Recidivism prediction

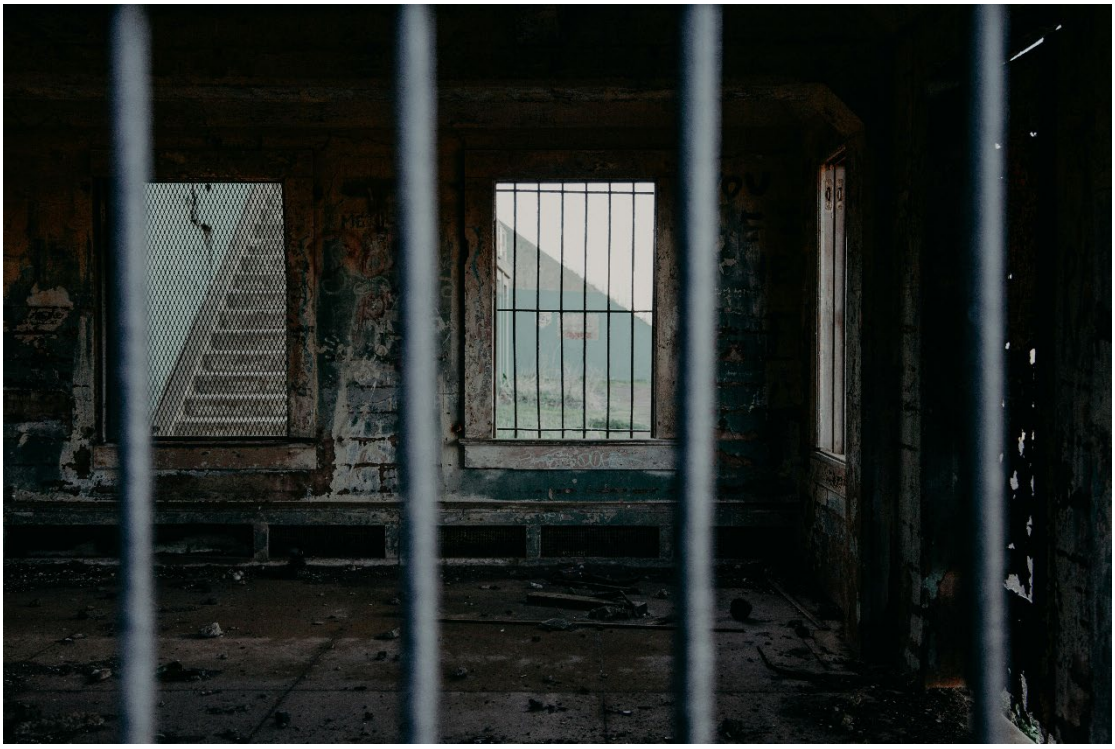
&

By Yael Weisman

Introduction

Recidivism, the tendency of previously convicted individuals to reoffend, is a major challenge faced by criminal justice systems worldwide.
It not only strains correctional facilities and public resources but also reflects deeper social and economic inequalities.
Understanding the factors that contribute to recidivism—such as prior criminal history, socio-economic background, and demographic conditions—can help improve rehabilitation programs and policy decisions.
In recent years, **Machine Learning** has emerged as a powerful tool to analyze these complex patterns and predict the likelihood of reoffending.
Such predictive models can support fairer, data-driven decision-making while reducing bias and improving public safety.

By integrating statistical analysis with ethical considerations, this project aims to identify meaningful insights that contribute to reducing recidivism rates.
Ultimately, the goal is to use data to promote a more effective and just criminal justice system.



Data Description

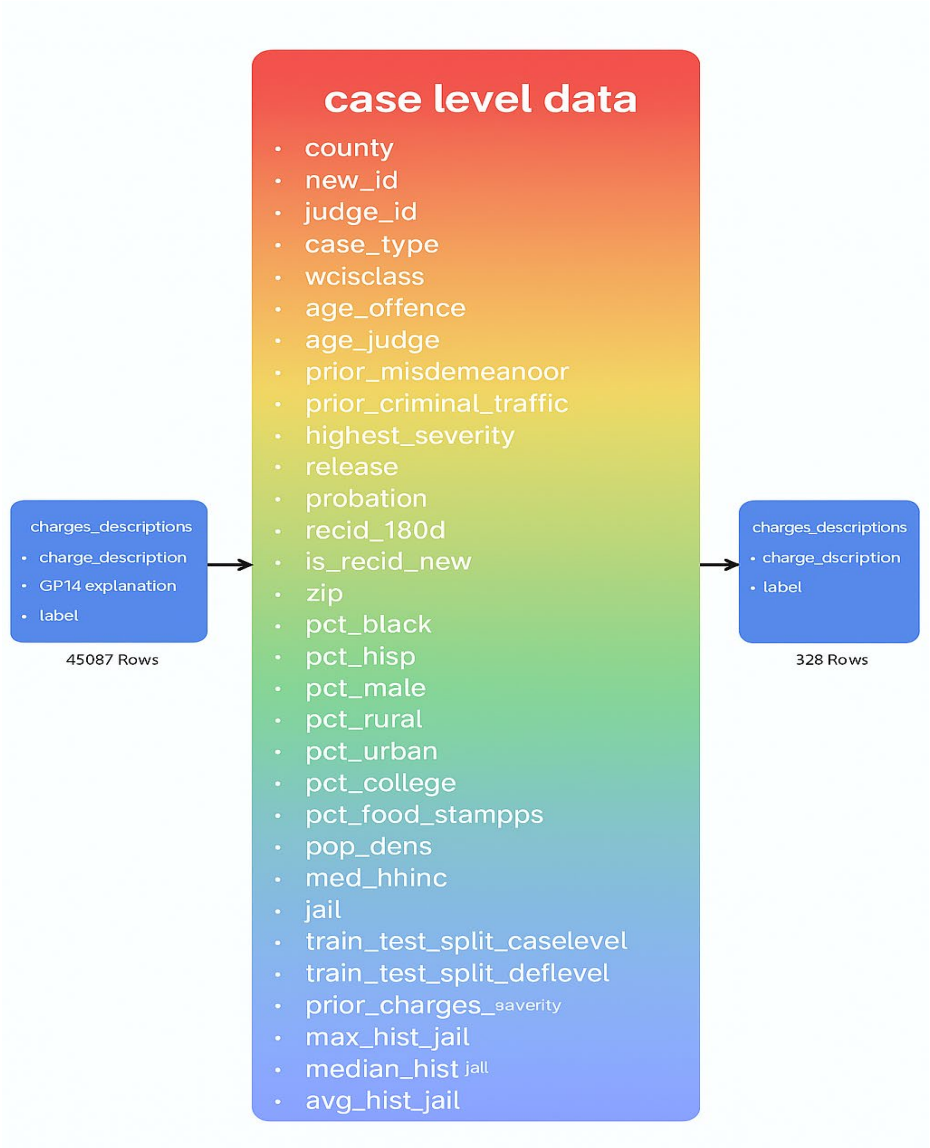
Source link-
https://arxiv.org/abs/2310.18724?utm_source=chatgpt.com

Data link-

https://drive.google.com/drive/folders/17S1VPYTyn6sKWMglWN2sb3CAfZveLGqO

The dataset used in this study contains approximately **1.43 million records** and **54 variables**, providing a comprehensive overview of criminal cases from **Wisconsin Circuit Courts**.
It integrates multiple dimensions of offender and case information, enabling both statistical and machine learning analysis.

Main feature categories:



Criminal history – number and types of prior offenses, conviction severity, sentencing details, and recidivism indicators.

Demographic attributes – age, gender, and race.

Socio-economic variables – neighborhood income indicators , and neighborhood-level conditions.

Social and behavioral factors –neighborhood education level, and community-related aspects.

Target variables (Recidivism definitions):

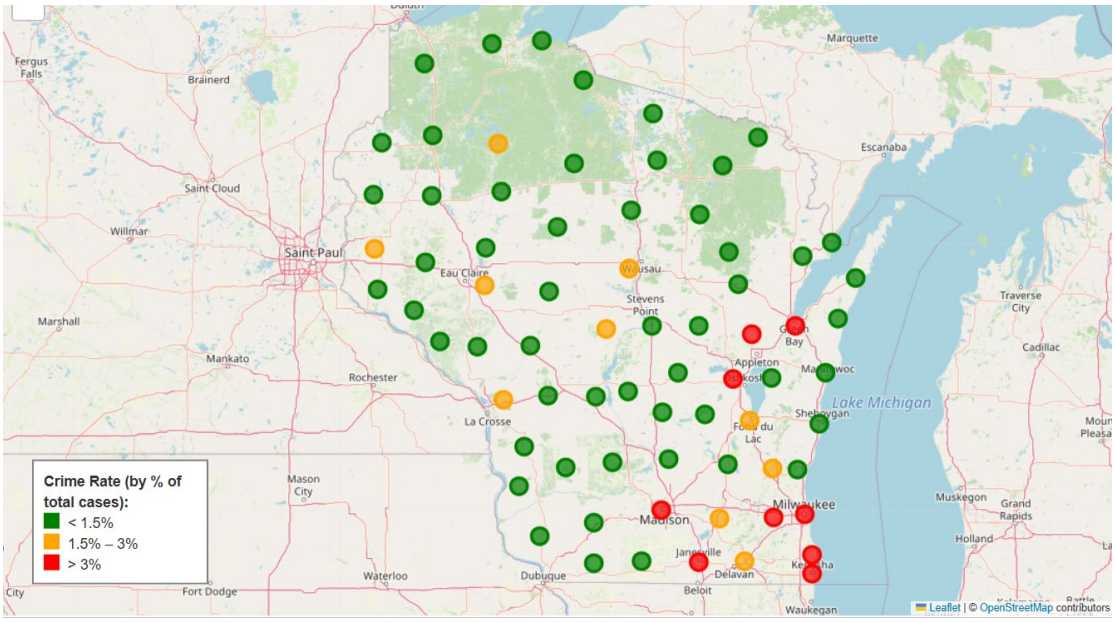
Recidivism within two years – whether the individual reoffended within two years after completing the sentence.

Recidivism after incarceration – whether the individual was re-incarcerated for up to two years after release, following an incarceration period of less than 180 days

Violent recidivism – whether the individual committed a violent offense after release.

The **selected target variable** for this study focuses on **predicting whether an individual will reoffend within two years after completing the sentence**.
This formulation captures the general likelihood of reoffending and

serves as a balanced and interpretable measure of recidivism for predictive modeling.



Factors Associated with Recidivism

The analysis shows that the likelihood of reoffending is influenced by several key factors:
younger age, extensive criminal history, offense severity, and low socio-economic status.
Additionally, low education levels, unemployment, and disadvantaged environments increase the risk.
The combination of personal, demographic, and socio-economic factors allows the model to identify a typical profile of individuals with a higher likelihood of recidivis.

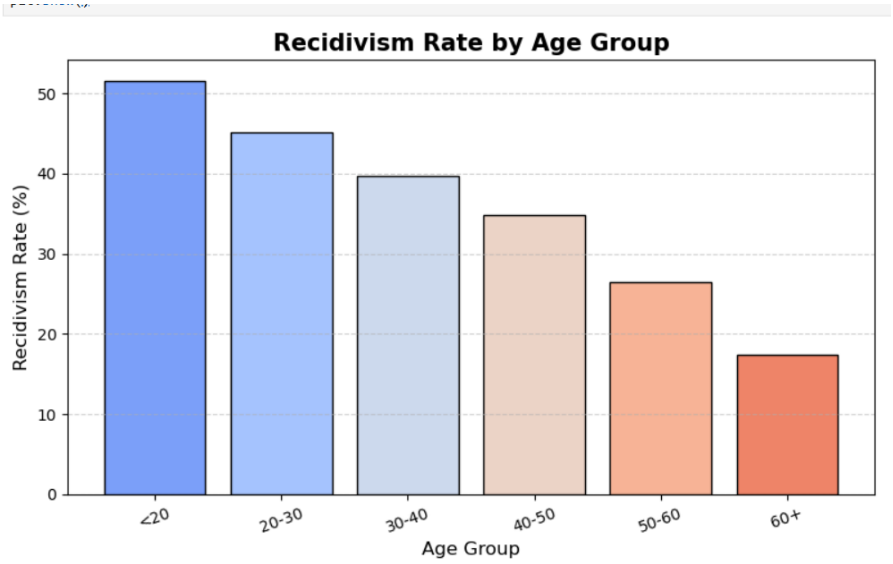
Reducing categories

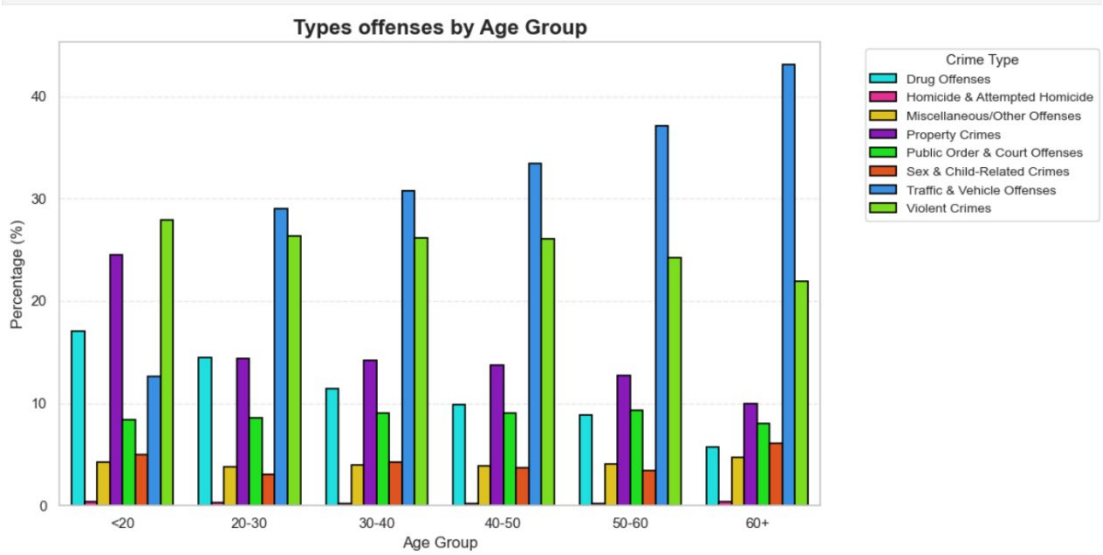
A significant reduction of minor categories was performed, consolidating them into a few main categories to minimize bias and centralize the data. This step ensures more reliable and comparable information across groups. Additionally, the consolidation simplifies statistical analysis and enhances model stability.

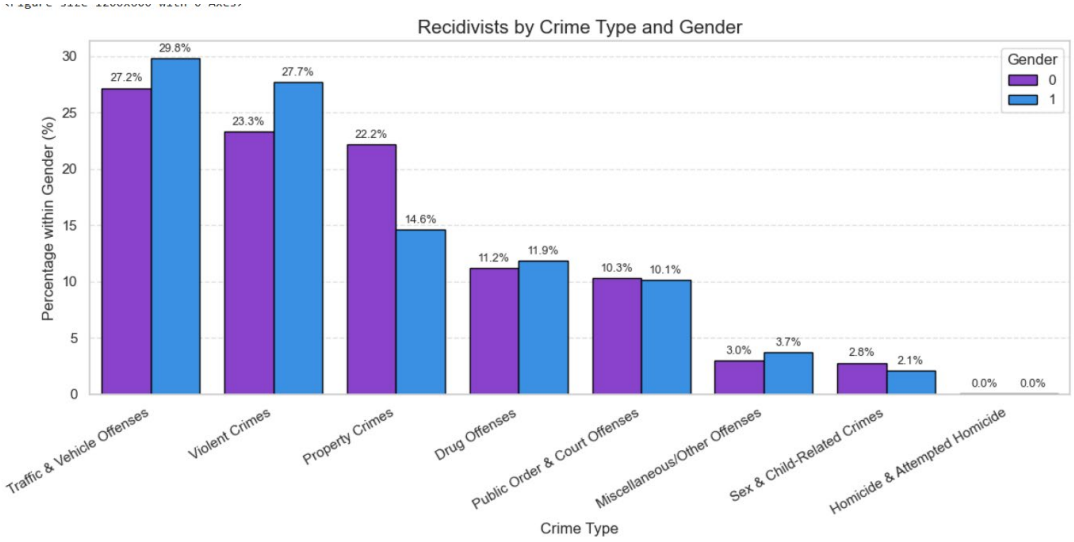
EDA

*Checked the popularity of recisivism by groups age.

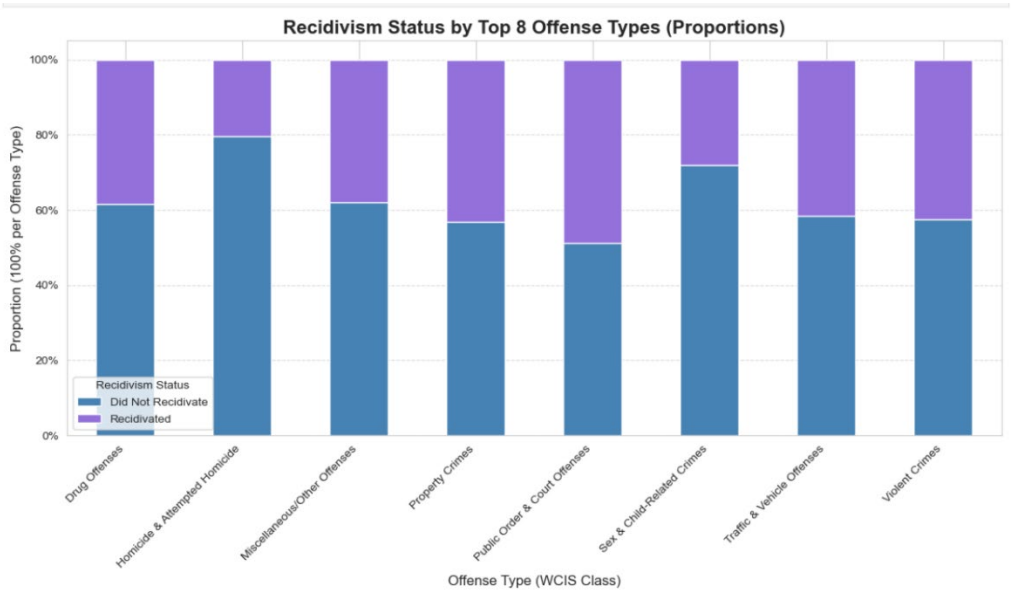
*What is the popularity of every offense by age group and recidivism?

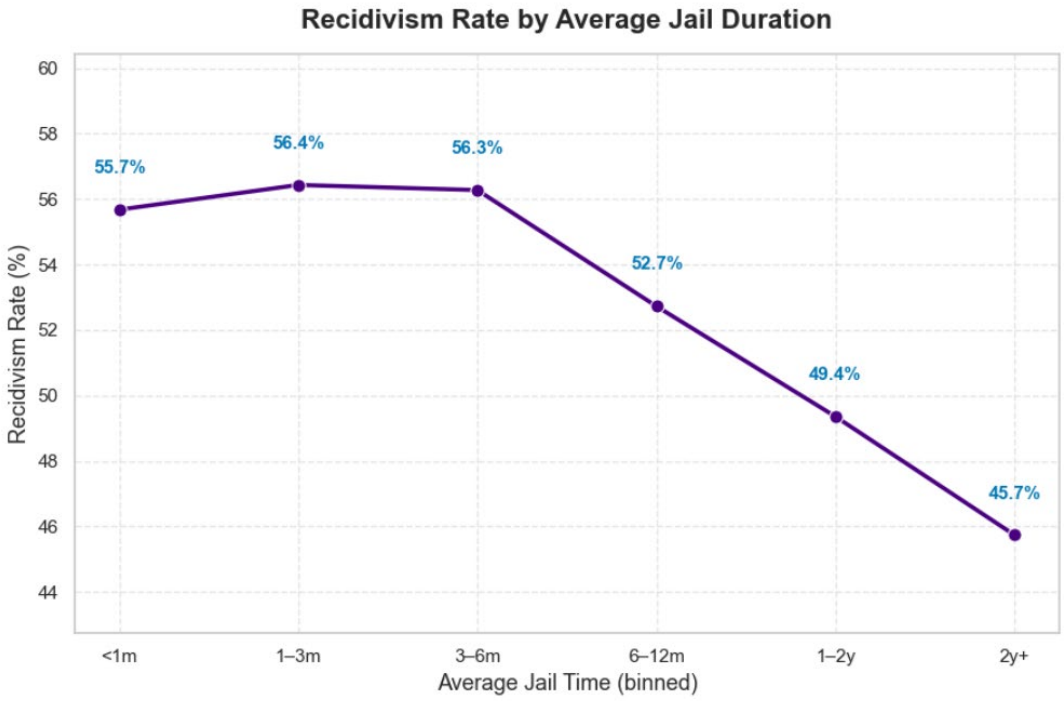






*which crime is the most popular in recidivism?





Outliers

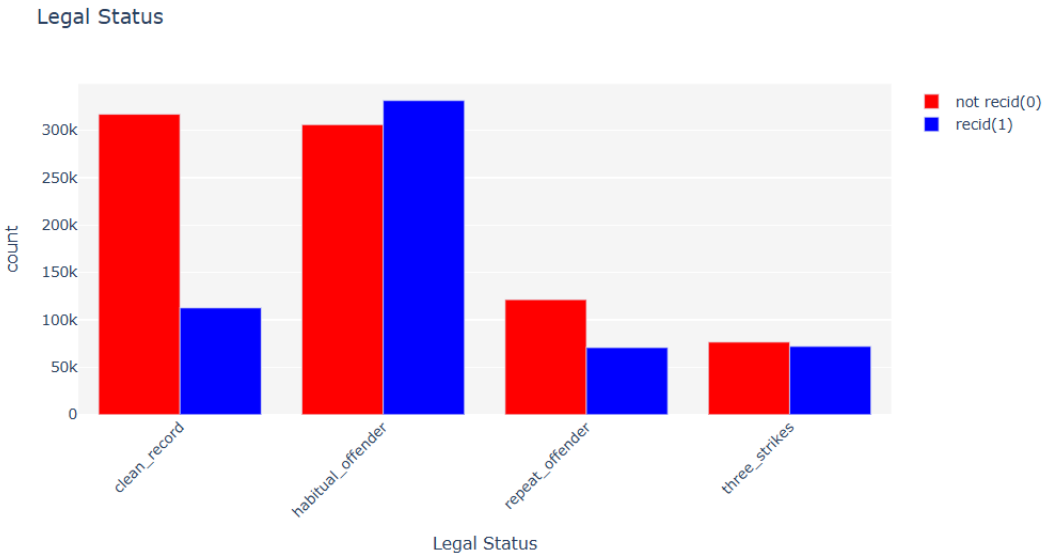
The dataset was highly skewed, with many columns dominated by a single extreme value. These values were flagged as outliers; however, since their dominance reflected the true data distribution, they were handled by grouping into **bins** rather than removal. The remaining outliers were treated using the **IQR (Interquartile Range)** method to ensure balanced and representative feature scaling.

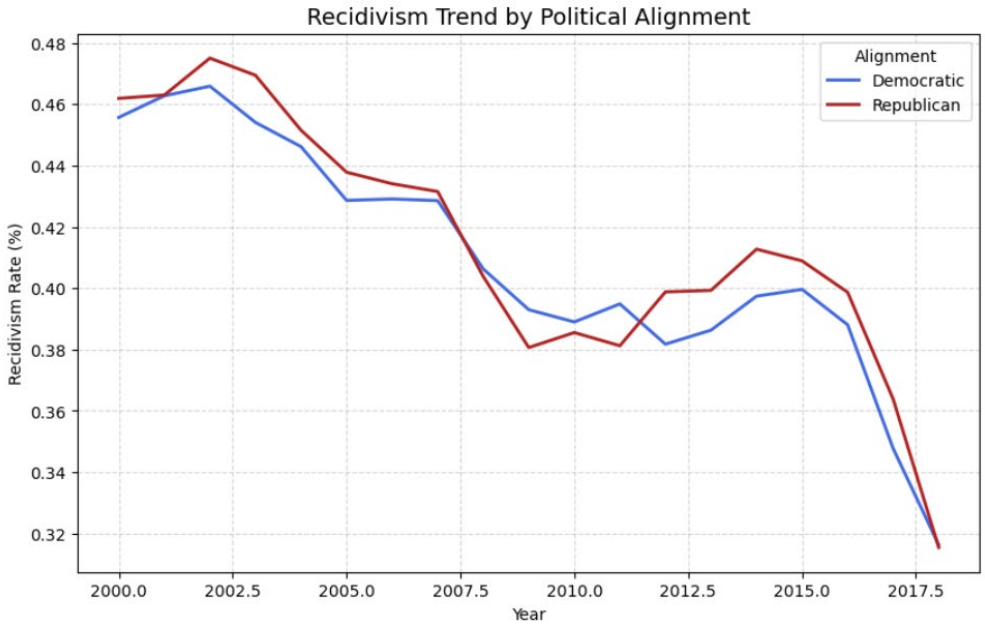
Missing values

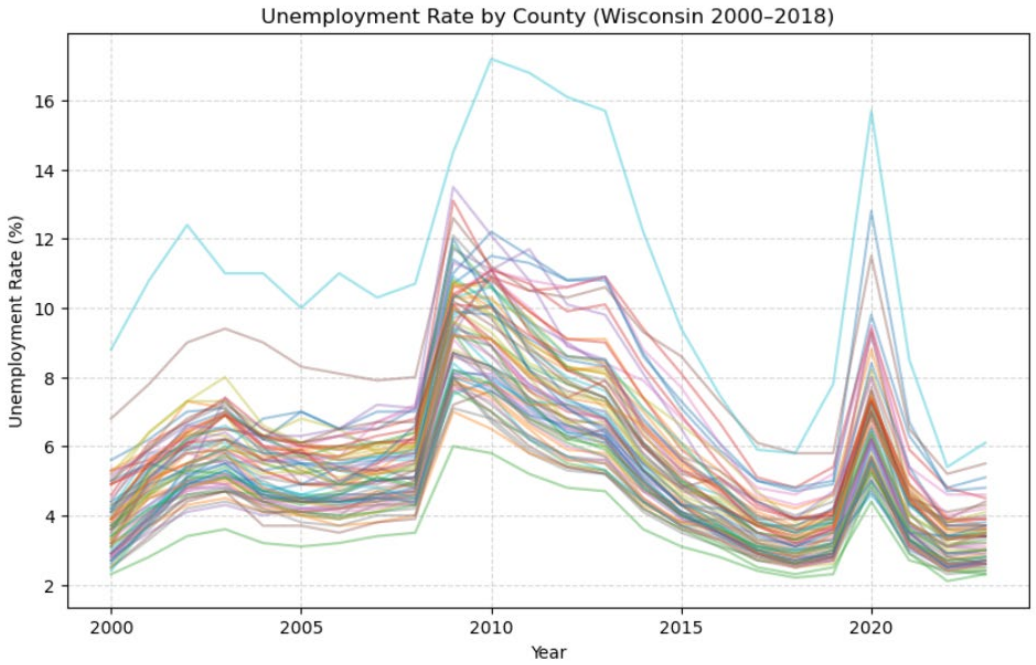
Missing values were imputed using the MICE (Multiple Imputation by Chained Equations) method. This approach models each feature with missing data as a function of other variables, ensuring statistically consistent imputations. It effectively handles both categorical and numerical features, preserving relationships within the dataset.

Feature engineering

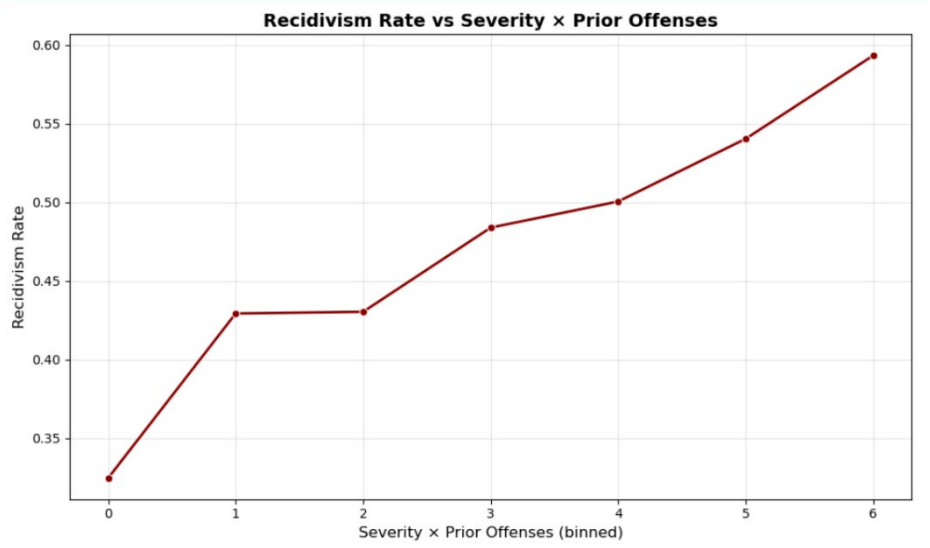
During the feature engineering stage, internal relationships within the dataset were emphasized through the construction of interaction and composite variables. Additionally, scaling was applied to features with large value disparities to ensure uniformity across variables. These transformations captured latent dependencies and improved the model's predictive accuracy and stability.Adding information including year and county of employment rate and unemployment rate, status of offense in law according to his past ,political alignment and precent crimes for every and county/rac.







The next plot introduce recidivism by count prior offenses* highest severity

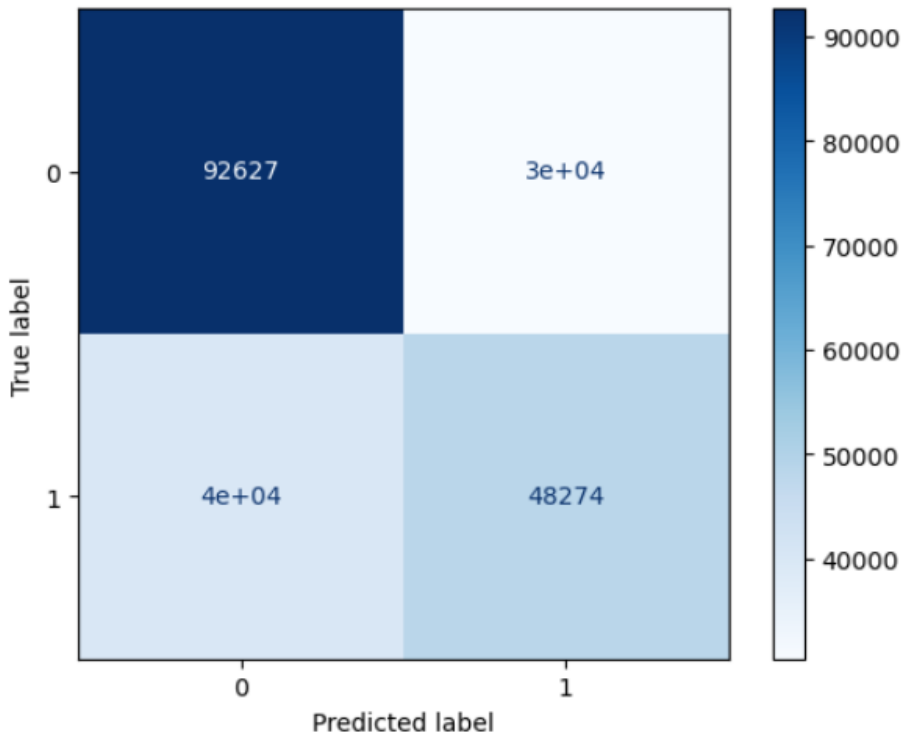


Models:

Using SVC, XGboots ,RandomForest,CatBoots, SGDClassifier models .

The chosen base model is RandomForest

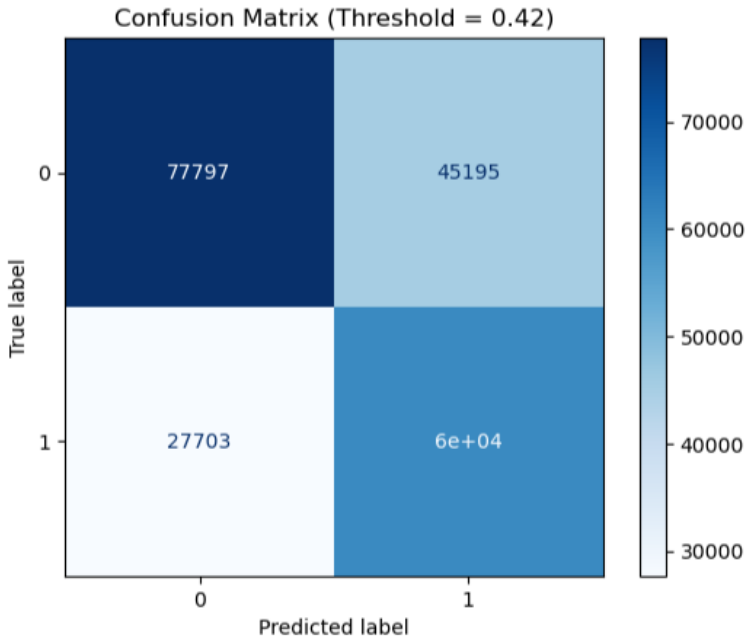
	model	Accuracy	Precision	Recall	f1-score	Log-loss	AUC
0	catboots	0.67	0.63	0.51	0.56	0.60	0.72
1	RandomForest	0.67	0.61	0.55	0.58	0.65	0.71



- Run RandomizedSearchCV to find hyperparameters
- Run model with best parameters on X_Dev
- The results indicate that the model is overly cautious in its predictions, suggesting the need to adjust the **decision threshold** based on the **F1 score**. This optimization balances **precision** and **recall**, leading to a more effective classification of recidivism outcomes.
- Run RandomForst with/out on test
- Choose the model with hyperparameters:

Metric	Score
Accuracy	0.655
Balanced Accuracy	0.659
AUC	0.719

	precision	recall	f1-score	support
0	0.737	0.633	0.681	122992.000
1	0.572	0.686	0.624	88110.000
accuracy	0.655	0.655	0.655	0.655
macro avg	0.655	0.659	0.652	211102.000
weighted avg	0.668	0.655	0.657	211102.000



Conclusion

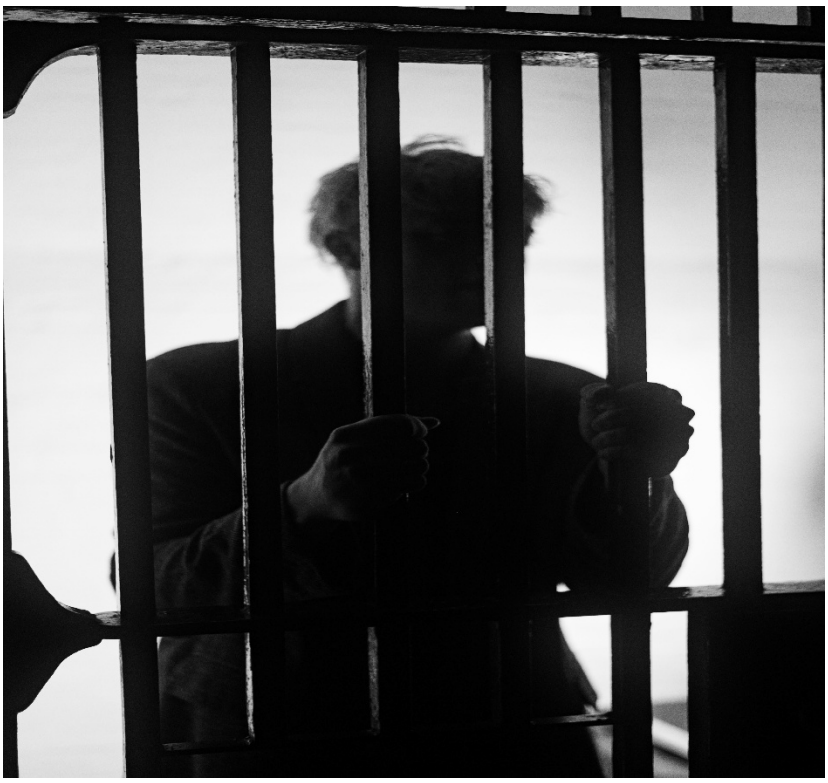
The project presents an advanced application of Machine Learning for **recidivism prediction**, integrating comprehensive data processing that included category reduction, outlier handling, and missing value imputation using **MICE**.

Through **feature engineering** and **scaling**, meaningful social, economic, and legal relationships were captured.

The **Random Forest** model achieved strong and reliable performance, demonstrating the potential for accurate and fair prediction.

Incorporating **psychological, educational, familial, and social-context data** in future work is expected to further enhance accuracy and fairness.

Overall, the model represents a significant step toward responsible, data-driven improvement of rehabilitation and justice systems.



Thank you !

Yael Weisman