

---

# PREDICTING RECIDIVISM IN THE CRIMINAL JUSTICE SYSTEM

---

**Jackson Le**  
wnj9tx

**Austin Shi**  
szb8fc

**Emily Chen**  
qeq3cf

December 13, 2024

## 1 Abstract

Recidivism, the tendency of convicted individuals to reoffend, presents significant challenges within criminal justice systems. In this project, we aim to predict the likelihood of recidivism within one, two, or three years following parole, as well as examine which features contribute the most to this likelihood. Using a dataset from the Virginia criminal justice system, which includes demographic and criminal history details, we apply machine learning classification techniques to track trends and provide predictive insights. By identifying the key factors that contribute to reoffending, this analysis seeks to inform more effective decision-making processes in the criminal justice system, ultimately supporting rehabilitative efforts and reducing recidivism.

## 2 Introduction and Motivation

Recidivism rates serve as indicators of the effectiveness of the rehabilitation and reintegration efforts of the Virginia criminal justice system. High recidivism rates often suggest gaps in support systems and may highlight the need for policy adjustments. Through this project, we will explore how accurately we can predict recidivism using demographic and criminal history data, such as gender, race, type of offense, and drug test results. Through this analysis, we hope to reveal critical insights about the factors that most strongly correlate with the risk of recidivism. By identifying patterns in these data, the project seeks to support improvements in parole management and resource allocation, enabling the justice system to better address the needs of at-risk individuals and reduce the likelihood of reoffending.

## 3 Hypothesis

With the immense recency of concern for model biases involving stereotypical natures relating to factors such as race and gender, it is all-too-simple to attribute such sentiments towards models dealing with data involving crime, a field that very much holds its stereotypes in modern day. We would like to challenge such narratives revolving around certain genders or races being more likely to commit/recommit crimes, thus leading to the following hypothesis: we believe that neither gender nor race will be among the top five key features correlating to how likely it is for a criminal to reoffend. Instead, other factors such as highest-achieved education level and past employment status will serve as notable features in recidivism indications. We believe that successfully achieving a higher level of education and greater occupational proficiency indicates greater moral development in one's character as individuals with such traits are more likely to work for their success. By examining these factors, we aim to identify significant predictors of recidivism and propose ideas that can help solve this problem.

## 4 Related work

Our project focuses on predicting recidivism by looking at prison data. This includes information like drug use, job security, and access to education. Other projects have also used machine learning to predict recidivism, often working on improving fairness and accuracy. For example, one study used lasso regression to combine community factors

with individual data, helping to reduce bias and make better predictions [1]. Another journal stated that algorithms like Logistic Regression are mostly used for recidivism [2]. Like our project, they showed that including community information makes recidivism prediction more fair and reliable.

## 5 Method

For preliminary analysis, we imported the dataset "recidivism.csv" that we got from the criminal justice system (link to dataset), into a Pandas Dataframe Object and labeled the object as df. Then we used the .head() and .info() command to look at the first few rows to check if it is valid. The raw data set has 18028 entities, with 53 total columns. We dropped a few features, including "ID" and "Recidivism in 1, 2, and 3 years" because it correlates with what we want to predict, which is Recidivism within 3 years.

For cleaning, we first checked out which features contained empty values and cleaned it as follows: for null Gang\_Affiliated value, we opted to default it to False because we believed if they belonged to a gang, they would have had an entry. For Supervision\_Risk\_Score\_First, we opted to use the average value as there are many different values. For the feature Supervision\_Level\_First we chose to have it be labeled "Standard" as that is the equivalent of the most normal value present. Prison\_Offense will be labeled as "Other" to avoid the complexity of introducing random crimes. Avg\_Days\_per\_DrugTest was also the average value. Any DrugTest\_Positive related feature will be 0 by default because we believed that if an offender did drugs, it would be noted. With that same logic, we opted to fill the jobs per year as 0 for empty space.

We opted to use Logistic Regressions, Decision Tree Classifier, and Forest Classifier first because they can be used to classify true or false, or classify the value for our Recidivism within 3 years value. Then later we attempted to do K-means clustering to see if there is any similarity between grouping that can help determine the relationship between features.

## 6 Experiments

With our model, we attempted to predict Recidivism within 3 years feature.

We set up a pipeline system with imputers that dealt with the feature listed above. For Jobs\_per\_year we directly engineered the values to be rounded up to an integers and filled empty row with 0 so that it can be used along the numerical pipeline. For any other category pipeline or other object we simply OneHotEncoded it with a value because there are many different responses.

Once the pipeline was set up, we made a train set, test set, and validation set ready to be used for our experiments. We then transformed the dataset with our pipeline and imported all of our models to be trained. We called fit\_transform() on the training and testing data, respectively.

For our Logistic Regression Model, we got an accuracy of about 72%, a precision of 74%, a recall of 81%, and an F1 Score of 77%. For the feature importance, the 5 features that influenced whether or not our model made the prediction on whether a former criminal would recidivate were the percentage of days employed, True for Gang Affiliation, when they are released at age 48+, when they are released at 18-22 years old, and when they have a Delinquency Report of 4 or more. We believe this model does pretty well and we were able to isolate the important features in predicting recidivism. The Logistic Regression model put more emphasis on features relating to age, indicating a likelihood of that younger people to commit crimes again compared to older ones.

For the Decision Tree Model, we observed an accuracy of about 62.23%, a precision of 68.93%, a recall of 66.29%, and an F1 score of 67.59%. For the feature importance, the 5 that had the highest importance were percentage of days employed, Average Days per Drug Test, Residence PUMA (Public Use Microdata Area which is a code for a geographical area), First Risk Score, and True for Gang Affiliation. We think that this tree model was not as accurate because some of these features did not show a clear pattern when we originally graphed it out in our notebook, but it did point out some notable features like Gang Affiliation, average day per drug test, and gang affiliation being a big

outlier for deciding recidivism.

For the Random Forest Model, we achieved an accuracy of about 71.38%, a precision of 73.44%, a recall of 81.19%, and an F1 score of 77.12%. The 5 features that stood out were the percentage of days employed, Average Days per Drug Test, First Supervision Score First, Residence PUMA, and Job Per Year. We believe that this is one of the better models we have because its accuracy is similar to our Logistic Regression Model, but Random Forest is known to handle more complex data. Similarly, it also pointed out similar features with the previous model we tried, this one put more emphasis on the employment of a prisoner, indicating that having incomes may likely decrease their chance of recidivism.

We settled with Random Forest because it is more suited for the kind of task we were trying to carry out, and we tuned it with 5 folds and optimized with F1 score using GridSearchCV. After tuning, we found that the best parameters were 10 for max depth,  $\log_2$  for max features, 2 for min samples leaf, 100 for  $n$  estimators and 5 for min sample splits. These parameters scored the best F1 score of around 76%.

Our best model from GridSearchCV gave us 71% for accuracy, 72% for precision, 85% for recall, and 78% for F1 score. Its most important features were Percent Days Employed, First Supervision Risk Score, True for Gang Affiliation, Average Day per DrugTest, and Prior Arrest Episodes PPViolationCharges.

We also implemented means clustering to see which features are prominent while sorting. First, we graphed out the elbow method to determine the best amount of cluster to use, and we ended up using  $k = 4$ , which gave us a silhouette score of 0.618, which is a reasonable score. We then got the recidivism rate, which is between 57-59%. While the results are close, it still revealed a tiny bit of nuance that can be pointed out.

Cluster 1 and Cluster 2 have the highest Recidivism rate, and Cluster 0 and 3 are slightly lower in the rate. For Cluster 0, the low recidivism most likely benefits from its frequent drug check-ins with its average day between checks being the lowest, it also has the highest employment rate compared to another cluster, indicating that individuals with stable employment are less likely to commit crimes. Cluster 1, has a longer drug test period, which indicates limited oversight. It has one of the employment and education statuses too, with those features not making the top 6 of the feature relevancy. Cluster 2 is similar in that it also has a longer drug test period, low employment and low education, indicating a potential need to improve in that area. Cluster 3 has a more frequent drug test period too; one thing worth noting is that it has the lowest THC level with it being around 0.15, which is the lowest compared to the other cluster, indicating that low drug use would make people less likely to commit a crime again.

## 7 Results

Link to code: <https://github.com/Quole0812/ML4VA>

### Logistic Regression Model:

| Metrics   | Training Sets |
|-----------|---------------|
| Accuracy  | 0.722684      |
| Precision | 0.7442254     |
| Recall    | 0.812324      |
| F1 Score  | 0.776785      |

### Feature Importance for Logistic Regression Model:

| Feature                          | Importance |
|----------------------------------|------------|
| Percent_Days_Employed            | 1.772146   |
| Gang_Affiliated_True             | 0.890877   |
| Age_at_Release_18-22             | 0.712471   |
| Age_at_Release_48 or older       | 0.710960   |
| Delinquency_Reports <sub>1</sub> | 0.496019   |

### Decision Tree Model:

| Metrics   | Training Sets |
|-----------|---------------|
| Accuracy  | 0.626178      |
| Precision | 0.691787      |
| Recall    | 0.668534      |
| F1 Score  | 0.679962      |

### Feature Importance for Decision Tree Model:

| Feature               | Importance |
|-----------------------|------------|
| Percent_Days_Employed | 0.130948   |
| Avg_Days_per_DrugTest | 0.076618   |
| Residence_PUMA        | 0.059841   |
| Risk_Score_First      | 0.041820   |
| Gang_Affiliated_True  | 0.025029   |

**Random Forest Model:**

| Metrics   | Training Sets |
|-----------|---------------|
| Accuracy  | 0.715751      |
| Precision | 0.731262      |
| Recall    | 0.824463      |
| F1 Score  | 0.775071      |

**Feature Importance for Random Forest Model:**

| Feature               | Importance |
|-----------------------|------------|
| Percent_Days_Employed | 0.080020   |
| Avg_Days_per_DrugTest | 0.050255   |
| Residence_PUMA        | 0.042861   |
| Risk_Score_Firs       | 0.041544   |
| Jobs_Per_Year         | 0.020216   |

**Best Random Forest Classifier:**

| Metrics   | Training Sets |
|-----------|---------------|
| Accuracy  | 0.715474      |
| Precision | 0.719685      |
| Recall    | 0.853408      |
| F1 Score  | 0.780862      |

**Feature Importance for Best Random Forest Classifier:**

| Feature               | Importance |
|-----------------------|------------|
| Percent_Days_Employed | 0.154333   |
| Risk_Score_First      | 0.052958   |
| Gang_Affiliated_True  | 0.049428   |
| Avg_Days_per_DrugTest | 0.038190   |
| PPViolationCharges_0  | 0.037956   |

**Cluster 0 Feature Results**

| Feature                         | Mean  |
|---------------------------------|-------|
| Avg_Days_per_DrugTest           | 34.14 |
| Supervision_Risk_Score_First    | 6.32  |
| Gender_M                        | 0.87  |
| Program_UnexcusedAbsences_0     | 0.73  |
| Delinquency_Reports_0           | 0.64  |
| Percent <sub>DaysEmployed</sub> | 0.54  |

**Cluster 1 Feature Results**

| Feature                      | Mean   |
|------------------------------|--------|
| Avg_Days_per_DrugTest        | 661.08 |
| Supervision_Risk_Score_First | 6.05   |
| Gender_M                     | 0.97   |
| Program_UnexcusedAbsences_0  | 0.87   |
| Race_BLACK                   | 0.75   |
| Delinquency_Reports_0        | 0.71   |

**Cluster 2 Feature Results**

| Feature                      | Mean   |
|------------------------------|--------|
| Avg_Days_per_DrugTest        | 280.39 |
| Supervision_Risk_Score_First | 6.02   |
| Gender_M                     | 0.91   |
| Program_UnexcusedAbsences_0  | 0.84   |
| Delinquency_Reports_0        | 0.71   |
| Race_BLACK                   | 0.64   |

**Cluster 3 Feature Results**

| Feature                      | Mean   |
|------------------------------|--------|
| Avg_Days_per_DrugTest        | 103.35 |
| Supervision_Risk_Score_First | 5.82   |
| Gender_M                     | 0.88   |
| Program_UnexcusedAbsences_0  | 0.87   |
| Delinquency_Reports_0        | 0.71   |
| Program_Attendances_0        | 0.67   |

**Cluster Recidivism Rate**

| Cluster number | Recidivism rate |
|----------------|-----------------|
| Cluster 0      | 0.578440        |
| Cluster 1      | 0.588629        |
| Cluster 2      | 0.592985        |
| Cluster 3      | 0.575113        |

## 8 Conclusion

We utilized machine learning techniques to predict recidivism within 3 years of parole using a dataset from a criminal justice system, which concluded that Random Forest was the best model. Random Forest identified the key characteristics that influenced the risk of a parolee recidivating, which were Percent Days Employed, Average Days per DrugTest, First Supervision Score First, Residence PUMA, and Job Per Year. Kmeans cluster pointed out that the features highly relevant to recidivism are the Employment rate, Average day per drug test, and Education rate. Consistent across these models were gang affiliation, age at release (with younger individuals more likely to commit a crime again), drug test frequency with frequent drug test users being the ones who re-offend less, and stable employment being the best for low recidivism too. This strongly supports our hypothesis that employment and education are significant indicators of recidivism in Virginia while race and gender are not. There were a few slip-ups where in our clustering model, race was a factor that was in the top 5 of the key features, but overall our hypothesis remains true for other models.

Virginia policymakers can utilize such data and analysis to adapt the policies they propose to allocate resources more effectively in order to create a safer society. These models can support parole officers in making data-based decisions to maximize the impacts of rehabilitation, prioritizing cases where individuals are at the highest risk of recidivism. For example, investing in employment programs for parolees or targeted interventions for gang-affiliated individuals could significantly reduce reoffending rates. These impact the well-being of all Virginians because reducing the statistic of convicted offenders who reoffend gives more opportunities for them to reunite with their loved ones and restart their lives, as well as reducing crime rates overall which protects public safety.

One shortcoming of this data is that there are other factors that impact recidivism rates – access to effective resources such as mental health counseling and community support – that are not easily quantifiable to be put in a dataset. Adding features like this to the dataset in the future could uncover more findings about the criminal justice system that would further benefit policymakers.

## 9 Contributions

Jackson Le worked on data preparation, data visualization, and model training, as well as training a random forest model. He also trained the cluster algorithm and analyzed it. Emily Chen wrote the checkpoint paper and performed model comparison analysis. Austin Shi trained the logistic regression and decision tree models, tuned hyperparameters for all models, visualized the data, and synthesized the final report.

## 10 References

- [1] E. Sevigny, T. Johnson, and J. Greathouse, Predicting Recidivism Fairly: A Machine Learning Application Using Contextual and Individual Data (accessed Nov. 5, 2024)
- [2] G. V. Travaini, F. Pacchioni, S. Bellumore, M. Bosia, and F. De Micco, "Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction," *Int. J. Environ. Res. Public Health*, vol. 19, no. 17, p. 10594, Aug. 2022. doi: 10.3390/ijerph191710594. (accessed Nov. 5, 2024).