Data-Driven Insights into Juvenile Recidivism: Leveraging Machine Learning for Rehabilitation Strategies

Saiakhil Chilaka *

Enloe High School

DOI: https://doi.org/10.5281/zenodo.13933479

Published Date: 15-October-2024

***Abstract:* Juvenile recidivism presents a significant challenge to the criminal justice system, impacting both the individuals involved and broader societal safety. This study aims to identify the key factors influencing recidivism and suc-cessful rehabilitation outcomes by utilizing a dataset of over 25,000 individ-uals from the NIJ Recidivism Challenge. We employed machine learning techniques, particularly Random Forest Classification, combined with SHAP (SHapley Additive exPlanations) for model interpretability. Our findings indicate that Supervision Risk Score, Percent Days Employed, and Education Level are critical factors affecting recidivism, with higher levels of supervision, successful employment, and education contributing to lower recidivism rates. Conversely, Gang Affiliationemerged as a significant risk factor for reoffending. The model achieved an accuracy of 68.8%, highlighting its utility in identifying high-risk individuals and informing targeted interventions. These results suggest that a comprehensive approach involving personalized supervision, vocational training, educational support, and anti-gang initiatives can significantly reduce recidivism and enhance rehabilitation outcomes for juveniles, providing critical insights for policymakers and juvenile justice practitioners.**

***Keywords:* Juvenile recidivism, criminal justice system, broader societal safety. juvenile justice practitioners.**

1. INTRODUCTION

Juvenile recidivism is a persistent issue that affects not only the lives of young offenders but also the broader community and society at large. Re-cidivism, defined as the tendency of previously convicted individuals to re-offend, poses challenges for rehabilitation systems across the globe, and par-ticularly in the United States. Juveniles, as a vulnerable population, require specific attention, since the factors influencing their criminal behavior can differ significantly from those affecting adult offenders.[1]Understanding the underlying causes of juvenile recidivism and identifying key factors that contribute to either successful rehabilitation or reoffending are crucial to designing effective interventions.

This research aims to investigate what influences juvenile rehabilitation and lowers recidivism rates by utilizing a comprehensive dataset from the NIJ Recidivism Challenge. Through the application of machine learning techniques, such as Random Forest Classification, and leveraging SHAP (SHapley Additive exPlanations) for model interpretability, we analyze fac-tors such as education, employment, gang affiliation, supervision levels, and other demographic features. This study's findings aim to inform policymak-ers and juvenile justice institutions about key areas for improvement, with the ultimate goal of reducing recidivism and promoting long-term rehabili-tation for juveniles.

2. DATASET

The dataset used in this study comes from the NIJ Recidivism Challenge and consists of over 25,000 records of individuals tracked over a 3-year pe-riod post-release. These individuals were drawn from various correctional facilities, and their outcomes were recorded across multiple dimensions. The dataset provides a rich source of information covering demographic, social, criminal, and institutional factors, enabling a comprehensive analysis of re-cidivism patterns.

2.1 Key Features of the Dataset

• Demographic Data: Gender, Race, Age at Release. This infor-mation helps us understand whether specific demographic groups are more prone to recidivism or more likely to rehabilitate successfully.

• Institutional Data: Supervision Level, Gang Affiliation, and Su-pervision Risk Score. These features help us assess how institutional practices, such as levels of supervision or involvement in gang activi-ties, influence outcomes.

• Rehabilitation Indicators: Education Level, Percent Days Em-ployed, Jobs Per Year, and Dependents. These are considered rehabil-itation factors, providing insights into how well individuals reintegrate into society post-release.

• Criminal History: Prior arrest episodes, including felony, misde-meanor, and drug-related charges, indicate the criminal background of individuals and help to assess their likelihood of reoffending.

• Recidivism Outcomes: The dataset tracks whether individuals were rearrested within 1, 2, or 3 years of their release. This binary outcome (recidivism or not) is used as the target variable for the predictive model.

2.2 Data Cleaning and Pre-Processing

Before applying machine learning models, several steps were taken to clean and preprocess the dataset. Missing values were handled through imputa-tion. For numerical columns, such as ”Supervision Risk Score” and ”Per-cent Days Employed”, missing values were filled with median values to avoid skewing the results. Categorical variables, such as ”Supervision Level” and ”Education Level”, were encoded into numerical values using LabelEncoder, allowing machine learning algorithms to interpret them. Additionally, boolean columns like ”Gang Affiliation” were converted to binary representations (0 or 1), and string-based fields like ”Dependents” (which had values such as ”3 or more”) were also converted into numerical form.

The recidivism outcome, Recidivism_Within_3years, was transformed into a binary variable where 1 represents recidivism and 0 represents no recidivism. The data was split into training (70%) and testing (30%) sets to ensure a fair evaluation of the model’s performance.

3. METHODOLOGY

The primary goal of this study is to identify factors that predict juvenile recidivism using machine learning, specifically the Random Forest algo-rithm. This section outlines the model selection, data preprocessing steps, and evaluation metrics used in the analysis.

3.1 Model Selection

The Random Forest Classifier was chosen for this study because of its ability to handle complex datasets with a mix of categorical and numerical features, as well as its robustness to overfitting due to the ensemble nature of the model.[2] Random Forests work by constructing multiple decision trees during training and outputting the class (recidivism or no recidivism) that is the mode of the classes predicted by individual trees.

In addition to Random Forest, Logistic Regression was tested as a benchmark for interpretability. While logistic regression provides simpler and more interpretable models, its linear nature makes it less capable of capturing complex interactions between features compared to tree-based methods like Random Forest.[3]

3.2 Data Pre-Processing

Prior to training the models, we processed the dataset to ensure compatibil-ity with machine learning algorithms. Key steps in preprocessing included:

• Label Encoding: For categorical variables like ”Supervision Level First” and ”Education Level”, we employed label encoding to convert these text categories into numeric codes.

• Missing Data Handling: Missing values in numerical columns, such as ”Supervision Risk Score First” and ”Percent Days Employed”, were replaced with the median value of the respective column. For categor-ical fields like ”Gang Affiliated”, missing values were replaced with 0, assuming no gang affiliation where data was not available.

• Train-Test Split: The data was split into 70% training data and 30% testing data to validate the model’s performance on unseen data.

3.3 Model Evaluation

We evaluated model performance using the following metrics:

- Accuracy: Measures the percentage of correct predictions made by the model.

- Precision: The proportion of true positive predictions among all pre-dicted positive outcomes (e.g., individuals predicted to recidivate who actually did).

- Recall: The proportion of true positive predictions among all actual positive outcomes (e.g., individuals who recidivated that were correctly predicted).

- F1-Score: The harmonic mean of precision and recall, providing a balanced metric for evaluating the model when class distribution is skewed.

4. RESULTS

The Random Forest Classifier achieved an overall accuracy of 68.8% in predicting juvenile recidivism. This means that approximately 69% of the predictions made by the model were correct. Below is a breakdown of the results for both classes (recidivists and non-recidivists):

- Precision for Non-Recidivists (Class 0): 0.65

- Recall for Non-Recidivists (Class 0): 0.58

- Precision for Recidivists (Class 1): 0.71

- Recall for Recidivists (Class 1): 0.77

- F1-Score: 0.69

The higher recall for recidivists (Class 1) suggests that the model is more effective in identifying individuals who are likely to reoffend. This is critical in the context of juvenile justice, where preventing future offenses is a key objective. However, the lower recall for non-recidivists suggests that there is room for improvement in correctly identifying individuals who will not reoffend.

4.1 SHAP Analysis

To further interpret the model, we used SHAP (SHapley Additive ex-Planations) values to explain the contribution of each feature to the model’s predictions. The SHAP analysis revealed the following key insights:

- Supervision Risk Score: The most influential factor, with higher risk scores strongly correlating with higher recidivism rates.

- Percent Days Employed: Employment plays a critical role in reha-bilitation, with individuals who were employed for a higher percentage of days being less likely to reoffend.

- Supervision Level: Individuals who were placed under more inten-sive supervision (e.g., specialized supervision) had better rehabilitation outcomes compared to those under standard supervision.

- Gang Affiliation: Gang-affiliated individuals had a significantly higher probability of recidivism, highlighting the need for targeted interven-tions in this area.

- Education Level: Education emerged as an important predictor of successful rehabilitation, with individuals who had attained higher lev-els of education being less likely to recidivate.

5. DISCUSSION

The results of this study offer important insights into the factors that con-tribute to juvenile recidivism. One of the key findings is the impact of Supervision Risk Score on recidivism. Juveniles who were assessed as high-risk were more likely to reoffend, suggesting that accurate risk assessments are essential for tailoring interventions. Moreover, the Percent Days Em-ployed feature highlights the importance of economic stability in rehabilita-tion. Juveniles who were able to maintain employment after release were less likely to fall back into criminal behavior, underlining the need for vocational training and job placement programs in juvenile rehabilitation facilities.

Supervision Level also plays a crucial role. Juveniles under more inten-sive supervision programs, such as those that provide specialized support, experienced lower recidivism rates compared to those under standard super-vision. This finding suggests that one-size-fits-all approaches to supervision may not be effective, and that higher-risk juveniles benefit from more indi-vidualized and intensive programs.

Another important observation is the impact of Gang Affiliation on re-cidivism. Gang-affiliated juveniles were much more likely to reoffend, em-phasizing the need for interventions specifically designed to disengage youths from gang activities and provide alternative support systems. Educational attainment was also found to be a protective factor, reinforcing the impor-tance of academic and vocational education in reducing recidivism.

6. CONCLUSION

This study demonstrates that juvenile recidivism is influenced by a range of factors, including supervision levels, employment stability, gang affilia-tion, and education. The findings underscore the need for a multi-faceted approach to juvenile rehabilitation that addresses not only the risk of re-offending but also the underlying socio-economic and environmental condi-tions that contribute to it. The Supervision Risk Score emerged as a critical predictor, highlighting the importance of thorough and accurate risk assess-ments in shaping the rehabilitation strategies for each individual.

Employment stability, measured through Percent Days Employed, was shown to be a powerful rehabilitative factor, suggesting that programs that focus on job training and placement may have a profound impact on reduc-ing recidivism rates. Moreover, Supervision Level and Gang Affiliation were also shown to be significant determinants of recidivism, further indicating that individualized, intensive supervision programs, and targeted anti-gang interventions are crucial for improving outcomes in the juvenile justice sys-tem.

Finally, the role of Education cannot be understated, as higher levels of educational attainment were linked to lower rates of recidivism. This sug-gests that providing educational opportunities during and after incarceration is essential for setting juveniles on a path towards successful reintegration into society.

In conclusion, the analysis shows that there is no single solution to re-ducing juvenile recidivism. Instead, a combination of personalized risk as-sessments, targeted interventions (such as job training and gang preven-tion), and support systems focusing on education and employment are key to improving rehabilitation outcomes. Future research should continue to explore these factors, particularly by integrating mental health and familial dynamics into predictive models. As machine learning techniques become increasingly sophisticated, they hold the potential to offer even more precise and actionable insights into the juvenile justice system, ultimately guiding policymakers and practitioners in developing more effective programs for reducing recidivism and fostering long-term rehabilitation.

REFERENCES

[1] Katherine S L Lau Anthony Perkins Patrick Monahan Thomas Grisso Matthew C Aalsma 1, Laura M White. Behavioral health care needs, detention-based care, and criminal recidivism at community reentry from juvenile detention: A multisite survival curve analysis. American journal of public health, 10(2):1372, 2022.

[2] Leo Breiman. Random forests. Machine Learning, 2001.

[3] Ryan G. McClarren. Decision trees and random forests for regression and classification. Machine Learning, 2021.