1. Correctional Education and Recidivism: Toward a Tool for Reduction​

Findings:

Correctional education significantly reduces recidivism. The study reviewed 10 empirical studies from 1995 to 2010 and found a negative correlation between education and recidivism.

Major recidivism risk factors include age, race, gender, marital status, socioeconomic status, and employment.

Education improves post-release employment opportunities, which is linked to lower recidivism rates.

Financial benefits: Every $962 spent on education results in a $5,306 savings per inmate in reduced criminal justice costs.

Challenges: Political and societal resistance due to the "principle of least eligibility," where the public resents inmates receiving educational benefits.

Relevance to Your Project:

Your project focuses on Evidence-Based Recidivism Reduction (EBRR) programs, and education is one of the most effective interventions. You can incorporate educational programming data into your machine learning models to classify programs and predict effectiveness. The paper also highlights the need for funding and validation, which aligns with your project's goal of providing an automated validation platform for recidivism reduction programs.

2. Does Incarceration-Based Drug Treatment Reduce Recidivism? (Meta-Analysis Study)​

Findings:

Therapeutic Communities (TCs) and Residential Substance Abuse Treatment (RSAT) were the most effective incarceration-based treatments.

Group counseling programs also showed a moderate reduction in recidivism.

Boot camps for drug offenders were ineffective.

Narcotic maintenance programs had mixed results, depending on post-release treatment engagement.

Effectiveness varies based on methodology, sample size, and program features.

Relevance to Your Project:

Your project involves evaluating rehabilitation programs, particularly substance abuse treatment. The findings can help rank programs based on effectiveness in your validation platform. Therapeutic Communities (TCs) and RSAT programs should be prioritized for inclusion as high-impact interventions in your model. The failure of boot camps suggests that policy recommendations should discourage their funding.

3. Drug Court Meta-Analysis: Do They Reduce Recidivism?​

Findings:

Drug courts reduce recidivism by approximately 14% compared to traditional court systems.

Effectiveness is influenced by:

Participant age (younger participants benefit more).

Program length (longer programs have better outcomes).

Follow-up period (longer follow-ups show stronger long-term effects).

Randomized studies showed stronger positive effects than observational studies.

Programs integrating housing, employment, and mental health services showed the best results.

Relevance to Your Project:

Your project involves evaluating recidivism reduction programs. Drug treatment courts (DTCs) are validated interventions, so they should be included in your database of effective programs. Since your AI model will classify programs based on effectiveness, these findings can help fine-tune predictive analytics and success criteria. The study also highlights the importance of integrated services, which aligns with your goal of identifying multi-faceted approaches to reducing recidivism.

Data set: NIJ's Recidivism Challenge Full Dataset

|  |  |
| --- | --- |
| Attribute | Details |
| Dataset Name | NIJ's Recidivism Challenge Full Dataset |
| Dataset Owner | Office of Justice Programs |
| Dataset Type | Open source |
| Dataset Size | ~7.3MB(54 columns,25.8k rows) |
| Dataset License | Public |
| Dataset Location | Internet |
| Dataset Access | https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk/about\_data |
| Dataset Restrictions | N/A |
| Dataset Time Range | Jan 1st 2013 to Dec 31st 2015 |
| Dataset Collection Process | Not specified |
| Analytic/Algorithm that will use dataset |  |

**Dataset Overview**

* **Total Records**: 25,360
* **Total Columns**: 54
* **Time Frame**: 2013-2015

Missing Values:

* There are missing values in 11 columns
* Handling them with imputation / dropping

Consistency:

1. High Consistency Areas:
   * Gender, Race, Age at release, Education level, and Recidivism data show 100% consistency, indicating highly reliable information in these categories.
   * Employment percentage (98.21%) and Supervision level (93.34%) also demonstrate very high consistency.
2. Moderate Consistency:
   * Prison offense data shows 87.32% consistency, which is good but suggests some discrepancies exist.
   * Drug test result consistency is at 79.98%, indicating a notable level of inconsistency that may require further investigation.
3. Low Consistency:
   * Prior arrest episodes data shows the lowest consistency at 42.63%, suggesting significant discrepancies or missing information in this category.
4. Implications:
   * The high consistency in demographic data (gender, race, age) and recidivism outcomes provides a solid foundation for analysis.
   * Employment and supervision level data appear reliable for use in predictive modeling.
   * The moderate consistency in prison offense and drug test results suggests caution when using these variables and may require data cleaning or imputation.
   * The low consistency in prior arrest episodes data indicates this variable may not be reliable for analysis without significant data quality improvements.
5. Recommendations:
   * Investigate the reasons for low consistency in prior arrest episodes data.
   * Review and potentially clean or impute data for prison offenses and drug test results.
   * Leverage the highly consistent variables as primary factors in any predictive modeling or analysis.
   * Consider excluding or giving less weight to the prior arrest episodes data in current analyses until its quality can be improved.

Duplicates : There are no duplicates

Conformity:

1. High Conformity Areas:
   * Gender, Race, Age at release, and Education level all show 100% conformity. This indicates that these variables are consistently formatted and adhere to expected standards across the dataset.
2. Moderate to High Conformity:
   * Supervision risk score has a 98.16% conformity rate. While this is still very high, it suggests that there might be a small number of records (about 1.84%) where the supervision risk score doesn't conform to the expected format or range.
3. Lower Conformity:
   * Prison offense shows 87.32% conformity. This indicates that about 12.68% of the prison offense records may have inconsistencies in formatting, categorization, or data entry.

Implications:

1. The demographic data (gender, race, age) and education level information appear to be highly reliable and consistent, providing a solid foundation for analysis.
2. The supervision risk score data is generally reliable, but may require a small amount of cleaning or validation for the non-conforming entries.
3. Prison offense data shows the lowest conformity among the checked variables. This suggests that:
   * There may be inconsistencies in how offenses are categorized or recorded.
   * Some data cleaning or standardization might be necessary for this variable.
   * Analysts should exercise caution when using this variable and consider potential impacts of the non-conforming data on their analyses.

Recommendations:

1. Investigate the non-conforming entries in the supervision risk score and prison offense variables to understand the nature of the inconsistencies.
2. Develop a data cleaning strategy to address the non-conforming entries, especially for the prison offense variable.
3. Consider creating standardized categories for prison offenses if the non-conformity is due to varied categorization methods.
4. Document any changes made to the data during the cleaning process for transparency and reproducibility.

Accuracy:

1. High Accuracy Areas:
   * Recidivism consistency: 100% accuracy, indicating highly reliable recidivism data.
   * Employment percentage accuracy: 98.21%, suggesting very reliable employment data.
2. Moderate Accuracy:
   * Prior arrests >= prior convictions: 76.58% accuracy, indicating some inconsistencies in arrest and conviction records.
   * Drug test results (THC, cocaine, meth, other): All at 79.98% accuracy, suggesting a consistent level of reliability across different drug tests, but with room for improvement.
3. Low Accuracy:
   * Age and recidivism logical consistency: Only 6.04% accuracy, indicating significant issues with the logical relationship between age and recidivism data.

Implications and Recommendations:

1. Recidivism and employment data appear highly reliable and can be confidently used in analyses and modeling.
2. Drug test results show moderate reliability. Consider investigating the ~20% of cases where accuracy is questionable. This might involve:
   * Checking for data entry errors
   * Verifying testing procedures and reporting methods
   * Potentially flagging or excluding unreliable entries from analysis
3. The relationship between prior arrests and convictions needs attention. The 23.42% of cases where this logical relationship doesn't hold should be investigated for:
   * Data entry errors
   * Misclassification of arrests or convictions
   * Potential jurisdictional differences in recording practices
4. The extremely low accuracy in age and recidivism logical consistency (6.04%) is a critical issue that requires immediate attention:
   * This could indicate severe data quality problems in age recording or recidivism tracking
   * A thorough review of how age and recidivism data are collected and recorded is necessary
   * Consider temporarily excluding age from recidivism analyses until this issue is resolved
5. Next steps:
   * Prioritize investigating and correcting the age and recidivism logical inconsistencies
   * Review and clean drug test result data
   * Verify and correct discrepancies in prior arrest and conviction records
   * Document all data cleaning processes for transparency
   * Consider developing more robust data entry and validation procedures to prevent future inconsistencies

Take Awat from EDA:

1. The dataset shows significant demographic disparities in gender, race, age, and recidivism rates.
2. Supervision risk scores are reliable indicators of offender risk levels and can guide intervention strategies.
3. Younger offenders and those involved in property crimes represent key target groups for rehabilitation efforts.
4. Employment support programs are essential given the low average employment rates among offenders.
5. Substance abuse interventions should focus on THC-related cases due to its higher prevalence.

Some plots from eda

A graph of age distribution

AI-generated content may be incorrect.A graph of different colored rectangles

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.

A graph with blue lines

AI-generated content may be incorrect.A graph of blue rectangular objects

AI-generated content may be incorrect.

AI/ML for future:

Supervised Learning:

**1. Logistic Regression (LR)**

* **Why Use**: Logistic regression is simple, interpretable, and effective for binary classification tasks like predicting whether an offender will reoffend within three years.
* **Use Case**: Predict recidivism (recidivism\_within\_3years) based on features like supervision risk score, employment percentage, drug test results, and age at release.
* **Benefits**:
  + Provides probabilities for predictions.
  + Easy to interpret feature importance (e.g., which factors most influence recidivism).

**2. Random Forest (RF)**

* **Why Use**: Random forests are robust and handle non-linear relationships well. They also provide feature importance rankings.
* **Use Case**: Predict recidivism or classify offenders into supervision levels (e.g., standard, specialized, high).
* **Benefits**:
  + Handles missing data effectively.
  + Reduces overfitting by averaging multiple decision trees.
  + Identifies the most important predictors (e.g., supervision risk score or employment percentage).

**3. Gradient Boosting Models (e.g., XGBoost, LightGBM)**

* **Why Use**: Gradient boosting models are highly accurate for tabular data and excel in handling complex relationships between features.
* **Use Case**: Predict recidivism with high accuracy while accounting for interactions between variables like drug test results and employment rates.
* **Benefits**:
  + Superior performance compared to other models on structured datasets.
  + Tunable hyperparameters for optimized performance.

**4. Support Vector Machines (SVM)**

* **Why Use**: SVMs are effective for classification tasks with complex boundaries between classes.
* **Use Case**: Classify offenders as high-risk or low-risk based on supervision risk scores and other features.
* **Benefits**:
  + Works well with smaller datasets.
  + Can handle non-linear decision boundaries using kernels.

Unsupervised Learning:

**1. Clustering Algorithms (e.g., K-Means, DBSCAN)**

* **Why Use**: Clustering helps group similar offenders based on characteristics like age at release, offense type, and drug test results.
* **Use Case**:
  + Identify offender profiles (e.g., low-risk vs. high-risk groups).
  + Segment offenders into groups for tailored rehabilitation programs.
* **Benefits**:
  + Uncovers hidden patterns in the data.
  + Useful for exploratory analysis.

**2. Principal Component Analysis (PCA)**

* **Why Use**: PCA reduces dimensionality while retaining essential information, making it easier to visualize and analyze high-dimensional data.
* **Use Case**:
  + Simplify the dataset by reducing correlated features like drug test results into fewer components for easier interpretation.
  + Prepare data for clustering or supervised models by removing noise.
* **Benefits**:
  + Improves computational efficiency.
  + Helps visualize relationships between offenders.