairness improvement is performed by picking different threshold values for each protected and unprotected group to achieve equality of opportunity [Lee and Kizilcec, 2020].

2.2.3 Quantifying Relationships between Treatments and Outcomes

Some features act as a treatment affecting the outcome risk and their effects needs to be measured. In the criminal justice domain, the effects of punishment or treatment on recidivism have been widely studied using different methods, among them we focus on causal inference methods such as Instrumental Variables (IV) [Angrist et al., 1996], Regression Discontinuity (RD) [Thistlethwaite and Campbell, 1960], and other statistical methods [Rosenbaum and Rubin, 1983, Bray et al., 2019, Glynn and Quinn, 2010]. In this regard, we consider two categories of studies on incarceration effects on recidivism and on alternatives to prison effects on recidivism.

There are several studies on the effect of incarceration on recidivism [Loeffler and Nagin, 2022]. The IV method is the most used approach in these studies. It estimates the causal impact of incarceration on recidivism by controlling for an exogenous variation in the assignment of cases [Green and Winik, 2010, Loeffler, 2013, Mueller-Smith, 2015, Gupta et al., 2016, Harding et al., 2017, Bhuller et al., 2020]. RD is another approach that is applied in estimating the effects of incarceration on recidivism [Chen and Shapiro, 2007, Loeffler and Grunwald, 2015, Mitchell et al., 2017]. In this method, program assignment is formed on a score-based system,

when the assignment is discontinuous and deterministic at some threshold value along the score, any sudden changes in the outcome of interest can be causally attributed to the effects of program [Loeffler and Nagin, 2022]. Statistical methods are also used to examine the effect of incarceration on recidivism [Mears and Bales, 2009, Jolliffe and Hedderman, 2015]. These methods are developed methodologies encompassing regression models and inverse probability weighting.

Several studies have explored the effects of programs providing an alternative to prison on recidivism [Vass, 1990, Dynia and Sung, 2000, Cid, 2009]. Some research use IV method to estimate the effect on recidivism of alternative programs such as electronic monitoring and parole [Henneguelle et al., 2016, Andersen and Telle, 2022, Meier et al., 2020]. The effect of programs such as early release and prison length reduction has been obtained in some studies using RD method [Marie, 2009, Rhodes et al., 2018]. There are few studies that obtain the effect of treatment programs on recidivism using statistical methods such as RA, IPW, and AIPW, and IPWRA [Sondhi et al., 2020, Gilman and Walker, 2020].

Results from above mentioned studies on the effect of incarceration and alternative programs on recidivism suggest that custodial sanctions have no effect or even a criminogenic effect on recidivism, except for rehabilitation-focused incarceration. However, non-custodial alternative programs to prison mostly show preventative effects and to a small extent show no effect on recidivism.

In the education domain, there are several research that study the effect of some features on students' performance and dropout risk. In a study, by identifying factors contributing to students' dropout risk and presenting them to academic institutes, they help distinguish more accurately students that may need further support [Chounta et al.,]. Another study performs a correlation analysis on university data collected over 11 years to investigate the potential relationship between some academic features and dropout risk over time [Tanvir and Chounta,]. In a research, using subgroup discovery, effective factors on student success are identified among important combinations of features known before students start their studies [Lemmerich et al., 2010].

Statistical methods including causal inference are also used to examine the effect of some features on students' performance. In a study, using propensity score matching (PSM), it is shown that participation in social media groups created by students has a negative effect on dropout risk (causes a reduction in dropout rate) [Masserini and Bini, 2021]. Another study investigates the effect of grants on university dropout rates by blocking on the propensity score with regression adjustment. Their results show that grants have a positive effect on the probability of completing college education [Modena et al., 2020].

2.2.4 Our Contribution with Respect to Previous Work

With the aim of performing fewer evaluations by RisCanvi risk assessment tool which in turn leads to save time, expenses and staff in the evaluations, we study the time series of evaluations and perform a simulation in which only those inmates with the highest probabilities of violent recidivism risk change are selected for the next assessment (Chapter 3) [Karimi-Haghighi and Castillo, 2021a]. We also evaluate the potential algorithmic bias introduced by this method along nationality and age. This evaluation is investigated along four metrics: accuracy differences, inequality in the missed changes which can be considered analogous to a notion of disparate mistreatment [Zafar et al., 2017], and disparate impact in the rate of evaluations or fraction of unnecessary evaluations. Since simultaneous satisfaction of all fairness measures is impossible, we try to mitigate the disparate impact in the rate of evaluation across groups while keeping the fraction of missed changes small. Thus, we address the disparity through an algorithmic bias mitigation procedure by moving the decision boundary and equalize evaluation rates across nationality and age.

We also seek to obtain a more efficient risk assessment of recidivism in criminal justice and dropout in higher education domains that at the same time has good algorithmic fairness properties. In recidivism risk assessment, we try to enhance the RisCanvi tool, which is an expert-based risk assessment tool, in terms of effectiveness and algorithmic fairness by proposing a new machine learning model to replace it (Chapter 4)

[Karimi-Haghighi and Castillo, 2021b]. To mitigate potential algorithmic discrimination along nationality and age groups, we use a relaxation method which seeks to satisfy equalized odds or parity in a single error rate in terms of generalized false positive rate while preserving calibration in each sub-group of nationality and age. The effects of algorithmic bias mitigation along groups are described on both the RisCanvi tool and the machine learned model. Similarly, in the education domain, we try to minimize bias in a predictive ML model of dropout risk by equalizing error in generalized false positive rate along some sensitive groups while preserving calibration in each group (Chapter 5) [Karimi-Haghighi et al., 2021].

Using statistical causal inference methods, we measure the effects of some important features or treatments on recidivism as a criminal risk and dropout as a risk in higher education system. We measure the effect of the most important features (the number of credits in the first year, age, and study access type) on the early risk of dropout in undergraduate studies (Chapter 6) [Karimi-Haghighi et al., 2022]. This effect is obtained for combinations of these features and the Average Treatment Effect (ATE) is measured using multiple causal inference methods. In criminal justice domain, we compute the ATE of conditional release on recidivism and compare this effect in different recidivism risk levels estimated by RisCanvi risk assessment tool (Chapter 7, submitted to Journal of Quantitative Criminology).

Part II

Algorithmic Fairness

Chapter 3

EFFICIENCY AND FAIRNESS IN RECURRING DATA-DRIVEN RISK ASSESSMENTS OF VIOLENT RECIDIVISM

3.1 Introduction

Risk assessment is a necessary process in many important decisions such as public health, information security, project management, auditing, and criminal justice. Since the 1920s, violence risk assessment tools have been progressively used in criminal justice [Kehl and Kessler, 2017]. These tools are used by probation and parole officers, police, and psychologists to assess the risk of harm, sexual, criminal, and violent offending in more than 44 countries [Singh et al., 2014]. The main purpose of violence risk assessment tools is to prevent criminal violence and its consequences, but they also help prison management identify offenders with a greater risk of recidivism and allocate rehabilitation efforts accordingly. Ideally, accurate risk assessment may help place low-risk defendants in alternative programs

to prison [Andrés-Pueyo et al., 2018].

In comparison to traditional prediction methods and unstructured clinical judgments, risk assessment tools offer superior accuracy and performance [Grove et al., 2000]. In this regard, factors such as the availability of large databases, inexpensive computing power, and developments in statistics and computer science have brought an increase in the accuracy and applicability of these structured tools [Berk, 2012]. Such advances have effectively increased the use of tools based on Machine Learning (ML) in criminal justice decisions for risk forecasting [Berk and Hyatt, 2015, Berk et al., 2016, Berk, 2017]. ML-based systems provide automatic methods that can improve accuracy and efficiency by discovering and exploiting regularities in historical (training) data [Langley and Simon, 1995].

Today, various semi-structured protocols for assessing risk of recidivism can be found in different countries including the U.S. [Desmarais and Singh, 2013], the U.K. [Howard and Dixon, 2012], Canada [Kröner et al., 2007], Austria [Rettenberger et al., 2010b], and Germany [Dahle et al., 2014]. In Spain, among current violence risk assessment tools including SAVRY, PCL-R, HCR-20, SVR-20, and SARA, RisCanvi is a relatively new tool for risk assessment of recidivism. It was originally developed in 2009 in response to concerns of Catalan prison system officials regarding violent recidivism among offenders after their sentences. In the RisCanvi protocol, each inmate is evaluated every six months and each evaluation results in four scores predicting four outcomes: (i) violent recidivism, (ii) self-directed harm/violence, (iii) violence within the prison facilities, and (iv) breaking of prison permits [Andrés-Pueyo et al., 2018].

Our contribution. Performing risk evaluations for all of the inmates every six months is an expensive and time-consuming task. We observe that the risk score for "Violent Recidivism" (hereinafter referred to as REVI) changes differently over time depending on the initial risk. Therefore, we study a scenario in which a decision on the next evaluation for each inmate is taken using an ML-based prediction of the risk change. To this purpose, three ML models are generated for the prediction of change within 6, 12, and 18 months. The ML models are created using risk factors (details

in Section 3.3), the current risk score (REVI) generated by the RisCanvi protocol using those risk factors, and demographic factors.

We perform a simulation in which only those inmates with the highest probabilities of risk change are selected for the next assessment. We show that this can halve the number of evaluations with a relatively small number of missed/undetected risk changes.

We also perform an evaluation of potential algorithmic bias introduced by this method. Given that ML-based models may lead to unfairness [Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Tolan et al., 2019], we compare the impact of our data-driven scheduling of risk assessment along nationality and age. This impact is investigated along four metrics: accuracy differences, inequality in the missed changes which can be considered analogous to a notion of disparate mistreatment [Zafar et al., 2017], and disparate impact in the rate of evaluations or fraction of unnecessary evaluations. We address these disparities through an algorithmic discrimination mitigation procedure, which equalizes evaluation rates across nationality and age. As expected, in exchange of evaluations rate parity, there is an increase in missed changes. As we will show, this increase is small.

We remark that there are many similar domains in which professionals need to perform periodic appraisals, potentially with the assistance of an algorithm, including education, public health, allocation of social benefits, and information security. In all cases where recurring data-driven risk assessment is used to make these kinds of decisions for individuals, the frequency of these assessments is key to achieve a balance of costs and benefits, and it is important to consider and mitigate the potential algorithmic bias that may be inadvertently introduced when seeking to reduce such frequency.

The rest of this paper is organized as follows. In Section 3.2, a brief description of the related work is presented. In Section 3.3, some explanations regarding the RisCanvi risk assessment tool, the data set used in this study, and violent recidivism are provided. The methodology including the model-level evaluation, system-level evaluation, and fairness evaluation are presented in Section 3.4. The results related to each of the evaluation

metrics are given in Section 3.5. To mitigate the discrimination, a procedure is suggested in Section 3.6. Finally, the obtained results are discussed in Section 3.7 and the paper is concluded in Section 3.8.

3.2 Related Work

The introduction of algorithms for risk assessment in criminal justice is a controversial topic, and perhaps one that has motivated a great deal of research on algorithmic fairness.

In the US, a widely-used program named Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) has been found to have biases across races and genders. In seminal research done by investigative journalism organization ProPublica [Angwin et al., 2016, Larson et al., 2016] it was concluded that the COMPAS risk assessment tool is biased against African American defendants. A follow-up study [Hamilton, 2019] analyzed the fairness of COMPAS in terms of predictive parity, and found that COMPAS outcomes systematically over-predict risk for women, thereby indicating systemic gender bias. However, the findings of the ProPublica study were rejected by Northpointe (COMPAS developer), claiming their algorithm is fair because it is well calibrated [Dieterich et al., 2016]. Moreover, in this report it is shown that the COMPAS risk scales exhibit accuracy equity and predictive parity.

In contrast to the case of COMPAS, other studies have shown that other risk assessment tools such as the Post Conviction Risk Assessment (PCRA) do not exhibit racial bias in the recidivism prediction [Skeem and Lowenkamp, 2016]. Similarly, in risk assessment tools used in juvenile probation decisions, such as the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI), no significant racial bias has been found in the prediction of re-offending, except for a higher score in African American youth compared to White youth in the YLS/CMI scale related to official juvenile history [Perrault et al., 2017]. In a more recent study focused on SAVRY [Tolan et al., 2019], it is shown that although ML models could be

more accurate than the simple summation used to compute SAVRY scores, they would introduce discrimination against some groups of defendants compared to the current method.

In general, it is impossible to maximize fairness and accuracy at the same time [Berk, 2019, Berk et al., 2018]. There are many different definitions of algorithmic fairness [Narayanan, 2018], some of which are incompatible with one another. It is impossible to satisfy all of them simultaneously, hence, there are necessary trade-offs between different metrics [Berk et al., 2018, Chouldechova, 2017, Kleinberg et al., 2016]. In this regard, some studies [Hardt et al., 2016, Zafar et al., 2017, Woodworth et al., 2017] try to mitigate potential algorithmic discrimination by satisfying equalized odds or in other words avoiding disparate mistreatment along different sensitive groups. In the methodology used by Zafar et al. disparate treatment can also be avoided simultaneously with disparate mistreatment since sensitive feature information is not used while making decisions, which make it more applicable for the scenarios when the sensitive attribute information is not available. Also, several studies [Zemel et al., 2013, Johndrow et al., 2019, Kamiran and Calders, 2009, Kamishima et al., 2011] tried to approach statistical parity in which the same probability of receiving a positive-class prediction is considered for different groups. In addition, due to the importance of the calibration in risk assessment tools [Dieterich et al., 2016, Berk et al., 2018], some previous work has also tried to minimize error disparity across groups while maintaining calibrated probability estimates [Pleiss et al., 2017].

As explained, our work is based on a cost-benefit analysis which results in fewer evaluations in exchange for some missed (undetected) changes. Thus, to mitigate potential algorithmic bias there is a trade-off between some fairness metrics; mitigating disparate impact in the evaluation rates and disparate mistreatment regarding undetected risk changes along groups. Since simultaneous satisfaction of both measures is impossible we try to mitigate the disparate impact in the rate of evaluation across groups while keeping the fraction of missed changes small.

3.3 RisCanvi Dataset

3.3.1 The RisCanvi Risk Assessment Tool

RisCanvi was introduced as a multi-level risk assessment protocol for violence prevention in the Catalan prison system in 2009 [Andrés-Pueyo et al., 2018]. It was designed jointly by professionals working in the prison system, including lawyers, social workers, criminologists, and psychologists, similarly to other risk assessment tools [Brennan and Dieterich, 2018, Borum, 2006]. RisCanvi is not a questionnaire. Instead, each inmate is interviewed by professionals, who are responsible for analyzing different areas of the inmate's progress through the lens of some risk factors. Each evaluation requires multiple interviews by several professionals spread along several days. RisCanvi interviews are coded by trained professionals and a system generates a risk score; a committee accepts or modifies this score and decides the next action, intervention, or program.

In the RisCanvi protocol, risk is determined for each inmate relative to four possible outcomes: self-directed violence, violence in the prison facilities, committing further violent offenses, and breaking prison permits. To determine the probability of each outcome, a unique predictive algorithm was designed. Each predictive algorithm incorporates various risk factors and three additional variables: age, gender, and country of origin (Spanish or foreign).

Two versions of the RisCanvi protocol were created, an abbreviated one of 10 items for screening (RisCanvi-S), and a complete one of 43 items (RisCanvi-C). Risk items for both versions are shown on Table 3.1.

Risk items are grouped into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, and Attitudes/Personality. These items can also be divided into static factors (such as "criminal history in their family" and "age at first violent offense") and dynamic factors (such as "member of socially vulnerable groups" and "pro-criminal or antisocial attitudes"). In the screening version RisCanvi-S, some risk items are directly taken from RisCanvi-C and others are a combination of items [Andrés-Pueyo et al., 2018].

Table 3.1: RisCanvi Risk Factors, with Items Related to Violent Recidivism Marked in Boldface

RisCanvi Complete items (S = shared with Riscanvi Screening)
(1) **Violent base offense**
(2) Age at the time of the base offense
(3) **Intoxication during performing the base offense**
(4) Victims with injuries
(5) Length of criminal convictions
(6) Time served in prison
(7) **History of violence** (S)
(8) Start of the criminal or violent activity (S)
(9) **Increase in frequency, severity and diversity of crimes**
(10) **Conflict with other inmates**
(11) Failure to accomplishment of penal measures
(12) **Disciplinary reports**
(13) Escape or absconding
(14) Grade regression
(15) Breaching prison permit
(16) **Poor childhood adjustment**
(17) **Distance from residence to prison**
(18) **Educational level**
(19) Problems related with employment
(20) **Lack of financial resources** (S)
(21) **Lack of viable plans for the future**
(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 (S)
(25) Criminal or antisocial friends
(26) Member of social vulnerable groups
(27) **Relevant criminal role**
(28) **Gender violence victims (only women)**
(29) Responsibility for the care of family
(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** (S)
(35) Personality disorder related to anger or violent behaviour
(36) Poor stress coping
(37) **Self-injury attempts or behaviour** (S)
(38) **Pro criminal or antisocial attitudes**
(39) **Low mental ability**
(40) **Recklessness**
(41) Impulsiveness and emotional instability
(42) **Hostility**
(43) **Irresponsibility**
Other RisCanvi Screening items
(1) Institutional/prison misconduct
(2) Escapes or breaches of permits
(3) Problems with drugs or alcohol use
(4) Hostile or pro criminal attitudes

RisCanvi is applied multiple times during an inmate's period in prison; the official recommendation is to do so every six months or at the discretion of the case manager. Generally, the screening version RisCanvi-S is applied to all inmates when they enter the prison. The outcome of RisCanvi-S can be "high-risk" or "low-risk." If the outcome is low-risk for all four criteria, the same RisCanvi-S protocol is repeated after six months. Otherwise, in the case of high-risk levels or significant change in an inmate's situation, the complete version RisCanvi-C is applied. The outcome of RisCanvi-C can be "high-risk," "medium-risk," or "low-risk." When the risk levels measured by RisCanvi-C are medium or high, the next evaluation is again a RisCanvi-C; otherwise, the RisCanvi-S is used [Andrés-Pueyo et al., 2018].

3.3.2 Dataset

The anonymized dataset used on this research comprises 7,239 offenders who first entered the prison between 1989 and 2012 and who were evaluated with the RisCanvi protocol between 2010 and 2013. We kept only offenders for which nationality information was recorded, that comprises 2,634 offenders. Among this population, 256 inmates had violent recidivism after being released. The data includes all of the information for the two RisCanvi versions (RisCanvi-S and RisCanvi-C). All inmates were evaluated at least once, and depending on the time they spend in prison, 46% had a second evaluation, 18% a third one, and less than 6% had four to eight evaluations. On average, inmates with only one evaluation remain for about three months in prison, while inmates with four evaluations on average spend two years before being released on parole or regaining freedom. There is no evaluation after an inmate's release.

Handling missing items. In the RisCanvi data, there were some missing items. Static items were replaced with the value of their counterparts from the previous or next evaluations. These static items were 7 items from the 43 RisCanvi-C risk factors in Table 3.1 (items 8, 16, 22, 23, 32, 33, and 39). Moreover, items with a yes/no/uncertain answer in which there was a missing item, had missing values replaced with "uncertain."

After the above replacements, we removed the cases with irreplaceable missing items from the sample, leaving 2,582 people in the final data set.

3.3.3 Violent Recidivism (REVI)

Violent crimes are more costly to victims and the criminal justice system compared to other crimes [Rubin et al., 2008]. Also, violent recidivism can be more clearly established and hence the ground truth is more reliable. Therefore, in this work we focus on RisCanvi to assess Violent Recidivism ("REVI" in the protocol) risk in sentenced inmates. REVI risk is an outcome predicted using a sub-set of risk factors shown in boldface on Table 3.1 (23 out of the 43 risk factors of the RisCanvi-C version), plus two demographic features (gender and nationality). In RisCanvi-C, the REVI score has been computed by applying the summation of these features in a hand-crafted formula, then using two cut-offs, obtaining three REVI risk levels (details in [Andrés-Pueyo et al., 2018]).

First, we compare the distribution of REVI risk scores by nationality and age groups. We do not consider a grouping by gender as the number of women in our sample is too small to draw robust conclusions. Figure 3.1 shows the distribution of REVI risk scores per group, while the average recidivism rate per group is shown on Table 3.2. For age groups we use 30 years old as a cut-off, as criminology research suggests that the types of offense and context are different for people under 30 and over 30 (see, e.g., [Ulmer and Steffensmeier, 2014]). This age is also used as a cut-off for young and old people in the design of the RisCanvi protocol. In our dataset, the majority of the population are Spanish nationals (68%) and older than 30 years old (67%). As can be seen in Figure 3.1, foreigners tend to have lower REVI risk scores compared to Spanish. Also, the distribution of REVI score by age shows that old and young inmates have similar risk score distributions. In this dataset, we observe that foreigners, which have lower risk, have less tendency to change in REVI risk compared to Spanish nationals. For the same reason, inmates older than 30 years old are slightly less likely to change in REVI risk compared to younger offenders.

Second, given our goal is to study the consequences of selectively

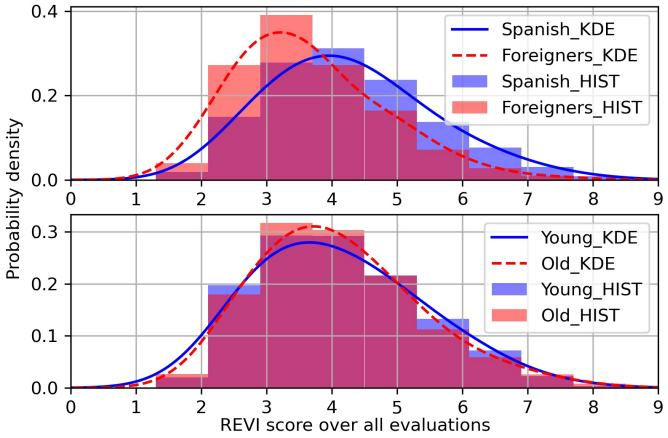


Figure 3.1: Violent recidivism score (REVI) distribution by nationality and age. Foreigners tend to have slightly lower REVI scores than nationals. Both “young” ($\leq$ 30 years old) and “old” ($>$ 30 years old) have similar REVI scores. Smooth curves are obtained by Kernel Density Estimation (KDE).

Table 3.2: Violent Recidivism Rate (Average)

Group	Spanish	Foreigners	“Young” (age $\leq$ 30)	“Old” (age $>$ 30)
Violent recidivism	12.4%	8.9%	12.7%	10.8%

re-evaluating to reduce the overall number of evaluations, we look at how risk changes. Figure 3.2 depicts REVI risk changes in RisCanvi evaluations separated by 6, 12, and 18 months intervals. In general, we note a larger probability of REVI risk changes when the interval is longer. Also, when the risk changes, there is more tendency to decrease. For medium- and high-risk inmates we observe a tendency to lower risk levels in the

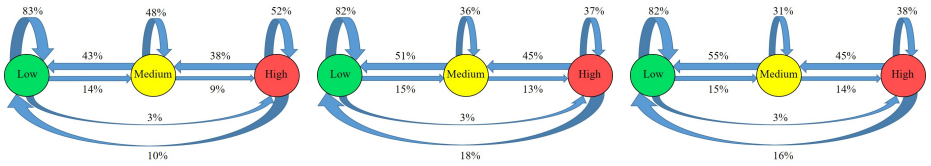


Figure 3.2: REVI variations in 6 months (left), 12 months (center), and 18 months (right). Low-risk inmates tend to have the same risk in successive evaluations, whereas medium- and high-risk inmates tend to exhibit less risk.

next evaluation, and for low-risk inmates a tendency to continue being evaluated as low-risk. This can be due to the effects of the rehabilitation and other interventions done while in the prison and it goes contrary to the incarceration effects noted in some works [Gendreau et al., 2000].

Third, to have more insights on the RisCanvi dataset and REVI risk scores, we create a new machine learning classifier using 43 risk factors, three demographic features (gender, age, and nationality), and REVI risk level (low, medium, and high). We use an off-the-shelf multi-layer perceptron as learning scheme, which performed better than other methods we tested for this task (including logistic regression and support vector machines). The cases considered in this model are 2,028 (out of 2,582) who are sentenced (not awaiting trial), that were released at most 9 months after their last RisCanvi evaluation, and for which violent recidivism (or its absence) was recorded at most two years after their freedom. Using 5-fold cross validation, the average AUC of the model is 0.69. In comparison, RisCanvi-C obtains an AUC of 0.68. These values are in line with that of similar tools used in other countries, which tend to have AUC values in the range of 0.57-0.74 [Brennan et al., 2009, DeMichele et al., 2018, Desmarais et al., 2016].

3.4 Methodology

Normally, each inmate is evaluated every six months; we test the effects of performing less RisCanvi evaluations by selectively postponing the

evaluation of an inmate for two periods or three periods (i.e., 12 months or 18 months). As ground truth, the cases over which we test are only inmates who actually received four evaluations regularly in an 18 month period, so we know whether their risk changed or not.

Three ML models corresponding to periods of 6, 12, and 18 months, are created to predict the necessity for a new evaluation at the end of a period. We use different ML methods, such as logistic regression, multi-layer perceptron (MLP), and support vector machines (SVM). The features used for the time prediction models are Violent Recidivism (REVI) items (boldface in Table 3.1), REVI score, gender, nationality, and age at the time of evaluation. Additionally, in 12- and 18-month models the output(s) from the shorter-period model(s) are used as additional features.

The whole data is split into two sets. The first set is divided into training and validation and used to create the 6, 12, and 18 months risk change models and for performing model-based evaluation. The second set is used for both model- and system-level evaluations. In the system-level evaluation, this set is used to schedule the evaluation of inmates using the prediction models as explained next. In the simulation, every six months, a fraction σ (the selection rate) of the inmates with the highest probability of REVI risk change (obtained in the previous six months period using ML models) are selected for evaluation. Those evaluated have their REVI risk change probabilities recomputed for the next six months.

The split for model-level and system-level evaluation is done k times using k-fold cross-validation, reporting average results. The part for model-based evaluation is also split using k-fold cross validation.

3.4.1 Model-Level Evaluation Methodology

We consider changes in REVI risk level between two evaluations separated by a time interval (6, 12, or 18 months). This is modeled as a binary classification task in which we have to predict whether there will be a change or not at the end of the period. If risk changes we have a positive example, if risk does not change we have a negative example. The predictive performance of the ML models is evaluated using Area Under

the ROC curve (AUC-ROC).

3.4.2 System-Level Evaluation Methodology

System-level evaluation is done through a simulation of 18 months. In the simulation, at time 0, the three ML models (6-, 12- and 18-month models) are applied to the second set (introduced in Section 3.4) and three series of predictions for the next 6, 12, and 18 months are obtained for each inmate. Then in each six months period, a fraction σ of the inmates with the highest probabilities of REVI risk change are selected for evaluation and the rest have their evaluation postponed. Whenever selected individuals are evaluated, we apply the ML models over their actual RisCanvi evaluation (which is known), and based on the new obtained predictions, the old predictions are updated.

By selecting only a part of inmates each time, there will be some omissions or *missed changes*: cases who experience REVI risk change but are not evaluated and hence not detected. As risk change leads to a positive class, *Missed changes* can be interpreted as False Negative Rate (FNR) and formulated as *undetected changes* divided by *total changes*.

Thus, we undertake a cost-benefit analysis. The *cost* is the fraction of inmates who experience an undetected risk change. The *benefit* (equal to $1 - \sigma$) is the fraction of evaluations that are not done, i.e., the resources saved because not all inmates have to be evaluated. The cost of the baseline (current method) is 0, as all risk changes are detected, and its benefit is also 0, as this is equivalent to have a selection rate $\sigma = 1$.

We compute the *cost* (*missed changes*) in two ways: cases with undetected REVI risk increase and cases with undetected REVI risk decrease. Studying missed risk increases is important since the outcome can be dangerous to society. Also, postponing the evaluation of inmates who have less risk now may have negative psychological effects on the inmates, and can have a negative impact on their rehabilitation. Furthermore, to study people with REVI risk increase and decrease more precisely, we create ML models for each group separately.

An additional metric we compute is the *average number of evaluations*

per inmate, a figure that we compute globally as well as per-group as explained next. This is a number between 1 (inmate is evaluated at the beginning and at some point in the next 18 months) and 3 (inmate is evaluated at the beginning, and then at 6, 12, and 18 months). Note we do not count the initial evaluation in this computation because it is shared among all settings.

Finally, we also compute the *average number of unnecessary evaluations*, which are REVI risk evaluations in which the outcome is the same risk level as the previous evaluation. Only a perfect predictive model (an oracle) could reduce this number to zero.

3.4.3 Algorithmic Fairness Evaluation Methodology

Finally, we consider *algorithmic fairness* by comparing metrics across groups. First, we study whether ML models show any discrimination in the prediction of REVI change against "Spanish" or "foreigners" and "young" or "old" inmates. Second, we check the disparate impact in the average number of evaluations along nationality and age. Third, for the obtained rate of undetected changes, we study if there is a disparate mistreatment (FNR discrepancy) between nationality and age sub-groups. Finally, we study if there is a disparate effect in terms of the average number of unnecessary evaluations between these groups.

3.5 Results

3.5.1 Model-Level Evaluation

To evaluate the predictive performance of the ML models, the validation data of the first set and the whole second set (introduced in Section 3.4) are used. Among MLP (Multi-Layer Perceptron), LR (Logistic Regression) and SVM (Support Vector Machines), the best results in terms of the AUC-ROC were obtained using MLP with a single hidden layer having 100 neurons. In Table 3.3, the results in terms of the AUC-ROC are presented. According to the AUC values, we can see that the three ML models (6-,12-,

Table 3.3: AUC of Risk Change Prediction Models

Model	6-month	12-month	18-month
MLP	0.78	0.75	0.74
LR	0.77	0.76	0.72
SVM	0.75	0.76	0.74

and 18-month) have good accuracy. We remark that AUC values are mostly dominated by low-risk individuals, who are the majority in this data (the average percentage of low-risk people is 70%).

3.5.2 System-Level Evaluation

Missed/Undetected Changes

As mentioned, given only a fraction σ of inmates is evaluated, there are missed or undetected changes. In Figure 3.3, the fraction of REVI missed changes (increase or decrease), REVI missed increases and REVI missed decreases are shown for different selection rates of the inmates. We see a much smaller number of missed changes compared to selecting inmates at random (the result for random selection of the inmate is shown by "Chance" curve in Figure 3.3). Also, the missed values for REVI change, increase, and decrease are very similar. This curve represents a series of trade-offs, and the specific trade-off should be chosen by the experts depending on the cost they assign to different aspects. We consider a selection rate of 50% in the following for concreteness, but remark that other selection rates could be chosen and would be analyzed in the same manner. Thus, by selecting 50% of the inmates with the highest probability of REVI change each time, we would miss about 12% to 15% of changes. Moreover, we note that we are more accurate at avoiding missed changes in a short time frame (6 months) compared to longer periods (12 or 18 months) and by selecting more than 50% of the inmates in 6-month model, there would

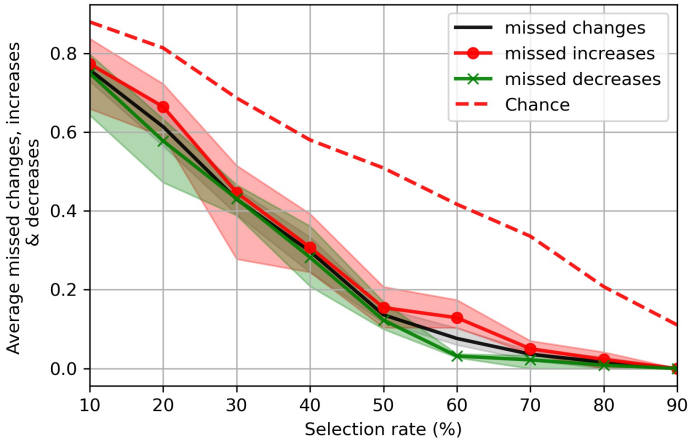


Figure 3.3: REVI missed changes. There is a much smaller number of missed changes compared to selecting inmates at random (“Chance”).

be zero missed changes, because REVI tends to change in longer time intervals, as explained in Section 3.3.3.

In addition, by evaluating the ML models created separately for the two groups with REVI risk increase and decrease as shown on Table 3.4, we conclude that the REVI risk decrease model shows more accuracy and almost less missed values (according to Figure 3.3) compared to REVI change and REVI increase models.

Evaluations Per Inmate

Our goal is to reduce the average number of evaluations performed for each inmate. According to our results in Figure 3.4, the average number of evaluations performed by our method is smaller than the 3 evaluations required by standard RisCanvi in an 18 months period. For instance, by selecting $\sigma \approx 50\%$ of inmates (those with the highest probability of REVI change), there would be 1.5 evaluations per inmate on average.

Unnecessary evaluations are situations where an evaluation is per-

Table 3.4: AUC of Risk Decrease and Risk Increase Prediction Models

Model	6-month		12-month		18-month	
Risk	Decrease	Increase	Decrease	Increase	Decrease	Increase
MLP	0.88	0.63	0.84	0.61	0.83	0.59
LR	0.87	0.60	0.87	0.58	0.85	0.53
SVM	0.87	0.63	0.88	0.63	0.89	0.58

formed and yields the same risk score as the previous evaluation. Our models lead to less unnecessary evaluations on average compared to the RisCanvi as shown on Figure 3.5. Again, by selecting $\sigma \approx 50\%$ of the inmates for evaluation, the average number of unnecessary evaluations per inmate would be close to 1.0, which is less than standard RisCanvi (2.4 unnecessary evaluations per inmate on average).

3.5.3 Algorithmic Fairness Evaluation

In Table 3.5, the results for the analysis of equity in accuracy (AUC) are shown for nationality (Spanish and foreigners), and age (young and old inmates) groups. The AUC results of the ML models show more accuracy for foreigners than for Spanish nationals in general, despite the latter comprising about 68% of this sample. For the age groups, the difference is small.

Next we check if there is parity in the average number of evaluations per Spanish and foreigner in Figure 3.4, for various selection rates. We observe that on average Spanish nationals and foreigners receive 1.69 (with a **spread** of 0.12 between the min and max among folds) and 1.08 (spread: 0.21) evaluations respectively for the selection rate of 50%. These results show more average number of evaluations per Spanish compared to foreigners. The same analysis for "young" vs "old" inmates is shown in Figure 3.4. The results show that for the selection rate of 50%, there are

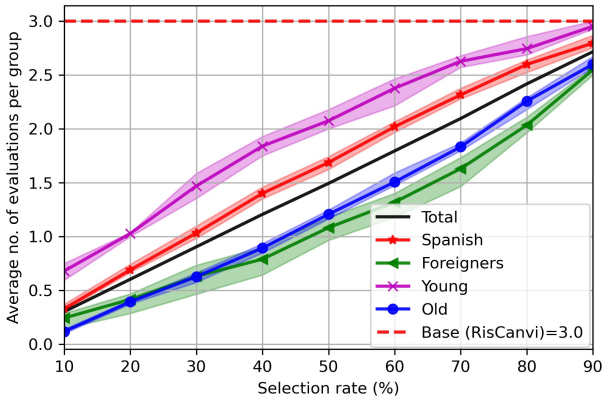


Figure 3.4: Average number of evaluations per person. Our method leads to a smaller number compared to the standard RisCanvi which requires 3 evaluations in an 18 months period. However, without mitigation measures for algorithmic bias, the evaluation rate is different across groups.

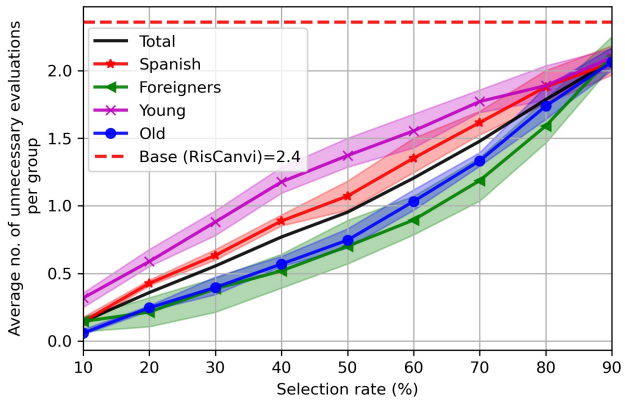


Figure 3.5: Average number of unnecessary evaluations per person. Our method leads to a smaller number compared to the standard RisCanvi which performs 2.4 unnecessary evaluations in an 18 months period. There are different rates of unnecessary evaluations across groups which is due to their different evaluation rates.

Table 3.5: AUC of Risk Change Prediction Models per Group

Model	6-month	12-month	18-month
Spanish	0.75	0.70	0.70
Foreigners	0.85	0.85	0.78
Young (age $\leq$ 30)	0.77	0.74	0.76
Old (age $>$ 30)	0.79	0.75	0.73

2.08 (spread: 0.18) and 1.20 (spread: 0.08) average number of evaluations per young and old respectively which represent more average number of evaluations per young than old inmate.

According to the results obtained for the average number of unnecessary evaluations in Figure 3.5, for the selection rate of 50%, on average Spanish with the value of 1.07 have more unnecessary evaluations than foreigners with the value of 0.7, since they have more average number of evaluations (Figure 3.4). Also considering a selection rate of 50%, the results for "young" and "old" inmates are 1.37 and 0.75 respectively, which shows that on average there are more unnecessary evaluations for younger inmates; this is consistent with the results of the average number of evaluations in these sub-groups (Figure 3.4).

Furthermore, the missed changes (FNR) for each sub-group of nationality and age are shown in Figure 3.6. For the selection rate of σ =50% missed changes for Spanish and foreigners are 0.14 (spread: 0.07) and 0.13 (spread: 0.13) respectively. For young and old sub-groups, the results show 0.16 (spread: 0.10) and 0.12 (spread: 0.04) missed changes respectively. For this particular cut-off value, and in general for selection rates larger than 50%, differences in missed changes are relatively small.

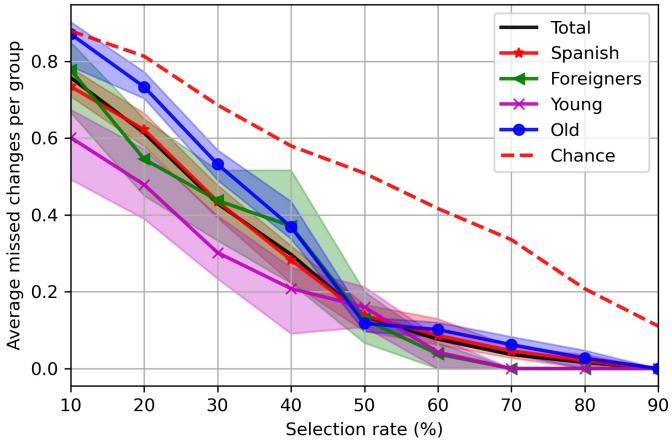


Figure 3.6: Average missed changes per group. For the selection rate of 50%, the missed changes difference in nationality (Spanish vs foreigners) and age (young vs old) is too small.

3.6 Mitigating Algorithmic Bias

The method we have described could introduce a disadvantage for a group of inmates if that group is consistently evaluated more often or less often than another. We would prefer to select inmates for evaluation at the same rate σ independently of their nationality, age, or other characteristics. In our experiments, by moving the decision boundary we select inmates so that the selection rate is similar for different nationality and age groups. First, we select a fraction σ of inmates having the highest probability of Violent Recidivism (REVI) change from both groups by nationality (nationals and foreigners). Second, we add cases with high probability of REVI change and remove cases with low probability of REVI change until both age groups (“young” and “old”) are equalized.

The rate of missed changes after applying the mitigation process increases by about three percentage points as shown on Figure 3.7. By selecting $\sigma \approx 50\%$ of the inmates with the highest probability of REVI

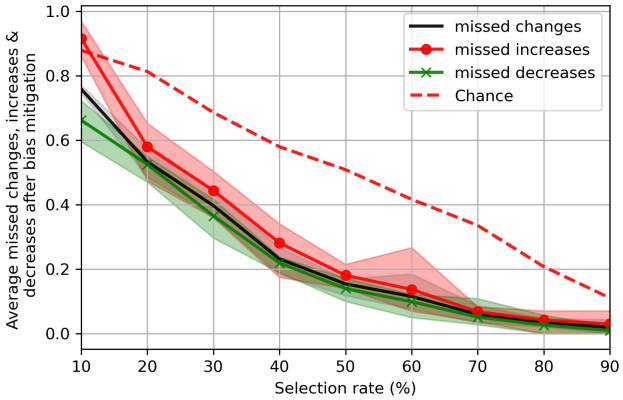


Figure 3.7: REVI missed changes per selection rate after bias mitigation. For the selection rates of 50%, missed change increases by less than three percentage points compared to its value before bias mitigation.

change, we would miss between 14% to 18% in REVI changes, increases, and decreases, compared to 12%-15% missed changes before bias mitigation. Also, according to Figure 3.8, the obtained results for the selection rate of 50% show missed changes of 0.16 and 0.13 for Spanish and foreigners respectively which represents a small range difference. These results for young and old inmates are 0.22 and 0.10 respectively which shows more missed changes for younger inmates.

The average number of evaluations per inmate after applying the bias mitigation procedure is shown on Figure 3.9. This number increases in the case of foreigners, and decreases in the case of Spanish nationals. This also decreases for young inmates and increases for old inmates. This is because we are correcting a disparity that was present before applying the mitigation. Our results show that for the selection rate of 50% of the inmates with the highest probability of REVI change, on average there are about 1.7 evaluations per inmate (same value for young and old inmates, 1.6 for Spanish nationals and 1.9 for foreigners); compare this to 1.5 evaluations per inmate before bias mitigation.

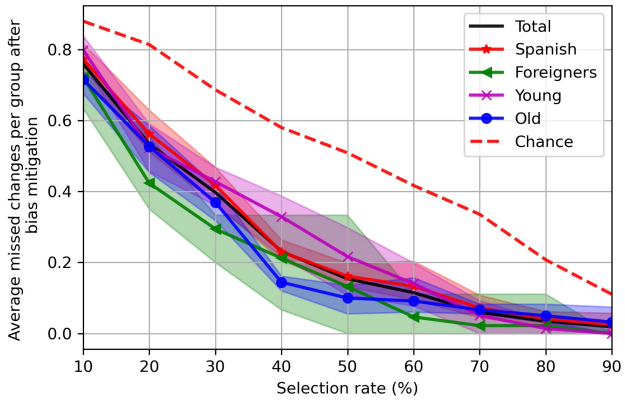


Figure 3.8: Average missed changes per group after bias mitigation. For the selection rate of 50% there is a small difference in missed changes of nationality groups and in age groups there are more missed changes for younger inmates.

The average number of unnecessary evaluations per person after bias mitigation is shown on Figure 3.10. This number is 1.2 (1.0 for young inmates, 1.2 for old inmates, 1.0 for Spanish nationals and 1.5 for foreigners) for the selection rate of 50%; compare this to about 1.0 unnecessary evaluations per inmate before bias mitigation. The reason is that balancing the evaluation rate caused less evaluations and accordingly less unnecessary evaluations for Spanish nationals who were evaluated more often in the scenario without bias mitigation. Something similar happens in the case of the unnecessary evaluations of “young” vs “old” inmates: the values are lower for young inmates and higher for old inmates.

Finally, if we wanted to ensure a specific bound on the number of missed changes, this would require a particular minimum selection rate. Figure 3.11, shows the evaluation rate needed to have on expectation a certain amount of missed changes before and after the mitigation. According to the results, missed change differences are small before and after the mitigation, and in particular for the selection rate of 50%, the difference is almost zero.

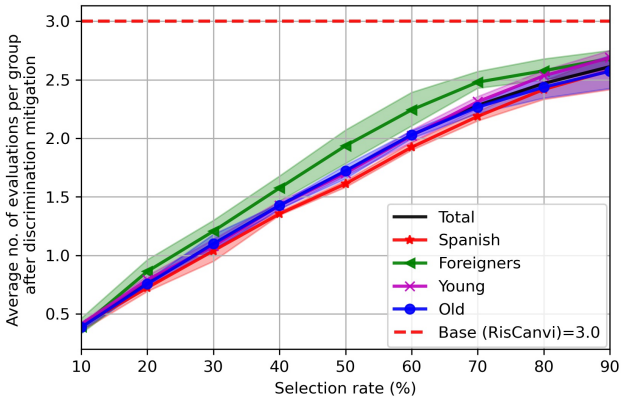


Figure 3.9: Average number of evaluations per person after bias mitigation. For the selection rate of 50%, there are about 1.7 evaluations per inmate which shows a small increase of two percentage points compared to 1.5 evaluations per inmate before bias mitigation.

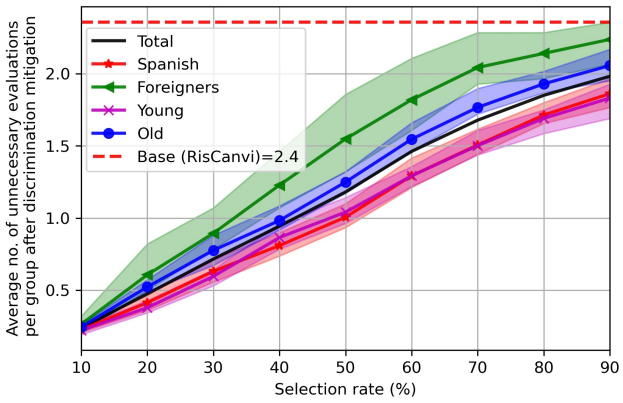


Figure 3.10: Average number of unnecessary evaluations per person after bias mitigation. For the selection rate of 50%, this number is 1.2 which has a small increase of two percentage points compared to about 1.0 unnecessary evaluations per inmate before bias mitigation.

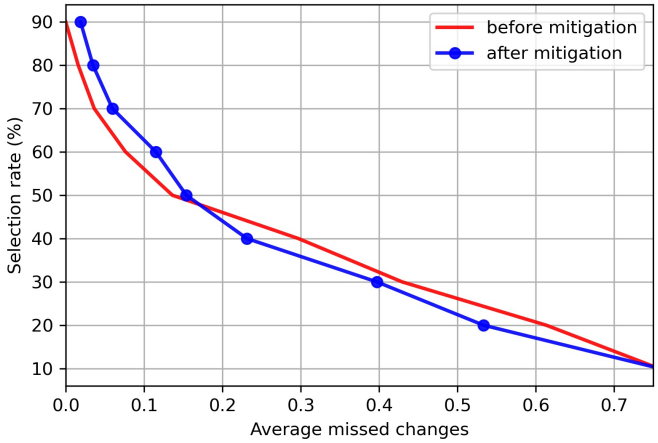


Figure 3.11: Evaluation rates per missed change before and after the mitigation. For the selection rates more than 50%, less than 5% more evaluations are needed to have no variation in the missed changes after the mitigation.

3.7 Discussion

We used ML models to predict changes of violent recidivism risk, these models have AUC in the range of 0.74-0.78. In the cost-benefit analysis of selecting the inmates for the next RisCanvi evaluation, we obtained a cost, in terms of missed changes, of nearly 14% when selecting the top σ =50% of the inmates with the highest probability of Violent Recidivism (REVI) change. The benefit is that the number of evaluations is halved. Other points in the cost-benefit trade-off curve can be used. Marginal benefits (further drops in missed changes) are decreasing, showing some saturation effect after reaching about σ =70% selection rate.

We observe this method introduces some differential treatment across groups such as disparate impact in the evaluation rates and disparate mistreatment with regard to undetected risk changes (false negative rates). Specifically, as the results showed in Figure 3.4, the average number of evaluations that a Spanish national must undergo is more than a foreigner.

The source of this difference is that according to the results obtained in Section 3.3.3, foreigners are less likely to change in REVI risk, so they should expect to be less selected for the next evaluation. Similarly, the difference in the average number of evaluations along age (Figure 3.4), can be traced to the same reasons, a lower tendency in old offenders to change in REVI risk.

Since there is a trade-off in mitigating both disparate impact in evaluation rate and disparate mistreatment in missed changes simultaneously, by moving the decision boundary, we mitigated the disparity in the evaluation rates along both nationality and age groups with a small additional loss of missed changes.

3.8 Conclusions and Future Work

In this paper, we employ ML-based methods to select the inmates for the next evaluation of the Violent Recidivism (REVI) risk in the RisCanvi protocol. These models showed good results in terms of AUC (0.74-0.78), which resulted in fewer evaluations per inmate compared to the standard RisCanvi, which in turn leads to save time, expenses and staff in the evaluations. This benefit has been obtained in exchange for some missed changes (about 14% when selecting 50% of the inmates with the highest probability of REVI change).

Furthermore, analyzing the fairness of the ML models along nationality (Spanish and foreigners) and age (young and old) led to the following results: in terms of AUC, the models are more accurate for foreigners than Spanish nationals and there is no significant difference in age sub-groups. In terms of missed changes (false negative rates), for the selection rate of 50%, the disparate mistreatment is less than 0.04 among both nationality and age sub-groups. There is a disparate impact in the average number of evaluations which shows lower number of evaluations in foreigners and older inmates on average. This also translates to a difference in the average number of unnecessary evaluations per group.

Applying a mitigation method to gain parity in the rate of evaluations

along nationality and age leads to a small increase in missed changes which is less than one percentage point for the selection rate of 50%. We obtained parity in the average number of evaluations per inmate along both nationality and age which is 1.7; about half of the evaluations done in RisCanvi.

The method used in this study can also be used for other RisCanvi criteria: self-directed violence, violence to other inmates or prison staff, and breaking prison permits to see if there is still such a possibility to perform less evaluations in exchange for a small number of missed changes, while preserving equality between different groups. We must note, however, that our work is validated on data from inmates that have four evaluations and spend on average two years (or more) in prison, and might not be applicable for people receiving shorter sentences.

The freed staff time of using this method can go to programs focused on reducing the likelihood of recidivism instead of merely predicting it, something that have been advocated by researchers critical of current ML-based risk assessments [Barabas et al., 2017].

The problem and approach raised in this work is general enough to be applicable in other areas where appraisals and predictions about individuals are done (e.g., education, public health, information security, immigration, social benefits, and so on).

Chapter 4

ENHANCING A RECIDIVISM PREDICTION TOOL WITH MACHINE LEARNING: EFFECTIVENESS AND ALGORITHMIC FAIRNESS

4.1 Introduction

Risk assessment is a necessary process in many important decisions such as public health, information security, project management, auditing, and criminal justice. Since the 1920s, violence risk assessment tools have been progressively used in criminal justice by probation and parole officers, police, and psychologists to assess the risk of harm, sexual, criminal, and violent offending in more than 44 countries [Singh et al., 2014, Kehl and Kessler, 2017]. In comparison to traditional prediction methods and unstructured clinical judgments, risk assessment tools offer superior accuracy and performance [Grove et al., 2000]. In this regard, factors such as the availability of large databases, inexpensive computing power, and developments in statistics and computer science have brought an increase

in the accuracy and applicability of these structured tools [Berk, 2012]. Such advances have effectively increased the use of tools based on Machine Learning (ML) in criminal justice decisions for risk forecasting [Berk and Hyatt, 2015, Berk et al., 2016, Berk, 2017]. Today, various semi-structured protocols for assessing risk of recidivism can be found in different countries including the U.S. [Desmarais and Singh, 2013], the U.K. [Howard and Dixon, 2012], Canada [Kröner et al., 2007], Austria [Rettenberger et al., 2010b], and Germany [Dahle et al., 2014]. In Spain, among current violence risk assessment tools including SAVRY, PCL-R, HCR-20, SVR-20, and SARA, RisCanvi is a relatively new tool for risk assessment of recidivism. It was originally developed in 2009 in response to concerns of Catalan prison system officials regarding violent recidivism among offenders after their sentences.

Research contribution. In this study, the effectiveness and algorithmic fairness of RisCanvi risk assessment tool are evaluated in comparison to ML models such as logistic regression, perceptron, and support-vector machines, in violent and general recidivism prediction. The effectiveness of the ML models are evaluated and compared to RisCanvi in terms of various metrics including AUC, Generalized False Positive (GFPR), and Generalized False Negative (GFNR). Also, potential algorithmic bias introduced by the ML methods is evaluated in both violent and general recidivism prediction. Given that model learning may lead to unfairness [Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Tolan et al., 2019], the impact of the obtained ML models is compared along nationality (national origin vs foreign origin) and age (young vs old). Then some differences are addressed through a mitigation procedure [Pleiss et al., 2017], which try to equalize GFPR across nationality and age groups while preserving the calibration in each group.

The rest of this paper is organized as follows. Section 4.2 outlines related work. In Section 4.3, the RisCanvi risk assessment tool and the dataset used in this study are described. The methodology including the ML models and algorithmic fairness analysis are presented in Section 4.4. Results are given in Section 4.5, and a procedure to mitigate algorithmic discrimination is used in Section 4.6. Finally, the results are discussed and

the paper is concluded in Section 4.7.

4.2 Related Work

The introduction of algorithms for risk assessment in criminal justice is a controversial topic, and perhaps one that has motivated a great deal of research on algorithmic fairness.

In seminal research done by investigative journalism organization ProPublica [Angwin et al., 2016, Larson et al., 2016] it was concluded that a widely-used program named Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) is biased against African American defendants. A follow-up study [Hamilton, 2019] found that COMPAS outcomes systematically over-predict risk for women, thereby indicating systemic gender bias. However, the findings of the ProPublica study were rejected by Northpointe (COMPAS developer), claiming their algorithm is fair because it is well calibrated [Dieterich et al., 2016]. Moreover, in this report it is shown that the COMPAS risk scales exhibit accuracy equity and predictive parity.

In contrast to the case of COMPAS, other studies have shown that other risk assessment tools such as the Post Conviction Risk Assessment (PCRA), the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI) do not exhibit racial bias in the recidivism prediction [Skeem and Lowenkamp, 2016, Perrault et al., 2017]. In a more recent study focused on SAVRY [Tolan et al., 2019, Miron et al., 2020], it is shown that although machine learning models could be more accurate than the simple summation used to compute SAVRY scores, they would introduce discrimination against some groups of defendants.

There are many different definitions of algorithmic fairness [Narayanan, 2018], some of which are incompatible with one another. It is impossible to satisfy all of them simultaneously except in pathological cases (such as a perfect classifier), and in general it is impossible to maximize algorithmic fairness and accuracy at the same time [Berk, 2019, Berk et al., 2018].

Hence, there are necessary trade-offs between different metrics [Berk et al., 2018, Chouldechova, 2017, Kleinberg et al., 2016]. In this regard, some studies [Hardt et al., 2016, Zafar et al., 2017, Woodworth et al., 2017] try to mitigate potential algorithmic discrimination by satisfying equalized odds or in other words avoiding disparate mistreatment along different sensitive groups. In addition, due to the importance of the calibration in risk assessment tools [Berk et al., 2018, Dieterich et al., 2016], some previous work has also tried to minimize error disparity across groups while maintaining calibrated probability estimates [Pleiss et al., 2017].

The most closely related previous work is Pleiss et al. [Pleiss et al., 2017], where algorithmic bias in a machine learned risk assessment (COMPAS) is minimized by equalizing generalized false positive rates along different races, finding this equalization to be incompatible with calibration. In contrast, in the work presented on this paper, we start from an expert-based risk assessment method, which is not machine learned, and propose a new machine learning model to replace it, describing the effects of algorithmic bias mitigation on both the original and the machine learned model. Additionally, we find that in RisCanvi equalization along nationality and age groups is not entirely incompatible with calibration.

4.3 RisCanvi Dataset

4.3.1 The RisCanvi Risk Assessment Tool

RisCanvi was introduced as a multi-level risk assessment protocol for violence prevention in the Catalan prison system in 2009 [Andrés-Pueyo et al., 2018]. This protocol is applied multiple times during an inmate's period in prison; the official recommendation is to do so every six months or at the discretion of the case manager. RisCanvi is not a questionnaire. Instead, each inmate is interviewed by professionals. In the original RisCanvi protocol, risk is determined for each inmate relative to four possible outcomes: self-directed violence, violence in the prison facilities, committing further violent offenses, and breaking prison permits. A fifth risk score was introduced more recently for general recidivism [Singh

et al., 2018].

Two versions of the RisCanvi protocol were created, an abbreviated one of 10 items for screening (RisCanvi-S), and a complete one of 43 items (RisCanvi-C). Risk items can be categorized into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, Attitudes/ Personality. These items can also be divided into static factors (such as “criminal history of family” and “age of starting violent activity”) and dynamic factors (such as “member of socially vulnerable groups” and “pro-criminal or antisocial attitudes”).

4.3.2 Dataset

The anonymized dataset used on this research comprises 7,239 offenders who first entered the prison between 1989 and 2012 and who were evaluated with the RisCanvi protocol between 2010 and 2013. Only offenders for which nationality information was recorded were kept that comprises 2,634 offenders. The result population was filtered in terms of their violent/general recidivism, freedom and last RisCanvi evaluation dates considering the following conditions: inmates who were released at most 9 months after their last RisCanvi evaluation, and for which violent/-general recidivism (or its absence) was recorded at most two years after their release. Finally, samples with the size of 2,027 (out of 2,634) were reached. Among this population, 146 committed a violent offence (violent recidivism) and 310 committed a violent or non-violent offence (general recidivism) after being released. The data includes all of the information for the two RisCanvi versions (RisCanvi-S and RisCanvi-C). This study is focused on the RisCanvi-C protocol which is the complete version done after RisCanvi-S and it consists of more risk factors which results in three risk levels (low, medium, and high).

4.3.3 Violent and General Recidivism

This work is focused on RisCanvi protocol to assess Violent Recidivism (“REVI” in the RisCanvi manual) and General Recidivism (“REGE” in

the RisCanvi manual) risks in sentenced inmates. REVI and REGE risks are outcomes predicted using two different sub-sets of risk factors. REVI risk is obtained using 23 items out of the 43 risk factors of the RisCanvi-C version plus two demographic features (gender and nationality) and to compute REGE risk, 14 items (out of 43 risk factors of the RisCanvi-C version) are used. In RisCanvi-C, each of the REVI and REGE scores has been computed by applying the summation of their related features in a hand-crafted formula, then using two cut-offs, obtaining three risk levels (details in [Andrés-Pueyo et al., 2018]).

The distribution of REVI and REGE risk scores in the last RisCanvi evaluation is compared by nationality and age groups in Figure 4.1 and Figure 4.2 respectively. Grouping by gender is not considered as the number of women in the sample is too small to draw robust conclusions. The comparison shows that recidivism risk scores have approximately similar distributions along nationality and age group except for the REVI score in nationality group which shows that foreigners tend to have lower REVI risk scores compared to Spaniards. For age groups, 30 years old is used as a cut-off, as criminology research suggests that the types of offense and context are different for people under 30 and over 30 (see, e.g., [Ulmer and Steffensmeier, 2014]). This age is also used as a cut-off for young and old people in the design of the RisCanvi protocol. In the present dataset, the majority of the population are Spanish nationals (70%) and older than 30 years old (74%).

According to the average violent and general recidivism rates shown on Table 4.1 for nationality and age groups, it can be seen that in general, foreigners and older offenders have a lower recidivism rate.

4.4 Methodology

The goal of this study is to compare the effectiveness and fairness of Machine Learning (ML) models and the RisCanvi risk assessment tool in the prediction of violent and general recidivism.

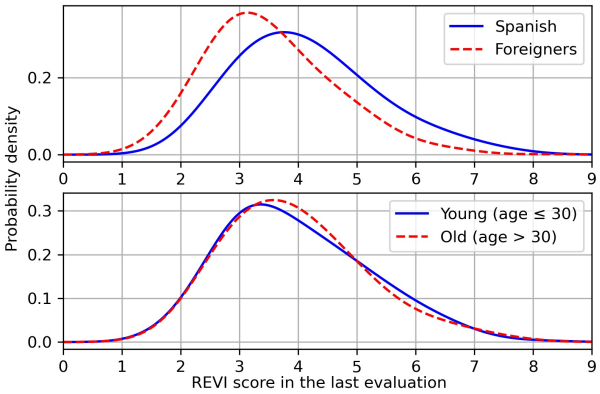


Figure 4.1: Distribution of REVI risk scores in the last RisCanvi evaluation for gender and nationality groups. REVI distribution is approximately similar along age group but in nationality group, lower REVI risk scores are found for foreigners compared to Spaniards.

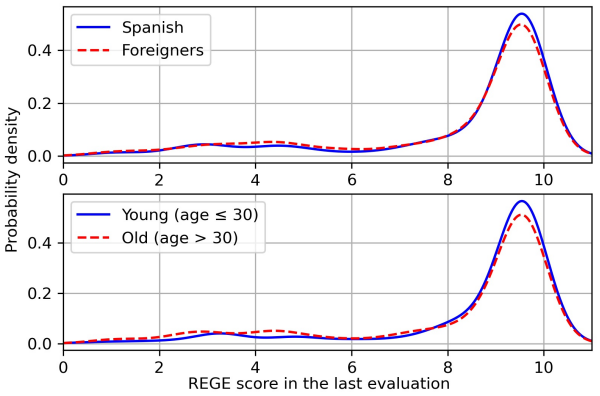


Figure 4.2: Distribution of REGE risk scores in the last RisCanvi evaluation for gender and nationality groups. REGE risk scores have approximately similar distributions along both nationality and age groups.

Table 4.1: Recidivism rates in nationality and age groups.

Group	Spaniards	Foreigners	Young	Old
Violent recidivism rate	8%	5%	9%	7%
General recidivism rate	17%	13%	19%	14%

4.4.1 ML-based Models

Different ML methods, such as logistic regression, multi-layer perceptron (MLP), and support vector machines (SVM) are used. The ground truth is the violent/general recidivism, which is recorded at most two years after the inmate’s release.

Different sub-sets of features are tested as input to the ML models, such as 43 RisCanvi-C items, Violent Recidivism (REVI)/General Recidivism (REGE) risk items, and a set of features selected from 43 risk items using a feature selection method. In addition, three demographic features (gender, nationality, and age) are used as general input features. Finally, the average of REVI/REGE risk scores over all of the RisCanvi evaluations from the first to the last evaluation is added.

The split of the two sets is done k times using stratified k-fold cross-validation, reporting average results.

4.4.2 Algorithmic Fairness

Algorithmic fairness is evaluated by comparing the impact of the risk prediction method across nationality and age groups.

As it is known, model calibration is a necessary condition, especially in criminal justice risk assessments [Berk et al., 2018, Dieterich et al., 2016]. If the risk tool is not calibrated with respect to different groups, then the same risk estimate carries different meanings and cannot be interpreted equally for different groups. Furthermore, creating parity in the error rates of different groups (“equalized odds”) is a well-established

method to mitigate algorithmic discrimination in automatic classification. Previous work has also emphasized the importance of this algorithmic fairness metric for this particular application [Hardt et al., 2016, Zafar et al., 2017, Woodworth et al., 2017]. Hence, to mitigate potential algorithmic discrimination, a relaxation method [Pleiss et al., 2017] is used in this paper which seeks to satisfy equalized odds or parity in the error rates (generalized false positive rate and generalized false negative rate) while preserving calibration in each sub-group of nationality and age. In most cases, calibration and equalized odds are mutually incompatible goals [Chouldechova, 2017, Kleinberg et al., 2016], so in this method it is sought to minimize only a single error disparity across groups while maintaining calibration probability estimates.

Generalized False Positive Rate (GFPR) and Generalized False Negative Rate (GFNR) are the standard notions of false-positive and false-negative rates that are generalized for use with probabilistic classifiers [Pleiss et al., 2017]. If variable x represent an inmate's features vector, y indicates whether or not the inmate recidivists, G_1, G_2 are the two different groups, and h_1, h_2 are binary classifiers which classify samples from G_1, G_2 respectively, GFPR and GFNR are defined as follows [Pleiss et al., 2017]: the GFPR of classifier h_t for group G_t is $c_{fp}(h_t) = E_{(x,y)\sim G_t}[h_t(x) \mid y = 0]$. GFPR is the average probability of being recidivist that the classifier estimates for people who actually do not recidivate. Conversely, the GFNR of classifier h_t is $c_{fn}(h_t) = E_{(x,y)\sim G_t}[(1 - h_t(x)) \mid y = 1]$. So the two classifiers h_1 and h_2 show probabilistic equalized odds across groups G_1 and G_2 if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$.

Classifier h_t is said to be *well-calibrated* if $\forall p \in [0, 1]$, $\mathrm{P}_{(x,y)\sim G_t}$ $[y = 1 \mid h_t(x) = p] = p$. To prevent the probability scores from carrying group-specific information, both classifiers h_1 and h_2 are calibrated with respect to groups G_1 and G_2 [Berk et al., 2018, Dieterich et al., 2016].

4.5 Results

4.5.1 Effectiveness Evaluation

Among logistic regression (LR), multi-layer perceptron (MLP) and support vector machines (SVM), the best results were obtained using LR for both violent and general recidivism predictions. The final set of features used for the model consists of a sub-set of the 43 risk items of the RisCanvi evaluation selected using a feature selection method (based on a linear model with L1-based penalization to yield sparse coefficients), the average Violent Recidivism (REVI)/General Recidivism (REGE) score (from the first to the last RisCanvi evaluation), gender, nationality, and age at the time of the last evaluation.

Results in terms of AUC-ROC, GFNR, and GFPR are presented and compared with the existing RisCanvi protocol in Table 4.2 for both violent and general recidivism prediction. These results are compared against RisCanvi_score, which is a number resulting from the application of the RisCanvi formula.

In both violent and general recidivism prediction, LR yields better results than RisCanvi in terms of all metrics. However, the results are close to RisCanvi. In general, the LR model is more accurate than RisCanvi, although by a small amount, which is surprising considering that RisCanvi was not computationally optimized for predictive accuracy.

4.5.2 Algorithmic Fairness Evaluation

The results for the analysis of algorithmic fairness in all metrics along nationality (national and foreigner), and age groups (young and old inmates) are shown in Table 4.3 for violent and general recidivism prediction. In the LR_calibrated model, the predictions have been calibrated with respect to each of the two sub-groups in nationality and age.

For violent recidivism, all models show a bias against nationals in terms of GFPR. The difference is less noticeable in RisCanvi. In LR model, we can also observe higher GFPR for young inmates compared to

Table 4.2: Effectiveness of models in violent and general recidivism prediction

Risk	Violent Recidivism			General Recidivism		
Model	AUC	GFNR	GFPR	AUC	GFNR	GFPR
LR	0.76	0.82	0.06	0.73	0.75	0.14
MLP	0.74	0.83	0.07	0.72	0.74	0.14
SVM	0.50	0.92	0.08	0.50	0.85	0.15
RisCanvi_score	0.72	0.87	0.07	0.70	0.79	0.14

old offenders. In general, LR_calibrated and RisCanvi models lead to more algorithmically fair results along both nationality and age in terms of all metrics, except for the metrics in which all the models show discrimination.

The results for general recidivism prediction show higher AUC for nationals compared to foreigners in RisCanvi. In terms of GFPR, the LR and LR_calibrated models show discrimination against national group. In age group, LR and LR_calibrated models show higher GFPR along young compared to old group. In terms of AUC, we can see more discrimination against young inmates in RisCanvi compared to other models. As a result, LR_calibrated model shows better algorithmic fairness properties across nationality and more balanced values can be observed along age group in RisCanvi.

Table 4.3: Effectiveness of models in violent and general recidivism prediction per group

Risk	Violent Recidivism									General Recidivism								
Model	LR			LR_Calibrated			RisCanvi			LR			LR_Calibrated			RisCanvi		
Group/Metrics	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR
National	0.81	0.77	0.07	0.81	0.81	0.06	0.76	0.85	0.08	0.78	0.70	0.15	0.77	0.73	0.13	0.72	0.78	0.14
Foreigner	0.85	0.87	0.05	0.85	0.85	0.04	0.72	0.91	0.05	0.68	0.80	0.11	0.72	0.77	0.11	0.59	0.83	0.13
$\frac{\text{National}}{\text{Foreigner}}$ (Ratio)	(0.95)	(0.88)	(1.64)	(0.95)	(0.95)	(1.50)	(1.05)	(0.93)	(1.44)	(1.14)	(0.87)	(1.30)	(1.07)	(0.95)	(1.20)	(1.22)	(0.94)	(1.08)
Young	0.84	0.78	0.08	0.84	0.83	0.06	0.79	0.86	0.07	0.67	0.74	0.17	0.72	0.75	0.15	0.58	0.82	0.14
Old	0.83	0.78	0.06	0.83	0.83	0.06	0.76	0.85	0.07	0.78	0.71	0.12	0.75	0.74	0.11	0.75	0.78	0.14
$\frac{\text{Young}}{\text{Old}}$ (Ratio)	(1.02)	(1.00)	(1.26)	(1.02)	(1.01)	(1.11)	(1.04)	(1.00)	(1.03)	(0.85)	(1.04)	(1.38)	(0.96)	(1.01)	(1.37)	(0.77)	(1.06)	(1.03)

Table 4.4: Equalized GFPR while preserving calibration in violent and general recidivism prediction

Risk	Violent Recidivism									General Recidivism					
Model	LR-Equalized			LR_Calib-Equalized			RisCanvi-Equalized			LR-Equalized			LR_Calib-Equalized		
Group/Metrics	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR
National	0.81	0.77	0.07	0.81	0.81	0.06	0.76	0.85	0.08	0.78	0.70	0.15	0.67	0.78	0.14
Foreigner	0.64	0.92	0.05	0.61	0.91	0.05	0.62	0.92	0.06	0.61	0.81	0.12	0.53	0.88	0.12
$\frac{\text{National}}{\text{Foreigner}}$ (Ratio)	(1.27)	(0.83)	(1.38)	(1.32)	(0.89)	(1.42)	(1.23)	(0.93)	(1.28)	(1.28)	(0.86)	(1.27)	(1.26)	(0.89)	(1.16)
Young	0.84	0.78	0.08	0.71	0.86	0.06	-	-	-	0.67	0.74	0.17	0.72	0.75	0.15
Old	0.62	0.88	0.07	0.60	0.89	0.06	-	-	-	0.63	0.78	0.14	0.53	0.86	0.13
$\frac{\text{Young}}{\text{Old}}$ (Ratio)	(1.36)	(0.89)	(1.17)	(1.19)	(0.97)	(1.05)	-	-	-	(1.06)	(0.95)	(1.22)	(1.35)	(0.88)	(1.18)

4.6 Equalized odds and calibration

In this section, it is tried to achieve parity along nationality and age groups in terms of two fairness metrics simultaneously. For this purpose, the method introduced by Pleiss et al. [Pleiss et al., 2017] is used that seeks parity in Generalized False Positive Rate (GFPR) or Generalized False Negative Rate (GFNR) while preserving calibration in each sub-group of nationality and age. The conclusion from the previous section based on the results obtained per group in Table 4.3, is that in both violent and general recidivism predictions, machine learning models show inequality in terms of GFPR along nationality and age. RisCanvi also shows an imbalance in GFPR values along nationality groups in violent recidivism prediction.

Hence, it is tried to create parity in this metric while preserving calibration in each group. The results after bias mitigation is presented in Table 4.4 for violent and general recidivism prediction. The obtained models are referred to in the following as LR-Equalized, LR_Calibrated-Equalized, and RisCanvi-Equalized.

By comparing the results before and after this bias mitigation (Table 4.3 and Table 4.4 respectively) in violent recidivism, it can be seen that the discrimination in GFPR has decreased in the order of 0.08-0.26 and 0.06-0.09 along nationality and age groups respectively. Also, comparing the results before and after bias mitigation in general recidivism shows that there are reductions in GFPR disparity in the orders of 0.03-0.04 and 0.16-0.19 along nationality and age groups respectively. However, in both violent and general recidivism prediction, the decline in GFPR bias is obtained at the expense of further inequalities in other metrics.

4.7 Discussion and Conclusions

The effectiveness and fairness of Machine Learning (ML) models in violent and general recidivism prediction were compared to the RisCanvi risk assessment tool, an in-use model created by experts. ML models were generated with AUC of 0.76 and 0.73 in violent and general recidivism prediction respectively which shows slightly better results compared to the

AUC of RisCanvi protocol which is 0.72 and 0.70. It is noteworthy that in this type of task, predictions are not very accurate in general (existing recidivism prediction tools typically have AUC in the range of 0.57-0.74 [Brennan et al., 2009, DeMichele et al., 2018, Desmarais et al., 2016]), and it is found that a hand-crafted formula created by experts is quite comparable to a machine-learned one. Although the improvement in accuracy by ML would be insufficient on its own to support its introduction as a risk assessment tool, a key element of ML models is their flexibility. An ML model can be re-trained with newer data, and incorporate new factors as the population of inmates changes and more data on recidivism becomes available.

By studying differential treatment of RisCanvi and ML models across different groups, it can be stated that depending on the desired metric and groups, machine learning and human expert can lead to different but comparable results. An advantage of ML models is that the emphasis on different metrics can be changed during the modeling as legal or policy changes are introduced. In this study, results in Table 4.3 showed that in both violent and general recidivism predictions, there is an inequality in terms of Generalized False Positive Rate (GFPR) metric along nationality and age groups. So using a relaxation method [Pleiss et al., 2017], it was tried to set parity in GFPR while preserving calibration in each sub-group of nationality and age. The results after bias mitigation (in Table 4.4) showed that GFPR disparity in violent and general recidivism has been respectively decreased at most 0.26 and 0.04 along nationality and 0.09 and 0.19 along age, however, in exchange for inequalities in some other metrics.

A robust conclusion from this work is that in a context in which predictive factors neither determine nor yield a clear signal of low/medium/high recidivism risk, ML cannot be considered a silver bullet. At the very least, improvements in accuracy need to be carefully contrasted with potential issues of algorithmic fairness when introducing ML, and calibration and some bias mitigation method (such as equalized odds in this study) needs to be used.

Chapter 5

PREDICTING EARLY DROPOUT: CALIBRATION AND ALGORITHMIC FAIRNESS CONSIDERATIONS

5.1 Introduction

About 36% of university students in the European Union, 39% in the US, 20% in the Australia and New Zealand, and 52% in Brazil discontinue their studies before graduation [Vossensteyn et al., 2015, Shapiro et al., 2017, OECD, 2012]. Reducing the rate of dropout and underperformance is crucial as these lead to social and financial losses. In addition, detecting students at risk as early as possible is necessary to improve learning and prevent them from quitting and failing their studies.

Research on actionable indicators that can lead to interventions to reduce dropout has received increased attention in the last decade, especially in the Learning Analytics (LA) field [Siemens, 2013, Viberg et al., 2018, Sclater et al., 2016, Leitner et al., 2017]. These indicators can help provide effective prevention strategies and personalized intervention actions [Romero and Ventura, 2019, Larrabee Sønderlund et al.,

2019]. Machine Learning (ML) methods, which identify patterns and associations between input variables and the predicted target [Pal, 2012], have been shown to be effective at this predictive task in many LA studies [Plagge, 2013, Kemper et al., 2020, Aulck et al., 2016, Nagy and Molontay, 2018, Del Bonifro et al., 2020].

We remark that among students who discontinue their studies, some sub-groups are over-represented, something that needs to be considered when developing ML methods. For example, in the UK elder students at point of entry (over 21 years) are more likely to drop out after the first year compared to younger students who enter university directly from high school [Larrabee Sønderlund et al., 2019]. In the US, graduation rate among ethnic minority university students is lower than among White students [Shapiro et al., 2017]. Disparities in risks have been studied in previous work [Gardner et al., 2019, Hutt et al., 2019, Kizilcec and Lee, 2020] and are addressed in our work by performing per-group analysis of dropout risk and algorithmic bias mitigation of the risk predictions across different groups.

Our contribution. We observe a high dropout rate (43%) among computer engineering undergraduate students at our university in comparison to the average EU university students' dropout rate (36%) [Vossensteyn et al., 2015]. In this work, we predict the risk of university dropout and underperformance in this engineering school. Calibrated ML models, having outputs that can be directly interpreted as probabilities for dropout or underperformance, are created using student's features available at the time of enrolment (before students start their studies). It is notable that dropout can also be due to the lack of some qualitative variables in the engineering field, such as motivation or vocational changes [Salas-Morera et al., 2019] in addition to the institutional rules. We evaluate our models for accuracy and fairness, as model learning may lead to unfairness for some sensitive groups [Corbett-Davies and Goel, 2018, Chouldechova and Roth, 2018, Barocas et al., 2017, Mehrabi et al., 2021, Zou and Schiebinger, 2018]. Some of the disparities found are addressed through a mitigation procedure [Pleiss et al., 2017], which seeks to equalize error rates (generalized false positive rate or generalized false negative rate) across groups

while preserving the calibration in each group.

The rest of this paper is organized as follows. Section 5.2 outlines related work. In Section 5.3, the dataset used in this study is described. The methodology including the ML models and algorithmic fairness analysis are presented in Section 5.4. Results are given in Section 5.5, and a procedure to mitigate algorithmic discrimination is used in Section 5.6. Finally, conclusions and recommendations are presented in Section 5.7.

5.2 Related Work

Machine Learning (ML) methods have been used to predict dropout in higher education. In a paper [Aulck et al., 2016], the impact of ML on undergraduate student retention is investigated by predicting students dropout (defined as not completing at least one undergraduate degree within 6 calendar years of first enrollment). Using students' demographics and academic transcripts, different ML models result in AUCs between 0.66 and 0.73. In another study [Nagy and Molontay, 2018], an early university dropout is predicted based on available data at the time of enrollment (personal data and secondary school performance) using several ML models with AUCs from 0.62 to 0.81. Similarly, in a recent study [Del Bonifro et al., 2020], several ML methods are used to predict the dropout of first-year undergraduate students before the student starts the course or during the first year.

Several studies [Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Barocas et al., 2017, Mehrabi et al., 2021, Zou and Schiebinger, 2018], have shown that ML models may lead to discriminatory outcomes for some sensitive groups. There are many different definitions of algorithmic fairness [Narayanan, 21], some of which are incompatible with one another. It is impossible to satisfy all of them simultaneously except in pathological cases (such as a perfect classifier), and in general it is impossible to maximize algorithmic fairness and accuracy at the same time [Berk, 2019]. Hence, there are necessary trade-offs between different metrics [Kleinberg et al., 2016]. Some studies [Hardt et al., 2016, Zafar et al.,

2017, Woodworth et al., 2017] try to mitigate potential algorithmic discrimination by introducing a penalization term for unfairness in an objective function to be optimized. Also, several studies [Zemel et al., 2013, Kamiran and Calders, 2009, Kamishima et al., 2011] tried to approach statistical parity in which the same probability of receiving a positive-class prediction is considered for different groups.

One of the closest studies to ours [Gardner et al., 2019], considers algorithmic fairness of predictive models of students dropout in MOOCs in terms of accuracy equity using the Absolute Between-ROC Area (ABROCA) metric. The method to improve algorithmic fairness is slicing analysis, which is also used in another study [Hutt et al., 2019] to analyze fairness across sociodemographic groups in a predictive ML modeling of on-time college graduation. In comparison, in this study we create calibrated ML models that can predict dropout and underperformance risks solely from information available at the time of enrollment, and that have passed through a bias mitigation procedure to avoid error disparities while keeping calibration.

Calibration means that the output of the classifier is not merely a score, but an estimate of the probability of the (adverse) outcome. When we talk about fairness across two groups, we would like this calibration condition to hold for the cases within each of these groups as well. Due to the importance of calibration in risk assessment tools [Berk et al., 2021, Dieterich et al., 2016], some previous work has tried to minimize error disparity across groups while maintaining calibration [Pleiss et al., 2017]. In their work, which is closely related to ours but for a different domain, algorithmic bias in a machine learned risk assessment task (criminal recidivism) is minimized by equalizing generalized false positive rates along different racial backgrounds, finding this equalization to be incompatible with calibration. In contrast, in the work presented on this paper, we try to minimize bias in dropout predictive ML models by equalizing error rates (generalized false positive rate or generalized false negative rate) along some sensitive groups while preserving calibration in each group. Finally, we find that equalization along some groups is not entirely incompatible with calibration.

5.3 Dataset

The anonymized dataset used in this research have been provided by Universitat Pompeu Fabra and consists of 881 computer engineering undergraduate students who first enrolled between 2009 and 2017. From this population, 31 cases who did not enroll for the first trimester, 33 students without admission grade, and 150 students without university grade information (students who first enrolled in 2015 are in this group) were removed and finally 667 cases were remained. Two outcome categories are defined; one is dropout and consists of students who enroll in the first year but do not show up in the second year, the other one is underperformance and consists of students who fail 4 or more of the 12 subjects offered in the first year. Out of 667 cases, 286 students drop out and an additional 62 students underperform.

5.3.1 Per-Group Analysis

The average (base) risk rates of different groups are shown on Table 5.1. Foreign students have more risks compared to nationals, and the risk of students with lower admission grades is higher than the risk of students with higher admission grades. Naturally, students who fail more subjects and/or who have to take re-sit exams exhibit more risk than their counterparts. There have been two study programs for the total of 60 credits in the first year; plan A (older) with 10 courses and plan B (newer) including 12 courses. In the newer plan, with the aim of improving learning process, there is a course reorganization so that students can experience their first programming course in the first trimester and as can be seen, this change caused lower dropout and dropout/underperformance compared to the older plan.

[1]For the purposes of this work, these are students who were born and are resident in the country.

[2]Number of credits passed over total credits during the first year

Table 5.1: Per-group risk rates. Groups with 10 percentage points or more of risk compared to their counterparts are marked with an asterisk (*).

Group	Size	Risk of Dropout	Risk of Dropout or Underperformance
Female	9%	41%	54%
Male	91%	43%	52%
Nationals[1]	88%	41%	50%
Foreigners	12%	58% *	69% *
Age $\leq$ 19 (median age)	55%	44%	55%
Age $>$ 19	45%	41%	48%
State high school	76%	44%	53%
Out-of-State High school	24%	41%	49%
Public high school	42%	44%	55%
Non-public high school	58%	42%	50%
Avg. admission grade $\leq$ median	50%	49% *	59% *
Avg. admission grade $>$ median	50%	37%	45%
Exam retake (at least once in first year)	87%	47% *	58% *
No exam retake	13%	13%	13%
Course failure (at least once in first year)	85%	47% *	58% *
No course failure	15%	17%	17%
Plan A (older)	74%	46% *	53%
Plan B (newer)	26%	35%	50%
Passed credits ratio [2] $\leq$ median	50%	70% *	84% *
Passed credits ratio $>$ median	50%	16%	21%

5.4 Methodology

We consider two predictive tasks: predicting dropout and predicting dropout or underperformance.

5.4.1 ML-based Models

According to the two ground truths (dropout, and dropout or underperformance), separate ML models are created. The feature set for the models consists of demographics (gender, age, and nationality), high school type and location, and average admission grade. Different ML algorithms: logistic regression, multi-layer perceptron (MLP), and support vector machines (SVM) are used to predict dropout risks. ML models are trained using cases enrolled between 2009 to 2013 (409 cases), then tested on students enrolled in 2014, 2016 and 2017 (258 cases). To mitigate the gender imbalance (only 9% of students are women), we use the SMOTE[1] algorithm [Chawla et al., 2002]. We only apply SMOTE on the training set and keep the original class distributions in the test set to ensure valid results.

5.4.2 Algorithmic Fairness

Parity in the error rates of different groups ("equalized odds") is a well-established method to mitigate algorithmic discrimination in automatic classification [Hardt et al., 2016, Zafar et al., 2017, Woodworth et al., 2017]. At the same time, we want to maintain model calibration [Dieterich et al., 2016, Berk et al., 2021], as otherwise the same risk estimate carries different meanings and cannot be interpreted equally for different groups. Hence, a relaxation method [Pleiss et al., 2017] is used in this paper which seeks to satisfy equalized odds or parity in the error rates while preserving calibration. In most cases, calibration and equalized odds are mutually incompatible goals [Chouldechova, 2017, Kleinberg et al., 2016], so in

[1] Synthetic Minority Oversampling Technique

this method it is sought to minimize only a single error disparity across groups while maintaining calibration probability estimates.

If variable x represents a student's features vector, y indicates whether or not the student drops out, G_1, G_2 are the two different groups, and h_1, h_2 are binary classifiers which classify samples from G_1, G_2 respectively, Generalized False Positive Rate (GFPR) and Generalized False Negative Rate (GFNR) are defined as follows [Pleiss et al., 2017]: the GFPR of classifier h_t for group G_t is $c_{fp}(h_t) = E_{(x,y)\sim G_t}[h_t(x) \mid y = 0]$. This is the average probability of dropout that the classifier estimates for students who do not drop out. Conversely, the GFNR of classifier h_t is $c_{fn}(h_t) = E_{(x,y)\sim G_t}[(1 - h_t(x)) \mid y = 1]$. So the two classifiers h_1 and h_2 show probabilistic equalized odds across groups G_1 and G_2 if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$. Classifier h_t is said to be *well-calibrated* if $\forall p \in [0, 1]$, $\mathrm{P}_{(x,y)\sim G_t}[y = 1 \mid h_t(x) = p] = p$. To prevent the probability scores from carrying group-specific information, both classifiers h_1 and h_2 are also calibrated with respect to groups G_1 and G_2 [Berk et al., 2021, Dieterich et al., 2016].

5.5 Results

5.5.1 Effectiveness Evaluation

Predictive performance in terms of the AUC-ROC, GFNR, GFPR, and F-score (the harmonic mean of precision and recall, which unlike the other metrics, requires to establish an optimal cut-off for classification) are presented in Table 5.2 for MLP (Multi-Layer Perceptron) and LR (Logistic Regression) models. The best results for both dropout risk predictions were obtained using an MLP. We used a single hidden layer having 100 neurons. According to the results, the models lead to good performance in terms of AUC and F-score in both prediction tasks. With a little information at the time of students' enrollment, these models show good AUC in comparison to previous work [Aulck et al., 2016, Nagy and Molontay, 2018] which showed AUC in the order of 0.62-0.81. Also, comparing calibrated and non-calibrated predictions we can see that calibrated model leads to lower

Table 5.2: Effectiveness of models in risk prediction. ”cal”:calibrated.

Risk	Dropout				Dropout or Underperformance			
Model	AUC	GFNR	GFPR	F-score	AUC	GFNR	GFPR	F-score
MLP	0.77	0.73	0.19	0.76	0.78	0.69	0.19	0.83
MLP cal.	0.77	0.36	0.42	0.76	0.78	0.27	0.49	0.83
LR	0.71	0.68	0.25	0.75	0.77	0.63	0.23	0.82
LR cal.	0.71	0.36	0.46	0.71	0.77	0.27	0.50	0.83

GFNR and non-calibrated results in lower GFPR.

5.5.2 Algorithmic Fairness Evaluation

The results for the analysis of algorithmic fairness are shown on the left side of Table 5.3. In dropout prediction, we can observe accuracy equity (less than 20% discrepancy) in terms of AUC in both models, even if results are slightly more accurate for male students. AUC is also higher for students with lower admission grades compared to their counterparts. In the calibrated model, males, foreigners, and lower admission grade students experience lower GFNR compared to their counterparts. However, non-calibrated model shows fairer results for GFNR along these groups. Regarding GFPR, there can be seen more false positive errors (higher risk scores for students who do not dropout or underperform) for males compared to females, students of out-of-State high schools than in-State high schools, and lower admission grade students compared to their counterparts in the non-calibrated model. In the calibrated model, this metric shows more errors for foreigners and for lower admission grade students compared to their counterparts.

Similar results are shown for predicting dropout or underperformance. In terms of AUC, MLP shows equity (less than 20% discrepancy) across groups except for more accuracy for students from in-State high schools. In

the calibrated model, higher AUC can be observed in nationals compared to foreigners and higher admission grade students. Also, both models show parity across all groups in terms of GFNR except for students with lower admission grade who experience lower errors compared to their counterparts, however, non-calibrated model shows lower discrimination to this groups compared to the calibrated one. In terms of GFPR, we can see more errors of the model for foreigners than nationals, out-of-State high school than in-State high school students, males than females, and cases with lower admission grades compared to their counterparts. In the calibrated model, this metric also shows more error for foreigners than national and students with lower admission grade compared to their counterparts, but it reveals more errors for females than males.

5.6 Equalized odds and calibration

In this section, parity is sought along groups in terms of two fairness metrics. For this purpose, the method introduced by Pleiss et al. [Pleiss et al., 2017] is used, which seeks parity in Generalized False Positive Rate (GFPR) or Generalized False Negative Rate (GFNR) while preserving calibration. In both prediction tasks, the models before mitigation exhibit in general better parity in terms of AUC and GFNR and more inequality in terms of GFPR. The results after bias mitigation are presented in the right side of the Table 5.3. By comparing the results before and after GFPR bias mitigation in dropout we can see that the disparity in GFPR has decreased in the order of 0.03-0.71 in MLP and 0.02-0.30 in MLP calibrated across all groups. Also, comparing the result before and after GFPR bias mitigation in dropout or underperformance show that bias in MLP and MLP calibrated models has been respectively reduced by the order of 0.08-1.15 and 0.14-0.59 across all groups.

Table 5.3: Effectiveness (AUC) and fairness (GFPR and GFNR ratios) of models for the two risk prediction tasks, before and after bias mitigation. Values in boldface should, ideally, be close to 1.0 to indicate perfect equity among groups.

	Before bias mitigation												After bias mitigation											
Risk	Dropout						Dropout or Underperformance						Dropout						Dropout or Underperformance					
Model	MLP			MLP calibrated			MLP			MLP calibrated			Equalized GFPR MLP			Equalized GFPR MLP calibrated			Equalized GFPR MLP			Equalized GFPR MLP calibrated		
Group / Metric	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR
Nationals	0.77	0.74	0.19	0.76	0.34	0.42	0.77	0.72	0.19	0.81	0.23	0.45	0.77	0.74	0.19	0.50	0.46	0.54	0.64	0.68	0.28	0.50	0.35	0.65
Foreigners	0.82	0.73	0.17	0.69	0.26	0.61	0.67	0.66	0.25	0.55	0.21	0.67	0.82	0.72	0.17	0.69	0.26	0.61	0.67	0.66	0.25	0.55	0.21	0.67
$\frac{\text{Nationals}}{\text{Foreigners}}$ (Ratio)	**0.93**	**1.01**	**1.13**	**1.10**	**1.33**	**0.68**	**1.15**	**1.08**	**0.76**	**1.45**	**1.07**	**0.67**	**0.93**	**1.03**	**1.13**	**0.73**	**1.79**	**0.88**	**0.95**	**1.03**	**1.14**	**0.90**	**1.63**	**0.98**
State_Highschool	0.79	0.74	0.18	0.74	0.35	0.46	0.79	0.72	0.17	0.78	0.22	0.47	0.65	0.71	0.25	0.74	0.35	0.46	0.62	0.62	0.32	0.72	0.27	0.54
NonState_Highschool	0.71	0.70	0.23	0.77	0.31	0.43	0.59	0.67	0.30	0.74	0.25	0.54	0.71	0.70	0.23	0.78	0.30	0.43	0.59	0.67	0.30	0.74	0.25	0.54
$\frac{\text{State\_Highschool}}{\text{NonState\_Highschool}}$ (Ratio)	**1.11**	**1.06**	**0.77**	**0.96**	**1.12**	**1.07**	**1.35**	**1.08**	**0.57**	**1.05**	**0.88**	**0.87**	**0.92**	**1.02**	**1.07**	**0.94**	**1.13**	**1.09**	**1.06**	**0.93**	**1.08**	**0.97**	**1.06**	**0.99**
Pub_Highschool	0.82	0.74	0.17	0.77	0.36	0.39	0.79	0.71	0.18	0.78	0.22	0.44	0.80	0.71	0.20	0.63	0.41	0.50	0.81	0.68	0.18	0.75	0.24	0.51
NonPub_Highschool	0.72	0.73	0.21	0.72	0.31	0.47	0.72	0.70	0.21	0.75	0.20	0.49	0.72	0.73	0.21	0.72	0.31	0.47	0.72	0.70	0.21	0.75	0.20	0.49
$\frac{\text{Pub\_Highschool}}{\text{NonPub\_Highschool}}$ (Ratio)	**1.14**	**1.01**	**0.85**	**1.07**	**1.19**	**0.82**	**1.09**	**1.01**	**0.88**	**1.03**	**1.07**	**0.91**	**1.10**	**0.96**	**0.96**	**0.88**	**1.33**	**1.05**	**1.11**	**0.97**	**0.88**	**0.99**	**1.18**	**1.05**
Low_AdmissionGrade	0.69	0.70	0.27	0.67	0.24	0.58	0.66	0.65	0.33	0.55	0.15	0.77	0.69	0.70	0.27	0.67	0.24	0.58	0.66	0.65	0.33	0.55	0.15	0.77
High_AdmissionGrade	0.52	0.84	0.16	0.71	0.50	0.37	0.64	0.82	0.15	0.74	0.41	0.40	0.52	0.73	0.27	0.50	0.57	0.43	0.49	0.67	0.34	0.50	0.51	0.49
$\frac{\text{Low\_AdmissionGrade}}{\text{High\_AdmissionGrade}}$ (Ratio)	**1.32**	**0.83**	**1.68**	**0.95**	**0.49**	**1.59**	**1.03**	**0.79**	**2.16**	**0.74**	**0.37**	**1.95**	**1.32**	**0.96**	**1.01**	**1.35**	**0.42**	**1.36**	**1.33**	**0.97**	**0.97**	**1.10**	**0.30**	**1.57**
Male	0.78	0.73	0.20	0.75	0.33	0.45	0.77	0.70	0.20	0.78	0.23	0.47	0.78	0.73	0.20	0.75	0.33	0.45	0.77	0.70	0.20	0.51	0.34	0.65
Female	0.62	0.84	0.13	0.57	0.51	0.44	0.67	0.80	0.15	0.71	0.23	0.64	0.62	0.76	0.20	0.67	0.55	0.45	0.56	0.80	0.22	0.71	0.23	0.64
$\frac{\text{Male}}{\text{Female}}$ (Ratio)	**1.26**	**0.87**	**1.51**	**1.33**	**0.66**	**1.02**	**1.15**	**0.87**	**1.37**	**1.10**	**0.99**	**0.73**	**1.26**	**0.96**	**1.00**	**1.13**	**0.61**	**1.01**	**1.38**	**0.87**	**0.91**	**0.72**	**1.45**	**1.02**

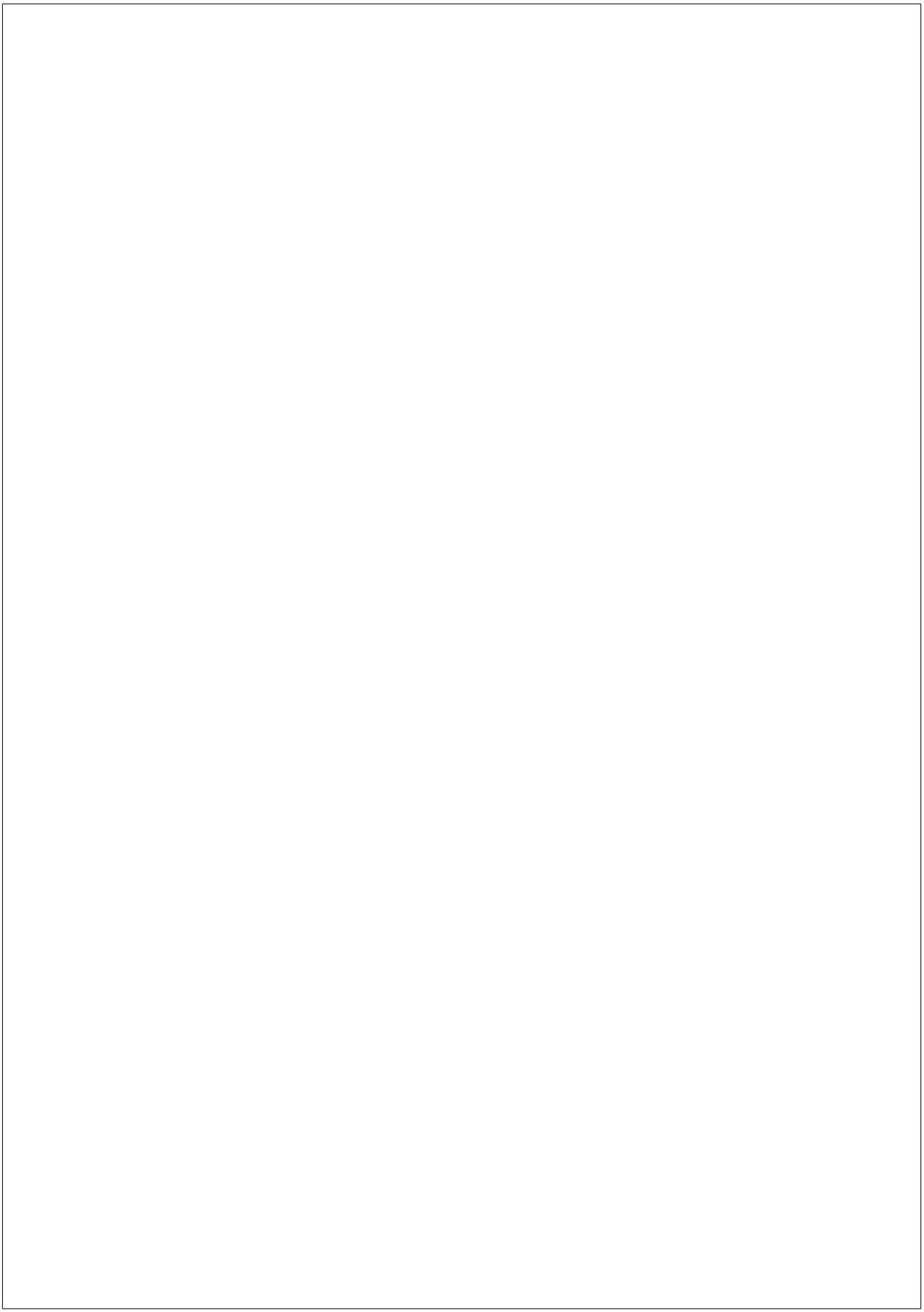
5.7 Conclusions and Recommendations

The effectiveness and fairness of Machine Learning (ML) models in the early prediction of university dropout and underperformance was evaluated. Using only information at the time of enrollment, calibrated ML models were created with AUC of 0.77 and 0.78 which can help reliably identify students at risk to trigger interventions that can help increase their success and ultimately reduce social and economic costs.

When introducing ML models, improvements in accuracy need to be carefully contrasted with potential algorithmic discrimination. Thus, we evaluated the algorithmic fairness of the ML models in terms of AUC and error (GFNR and GFPR) across five groups defined by nationality, gender, high school type and location, and admission grade. According to the results, our modeling has parity in terms of AUC and GFNR but disparities in GFPR. These disparities in GFPR are larger among groups defined by admission grade, and the bias is against students with lower admission grades. The predicted probability of dropout for the students of this sub group who do not actually drop out is larger than that of their counterparts (students of higher admission grade sub group). Using a relaxation method [Pleiss et al., 2017], we tried to obtain parity in GFPR while preserving calibration. By maintaining the calibration among subgroups, we prevent the probability scores from needing group-dependent interpretation. The results after bias mitigation show that GFPR ratio in both dropout and dropout or underperformance predictions has been changed to a perfect value close to 1 across most of the groups. This bias mitigation also caused better parities in other metrics (AUC and GFNR) along majority of the groups compared to the non-mitigated model. Studying algorithmic discrimination means addressing unfair decisions not only to the identification of students that would require preventive mentoring programs, but also to the identification of potentially successful students that would benefit from e.g. additional educational opportunities or to the formulation of pedagogical interventions related to changes in the study plans or in pedagogical methods suiting specific students' profiles.

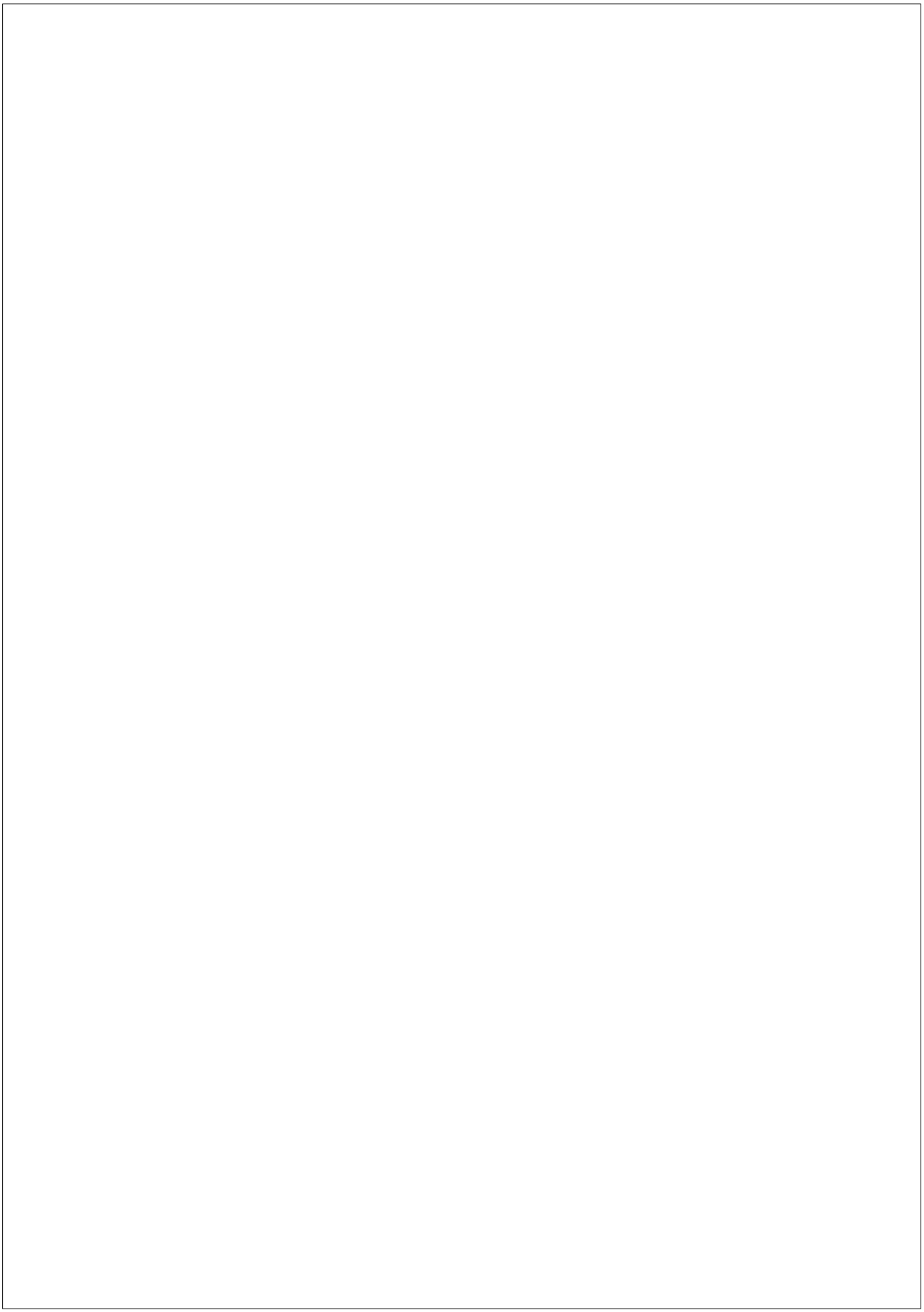
In terms of contributions to learning analytics, in addition to creating

ML models for dropout and underperformance that exhibit high accuracy, we evaluated algorithmic fairness of the models across different groups in terms of several metrics and applied a bias mitigation method to set parity for subgroups with unfair results. For the students at high risk of dropout or underperformance, different interventions can be considered such as tutoring, counselling and mentoring. A suggested beneficial intervention [Lowis and Castley, 2008] is interviewing with the students in informal discussion and asking for their perceptions and experiences at the university which can help with the planning process for their subsequent academic years. Also, a preventive mentoring program [Larose et al., 2011] showed high levels of motivation and more positive career decision profiles for the newcomer students who participated in bimonthly meetings with students completing their undergraduate degree. Both require early prediction models with equity among groups, which the methods we have described can provide in a real-world setting.



Part III

Causal Inference



Chapter 6

A CAUSAL INFERENCE STUDY ON THE EFFECTS OF FIRST YEAR WORKLOAD ON THE DROPOUT RATE OF UNDERGRADUATES

6.1 Introduction

Research on actionable indicators that can lead to interventions to reduce dropout has received increased attention in the last decade, especially in the Learning Analytics (LA) field [Syed et al., 2019, Viberg et al., 2018, Siemens, 2013, Sclater et al., 2016, Leitner et al., 2017]. These indicators can help provide effective prevention strategies and personalized intervention actions [Romero and Ventura, 2019, Larrabee Sønderlund et al., 2019]. Machine Learning (ML) methods, which identify patterns and associations between input variables and the predicted target [Pal, 2012], have been shown to be effective at this predictive task in many LA studies [Plagge, 2013, Kemper et al., 2020, Aulck et al., 2016, Nagy and Molontay, 2018, Del Bonifro et al., 2020, Albreiki et al., 2021].

Dropout is a serious problem especially in higher education, leading to social and financial losses impacting students, institutions, and society [Bukralia et al., 2015]. In particular, the early identification of vulnerable students who are prone to fail or drop their courses is necessary to improve learning and prevent them from quitting and failing their studies [Márquez-Vera et al., 2016].

We remark that among students who discontinue their studies, some sub-groups are over-represented, something that needs to be considered when designing dropout-reduction interventions. For example, in the UK, older students at point of entry (over 21 years) are more likely to drop out after the first year compared to younger students who enter university directly from high school [Larrabee Sønderlund et al., 2019], something that we also observe in our data. In the US, graduation rates among ethnic minority university students are lower than among White students [Shapiro et al., 2017]. Disparities in dropout risk have been studied in previous work [Gardner et al., 2019, Hutt et al., 2019, Kizilcec and Lee, 2020, Karimi-Haghighi et al., 2021]. Recent studies [Modena et al., 2020, Choi, 2018, Olaya et al., 2020, Masserini and Bini, 2021] look at the influence on student's performance and dropout of factors such as having a scholarship or being employed. In our work, we consider the increased dropout risk of older students and of students who do not enter university immediately after high school, and we study the effects of some features such as age and workload (i.e., number of credits taken on the first year).

Research contribution. In this work, we use causal inference methods to study the effects of several features on the risk of early dropout in undergraduates students. We consider students enrolled between 2009 and 2018 in eight centers at our university. The average dropout rate we observe among these students is 15.3%, which is lower than the European average (36%) [Vossensteyn et al., 2015]. The originality of our contribution relies on its focus on students who have higher risk, the combination of features, the use of causal inference methods, and the size and scope of our dataset.

Specifically, we predict the risks of early dropout (i.e., not enrolling

on the second year) and underperformance (failing to pass two or more subjects in the first year in the regular exams[1]) using Machine Learning (ML) methods. ML models are created using features available at the time of enrolment and the predictive performance of the models is evaluated in terms of AUC-ROC (Area Under ROC Curve). For the sake of space, we focus our exposition on dropout.

Among features available at the time of enrolment, we obtain the most important features for predicting dropout in our setting, which are the workload (number of credits taken) in the first year, admission grade, age, and study access type. Focusing on the workload, which is the most important feature and one over which first-year students have some level of control (only a minimum number of credits is established), we compute its effect on dropout risk in different age and study access type groups. We use causal inference methods to test the effects of combinations of theses features, and calculate the average treatment effect on dropout; the methods we use are the most used in the literature [Athey, 2015, Athey and Wager, 2019] including Propensity Score Matching (PSM) [Rosenbaum and Rubin, 1983], Inverse-Propensity score Weighting (IPW) [Bray et al., 2019], Augmented Inverse-Propensity Weighted (AIPW) [Glynn and Quinn, 2010], and Doubly Robust Orthogonal Forest Estimation (DROrthoForest) [Battocchi et al., 2019] methods.

The rest of this paper is organized as follows. After outlining related work on Section 6.2, the dataset used in this study is described and analysed in Section 6.3. The methodology is presented in Section 6.4. Results are given in Section 6.5, and finally, the results are discussed and the paper is concluded in Section 6.6.

[1]These students have an opportunity of taking a resit exam which may finally result in passing or failing the subject, but given that passing the regular exam at the end of the course is expected, we consider failing the regular exam as underperforming.

6.2 Related work

Machine Learning (ML) methods have been used to predict dropout and detect students at risk in higher education and play essential roles in improving the students' performance [Albreiki et al., 2021]. In a reference [Aulck et al., 2016], the impact of ML on undergraduate student retention is investigated by predicting students dropout. Using students' demographics and academic transcripts, different ML models result in AUCs between 0.66 and 0.73. Another reference [Bukralia et al., 2015] develops a model to predict real-time dropout risk for each student during an online course using a combination of variables from the Student Information Systems and Course Management System. Evaluating the predictive accuracy and performance of various data mining techniques, the study results show that the boosted C5.0 decision tree model achieves 90.97% overall predictive accuracy in predicting student dropout in online courses. In a study [Nagy and Molontay, 2018], early university dropout is predicted based on available data at the time of enrollment using several ML models with AUCs from 0.62 to 0.81. Similarly, in a recent study [Del Bonifro et al., 2020], several ML methods are used to predict the dropout of first-year undergraduate students before the student starts the course or during the first year.

Some studies look at the features driving dropout. A reference [Chounta et al.,] identifies factors contributing to dropout and estimates the risk of dropout for a group of students. By presenting the computed risk and explaining the reasons behind it to academic stakeholders, they help identify more accurately students that may need further support. In a research [Tanvir and Chounta,], the potential relationship between some features (academic background, students' performance and students' effort dimensions) and dropout is investigated over time by performing a correlation analysis on a longitudinal data collected spanning over 11 years. The results show that the importance of features related to the academic background of students and the effort students make may change over time. On the contrary, performance measures are stable predictors of dropout over time. Influential factors on student success are identified in a reference

[Lemmerich et al., 2010] using subgroup discovery; this uncovers important combinations of features known before students start their degree program, such as age, sex, regional origin or previous activities.

Recent work uses sophisticated statistical methods including causal inference. In a very recent paper [Masserini and Bini, 2021], using propensity score matching (PSM) it is investigated whether university dropout in the first year is affected by participation in Facebook groups created by students. The estimated effect indicates that participation in social media groups reduces dropout rate. Another recent paper [Olaya et al., 2020], implements an uplift modeling framework to maximize the effectiveness of retention efforts in higher education institutions, i.e., improvement of academic performance by offering tutorials. Uplift modeling is an approach for estimating the incremental effect of an action or treatment on an outcome of interest at the individual level (individual treatment effect). They show promising results in tailoring retention efforts in higher education over conventional predictive modeling approaches. In a study, the effect of grants on university dropout rates is studied [Modena et al., 2020]. The average treatment effect is estimated using blocking on the propensity score with regression adjustment. According to their results, grants have a relevant impact on the probability of completing college education.

In our paper, we carefully measure the effect of the most important features (the number of credits in the first year, age, and study access type) on the early risk of dropout in undergraduate studies. This effect is obtained for combinations of these features. The Average Treatment Effect (ATE) is measured using multiple causal inference methods [Athey, 2015, Athey and Wager, 2019] as discussed in the introduction. It is noteworthy that according to a recent survey, the methods we use in this paper have not been applied in related studies so far [Albreiki et al., 2021].

6.3 Dataset

The anonymized dataset used in this study has been provided by Universitat Pompeu Fabra and consists of 24,253 undergraduate students who enrolled

between 2009 to 2018 to 21 different study programs offered by eight academic centers. From this population, about 5% of cases were discarded for various reasons: 54 had an external interruption in their education between the first and second study year, 469 students did not have grade records (dropped out before starting), 560 students were admitted but did not enroll for the first trimester, and 74 cases did not have a study access type. Finally, 23,096 cases remained.

Students were admitted to university through four access types: type I students took a standard admission test (81%), type II students moved from incomplete studies in another university or were older than 25 (10%), type III students completed vocational training before (7%), and type IV students completed a different university degree before (2%). First year courses add up to a total of 60 credits across all study programs, this is also the median number of credits taken by first year students. However, students are also free to take additional credits out of different educational offers at the university such as languages, sports, and solidarity action.

The main studied outcome is dropout and consists of students who enroll in the first year but not in the second year. We also studied underperformance, which we defined as failing two or more subjects of the first year in the regular exams. Out of 23,096 cases, 3,531 students drop out (15.3%) and 6,652 students underperform (28.8%). Per-center dropout, underperformance, and other features are shown in Table 6.1.

Table 6.1: Per-center statistics: number of students, drop-out rate, underperformance rate, percentage of national students, percentage of men, average age, average first year credits, average grade on the first year, and percentage of students in access type I.

Center	N	Dropout rate	Underperf. rate	National %	Male %	Avg. age	Avg. credits	Avg. grade	Access type I
ENG	2,444	41%	56%	89%	79%	19.4	63.4	4.6	65%
HUM	1,749	22%	33%	90%	32%	20.3	63.1	5.9	76%
TRA	2,292	16%	28%	88%	18%	19.3	62.9	6.3	83%
POL	1,683	14%	27%	94%	55%	18.8	63.1	6.2	87%
HEA	1,206	14%	16%	93%	25%	19.0	60.2	7.2	82%
LAW	5,479	12%	32%	92%	33%	19.3	62.5	6.0	79%
ECO	5,707	9%	26%	93%	47%	18.5	62.9	6.3	88%
COM	2,536	7%	7%	96%	27%	18.8	61.7	7.5	84%
All	**23,096**	**15%**	**29%**	**92%**	**40%**	**19.1**	**62.6**	**6.2**	**81%**

There are various differences among centers.[1] The students in the School of Engineering and Faculty of Humanities have the highest dropout and underperformance rates and the Faculty of Communication has the lowest dropout rate and the best performance. In the Faculty of Communication, which has the lowest dropout and underperformance rates, there are more national students compared to other schools. In the School of Engineering, with the highest dropout and underperformance rates, males are in the majority. The average age in the two centers with the highest dropout and underperformance rates (School of Engineering and Faculty of Humanities) is higher compared to other faculties. In these two centers, the percentage of students admitted through a standard test (study access type I) is lower than other centers, and we can observe higher average number of credits and lower average grades in their first year compared to others. In the Faculty of Humanities, 22% of the students drop out (that includes 38% of those who underperform), while in the Faculty of Law, with almost the same underperformance rate, only 12% of the students drop out (including 18% of those who underperform). This might be partially explained because in the Faculty of Law, students are one year younger (19.3 vs 20.3 years old on average) and are also slightly more likely to come directly from high school (study access type I: 79% vs 76%).

6.4 Methodology

Our study focuses on modeling dropout and underperformance risks using data available at the time students enrol. The feature set for our two models consists of demographics (gender, age, and nationality), study access type, study program, number of first year credits, and average admission grade. Different ML algorithms: logistic regression (LR), multi-layer perceptron (MLP), and decision trees are used to predict the risks. Both ML models are trained using students enrolled between 2009 to 2015 (16,273 cases),

[1] ENG:Engineering, HUM:Humanities, TRA:Translation and Language Sciences, POL:Political and Social Sciences, HEA:Health and Life Sciences, ECO:Economics and Business, COM:Communication.

and tested on students enrolled in 2016, 2017, and 2018 (6,823 cases). Due to space consideration and because of the severity of dropout, we mainly focus on this risk. Using a feature selection method based on decision trees (CART), we find that among the features available at the time of enrolment, the most important features in predicting dropout risk are the number of credits in the first year (workload), admission grade, age, and study access type.

In Table 6.2, we compare the dropout rate of different student groups in terms of these features and some of their combinations. This comparison shows the following results. Students older than the average age have higher rate of dropout than younger students, across all centers except the Faculty of Health and Life Sciences (HEA). Students admitted through study access types III and IV have a higher dropout rate compared to the cases admitted through access types I and II; students with admission grades less than the average admission grade have higher dropout rate compared to the cases with higher admission grades, and students taking more credits than the median also have higher dropout rate. Considering combinations of these features, we can see that mostly older students with a number of credits larger than the median, students admitted through access types III and IV who take a larger number of credits than the median, as well as student with admission grades less than average who take credits larger than the median have higher dropout rates. Results for underperformance rates are shown for some combinations of the features on Table 6.3. The results are mostly similar to dropout rates, except in two senses: they do not hold for Engineering (ENG) and Humanities (HUM), possibly in part due to the overall lower grades in these centers compared to all others (Table 6.1), as well for the school of Economics and Business (ECO) but the group difference is small (mostly less than 5 percentage points) in this school, and they do not hold for credits alone, but for credits in combination with other features.

Table 6.2: Dropout rate (%) across groups defined by age, workload (number of credits), access type, and admission grade. Differences of ten percentage points or more appear in **boldface**.

Center	ENG	HUM	TRA	POL	HEA	LAW	ECO	COM	**All**
Age $>$ Avg. age	45	28	24	**21**	13	**21**	16	12	**26**
Age $\leq$ Avg. age	39	21	15	**12**	15	**10**	8	6	**13**
Access types III/IV	44	28	**27**	23	16	18	**21**	11	**24**
Access types I/II	40	22	**16**	14	14	11	**9**	6	**14**
Access type I	40	23	15	13	14	10	8	6	14
Other access types	42	20	22	22	14	19	17	11	23
Credits $>$ 60	47	29	22	22	21	19	11	7	18
Credits $\leq$ 60	39	20	15	13	13	10	9	6	14
Admission grade $>$ Avg. adm. grade	**24**	**12**	**9**	**9**	14	9	7	6	**10**
Admission grade $\leq$ Avg. adm. grade	**53**	**30**	**23**	**19**	15	15	12	10	**22**
Age $>$ Avg. age & credits $>$60	**53**	29	**33**	**29**	13	**33**	18	13	**32**
Others	**39**	22	**16**	**13**	14	**11**	9	6	**14**
Acc. types III/IV & credits $>$ 60	**51**	**36**	**61**	**27**	15	**33**	**23**	10	**30**
Others	**40**	**22**	**16**	**14**	14	**11**	**9**	7	**15**
Adm. grade $\leq$ Avg. & credits $>$ 60	**50**	**32**	**26**	**23**	**23**	**23**	13	12	**24**
Others	**38**	**20**	**15**	**13**	**13**	**11**	9	6	**14**

Table 6.3: Underperformance rate (%) across groups defined by age, workload (number of credits), access type, and admission grade. Differences of ten percentage points or more appear in **boldface**.

Center	ENG	HUM	TRA	POL	HEA	LAW	ECO	COM	**All**
Age $>$ Avg. age	**47**	29	**37**	31	**29**	36	26	**16**	35
Age $\leq$ Avg. age	**61**	34	**26**	26	**12**	31	26	**4**	28
Access types III/IV	**48**	30	35	34	**31**	35	22	15	33
Access types I/II	**58**	33	27	27	**14**	31	26	6	28
Access type I	**60**	35	27	27	**12**	31	26	**5**	28
Other access types	**49**	28	33	28	**32**	34	26	**16**	33
Admission grade $>$ Avg. adm. grade	58	**21**	**15**	**16**	15	28	29	**4**	**24**
Admission grade $\leq$ Avg. adm. grade	55	**44**	**39**	**38**	17	36	22	**14**	**35**
Age $\geq$ Avg. age & credits ≥ 60	**38**	30	**38**	33	**34**	33	24	**16**	31
Others	**60**	34	**26**	26	**14**	32	27	**5**	28
Acc. types II/III/IV & credits ≥ 60	**42**	30	34	31	**40**	31	24	**17**	30
Others	**60**	34	27	27	**14**	32	26	**5**	29
Adm. grade $\leq$ Avg. & Age $>$ Avg. age	**48**	29	**39**	**37**	**30**	37	22	**23**	36
Others	**59**	34	**25**	**25**	**14**	31	27	**5**	28

We aim to determine the causal effects on dropout of the features we studied by the following intervention: taking a workload in the first year of less credits than the median. The number of credits taken is a feature over which students have some degree of control at the enrolment time. Since higher dropout rates are observed among older students and students with access types III and IV, we are interested in the following scenarios:

- Scenario 1: in this scenario, the study group is limited to the first-year students who are older than the mean. Among these, those with less workload (credits $<$ median) are considered as treated and those with more workload (credits $\geq$ median) are regarded as a control group.

- Scenario 2: in this scenario, the study group are all students. Older students taking less workload (credits $<$ median) plus all younger students are considered as treated, and older students with more workload (credits $\geq$ median) are regarded as a control group.

- Scenario 3: in this scenario, the study group is limited to students from access types III and IV. Among these, students with less workload (credits $<$ median) are considered as treated and students with more workload (credits $\geq$ median) are regarded as a control group.

The propensity of treatment is estimated in each scenario using Machine Learning (ML) models and input features including demographics (gender and nationality), study programs, and average admission grade. In scenarios 1 and 2, study access type is also added as a feature, and in scenario 3, age is added as a feature. We compute the Average Treatment Effect (ATE) of each treatment on the dropout probability using various causal inference methods:

- The propensity score matching method [Rosenbaum and Rubin, 1983], in which data is sorted by propensity score and then stratified into buckets (five in our case). In our work, we obtain ATE by subtracting the mean dropout of non-treated (control) cases from treated ones in each bucket.

- Inverse-Propensity score Weighting (IPW) [Bray et al., 2019]: The basic idea of this method is weighting the outcome measures by the inverse of the probability of the individual with a given set of features being assigned to the treatment so that similar baseline characteristics are obtained. In this method, the treatment effect for individual i is obtained using the following equation:

$$TE_i = \frac{W_i Y_i}{p_i} - \frac{(1 - W_i)Y_i}{1 - p_i} \tag{6.1}$$

  W_i shows treatment (1 for treated and 0 for control cases), p_i represents probability of receiving treatment (propensity score of treatment), and Y_i shows dropout (1 if drop out and 0 if not drop out) for individual i.

- Augmented Inverse-Propensity Weighted (AIPW) [Glynn and Quinn, 2010]: This method combines both the properties of the regression-based estimator and the IPW estimator. It has an augmentation part $(W_i - p_i)\widehat{Y}_i$ to the IPW method, in which $\widehat{Y}_i$ is the estimated probability of dropout using all features applied to the propensity score model plus the treatment variable. So, this estimator can lead to doubly robust estimation which requires only either the propensity or outcome model to be correctly specified but not both. We can compute the treatment effect on individual i as:

$$TE_i = \frac{W_i Y_i - (W_i - p_i)\widehat{Y}_i}{p_i} - \frac{(1 - W_i)Y_i - (W_i - p_i)\widehat{Y}_i}{1 - p_i} \tag{6.2}$$

- Causal forests from EconML package [Battocchi et al., 2019]: This method uses Doubly Robust Orthogonal Forests (DROrthoForest) which are a combination of causal forests and double machine learning to non-parametrically estimate the treatment effect for each individual.

In IPW, AIPW, and DROrthoForest, we obtain the individual treatment effect TE_i, which is the difference between the outcomes if the person is

treated (treatment) and not treated (control). In other words, this effect is the difference of dropout probability when the student is treated and not treated; a negative value shows a reduced dropout risk and a positive value indicates an increased dropout risk. The resulting ATE is the average over individual treatment effects.

6.5 Results

The ML-based models of dropout and underperformance obtained using an MLP (Multi-Layer Perceptron) with 100 hidden neurons show the best predictive performance, with AUC-ROC of 0.70 and 0.74 for each risk respectively. Table 6.4 shows the AUC-ROC per center, and we observe that the AUC-ROC is in general higher for centers with higher dropout and underperformance rates. We also observe that dropout and underperformance predictions are not reliable for some centers, particularly Health and Life Sciences (HEA), and Law, where the AUC is less than 0.65.

For the three scenarios introduced in section 6.4, the best predictive performance results obtained for the propensity score of the related treatment are shown on Table 6.5 in terms of AUC-ROC. Propensity is better predicted for scenarios 1 and 2 with the Multi-Layer Perceptron (MLP) and for scenario 3 with the Logistic Regression (LR). In each scenario, we removed study programs with relatively low predictive performance. According to the AUC values, ML models show accurate results in all of the scenarios, especially in scenario 2. The distribution of propensity scores in treatment and control groups of each scenario is shown in Figure 6.1, Figure 6.2, and Figure 6.3. In all scenarios, there is an overlap in the distribution of the propensity scores of treatment and control groups to find adequate matches. This is a necessary condition to be able to apply some of our methods.

Table 6.4: AUC-ROC of the prediction of dropout and underperformance across centers. Centers are sorted left-to-right by decreasing dropout rate.

Center	**All**	ENG	HUM	TRA	POL	HEA	LAW	ECO	COM
Dropout	**0.70**	0.72	0.72	0.68	0.67	0.57	0.64	0.67	0.68
Underperformance	**0.74**	0.82	0.80	0.73	0.69	0.53	0.64	0.69	0.76

Table 6.5: AUC-ROC of propensity score prediction.

	Scenario 1	Scenario 2	Scenario 3
N	3,866	23,096	1,963
Model	MLP	MLP	LR
AUC	0.75	0.91	0.75

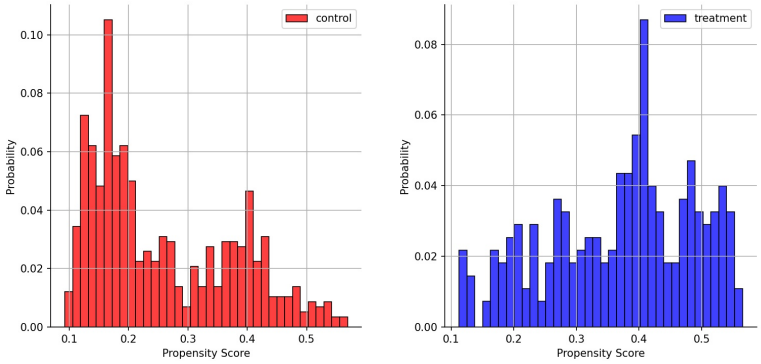


Figure 6.1: Propensity score distribution in control and treatment groups of scenario 1. There is an overlap in the distribution of the propensity scores of treatment and control groups.

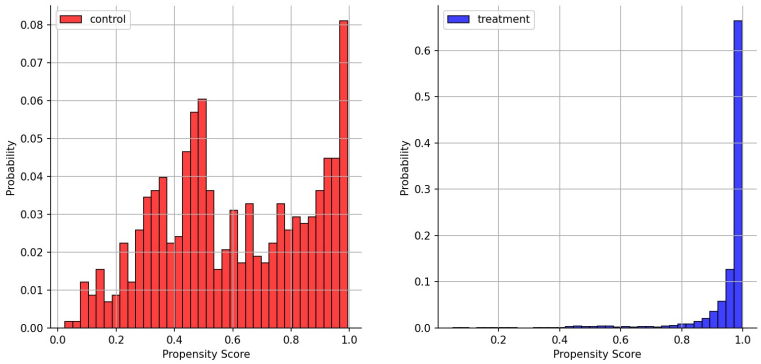


Figure 6.2: Propensity score distribution in control and treatment groups of scenario 2. There is an overlap in the distribution of the propensity scores of treatment and control groups.

Our goal is to determine whether these “treatments,” which have a common feature of involving less workload, reduce dropout rate. The Average Treatment Effect (ATE) obtained using propensity score matching is shown on Table 6.6. Across all three scenarios we can see mixed results,

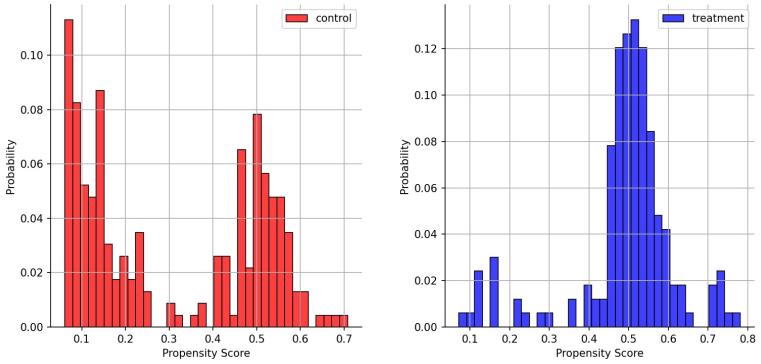


Figure 6.3: Propensity score distribution in control and treatment groups of scenario 3. There is an overlap in the distribution of the propensity scores of treatment and control groups.

as in some propensity buckets the treatment increases the risk of dropout (scenario 1, bucket “1. Low”; scenario 3, bucket “4. Med-high”) while in other cases the results are neutral or large reduction. In general, the results suggest that in high propensity to treatment conditions (bucket “5. High” i.e., students who are already likely to take less workload) there is a substantial reduction of the probability of dropout, particularly in scenarios 2 and 3.

The ATE values obtained from IPW, AIPW, and DROrthoForest methods are shown in Table 6.7 for all scenarios. In the case of IPW and AIPW, we can see that the 95% confidence intervals (from “lower-ci” to “upper-ci” in the table) contain the value zero. This means that the uncertainty in these methods is large and we cannot establish with them whether there is a change in the dropout risk due to the treatment. However, the results with the DROrthoForest method, which is a combination method of causal forests and doubly robust learner, are all negative with confidence intervals that do not contain the zero; indeed, they show a reduction of the probability of dropout of about 5 percentage points in all three scenarios because of the treatment.

Table 6.6: ATE obtained using Propensity Score Matching with five buckets.

Propensity	1. Low	2. Med-low	3. Med	4. Med-high	5. High
Scenario 1	0.18	0.04	-0.05	0.02	-0.08
Scenario 2	-0.04	0.03	0.00	0.02	-0.42
Scenario 3	0.04	-0.08	-0.17	0.30	-0.22

Table 6.7: IPW, AIPW, and DROrthoForest results estimating the Average Treatment Effect (ATE) and its 95% confidence interval [lower-ci, upper-ci] in three scenarios.

	IPW			AIPW			DROrthoForest		
Scenario	lower-ci	ATE	upper-ci	lower-ci	ATE	upper-ci	lower-ci	ATE	upper-ci
Scenario 1	-0.06	0.02	0.11	-0.01	0.07	0.15	-0.07	-0.06	-0.05
Scenario 2	-0.03	0.03	0.09	-0.06	0.01	0.08	-0.04	-0.04	-0.03
Scenario 3	-0.12	-0.01	0.10	-0.10	0.01	0.12	-0.07	-0.05	-0.03

6.6 Discussion, Conclusions, and Future Work

In this study, we first created ML models to predict dropout (students who enroll in the first year but do not show up in the second year) and underperformance (failing two or more subjects in the regular exams of the first year), using only information available at the time of enrollment. The obtained AUC-ROC of our models were 0.70 and 0.74 for dropout and underperformance risks respectively, which shows a relatively reliable prediction of students at risk. This is particularly true for centers having large risk of dropout or underperformance, while the performance of the same models for centers having lower risk is lower. This is to some extent expected and in those cases we are modeling a phenomenon that is more rare.

Next, we focused in dropout risk prediction and found that workload (first year credits) was an important feature. We also compared dropout risk across various groups of students. The comparison showed that to a large extent there is higher probability of dropout in older students (age $>$ average-age), in students taking a higher workload (more first year credits than the established minimum and the median), and in students admitted through access types III and IV.

We considered three scenarios using a combination of these features. In these scenarios, interventions were designed having the common characteristic of a reduced workload for students. In each scenario, the propensity score of the treatment was obtained with AUC-ROC of $0.75 \sim 0.91$ using ML-based models. Then, for each scenario, the Average Treatment Effect (ATE) on dropout was computed using causal inference methods. The results suggest a negative effect, i.e., a reduction of risk of dropout, following a lower number of credits taken on the first year. An actionable recommendation that these results suggest is to ask students at risk (in this study, older students and students admitted through access types III and IV) to consider taking a reduced workload (e.g., the minimum established), or to ask educational policy makers to consider revising the regulations that establish the minimum number of credits (e.g., to reduce the current minimum).

In addition to creating ML models for early prediction of dropout and underperformance risks that exhibit high predictive performance, the originality of this contribution is focusing on the vulnerable groups of students prone to dropout, studying combinations of different features such as workload, age, and study access type, and using different causal inference models to calculate the effects of these features on dropout in terms of ATE. Causal inference methods such as the ones we used provide a path towards effectively supporting the students. They also allow to perform observational studies, as education is a domain in which some types of direct experimentation might be unethical or harmful. We also used a large dataset and our results hold across substantially diverse study programs. We stress that the methodology we described is broadly applicable. Our findings are likely to be specific to this particular dataset, but show the general effectiveness of the methodology in this setting.

More scenarios can be defined in terms of other combinations of the relevant features, to determine their effects on dropout or underperformance. Additionally, the causal inference methods used in this study can also be applied to other risks faced by higher education students.

Chapter 7

EFFECT OF CONDITIONAL RELEASE ON VIOLENT AND GENERAL RECIDIVISM: A CAUSAL INFERENCE STUDY

7.1 Introduction

Studies on US prison population show that almost 1% of the US adult population is incarcerated which is 5 to 10 times higher than the rates in Western European and other liberal democracies [Travis et al., 2014, Loeffler and Nagin, 2022]. Looking at the latest European prison population rate reported in January 2021, there is a rate of 102 incarcerated persons per 100,000 inhabitants which is about 0.1% of the European population [Marcelo F. Aebi, 2022]. Although incarceration rates in European countries are not nearly as high as in the United States, in all countries the rise of ”mass incarceration” over the last half century has caused an increasing attention to assessing the effects on crime rates as well as the social and economic costs [Raphael and Stoll, 2009, Durlauf and Nagin, 2011, Spelman, 2020, Loeffler and Nagin, 2022].

Measures to address mass incarceration, e.g. by releasing a share of

the prison population, have to consider the tradeoff between the social and economic costs of potentially unnecessary incarceration and a potentially increased risk to public safety as more people with a probably higher likelihood of re-offense are released. However, from a mid-to-long term perspective and according to the literature of the past decades, this trade-off is not as strong as it initially seems, especially if we go beyond the simplistic view of its underlying dichotomous decision (release or no release) and consider a variety of alternative rehabilitation focused interventions. In fact, an extended period of incarceration as opposed to rehabilitation-focused early release programmes may yield a net-positive effect on overall reoffense rates. Understanding the effects that such programmes may have on overall recidivism rates can help navigate these considerations.

In this paper, we study the effect of conditional release on recidivism. Conditional Release (C.R.) is an early release from prison that an incarcerated person can obtain before fulfilling the complete prison sentence. It needs satisfying some requirements. C.R. can, by definition, help reduce the number of people who are incarcerated. However, it is unclear how increased availability of C.R. will impact recidivism and public safety.

Some studies suggest that incarceration or the length of incarceration has a deterrent effect on recidivism, whereby people refrain from committing crimes for fear of the resulting sanctions. Generally, crime prevention avoids both the costs of crime and the costs of punishment [Marchese di Beccaria, 1819, Becker, 1968, Cotter, 2020]. However, incarceration often fails to achieve deterrence from recidivism, and triggers punishment and costs [Loeffler and Nagin, 2022]. Incarceration punishment may reduce crime during incapacitation, when the person is physically separated from free society, however, beyond that it has a chastening impact on the incarcerated person. Punishment may affect future criminality of a person through different mechanisms, some of which such as rehabilitation may reduce future criminal involvement, whereas others such as social stigma may increase criminal involvement.

Prison conditions and prison experience are very important in the determination of the direction and magnitude of the incarceration effect. The

effects are heterogeneous as we can see when comparing the findings of studies on Scandinavia-based prisons, which are more orientated towards rehabilitation [Benko, 2018, Bhuller et al., 2020, Lappi-Seppälä, 2012, Hjalmarsson and Lindquist, 2020], with studies on US-based prisons, which are more orientated towards punishment [Cullen et al., 2000, Beckett and Sasson, 2003, Weaver, 2007, Cullen and Gilbert, 2012]. Other studies have shown that prisons do not reduce recidivism, but in fact, act as "schools for crime" which has a criminogenic effect that increases the risk of recidivism. In addition, the effectiveness of intermediate sanctions (penalties that exist between prison and probation) is mediated only through the provision of appropriate cognitive-behavioral treatments [Gendreau et al., 2000, Cullen et al., 2011]. Prison incarceration with a focus on rehabilitation can be largely crime preventive. Rehabilitation programs such as employment training services during sentence, which is common in Scandinavian countries, correctional substance abuse treatments, and generally high quality prisons can decrease future criminal involvement, including recidivism [Bhuller et al., 2020, Sondhi et al., 2020, Tobón, 2020, Andrews and Bonta, 2010]. In addition to rehabilitation during sentence, noncustodial sanctions which are partially or totally alternative to prison such as community sentences, electronic monitoring and semi-liberty [Cid, 2009, Henneguelle et al., 2016, Yukhnenko et al., 2019b, Statistics and Agency, 2019, Monnery et al., 2020, Williams and Weatherburn, 2022, Andersen and Telle, 2022], as well post-prison interventions such as employment, housing, and social reintegration support can also help reduce recidivism risk [Western, 2018, Kirk, 2020, Harding and Harris, 2020].

In hopes of reducing incarceration rates without substantially increasing crime, decision makers commonly use violence risk assessment tools when making noncustodial decisions such as conditional release. The main purpose of such tools is to prevent criminal violence and its consequences, but they also help prison management identify people with a greater risk of recidivism and allocate rehabilitation efforts accordingly. Ideally, accurate risk assessment may help place low-risk defendants into alternative programs to prison [Andrés-Pueyo et al., 2018]. Accordingly,

cases with an assigned low risk level have higher chances of receiving conditional release compared to the cases who are assessed as high risk. In this regard, these tools have to diagnose correctly and target the proper person to be released early, which in turn may reduce the recidivism rate [Austin, 2006].

Research contribution. In this work, we use machine learning (ML) supported causal inference methods to study the effect of Conditional Release (C.R.) on general and violent recidivism risk within 2 to 5 years of a person's release from prison. We consider 22,726 people released between 2010 and 2016 from 87 prison centers in Catalonia. Among them, 28% are released under C.R., while the remaining 72% are released after completing their entire sentences, something called Definitive Release (D.R.).[1]

We look at both general and violent recidivism rates within 2 to 5 years of each release year (between 2010 and 2016). The average rates in general and violent recidivism within 5 years of all released cases are 17.1% and 4.7% respectively. A comparison of the means of several demographic and penitentiary features between C.R. cases and D.R. cases and between men and women reveal relevant differences between the former and striking differences between the latter groups. Therefore, we conduct separate analyses for men and women, by also creating two independent models. A diagram for the methodology steps used in this paper is shown on Figure 7.1.

We use causal inference methods, which involve several computational steps, i.e. the creation of a predictive model of C.R. propensity and of predictive models of general and violent recidivism risks. The models are obtained using different ML methods. General input features to the models include demographics, penitentiary variables, and risk items and computed risk scores and levels of a risk assessment tool named RisCanvi (name changed for double-blind review). The best predictions in terms

[1]Additionally, a small number of cases, not included in our dataset, are released due to other reasons, including being pardoned or successfully asking for a retrial finding them not guilty.

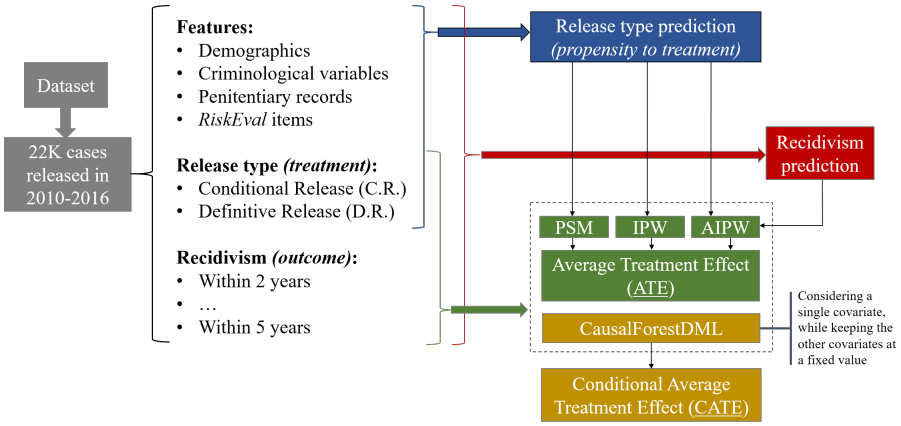


Figure 7.1: Methodology diagram

of AUC-ROC (Area Under ROC Curve) are applied to the causal inference methods to calculate the Average Treatment Effect (ATE) of C.R. on general and violent recidivism. The causal methods we use are popular impact evaluation methods in economics and social science in case of non-random assignment of individuals to alternative policies in observational studies [Athey, 2015, Athey and Wager, 2019], the methods include Propensity Score Matching (PSM) [Rosenbaum and Rubin, 1983], Inverse-Propensity score Weighting (IPW) [Bray et al., 2019], and Augmented Inverse-Propensity Weighting (AIPW) [Glynn and Quinn, 2010].

We determine heterogenous effects of C.R. on recidivism by estimating the Conditional Average Treatment Effect (CATE) using Causal Forest and Double Machine Learning [Athey et al., 2019, Nie and Wager, 2021, Chernozhukov et al., 2016]. To observe the treatment effect consistency with the risk estimated by the RisCanvi risk assessment tool, we compare ATE values separately for cases that have been estimated by the RisCanvi tool to have low, medium, or high risk.

The rest of this paper is organized as follows. After outlining related work in Section 7.2, risk assessment tools and conditional release are described in Section 7.3. In Section 7.4, the dataset used in this study is described and analysed with respect to recidivism and conditional release

variables. The methodology is presented in Section 7.5. Results are given in Section 7.6, the results are discussed and the paper is concluded in Section 7.7. Finally, the imputation process we applied for the RisCanvi items and the list of features used in this study are presented in Section 7.8 and Section 7.9 respectively.

7.2 Related work

Criminologists have long studied the effects of punishment or treatment on recidivism, where they have drawn from various different methods. In this section, we focus on causal inference methods such as Instrumental Variables (IV), Regression Discontinuity (RD), and other statistical methods [VanderWeele, 2015].

IV methods are used to approximate the methodological conditions of randomized control trials, by conditioning on a variable that is exogenous to the treatment status and filtering out selection bias that normally contaminates the estimated impact of treatment on the outcome of interest [Angrist et al., 1996]. The most commonly used instrumental variable in the studies dealing with the effects of custodial/noncustodial sanctions is the random assignment of cases to judges [Loeffler and Nagin, 2022]. Some research suggests that extraneous factors such as hunger or a bad mood can influence judge decisions [Danziger et al., 2011, Cho et al., 2016, Eren and Mocan, 2018, Heyes and Saberian, 2019], despite a societal agreement that such factors should not influence judicial decisions. In particular, it has been observed that the proportion of favorable rulings decreases with serial order within a session but goes back to the initial level after a session break that includes eating a meal [Danziger et al., 2011]. Another study revisited this finding and claimed that their analyses do not provide conclusive evidence for the hypothesis that mood influences legal rulings [Glöckner, 2016]. In fact, the observed downward trend could be explained by selective dropout of favorable cases due to rational time management, censoring of data and autocorrelation .

Regression Discontinuity (RD) is another strategy that addresses selec-

tion bias in estimates of treatment effects on the outcome. RD was first developed in education research [Thistlethwaite and Campbell, 1960]. In RD research designs, the assignment of units to treatments is performed based on a score-based system. In these scoring systems, when the assignment is discontinuous and deterministic at some threshold value along the score, any sudden changes in the outcome of interest can be causally attributed to the effects of treatment [Loeffler and Nagin, 2022].

Other statistical methods include methodologies encompassing regression models and inverse probability weighting that have been developed for the estimation of a treatment effect on an outcome. These include methods such as propensity score matching (PSM) [Rosenbaum and Rubin, 1983], regression adjustment (RA) [Vansteelandt and Daniel, 2014], inverse probability weighting (IPW) [Bray et al., 2019], and augmented inverse probability weighting (AIPW) [Glynn and Quinn, 2010].

On Table 7.1, we summarize several causal inference studies on the effects of custodial and noncustodial sanctions on recidivism. We explain them in the following sections in two categories of studies on incarceration effects on recidivism and alternatives to prison effects on recidivism.

7.2.1 Effects of incarceration on recidivism

There are many studies on the effect of incarceration on recidivism [Loeffler and Nagin, 2022]. The most used method in these studies is an Instrumental Variables (IV) approach which is used to estimate the causal impact of incarceration on recidivism by controlling for an exogenous variation in the assignment of cases [Green and Winik, 2010, Loeffler, 2013, Mueller-Smith, 2015, Gupta et al., 2016, Harding et al., 2017, Bhuller et al., 2020]. In a study of the District of Columbia's Superior Court, drug-related persons are assigned randomly to different judicial calendars on which judges gave out sentences that varied substantially in terms of prison and probation time [Green and Winik, 2010]. Their results show that variations in prison and probation time have no noticeable effect on recidivism rates. Also, in another study on the cases from the state of Georgia, the causal effect of prison and parole time on recidivism is estimated by relying on

two instrumental variables [Zapryanova, 2020]. The results are consistent, and show that time in parole has no significant effect on recidivism and time in prison has a negative effect of 1.04 percentage points only if a person recidivates while on parole, which seems to have no effect on overall recidivism. However, the results are different in a different IV study on data from Norway, which shows that time spent in prison with a focus on rehabilitation can be preventive and reduces further criminal behavior [Bhuller et al., 2020]. One reason for this could be that the Norwegian prison system is successful in increasing participation in rehabilitation programs such as job training and encouraging employment. In an investigation on data from Texas, using IV estimates, it is found that incarceration generates modest incapacitation effects and sizable social costs to society [Mueller-Smith, 2015].

In addition to judge IV studies on imprisonment effects, Regression Discontinuity (RD) is applied in estimating the effects of incarceration on recidivism. An example of RD research includes estimating the causal effects of prison conditions (custodial security classification levels) on recidivism which suggests that harsher prison conditions lead to more post-release crime [Chen and Shapiro, 2007]. Another RD study shows that processing juveniles in the adult system may not uniformly increase offending and may reduce offending in some circumstances [Loeffler and Grunwald, 2015]. Also, using the RD approach, it is shown that prison has no effect on the reconviction rates of persons offended due to drug in Florida [Mitchell et al., 2017].

There are also various studies which examine the effect of incarceration on recidivism using statistical methods. In a study on data from the Florida Department of Corrections, the effects of ”supermax housing” (a highly restrictive type of incarceration) on 3-year recidivism outcomes is investigated using propensity score matching analysis. They show that supermax incarceration may increase violent recidivism [Mears and Bales, 2009]. Another study using propensity score matching on cases from several regions of the UK shows that incarceration slightly increases the chances of reoffending [Jolliffe and Hedderman, 2015].

7.2.2 Effects of alternatives to prison on recidivism

Multiple studies have sought to determine whether programs providing an alternative to prison reduce recidivism, and to measure the extent of this reduction [Vass, 1990, Dynia and Sung, 2000, Cid, 2009].

Most previous research uses IV methods. One study in France, by using IV estimates, shows that converting entire sentences into electronic monitoring (sentence at home under electronic monitoring instead of incarceration) has long-lasting beneficial effects on recidivism rates [Henneguelle et al., 2016]. The estimates suggest that this conversion can reduce the probability of reconviction by 6-7 percentage points after five years. Similarly, in another paper, the effect of an electronic monitoring program in Norway is evaluated on the recidivism rate using IV design [Andersen and Telle, 2022]. Their results show a reduction of about 15 percent in two-year recidivism rates and approximately 0.3 offences on average in the one-year recidivism frequency. In a study in Israel, it is shown that the parole requests of cases appearing further from the judges' last break are more often denied by the judges [Meier et al., 2020]. Exploiting this behavioral pattern in an instrumental variable, the authors estimate that early release decisions driven by exogenous factors reduce the propensity to recidivate.

In another study, semi-liberty is also introduced as a suitable alternative to prison which has a beneficial effect on recidivism [Monnery et al., 2020]. In this study, it is found that under treatment exogeneity and conditional independence, semi-liberty results in a reduction of 22% to 31% in cases' recidivism in the five years after release.

There are few Regression Discontinuity (RD) studies on the effect of alternatives to prison on recidivism. A RD study based on data from England and Wales shows that early release on electronic monitoring (EM) can reduce the probability of rearrest by 5 to 7 percent [Marie, 2009]. In another research using both IV and RD methods, it is found that average length of prison stay can be reduced by 7.5 months with a small impact on recidivism [Rhodes et al., 2018].

Regarding statistical methods, we find comparatively less research on

the effect of alternatives to prison on recidivism. The impact of multiple component treatments on reoffending of incarcerated people with an alcohol use disorder in England is investigated using multiple treatment effect estimators (RA, IPW, and AIPW, and IPWRA) [Sondhi et al., 2020]. The results show that a Risk Need Responsivity (RNR) programming is the most effective intervention compared to other treatments and represents a lower recidivism rate for treated cases compared to the control group. By contrast, pharmacological treatment results in a statistically significant higher level of reoffending in treated cases relative to the control group. In another study, a treatment program (named "*Step Up*") for youth and families experiencing Child to Family Violence (CFV) is evaluated and its effects on the three outcomes of general recidivism, assault-related recidivism and domestic violence-related recidivism are estimated using an IPW estimator [Gilman and Walker, 2020]. The results show that, even when including youth who did not fully complete the program, there is a significantly lower risk of general recidivism for treated cases compared to the control group; and for program completers, the effects are even more pronounced.

The evidence on the effect of incarceration and alternative programs on recidivism is mixed and seems to depend on location, differences in the objectives of the incarceration system, and type of offense. Comparing these effects in Table 7.1, we can see that generally the literature suggests that custodial sanctions have, at best, no effect or even a criminogenic effect on recidivism, except for rehabilitation-focused incarceration. However, non-custodial alternative programs to prison mostly show preventative effects and to a small extent show no effect on recidivism.

In the present study, we consider conditional release, which is an alternative to a part of prison sentence, as a treatment and measure its effect on two types of general and violent recidivism. The Average Treatment Effect (ATE) is measured using statistical causal inference methods such as PSM, IPW, and AIPW. It is noteworthy that the effect of conditional release on recidivism has not been evaluated so far by the methods we use in this paper.

Table 7.1: Causal inference studies on the effects of custodial and noncustodial sanctions on recidivism

Study	Num. of observations and location	Follow-up after release	Method	Sanction	Cause	Effect on recidivism
[Green and Winik, 2010]	1,003 (Washington D.C., USA)	4 years	IV	Custodial Noncustodial	Prison time Probation time	No effect No effect
[Zapryanova, 2020]	700,000 (Georgia, USA)	3 years	IV	Custodial Noncustodial	Prison time Parole time	No effect No effect
[Bhuller et al., 2020]	23,373 (Norway)	5 years	IV	Custodial	Prison time (rehabilitation)	Preventative
[Chen and Shapiro, 2007]	1,205 (USA)	1-3 years	RD	Custodial	Prison conditions (security levels)	Criminogenic
[Mitchell et al., 2017]	96,254 (Florida, USA)	3 years	RD	Custodial	Incarceration	No effect
[Mears and Bales, 2009]	58,752 (Florida, USA)	3 years	PSM	Custodial	Supermax incarceration	Criminogenic
[Jolliffe and Hedderman, 2015]	5,500 (ENG & WLS, UK)	1 year	PSM	Custodial	Incarceration	Criminogenic
[Henneguelle et al., 2016]	2,827 (France)	5 years	IV	Noncustodial	Electronic monitoring	Preventative
[Andersen and Telle, 2022]	48,636 (Norway)	1-3 years	IV	Noncustodial	Electronic monitoring	Preventative
[Meier et al., 2020]	804 (Israel)	1-6 years	IV	Noncustodial	Parole	Preventative
[Monnery et al., 2020]	1,445 (France)	5 years	IV	Noncustodial	Semi-liberty	Preventative
[Marie, 2009]	260,000 (ENG & WLS, UK)	1-2 years	RD	Noncustodial	Early release (EM)	Preventative
[Rhodes et al., 2018]	304,000 (USA)	3 years	IV RD	Noncustodial	Prison length reduction	No effect
[Sondhi et al., 2020]	59,150 (England, UK)	1 year	RA IPW AIPW IPWRA	Custodial	RNR Treatment Pharm. Treatment	Preventative Criminogenic
[Gilman and Walker, 2020]	1,478 (Washington, USA)	1 year	IPW	Custodial	Treatment (”Step Up”)	Preventative

7.3 Risk Assessment and Conditional Release

7.3.1 Risk Assessment Instrument

With public safety as one of the fundamental goals of intervention with incarcerated persons, the need for accurate risk assessments has intensified in recent decades. The adoption of structured risk assessment tools has made major progress in the past 40 years. Although these tools are far from perfect, they are more accurate compared to unguided professional judgement used to assess the risk for violence in the 1980s [Hanson, 2005]. These tools are used in many socially relevant contexts such as public health, information security, project management, auditing, and criminal justice [Raz and Michael, 2001, Alberts and Dorofee, 2003, Allen et al., 2006, Anenberg et al., 2016]. In the field of criminal justice, they are applied in different areas such as pre-trial risk assessment, sentencing, probation, and parole [Kehl and Kessler, 2017, Lowenkamp, 2009, Monahan and Skeem, 2016, Miron et al., 2021, Wright et al., 1984, Funk, 1999, Meredith et al., 2007]. The risk estimated by these tools can be linked to an intervention consistent with the computed risk. The expectation is that persons assessed with low risk should have lower rates of being sentenced to prison, shorter sentences, higher rates of being paroled and receive lower levels of supervision compared to high-risk cases [Austin, 2006]. However, the goals of community protection do not require an exclusive focus on low-risk cases, but can also be effectively promoted when more resources and services are directed towards higher risk cases [Hanson, 2005]. Indeed, in the present study, we show, using causal inference methods, that conditional release as an intervention can reduce recidivism, and that this reduction is more pronounced in people deemed higher risk.

RisCanvi was introduced as a multi-level risk assessment protocol for violence prevention in the prison system of a community in a European country in 2009 [Andrés-Pueyo et al., 2018]. This protocol is applied multiple times during a person's period in prison; the official recommendation is to do this every six months, or at the discretion of the case manager.

RisCanvi is not a questionnaire. Instead, each person is interviewed by trained professional case workers. Two versions of the RisCanvi protocol were created, an abbreviated one of 10 items for screening (RisCanvi-S), and a complete one of 43 items (RisCanvi-C). The risk items are listed in Table 3.1. Risk items can be categorized into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, Attitudes/Personality. These items can also be divided into static factors (which cannot be altered, such as ”age of starting violent activity”) and dynamic factors (which can change, such as ”pro-criminal or antisocial attitudes”). In the original RisCanvi protocol, risk is determined for each incarcerated person relative to four possible outcomes: self-directed violence, violence in the prison facilities, committing further violent offenses, and breaking prison permits. A fifth risk score was introduced later for general recidivism [Singh et al., 2018]. The outcome of RisCanvi-S can be ”high-risk” or ”low-risk”. If the outcome is low-risk for all five criteria, the same RisCanvi-S protocol is repeated after six months. Otherwise, in the case of high-risk levels or significant change in a person’s situation, the complete version RisCanvi-C is applied. The outcome of RisCanvi-C can be ”high-risk”, ”medium-risk”, or ”low-risk”. When the risk levels measured by RisCanvi-C are medium or high, the next evaluation is again a RisCanvi-C; otherwise, RisCanvi-S is used.

7.3.2 Conditional Release

“Conditional release” (abbreviated C.R. throughout this paper) is similar to “parole” in the USA and is described in detail in the legislation of the country under study. It occurs when an incarcerated person who meets some requirements is released before completing the full period of the sentence. Cases that are not granted C.R. and are only released at the end of the sentence are described as “definitive release” (abbreviated D.R. throughout this paper). Regarding the cases in our study, each penitentiary center of Catalonia is associated to one court (“Court of Penitentiary Oversight”) that includes one judge. In each center, C.R. requests are proposed by a treatment committee to the judge after the person has completed, in most

cases, 75% or, in some few cases, 67% of the sentencing time. When this committee believe that the person presents a low risk to society if released early, they prepare a request for C.R. and present it to the judge. This request does not explicitly include the computed RisCanvi level, but in the majority of cases, a high RisCanvi risk level makes it unlikely that a request for C.R. will be presented. Finally, the decision for C.R. of the cases is taken by the judge assigned to that center.

7.4 Dataset

The anonymized dataset used in this study has been provided by the prison centers of Catalonia and consists of 26,305 prison releases between 2010 and 2016 which are definitive (72%) or conditional (28%) releases among 22,726 individuals.[1] These cases have been evaluated with RisCanvi every 6 months. Persons who have only RisCanvi-S evaluation are low-risk cases and cases with both RisCanvi-S and RisCanvi-C evaluations are the ones who have been evaluated as high-risk in RisCanvi-S and then assigned to RisCanvi-C. For each case, we sought the latest RisCanvi (RisCanvi-S or RisCanvi-C) evaluation, considering it valid for the purposes of predicting recidivism if it was done at most 9 months before the release date. After imputing missing items in RisCanvi evaluations (the imputation process is explained in the Section 7.8) and removing cases with incomplete evaluations, we remained with 15,029 evaluations which are presented per release year in Table 7.2. As can be seen, the number of evaluations has increased with each passing year, as RisCanvi is adopted more consistently and thoroughly through the entire prison system.

[1]Note that there are more prison releases than studied individuals as the data can record more than one prison release per person.

Table 7.2: RisCanvi evaluations per release year

Release year	2010	2011	2012	2013	2014	2015	2016
No. of releases	3,494	3,766	4,152	4,010	3,999	3,596	3,288
No. of evaluations (fraction with a valid evaluation at most 9 months prior to release)	634 (18%)	1,776 (47%)	2,320 (56%)	2,501 (62%)	2,702 (68%)	2,582 (72%)	2,514 (76%)
No. of Screening evaluations	172	670	1,021	1,036	916	836	807
No. of Complete evaluations	462	1,106	1,299	1,465	1,786	1,746	1,707

Table 7.3: Average recidivism rates two to four years after release for people released in 2010-2016

Recidivism type	within 2 years	within 3 years	within 4 years	within 5 years
Any (General)	10.0%	12.8%	15.2%	17.1%
Violent	3.2%	3.9%	4.4%	4.7%

We consider 220 features (feature list is found on Table 7.17 of Section 7.9) including 23 demographic features, 146 penitentiary features, 43 RisCanvi items and 8 RisCanvi risk levels and scores (4 risk levels and 4 risk scores). After dropping cases with missing values, as well as very few special cases (twelve) that underwent RisCanvi evaluation but were not sentenced, 12,250 cases remain in the final dataset used in our analysis.

7.4.1 Recidivism

We obtained general and violent recidivism rates of the cases in four follow-up periods within release date. The rates are shown in Table 7.3. Logically, the probability of committing a new crime after being released from prison is non-decreasing over the span of time spent out of prison.

In Figure 7.2, we can also observe general and violent recidivism rates in the four follow-up periods for each release year. In all the follow-up periods, the highest rates of general and violent recidivism are observed in the persons released in earlier years of 2010 and 2011. Comparing recidivism rates of 2016 to recidivism rates of 2010, we find that there has been a minimum 6 and 2 percentage points decrease in general and violent recidivism rates, respectively. This decreasing rate is part of a global trend on the reduction of recidivism and crime rate [Velázquez, 2018, Tonry, 2014].

General recidivism rates have decreased in cases released between 2010 and 2014 in all follow-up periods, but the rates have been unchanged or increased for the cases released after 2014. There has been a decreasing rate of violent recidivism for the cases released between 2010 and 2015 within two years of their release. However, for when looking at the recidivism rates for more than 2 years follow-up, this decrease happens for persons released between 2010 and 2013 and in 2015, while we observe an increased rate of violent recidivism for release years 2014 (except for the follow-up period of two years) and 2016 (except for the follow-up period of five years).

The rate we obtained for general recidivism of the cases released in 2010 (3,494 cases) in the follow-up period of five years (28%) is almost

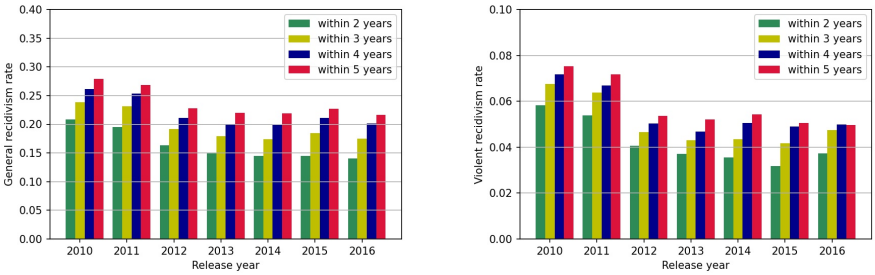


Figure 7.2: Recidivism rates in four follow-up periods within each release year

inline with what is estimated by the Center of Legal Studies and Specialized Training of the community [Capdevila et al., 2015]. They estimate a general recidivism rate of 30% within 5.5 years of the cases released in 2010 (3,414 cases that are almost the same cases of our study who were released in 2010). In their report, there is a rate drop of 10 percentage points in general recidivism of the persons released in 2010 compared to the cases released in 2002.

For different groups of people released between 2010 and 2016, we can observe general and violent recidivism rates within 5 years of their release in Table 7.4.

Most groups are self-explanatory:

- *Single/other* refers to their civil status.
- *With deportation* are cases in which, normally due to violations of immigration law, the person is expected to be removed from the country immediately after being released.
- *Base crime types* include violent or non-violent crimes against property, crimes against people, gender-based violence, crimes against sexual freedom, drug-related crimes, traffic-related crimes, and others.

- *Permission* is an ordinary short leave from prison during the base sentence due to some reasons such as death or serious illness of a direct family member of the person.

- *Prison degree* can be 1st: cases that are isolated from the general prison population, 2nd: cases who are in general prison population, and 3rd: cases who must spend 8 hours in prison every day but can be outside 16 hours per day.

- *With electronic surveillance* are cases who have an electronic surveillance mechanism, usually an ankle bracelet. They are in the 3rd degree, but instead of coming to prison at night, they can sleep at their own home, as the ankle bracelet can send an alert if they are not at home at night.

- *Dependent units* are special, managed housing units outside of prison that can be assigned to those in the 3rd degree.

Comparing recidivism rates of these groups against the base rate (overall prevalence), we can see that cases younger than 30 at the time of release, who are national, with single civil status, pending deportation, who are in prison because of a violent crime or crime against property, convicted to more than 5 years sentence, with more than one previous prison entry, with rejected permission or no permission request, with mostly degree regression during their sentence (i.e., mostly being moved to a more restricted environment instead of a freer one), having (very) severe rules violations within prison, with (no) lower points in the prison evaluations, who have been relocated to another module within prison multiple times, who at least once went to special supervision, psychiatry, or nursing modules, and persons who were classified in the 1st and 2nd degree before their release have relatively higher violent and general recidivism rates compared to the related base rates. There are lower general and violent recidivism rates for cases with electronic surveillance and persons who spent time in dependent units.

Table 7.4: Recidivism rates within five years of release for different groups

Group	Size	General recidivism within 5 years Base rate: 17.1%	Violent recidivism within 5 years Base rate: 4.7%
Male	93%	17.3%	4.9%
Female	7%	13.7%	2.0%
Age at release time ≤ 30	22%	23.0%	6.9%
Age at release time > 30	78%	15.4%	4.1%
National	62%	19.0%	5.4%
Foreigner	38%	13.9%	3.5%
Single	57%	20.2%	6.0%
Other	43%	13.0%	3.0%
With deportation	12%	22.5%	6.8%
Without deportation	88%	16.4%	4.4%
Violent base crime	35%	19.4%	7.6%
Non-violent base crime	65%	15.8%	3.2%
(Non) violent base crime against property	31%	27.0%	7.8%
Other types of base crime	69%	12.6%	3.3%
Base crime sentence < 5 years	77%	16.6%	4.0%
Base crime sentence ≥ 5 years	23%	18.6%	7.0%
Previous prison entries > 1	26%	30.2%	8.6%
Previous prison entries ≤ 1	74%	12.4%	3.3%
Permission rejection or no permission request	29%	26.0%	7.0%
Permission acceptance	71%	13.4%	3.8%
Mostly degree regression	15%	28.6%	9.2%
Others	85%	15.1%	3.9%
(Very) severe rules violations within prison	34%	26.3%	8.2%
Light/no rules violations within prison	66%	12.3%	2.9%
(No) lower evaluation points (level C and D)	37%	24.0%	7.0%
Higher evaluation points (level A and B)	63%	13.0%	3.4%
Module changes > 7 (median)	48%	20.7%	6.5%
Module changes ≤ 7	52%	13.8%	3.1%
Special supervision module ≥ 1	19%	30.1%	10.1%
No special supervision module	81%	14.1%	3.5%
Nursing module ≥ 1	18%	21.0%	7.0%
No nursing module	82%	16.2%	4.2%
Psychiatry module ≥ 1	4%	30.0%	11.6%
No psychiatry module	96%	16.5%	4.4%
Last prison degree before release: 1st & 2nd	45%	26.4%	7.7%
Last prison degree before release: 3rd	55%	9.4%	2.2%
With electronic surveillance	13%	5.2%	1.0%
Others	87%	18.8%	5.3%
In dependent units	3%	9.9%	2.9%
Others	97%	17.3%	4.8%

Table 7.5: Conditional release (C.R.) rate per year

Release year	2010	2011	2012	2013	2014	2015	2016
Conditional Release (C.R.) rate	22.6%	24.2%	26.5%	29.9%	30.4%	31.8%	34.5%

7.4.2 Conditional Release (C.R.) vs. Definitive Release (D.R.)

In line with the main objective of this study to estimate the effect of Conditional Release (C.R.) on recidivism, we first look at C.R. rates in comparison to recidivism rates over the years and at descriptive statistics of relevant features for both C.R. cases and Definitive Release (D.R.) cases separately.

Table 7.5 shows that the C.R. rate has increased yearly from 22.6% in 2010 to 34.5% in 2016. While there are year-by-year variations in the amount of this increase in C.R., there does not seem to be any discontinuity or sudden change during the observation period. This increase is part of a strong policy push applied by the Dept. of Justice to request C.R. for more cases, which has led treatment committees to request for C.R. more often. It does not reflect any change in C.R. legislation, or any change that we are aware of in the criteria applied by judges.

The increase in C.R. rates can also be a reason for the recidivism rate decrease during these years as shown previously in Figure 7.2. We can see that in the release years with the lowest C.R. rate (2010,2011) the recidivism rates within 2 to 5 years follow-up periods are the highest. Also, in release year 2014, which has a very small increase in C.R. rate, there has been no or a small recidivism rate decrease within 2 to 5 follow-up years. Even violent recidivism rates increased within follow-up periods higher than 3 years.

Considering conditional release as an intervention, or treatment that may reduce recidivism, we show some descriptive statistics of C.R. cases as treatment group in comparison to the ones of D.R. persons as control group. The comparison is presented in terms of some demographics, penitentiary features and recidivism within several years of release in

Table 7.6.

As can be seen, the observed mean difference of D.R. and C.R. cases is statistically significant in variables such as gender, age at release time, civil status, violent base crime, deportation, permission request, acceptance, and rejection, number of nursing, psychiatry, and supervision modules, degree regression and progress, (very) severe rules violations within prison, number of previous prison entry, base crimes against people or property, gender-based violence crime, base crimes related to drugs or traffic, having electronic surveillance, being in dependent units, RisCanvi risks of self-directed violence, violence in the prison facilities, violent recidivism, and breaking prison permits, general recidivism within 2 to 5 years, and violent recidivism within 2 to 5 years.

Comparing the mean value of the variables in C.R. and D.R. cases shows the following. Cases that are less likely to receive C.R. (and hence, more likely to be released under D.R.) tend to be: cases with single civil status, who are in prison because of violent crime or non-violent crimes against property, pending deportation, with no permission request or rejected permissions, spending time in nursing, psychiatry, or special supervision modules, with mostly degree regression, who remain in the 2nd prison degree, having (very) severe rules violations within prison, with previous prison entry, and having higher risk scores of self-directed violence, violence in the prison facilities, violent recidivism, and breaking prison permits. On the other hand, cases that are more likely to receive C.R. tend to be persons with married civil status, who were more often granted and enjoyed permissions, having always or mostly degree progressions, who are in prison because of a drug-related crime, with electronic surveillance, or lived in dependent units.

Both general and violent recidivism are more likely in cases released with D.R. compared to C.R. General recidivism within 2 to 5 years of prison release happens 12 to 17 percentage points more in cases with D.R. compared to ones with C.R. Violent recidivism within the same periods happens 3 to 5 percentage points more in cases with D.R. compared to C.R. We can also compare general and violent recidivism rates of the C.R. and D.R. cases within 2 to 5 years of their release per each release year

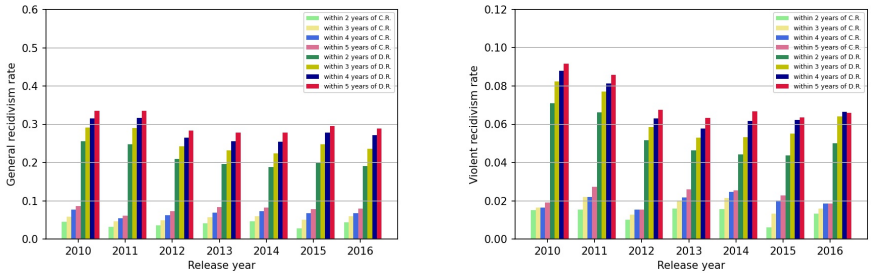


Figure 7.3: General and violent recidivism rates in C.R. (light-color bars) and D.R. (dark-color bars) cases

in Figure 7.3. We can see at least 14 percentage points and 3 percentage points higher general and violent recidivism rates respectively for D.R. cases compared to the C.R. persons in all follow-up periods within each release year.

From this observation alone, however, one cannot conclude that C.R. *causes* a reduction in recidivism risk. C.R. is granted almost exclusively to cases who are deemed to have lower risk, as the relevant legislation indicates that people who receive C.R. must have a “favorable individualized prognosis of social reintegration”. Therefore, we need to control for this selection through causal inference methods, if we want to study the causal effect of C.R. on recidivism.

7.5 Methodology

The distribution of the features (feature list is found on Table 7.17 of Section 7.9) by gender reveals significant differences between men and women which are explained in section 7.5.1. Considering conditional release as a treatment, we focus on the causal effect of this treatment on general and violent recidivism. The average treatment effect is obtained in section 7.5.4 using statistical methods such as Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse

Table 7.6: Descriptive statistics: control (D.R.) vs. treatment (C.R.)

Variable	Mean value in control (D.R.) (8,162 cases)	Mean value in treatment (C.R.) (4,088 cases)	P-value (D.R. vs. C.R.)
Male (0=no, 1=yes)	0.94	0.89	< 0.0001
Age at release	38.78	40.67	< 0.0001
Single	0.60	0.50	< 0.0001
Married	0.19	0.28	< 0.0001
Violent base crime (0=no, 1=yes)	0.39	0.28	< 0.0001
Deportation (0=no, 1=yes)	0.15	0.04	< 0.0001
Rejected permission (0=no, 1=yes)	0.24	0.03	< 0.0001
Accepted and enjoyed permission (0=no, 1=yes)	0.56	0.96	< 0.0001
No permission request (0=no, 1=yes)	0.18	0.01	< 0.0001
Number of nursing modules	0.49	0.39	0.001
Number of psychiatry modules	0.13	0.04	< 0.0001
Number of special supervision modules	0.82	0.18	< 0.0001
Mostly degree regression (0=no, 1=yes)	0.20	0.05	< 0.0001
Mostly degree progress (0=no, 1=yes)	0.24	0.52	< 0.0001
Remaining in the 2nd prison degree	0.48	0.00	< 0.0001
Always degree progress	0.20	0.51	< 0.0001
Severe prison rules violations (0=no, 1=yes)	0.25	0.15	< 0.0001
Very severe prison rules violations (0=no, 1=yes)	0.16	0.05	< 0.0001
Number of previous prison entry	1.55	1.11	< 0.0001
Base crime: Against people	0.15	0.12	< 0.001
Base crime: Gender-based violence	0.13	0.06	< 0.0001
Base crime: Against property (non-violent)	0.24	0.12	< 0.0001
Base crime: Drugs	0.10	0.38	< 0.0001
Base crime: Traffic	0.12	0.09	< 0.0001
With electronic surveillance	0.05	0.29	< 0.0001
In dependent unit	0.02	0.06	< 0.0001
Self-directed violence risk score	10.83	2.57	< 0.0001
Score of violence in the prison facilities	9.70	1.22	< 0.0001
Violent recidivism risk score	10.26	1.12	< 0.0001
Breaking prison permits risk score	0.05	-1.46	< 0.0001
General recidivism (0=no, 1=yes) within 2 years	0.14	0.02	< 0.0001
General recidivism (0=no, 1=yes) within 3 years	0.17	0.04	< 0.0001
General recidivism (0=no, 1=yes) within 4 years	0.20	0.05	< 0.0001
General recidivism (0=no, 1=yes) within 5 years	0.23	0.06	< 0.0001
Violent recidivism (0=no, 1=yes) within 2 years	0.04	0.01	< 0.0001
Violent recidivism (0=no, 1=yes) within 3 years	0.05	0.01	< 0.0001
Violent recidivism (0=no, 1=yes) within 4 years	0.06	0.01	< 0.0001
Violent recidivism (0=no, 1=yes) within 5 years	0.06	0.01	< 0.0001

Propensity Weighting (AIPW). All of these methods rely on the propensity to treatment which is estimated in Section 7.5.2 using Machine Learning (ML) models. In the AIPW method, we also need to obtain a model for the outcome (general/violent recidivism). We present the models for both general and violent recidivism prediction in section 7.5.3. Finally, in section 7.5.5, we determine the treatment effect heterogeneity by estimating Conditional Average Treatment Effect (CATE) using Generalized Random Forest and Double Machine Learning methods.

7.5.1 Gender Differences

We observe significant differences between men and women within our dataset, so we prefer to treat them differently and create separate ML models for these two groups [Skeem et al., 2016, Collins, 2010, Huebner et al., 2010]. The rest of this section describes these differences.

Table 7.7 shows descriptive statistics of relevant features of this study separately for men and women. The mean difference is statistically significant for almost all listed variables.

Comparing the mean value of the variables for men and women shows that the group of women comprises more nationals, cases with accepted and enjoyed permissions, number of prison activities, number of nursing and special supervision modules, number of light, severe, and very severe prison rules violations, cases with always or mostly degree progress, cases with electronic surveillance or spending in dependent units, cases in prison because of a non-violent crime against property or a drug-related crime, and cases receiving C.R.

In contrast, there are more cases with single civil status, who are in prison because of a violent crime, crime against people, gender-based violence, and traffic-related crime, pending deportations, with rejected permissions, who remain in the 2nd prison degree, with a previous prison entry, having RisCanvi risks of self-directed violence, violence in the prison facilities, and violent recidivism, and with general recidivism within 3 to 5 years and violent recidivism within 2 to 5 years among men than among women.

Table 7.7: Descriptive statistics: men vs. women

Variable	Mean value in men (11,335 cases)	Mean value in women (915 cases)	P-value (men vs. women)
National (0=no, 1=yes)	0.61	0.67	< 0.001
Single	0.57	0.51	< 0.001
Violent base crime (0=no, 1=yes)	0.36	0.20	< 0.0001
Deportation (0=no, 1=yes)	0.12	0.04	< 0.0001
Number of rejected permissions	0.84	0.64	0.1
Number of accepted and enjoyed permissions	9.68	10.66	< 0.1
Number of activities	16.80	19.90	< 0.0001
Number of nursing modules	0.41	1.03	< 0.0001
Number of special supervision modules	0.56	1.23	< 0.0001
Number of light prison rules violations	0.04	0.28	< 0.0001
Number of severe prison rules violations	0.56	0.80	< 0.0001
Number of very severe prison rules violations	0.46	0.62	0.001
Mostly degree progress	0.33	0.39	< 0.0001
Remaining in the 2nd prison degree	0.33	0.20	< 0.0001
Always degree progress	0.30	0.37	< 0.0001
With electronic surveillance (0=no, 1=yes)	0.11	0.30	< 0.0001
In dependent unit (0=no, 1=yes)	0.03	0.11	< 0.0001
Previous prison entry (0=no, 1=yes)	0.83	0.71	< 0.0001
Base crime: Against people	0.14	0.09	< 0.0001
Base crime: Gender-based violence	0.11	0.02	< 0.0001
Base crime: Against property (violent)	0.11	0.09	0.1
Base crime: Against property (non-violent)	0.20	0.28	< 0.0001
Base crime: Drugs	0.18	0.34	< 0.0001
Base crime: Traffic	0.11	0.05	< 0.0001
Self-directed violence risk score	8.49	2.93	< 0.0001
Score of violence in the prison facilities	7.25	2.14	< 0.0001
Violent recidivism risk score	7.84	-0.67	< 0.0001
Breaking prison permits risk score	-0.44	-0.68	< 0.1
C.R.	0.32	0.48	< 0.0001
General recidivism (0=no, 1=yes) within 2 years	0.10	0.08	< 0.1
General recidivism (0=no, 1=yes) within 3 years	0.13	0.10	0.01
General recidivism (0=no, 1=yes) within 4 years	0.15	0.12	< 0.01
General recidivism (0=no, 1=yes) within 5 years	0.17	0.14	< 0.01
Violent recidivism (0=no, 1=yes) within 2 years	0.03	0.01	< 0.001
Violent recidivism (0=no, 1=yes) within 3 years	0.04	0.01	< 0.0001
Violent recidivism (0=no, 1=yes) within 4 years	0.05	0.02	< 0.001
Violent recidivism (0=no, 1=yes) within 5 years	0.05	0.02	< 0.0001

Also, testing a global model of propensity to C.R. (which is trained using all population consisting both men and women) on women and men results in different predictive performances which can be interpreted as an algorithmic bias. So we prefer not to use a global model to prevent this algorithmic discrimination.

Based on these important differences between men and women, we do our analysis separately for these two groups but with the main focus on the men population which makes up the majority in the dataset (93%).

7.5.2 Propensity to Conditional Release (C.R.)

We consider Conditional Release (C.R.) as a treatment, and hence people released with C.R. are the treatment group, and people released with Definitive Release (D.R.) are the control group. Among releases between 2010 and 2016, there are 11,335 releases for men; 32% of them are C.R. (treatment) and 68% are D.R. (control). Similarly, there are 915 releases for women, 48% of them are C.R. and 52% are D.R.

For both men and women, we estimate the propensity to treatment (C.R.) using different Machine Learning (ML) models such as Logistic Regression (LR), Multi-Layer Perceptron (MLP) and Random Forest (RF). Input features (feature list is found on Table 7.17 of Section 7.9) to the models consist of 21 demographics, some penitentiary features (142 for men and 89 for women, the difference is due to the fact that some penitentiary centers include only men and some include only women), 43 RisCanvi items, 8 RisCanvi risk levels and scores (4 risk levels and 4 risk scores). In order to account for the fact that risk assessment tools are trained on historical data to predict the risk of recidivism in the future, we split into training and test set accordingly. In more detail, we use the cases with releases between 2010 and 2014 for training each model (7,482 cases in men model and 592 for women model) and test the models using cases released in 2015 and 2016 (3,549 cases in men model and 310 in women model). In each model, the test set does not include any cases of the training set, which is why the total size of the training and test sets are smaller than their related total population.

Finally, we ensure that the distribution of treatment cases in training and test set of each model is almost balanced. The percentage of C.R. cases in the training and test set of the model for men is 31% and 36%, respectively and C.R. cases in the training and test sets of the women is 47% and 50%, respectively.

7.5.3 General and Violent Recidivism Prediction

To compute the causal effect of conditional release (C.R.) on general and violent recidivism using the Augmented Inverse Propensity Weighted (AIPW) method, we construct models for both the outcomes (general and violent recidivism within 2 to 5 years of release) and the propensity to be assigned to C.R. Using different ML algorithms (such as LR, MLP, and RF) and the same training and test sets used in C.R. propensity models of men and women, we obtain eight prediction models of general and violent recidivism outcomes within 2 to 5 years of release for each group. The input features of these models are the same features used in the models for the propensity to C.R. plus the actual treatment variable (C.R.).

7.5.4 Average Treatment Effect (ATE)

We compute the Average Treatment Effect (ATE) of C.R. on general and violent recidivism using various causal inference methods. In order to obtain consistent estimates of the causal effect, the following conditions need to hold:

- Stable Unit Treatment Value Assumption (SUTVA) [Angrist et al., 1996]: We assume that the Stable Unit Treatment Value Assumption (SUTVA) holds such that the recidivism risk of a person is unaffected by the particular assignment of C.R. to other cases.
- Common Support (Overlap) [Caliendo and Kopeinig, 2008]: Common support means that there is complete 'overlap' in the distribution of propensity scores across treatment and control cases to find adequate matches. This condition is also satisfied in our study,

which is shown in the propensity score distribution plots (Figure 7.4 and Figure 7.5) in Section 7.6.

- Conditional Independence [Dawid, 1979]: Conditional independence or unconfoundedness requires, that conditional on all confounders used in the model, the assignment of C.R. is random. This condition cannot be tested but the different estimates of the propensity to treatment yield notably high AUC values which can be attributed to the amount and the criminological relevance of the confounders that we use. This can be seen in Section 7.6. This suggests that we include most of the relevant confounders to predict treatment assignment and the two recidivism outcomes.

When these conditions have been fulfilled, then there is ’strong ignorability’ of how an individual came to be treated relative to the outcome [Rosenbaum and Rubin, 1983]. Strong ignorability implies that no systematic, unobserved, pretreatment differences exist between treated and control subjects that are related to the response under study [Joffe and Rosenbaum, 1999].

In the following, we explain the causal inference methods of Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse Propensity Weighting (AIPW) that we use to obtain the Average Treatment Effect of conditional release on general and violent recidivism.

Propensity Score Matching (PSM)

In the method of propensity score matching [Rosenbaum and Rubin, 1983], we sort the data by propensity scores and then stratify it into buckets (four in our case). In our work, we obtain the ATE by subtracting the mean recidivism of non-treated (control) cases from treated ones in each bucket.

Inverse-Propensity Score Weighting (IPW)

The basic idea of this method is weighting the outcome measures by the inverse of the individual’s treatment propensity so that similar baseline

characteristics are obtained [Bray et al., 2019]. In this method, the treatment effect for individual i is obtained using the following equation:

$$TE_i = \frac{W_i Y_i}{p_i} - \frac{(1 - W_i)Y_i}{1 - p_i} \tag{7.1}$$

W_i indicates treatment (1 for treated and 0 for control cases), p_i represents probability of receiving treatment (propensity score of treatment), and Y_i indicates recidivism (1 if recidivate and 0 if not recidivate) for individual i.

The IPW method places more weights on observations from the control group with a high treatment propensity and vice versa for observations in the treatment group, improving on the covariate balance. In other words, the untreated units with higher estimated probability of being treated and the treated units with lower estimated probability of being treated receive higher weights. At last, the model is estimated using data of those that are more similar (thus more comparable) to each other. "Extracting" data on similar observation units mimics natural experiments.

If the propensity scores were known (which is the case here), then this estimator will be unbiased for the ATE [Tsiatis, 2006]. Furthermore, when the propensity scores are estimated consistently, then this estimator is consistent for the ATE. In our study, looking at the propensity score distributions of the treatment and control groups in Section 7.6 (Figure 7.4 and Figure 7.5), we can see the consistency of these estimates especially for men group. The IPW estimator is also widely believed to have poor small sample properties when the propensity score gets close to zero or one for some observations. Specifically, treatment cases with very low propensity scores and control cases with very high propensity scores will provide extreme contributions to the estimate [Glynn and Quinn, 2010]. However, in our study, according to the propensity score distribution of the treatment and control group which is shown in Section 7.6 (Figure 7.4 and Figure 7.5), the percentage of treatment cases with very low propensity scores and control cases with very high propensity scores are very low (less than 1%).

Augmented Inverse-Propensity Weighted (AIPW)

This method combines both the properties of the regression-based estimator and the IPW estimator. It has an augmentation part $(W_i - p_i)\widehat{Y}_i$ to the IPW method, in which $\widehat{Y}_i$ is the estimated probability of recidivism using all features applied to the propensity score model plus the treatment variable. In other words, in this method, two models are used; a binary regression model for the propensity score, and a regression model for the outcome variable. So, this estimator yields doubly robust estimations which requires only either the propensity or outcome model to be correctly specified but not both. Comparing this estimator to IPW and PSM estimators, it is shown that the AIPW estimator has comparable or lower mean square error than the other two estimators. When the propensity score and outcome models are both properly specified and, when one of the models is misspecified, the AIPW estimator is superior [Glynn and Quinn, 2010]. This double-robustness property gives the AIPW estimator a tremendous advantage over most other estimators in that with the AIPW estimator the researcher has more hope of getting a reasonable answer in complicated real-world situations where there is uncertainty about both the treatment assignment process and the outcome model. We can compute the AIPW treatment effect on individual i as:

$$TE_i = \frac{W_i Y_i - (W_i - p_i)\widehat{Y}_i}{p_i} - \frac{(1 - W_i)Y_i - (W_i - p_i)\widehat{Y}_i}{1 - p_i} \tag{7.2}$$

In IPW and AIPW, we obtain the individual treatment effect TE_i, which is the difference between the outcomes if the person is treated (treatment) and not treated (control). In other words, this effect is the difference of recidivism probability when the person is granted C.R. and not granted C.R. A negative value shows a reduced recidivism risk and a positive value indicates an increased recidivism risk. The resulting ATE is the average over all individual treatment effects.

7.5.5 Conditional Average Treatment Effect (CATE)

To determine heterogenous effects of C.R. on recidivism, we estimate the Conditional Average Treatment Effect (CATE) using Generalized Random Forest and Double Machine Learning [Athey et al., 2019, Nie and Wager, 2021, Chernozhukov et al., 2016]. Generalized Random Forests are flexible methods for estimating treatment effect heterogeneity with Random Forests. The specific application of this algorithm to estimate CATE is what researchers call Causal Forests. These estimators are used as final models for CATE estimation within the EconML [Battocchi et al., 2019] package. CATE is the ATE conditioned on membership in a subgroup. Using *econml.dml.CausalForestDML* in the EconML package, we obtain the CATE by considering a single covariate, while keeping all the other covariates at a fixed value (median). The *econml.dml.CausalForestDML* combines a Causal Forest with Double Machine Learning to residualize the treatment and outcome, which again yields doubly robust estimates.

7.6 Results

We observe the predictive performance of the Machine Learning (ML) models introduced in Section 7.5 for the propensity to receive conditional release (C.R.) and general and violent recidivism prediction of men and women groups in section 7.6.1. The computed Average Treatment Effect (ATE) using the three statistical methods of Propensity Score Matching (PSM), Inverse Propensity score Weighting (IPW), and Augmented Inverse Propensity Weighted (AIPW) are presented for both gender groups in section 7.6.2. In section 7.6.3, we compare the obtained ATE values in cases with three different risk levels (high, medium, and low) of the RisCanvi risk assessment tool. Finally, the results of Conditional ATE (CATE) on membership in different subgroups are given in section 7.6.4.

Table 7.8: AUC-ROC of propensity to conditional release (C.R.) prediction. LR stands for logistic regression.

Group	AUC-ROC	Model
Men	0.92	LR
Women	0.89	LR

7.6.1 Predictive Performance of ML models

We find that Logistic Regression yields the most accurate prediction of the propensity to receive conditional release (C.R.) for both men and women. The results are shown in Table 7.8 in terms of AUC-ROC which stands for ”Area under the ROC (Receiver Operating Characteristics) Curve”. This metric is used to measure the performance of the classification models at various threshold settings. ROC is a probability curve with TPR (True Positive Rate) against FPR (False Positive Rate) and AUC measures the entire two-dimensional area underneath the ROC curve. The Higher the AUC, the better the classification model is at distinguishing between positive and negative classes. Obtained estimates from our ML models are well calibrated for both groups (calibration curves omitted for brevity). According to the AUC values in Table 7.8, ML models show accurate results for both groups especially for men. The models will be used in the computation of the ATE. We can also observe the distribution of the C.R. propensity scores in treatment and control groups of men and women in Figure 7.4 and Figure 7.5. As can be seen, for both men and women there is an overlap in the distribution of the propensity scores of treatment and control cases to find adequate matches. This is a necessary condition to be able to apply our causal inference methods. Also, for both groups the distributions are well spread between 0 and 1.

We can also observe the predictive performance of the ML-based models of general and violent recidivism within 2 to 5 years of release in terms of AUC-ROC in Table 7.9. The algorithm used for all models is Random Forest with n_estimators=500 and max_depth=2. As can be seen, all models show high AUC for both risk outcomes and in both groups

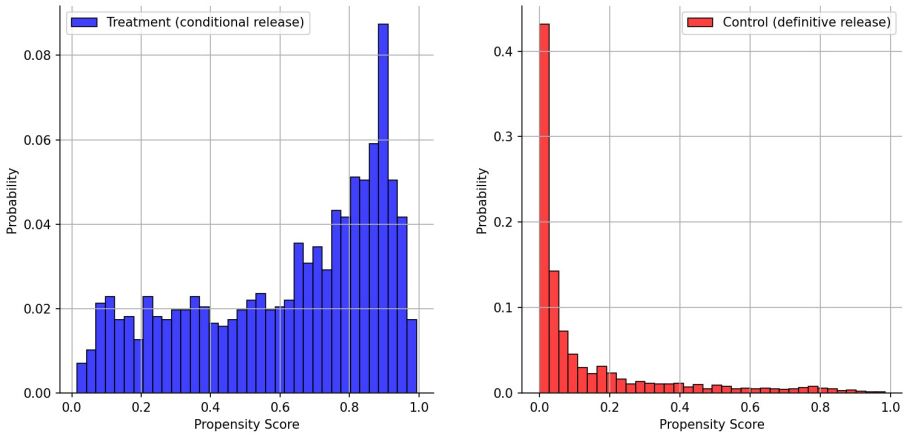


Figure 7.4: Distribution of the propensity to treatment (C.R.) for **men** in our sample

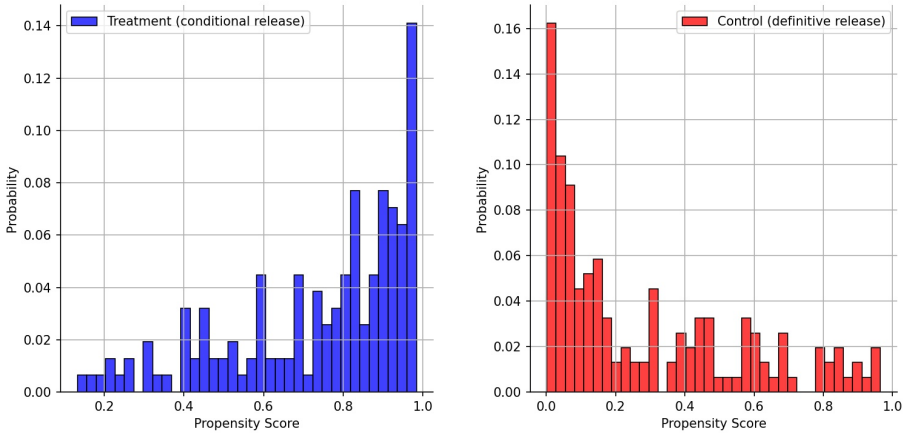


Figure 7.5: Distribution of the propensity to treatment (C.R.) for **women** in our sample

Table 7.9: AUC-ROC of general and violent recidivism prediction using Random Forests

Group	Recidivism within	2 years	3 years	4 years	5 years
Men	General	0.77	0.75	0.75	0.74
	Violent	0.80	0.80	0.78	0.77
Women	General	0.87	0.86	0.85	0.84
	Violent	0.83	0.81	0.78	0.79

especially for women. In general recidivism prediction, the AUC value is 0.74-0.77 and 0.84-0.87 for men and women groups respectively. The AUC results for violent recidivism prediction show values of 0.77-0.80 and 0.78-0.83 for men and women respectively. These outcome models will be used in the computation of ATE in the AIPW method.

7.6.2 Average Treatment Effect (ATE)

Our goal is to determine whether conditional release (C.R.) has a causal effect on general and violent recidivism within 2 to 5 years of a person's release. The Average Treatment Effect (ATE) obtained using propensity score matching is shown in Table 7.10 and Table 7.11 for men and women respectively.

A negative ATE indicates a reduction in the probability of recidivism due to the treatment (C.R.). For men with lower propensity to receive C.R., we can observe a more negative ATE of C.R. on both general and violent recidivism within all follow-up periods. This means that if men with low probability of getting C.R. (high risk cases) have chances to receive C.R., their general and violent recidivism probability within 5 years of their release would be respectively 34 percentage points and 11 percentage points lower than if they would not receive C.R.

Table 7.10: ATE obtained for **men** using Propensity Score Matching with four buckets and for all. ATE on general recidivism is in the columns marked "gen", and on violent recidivism is in the columns marked "vio". Negative numbers indicate that the probability of recidivism of those who treated (i.e., with C.R.) is lower.

Propensity	P(T|X)=low		P(T|X)=med-low		P(T|X)=med-high		P(T|X)=high		All	
Treatment size	0.2%		7.6%		32.8%		59.3%		35.8%	
ATE of C.R. on Recidivism type:	gen	vio	gen	vio	gen	vio	gen	vio	gen	vio
within 2 years	-0.21	-0.07	-0.08	-0.02	-0.04	-0.01	-0.02	-0.01	-0.11	-0.04
within 3 years	-0.26	-0.09	-0.11	-0.03	-0.05	-0.02	-0.02	-0.01	-0.14	-0.05
within 4 years	-0.31	-0.10	-0.12	-0.03	-0.05	-0.01	-0.01	-0.01	-0.16	-0.05
within 5 years	-0.34	-0.11	-0.14	-0.03	-0.05	-0.01	-0.03	-0.01	-0.17	-0.05

Table 7.11: ATE obtained for **women** using Propensity Score Matching with four buckets and for all

Propensity	P(T|X)=low		P(T|X)=med-low		P(T|X)=med-high		P(T|X)=high		All	
Treatment size	0.6%		19.9%		34.6%		44.9%		50.3%	
ATE of C.R. on Recidivism type:	gen	vio	gen	vio	gen	vio	gen	vio	gen	vio
within 2 years	-0.24	-0.05	-0.02	0.00	-0.09	0.00	0.00	0.00	-0.13	-0.03
within 3 years	-0.29	-0.07	0.01	0.00	-0.14	0.00	0.00	0.00	-0.16	-0.03
within 4 years	-0.33	-0.07	-0.03	-0.02	-0.14	0.00	0.00	0.00	-0.19	-0.04
within 5 years	-0.37	-0.08	-0.03	-0.02	-0.18	0.00	0.01	0.00	-0.21	-0.05

For women, the effects are similar for the group with the lowest propensity to receive C.R. but slightly different from the effects for men for the higher propensity groups. For violent recidivism of women, we find a stronger ATE in buckets with lower propensity to receive C.R. but in buckets with medium-high and high C.R. probability there is no significant effect of C.R. on violent recidivism. However, considering all cases together, the ATE of C.R. on general and violent recidivism (within all follow-up periods) is negative for both men and women.

The ATEs computed using Inverse-Propensity score Weighting (IPW) and Augmented Inverse-Propensity Weighting (AIPW) methods are shown in Table 7.12 and Table 7.13 for men and women respectively. As results show, all confidence intervals in ATE values obtained from the IPW method lie entirely in the negative region for general and violent recidivism of both men and women. This is a strong indication that C.R. reduces the risk of violent and general recidivism for men and women within 2 to 5 years of their release. We can also observe negative confidence intervals in all ATE values obtained from the AIPW method for general recidivism risk of men and women, which shows that C.R. reduces general recidivism risk for both groups within 2 to 5 years of their release. In the ATE values obtained from AIPW method for violent recidivism, negative confidence intervals are found for men within 3 years and women within 5 years of their release. However, in the AIPW results obtained for other follow-up periods of the two groups, which are shown in italics, the confidence intervals contain the value zero, from which we cannot establish whether there is a change in the violent recidivism risk due to C.R.

7.6.3 ATE by Risk Level

In this section, we explore the heterogeneity of the computed ATEs by three different risk levels (high, medium, and low) of violent recidivism risk (REVI risk) as obtained from the RisCanvi risk assessment tool. In Table 7.14 violent recidivism base rates are shown for different REVI risk levels. As can bee seen, for men there is a clear correlation between the base rates and the RisCanvi risk levels in all follow-up periods which

means that the estimated REVI risk level by RisCanvi is consistent with the violent recidivism rates within 2 to 5 years of release. This correlation is not as clear for women due to small sample size. Risk level for women is almost always ”low risk” in our sample so that 286 women have low risk, but only 17 have medium risk, and only 7 have high risk, which makes statistics relating REVI risk and recidivism unreliable. Hence, we only compare ATE values by different REVI risk levels for men.

On Table 7.15 and Table 7.16, the ATE of C.R. on violent recidivism of men (within 2 to 5 years of release) obtained respectively from IPW and AIPW is shown for three different REVI risk levels. ATE values with confidence intervals consisting value zero are not reliable and shown in italics. Comparing other ATE values (with confidence intervals not including value zeros), we can see the most negative ATE of C.R. on violent recidivism in cases with medium REVI risk level in both IPW and AIPW results. These results show that granting C.R. to men with higher REVI risk (medium) yields a stronger reduction in violent recidivism risk compared to granting C.R. only to the cases with a low REVI risk level. According to these results, we note that the risk estimated by a risk assessment tool should not be linked to treatment (C.R.) of a case. By dedicating more resources toward higher risk detected cases than lower risk ones the community protection can be effectively promoted [Hanson, 2005].

7.6.4 Conditional Average Treatment Effect (CATE)

We measured Conditional Average Treatment Effect (CATE) in different groupings according to base crime, and criminological features. This is important because it would be relevant to know if granting C.R. to specific groups could yield a stronger reduction in recidivism compared to other groups. However, we found no differences worth reporting (mostly one percentage point or less), and in all groups we studied the effect of C.R. is a reduction of recidivism, and the reduction is similar in magnitude.

Table 7.12: IPW and AIPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. Almost all of the confidence intervals, with the exception of those in italics, lie entirely in the negative region.

MEN	IPW on general recidivism			AIPW on general recidivism			IPW on violent recidivism			AIPW on violent recidivism		
	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.09	-0.08	-0.06	-0.06	-0.04	-0.02	-0.03	-0.02	-0.01	*-0.02*	*-0.01*	*0.002*
within 3 years	-0.11	-0.09	-0.07	-0.08	-0.05	-0.03	-0.04	-0.03	-0.02	-0.03	-0.02	-0.003
within 4 years	-0.13	-0.10	-0.07	-0.08	-0.06	-0.03	-0.04	-0.03	-0.01	*-0.03*	*-0.01*	*0.01*
within 5 years	-0.14	-0.11	-0.08	-0.09	-0.06	-0.03	-0.04	-0.02	-0.002	*-0.03*	*-0.01*	*0.01*

"output" — 2022/9/20 — 11:46 — page 147 — #173

Table 7.13: IPW and AIPW results estimating ATE and its 95% confidence interval [lower-ci, upper-ci]. Almost all of the confidence intervals, with the exception of those in italics, lie entirely in the negative region.

WOMEN	IPW on general recidivism			AIPW on general recidivism			IPW on violent recidivism			AIPW on violent recidivism		
	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
within 2 years	-0.12	-0.08	-0.04	-0.10	-0.06	-0.02	-0.03	-0.01	-0.0003	*-0.02*	*-0.01*	*0.004*
within 3 years	-0.16	-0.10	-0.05	-0.14	-0.08	-0.02	-0.03	-0.02	-0.002	*-0.03*	*-0.01*	*0.002*
within 4 years	-0.18	-0.12	-0.07	-0.16	-0.10	-0.04	-0.04	-0.02	-0.004	*-0.04*	*-0.02*	*0.001*
within 5 years	-0.20	-0.14	-0.08	-0.18	-0.12	-0.05	-0.05	-0.03	-0.01	-0.04	-0.02	-0.001

Table 7.14: Violent recidivism base rates per REVI level. Violent recidivism probability is higher for men having higher REVI risk assessments. Result can not be established for women due to the small sample size.

			Violent recidivism within:							
	Size		2 years		3 years		4 years		5 years	
REVI level	men	women	men	women	men	women	men	women	men	women
Low	2,594	286	1%	1%	2%	1%	2%	1%	3%	2%
Medium	609	17	6%	6%	8%	6%	10%	6%	10%	6%
High	346	7	9%	0%	11%	14%	12%	14%	13%	14%

Table 7.15: ATE-IPW of C.R. on violent recidivism and its 95% confidence interval [lower-ci, upper-ci] in different REVI risk levels of **men**. Confidence intervals shown in italics contain zero value and are not reliable. Granting C.R. to men with medium REVI risk yields a stronger reduction in violent recidivism risk compared to low REVI risk cases.

	IPW on violent recidivism within:											
	2 years			3 years			4 years			5 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (73%)	-0.02	-0.01	-0.002	-0.02	-0.01	-0.01	*-0.03*	*-0.003*	*0.02*	*-0.02*	*0.001*	*0.02*
Medium (17%)	-0.08	-0.06	-0.04	-0.10	-0.08	-0.05	-0.12	-0.09	-0.06	-0.12	-0.09	-0.06
High (10%)	*-0.14*	*-0.04*	*0.06*	*-0.16*	*-0.06*	*0.04*	*-0.17*	*-0.07*	*0.03*	*-0.18*	*-0.08*	*0.03*

Table 7.16: ATE-AIPW of C.R. on violent recidivism and its 95% confidence interval [lower-ci, upper-ci] in different REVI risk levels of **men**. Confidence intervals shown in italics contain zero value and are not reliable. Granting C.R. to men with medium REVI risk yields a stronger reduction in violent recidivism risk compared to low REVI risk cases.

	AIPW on violent recidivism within:											
	2 years			3 years			4 years			5 years		
REVI level	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci	lo-ci	ATE	up-ci
Low (73%)	-0.02	-0.01	-0.003	-0.03	-0.02	-0.01	*-0.03*	*-0.005*	*0.02*	*-0.02*	*-0.001*	*0.02*
Medium (17%)	-0.05	-0.03	-0.01	-0.07	-0.04	-0.01	-0.09	-0.05	-0.02	-0.08	-0.05	-0.02
High (10%)	*-0.05*	*0.04*	*0.13*	*-0.06*	*0.03*	*0.12*	*-0.06*	*0.03*	*0.12*	*-0.06*	*0.03*	*0.12*

"output" — 2022/9/20 — 11:46 — page 151 — #177

7.7 Discussion and Conclusion

In this paper we studied the effect of conditional release (C.R.) on violent and general recidivism of persons who were released from several prison centers in a European country between 2010 and 2016. Due to noticeable differences in men and women in our dataset with respect to some penitentiary features and the performance of the predictive models, we treated them differently by creating separate ML models for these two groups [Skeem et al., 2016, Collins, 2010, Huebner et al., 2010].

We computed the Average Treatment Effect (ATE) of C.R. on general and violent recidivism of men and women using statistical causal inference methods such as Propensity Score Matching (PSM) [Rosenbaum and Rubin, 1983], Inverse-Propensity score Weighting (IPW) [Bray et al., 2019], and Augmented Inverse-Propensity Weighted (AIPW) [Glynn and Quinn, 2010] methods. These methods require a precise prediction of the propensity to receive treatment (C.R.) and the probability of the studied outcome (recidivism). For both the treatment propensity and the outcome probability we obtain high predictive performance in terms of AUC. This suggests that our data explains most of the variations in treatment and outcome which supports our identification strategy. The obtained ATE values from all of the methods mostly show that C.R. reduces the risks of violent and general recidivism of men and women within 2 to 5 years of their release.

We compared ATE values of men with 3 different risk levels of violent recidivism (REVI risk) estimated in RisCanvi risk assessment tool. Comparison could not be established for women due to the small sample size. The comparison showed that granting C.R. to men with medium REVI risk can be more effective in reducing their violent recidivism probability compared to granting C.R. to the cases with low REVI risk level.

A recommendation that these results suggest is that the risk estimated by a risk assessment tool should not be the only basis for granting treatment (C.R.). In fact, our results show that granting C.R. to higher risk detected cases can yield improvements in community safety by reducing overall recidivism rates. However, risk assessment, as currently used, mainly serves

as a motivation to grant C.R. to low-risk incarcerated persons. This usage has two main problems: First, it assumes risk is static, but according to the ”third generation” of risk assessment tools, we should address dynamic factors that can be changed to reduce risk. Instead of determining risk, we should move towards needs assessment and intervention, based on the risk-need-responsivity (RNR) principle, and look at what needs an individual has that can be met to reduce their risk [Bonta, 1996, Barabas et al., 2018]. According to RNR principle, to achieve effective rehabilitation, risk instruments have to be evidence-based and level of rehabilitation service should go with the level of risk, type of criminogenic need, and learning style and motivations (responsivity) of the individual being treated [Bonta and Andrews, 2007]. Second, even if it is used to determine risk, it is unclear that the best for society is to grant C.R. only to low-risk cases, as a robust conclusion from our analysis is that C.R. greatly reduces the chances of recidivism for higher risk cases. This may seem to contradict the literature related to risk estimates for flight risk [Kleinberg et al., 2018], which uses such estimates to grant bail to low-risk defendants. However, pre-trial and C.R. applications of risk assessment instruments, which tend to be considered as two analogous settings by computer scientists, should not be treated in the same way. Hence, we believe that the connection between risk assessment and C.R. requires a deep examination in the light of these results.

The originality of this contribution is using different causal inference methods to calculate the effects of C.R. on two types of general and violent recidivism in terms of ATE and looking at this causal effect in relation to different estimated risk levels of a risk assessment tool. This causal inference study for C.R. application provides a path towards effectively supporting incarcerated persons, less incarceration, and prison systems with capacity of C.R. programs in which cost of C.R. is lower than the cost of incarceration. Also, causal inference methods such as the ones we used allow to perform observational studies, as criminal justice is a domain in which some types of direct experimentation might be unethical or harmful. We also used a large dataset and our results hold across substantially diverse prison centers. We stress that the methodology we described is

broadly applicable. Our findings are likely to be specific to this particular dataset, but show the general effectiveness of the methodology in this setting.

7.8 RisCanvi Items Imputation

The number of items in RisCanvi-S (10 items) is less than RisCanvi-C (43 items). However, 6 items of RisCanvi-S match 6 items in RisCanvi-C and the remaining 4 items are combinations of other RisCanvi-C items. To have 43 numbers of items, which is also more informative, in cases with only RisCanvi-S evaluation (these are low-risk cases), we imputed the 33 remaining items using low risk values of RisCanvi-C dynamic items and values of penitentiary or demographic features in case of static items. For the cases with RisCanvi-C as the latest valid evaluation (which is at most 9 months before the release date), if there is a valid RisCanvi-S evaluation before that, we use its non-empty items to impute the missing items in the RisCanvi-C.

7.9 Features

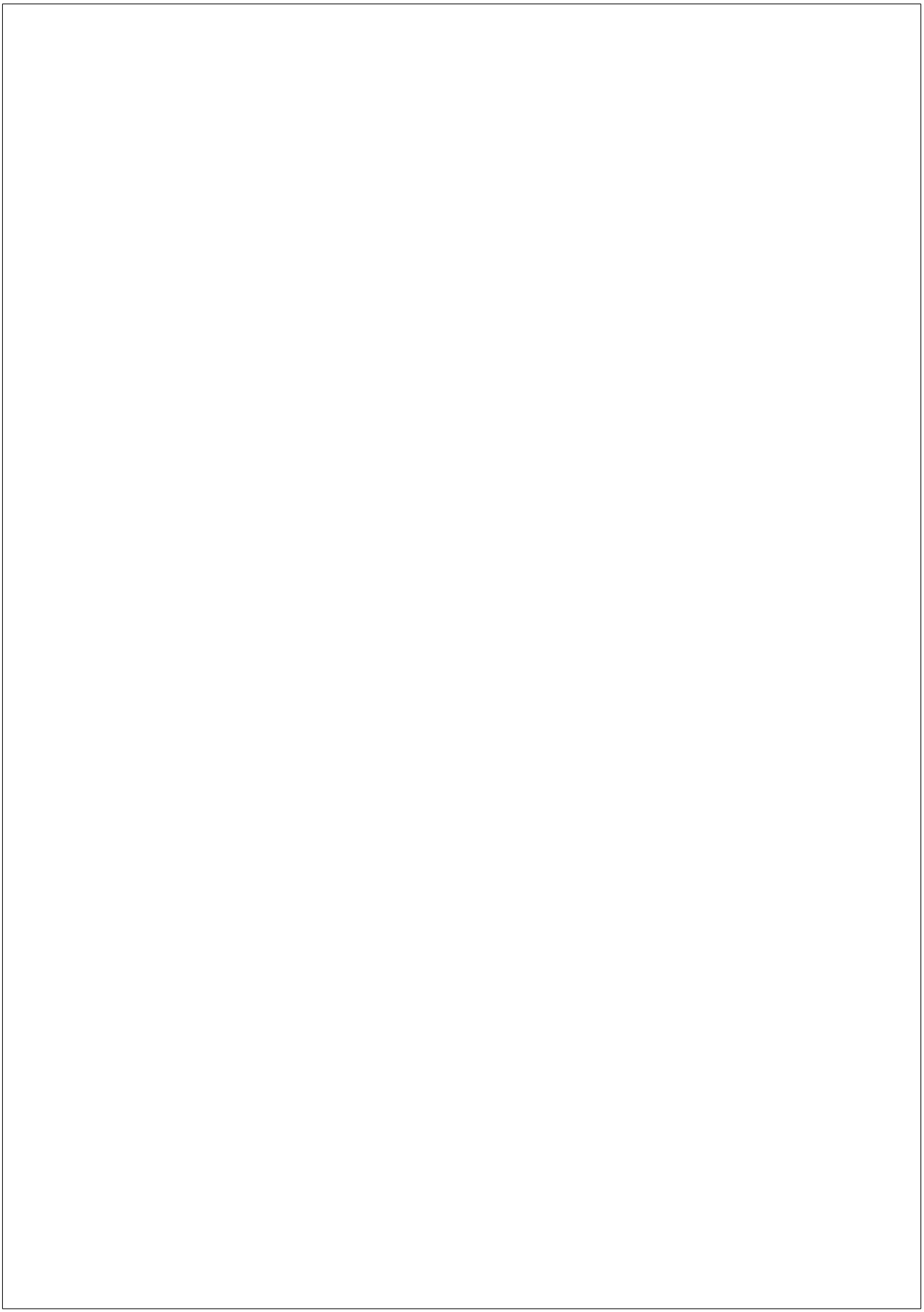
List of features used in this study is shown in three categories of demographics, penitentiary features, and RisCanvi features on Table 7.17.

[l]"Civil Liability" (CL) is a monetary compensation imposed in addition to time in jail. There are two basic cases: civil liability (with CL) and no civil liability (without CL). The former is further divided into sub-classes including whether civil liability was paid in full (fully-paid CL), not paid in full (not fully-paid CL), or the person declared him/herself unable to pay."

[2]Against people, Gender-based violence, Against sexual freedom, Against property (violent), Against property (non-violent), Drugs, Traffic, and others.

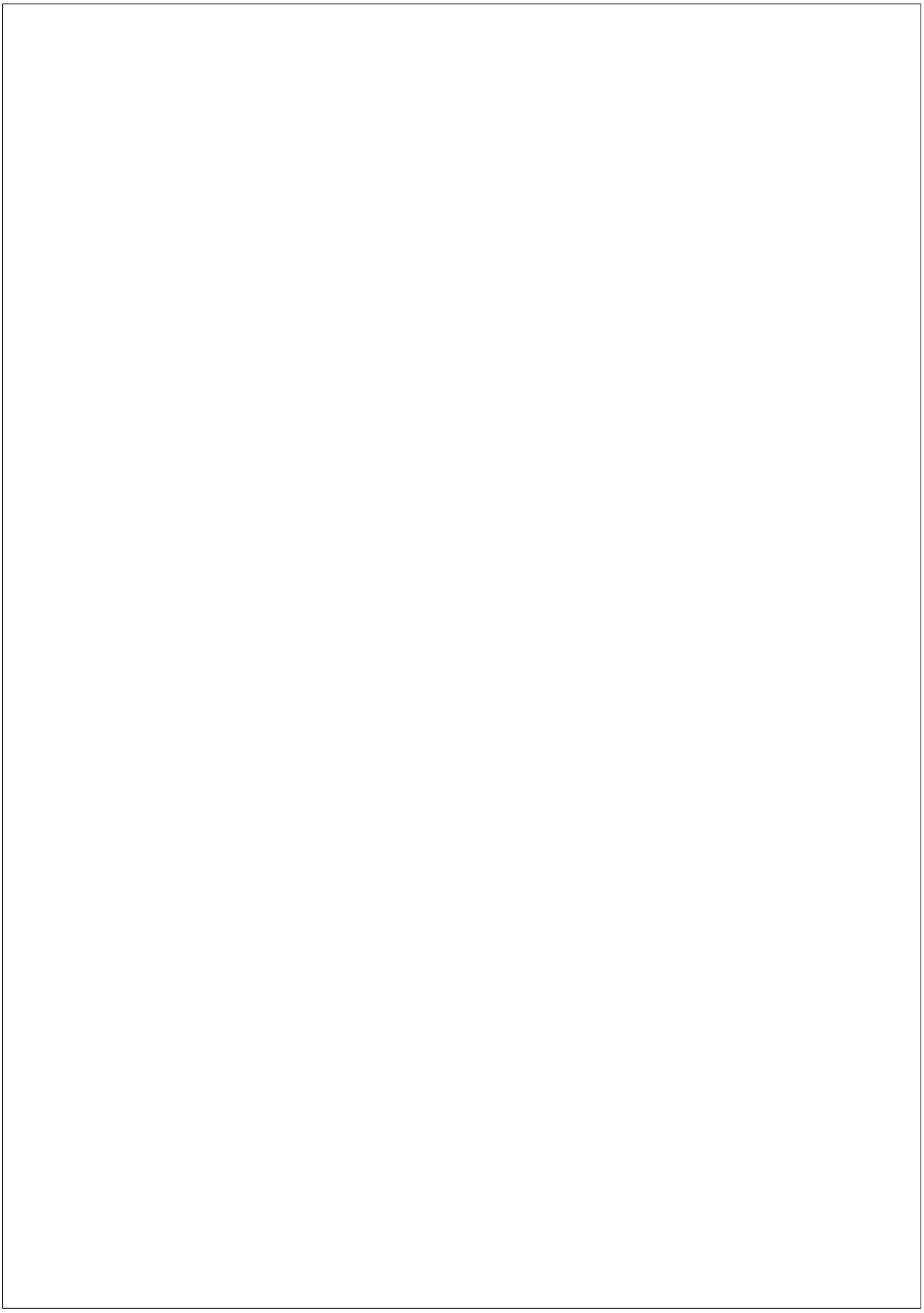
Table 7.17: List of features. “Y/N” are boolean features, and “Num” are numerical features.

Demographics		Penitentiary features		RisCanvi features	
Feature	Value	Feature	Value	Feature	Value
Age at release	Num	Prison center {1,...,87}	Y/N	Items {1,...,43} (Table 3.1)	Num
Male	Y/N	No permission request	Y/N	Risk level in self-directed violence	Num
Female	Y/N	Permission rejection	Y/N	Risk level in violence in the prison facilities	Num
Country birth	Y/N	Enjoyed permission	Y/N	Violent recidivism risk level	Num
Country residence	Y/N	Not enjoyed accepted permission	Y/N	Risk level in breaking prison permits	Num
Country nationality	Y/N	Num. of enjoyed permissions	Num	Risk score in self-directed violence	Num
Single	Y/N	Num. of rejected permissions	Num	Risk score in violence in the prison facilities	Num
Married	Y/N	Activity	Y/N	Violent recidivism risk score	Num
Divorced	Y/N	Num. of activities	Num	Risk score in breaking prison permits	Num
Separated	Y/N	Num. of module changes	Num		
Widow	Y/N	Num. of nursing modules	Num		
Deported	Y/N	Num. of psychiatry modules	Num		
Country language 1	Y/N	Num. of special supervision module	Num		
Country language 2	Y/N	Num. of regressions to 1st degree	Num		
Education level	Num	Num. of regressions to 2nd degree	Num		
Nationality {1,2,...,8}	Y/N	Num. of progresses to 2nd degree	Num		
		Num. of progresses to 3rd degree	Num		
		Num. of degree change	Num		
		Mostly degree regression	Y/N		
		Mostly degree progress	Y/N		
		Num. of degree regression	Num		
		Num. of degree progress	Num		
		First prison degree	Num		
		Last prison degree	Num		
		Degree evolution {0,...,4}	Y/N		
		Light rules violations	Y/N		
		Severe rules violations	Y/N		
		Very severe rules violations	Y/N		
		Num. of light rules violations	Num		
		Num. of severe rules violations	Num		
		Num. of very severe rules violations	Num		
		Violent base crime	Y/N		
		With CL[1]	Y/N		
		Without CL	Y/N		
		Fully-paid CL	Y/N		
		Not fully-paid CL	Y/N		
		Unable to pay CL	Y/N		
		Evaluation points	Num		
		Previous prison entries	Y/N		
		Num. of previous prison entries	Num		
		Sentence day	Num		
		Sentence duration class	Num		
		Penalty admission type {1,2,3}	Y/N		
		Base crime types {1,...,8}[2]	Y/N		
		With electronic surveillance	Y/N		
		In dependent units	Y/N		



Part IV

Conclusions



Chapter 8

CONCLUSIONS

In this thesis, we discussed and addressed some challenges of structured risk assessment systems including predictive accuracy, algorithmic bias, and effectiveness of these systems. We mainly focused on risks in two applications, recidivism risk in criminal justice and dropout risk in higher education. However, our methodologies can also be applied to address other types of risks in different application areas.

Considering the important relation of risk change to recidivism of offenders [Labrecque et al., 2014], we tried to improve the efficiency of RisCanvi tool which is an in-use risk assessment instrument in Catalonia. The studied dataset consists of 2,634 offenders who were evaluated with the RisCanvi protocol between 2010 and 2013. Normally, each inmate is evaluated by RisCanvi every six months. We used Machine Learning (ML) models to select the inmates for the next evaluation of the Violent Recidivism (REVI) risk in the RisCanvi protocol. In the selection process, only a fraction of the inmates, those with the highest probability of having changed risk, are chosen for the next evaluation. ML models showed good results in terms of AUC (0.74-0.78) and resulted in fewer evaluations per inmate compared to the standard RisCanvi (the number of evaluations is halved compared to RisCanvi), which in turn leads to save time, expenses and staff in the evaluations. This benefit has been obtained in exchange for the cost of a relatively small number of missed/undetected risk changes.

Using this method can help free staff time for programs which are more focused on reducing the likelihood of recidivism instead of merely predicting it, an idea that has been supported by researchers of current ML-based risk assessments [Barabas et al., 2017]. Furthermore, we analyzed if the ML models lead to discriminatory outcomes across nationality and age groups. In terms of AUC, the models show more accuracy for foreigners than Spanish nationals and there is no significant difference in age subgroups. In terms of missed changes (false negative rates), there is a small disparate mistreatment among both nationality and age sub-groups. There is a disparate impact in the average number of evaluations which shows lower number of evaluations in foreigners and older inmates on average compared to their counterparts. So we applied a mitigation method by adjusting decision boundaries and obtained equality in the rate of evaluation along both nationality and age with a small additional loss of missed changes.

We also studied the predictive accuracy and fairness of ML models compared to RisCanvi tool in predicting recidivism risk within two years of inmate's release. The data under study includes 2,027 inmates who were evaluated with the RisCanvi protocol between 2010 and 2013. The created ML models show AUC of 0.76 and 0.73 in violent and general recidivism prediction respectively which is slightly better result compared to the AUC of RisCanvi protocol which is 0.72 and 0.70. This shows that a hand-crafted formula created by RisCanvi experts is quite comparable to a machine-learned one. However, a key element of ML models is their flexibility as they can be re-trained with newer data, and incorporate new factors as the population of inmates changes and more data on recidivism becomes available. Studying differential outcomes of RisCanvi and ML models across some demographics show different but comparable results in both models, depending on the desired metric and studied group. However, an advantage of ML models is that as policy changes are introduced, the emphasis on different metrics can be changed during the modeling. The results show error disparity in terms of generalized false positive rate (GFPR) in both violent and general recidivism predictions along nationality and age groups. Using a bias mitigation method [Pleiss et al., 2017], we

could satisfy equalized odds and decrease GFPR disparity in both violent and general recidivism predictions, while preserving calibration in each sub-group of nationality and age. However, this is obtained in exchange for inequalities in some other metrics which can be acceptable based on the application, otherwise, bias mitigation can be applied in terms of a more critical metric for that application. Eventually, it should be noted that improving the predictive accuracy of ML models needs to be carefully contrasted with potential issues of algorithmic fairness, and bias mitigation methods have to be used.

We also addressed the effectiveness of ML models in the early prediction of university dropout and underperformance from a perspective of algorithmic fairness. The studied population consists of 667 computer engineering undergraduate students from Universitat Pompeu Fabra (UPF) who first enrolled between 2009 and 2017. We created calibrated ML models with AUC of 0.77 and 0.78 using only information at the time of enrollment. The models help reliably identify students at risk to trigger interventions that can help increase their success and ultimately reduce social and economic costs. We analyzed the discriminatory outcomes of the Ml models across some sensitive groups defined by nationality, gender, high school type and location, and admission grade. Our models show parity in terms of AUC and GFNR but disparities in GFPR which are larger among groups defined by admission grade, and the bias is against students with lower admission grades. By using the relaxation method [Pleiss et al., 2017], we could obtain a perfect parity in GFPR across most of the groups while preserving calibration. This bias mitigation also caused increases in parity along other metrics (AUC and GFNR) along majority of the groups compared to the non-mitigated model.

We expanded our study in the higher education field by evaluating the risk of early dropout using causal inference methods, and focusing on groups of students who have a relatively higher dropout risk. The population under study are 23,096 undergraduate students who enrolled between 2009 to 2018 to 21 different study programs offered by eight academic centers in UPF. We created ML models with AUC values of 0.70 and 0.74 in predicting dropout and underperformance risks respectively

using only information available at the time of enrollment. We found that workload (first year credits) is an important driver of dropout and it is a feature over which the first-year students have some control. Also, comparing dropout risk across various groups of students showed that to a large extent there is a higher probability of dropout in older students, in students taking a higher workload, and in students admitted through study access types III and IV. Using a combination of these features, we considered three scenarios in which interventions were designed having the common characteristic of a reduced workload for students. In each scenario, the propensity score of the treatment was obtained with AUC-ROC of 0.75-0.91 using ML-based models. Then, for each scenario, the Average Treatment Effect (ATE) on dropout was computed using causal inference methods. The results suggest a reduction of risk of dropout, following a lower number of credits taken on the first year. This result recommends to ask at-risk students to consider taking a reduced workload or asking educational policy makers to revise the regulations that establish the minimum number of credits. Some important aspects of this contribution is focusing on vulnerable groups of students prone to dropout and effectively supporting them, studying combinations of different features (such as workload, age, and study access type), applying different causal inference models to calculate ATE, and using a large dataset among diverse study programs.

We also studied the causal effect of conditional release (C.R.) on violent and general recidivism (within 2 to 5 years of release) of 22,726 persons who were released from several prison centers in Catalonia between 2010 and 2016. We studied men and women separately due to noticeable differences in their profiles. Using statistical causal inference methods, we computed the Average Treatment Effect (ATE) of C.R. on general and violent recidivism of men and women. These methods require a precise prediction of the propensity to receive treatment (C.R.) and the probability of the studied outcome (recidivism). We obtained ML models with high predictive performance for both treatment propensity (AUC-ROC values of 0.92 and 0.89 for men and women respectively) and the outcome probability (AUC of 0.74-0.77 and 0.84-0.87 in general

recidivism of men and women respectively and AUC of 0.77-0.80 and 0.78-0.83 in violent recidivism of men and women respectively). This high performance of the two models shows that most of the variations in treatment and outcome are explained in our data, which supports our identification strategy. Then obtaining the ATE values from causal inference methods shows that C.R. reduces the risks of violent and general recidivism of men and women within 2 to 5 years of their release. We also compared the ATE values in cases with 3 different risk levels of violent recidivism (REVI risk) estimated in RisCanvi risk assessment tool. The results show that granting C.R. to men with medium REVI risk can be more effective in reducing their violent recidivism probability compared to granting C.R. to the cases with low REVI risk level. Comparison could not be established for women due to the small sample size. These results suggest that the risk estimated by a risk assessment tool should not be the only basis for granting treatment (C.R.). However, currently used risk assessment mainly serves as a motivation to grant C.R. to low-risk inmates, while our results show that granting C.R. to higher risk detected offenders can yield improvements in community safety by reducing overall recidivism rates. This causal inference study on C.R. application provides a path towards effectively supporting incarcerated persons, less incarceration, and prison systems with capacity of C.R. programs in which cost of C.R. is lower than the cost of incarceration. Also, our results in this study hold across a large dataset consisting various prison centers.

Some important elements are worth mentioning looking at our studies in this thesis. Risk assessment related to human behavior is a delicate task. In such assessments, it is hard to improve the predictive performance beyond a certain point, hence, human intervention might be necessary. Also, algorithmic discrimination tends to happen along some dimensions, not along others, so it is hard to anticipate, but can be mitigated. Since maximizing fairness and accuracy at the same time is impossible, there is always a trade off in which improving each of them requires careful controlling to avoid worsening the other. Quantifying algorithmic discrimination is a highly context-sensitive task and requires to choose metrics carefully in a task-specific manner.

In some application areas such as criminal justice [Bonta and Andrews, 2007] and clinical decisions [Wand, 2011, Ryan et al., 2010], risk assessment should be cautiously used in relation to the treatment assignment. Hence, the connection between risk assessment and treatment requires a deep examination and accurate measurement in the light of results obtained in this thesis. Causal inference methods are applicable in such frameworks which do not assume a fixed future for a person. These methods can help understand how risk can be changed, and hence design effective interventions. This is consistent with the ”third generation” of risk assessment tools in which dynamic factors that can be changed to reduce risk are addressed, as well in line with the risk-need-responsivity (RNR) principle in which needs assessment and intervention are more focused than risk determination [Bonta, 1996, Barabas et al., 2018, Bonta and Andrews, 2007]. In the end, the objective of case workers is not to predict the future, but to reduce the risk, so we should move towards needs assessment and interventions to reduce risk, instead of solely determining risk.

In structured risk assessment, it is also noteworthy to focus on vulnerable groups who are prone to risk. Such groups can be identified from single variable or combinations of different features. Also, considering separate analysis and modeling for groups with significant differences can lead to more reliable and effective results compared to considering a global behavioural model for all population. More generalized results can be obtained from larger dataset and by applying different methods. We emphasize that the methodologies we described in this thesis are broadly applicable. Although our findings are likely to be specific to particular dataset, they show the general effectiveness of the methodologies used in this thesis. Observational studies, such as the ones we did in criminal justice and higher education areas, are appropriate for domains in which direct experimentation might be unethical or harmful.

Chapter 9

LIMITATIONS

In this thesis, we only examined risk assessment in two application areas. However, our methodologies may be applicable to assessment of other types of risk in different domains.

The study we applied to achieve fewer RisCanvi evaluations is validated on cases who have four evaluations and spend on average two years (or more) in prison. So in criminal risk assessment, this study might not be applicable for people receiving shorter sentences and in other domains, depending on the application, it might not be applied to people with less than four or three evaluations.

In algorithmic fairness studies, we had limited data in terms of cases and features in both criminal justice and higher education systems. Also, since some of algorithmic fairness definitions are incompatible with each other, it was impossible to satisfy all of them concurrently, so we tried to satisfy equalized odds while preserve calibration.

In causal inference studies, among conditions which are needed to be satisfied in order to obtain consistent estimates of the causal effect, conditional independence or unconfoundedness cannot be tested. This condition requires that, conditional on all confounders used in the model, the assignment of treatment is random. However, in our two causal inference studies (Chapter 6 and Chapter 7), the high AUC values we obtained in predicting treatment assignment and risk outcomes can be attributed to the

high amount and relevance of the confounders we used.

Possibly there are many blind spots we tried to address them by consulting with criminal justice domain experts from CEJFE and Department of Justice and TIDE education group at UPF during the thesis.

Chapter 10

FUTURE WORK

In applications with recurring data-driven risk assessment, we can reduce the number of assessments in a balance way of costs and benefits by predicting risk change, decrease, or increase. In this regard, the methodology we used for predicting violent recidivism risk change can be applied to many similar domains in which periodic appraisals are needed to be done by professionals, potentially with the assistance of an algorithm, such as education, immigration, information security, and public health. Also, risk score change, which is an important feature in risk prediction models can be easily utilized in such applications to obtain more accurate estimates. In addition, seeking to reduce the number of assessments has to be accompanied with algorithmic fairness considerations and bias mitigation procedures.

Algorithmic fairness examinations vary in terms of different metrics, so fairness improvements have to be performed based on more critical metrics depending on the application. Therefore, in addition to our methodology which is based on satisfying equalized odds while preserving calibration, other bias mitigation methods can be used to achieve predictive equality [Corbett-Davies et al., 2017] or statistical parity [Dwork et al., 2012].

In our causal inference studies, more scenarios can be defined to design the intervention using other features or different combinations of features. In criminal risk assessment, for instance being classified in the 3rd prison

degree (cases who must spend 8 hours in prison every day but can be outside 16 hours per day) can be considered as a treatment program. In higher education, other combinations of the three important features in dropout prediction can also be considered as a treatment variable, for example older students with study access types I and II who have less credits than median plus all younger students can be considered as a treated group and others can be regarded as a control group. In addition to statistical causal inference methods we used in this thesis, other causal inference methods can also be used to measure the effect of a treatment on an outcome risk such as as instrumental variables [Angrist et al., 1996] and regression discontinuity [Thistlethwaite and Campbell, 1960].

Other risks can also be investigated by the methodologies we used in this thesis, including sexual recidivism, domestic violence risk, flight risk, self-directed violence, violence to other inmates or prison staff, and breaking prison permits in criminal justice system, as well as students high or low performance in school, college or higher education system.

In algorithmic fairness studies, we can also focus on bias in socioeconomic factors such as education, employment, income and housing [Van Eijk, 2017].

Access to more features may lead to more accurate results and help finding more important predictors. In criminal risk assessment, penal and penitentiary features are significant variables in recidivism risk prediction, as using them in the causal study of Chapter 7 resulted in high AUC values. In the higher education domain, students’ interactions via the educational platform they use (e.g., Moodle) can be considered as additional features to the students’ performance prediction model, which may lead to more accurate results.

Bibliography

[Ægisdóttir et al., 2006] Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3):341–382.

[Alberts and Dorofee, 2003] Alberts, C. J. and Dorofee, A. J. (2003). *Managing information security risks: the OCTAVE approach*. Addison-Wesley Professional.

[Albreiki et al., 2021] Albreiki, B., Zaki, N., and Alashwal, H. (2021). A systematic literature review of studentâperformance prediction using machine learning techniques. *Education Sciences*, 11(9):552.

[Allen et al., 2006] Allen, R. D., Hermanson, D. R., Kozloski, T. M., and Ramsay, R. J. (2006). Auditor risk assessment: Insights from the academic literature. *Accounting Horizons*, 20(2):157–177.

[Andersen and Telle, 2022] Andersen, S. N. and Telle, K. (2022). Better out than in? the effect on recidivism of replacing incarceration with electronic monitoring in norway. *European Journal of Criminology*, 19(1):55–76.

[Anderson et al., 2019] Anderson, H., Boodhwani, A., and Baker, R. S. (2019). Assessing the fairness of graduation predictions. In *EDM*.

[Andrés-Pueyo et al., 2018] Andrés-Pueyo, A., Arbach-Lucioni, K., and Redondo, S. (2018). The RisCanvi: a new tool for assessing risk for violence in prison and recidivism. *Recidivism Risk Assessment: A Handbook for Practitioners*, pages 255–268.

[Andrews and Bonta, 2010] Andrews, D. A. and Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law*, 16(1):39.

[Andrews et al., 2006] Andrews, D. A., Bonta, J., and Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & delinquency*, 52(1):7–27.

[Andrews et al., 2000] Andrews, D. A., Bonta, J., and Wormith, S. (2000). *Level of service/case management inventory: LS/CMI*. Multi-Health Systems Toronto, Canada.

[Anenberg et al., 2016] Anenberg, S. C., Belova, A., Brandt, J., Fann, N., Greco, S., Guttikunda, S., Heroux, M.-E., Hurley, F., Krzyzanowski, M., Medina, S., et al. (2016). Survey of ambient air pollution health risk assessment tools. *Risk analysis*, 36(9):1718–1736.

[Angrist et al., 1996] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23:2016.

[Arnold and Pistilli, 2012] Arnold, K. E. and Pistilli, M. D. (2012). Course signals at purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 267–270.

[Asante-Duah, 2002] Asante-Duah, D. K. (2002). *Public health risk assessment for human exposure to chemicals*, volume 6. Springer.

[Athey, 2015] Athey, S. (2015). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 5–6.

[Athey et al., 2019] Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

[Athey and Wager, 2019] Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

[Aulck et al., 2016] Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

[Austin, 2006] Austin, J. (2006). How much risk can we take-the misuse of risk assessment in corrections. *Fed. Probation*, 70:58.

[Baker and Hawn, 2021] Baker, R. S. and Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41.

[Barabas et al., 2017] Barabas, C., Dinakar, K., Ito, J., Virza, M., and Zittrain, J. (2017). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv:1712.08238*.

[Barabas et al., 2018] Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*, pages 62–76. PMLR.

[Barocas et al., 2017] Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2.

[Battocchi et al., 2019] Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., and Syrgkanis, V. (2019). EconML: A Python

Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

[Becker, 1968] Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer.

[Beckett and Sasson, 2003] Beckett, K. and Sasson, T. (2003). *The politics of injustice: Crime and punishment in America*. Sage Publications.

[Beggs and Grace, 2010] Beggs, S. M. and Grace, R. C. (2010). Assessment of dynamic risk factors: An independent validation study of the violence risk scale: Sexual offender version. *Sexual Abuse*, 22(2):234–251.

[Bell and Mellor, 2009] Bell, I. and Mellor, D. (2009). Clinical judgements: Research and practice. *Australian Psychologist*, 44(2):112–121.

[Benko, 2018] Benko, J. (2018). The radical humaneness of norwayâs halden prison: The goal of the norwegian penal system is to get inmates out of it. *NY Times Mag. Retrieved*, 10.

[Berk, 2012] Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

[Berk, 2017] Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216.

[Berk, 2019] Berk, R. (2019). Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies*, 16(1):175–194.

[Berk et al., 2018] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533.

[Berk et al., 2021] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

[Berk and Hyatt, 2015] Berk, R. and Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, 27(4):222–228.

[Berk et al., 2016] Berk, R. A., Sorenson, S. B., and Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, 13(1):94–115.

[Bhuller et al., 2020] Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324.

[Bickley and Beech, 2001] Bickley, J. and Beech, A. R. (2001). Classifying child abusers: Its relevance to theory and clinical practice. *International Journal of Offender Therapy and Comparative Criminology*, 45(1):51–69.

[Bonta, 1996] Bonta, J. (1996). Risk-needs assessment and treatment.

[Bonta, 2002] Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal justice and behavior*, 29(4):355–379.

[Bonta and Andrews, 2007] Bonta, J. and Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation*, 6(1):1–22.

[Bonta et al., 1998] Bonta, J., Law, M., and Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: a meta-analysis. *Psychological bulletin*, 123(2):123.

[Bonta and Wormith, 2007] Bonta, J. and Wormith, S. J. (2007). Risk and need assessment. *Developments in social work with offenders*, pages 131–152.

[Borum, 2006] Borum, R. (2006). Manual for the structured assessment of violence risk in youth (savry).

[Borum et al., 2020] Borum, R., Lodewijks, H. P., Bartel, P. A., and Forth, A. E. (2020). The structured assessment of violence risk in youth (savry). In *Handbook of violence risk assessment*, pages 438–461. Routledge.

[Bray et al., 2019] Bray, B. C., Dziak, J. J., Patrick, M. E., and Lanza, S. T. (2019). Inverse propensity score weighting with a latent class exposure: Estimating the causal effect of reported reasons for alcohol use on problem alcohol use 16 years later. *Prevention Science*, 20(3):394–406.

[Brennan and Dieterich, 2018] Brennan, T. and Dieterich, W. (2018). Correctional offender management profiles for alternative sanctions (compas). *Handbook of Recidivism Risk/Needs Assessment Tools (2018)*, 49.

[Brennan et al., 2009] Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40.

[Brigell et al., 2020] Brigell, M. G., Chiang, B., Maa, A. Y., and Davis, C. Q. (2020). Enhancing risk assessment in patients with diabetic retinopathy by combining measures of retinal function and structure. *Translational vision science & technology*, 9(9):40–40.

[Bukralia et al., 2015] Bukralia, R., Deokar, A. V., and Sarnikar, S. (2015). Using academic analytics to predict dropout risk in e-learning courses. In *Reshaping Society through Analytics, Collaboration, and Decision Support*, pages 67–93. Springer.

[Caliendo and Kopeinig, 2008] Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72.

[Calkin et al., 2011] Calkin, D. E., Thompson, M. P., Finney, M. A., and Hyde, K. D. (2011). A real-time risk assessment tool supporting wildland fire decisionmaking. *Journal of Forestry*, 109(5):274–280.

[Capdevila et al., 2015] Capdevila, M., Blanch Serentill, M., Ferrer Puig, M., Andrés Pueyo, A., Framis Ferrer, B., Comas López, N., Garrigós Bou, A., Boldú Pedro, A., Batlle Manonelles, A., Mora Encinas, J., et al. (2015). Taxa de reincidència penitenciària 2014.

[Casanova et al., 2021] Casanova, J. R., Gomes, C. M. A., Bernardo, A. B., Núñez, J. C., and Almeida, L. S. (2021). Dimensionality and reliability of a screening instrument for students at-risk of dropping out from higher education. *Studies in Educational Evaluation*, 68:100957.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[Chen and Shapiro, 2007] Chen, M. K. and Shapiro, J. M. (2007). Do harsher prison conditions reduce recidivism? a discontinuity-based approach. *American Law and Economics Review*, 9(1):1–29.

[Chernozhukov et al., 2016] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

[Cho et al., 2016] Cho, K., Barnes, C. M., and Guanara, C. L. (2016). Sleepy punishers are harsh punishers: Daylight saving time and legal sentences. *Psychological science*.

[Choi, 2018] Choi, Y. (2018). Student employment and persistence: Evidence of effect heterogeneity of student employment on college dropout. *Research in Higher Education*, 59(1):88–107.

[Chouldechova, 2017] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

[Chouldechova and Roth, 2018] Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

[Chounta et al.,] Chounta, I.-A., Uiboleht, K., Roosimäe, K., Pedaste, M., and Valk, A. From data to intervention: Predicting students at-risk in a higher education institution.

[Cid, 2009] Cid, J. (2009). Is imprisonment criminogenic? a comparative study of recidivism rates between prison and suspended prison sanctions. *European Journal of Criminology*, 6(6):459–480.

[Clarke et al., 2017] Clarke, M. C., Peterson-Badali, M., and Skilling, T. A. (2017). The relationship between changes in dynamic risk factors and the predictive validity of risk assessments among youth offenders. *Criminal Justice and Behavior*, 44(10):1340–1355.

[Cohen, 2017] Cohen, A. (2017). Analysis of student activity in web-supported courses as a tool for predicting dropout. *Educational Technology Research and Development*, 65(5):1285–1304.

[Coleman, 2019] Coleman, J. S. (2019). *Equality and achievement in education*. Routledge.

[Collins, 2010] Collins, R. E. (2010). The effect of gender on violent and nonviolent recidivism: A meta-analysis. *Journal of Criminal Justice*, 38(4):675–684.

[Corbett-Davies and Goel, 2018] Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

[Corbett-Davies et al., 2017] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

[Cording et al., 2016] Cording, J. R., Beggs Christofferson, S. M., and Grace, R. C. (2016). Challenges for the theory and application of dynamic risk factors. *Psychology, Crime & Law*, 22(1-2):84–103.

[Cotter, 2020] Cotter, R. (2020). Length of incarceration and recidivism. Technical report, Washington, DC, USA.

[Cullen and Gilbert, 2012] Cullen, F. and Gilbert, K. (2012). *Reaffirming rehabilitation*. Routledge.

[Cullen et al., 2000] Cullen, F. T., Fisher, B. S., and Applegate, B. K. (2000). Public opinion about punishment and corrections. *Crime and justice*, 27:1–79.

[Cullen et al., 2011] Cullen, F. T., Jonson, C. L., and Nagin, D. S. (2011). Prisons do not reduce recidivism: The high cost of ignoring science. *The Prison Journal*, 91(3_suppl):48S–65S.

[Dahle et al., 2014] Dahle, K.-P., Biedermann, J., Lehmann, R. J., and Gallasch-Nemitz, F. (2014). The development of the crime scene behavior risk measure for sexual offense recidivism. *Law and human behavior*, 38(6):569.

[Dale et al., 2008] Dale, V. H., Biddinger, G. R., Newman, M. C., Oris, J. T., Suter, G. W., Thompson, T., Armitage, T. M., Meyer, J. L., Allen-King, R. M., Burton, G. A., et al. (2008). Enhancing the ecological risk assessment process. *Integrated environmental assessment and management*, 4(3):306–313.

[Danziger et al., 2011] Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

[Dawes et al., 1989] Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

[Dawid, 1979] Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

[Del Bonifro et al., 2020] Del Bonifro, F., Gabbrielli, M., Lisanti, G., and Zingaro, S. P. (2020). Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, pages 129–140. Springer.

[DeMichele et al., 2018] DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., and Misra, S. (2018). The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky. *Available at SSRN 3168452*.

[Denley, 2013] Denley, T. (2013). Degree compass: A course recommendation system. *EDUCAUSE Review Online,* `https://er.educause.edu/articles/2013/9/degree-compass-a-course-recommendation-system`.

[Desmarais and Singh, 2013] Desmarais, S. and Singh, J. (2013). Risk assessment instruments validated and implemented in correctional settings in the united states.

[Desmarais et al., 2016] Desmarais, S. L., Johnson, K. L., and Singh, J. P. (2016). Performance of recidivism risk assessment instruments in us correctional settings. *Psychological services*, 13(3):206.

[Dieterich et al., 2016] Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4).

[Douglas and Webster, 1999] Douglas, K. S. and Webster, C. D. (1999). The hcr-20 violence risk assessment scheme: Concurrent validity in a sample of incarcerated offenders. *Criminal justice and behavior*, 26(1):3–19.

[Dueñas-Espín et al., 2016] Dueñas-Espín, I., Vela, E., Pauws, S., Bescos, C., Cano, I., Cleries, M., Contel, J. C., de Manuel Keenoy, E., Garcia-Aymerich, J., Gomez-Cabrero, D., et al. (2016). Proposals for enhanced

health risk assessment and stratification in an integrated care scenario. *BMJ open*, 6(4):e010301.

[Durlauf and Nagin, 2011] Durlauf, S. N. and Nagin, D. S. (2011). Imprisonment and crime: Can both be reduced? *Criminology & Public Policy*, 10(1):13–54.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

[Dynia and Sung, 2000] Dynia, P. and Sung, H.-E. (2000). The safety and effectiveness of diverting felony drug offenders to residential treatment as measured by recidivism. *Criminal Justice Policy Review*, 11(4):299–311.

[Dyrbye et al., 2011] Dyrbye, L. N., Schwartz, A., Downing, S. M., Szydlo, D. W., Sloan, J. A., and Shanafelt, T. D. (2011). Efficacy of a brief screening tool to identify medical students in distress. *Academic Medicine*, 86(7):907–914.

[Eren and Mocan, 2018] Eren, O. and Mocan, N. (2018). Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3):171–205.

[Feldman et al., 2015] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

[Ferguson, 2012] Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6):304–317.

[Funk, 1999] Funk, S. J. (1999). Risk assessment for juveniles on probation: A focus on gender. *Criminal Justice and Behavior*, 26(1):44–68.

[Ganschow and Sparks, 1991] Ganschow, L. and Sparks, R. (1991). A screening instrument for the identification of foreign language learning problems. *Foreign Language Annals*, 24(5):383–398.

[Gardner et al., 2019] Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234.

[Gendreau et al., 2000] Gendreau, P., Goggin, C., Cullen, F. T., and Andrews, D. A. (2000). The effects of community sanctions and incarceration on recidivism. In *Forum on corrections research*, volume 12, pages 10–13. Correctional Service of Canada.

[Gilman and Walker, 2020] Gilman, A. B. and Walker, S. C. (2020). Evaluating the effects of an adolescent family violence intervention program on recidivism among court-involved youth. *Journal of family violence*, 35(2):95–106.

[Glöckner, 2016] Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision making*, 11(6):601.

[Glynn and Quinn, 2010] Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

[Goad et al., 2021] Goad, T., Jones, E., Bulger, S., Daum, D., Hollett, N., and Elliott, E. (2021). Predicting student success in online physical education. *American Journal of Distance Education*, 35(1):17–32.

[Grann et al., 1999] Grann, M., Långström, N., Tengström, A., and Kullgren, G. (1999). Psychopathy (pcl-r) predicts violent recidivism among

criminal offenders with personality disorders in sweden. *Law and Human Behavior*, 23(2):205–217.

[Green and Winik, 2010] Green, D. P. and Winik, D. (2010). Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology*, 48(2):357–387.

[Grove and Meehl, 1996] Grove, W. M. and Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, public policy, and law*, 2(2):293.

[Grove et al., 2000] Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19.

[Gupta et al., 2016] Gupta, A., Hansman, C., and Frenchman, E. (2016). The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies*, 45(2):471–505.

[Gutiérrez et al., 2020] Gutiérrez, F., Seipp, K., Ochoa, X., Chiluiza, K., De Laet, T., and Verbert, K. (2020). Lada: A learning analytics dashboard for academic advising. *Computers in Human Behavior*, 107:105826.

[Haarsma et al., 2020] Haarsma, G., Davenport, S., White, D. C., Ormachea, P. A., Sheena, E., and Eagleman, D. M. (2020). Assessing risk among correctional community probation populations: Predicting reoffense with mobile neurocognitive assessment software. *Frontiers in Psychology*, 10:2926.

[Hamilton, 2019] Hamilton, M. (2019). The sexist algorithm. *Behavioral sciences & the law*, 37(2):145–157.

[Hanson, 2005] Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. *Journal of interpersonal violence*, 20(2):212–217.

[Hanson and Bussiere, 1998] Hanson, R. K. and Bussiere, M. T. (1998). Predicting relapse: a meta-analysis of sexual offender recidivism studies. *Journal of consulting and clinical psychology*, 66(2):348.

[Harding and Harris, 2020] Harding, D. J. and Harris, H. M. (2020). *After prison: Navigating adulthood in the shadow of the justice system*. Russell Sage Foundation.

[Harding et al., 2017] Harding, D. J., Morenoff, J. D., Nguyen, A. P., and Bushway, S. D. (2017). Short-and long-term effects of imprisonment on future felony convictions and prison admissions. *Proceedings of the National Academy of Sciences*, 114(42):11103–11108.

[Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

[Hare, 2003] Hare, R. D. (2003). The psychopathy checklist–revised. *Toronto, ON*, 412.

[Harris et al., 1993] Harris, G. T., Rice, M. E., and Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal justice and behavior*, 20(4):315–335.

[Hart and Boer, 2011] Hart, S. D. and Boer, D. P. (2011). Structured professional judgment guidelines for sexual violence risk assessment: e sexual violence risk-20 (svr-20) and risk for sexual violence protocol (rsvp). In *Handbook of violence risk assessment*, pages 279–304. Routledge.

[Hart et al., 2007] Hart, S. D., Michie, C., and Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the âmargins of errorâof group v. individual predictions of violence. *The British Journal of Psychiatry*, 190(S49):s60–s65.

[Helmus et al., 2012] Helmus, L., Thornton, D., Hanson, R. K., and Babchishin, K. M. (2012). Improving the predictive accuracy of static-99 and static-2002 with older sex offenders: Revised age weights. *Sexual Abuse*, 24(1):64–101.

[Henneguelle et al., 2016] Henneguelle, A., Monnery, B., and Kensey, A. (2016). Better at home than in prison? the effects of electronic monitoring on recidivism in france. *The Journal of Law and Economics*, 59(3):629–667.

[Herbaut, 2021] Herbaut, E. (2021). Overcoming failure in higher education: Social inequalities and compensatory advantage in dropout patterns. *Acta Sociologica*, 64(4):383–402.

[Heyes and Saberian, 2019] Heyes, A. and Saberian, S. (2019). Temperature and decisions: evidence from 207,000 court cases. *American Economic Journal: Applied Economics*, 11(2):238–65.

[Hilton et al., 2010] Hilton, N. Z., Harris, G. T., and Rice, M. E. (2010). *Risk assessment for domestically violent men: Tools for criminal justice, offender intervention, and victim services.* American Psychological Association.

[Hirschi and Gottfredson, 1983] Hirschi, T. and Gottfredson, M. (1983). Age and the explanation of crime. *American journal of sociology*, 89(3):552–584.

[Hjalmarsson and Lindquist, 2020] Hjalmarsson, R. and Lindquist, M. J. (2020). The health effects of prison.

[Hoffman and Beck, 1974] Hoffman, P. B. and Beck, J. L. (1974). Parole decision-making: A salient factor score. *Journal of criminal justice*, 2(3):195–206.

[Howard and Dixon, 2012] Howard, P. D. and 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.

[Hu and Rangwala, 2020] Hu, Q. and Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. *International Educational Data Mining Society*.

[Huang et al., 2020] Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C., and Yang, S. J. (2020). Predicting studentsâ academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, 28(2):206–230.

[Huebner et al., 2010] Huebner, B. M., DeJong, C., and Cobbina, J. (2010). Women coming home: Long-term patterns of recidivism. *Justice Quarterly*, 27(2):225–254.

[Hutt et al., 2019] Hutt, S., Gardner, M., Duckworth, A. L., and D'Mello, S. K. (2019). Evaluating fairness and generalizability in models predicting on-time graduation from college applications. *International Educational Data Mining Society*.

[Joffe and Rosenbaum, 1999] Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology*, 150(4):327–333.

[Johndrow et al., 2019] Johndrow, J. E., Lum, K., et al. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

[Johnson et al., 2011] Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., and Robinson, C. R. (2011). The construction and validation of the federal post conviction risk assessment (pcra). *Fed. Probation*, 75:16.

[Jolliffe and Hedderman, 2015] Jolliffe, D. and Hedderman, C. (2015). Investigating the impact of custody on reoffending using propensity score matching. *Crime & Delinquency*, 61(8):1051–1077.

[Kamiran and Calders, 2009] Kamiran, F. and Calders, T. (2009). Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE.

[Kamishima et al., 2011] Kamishima, T., Akaho, S., and Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE.

[Karimi-Haghighi and Castillo, 2021a] Karimi-Haghighi, M. and Castillo, C. (2021a). Efficiency and fairness in recurring data-driven risk assessments of violent recidivism. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 994–1002.

[Karimi-Haghighi and Castillo, 2021b] Karimi-Haghighi, M. and Castillo, C. (2021b). Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 210–214.

[Karimi-Haghighi et al., 2022] Karimi-Haghighi, M., Castillo, C., and Hernández-Leo, D. (2022). A causal inference study on the effects of first year workload on the dropout rate of undergraduates. In *International Conference on Artificial Intelligence in Education*, pages 15–27. Springer.

[Karimi-Haghighi et al., 2021] Karimi-Haghighi, M., Castillo, C., Hernandez-Leo, D., and Oliver, V. M. (2021). Predicting early dropout: Calibration and algorithmic fairness considerations. *arXiv preprint arXiv:2103.09068*.

[Kehl and Kessler, 2017] Kehl, D. L. and Kessler, S. A. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing.

[Kemper et al., 2020] Kemper, L., Vorhoff, G., and Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, 10(1):28–47.

[Kemshall, 2003] Kemshall, H. (2003). *Understanding risk in criminal justice*. McGraw-Hill Education (UK).

[Kirk, 2020] Kirk, D. S. (2020). *Home free: Prisoner reentry and residential change after hurricane Katrina*. Oxford University Press, USA.

[Kirton and Kravitz, 2011] Kirton, S. B. and Kravitz, L. (2011). Objective structured clinical examinations (osces) compared with traditional assessment methods. *American journal of pharmaceutical education*, 75(6).

[Kizilcec and Lee, 2020] Kizilcec, R. F. and Lee, H. (2020). Algorithmic fairness in education. *arXiv preprint arXiv:2007.05443*.

[Kleinberg et al., 2018] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

[Kleinberg et al., 2016] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[Knight and Thornton, 2007] Knight, R. A. and Thornton, D. (2007). *Evaluating and improving risk assessment schemes for sexual recidivism: A long-term follow-up of convicted sexual offenders*. National Criminal Justice Reference Service Rockville, MD.

[Kröner et al., 2007] Kröner, C., Stadtland, C., Eidt, M., and Nedopil, N. (2007). The validity of the violence risk appraisal guide (vrag) in

predicting criminal recidivism. *Criminal Behaviour and Mental Health*, 17(2):89–100.

[Kropp and Hart, 2000] Kropp, P. R. and Hart, S. D. (2000). The spousal assault risk assessment (sara) guide: Reliability and validity in adult male offenders. *Law and human behavior*, 24(1):101–118.

[Krumm et al., 2014] Krumm, A. E., Waddington, R. J., Teasley, S. D., and Lonn, S. (2014). A learning management system-based early warning system for academic advising in undergraduate engineering. In *Learning analytics*, pages 103–119. Springer.

[Labrecque et al., 2014] Labrecque, R. M., Smith, P., Lovins, B. K., and Latessa, E. J. (2014). The importance of reassessment: How changes in the lsi-r risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation*, 53(2):116–128.

[Langley and Simon, 1995] Langley, P. and Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):54–64.

[Lappi-Seppälä, 2012] Lappi-Seppälä, T. (2012). Penal policies in the nordic countries 1960–2010. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 13(sup1):85–111.

[Larose et al., 2011] Larose, S., Cyrenne, D., Garceau, O., Harvey, M., Guay, F., Godin, F., Tarabulsy, G. M., and Deschênes, C. (2011). Academic mentoring and dropout prevention for students in math, science and technology. *Mentoring & Tutoring: Partnership in Learning*, 19(4):419–439.

[Larrabee Sønderlund et al., 2019] Larrabee Sønderlund, A., Hughes, E., and Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, 50(5):2594–2618.

[Larson et al., 2016] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9.

[Latessa et al., 2009] Latessa, E., Smith, P., Lemke, R., Makarios, M., and Lowenkamp, C. (2009). Creation and validation of the ohio risk assessment system: Final report. *Center for Criminal Justice Research, School of Criminal Justice, University of Cincinnati, Cincinnati, OH. Retrieved from* `https://www.uc.edu/content/dam/uc/ccjr/docs/reports/project_reports/ORAS_Final_Report.pdf`.

[Latessa et al., 2010] Latessa, E. J., Lemke, R., Makarios, M., and Smith, P. (2010). The creation and validation of the ohio risk assessment system (oras). *Fed. Probation*, 74:16.

[Lee and Kizilcec, 2020] Lee, H. and Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. *arXiv preprint arXiv:2007.00088*.

[Leitner et al., 2017] Leitner, P., Khalil, M., and Ebner, M. (2017). Learning analytics in higher educationâa literature review. *Learning analytics: Fundaments, applications, and trends*, pages 1–23.

[Lemmerich et al., 2010] Lemmerich, F., Ifl, M., and Puppe, F. (2010). Identifying influence factors on students success by subgroup discovery. In *Educational Data Mining 2011*.

[Liz-Domínguez et al., 2019] Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., and Mikic-Fonte, F. A. (2019). Systematic literature review of predictive analysis tools in higher education. *Applied Sciences*, 9(24):5569.

[Lloyd-Jones et al., 2019] Lloyd-Jones, D. M., Braun, L. T., Ndumele, C. E., Smith Jr, S. C., Sperling, L. S., Virani, S. S., and Blumenthal, R. S. (2019). Use of risk assessment tools to guide decision-making

in the primary prevention of atherosclerotic cardiovascular disease: a special report from the american heart association and american college of cardiology. *Circulation*, 139(25):e1162–e1177.

[Loeffler, 2013] Loeffler, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, 51(1):137–166.

[Loeffler and Grunwald, 2015] Loeffler, C. E. and Grunwald, B. (2015). Processed as an adult: A regression discontinuity estimate of the crime effects of charging nontransfer juveniles as adults. *Journal of research in crime and delinquency*, 52(6):890–922.

[Loeffler and Nagin, 2022] Loeffler, C. E. and Nagin, D. S. (2022). The impact of incarceration on recidivism. *Annual Review of Criminology*, 5:133–152.

[Lopez, 1989] Lopez, S. R. (1989). Patient variable biases in clinical judgment: Conceptual overview and methodological considerations. *Psychological Bulletin*, 106(2):184.

[Lowenkamp, 2009] Lowenkamp, C. T. (2009). The development of an actuarial risk assessment instrument for us pretrial services. *Fed. Probation*, 73:33.

[Lowis and Castley, 2008] Lowis, M. and Castley, A. (2008). Factors affecting student progression and achievement: prediction and intervention. a two-year study. *Innovations in education and teaching international*, 45(4):333–343.

[Loza, 2018] Loza, W. (2018). Self-appraisal questionnaire (saq): A tool for assessing violent and non-violent recidivism. *Handbook of recidivism risk/needs assessment tools*, pages 165–179.

[Lum, 2017] Lum, K. (2017). Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour*, 1(7):1–1.

[Marcelo F. Aebi, 2022] Marcelo F. Aebi, Edoardo Cocco, L. M. . M. M. T. (2022). Prisons and prisoners in europe 2021: Key findings of the space i report, https://wp.unil.ch/space/files/2022/05/Aebi-Cocco-Molnar-Tiago_2022_Prisons-and-Prisoners-in-Europe-2021_Key-Findings-SPACE-I_-220404.pdf.

[Marchese di Beccaria, 1819] Marchese di Beccaria, C. (1819). *An essay on crimes and punishments*. Number 47183. Philip H. Nicklin.

[Marie, 2009] Marie, O. (2009). The best ones come out first! early release from prison and recidivism a regression discontinuity approach. Technical report.

[Márquez-Vera et al., 2016] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124.

[Masserini and Bini, 2021] Masserini, L. and Bini, M. (2021). Does joining social media groups help to reduce studentsâ dropout within the first university year? *Socio-Economic Planning Sciences*, 73:100865.

[McDermott et al., 2008] McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., and Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable-and person-focused analyses. *Law and human behavior*, 32(4):325–338.

[Mears and Bales, 2009] Mears, D. P. and Bales, W. D. (2009). Supermax incarceration and recidivism. *Criminology*, 47(4):1131–1166.

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[Meier et al., 2020] Meier, A., Levav, J., and Meier, S. (2020). Early release and recidivism.

[Meredith et al., 2007] Meredith, T., Speir, J. C., and Johnson, S. (2007). Developing and implementing automated risk assessments in parole. *Justice Research and Policy*, 9(1):1–24.

[Miron et al., 2020] Miron, M., Tolan, S., Gómez, E., and Castillo, C. (2020). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, pages 1–37.

[Miron et al., 2021] Miron, M., Tolan, S., Gómez, E., and Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, 29(2):111–147.

[Mitchell et al., 2017] Mitchell, O., Cochran, J. C., Mears, D. P., and Bales, W. D. (2017). The effectiveness of prison for reducing drug offender recidivism: A regression discontinuity analysis. *Journal of Experimental Criminology*, 13(1):1–27.

[Modena et al., 2020] Modena, F., Rettore, E., and Tanzi, G. M. (2020). The effect of grants on university dropout rates: Evidence from the italian case. *Journal of Human Capital*, 14(3):343–370.

[Monahan and Skeem, 2015] Monahan, J. and Skeem, J. L. (2015). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, Forthcoming, Virginia Public Law and Legal Theory Research Paper*, (53).

[Monahan and Skeem, 2016] Monahan, J. and Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513.

[Monnery et al., 2020] Monnery, B., Wolff, F.-C., and Henneguelle, A. (2020). Prison, semi-liberty and recidivism: Bounding causal effects in a survival model. *International Review of Law and Economics*, 61:105884.

[Mueller-Smith, 2015] Mueller-Smith, M. (2015). The criminal and labor market impacts of incarceration. *Unpublished Working Paper*, 18.

[Myers and Nikoletti, 2003] Myers, H. and Nikoletti, S. (2003). Fall risk assessment: a prospective investigation of nursesâ clinical judgement and risk assessment tools in predicting patient falls. *International journal of nursing practice*, 9(3):158–165.

[Nagy and Molontay, 2018] Nagy, M. and Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, pages 000389–000394. IEEE.

[Narayanan, 2018] Narayanan, A. (2018). 21 fairness definitions and their politics. *presenterad på konferens om Fairness, Accountability, and Transparency*, 23.

[Narayanan, 21] Narayanan, A. (21). Fairness definitions and their politics. *Youtube: Arvind Naranayan, Available online: https://www. youtube. com/watch*.

[Nie and Wager, 2021] Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

[Nuffield, 1982] Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines*. Communication Division, Solicitor General of Canada Ottawa, Canada.

[OECD, 2012] OECD (2012). Education at a glance 2016. *Editions OECD*, 90.

[Olaya et al., 2020] Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., and Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, 134:113320.

[Pal, 2012] Pal, S. (2012). Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, 4(2):1.

[Pascarella and Terenzini, 2005] Pascarella, E. T. and Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2.* ERIC.

[Perrault et al., 2017] Perrault, R. T., Vincent, G. M., and Guy, L. S. (2017). Are risk assessments racially biased?: Field study of the savry and yls/cmi in probation. *Psychological assessment*, 29(6):664.

[Phillips et al., 2014] Phillips, S. M., Glasgow, R. E., Bello, G., Ory, M. G., Glenn, B. A., Sheinfeld-Gorin, S. N., Sabo, R. T., Heurtin-Roberts, S., Johnson, S. B., and Krist, A. H. (2014). Frequency and prioritization of patient health risks from a structured health risk assessment. *The Annals of Family Medicine*, 12(6):505–513.

[Plagge, 2013] Plagge, M. (2013). Using artificial neural networks to predict first-year traditional students second year retention rates. In *Proceedings of the 51st ACM Southeast Conference*, pages 1–5.

[Pleiss et al., 2017] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

[Quayle and Taylor, 2004] Quayle, E. and Taylor, M. (2004). *Child pornography: An internet crime*. Routledge.

[Quinlivan et al., 2017] Quinlivan, L., Cooper, J., Meehan, D., Longson, D., Potokar, J., Hulme, T., Marsden, J., Brand, F., Lange, K., Riseborough, E., et al. (2017). Predictive accuracy of risk scales following self-harm: multicentre, prospective cohort study. *The British Journal of Psychiatry*, 210(6):429–436.

[Raphael and Stoll, 2009] Raphael, S. and Stoll, M. A. (2009). *Do prisons make us safer?: the benefits and costs of the prison boom*. Russell Sage Foundation.

[Rausand, 2013] Rausand, M. (2013). *Risk assessment: theory, methods, and applications*, volume 115. John Wiley & Sons.

[Raz and Michael, 2001] Raz, T. and Michael, E. (2001). Use and benefits of tools for project risk management. *International journal of project management*, 19(1):9–17.

[Rettenberger et al., 2011] Rettenberger, M., Boer, D. P., and Eher, R. (2011). The predictive accuracy of risk factors in the sexual violence risk–20 (svr-20). *Criminal Justice and Behavior*, 38(10):1009–1027.

[Rettenberger et al., 2010a] Rettenberger, M., Matthes, A., Boer, D. P., and Eher, R. (2010a). Prospective actuarial risk assessment: A comparison of five risk assessment instruments in different sexual offender subtypes. *International Journal of Offender Therapy and Comparative Criminology*, 54(2):169–186.

[Rettenberger et al., 2010b] Rettenberger, M., Mönichweger, M., Buchelle, E., Schilling, F., and Eher, R. (2010b). Entwicklung eines screeninginstruments zur vorhersage der einschlägigen rückfälligkeit von gewaltstraftätern [the development of a screening scale for the prediction of violent offender recidivism]. *Monatsschrift für Kriminologie und Strafrechtsreform*, 93(5):346–360.

[Rhodes et al., 2018] Rhodes, W., Gaes, G. G., Kling, R., and Cutler, C. (2018). Relationship between prison length of stay and recidivism: A study using regression discontinuity and instrumental variables with multiple break points. *Criminology & Public Policy*, 17(3):731–769.

[Romero and Ventura, 2019] Romero, C. and Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies*, 12(2):145–147.

[Rosenbaum and Rubin, 1983] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[Rubin et al., 2008] Rubin, J., Gallo, F., and Coutts, A. (2008). Violent crime: Risk models, effective interventions and risk management.

[Ryan et al., 2010] Ryan, C., Nielssen, O., Paton, M., and Large, M. (2010). Clinical decisions in psychiatry should not be based on risk assessment. *Australasian Psychiatry*, 18(5):398–403.

[Salas-Morera et al., 2019] Salas-Morera, L., Molina, A. C., Olmedilla, J. L. O., García-Hernández, L., and Palomo-Romero, J. M. (2019). Factors affecting engineering students dropout: A case study. *The International journal of engineering education*, 35(1):156–167.

[Sampson and Laub, 2017] Sampson, R. J. and Laub, J. H. (2017). Life-course desisters? trajectories of crime among delinquent boys followed to age 70. In *Developmental and life-course criminological theories*, pages 37–74. Routledge.

[Sclater et al., 2016] Sclater, N., Peasgood, A., and Mullan, J. (2016). Learning analytics in higher education. *London: Jisc. Accessed February*, 8(2017):176.

[Shameli-Sendi et al., 2016] Shameli-Sendi, A., Aghababaei-Barzegar, R., and Cheriet, M. (2016). Taxonomy of information security risk assessment (isra). *Computers & security*, 57:14–30.

[Shapiro et al., 2017] Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., and Bhimdiwali, A. (2017). Completing college: A national view of student completion rates–fall 2011 cohort.

[Siemens, 2013] Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400.

[Singh et al., 2018] Singh, J., Kroner, D., Wormith, J., Desmarais, S., and Hamilton, Z. (2018). *Handbook of recidivism risk/needs assessment tools*. John Wiley & Sons.

[Singh et al., 2014] Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., Doyle, M., Folino, J. O., Godoy-Cervera, V., Grann, M., et al. (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44

countries. *International Journal of Forensic Mental Health*, 13(3):193–206.

[Singh et al., 2011] Singh, J. P., Grann, M., and Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical psychology review*, 31(3):499–513.

[Skeem et al., 2016] Skeem, J., Monahan, J., and Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5):580.

[Skeem and Lowenkamp, 2016] Skeem, J. L. and Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712.

[Sondhi et al., 2020] Sondhi, A., Leidi, A., and Best, D. (2020). Estimating a treatment effect on recidivism for correctional multiple component treatment for people in prison with an alcohol use disorder in england. *Substance Abuse Treatment, Prevention, and Policy*, 15(1):1–12.

[Spelman, 2020] Spelman, W. (2020). The limited importance of prison expansion. In *Crime, Inequality and the State*, pages 150–164. Routledge.

[Statistics and Agency, 2019] Statistics, N. I. and Agency, R. (2019). The enhanced combination order october 2015 to november 2018. Technical report, Northern Ireland.

[Syed et al., 2019] Syed, M., Anggara, T., Lanski, A., Duan, X., Ambrose, G. A., and Chawla, N. V. (2019). Integrated closed-loop learning analytics scheme in a first year experience course. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 521–530.

[Tanvir and Chounta,] Tanvir, H. and Chounta, I.-A. Exploring the importance of factors contributing to dropouts in higher education over time.

[Thistlethwaite and Campbell, 1960] Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6).

[Tinto, 2010] Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In *Higher education: Handbook of theory and research*, pages 51–89. Springer.

[Tinto, 2017] Tinto, V. (2017). Through the eyes of students. *Journal of College Student Retention: Research, Theory & Practice*, 19(3):254–269.

[Tobón, 2020] Tobón, S. (2020). Do better prisons reduce recidivism? evidence from a prison construction program. *The Review of Economics and Statistics*, pages 1–47.

[Tolan et al., 2019] Tolan, S., Miron, M., Gómez, E., and Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia.

[Tonry, 2014] Tonry, M. (2014). Why crime rates are falling throughout the western world. *Crime and justice*, 43(1):1–63.

[Torabi et al., 2016] Torabi, S. A., Giahi, R., and Sahebjamnia, N. (2016). An enhanced risk assessment framework for business continuity management systems. *Safety science*, 89:201–218.

[Travis et al., 2014] Travis, J., Western, B., and Redburn, F. S. (2014). The growth of incarceration in the united states: Exploring causes and consequences.

[Tsiatis, 2006] Tsiatis, A. A. (2006). Semiparametric theory and missing data.

[Turner et al., 2019] Turner, E., Medina, J., and Brown, G. (2019). Dashing hopes? the predictive accuracy of domestic abuse risk assessment by police. *The British Journal of Criminology*, 59(5):1013–1034.

[Ulmer and Steffensmeier, 2014] Ulmer, J. T. and Steffensmeier, D. J. (2014). The age and crime relationship: Social variation, social explanations. In *The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality*, pages 377–396. SAGE Publications Inc.

[Van Eijk, 2017] Van Eijk, G. (2017). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, 19(4):463–481.

[VanderWeele, 2015] VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

[Vansteelandt and Daniel, 2014] Vansteelandt, S. and Daniel, R. M. (2014). On regression adjustment for the propensity score. *Statistics in medicine*, 33(23):4053–4072.

[Vass, 1990] Vass, A. A. (1990). *Alternatives to prison: Punishment, custody and the community*. Sage London.

[Velázquez, 2018] Velázquez, A. G. . T. (2018). The changing state of recidivism: Fewer people going back to prison. Technical report, US.

[Verma and Rubin, 2018] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

[Viberg et al., 2018] Viberg, O., Hatakka, M., Bälter, O., and Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89:98–110.

[Vossensteyn et al., 2015] Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., and Wollscheid, S. (2015). Dropout and completion in higher education in europe: Main report.

[Wand, 2011] Wand, T. (2011). Investigating the evidence for the effectiveness of risk assessment in mental health care. *Issues in Mental Health Nursing*, 33(1):2–7.

[Weaver, 2007] Weaver, V. M. (2007). Frontlash: Race and the development of punitive crime policy. *Studies in American political development*, 21(2):230–265.

[Western, 2018] Western, B. (2018). *Homeward: Life in the year after prison*. Russell Sage Foundation.

[Williams and Weatherburn, 2022] Williams, J. and Weatherburn, D. (2022). Can electronic monitoring reduce reoffending? *Review of Economics and Statistics*, 104(2):232–245.

[Woodworth et al., 2017] Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

[Wright et al., 1984] Wright, K. N., Clear, T. R., and Dickson, P. (1984). Universal applicability of probation risk-assessment instruments: A critique. *Criminology*, 22(1):113–134.

[Yang et al., 2010] Yang, M., Wong, S. C., and Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological bulletin*, 136(5):740.

[Yu et al., 2021] Yu, R., Lee, H., and Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the eighth ACM conference on learning@ scale*, pages 91–100.

[Yukhnenko et al., 2019a] Yukhnenko, D., Sridhar, S., and Fazel, S. (2019a). A systematic review of criminal recidivism rates worldwide: 3-year update. *Wellcome Open Research*, 4.

[Yukhnenko et al., 2019b] Yukhnenko, D., Wolf, A., Blackwood, N., and Fazel, S. (2019b). Recidivism rates in individuals receiving community sentences: A systematic review. *PloS one*, 14(9):e0222495.

[Zafar et al., 2017] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

[Zapryanova, 2020] Zapryanova, M. (2020). The effects of time in prison and time on parole on recidivism. *The Journal of Law and Economics*, 63(4):699–727.

[Zemel et al., 2013] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.

[Zeng et al., 2007] Zeng, J., An, M., and Smith, N. J. (2007). Application of a fuzzy based decision making methodology to construction project risk assessment. *International journal of project management*, 25(6):589–600.

[Zottola et al., 2022] Zottola, S. A., Desmarais, S. L., Lowder, E. M., and Duhart Clarke, S. E. (2022). Evaluating fairness of algorithmic risk assessment instruments: The problem with forcing dichotomies. *Criminal Justice and Behavior*, 49(3):389–410.

[Zou and Schiebinger, 2018] Zou, J. and Schiebinger, L. (2018). Ai can be sexist and racistâitâs time to make it fair.