Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia

Songül Tolan*
Marius Miron*
Joint Research Centre,
European Commission
firstname.lastname@ec.europa.eu

Emilia Gómez
Joint Research Centre,
European Commission
Universitat Pompeu Fabra, Barcelona
emilia.gomez@upf.edu

Carlos Castillo
Universitat Pompeu Fabra, Barcelona
chato@acm.org

ABSTRACT

In this paper we study the limitations of Machine Learning (ML) algorithms for predicting juvenile recidivism. Particularly, we are interested in analyzing the trade-off between predictive performance and fairness. To that extent, we evaluate fairness of ML models in conjunction with SAVRY, a structured professional risk assessment framework, on a novel dataset originated in Catalonia. In terms of accuracy on the prediction of recidivism, the ML models slightly outperform SAVRY; the results improve with more data or more features available for training (AUCROC of 0.64 with SAVRY vs. AUCROC of 0.71 with ML models). However, across three fairness metrics used in other studies, we find that SAVRY is in general fair, while the ML models tend to discriminate against male defendants, foreigners, or people of specific national groups. For instance, foreigners who did not recidivate are almost twice as likely to be wrongly classified as high risk by ML models than Spanish nationals. Finally, we discuss potential sources of this unfairness and provide explanations for them, by combining ML interpretability techniques with a thorough data analysis. Our findings provide an explanation for why ML techniques lead to unfairness in data-driven risk assessment, even when protected attributes are not used in training.

CCS CONCEPTS

• **Information systems → Expert systems**; • **Applied computing → Law**; • **Social and professional topics**;

KEYWORDS

algorithmic fairness, algorithmic bias, machine learning, risk assessment, criminal recidivism

ACM Reference Format:
Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3322640.3326705

*Both authors contributed equally to this research.

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326705

1 INTRODUCTION

Machine learning (ML) systems detect patterns in data and are able to predict complex outputs under high uncertainty [37]. Medicine, finance and law, are a few domains where humans rely on an algorithms to solve expert tasks [28]. In these cases ML systems can surpass human capabilities, particularly when dealing with large datasets or a high number of input features. One example where ML algorithms and expert systems can better inform human decisions is predicting criminal recidivism [26], defined as the act of a person committing a crime after they have been convicted of an earlier crime [11]. However, the adoption of ML in this area is problematic, knowing that the decisions of ML models can often be biased and discriminate against certain minority groups or populations, becoming unfair [3, 4, 13]. This can be concerning for nontransparent models, such as deep neural networks, if the decision process is not made transparent. [29].

Here we propose a methodology to assess predictive performance and unfairness for the ML methods used in juvenile recidivism prediction, and to investigate the potential sources of unfairness. To that extent, we compare the ML models with an existing risk assessment tool in terms of predictive performance and fairness and we use ML interpretability [29] combined with a thorough data analysis to find explanations of disparity.

The literature on fair algorithms mainly derives its (group) fairness concepts from a legal context. Generally, a process or decision is considered fair if it does not discriminate against people on the basis of their membership to a protected group. For instance, Article 14 of the European Convention on Human Rights states as protected groups "sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status". In practice, the legal context leaves us with more than one definition of fairness. Computer science researchers talk of at least 21 definitions of fairness (see, e.g., [5, 32, 41] for an overview on different definitions of algorithmic fairness.), which effectively contradict each other [13, 27]. Moreover, the literature distinguishes between two categories of fairness: individual fairness [19] and group fairness. For applicability, we focus on group fairness, the more extensively studied type of fairness, across two dimensions: sex and nationality. It should be noted that generally, fairness is a value-driven concept, not a technical one. Fulfilling specific fairness constraints with an algorithmic risk-assessment tool, does not preclude the violation of other aspects of

fairness. Nevertheless, our chosen fairness criteria, as derived from a legal context, are appropriate in the context of criminal justice.

Risk assessment tools are globally well established to inform judges about a defendant's risk of recidivism [39]. These instruments vary strongly in their degree of structure and involvement of (human) experts. One such instrument, the Structured Assessment of Violence Risk in Youth (SAVRY) is used to assess the risk of violence in juvenile justice [8]. As a "Structured Professional Judgment" (SPJ), SAVRY leaves a high degree of involvement in the risk assessment to professionals. Being designed to inform intervention planning, such as clinical treatment plans or release and discharge decisions [8] SAVRY plays an essential role in the course of a juvenile defendant in the justice system.

There are good reasons to make use of such instruments. Naturally, judges are not free from subjective biases that affect our decision making [16]. Moreover, meta-studies show that structured assessments outperform individual experts in the prediction of criminal behavior [1]. Besides, structured assessments have been shown to reduce the use of more severe punishments [42]. Thus, structured assessments can potentially improve decision making efficiency in criminal justice. However, it remains unclear whether a SPJ like SAVRY can be discriminatory. Unlike the discriminatory tendencies in risk assessments, like Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) [10], the literature on discrimination in SPJs is still scarce. An analysis of SAVRY for racial bias against blacks in Pennsylvania found that SAVRY did not predict significantly different risk scores as a function of race [35].

The contributions of this paper are three-fold. First, we compare the predictive performance of SAVRY against a risk assessment generated by ML methods based on information on defendant demographics and criminal history. We assume that expert assessment is laborious and performing the SAVRY assessment is expensive. We use a Catalonian dataset on recidivism in juvenile justice comprising observations of 4753 Catalan adolescents who committed offences between 2002 and 2010 and whose recidivism behavior was recorded in 2013 and 2015. The SAVRY assessment is available solely for a subset of 855 defendants. Therefore, we study whether ML models taking as input solely demographic and criminal history data, are able to have better predictive performance at lower input costs.

Second, we are interested in assessing whether SAVRY and ML models show any discrimination along sex or nationality. Third, we combine ML interpretability techniques combined with a thorough data analysis to discuss and elaborate on the possible causes of disparity in ML methods. Finally, we discuss the consequences and limitations of potential unfairness mitigation techniques, justifying not deploying such techniques in the present paper.

The remainder of this paper is structured as follows. We provide background information on SAVRY in Section 2. In Section 3 we introduce the dataset used in our experiments. We present the methodology in Section 4. The evaluation of the methods is introduced in Section 5, the experimental design in Section 5.1 and the results in Section 5.2. In Section 6 we discuss and elaborate on potential sources of discrimination and ways to mitigate it. The conclusions and future directions of this research are presented in Section 7.

2 THE SAVRY RISK ASSESSMENT TOOL

This section gives a general overview of SAVRY, the SPJ. Its use in the underlying dataset is described in the subsequent section. SAVRY is a violence risk assessment tool designed as a SPJ [9]. That is, as opposed to COMPAS, SAVRY is an open and interpretable assessment that actively guides the evaluating expert through the individual features that make up the overall assessment. As such it leaves a high degree of involvement by individual expert assessments. With SAVRY, juvenile justice professionals assign scores on a three-level coding structure for severity (low, moderate high) to a list of 24 risk factors and six protective factors. These risk factors are divided into three categories: Historical, Individual, and Social/Contextual. The SAVRY manual provides an example on how the categorization of the individual risk factors is conducted: "[...][I]n coding the History of Violence item, a youth would be coded as "Low" if he had committed no prior acts of violence, "Moderate" if he was known to have committed one or two violent acts, and "High" if there were three or more. Protective factors are simply coded as present or absent."[8]. Thus, the manual provides measurable benchmarks in the assessment of each individual risk item. The 24 risk factors are summed up to a total risk score (SAVRY sum). The six protective factors are recorded as present/absent. After the assessment of the individual items an expert assigns a final overall score (low, moderate or high risk) that indicates the defendants risk of violent recidivism (Expert). This final evaluation is a professional judgment, not algorithmic.

Note that experts are aware that empirical assessments of SAVRY are mostly based on cases of male defendants. In addition, they are informed about substantial sex-differences in the response to specific risk factors [6, 38, 43]. Therefore, the current SAVRY manual indicates risk factors that may apply differently to males and females [8].

Meta studies show a good predictive validity for the SAVRY expert evaluation with a median AUCROC of 0.71 [34, 39] and the SAVRY sum with mean weighted AUCROC values of 0.71 [23].

3 DATASET AND DATA PRE-PROCESSING

For this research we start with a dataset of all juvenile offenders who in 2010 finished a sentence in the juvenile justice system of Catalonia (N=4753) [1]. The corresponding crimes were committed between 2002 and 2010 when the offenders were aged 12-17 years. To observe recidivism behavior, their status was followed up on December 31, 2013 and December 31, 2015 (independent of their association with the juvenile justice system). We focus our research on the sub-sample of 855 who were subject to a SAVRY assessment. The outcome variable indicates the recidivism status by December 31, 2015. All SAVRY assessments were conducted towards the end of the sentences in 2010. That is, the SAVRY assessment did not impact the sentence that the defendant received for the main crime committed. In this research we use a pre-processed version of the data. The pre-processing code and the resulting are available on the repository described in Section 5.1.5. As a representation of the SAVRY assessment we look at both the SAVRY sum of 24 risk

[1]Provided by the Centre for Legal Studies and Specialised Training [7], available at http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html.

factors as well as the final expert assessment. We elaborate on their dependency in Section 4.2.

Table 2 shows descriptive statistics of most features contained in the analysis by recidivism status in 2015. We distinguish between two sets of input features: static features that are not encoded in SAVRY, including protected features, and features that are encoded in SAVRY. Note that SAVRY features are not restricted to the 24 risk factors. The protected features are listed in the top panel. Given our analysis of unfairness across sex and nationality, we distinguish with the indicators male/female, as well as Spanish/foreign, where we also look at the subgroup of Latin Americans and Maghrebi. Due to sample size restrictions we exclude further analysis on (non-Spanish) Europeans or other national groups. The table shows that many static features as well as almost all SAVRY features significantly differ between the group of recidivists and non-recidivists. This emphasizes the empirical relevance of the input features used in this analysis. Further, we also see that almost all protected features differ significantly between the compared groups. Finally, it is important to note that we observe substantial differences in the base rates (the prevalence of recidivism within each group) across protected group features.

4 METHODOLOGY

4.1 Learning algorithms

Predicting recidivism from demographics, criminal history and SAVRY features can be modeled as a binary classification task. Data for the time between 2002 and 2010 is used as input and recidivism is predicted for a period between release in 2010 and December 31, 2015.

The present data does not have a separate test set. To split the data between training and testing we use k-fold cross validation [37]. In a consequent split, the validation data is chosen from the training set, comprising 10% random elements. The validation set is used to tune the ML model's hyper-parameters and to pick the binarization threshold for the prediction of the ML models.

We test several machine learning algorithms for supervised learning: logistic regression (logit), multi-layer perceptron (mlp), support vector machine with a linear (lsvm) or radial (rsvm) kernel, K-nearest neighbors (knn), random forest (rf), and naive bayes (nb) [37]. For brevity, we report fairness metrics solely for the learning algorithms that achieved the best predictive results in terms of area under the curve (AUC, defined in Section 5.1.2), which correspond to logistic regression ("*logit*" in the following tables and figures) and multi-layer perceptron ("*mlp*").

4.2 Feature sets

Depending on the selected features and the amount of training data, we design four experiments. The first setting, "**Static ML**" corresponds to static features such as demographics and criminal history, such as sex, nationality, the number of prior crimes, the type of crime (full list in Table 2).

The second setting, "**SAVRY ML**" corresponds to all SAVRY features as input features, namely the final expert evaluation, the 24 risk items, the corresponding summary scores, the six protective features, the five average scores on individual characteristics as well as the program that the defendant was in (internment or probation) during the SAVRY assessment. Example SAVRY features include among others: home violence, school achievement, personality (full list in Table 2).

The third setting, "**Static+SAVRY ML**" corresponds to jointly using demographics, criminal history, and SAVRY features.

As baselines, we choose the summed score of all SAVRY risk items, using no machine learning, denoted in the following by "**SAVRY Sum**", in addition to the expert evaluation, denoted by "**Expert**". While SAVRY sum does not represent the final professional judgment, it is a good proxy as a meta-study shows that there is no significant difference between summed scores and professional judgments in risk assessments [12]. Figure 1 supports this finding as it highlights a correlation between the expert assessment and SAVRY sum.

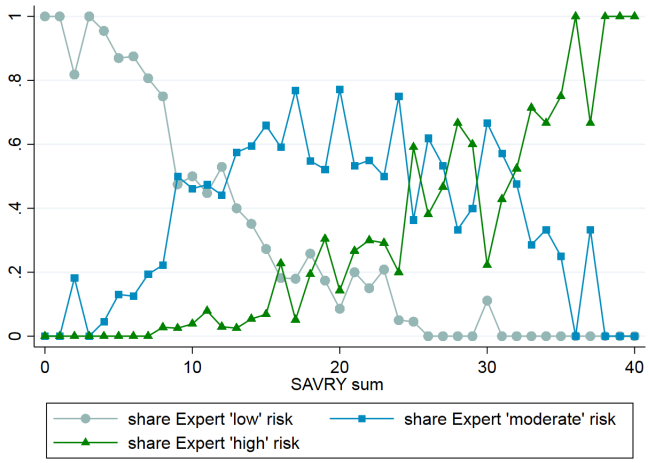

Figure 1: Plot of expert assessment, represented as shares of "low," "moderate," and "high" risk categorization against "SAVRY Sum", the summed score of all 24 SAVRY risk factors. We observe that in general people with a low SAVRY sum get a "low risk" expert evaluation and people with a high SAVRY sum get a "high risk" expert evaluation.

For a fair comparison between the experiments and baselines, the recidivism prediction results are limited to the 855 people for which the SAVRY items are available.

4.3 Feature importance

Trusting the predictions of a ML model is related to its property to explain its decisions, a research field known as ML interpretability [29]. Interpretable ML models have the advantage of explaining how items are classified. In the case of fairness we get more insight on where the ML models go wrong, particularly the features which contribute to unfair classification.

While logit is interpretable and the importance is derived from the coefficients learned by the model, other black-box models, such as mlp, lack this feature. In the latter case, feature importance is obtained with an interpretability framework, LIME [36]. This framework fits a linear model for each data point and offers individual explanations for each data point. Hence, these explanations are

approximations. For further exploration of the ML model, the local explanations can be aggregated to derive the global importance of each feature.

5 EVALUATION

5.1 Experimental setup

5.1.1 Data encoding. To ensure compatibility of data with different ML algorithms, data is encoded numerically. We have various types of input features which have to be treated differently. Numerical values are normalized to have a mean of 0 and standard deviation of 1. Categorical features which have two unique values are encoded as binary. The other categorical features are encoded either numerically if the values represent a scale (e.g. High, Medium, Low are encoded as $1, 0.5, 0$) or using one-hot encoding. The former encodes features as combinations of dummy features and removes any undesired numerical relation between the categories.

5.1.2 Performance evaluation metrics. ML models produce as output a probability of recidivism. To obtain recidivist/non-recidivist labels, a classification threshold has to be applied to this probability. ML systems have different ways of setting the threshold, from the simple 0.5 to more complex objectives which depend on the context: cost-benefit trade-off, maximizing accuracy or any other metrics. Since we are interested in the performance of the metrics we use the threshold values which maximizes balanced accuracy on the validation set defined as $BA(t) = 0.5(TPR(t) + TNR(t))$, where $t = (0, 1)$ is the varying threshold, TPR is the true positive rate, and TNR is the true negative rate. The best threshold t_{max} is obtained for $max(BA)$ on the validation set.

To measure predictive performance we use the area under the ROC curve (AUCROC) which trades-off false positive rate and true positive rate for all the thresholds $t = (0, 1)$.

Note that "SAVRY Sum" comprises a total score of the items, ranged between $(0, 40)$. This score is normalized in the range $(0, 1)$. This value is used to obtain a ROC curve. For a fair comparison with the ML models, this score is thresholded similarly as the probability output of the ML models.

The "Expert" evaluation has three possible risk values: High, Moderate, Low which need to be transformed into binary values. To maximize balanced accuracy, we choose to assign a non-recidivist classification to Moderate/Low expert evaluation label.

5.1.3 Fairness evaluation metrics. We present the results of two fairness measures that are based on different aspects of the classification: demographic parity, and error-rate balance.[2]

All group fairness measures are reported for each protected group $g(a_i)$ with respect to the reference group $g(a_r)$, where $g(a_i)$, $g(a_r)$ represent the group of all defendants with the same protected features a_i or a_r (such as sex or nationality), where $a_i, a_r \in A$. We denote the outcome recidivism as Y, where $Y = 1$ if the defendant recidivated. We denote the number of defendants of group i labeled with $Y = 1$ as LP_i. We denote the number of defendants of group i labeled with $Y = 0$ analogously as LN_i. The predicted outcome is represented by $\hat{Y}$. The ML algorithm classifies someone as high risk for recidivism, i.e. $\hat{Y} = 1$ if the risk score R surpasses a predefined threshold (t), i.e. $R > t$. We denote the number of defendants of group i predicted positive for recidivism as PP_i. We denote the number of defendants of group i predicted negative for recidivism as PN_i. Equivalently, we denote the number of group-specific false positives (FP_i), false negatives (FN_i), true positives (TP_i), and true negatives (TN_i).

Demographic parity [19, 45] means that each person with a protected attribute i has the same likelihood of being classified as recidivist as someone from the reference group with attribute r. We compute demographic disparity (DD) of group i with respect to the reference group

$$DD_i = \frac{PP_i/g(a_i)}{PP_r/g(a_r)} \tag{1}$$

$DD_i = 1$ means that someone from group i is just as likely to be classified as recidivist as someone from the reference group. $DD_i = 2$ means that someone with attribute a_i is twice as likely to be classified as recidivist as someone from the reference group with attribute a_r.

Error rate balance [13] comprises two measures of fairness: equal false negative rates and equal false positive rates. This means that each person with a protected attribute i has the same likelihood of falsely being classified as recidivist (or non-recidivist) as someone from the reference group with attribute a_r. We compute the false positive rate and false negative rage of group i (FPR_i,FNR_i), from which we derive false positive rate disparity and false negative rate disparity of group i ($FPRD_i$,$FNRD_i$):

$$\begin{aligned} FPR_i &= FP_i/LN_i & FPRD_i &= FPR_i/FPR_r \\ FNR_i &= FN_i/LP_i & FNRD_i &= FNR_i/FNR_r \end{aligned} \tag{2}$$

As an example, $FPRD_i = 2$ means that someone with attribute a_i is twice as likely to be wrongly classified as recidivist as someone from the reference group with attribute a_r.

5.1.4 Parameters and model selection. We pick with $k = 10$ for k-fold cross validation. Each fold is replicated 50 times for a different seed which controls the initialization of the parameters and the random split between training, validation, and testing.

For each seed we determine the best hyper-parameters for the ML algorithms. We train 30 models for each ML algorithm representing different random combinations of hyper-parameters. For logit we pick the inverse of regularization strength from an uniform distribution $\mathcal{U}(0.1, 10)$. For mlp, we use a two layer network with the sizes $(F, L * F), (L * F, (L + 1) * F), (L * F, 1)$, where F is the number of input features and L is chosen randomly from an uniform distribution $\mathcal{U}(1, 10)$. In addition we experimentally determined the batch size to be 64, we update parameters using the stochastic gradient descent for 100 epochs. The cost function for mlp classification is binary cross entropy, with an $\mathcal{L}_2$ penalty on weights of 0.01 to avoid over-fitting. For knn the number of neighbors and the distance metrics are picked randomly between $(3, 20)$ and between Minkowski, Euclidean and Manhattan. For the svm we trained a linear and circular kernel separately. The kernel radius and gamma are drawn from uniform distributions $\mathcal{U}(0.1, 10)$. For the rf we randomly pick the number of estimators to be between $(10, 50)$, the

[2]We computed and looked at eleven further measures of fairness but found these to be highly correlated. This is in line with findings of [20].

Table 1: AUCROC for each experiment and for the ML models, including mean and standard deviations aggregated across 50 random seeds. For comparison, the baselines "SAVRY Sum" and 'Expert" achieve AUCROC of .64 and .66, . The scores for the top two ML methods are marked in boldface.

	logit		**mlp**		knn		lsvm		rsvm		nb		rf	
	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.	mean	std.dev.
SAVRY ML	**.66**	.0058	**.66**	.0058	.60	.0121	.65	.0082	.52	.0197	.65	.0015	.65	.0110
Static ML	**.70**	.0055	**.70**	.0068	.62	.0122	.61	.0119	.56	.0149	.69	.0040	.66	.0110
Static+SAVRY ML	**.71**	.0064	**.70**	.0053	.64	.0129	**.71**	.0074	.50	.0058	.69	.0018	.69	.0121

maximum depth between $(5, 50)$ and the minimum number of samples per leaf between $(1, 10)$. The best model for each ML algorithm is the one having the highest AUCROC on the validation set.

5.1.5 Software implementation details. The experiments are replicated 50 times for different seeds to ensure robustness and reproducibility. The code is implemented in Python using libraries such as pandas and sklearn-pandas for data processing, sklearn and pytorch for machine learning, numpy and scipy for numerical processing. This research complies with research reproducibility principles. Code and data are made available as a part of a framework [3].

5.2 Results

5.2.1 Predictive Performance. The results in terms of AUCROC are presented in Table 1. The values of AUCROC of the best performing ML methods are typical of recidivism prediction using ML methods as reported in the literature: 0.67 for a 5-variables random forest classifier [21], 0.68-0.71 for COMPAS [33], 0.65-0.66 for the Public Safety Assessment [17], 0.57-0.74 in a meta-study of various risk assessment used in the US [18].

We compare "SAVRY ML" experiment with the other experiments with and without non-SAVRY and SAVRY features. We observe that not including demographic and criminal history features decreases the accuracy across all methods with values between $(.01, 0.05)$ points. Although informative for an evaluator, the SAVRY features are less useful for ML methods in determining if a person will recidivate.

Combining features derived from SAVRY items with static demographics and criminal history, or increasing the size of the training set yields better AUC across several learning algorithms (logit, mlp, knn, lsvm, rf). As expected, data-driven methods benefit from including more features or more data in training.

5.2.2 Fairness. Besides predictive performance we are interested in measuring if the ML methods are discriminating. We analyse fairness across the protected features sex and nationality. We select the top two ML models across all experiments to compare their fairness performance with "SAVRY sum" and "Expert". In Figures 2, 3 and 4 we report the group metrics described in Section 5.1.3. Error bars represent 95% confidence intervals across the 50 seeds. The experiments "SAVRY ML", "Static ML", "Static+SAVRY ML" are separated by vertical blue lines. We delimit thresholds for discrimination with horizontal red lines between $(0.8, 1.2)$ similarly to Propublica's COMPAS analysis[11]. Note that these thresholds are purely informative. They are derived from US laws and they do not hold any legal status in Catalonia.

[3]HUMAINT repository: https://gitlab.com/HUMAINT/humaint-fatml.

Since the disparity measures are obtained by dividing the corresponding measures for the groups with the ones of the reference group, we do not plot the results for the reference group.

Figure 2: Comparison of group fairness metrics using sex as the protected attribute. The reference group are men.

The results in terms of sex are displayed in Figure 2. In this case, the reference group are men. While "SAVRY Sum" is within the fairness bounds in terms of FPRD, the expert evaluation, and the ML methods using mlp and logit as learning algorithms are less likely to erroneously label females as recidivists than men. In addition, in "SAVRY ML", women are less likely to be wrongly classified as non-recidivists, having lower FNRD, just above the acceptable 0.8 threshold. The ML methods, while within the acceptable range when using SAVRY features, become discriminatory when using demographic features, with women being more likely to be classified as non-recidivists.

Across all the three metrics we observe that training on static non-SAVRY features accentuates the disparity between the two groups, with small differences depending on the learning algorithm used (logit slightly improves if SAVRY features are added to demographics while this does not happen for mlp). Further, because it considers all the people in a group labeled as recidivists, and not only the falsely labeled, DD is stricter than FPRD. Consequently, DD follows the same trend as FPRD across all experiments with lower results.

The results for using nationality as a protected attribute are displayed in Figure 3. In this case, the protected group are foreigners

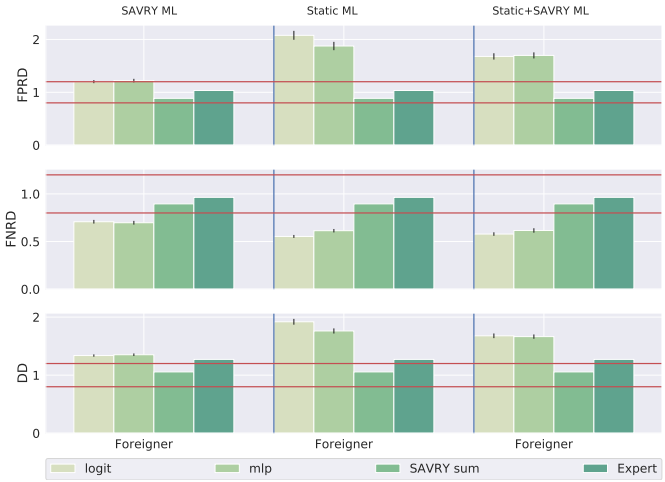

Figure 3: Comparison of group fairness metrics in terms of nationality. The reference group are Spanish nationals.

and the reference group are Spanish nationals. We observe that ML methods have higher disparity than the "SAVRY Sum" and the expert evaluation across all metrics. Foreigners are more likely to be falsely labeled as recidivist (FPRD), they are less likely be labeled as non-recidivists (FNRD) and their proportion of individual labeled as recidivists is higher (DD).

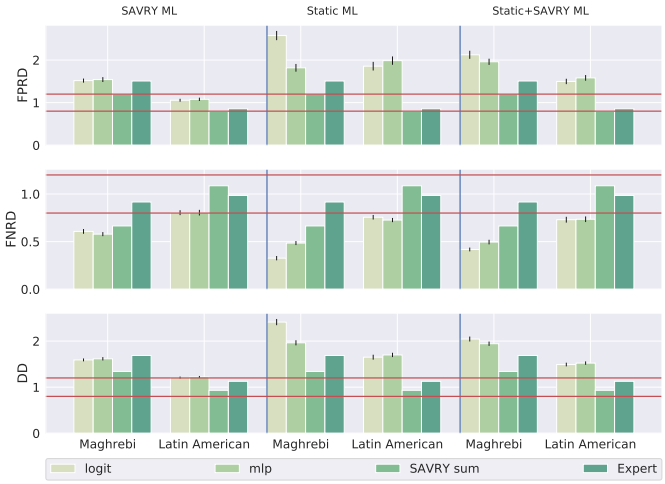

Figure 4: Comparison of group fairness metrics in terms of national groups. The reference group are Spanish nationals.

In this data foreigners are subsequently split into subgroups: Maghrebi, Latin American, European and Other. While a system is fair towards a group, it can discriminate towards a particular subgroup and positively discriminate towards other subgroups. Thus, in Figure 4 we look at fairness for the nationality subgroups. Since European (37 people) and Others (13 people) are relatively small groups and flipping the label on one individual drastically changes the fairness outcomes, we exclude these groups from the analysis, noting that they are more likely to be positively discriminated across all metrics. We observe that for "SAVRY Sum" the Maghrebi are discriminated in terms of FPRD, DD, FNRD, across all experiments, while this assessment is more fair towards Latin Americans. Moreover, "SAVRY Sum" has higher disparity towards Maghrebi which are less likely to be falsely labeled as non-recidivists, and have a higher proportion of individuals labeled as recidivists. The "Expert" evaluation has more disparity with respect to Maghrebi than the "SAVRY Sum"(higher FPRD and DD) , while it is within the acceptable boundaries when looking at FNRD.

ML methods yield more disparity towards Maghrebi and Latin Americans for all metrics when including non-SAVRY features in training. Training with savry items has slightly higher disparity than the "SAVRY Sum". This disparity is within the acceptable bounds for Latin Americans and surpassing the bounds for Maghrebi. For all experiments and all protected groups logit is more unfair than mlp when trained on non-SAVRY features.

6 DISCUSSION

Throughout the analysis we are faced with a tradeoff between predictive performance and group fairness: the application of ML over the risk factors yields a more accurate prediction. Yet, it introduces issues of group fairness that a simple SAVRY sum does not have. This is more pronounced if we also include non-SAVRY factors. In this case, the accuracy further increases, but problems of group fairness become more severe as the error rate disparity increases.

These findings have to be interpreted in light of some limitations. First, despite the relatively random selection of the sample by release year in 2010, we find that the selection into the SAVRY assessment with 855 defendants is on average targeted to defendants with a higher violence risk. Still, the sample remains fairly heterogeneous. The second problem is sample bias. The outcomes of the ML analysis could mostly be driven by the largest protected group, in this case Spanish males. However, we repeated ML analysis allowing for group-specific features and found no substantial differences to the baseline analysis. Finally, we have to consider the potential for measurement error because we measure recidivism with rearrests. Yet, the base rate of a particular minority group could be upwards biased if policing tends to be more strict with this protected group. Despite these data limitations it is important to understand how ML methods can introduce unfairness in seemingly fair assessments and potential mitigation measures.

6.1 Sources of Group Unfairness

Predicting recidivism is not a trivial problem and no ML method achieves perfect accuracy. The selected top two ML models trade off predictive performance for fairness. One explanation can be that the base rates (i.e. the prevalence of recidivism) differ between various groups, as seen in the top panel of Table 2. The literature has shown extensively, that base rates substantially affect the outcomes of group-fairness measures [5]. Most prominently, [27]'s and [13]'s proofs show that, when base rates differ, it is mathematically impossible to fulfill multiple measures of group fairness simultaneously. In this dataset, the recidivism rate for men is 40%, while the recidivism rate for women is 20%. Also, the recidivism rate for foreigners is 46%, (specifically for Latin Americans it's 44.5% and for Maghrebi it's even 55%), while for Spanish nationals it is 32%.

This is emphasized by the differences in the group composition between recidivists and non-recidivists. In detail, Table 2 shows that compared to non-recidivists, recidivists are significantly more likely to be male, foreign and specifically Maghrebi or Latin American but less likely to be female or Spanish. ML methods pick up on these empirical correlations when producing predictions on recidivism. Under these conditions, it is clearly difficult to achieve similar classification rates for both groups.

As discussed in Section 4.3, we get additional insight on the possible sources of ML unfairness if we look at which features are important for the ML models for each of the experiments "SAVRY ML", "Static ML", "Static+SAVRY ML". To derive the importance of the features used for training we use the LIME framework [36] for mlp or the coefficients learned by the model in the case of logit. Here we present the first ten highly ranked features in terms of coefficients for logit (Table 3) and LIME importance for mlp (Table 4). The results are averaged for 50 seeds for which mean and standard deviation are reported.

As shown in Table 3, the most important features for logit in the "Static + SAVRY ML" setting are almost all static features, including sex and whether the defendant is Maghrebi or Latin American. The only relevant SAVRY feature in this column is the final overall evaluation by the expert which is also the only significant feature in the "SAVRY ML" setting. Similarly, in Table 4 for "Static+SAVRY ML" all static features are more important than SAVRY features. That is, although the SAVRY features are empirically relevant in the prediction of recidivism (as shown in Table 2), we find that they are not as predictive of recidivism in comparison with static non-SAVRY features.

We further explore the reason why "SAVRY Sum" does not seem to exhibit large differences between groups, by looking at differences in the 24 risk factors and in SAVRY sum, as shown on Table 5 for the case of Spaniards versus foreigners (the case of men and women is similar). For SAVRY Sum, on average foreigners obtain slightly higher risk scores than Spaniards, but the difference is small: only 1 point on average, out of a maximum of 40 points within the sample and a potential maximum of 48 points. Additionally, the 24 risk items show only small differences and there is a mixture of items for which Spaniards get higher scores and items for which foreigners get higher scores.

6.2 Potential Mitigation Measures

A consequential next step to this analysis is to look for methods that mitigate unfairness in the ML methods and at the same time maintain the accuracy gains. It is important to note here, that 'mitigating unfairness' in this case means mitigating the unfairness according to the definitions used in this context. However, the fact that base rates differ so substantially can already be the result of years of structural discrimination long before the present data was recorded [22]. This type of discrimination would not be resolved even if we mitigate differences in error rates. Even if we apply mitigation measures, they will each bring their own problems.

6.2.1 "Color blind" methods: A first potential mitigation measure is to remove the protected attributes from the input. This do not affect the difference between "SAVRY Sum" and "SAVRY ML," as none of these settings use protected attributes, but can help in the case of "Static ML" and "Static + SAVRY ML." However, in general methods designed to avoid disparate treatment (i.e., being blind to sensitive attributes) do not guarantee that disparate impact will be absent [14, 30, 46]. Even if we remove the protected attributes from the training of the algorithm, there are still multiple other attributes in the feature set that can be correlated with the protected attributes [4, 24]. That is, if the protected attributes and the outcome of recidivism correlate (see Table 2), this correlation will not disappear if you remove the protected features. Besides, the results for "Savry ML" in Figure 3 and Figure 4 for Maghrebi show that just using ML methods, without including protected attributes, produces unfair outputs.

6.2.2 Different models or different thresholds: Using the same model with the same threshold for two different populations might not be advisable, as this could be harmful for specific groups. Particularly with respect to sex this could be problematic [40], as it has been shown that in juvenile justice females react differently from males to specific risk factors [43]. However, having different thresholds for different protected groups can be problematic, too. For instance, trying to equalize error rates by applying different thresholds to the same risk score could yield higher error rates. More specifically, in this case it could result in classifying potentially high risk people as low risk (and not prosecuting them) or potentially classifying low risk people as high risk (and consequently assigning to interventions [13]. This action produces public costs that will either have to be paid in the form of sacrificing public safety or making innocent people subject to costly criminal justice interventions [15].

6.2.3 Algorithm adjustments: In-processing methods to achieve fairness in machine learning modify the objective that is optimized during learning. In this case, one could introduce an extra term that penalizes classifiers that yield different error rates [2, 24, 44]. However, 'adjusting' a classifier without understanding its underlying mechanisms can in turn lead to discriminatory predictions. For instance, enforcing demographic parity might not be compatible with notions of justice, particularly if we consider that base recidivism rates are different. that is, if the adjusted method classifies more people from a racial minority group with a high base rate as low risk, these people remain without any criminal justice intervention, and then they go on to recidivate. These mechanisms could lead to a bad track record for specific minority groups, creating downwards dynamics in terms of risk assessment [25, 31]. Therefore, it is crucial to take these general equilibrium effects into account [14].

7 CONCLUSIONS AND FUTURE WORK

We discussed the problems with using ML in recidivism prediction juvenile justice in Catalonia. Having better predictive performance, the ML methods, logit and mlp, are discriminating across sex and national groups across the chosen fairness benchmarks. The disparity associated with ML models is observed regardless of the training features and increases when using demographic and personal history features. This is strongly related to having different base rates for different groups in the training data. Our analysis of feature importance shows that when combining static features with SAVRY features, ML models rely more on the former.

This research does not propose a method to mitigate disparity. However, we discuss the possible mitigation methods in line with the sources of unfairness in Section 6.2. As a future task, we plan to address mitigation methods that consider particular issues of this dataset and the respective domain. We plan to include in future experiments datasets containing SAVRY assessments which are obtained in other countries. Furthermore, we consider extending our work to predicting adult criminal recidivism.

ACKNOWLEDGEMENTS

We would like to thank Prof. Antonio Andres Pueyo for his insight on the youth criminal justice system in Catalonia.

REFERENCES

[1] S Ægisdóttir 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 (2006), 341–382.
[2] A Agarwal, A Beygelzimer, M Dudík, J Langford, and H Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).
[3] J Angwin, J Larson, S Mattu, and L Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).
[4] S Barocas and A Selbst. 2016. Big Data ' s Disparate Impact. *California law review* 104, 1 (2016), 671–729. https://doi.org/10.15779/Z38BG31
[5] R Berk, H Heidari, S Jabbari, M Kearns, and A Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. (2017), 1–42. arXiv:1703.09207 http://arxiv.org/abs/1703.09207
[6] K Björkqvist, KM Lagerspetz, and A Kaukiainen. 1992. Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive behavior* 18, 2 (1992), 117–127.
[7] M Blanch, M Capdevila, M Ferrer, B Framis, Ú Ruiz, J Mora, A Batlle, and B López. 2017. La reincidència en la justícia de menors. CEJFE.
[8] R Borum. 2006. Manual for the structured assessment of violence risk in youth (SAVRY). (2006).
[9] Randy Borum, Patrick A Bartel, and Adelle E Forth. 2005. Structured assessment of violence risk in youth. *Mental health screening and assessment in juvenile justice* (2005), 311–323.
[10] T Brennan and W Dieterich. 2018. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). *Handbook of Recidivism Risk/Needs Assessment Tools* (2018), 49.
[11] T Brennan and WL Oliver. 2013. The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. *Criminology & Public Policy* 12, 3 (2013), 551–562.
[12] C S Chevalier. 2017. *The Association between Structured Professional Judgment Measure Total Scores and Summary Risk Ratings: Implications for Predictive Validity.* Ph.D. Dissertation.
[13] A Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[14] S Corbett-Davies and S Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
[15] S Corbett-Davies, E Pierson, A Feller, S Goel, and A Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.
[16] S Danziger, J Levav, and L Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.
[17] M DeMichele, P Baumgartner, M Wenger, K Barrick, M Comfort, and S Misra. 2018. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky. (2018).
[18] S L Desmarais, K L Johnson, and J P Singh. 2016. Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services* 13, 3 (2016), 206.
[19] C Dwork, M Hardt, T Pitassi, O Reingold, and R Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[20] S A Friedler, C Scheidegger, S Venkatasubramanian, S Choudhary, EP Hamilton, and D Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422* (2018).
[21] B Green and Y Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *ACM Conference on Fairness, Accountability, and Transparency*.
[22] B Green and L Hu. 2018. The myth in the methodology: towards a recontextualization of fairness in machine learning. In *Proceedings of the Machine Learning: The Debates Workshop*.
[23] L S Guy. 2008. *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey.* Ph.D. Dissertation. Dept. of Psychology-Simon Fraser University.
[24] M Hardt, E Price, N Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[25] N Kallus and A Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. *arXiv preprint arXiv:1806.02887* (2018).
[26] J Kleinberg, H Lakkaraju, J Leskovec, J Ludwig, and S Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2017), 237–293.
[27] J Kleinberg, S Mullainathan, and M Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.
[28] P Langley and H A Simon. 1995. Applications of machine learning and rule induction. *Commun. ACM* 38, 11 (1995), 54–64.
[29] ZC Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).
[30] Za Lipton, J McAuley, and A Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*. 8136–8146.
[31] LT Liu, S Dean, E Rolf, M Simchowitz, and M Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).
[32] A Narayanan. 2018. *21 fairness definitions and their politics.* Technical Report. Conference on Fairness, Accountability and Transparency 2018, Tutorial.
[33] Northpoint, Inc. 2012. *COMPAS Risk and Need Assessment System.* Technical Report. Northpoint, Inc.
[34] M E Olver, K C Stockdale, and JS Wormith. 2009. Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior* 36, 4 (2009), 329–353.
[35] R T Perrault, G M Vincent, and L S Guy. 2017. Are risk assessments racially biased?: Field study of the SAVRY and YLS/CMI in probation. *Psychological assessment* 29, 6 (2017), 664.
[36] M T Ribeiro, S Singh, and C Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.
[37] C Robert. 2014. Machine learning, a probabilistic perspective.
[38] D C Rowe, A T Vazsonyi, and D J Flannery. 1995. Sex differences in crime: Do means and within-sex variation have similar causes? *Journal of research in Crime and Delinquency* 32, 1 (1995), 84–100.
[39] J P Singh 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 (2014), 193–206.
[40] J Skeem, J Monahan, and C Lowenkamp. 2016. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior* 40, 5 (2016), 580.
[41] S Tolan. 2018. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. *JRC Digital Economy Working Paper* 10 (2018).
[42] G M Vincent, L S Guy, B G Gershenson, and P McCabe. 2012. Does risk assessment make a difference? Results of implementing the SAVRY in juvenile probation. *Behavioral sciences & the law* 30, 4 (2012), 384–405.
[43] E M Wright, E J Salisbury, and P Van V. 2007. Predicting the prison misconducts of women offenders: The importance of gender-responsive needs. *Journal of Contemporary Criminal Justice* 23, 4 (2007), 310–340.
[44] M B Zafar, I Valera, M Gomez Rodriguez, and K P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.
[45] I Žliobaitė. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723* (2015).
[46] I Žliobaitė and B Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (2016), 183–201.

Table 2: Descriptive Statistics

		Not Recidivated		Recidivated		Difference	
	Base rate	Mean	Std.Dev.	Mean	Std.Dev.	Diff	Std.Dev.
protected features							
male	40.03%	0.839	0.368	0.931	0.253	0.093***	0.021
female	20.37%	0.161	0.368	0.069	0.253	-0.093***	0.021
Spanish	32.06%	0.667	0.471	0.523	0.499	-0.143***	0.035
foreign	46.22%	0.333	0.471	0.477	0.499	0.143***	0.035
Latin American	44.52%	0.161	0.368	0.215	0.411	0.054*	0.028
Maghrebi	55.12%	0.107	0.309	0.218	0.413	0.111***	0.027
European	32.35%	0.043	0.203	0.034	0.182	-0.009	0.013
other	20.00%	0.022	0.148	0.009	0.096	-0.013	0.008
Static features (not SAVRY)							
age maincrime		16.011	1.009	15.720	1.060	-0.292***	0.074
prior crimes		0.700	0.458	0.863	0.344	0.163***	0.028
prior crimes frequency							
1 incident		0.586	0.493	0.604	0.489	0.018	0.035
2 incidents		0.243	0.429	0.215	0.411	-0.028	0.030
3 or more incidents		0.170	0.376	0.181	0.385	0.010	0.027
maincrime violent		0.609	0.488	0.611	0.488	0.002	0.034
maincrime category							
nonviolent against property		0.251	0.434	0.265	0.441	0.014	0.031
violent against property		0.264	0.441	0.293	0.455	0.029	0.032
against persons		0.345	0.475	0.318	0.466	-0.027	0.033
other		0.140	0.347	0.125	0.330	-0.016	0.024
maincrime program sentence							
technical sentence		0.060	0.237	0.262	0.440	0.202***	0.027
mediation and reparation		0.021	0.142	0.025	0.156	0.004	0.011
enforcement measure		0.919	0.272	0.713	0.452	-0.206***	0.028
internment (no probation)		0.142	0.349	0.265	0.441	0.122***	0.029
days to sentence start		481.803	269.611	364.579	276.429	-117.224***	19.343
sentence duration (days)		285.058	190.887	235.536	233.089	-49.522***	15.411
Year of main crime							
2006 or earlier		0.064	0.244	0.069	0.253	0.005	0.018
2007/2008		0.672	0.469	0.449	0.497	-0.224***	0.034
2009/2010		0.264	0.441	0.483	0.5	0.219***	0.034
SAVRY							
final expert evaluation		0.315	0.330	0.530	0.359	0.215***	0.025
SAVRY summary scores							
total (automatic)		14.150	8.424	18.262	8.712	4.112***	0.608
historical factors		5.762	3.941	7.084	3.890	1.322***	0.276
social factors		3.803	2.562	5.050	2.701	1.246***	0.187
individual factors		4.584	3.438	6.128	3.703	1.543***	0.255
protective factors		2.152	1.861	2.956	1.877	0.805***	0.132
SAVRY 24 risk items							
previous violent offenses		0.428	0.404	0.536	0.402	0.108***	0.028
history nonviolent offending		0.344	0.368	0.472	0.386	0.128***	0.027
early violence (below 14)		0.182	0.326	0.254	0.364	0.072***	0.025
past intervention failures		0.184	0.317	0.280	0.365	0.097***	0.025
self-harm/suicide history		0.099	0.248	0.132	0.268	0.033*	0.018
home violence		0.254	0.383	0.263	0.379	0.009	0.027
childhood mistreatment		0.239	0.352	0.290	0.379	0.051**	0.026
criminal parent/caregiver		0.163	0.310	0.196	0.347	0.033	0.024
childhood care giving disruption		0.285	0.389	0.335	0.402	0.050*	0.028
poor school achievement		0.705	0.351	0.783	0.319	0.078***	0.023
delinquency in peer group		0.364	0.362	0.525	0.365	0.161***	0.026
rejection by peer group		0.110	0.230	0.154	0.282	0.045**	0.019
stress and poor coping		0.390	0.350	0.438	0.373	0.048*	0.026
poor parental management		0.456	0.351	0.578	0.368	0.122***	0.026
lack of personal/social support		0.286	0.340	0.419	0.380	0.133***	0.026
community disorganization		0.297	0.381	0.411	0.394	0.114***	0.027
negative attitudes		0.279	0.304	0.397	0.326	0.118***	0.022
risk taking/impulsive		0.369	0.343	0.469	0.349	0.100***	0.025
substance abuse		0.317	0.347	0.416	0.371	0.098***	0.026
anger management issues		0.334	0.340	0.410	0.341	0.075***	0.024
low empathy		0.282	0.326	0.393	0.342	0.111***	0.024
attention deficit hyperactivty		0.202	0.297	0.262	0.323	0.059***	0.022
poor compliance		0.202	0.294	0.313	0.348	0.111***	0.023
low commitment to school		0.306	0.366	0.405	0.402	0.099***	0.027
SAVRY 6 protective factors							
pro-social activities		0.485	0.500	0.333	0.471	-0.152***	0.034
pro-social support		0.700	0.458	0.536	0.499	-0.165***	0.034
pro-social support (by adult)		0.676	0.468	0.592	0.491	-0.084**	0.034
positive attitude		0.837	0.369	0.741	0.438	-0.096***	0.029
high interest in school/work		0.669	0.471	0.508	0.500	-0.161***	0.035
positive/resilience characteristics		0.481	0.500	0.333	0.471	-0.148***	0.034
SAVRY 5 factors model							
antisocial behavior		0.548	0.449	0.747	0.445	0.199***	0.032
family dynamics		0.470	0.527	0.542	0.558	0.072*	0.039
personality		0.561	0.422	0.720	0.446	0.159***	0.031
social support		0.540	0.425	0.738	0.468	0.198***	0.032
treatment susceptibility		0.565	0.351	0.727	0.377	0.162***	0.026
N		534		321			

Descriptive statistics of input features by recidivism status. Not displayed: province of residence, province of sentencing

Table 3: Feature importance for logistic regression

SAVRY ML			Static			Static+SAVRY		
final expert evaluation	0.370***	(0.076)	✓crime in 07-08	-0.298**	(0.118)	✓crime in years 07-08	-0.272**	(0.133)
SAVRY sum	0.183	(0.910)	✓crime in year 09	-0.259**	(0.121)	✓crime in year 09	-0.255*	(0.132)
personality	-1.362	(7.061)	✓age maincrime	-0.109***	(0.021)	✓days to program start (norm)	-0.117***	(0.044)
treatment susceptibility	-1.340	(6.336)	✓days to program start (norm)	-0.105***	(0.040)	✓age maincrime	-0.115***	(0.022)
total score (social)	-0.141	(0.909)	✓crime in year 10	-0.275***	(0.098)	final expert evaluation	0.291***	(0.091)
total score (protective)	0.191	(0.902)	✓days in program (norm)	-0.087*	(0.048)	✓crime in year 10	-0.256**	(0.115)
previous violent offenses	-0.601	(2.533)	✓prog: enforcement measure	-0.248**	(0.103)	✓female	-0.196***	(0.053)
total score (historic)	0.056	(0.045)	✓prior crimes frequency	0.059*	(0.033)	✓enforcement measure	-0.206*	(0.122)
home violence	-0.543	(1.816)	✓female	-0.187***	(0.046)	✓Maghrebi	0.152**	(0.069)
past intervention failures	-0.598	(2.530)	✓Maghrebi	0.158***	(0.058)	✓Latin American	0.135**	(0.060)
			✓Latin American	0.105**	(0.052)			
			✓prog: mediation/reparation	-0.178*	(0.103)			
Pseudo R^2	0.096		Pseudo R^2	0.146		Pseudo R^2	0.199	

Marginal effects of relevant features in logisitic regression. Standard errors in parentheses. *, ** and *** denote significance level of 10%, 5% and 1%, respectively. ✓indicates static features not in SAVRY. Bootstrapped regression with 200 repetitions. Features are ranked in two steps: (1) a minimum confidence level of 90%, (2) fully standardized coefficients by features and outcome variable.

Table 4: Feature importance for MLP using LIME

SAVRY ML			Static ML			Static+SAVRY ML		
feature	importance		feature	importance		feature	importance	
	Mean	StdDev		Mean	StdDev		Mean	StdDev
probation/internment	147.43	24.85	✓province of residence	219.21	28.44	✓foreigner	199.80	11.37
total score (social)	117.93	9.71	✓age maincrime	202.83	25.72	✓sex	188.07	8.35
total score (personality)	117.63	9.83	✓foreigner	178.38	19.06	✓national group	117.40	23.09
total score (protective)	115.76	8.56	✓year of maincrime	168.96	13.86	✓maincrime category	150.90	16.44
total score (historic)	116.59	10.25	✓prior crimes	175.11	22.56	✓prior crimes frequency	151.53	18.26
history non-violent offending	112.17	7.44	✓national group	181.68	32.23	✓maincrime program sentence	143.29	10.50
positive/resilience characteristics	111.62	7.32	✓prior crimes frequency	156.15	20.98	✓year of maincrime	141.88	9
previous violence	113.22	8.93	✓maincrime category	144.27	18.26	✓maincrime violent	148.92	16.23
early violence	111.42	7.17	✓maincrime violent	137.20	14.95	✓province of execution	146.07	13.76
pro-social activities	109.82	5.57	✓prior crimes	131.53	12.66	✓prior crimes	146.97	14.71

Table 5: SAVRY items by foreigner status

	Spaniard		Foreigner		Difference	
	Mean	Std.Dev.	Mean	Std.Dev.	Diff	Std.Dev.
total score	15.281	8.387	16.347	9.289	1.067*	0.628
previous violent offenses	0.440	0.399	0.514	0.414	0.074**	0.029
history nonviolent offending	0.385	0.377	0.402	0.383	0.016	0.027
early violence (below 14)	0.196	0.327	0.230	0.365	0.034	0.025
past intervention failures	0.200	0.327	0.251	0.355	0.050**	0.024
self-harm/suicide history	0.102	0.244	0.127	0.273	0.025	0.018
home violence	0.262	0.385	0.249	0.376	-0.013	0.027
childhood mistreatment	0.244	0.354	0.279	0.376	0.035	0.026
criminal parent/caregiver	0.193	0.340	0.148	0.297	-0.045**	0.022
childhood care giving disruption	0.252	0.376	0.385	0.408	0.133***	0.028
poor school achievement	0.745	0.335	0.718	0.351	-0.028	0.024
delinquency in peer group	0.373	0.345	0.506	0.396	0.133***	0.026
rejection by peer group	0.125	0.253	0.128	0.251	0.003	0.018
stress and poor coping	0.413	0.360	0.399	0.360	-0.014	0.025
poor parental management	0.507	0.358	0.494	0.369	-0.013	0.026
lack of personal/social support	0.318	0.345	0.364	0.384	0.046*	0.026
community disorganization	0.318	0.384	0.375	0.397	0.057**	0.028
negative attitudes	0.309	0.304	0.346	0.337	0.037	0.023
risk taking/impulsive	0.400	0.337	0.417	0.367	0.017	0.025
substance abuse	0.344	0.337	0.372	0.392	0.028	0.026
anger management issues	0.385	0.336	0.326	0.349	-0.059**	0.024
low empathy	0.300	0.318	0.361	0.361	0.061**	0.024
attention deficit, hyperactivity issues	0.260	0.320	0.168	0.280	-0.093***	0.021
poor compliance	0.226	0.304	0.272	0.341	0.046**	0.023
low commitment to school	0.343	0.374	0.344	0.397	0.002	0.027