A DEEP DIVE INTO GEORGIA'S RECIDIVISM

- *Abhishek Anumalla*
- *Freny Patel*
- *Lakshman Kushal Bogi*
- *Madhuri Muppa*
- *Nivedita J*

AIT614 Big Data Essentials

Dr. Liao

George Mason University

Introduction

What is Recidivisim?

Recidivism is the likelihood of an individual to commit new crimes after serving a sentence, measured by post-release arrests or convictions.

Importance of Recidivism:

Studying recidivism is crucial for understanding factors influencing post-criminal justice crime likelihood. This enhances public safety and supports rehabilitation for societal reintegration.

Exploratory Data Analysis:

Thoroughly explore the dataset to understand variable distributions, connections, and potential correlations.

Objective:

Develop new features to enhance prediction models and utilize ML approaches to predict recidivism across different time periods in the state of Georgia.

Abstract

This research navigates the intricacies of recidivism in Georgia, employing data cleaning, exploratory data analysis (EDA), and predictive modeling. Focusing on localized insights, it examines demographic, behavioral, and systemic factors to comprehend reoffending. Visualizations and machine learning models uncover patterns, aiding targeted interventions and contributing to evidence-based decisions.

Exploring age, gender, ethnicity, and education, the EDA phase uses Databricks DBFS for efficient data management. PySpark handles data preprocessing, while R produces visualizations like bar and line charts. These visuals illuminate relationships between demographic factors, behaviors, and recidivism, providing nuanced insights crucial for targeted interventions.

The predictive modeling phase embraces machine learning algorithms, including logistic regression, random forests, support vector machines, and gradient boosting. Hyperparameter tuning and cross-validation refine models, ensuring robust predictions. Chi-square testing for feature selection identifies influential factors, contributing to a comprehensive understanding of the driving forces behind recidivism.

Objective

- **The primary objective of this research is to investigate the multifaceted factors and underlying causes contributing to high rates of recidivism among individuals involved in the Georgia State criminal justice system during the years 2013-2015. To achieve this overarching goal, the study is guided by the following specific sub-questions:**
- **Examine the influence of demographic factors, including gender, race, and age at release, on recidivism rates, with the aim of understanding any disparities in the likelihood of reoffending.**
- **Investigate the correlation between prior criminal history, such as the number of prior arrests, convictions, and types of offenses, and recidivism, in order to assess the impact of an individual's criminal past.**
- **Analyze the role of behavioral factors, including substance abuse, mental health conditions, education level, employment history, and social support networks, in predicting recidivism, aiming to unveil the behavioral contributors to reoffending.**
- **Assess the effectiveness of supervision levels and risk scores in predicting recidivism and explore whether specific supervision strategies impact the likelihood of reoffending.**
- **Explore the impact of violations, program attendance, residence changes, and employment status on recidivism rates, to understand the effects of compliance and participation in rehabilitation and support programs.**
- **Investigate the patterns and trends in recidivism over specified time frames (1 year, 2 years, and 3 years), aiming to identify critical periods in which interventions may be most effective.**
- **Examine systemic and societal factors, including the availability of social services and societal support, to understand their role in recidivism and to provide insights into potential systemic improvements.**

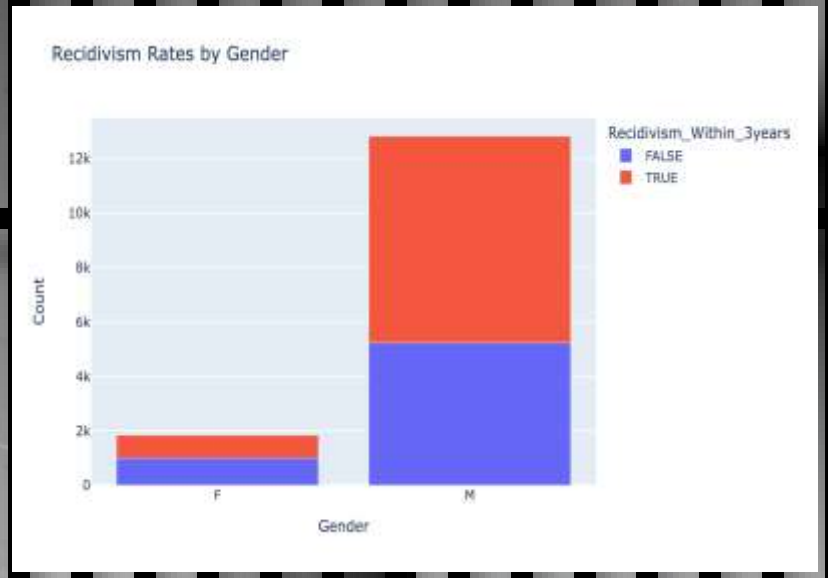
Data Cleaning

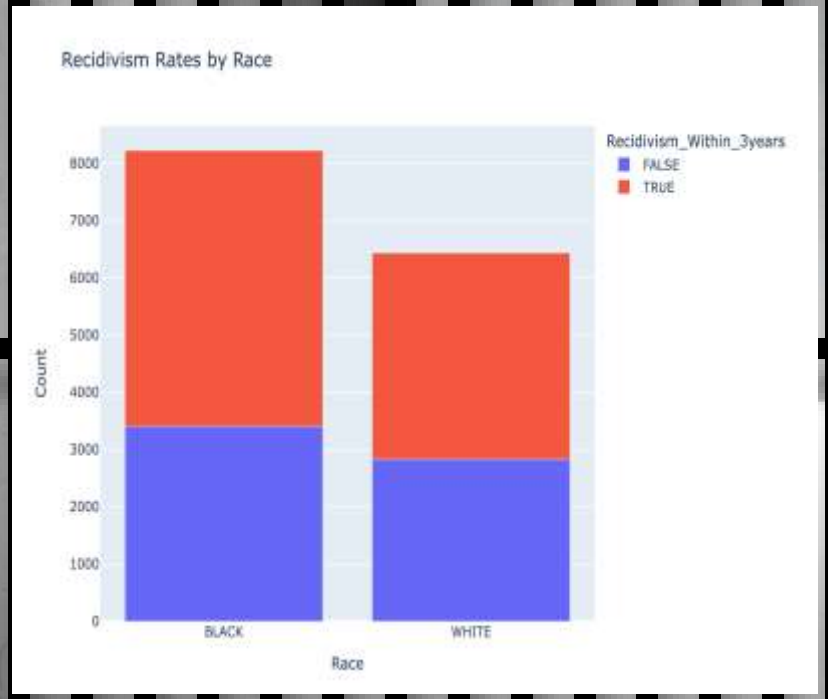
- Checking for Null Values: We identified missing data to understand gaps in the information.
- Handling Missing Values: We used interpolation method to fill in missing values, ensuring accuracy in columns like 'Supervision_Risk_Score_First' and drug test data.
- Filling Averages: For columns such as 'Avg_Days_per_DrugTest', 'Percent_Days_Employed', and 'Jobs_Per_Year', we applied the average value to fill in the missing data points.
- Dropping Columns: Columns with a high volume of missing values were dropped to maintain data integrity and quality in our analysis.

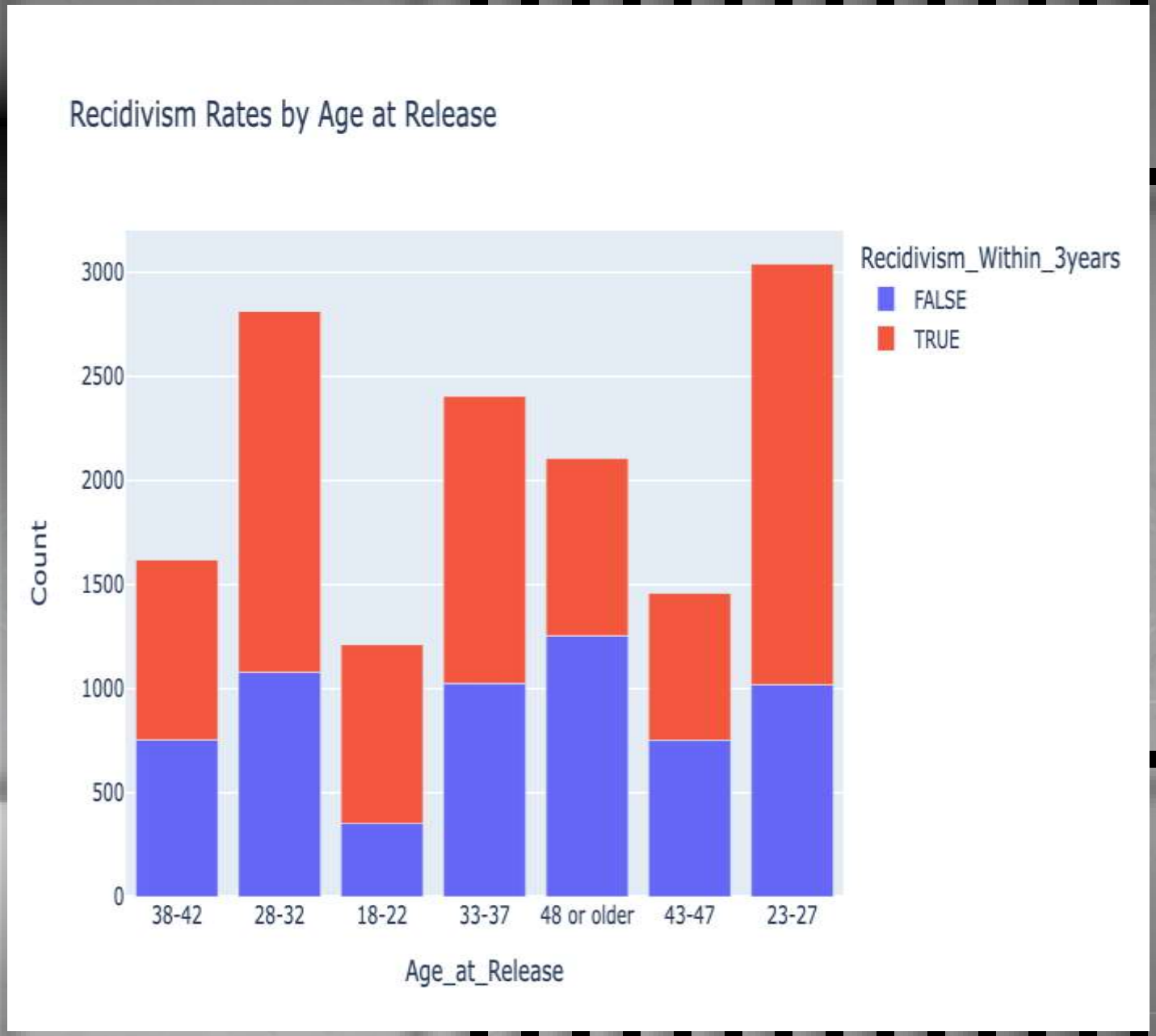
```
{'ID': 0,
'Gender': 0,
'Race': 0,
'Age_at_Release': 0,
'Residence_PUMA': 0, 'Gang_Affiliated': 0,
'Supervision_Risk_Score_First': 330,
'Supervision_Level_First': 1212,
'Education_Level': 0,
'Dependents': 0,
'Prison_Offense': 2321,
'Prison_Years': 0,
'Prior_Arrest_Episodes_Felony': 0,
'Prior_Arrest_Episodes_Misd': 0,
'Prior_Arrest_Episodes_Violent': 0,
'Prior_Arrest_Episodes_Property': 0,
'Prior_Arrest_Episodes_Drug': 0,
'Prior_Arrest_Episodes_PPViolationCharges':
0,
'Prior_Arrest_Episodes_DVCharges': 0,
'Prior_Arrest_Episodes_GunCharges': 0,
'Prior_Conviction_Episodes_Felony': 0,
.
.
}
```

Exploratory Data Analysis

Recidivism rates by Gender, Race and Age at Release





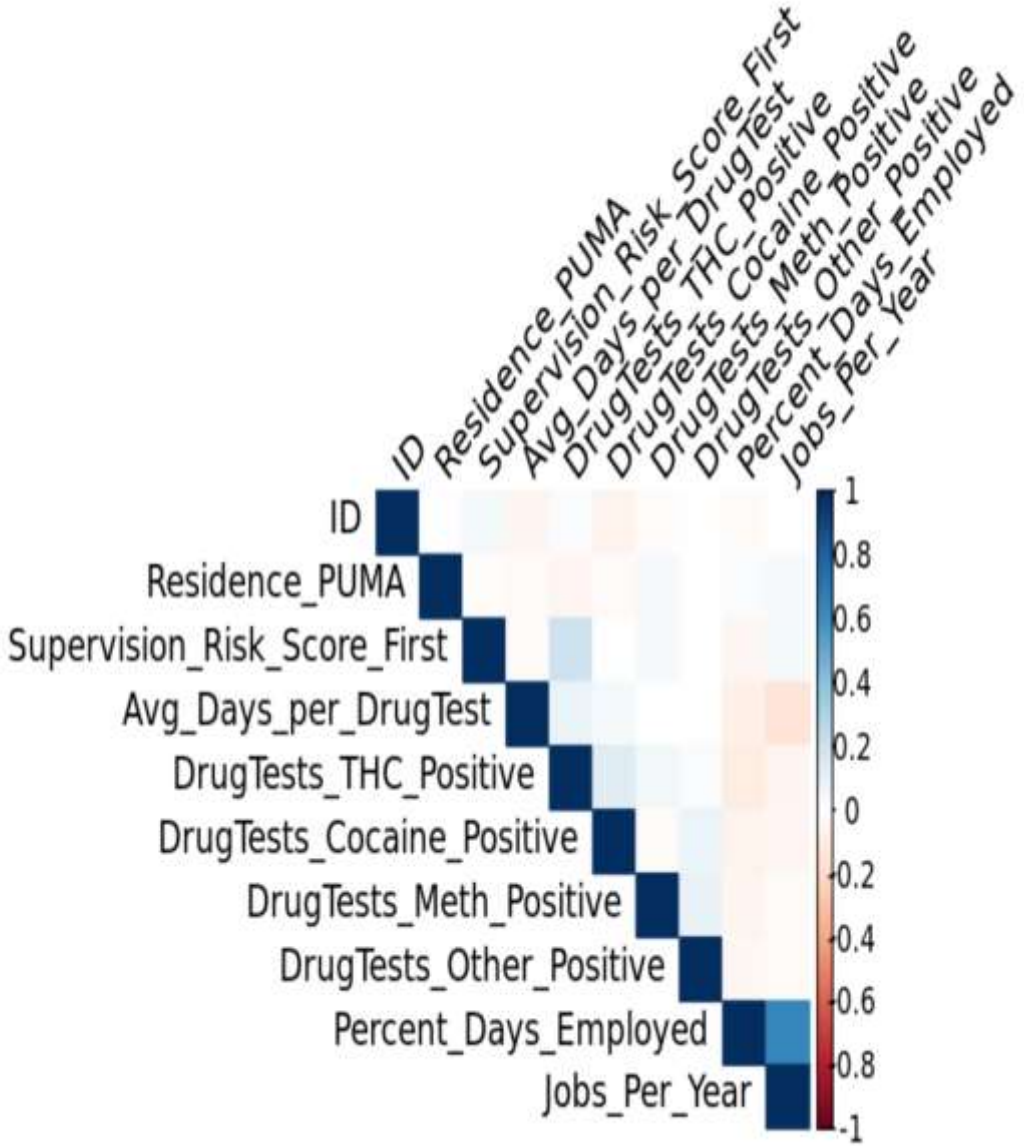


Correlation Matrix

	ID	Residence_PUMA
ID	1.0000000000	-0.0004423304
Residence_PUMA	-0.0004423304	1.0000000000
Supervision_Risk_Score_First	0.0457225191	-0.0155652436
Avg_Days_per_DrugTest	-0.0560999015	-0.0249946226
DrugTests_THC_Positive	0.0296710035	-0.0461982985
DrugTests_Cocaine_Positive	-0.0639725274	-0.0269794676
DrugTests_Meth_Positive	-0.0184012535	0.0302760316
DrugTests_Other_Positive	0.0095256535	-0.0023734948
Percent_Days_Employed	-0.0281152065	0.0280970981
Jobs_Per_Year	0.0026831444	0.0372343533

	SupervisionRiskScoreFirst	AvgDaysperDrugTest
ID	0.045722519	-0.056099901
Residence_PUMA	-0.015565244	-0.024994623
Supervision_Risk_Score_First	1.000000000	-0.027985731
Avg_Days_per_DrugTest	-0.027985731	1.000000000
DrugTests_THC_Positive	0.203461572	0.087865271
DrugTests_Cocaine_Positive	0.007238143	0.041602012
DrugTests_Meth_Positive	0.034827079	-0.005952553
DrugTests_Other_Positive	0.006583339	-0.003999103
Percent_Days_Employed	-0.041981807	-0.071880538
Jobs_Per_Year	0.041175915	-0.146898924

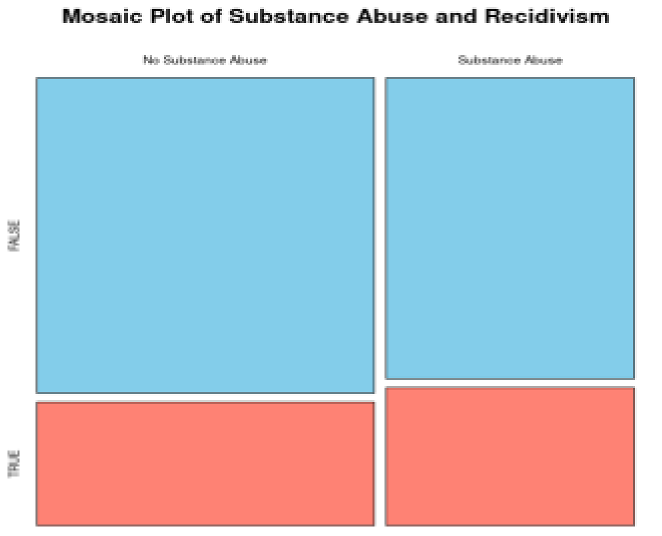
	DrugTests_THC_Positive	DrugTests_Cocaine_Positive
ID	0.02967100	-0.063972527
Residence_PUMA	-0.04619830	-0.026979468
Supervision_Risk_Score_First	0.20346157	0.007238143
Avg_Days_per_DrugTest	0.08786527	0.041602012
DrugTests_THC_Positive	1.00000000	0.135012281
DrugTests_Cocaine_Positive	0.13501228	1.000000000
DrugTests_Meth_Positive	0.05034512	-0.022388028
DrugTests_Other_Positive	0.02937612	0.074004412
Percent_Days_Employed	-0.10611484	-0.066846586
Jobs_Per_Year	-0.05278347	-0.048920213

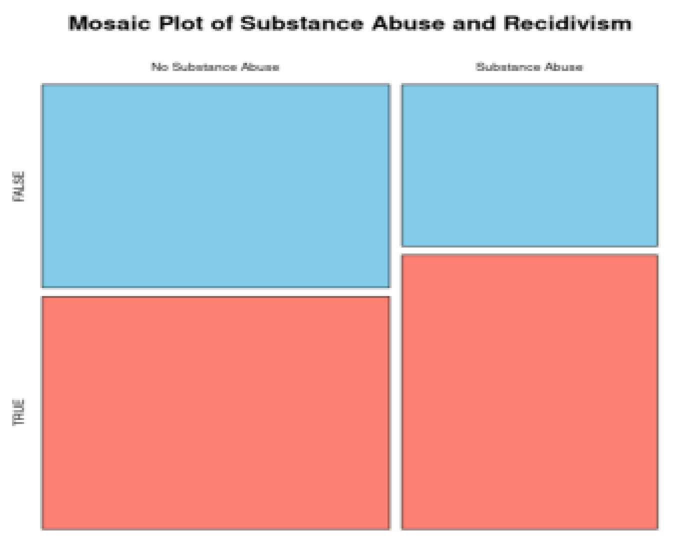


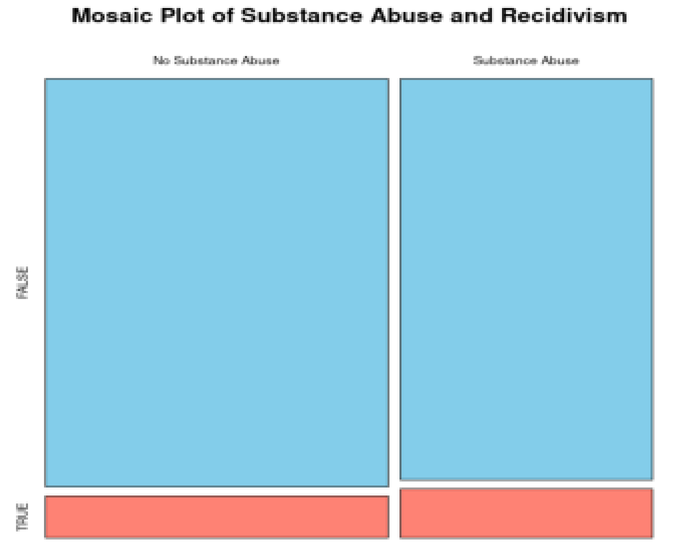
	DrugTests_Meth_Positive	DrugTests_Other_Positive
ID	-0.018401254	0.009525653
Residence_PUMA	0.030276032	-0.002373495
Supervision_Risk_Score_First	0.034827079	0.006583339
Avg_Days_per_DrugTest	-0.005952553	-0.003999103
DrugTests_THC_Positive	0.050345124	0.029376119
DrugTests_Cocaine_Positive	-0.022388028	0.074004412
DrugTests_Meth_Positive	1.000000000	0.091207804
DrugTests_Other_Positive	0.091207804	1.000000000
Percent_Days_Employed	-0.052962558	-0.034088930
Jobs_Per_Year	-0.029886922	-0.022995269

	Percent_Days_Employed	Jobs_Per_Year
ID	-0.02811521	0.002683144
Residence_PUMA	0.02809710	0.037234353
Supervision_Risk_Score_First	-0.04198181	0.041175915
Avg_Days_per_DrugTest	-0.07188054	-0.146898924
DrugTests_THC_Positive	-0.10611484	-0.052783470
DrugTests_Cocaine_Positive	-0.06684659	-0.048920213
DrugTests_Meth_Positive	-0.05296256	-0.029886922
DrugTests_Other_Positive	-0.03408893	-0.022995269
Percent_Days_Employed	1.00000000	0.641602601
Jobs_Per_Year	0.64160260	1.000000000

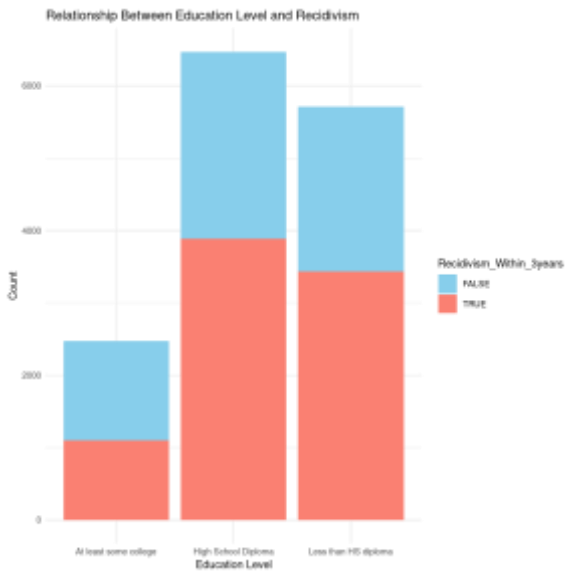
Impact of Substance Abuse on Recidivism

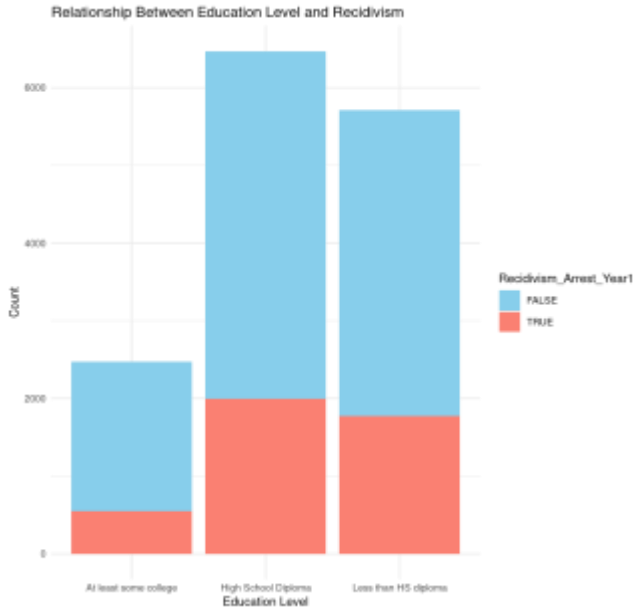


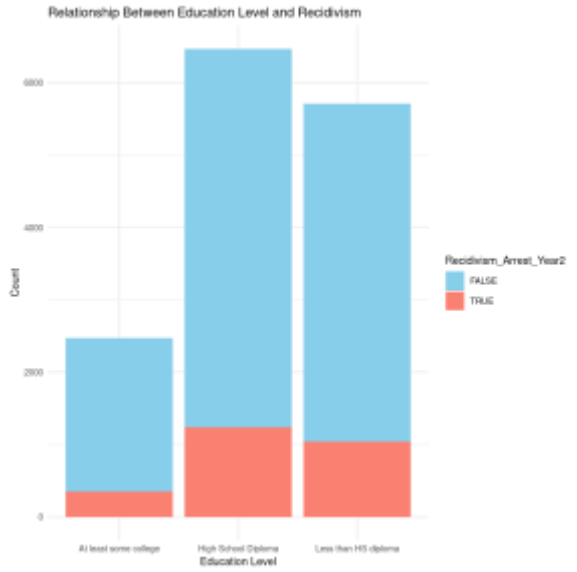


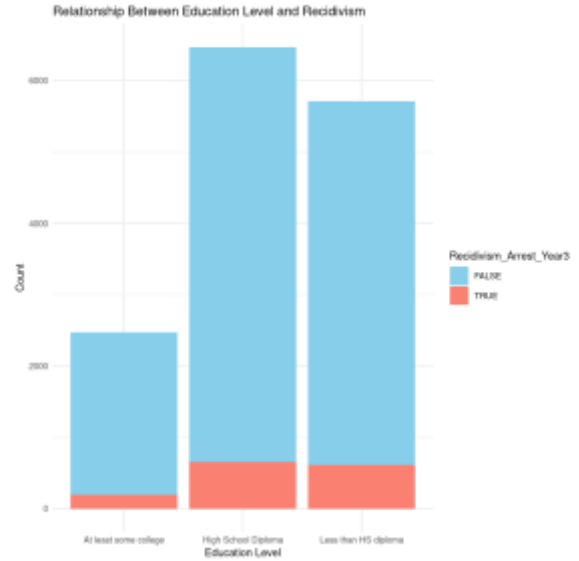


Relationship b/w Education Level and Recidivism

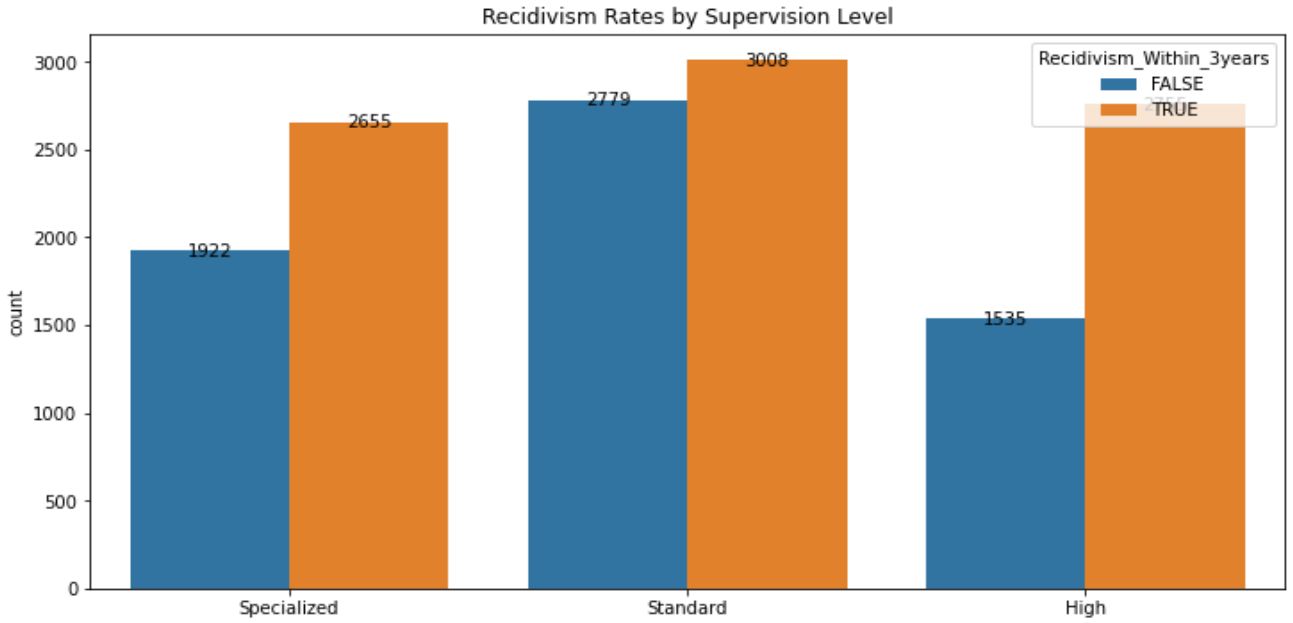


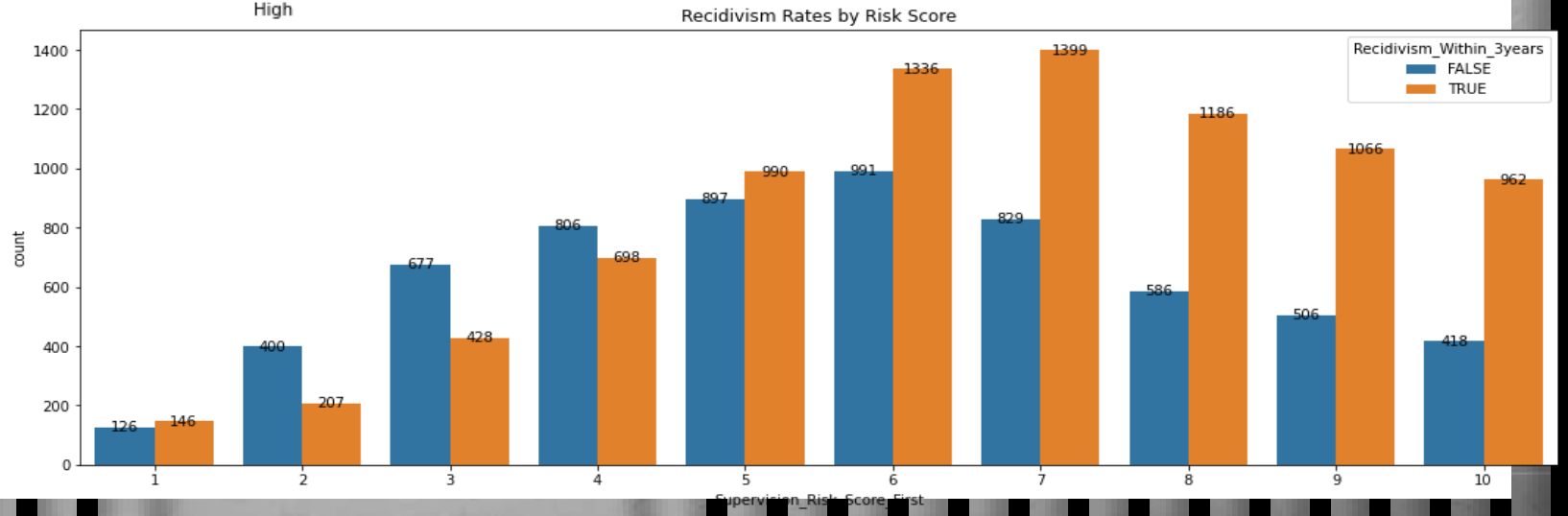




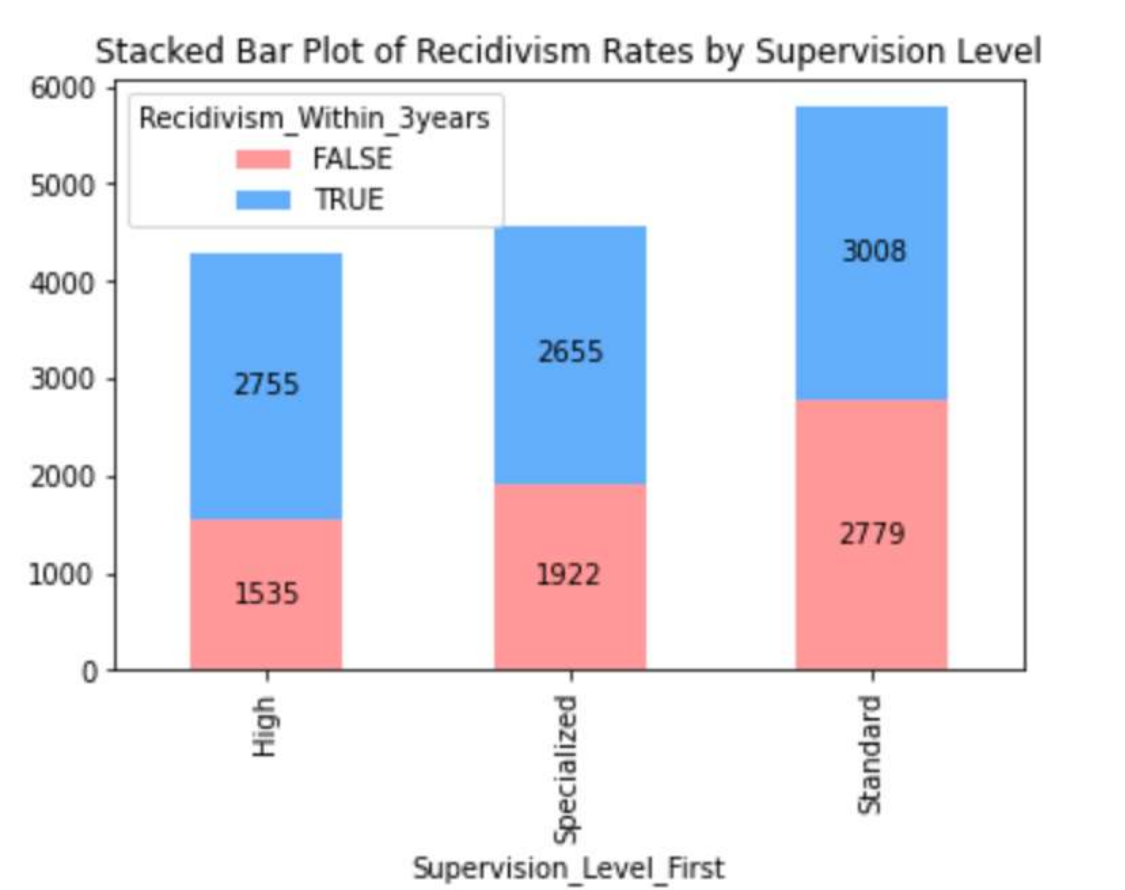


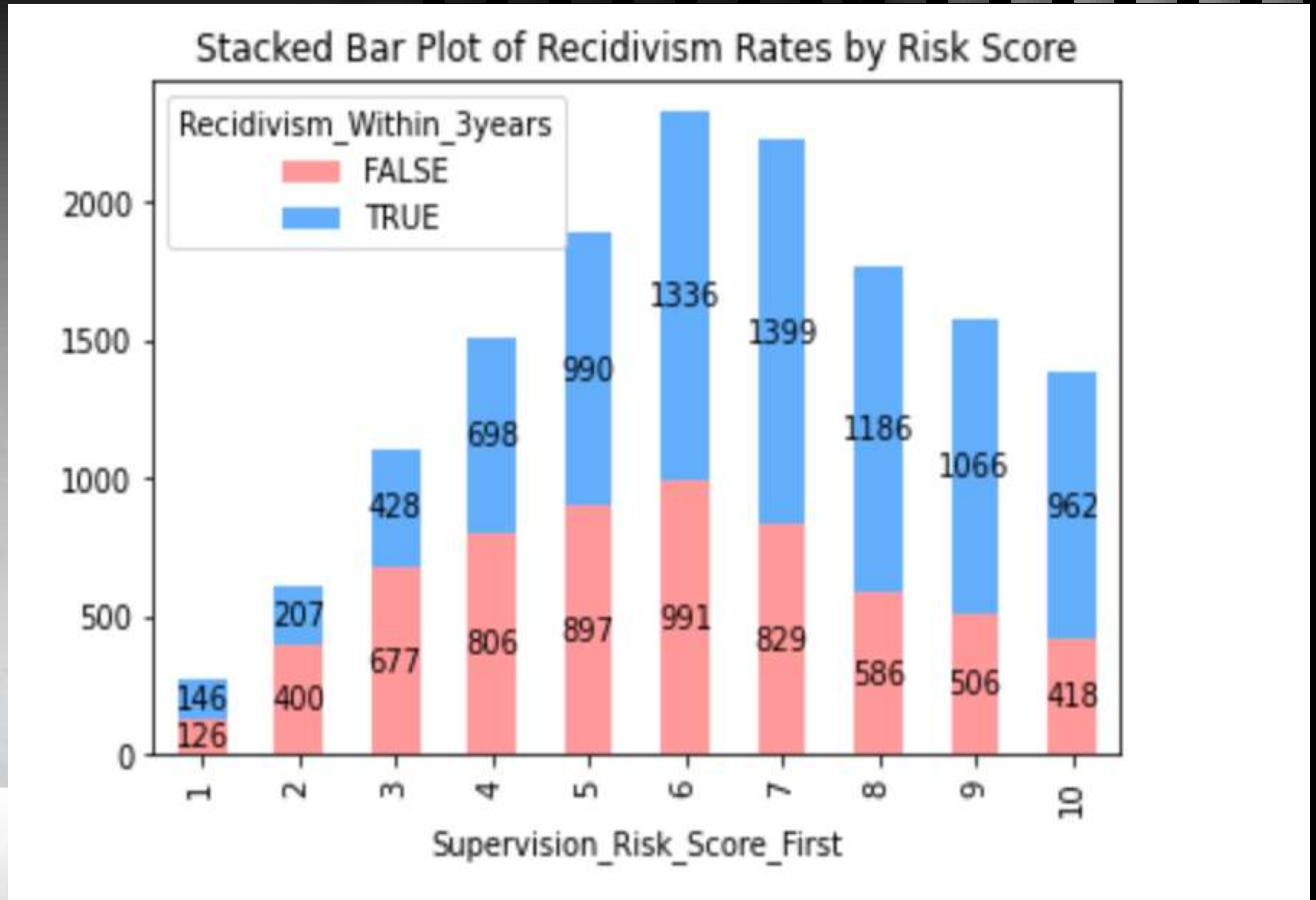
Recidivism Rate vs Supervision Level, Risk Score





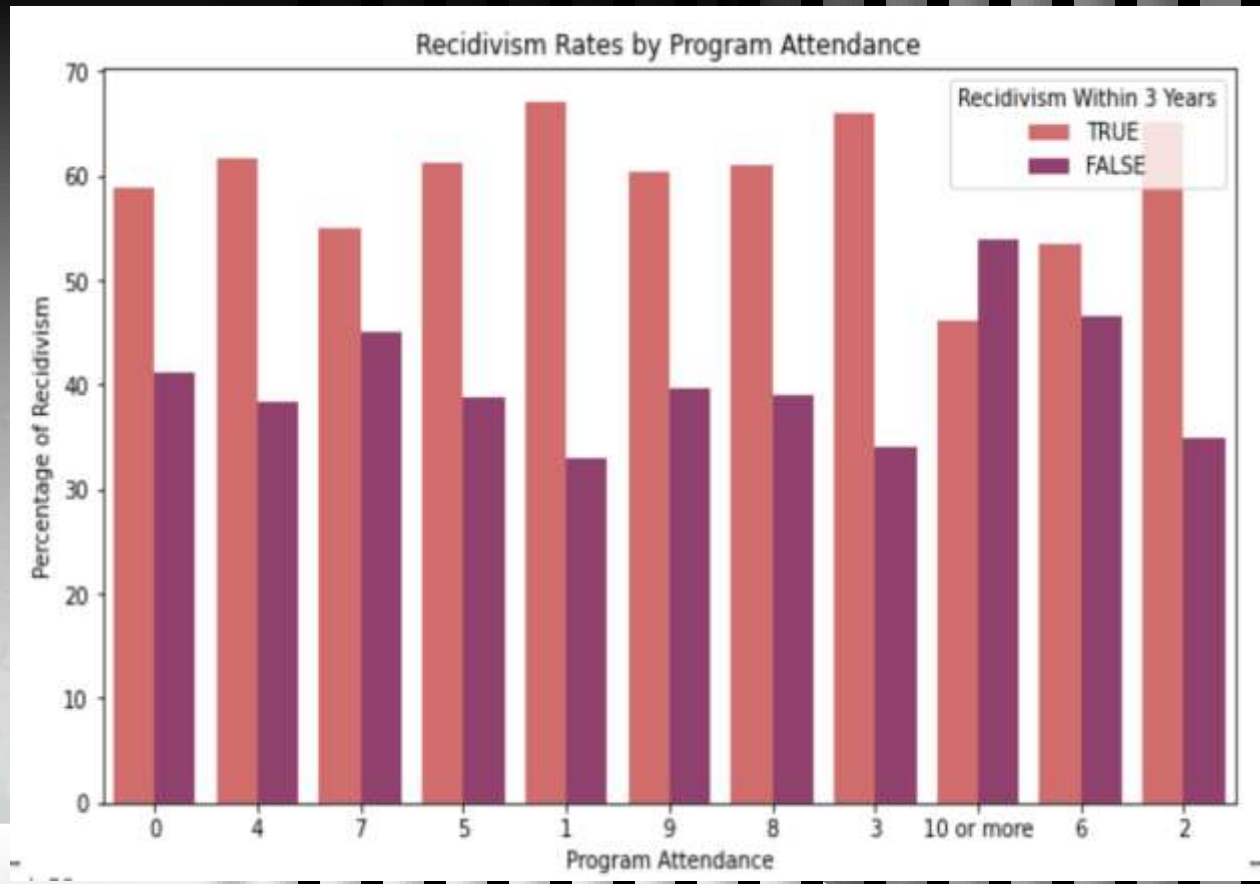
Recidivism Rate vs Supervision Level, Risk Score





Contingency Table for Program Attendance

	Program_Attendances_Recidivism_Within_3years	FALSE	TRUE
1	7	214	262
2	10 or more	1031	885
3	3	125	243
4	8	102	159
5	0	3151	4502
6	5	251	396
7	6	724	829
8	9	102	155
9	1	194	394
10	4	172	276
11	2	170	317



Chi Squared Test correlation b/w Prior Criminal History and Recidivism

X-squared3	Gang_Affiliated	2.613814e+02
X-squared5	Education_Level	2.071200e+02
X-squared27	Condition_MH_SA	1.883530e+02
X-squared35	Program_Attendances	1.706945e+02
X-squared17	Prior_Conviction_Episodes_Felony	1.663232e+02
X-squared4	Supervision_Level_First	1.518615e+02
X-squared13	Prior_Arrest_Episodes_Drug	1.479037e+02
X-squared22	Prior_Conviction_Episodes_PPViolationCharges	1.429879e+02
X-squared	Gender	1.229641e+02
X-squared34	Delinquency_Reports	1.092703e+02
X-squared21	Prior_Conviction_Episodes_Drug	8.973003e+01
X-squared15	Prior_Arrest_Episodes_DVCharges	7.539186e+01
X-squared11	Prior_Arrest_Episodes_Violent	7.255064e+01
X-squared36	Program_UnexcusedAbsences	6.041873e+01
X-squared31	Violations_Instruction	5.891359e+01
X-squared25	Prior_Revocations_Parole	5.873951e+01
X-squared38	Employment_Exempt	5.207081e+01
X-squared37	Residence_Changes	4.911948e+01
X-squared23	Prior_Conviction_Episodes_DomesticViolenceCharges	4.557649e+01
X-squared16	Prior_Arrest_Episodes_GunCharges	4.294193e+01
X-squared19	Prior_Conviction_Episodes_Viol	3.383475e+01
X-squared26	Prior_Revocations_Probation	3.362798e+01
X-squared6	Dependents	2.943423e+01
X-squared33	Violations_MoveWithoutPermission	2.549155e+01
X-squared28	Condition_Cog_Ed	2.248716e+01
X-squared32	Violations_FailToReport	1.077630e+01
X-squared1	Race	1.000292e+01
X-squared24	Prior_Conviction_Episodes_GunCharges	9.303368e+00
X-squared30	Violations_ElectronicMonitoring	6.840415e-01
X-squared29	Condition_Other	6.085710e-01

	PValue
X-squared39	0.000000e+00
X-squared40	0.000000e+00
X-squared41	0.000000e+00
X-squared42	3.456709e-263
X-squared14	2.269337e-179
X-squared9	7.728068e-148
X-squared10	8.605941e-112
X-squared12	5.782140e-110
X-squared2	7.621152e-109
X-squared18	3.759189e-108
X-squared20	1.162183e-77
X-squared8	2.158102e-68
X-squared7	1.083367e-60
X-squared3	8.578287e-59
X-squared5	1.057927e-45
X-squared27	7.275135e-43
X-squared35	1.991999e-31
X-squared17	7.913872e-36
X-squared4	1.056087e-33
X-squared13	3.729748e-30
X-squared22	5.913800e-33
X-squared	1.419923e-28
X-squared34	1.041390e-22
X-squared21	3.276203e-20
X-squared15	3.859778e-18
X-squared11	1.213268e-15
X-squared36	4.783910e-13
X-squared31	1.647506e-14
X-squared25	1.799908e-14
X-squared38	5.353491e-13
X-squared37	1.230248e-10
X-squared23	1.467930e-11
X-squared16	5.638893e-11
X-squared19	5.999749e-09
X-squared26	6.672595e-09
X-squared6	1.814983e-06
X-squared33	4.443246e-07
X-squared28	2.115527e-06
X-squared32	1.028082e-03
X-squared1	1.562924e-03
X-squared24	2.287329e-03
X-squared30	4.081985e-01
X-squared29	4.353264e-01

Machine Learning Models

#LOGISTIC REGRESSION without hyperparameter tuning and cross validation	Area under ROC curve: 0.7484329712396544 Accuracy: 0.6860830136030694
#RANDOM FOREST without hyperparameter and cross validation	Area under ROC curve: 0.7588282085063021 Accuracy: 0.6927101499825602
#RANDOM FOREST with hyperparameter and cross validation	Area under ROC curve: 0.7786800566755349 Accuracy: 0.7084059993024067
# GRADIENT BOOSTING without hyperparameter and cross validation	Area under ROC curve: 0.7727870441721143 Accuracy: 0.7160795256365539
#SUPPORT VECTOR MACHINE with hyperparameter tuning and cross validation	Tuned model - Area under ROC curve: 0.7839506326282045 Tuned model - Accuracy: 0.7185211021974189
#SUPPORT VECTOR MACHINE without hyperparameter tuning and cross validation	Area under ROC curve: 0.7847301697780221 Accuracy: 0.7185211021974189
#LOGISTIC REGRESSION with hyperparameter tuning and cross validation	Area under ROC curve: 0.7832108677819503 Accuracy: 0.7220090687129403

Conclusion

In our comprehensive evaluation of diverse machine learning models for recidivism prediction, we emphasize the significance of refining models through hyperparameter tuning and cross-validation. Notably, Logistic Regression, subjected to these optimization techniques, emerges as the optimal choice, exhibiting robust performance with the highest Area under the ROC curve (0.7832) and accuracy (0.7220). While Support Vector Machine competes effectively, our generic conclusion underscores the universal importance of fine-tuning model parameters for enhanced predictive accuracy, offering insights applicable beyond recidivism prediction.

Beyond its promising predictive prowess, the optimized Logistic Regression model, resulting from hyperparameter tuning and cross-validation, holds paramount implications for recidivism control. The heightened accuracy and precision empower criminal justice professionals to discern individuals with a higher likelihood of reoffending more accurately. This targeted identification facilitates the implementation of personalized intervention and rehabilitation strategies, optimizing resource allocation. By leveraging advanced machine learning techniques, our model contributes to a data-driven and proactive approach in the criminal justice system, paving the way for more effective programs and policies to curb recidivism rates.

References:

“NIJ’s Recidivism Challenge Full Dataset | Office of Justice Programs,” data.ojp.usdoj.gov.

https://data.ojp.usdoj.gov/Courts/NIJ-s-Recidivism-Challenge-Full-Dataset/ynf5-u8nk (accessed Oct. 24, 2023).

Cannonier, C., Galloway Burke, M., & Mitchell, E. (2020). The Impact of a Reentry and Aftercare Program on Recidivism. The Review of Black Political Economy, 48(1), 003464462097393.

https://doi.org/10.1177/0034644620973931. (accessed Oct. 24, 2023)

Khan, Shahid & Lucas, Mia. (2023). The effectiveness of restorative justice programs in reducing/recidivism. https://www.researchgate.net/publication/372750946_The_effectiveness_of_restorative_justice_programs_in_reducing_recidivism. (accessed Oct. 24, 2023)

Yukhnenko D, Blackwood N, Fazel S. Risk factors for recidivism in individuals receiving community sentences: a systematic review and meta-analysis. CNS Spectr. 2020 Apr;25(2):252-263.

https://doi.org/10.1017/S1092852919001056. Epub 2019 Jun 20. PMID: 31218975; PMCID: PMC7183820. (accessed Oct. 24, 2023)

Z. Zhang, Z. Huang, Z. Wan and L. Meng, "Comparative Analysis of Machine Learning Models for Recidivism Prediction Based on Chi-square Test," 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), Nanjing, China, 2021, pp. 21-26, https://doi.org/10.1109/ICAA53760.2021.00012. (accessed Oct. 24, 2023)

“Recidivism Forecasting Challenge.” National Institute of Justice,

nij.ojp.gov/funding/recidivism-forecasting-challenge Accessed 24 Oct. 2023.


Thank You