CRIMINAL RECIDIVISM. TOWARDS RELIABLE AND TRANSPARENT PREDICTIVE MODELS

Flavia Tagliafierro, Claudio Caterino

Abstract. The prediction of criminal recidivism through machine learning (ML) models raises significant ethical, legal, and methodological challenges. This article promotes for a transparency and explainability-oriented approach by comparing three predictive models – logistic regression, random forest, and neural networks – applied to the COMPAS dataset of criminal history data, released by ProPublica a non-profit journalism organization in USA. To assess the coherence and readability of algorithmic decisions interpretability techniques such as SHAP values are employed. The analysis also considers the implications of adjusting the decision threshold to increase false positives for supportive – rather than punitive – purposes, emphasizing the greater ethical and social acceptability of such a strategy. The discussion is complemented by an overview of the regulatory developments in Italy and the European Union regarding the use of predictive technologies in the criminal justice system.

1. Introduction

Recidivism is broadly defined as the tendency of previously convicted individuals to reoffend, serving as a key indicator of risk, social dangerousness, and the effectiveness of penal and rehabilitative measures (Baratta, 1998; Lappi-Seppälä, 2003). Identifying the factors associated with reoffending as well as understanding their interactions can support prevention strategies, resource allocation, and evidence-based decision-making in the justice system.

Definitions and measurements of recidivism vary across jurisdictions. Some systems consider only new convictions while others include arrests or reports (Weatherburn *et al.*, 2003). In Italy, recidivism is regulated by articles 99–105 of the Penal Code which influence sentencing and enforcement. International organizations such as the United Nations Office on Drugs and Crime (UNODC, 2018) and the Council of Europe (SPACE, 2023) adopt different indicators depending on whether the focus is on policy evaluation, reintegration, or risk assessment.

Initial predictive efforts relied on static models using demographic and criminal history data (Burgess, 1928; Glueck and Glueck, 1950). Since the 1970s, attention shifted toward dynamic models that incorporate contextual and modifiable variables (Andrews and Bonta, 2010). In the last two decades, the development of machine

learning (ML) has enabled more complex, yet less transparent, systems (Berk *et al*., 2018), underscoring the need for interpretable tools to guarantee transparency for legal actors and the public.

From a legal perspective, there is a clear divergence between common law and civil law systems. In Italy, Article 220 of the Code of Criminal Procedure prohibits expert assessments on personality or criminal propensity during trial, thereby restricting the use of AI tools to post-sentencing phases. Conversely, in common law systems such as the U.S., tools like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) are routinely integrated into judicial decision-making in several States. Common law jurisdictions are increasingly adopting principles of fairness, accountability, and transparency (Citron and Pasquale, 2014), whereas civil law countries require a more cautious, rights-based approach (Floridi *et al.*, 2018). These institutional differences shape how ML systems are introduced into the justice sector and raise important questions about balancing predictive efficiency with the protection of individual rights.

The European Union's AI Act (Regulation EU 2024/1689), adopted on August 1, 2024, establishes a harmonized framework based on four levels of risk. Recidivism prediction tools categorized as "high risk" are subject to rigorous standards concerning transparency, reliability, fairness, and continuous oversight. While the AI Act seeks to facilitate the European digital single market, it also allows national authorities to tailor implementation to local legal traditions, ensuring respect for fundamental rights.

2. Data and Methods

This study relies on the COMPAS dataset published by a non-profit investigative journalism organization (Propublica, 2016), which contains information more than 7,000 individuals arrested in Broward County, Florida. It includes demographic data, criminal history, risk scores and a binary recidivism outcome within two years. We focused on key features: age, decile score, prior offenses, ethnicity (origins), gender, and offense type. Moderate linear correlations are observed between age and decile score (0.39), prior offenses and decile score (0.43), decile score and recidivism (0.35). The decile score, a composite risk indicator of recidivism correlates with reoffending as expected. The dataset showed issues of racial bias (Angwin *et al.*, 2016) and misclassification risks (Venkataraman, 2025). Instead of predicting individual risk scores, our aim is to investigate the relative importance of features and their contribution to the overall decision-making process. Despite ethical and legal constraints (Rudin *et al*., 2020), we argue that ML can elucidate complex crime patterns in data on convicts/detainees and guide targeted interventions.

Our approach combines a data-driven methodology using machine learning models like Random Forests and Neural Networks for predictive pattern detection, alongside interpretable models such as Logistic Regression to enhance transparency. Our goal is to critically evaluate the reliability and interpretability of these models. All models were implemented in Python using specific packages: *sklearn.preprocessing, model_selection, sklearn.metrics, sklearn.linear_model, sklearn.ensemble, keras, sklearn.cluster, kmodes.prototypes, shap*.

Because the dataset contains features with different units of measure, variables were standardized to ensure comparability. In our preprocessing, we applied min-max normalization. Given potential heterogeneity in model performance, maximizing accuracy supports the identification of local explanations and the ranking of feature importance.

For the Logistic Regression model, we used default parameters from the "linear_model" package. The Random Forest model employed 100 trees from the "ensemble" package, with other hyperparameters left unchanged; increasing this number could improve performance but would increase model complexity. For example, a randomized search ("RandomizedSearchCV") might identify a model with similar accuracy using 385 trees. The Neural Network is sequential, with three layers consisted of 16, 16, and 1 neuron(s); the first two layers used *Rectified Linear Unit* (ReLU) activation, while the final layer used *sigmoid function*. The optimizer was *Adam*, and the loss function was *binary_crossentropy*. The same test set served as validation.

We split the COMPAS dataset into training and test sets to evaluate model performance and to compute accuracy on the test data, in order to reduce the risk of overfitting. We focused on supervised learning approach for classification, where input features (labelled data) where mapped to discrete outcome categories. We used the test set to assess model capacity in the population (Hold-out procedure). We fitted our data with Random Forest, a non-parametric supervised learning method for non-linear relationships defined as an ensemble model, the result of aggregating a set of Decision Trees; it avoids local optima and correlation, Decision Trees limitations, through bagging and features selection. We also applied Neural networks (thousands of simple nonlinear models that work together, very difficult to interpret) and Logistic regression for comparing classification results, as they are representative in the trade-off between interpretability and accuracy (Wang *et al*., 2023).

With a binary target we analysed the classification report and the confusion matrix, focusing on the most important metrics: *accuracy, recall, precision, F1-score* and *specificity*. The confusion matrix is created by setting an assignment threshold. The decision threshold or cutoff point, generally set at 0.5, is a critical value used to convert the output of a classification model into a class prediction,

since many algorithms return a probability score indicating the likelihood that an input belongs to the positive class, the threshold determines the cutoff point for classification. We obtained feature importance through the SHAP values related to each classifier, a powerful method for explaining the predictions of any machine learning model representing how much each feature contributes to a particular prediction for a given instance, given the expected baseline. We also analysed the classifiers behaviour in subgroups, to verify if accuracy is homogeneous in these subsets. To this end, we applied unsupervised classification to group the N units into clusters, to ensure that units within the same cluster displayed homogeneity.

We employed K-means to partition observations by minimizing within-cluster variance and maximizing separation between groups. To avoid instability associated with random initialization, we adopted the more robust K-means++ method. This approach improves convergence and clustering quality by selecting initial centroids with probabilities proportional to their squared distance from those already chosen, ensuring more diverse and informative starting points.

In order to select the optimal number of clusters K, we computed the Silhouette Score, which evaluates how well each unit fits its assigned cluster (cohesion) compared to the nearest alternative cluster (separation). The score ranges from –1 to 1, with higher values indicating better-defined clusters (1).

$$s(i) = b(i)\text{-}a(i) / \max\{b(i), a(i)\} \quad (1)$$

where a(i) is the average distance of each unit i to all the other units in its assigned cluster, and b(i) is the average distance of unit i to the units in the nearest cluster to which it was not assigned. The optimal K was selected by maximizing the silhouette score. A hierarchical solution with Ward's method (Ward, 1963) confronted at the same K produced similar values, supporting robustness, while K-means++ achieved slightly higher silhouette values, indicating improved stability and separation. By clustering similar instances, we identified subgroups differentiated by accuracy, and we analysed each cluster in terms of label imbalance, evaluation metrics, and SHAP-based feature importance (Lundberg and Lee, 2017). SHAP provides a linear explanation model in which binary variables attribute an effect to each feature; the sum of these contributions approximates the original model's output f(x) (2).

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \quad (2)$$

where z'_i in $\{0,1\}$, M is the number of simplified input features, i ∈ R, f is the original prediction model to be explained and g the explanation model.

We compared subgroup metrics with the overall situation on the entire test set to underline differences and define each subgroup peculiarity: the idea is that the defined subgroup may highlight regions where the problem is more or less accurate.

3. Outcomes

The models show (Table1) accuracies of 0.68 (Logistic Regression), 0.67 (Neural Networks), and 0.63 (Random Forest), with the latter performing worst.

To assess overfitting, performance variability was evaluated via Cross-Validation (CV), which revealed a standard deviation (std) of 0.02 across all models, thus excluding significant overfitting.

Table 1 – *Classification report, main metrics for each class (negative, positive). Logistic Regression, Neural Networks, Random Forest models; COMPAS dataset.*

Models	Precision	Recall	F1-score	Support
Logistic Regression				
negative	0.69	0.78	0.73	1,189
positive	0.68	0.57	0.62	976
Accuracy			**0.68**	2,165
Macro avg	0.68	0.67	0.67	2,165
Weighted avg	0.68	0.68	0.68	2,165
Neural Networks				
negative	0.66	0.82	0.73	1,189
positive	0.69	0.50	0.58	976
Accuracy			**0.67**	2,165
Macro avg	0.68	0.66	0.66	2,165
Weighted avg	0.68	0.68	0.67	2,165
Random Forest				
negative	0.66	0.69	0.67	1,189
positive	0.60	0.56	0.58	976
Accuracy			**0.63**	2,165
Macro avg	0.63	0.63	0.63	2,165
Weighted avg	0.63	0.63	0.63	2,165

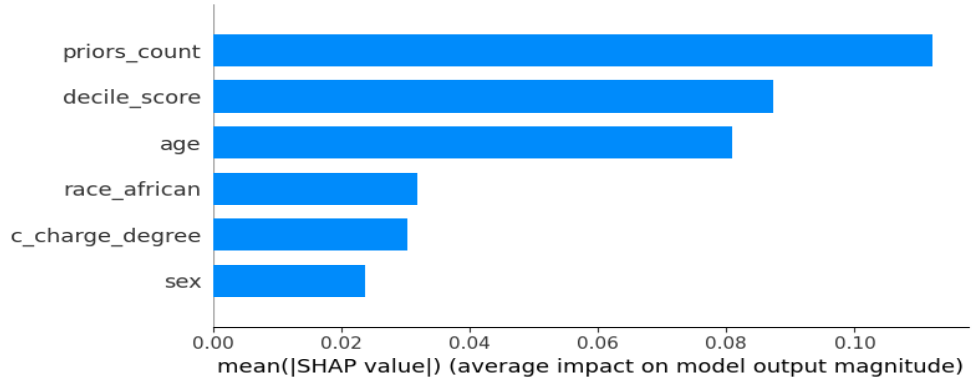
Note: negative (N) corresponds to non-recidivist; positive (P) corresponds to recidivist

All models show higher F1-scores for the negative class, as recall is greater for non-recidivists. Conversely, precision exceeds recall for the positive class, indicating

better control of false positives than false negatives. Given the higher risk of false positives in criminal recidivism, prioritizing precision reduces the chance of mislabelling non-recidivists as recidivists. In the COMPAS tool's trade-off, avoiding missed recidivists is considered more acceptable than penalizing non-recidivists. The *Receiver Operating Characteristic-Area under the Curve (*ROC-AUC) values (0.74 for Logistic Regression and Neural Networks, 0.68 for Random Forest) support these findings; 0.74 is considered excellent for complex classification tasks. Selecting an optimal threshold requires analysing the confusion matrix to balance FP and TP where T and F denote true and false predicted values, respectively.

We obtained feature importance through the SHAP values (Figure 1), in general model agnostic while the actual computation for different model types often uses specific approximations or exact algorithms. The explainer for Random Forest is specifically designed for tree-based models and ensembles of trees (model specific) and can calculate SHAP values exactly or with high accuracy and efficiency, taking into account all possible paths. Due to the nature of tree models, Tree SHAP is particularly effective at capturing and quantifying interaction effects among features.

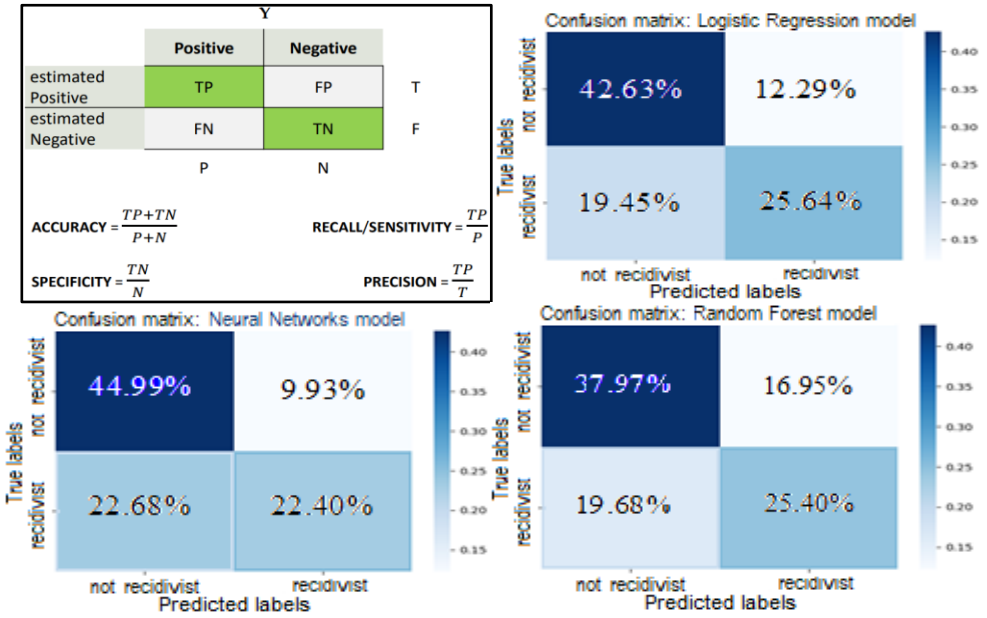
Figure 1 – *Feature importance based on the mean absolute SHAP values of the Random Forest model for the COMPAS dataset.*



The confusion matrix (Figure 2) shows consistent patterns across models, except for Random Forest. It has a higher proportion of FPs (16.95%) and a lower TNs (37.97%), indicating a greater risk of overestimate the recidivism (false positive errors).

In complex scenarios characterized by weak feature-target relationships or nearly indistinguishable classes, an AUC of 0.68 (as achieved by the Random Forest model) may still be acceptable. Further analysis is required: evaluation of feature selection, hyperparameters optimization and classifier behaviour across specific subsets.

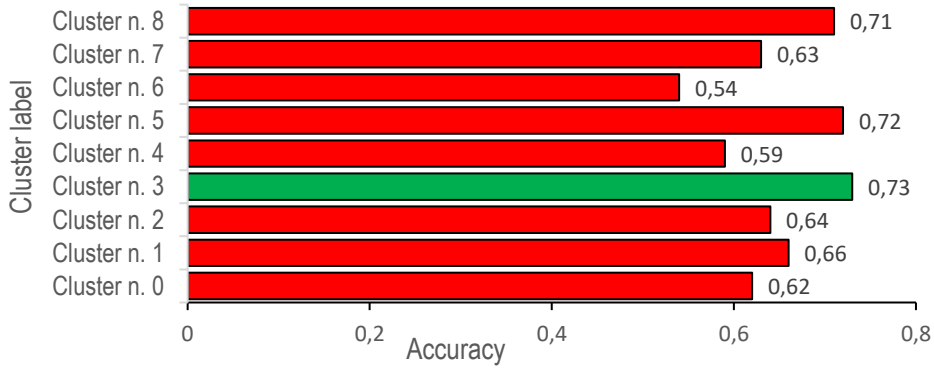
Figure 2 – *Confusion matrix: comparison between actual (rows) and predicted values (columns). Logistic Regression, Neural Network and Random Forest models; COMPAS dataset.*



We grouped similar units, selecting a nine-cluster solution based on the peak Silhouette Score (k = 9, 0.57), a value considered reasonable (Figure 3).

Figure 3 – *Silhouette score and number of clusters for COMPAS dataset.*

Figure 4 – *Accuracy of the Random Forest model calculated for each cluster identified using the K-Means method (k = 9) on the COMPAS dataset.*



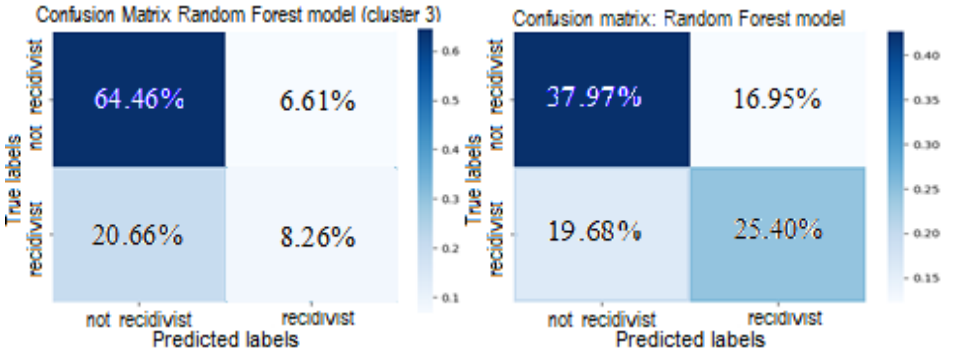
We compared K-means and K-prototypes for clustering numerical and binary features from the COMPAS dataset. The results showed no significant differences in classification accuracy across clusters – an essential focus of this study – nor in overall clustering quality (k = 8, silhouette score = 0.60). We selected K-means, which produced the widest accuracy range, and applied a Random Forest classifier to the resulting clusters (Figure 4). Cluster centroids and sizes were computed. Accuracy ranged from 0.54 (cluster 6) to 0.73 (cluster 3). Comparing the global metrics with those of cluster 3 with the highest accuracy, patterns emerged that helped identify factors potentially enhancing classification performance (Table 2).

Table 2 – *Classification report, main metrics for each class, cluster n.3 (left side) and whole dataset COMPAS (right side) based on the Random Forest model*.

	Random Forest (cluster n.3)					**Random Forest (whole dataset)**				
Random Forest	Precision	Recall	F1-score	Support	Balance	Precision	Recall	F1-score	Support	Balance
negative	0.76	0.91	0.83	86	0.71	0.66	0.69	0.67	1,189	0.55
positive	0.56	0.29	0.38	35	0.29	0.60	0.56	0.58	976	0.45
Accuracy			**0.73**	121				**0.63**	2,165	
Macro avg	0.66	0.60	0.60	121		0.63	0.63	0.63	2,165	
Weighted avg	0.70	0.73	0.70	121		0.63	0.63	0.63	2,165	

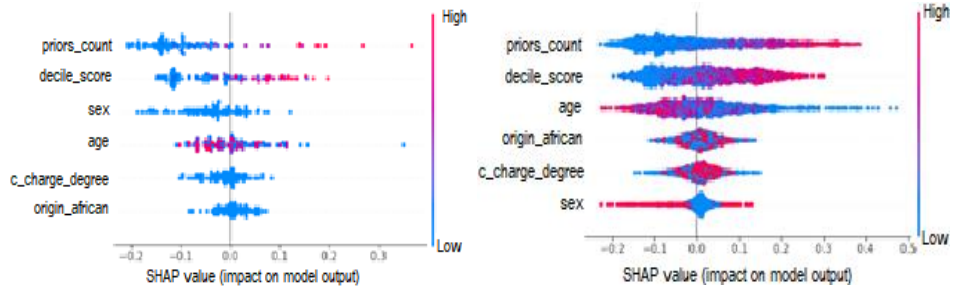
The confusion matrix for cluster 3 versus the whole dataset (Figure 5) indicates improved recall for the negative class but decreased recall for the positive class, reflecting class imbalance (>70% non-recidivists). Consequently, FPs decrease (6.61 vs. 16.95), TNs increase (64.46 vs. 37.97), while TPs decline (8.26 vs. 25.40).

Figure 5 – *Random Forest model's confusion matrix. Comparison between actual (rows) and predicted values (columns). Cluster n.3 (left) and the COMPAS dataset (right).*



SHAP values in the most accurate clusters identified key features influencing model behaviour. In cluster 3, gender – specifically female – was more important than age, while origin was not relevant; this cluster mainly include Caucasian women with minor offenses and low risk scores (Figure 6).

Figure 6 – *Random Forest model's SHAP values for cluster no. 3 (left) and COMPAS dataset (right).*



Cluster 8 (not shown), with similar profiles but of African origin, also demonstrated high accuracy (0.71). Conversely, cluster 6 (not shown): Caucasian men with serious offenses and high-risk scores, had the lowest accuracy (0.54), likely due to recidivism imbalance. Importantly, *ceteris paribus*, individuals of African

origin (cluster 0) outperformed their Caucasian counterparts: accuracy increased from 0.54 to 0.62 (not shown).

4. Conclusions and future perspective

This study examines the predictive potential of ML models in assessing criminal recidivism risk, using the COMPAS dataset as a case study. Three supervised classification models – logistic regression, neural networks, and random forest – were tested and evaluated in terms of accuracy, interpretability, and subgroup variability. Although the overall predictive performance was moderate (maximum accuracy 0.68), logistic regression and neural networks outperformed random forest, particularly with respect to the area under the ROC curve (AUC) and F1-score.

This study shows that clustering reveals substantial accuracy heterogeneity in the COMPAS data: K-means++ identified stable subgroups with distinct performance levels, a structure confirmed by a Ward hierarchical solution and supported by slightly higher silhouette scores. SHAP values clarified why accuracy varies, with high-performing clusters showing consistent feature contributions and low-performing ones affected by heterogeneity and class imbalance. These results show that overall accuracy can hide substantial performance differences across subgroups. Cluster-based evaluation and interpretability methods therefore provide a more reliable and transparent assessment of model behaviour in criminal-risk prediction.

From a legal perspective, the study highlights the divergence between civil and common law systems: while tools like COMPAS are used in the U.S., In the Italian legal system, predictive assessments of individual risk are restricted by procedural norms (Article 220 of the Italian Code of Criminal Procedure), effectively limiting the use of AI tools to the post-sentencing phase. The EU AI Act (Regulation 2024/1689) classifies such tools as high-risk, subjecting them to strict requirements of transparency, reliability, and protection of fundamental rights

Despite their current limitations, ML models could serve a complementary but important role in public policy and correctional planning. Their use should not be confined to making decisions about alternative sanctions or sentencing. Instead, they may be strategically employed to identify individuals at higher risk of reoffending, thereby enabling targeted interventions and the efficient allocation of social support services. We suggest adjusting the classification threshold; in this context, prioritizing a slightly higher false positive rate may be ethically acceptable if it serves to initiate rehabilitative or generally supportive interventions for the individual – rather than punitive measures.

In conclusion, while machine learning holds promise for enhancing our understanding of recidivism patterns and informing data-driven policies, its

application in the criminal justice domain must remain grounded in legal safeguards, methodological transparency, and a strong commitment to human rights. Future research should further explore the integration of richer contextual variables, investigate algorithmic fairness across demographic groups.

Acknowledgements

Deep gratitude is expressed to Prof. Donatella Firmani, for her exceptional guidance, unwavering support, and valuable feedback. We also express our gratitude to the RIEDS reviewers for their thoughtful and valuable feedback.

References

ANDREWS, D.A., BONTA, J. 2010. *The Psychology of Criminal Conduct*. New York: Routledge.

ANGWIN J., LARSON J, MATTU S, KIRCHNER L, 2016. *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica, May 23.

BARATTA A. 1998. *Criminologia critica e critica del diritto penale.* Napoli: Ed. Scientifiche italiane.

BERK, R., BURNETT, R., NELSON, C., SHEARER, C., BERGMAN, M. 2018. Fairness and Transparency in Risk Assessments, *Criminology & Public Policy*, Vol. 17, No. 4, pp. 857-872.

BURGESS E.W. 1928. Factors Determining Success or Failure on Parole. In *Illinois Committee on Indeterminate-Sentence Law and Parole*, Springfield: State Board of Parole, pp. 221–234.

CABRERA Á.A., EPPERSON W., HOHMAN F., KAHNG M., MORGENSTERN J., CHAU D.H. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 46–56.

CITRON D.K., PASQUALE F. 2014. The Scored Society: Due Process for Automated Predictions, *Washington Law Review*, Vol. 89, No. 1, pp. 1–33.

FLORIDI L., COWLS J., BELLI L., TURILLI M., TADDEO M. 2018. AI4People – An Ethical Framework for a Good AI Society, Vol. 28, No. 4, pp. 689–707. https://doi.org/10.1007/s11023-018-9482-5

GLUECK S., GLUECK E. 1950. *Unraveling Juvenile Delinquency*. Cambridge: Harvard University Press.

LAPPI-SEPPÄLÄ T, 2003. Techniques in enhancing community based alternatives to incarceratio. In *Resource Material Series*, No. 61. Tokyo UNAFEI, pp. 61-87.

LUNDBERG S.M., LEE S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In GUYON I. *et al.* (Eds.), *Advances in Neural Information Processing Systems,* No. 30, Curran Associates, Inc., pp. 4765–4774.

PROPUBLICA, 2016. Compas-Dataset https://github.com/propublica/Compas-analysis.

RUDIN C., WANG C., COKER B. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction, *Harvard Data Science Review*, Vol. 2, No. 1 (31 March 2020). https://hdsr.mitpress.mit.edu/pub/7z10o269

SPACE II, 2023. Annual penal statistics: recidivism rates. Council of Europe.

UNODC, 2018. Manual on the measurement of juvenile justice indicators. United Nations Office on Drugs and Crime.

VENKATARAMAN N. 2025. *Quantitative Criminology Handbook*. (e-book) Educohack Press.

WANG C., HAN B., PATEL B., RUDIN C. 2023. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, *Journal of Quantitative Criminology*, Vol. 39, No. 2.

WARD J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, Vol. 58 No. 301, pp. 236-244.

WEATHERBURN D., FITZGERALD J., HUA J. 2003. Reducing Recidivism: Does Imprisonment Have Any Effect? *Crime and Justice Bulletin*, No. 58, NSW Bureau of Crime Statistics and Research, Sydney.

Flavia TAGLIAFIERRO, Sapienza University, flavia.tagliafierro@uniroma.it
Claudio CATERINO, Istat, caterino@istat.it