Crime Rate Knowledge Mining in U.S. Communities

Submitted in partial fulfillment of the
requirements of the course

Business Intelligence Lab (ITL601)

In

T. E. Information Technology

By

Tanmay Bhatkar	**221009**
Mazin Bangi	**221008**
Sahil Bangera	**221007**
Shannen Anthony	**221004**

Supervisor:

Ms. Shree Jaswal
Assistant Professor

Department of Information Technology
St. Francis Institute of Technology
(Engineering College)

University of
Mumbai 2024-2025

CERTIFICATE

This is to certify that the project entitled **"Crime Rate Knowledge Mining in U.S. Communities "** is a Bonafide work of **Tanmay Bhatkar (09), Mazin Bangi (08), Sahil Bangera (07) and Shannen Anthony (04)** submitted in partial fulfillment of the requirements of the course **Business Intelligence Lab (ITL601)**

Ms. Shree Jaswal

Supervisor/Guide

St. Francis Institute of Technology, Mumbai-400 103
Department Of Information Technology

A.Y. 2024-2025
Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

Business Intelligence Mini Project

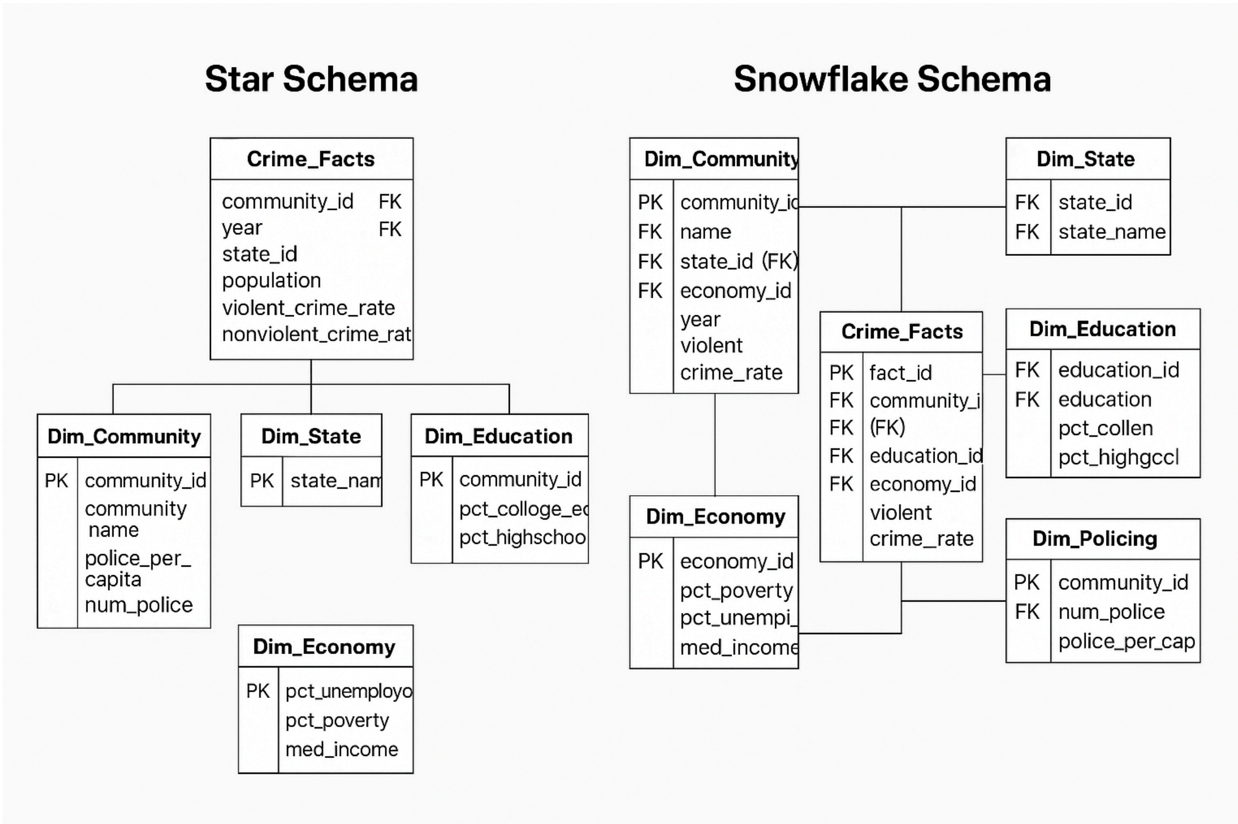
1. **Aim:** Develop a Business Intelligence Mini Project for a particular case study
0. **Objectives:** After study of this experiment, the students will be able to develop mini project
0. **Outcomes:**
   CO6: Apply BI to solve practical problems: Analyze the problem domain, use the data collected in enterprise apply the appropriate data mining technique, interpret and visualize the results and provide decision support
0. **Prerequisite:** Study of DM&BI Tools.
0. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool, Orange tool, BI Tool.
0. **Theory: Nil**
0. **Laboratory Exercise:** Each group select one case study for this. A BI report (Initial pages shared separately) must be prepared outlining the following steps:
   a. Write the problem statement for your case study
   a. Draw star schema and snowflake schema
   b. Give dataset details, identifying which data mining task is needed
   a. Download and use a standard data mining dataset available for the problem. Some links for data mining datasets are WEKA site, UCI Machine Learning Repository, KDD site, KDD Cup etc.
   a. Implement the data mining algorithm using Weka and Orange
   a. Interpret and visualize the results using BI tool like Qlikview & Tableau
   a. After interpretation clearly provide the BI decision that is to be taken
0. **Post-Experiment Exercise:**
   a. Conclusion:
   - o Summary of mini project
0. **Reference:** Business Intelligence: Data Mining and Optimization for Decision Making by Carlo Vercellis, Wiley India Publication

Laboratory Exercise:

a. Problem Statement

"To analyze socio-economic, demographic, and law enforcement factors influencing crime rates in U.S. communities and build a predictive model to identify high-risk areas for proactive resource allocation."

b. Draw star schema and snowflake schema



c. Dataset Details

Source: UCI Machine Learning Repository: Communities and Crime

Link: https://archive.ics.uci.edu/dataset/183/communities+and+crime

Size: 1,994 instances, 128 attributes (socio-economic factors + crime statistics).

Target Variable: ViolentCrimesPerPop (numerical).

Data Mining Task: Regression (predict crime rate) and Classification (identify high-risk communities).

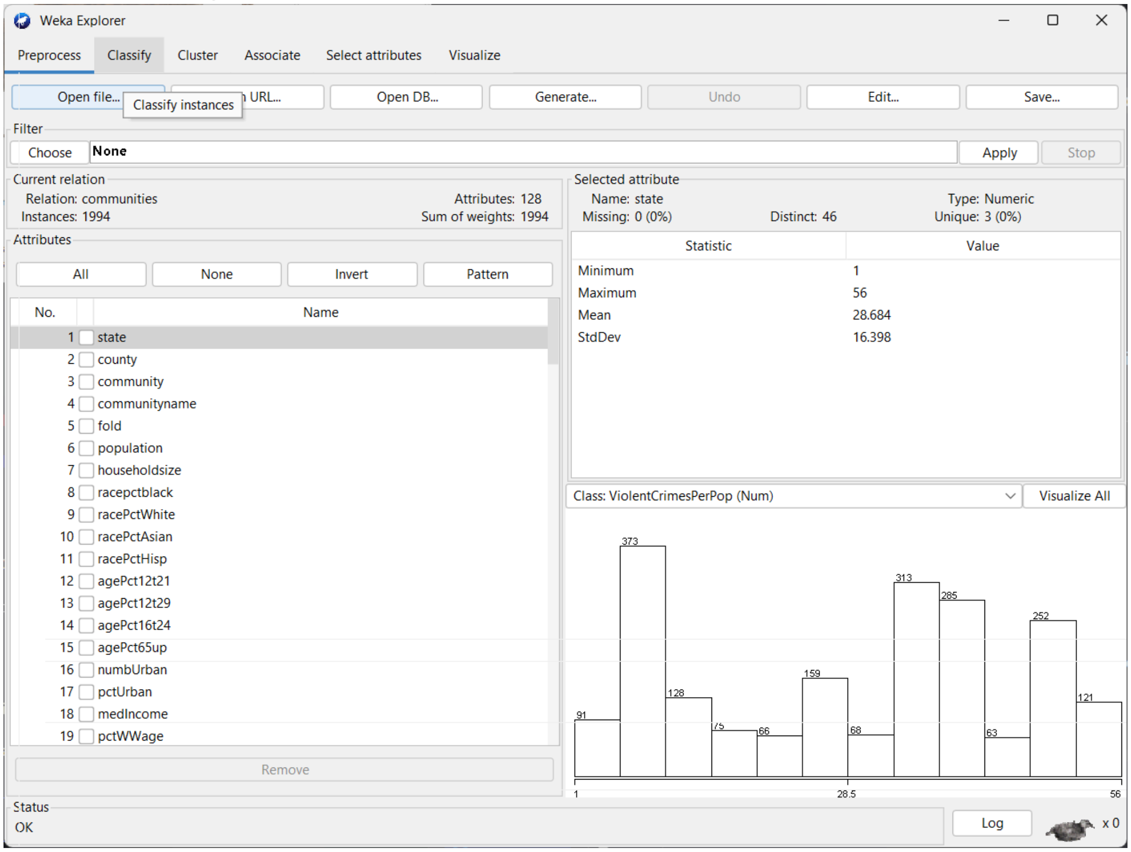
Key Attributes: population, medIncome, pctUnemployed, pctHousNoPhone, policePerPop.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	state	county	communit	communit	fold	populatior	household	racepctbla	racePctWh	racePctAsi	racePctHis	agePct12t	agePct12t	agePct16t	agePct65u	numbUrba	pctUrban	medIncom	pctWWage	pctWFarm	pctWInvIn	pctWSocS	pctWPubA
2	8	?	?	Lakewood	1	0.19	0.33	0.02	0.9	0.12	0.17	0.34	0.47	0.29	0.32	0.2	1	0.37	0.72	0.34	0.6	0.29	0.15
3	53	?	?	Tukwilacit	1	0	0.16	0.12	0.74	0.45	0.07	0.26	0.59	0.35	0.27	0.02	1	0.31	0.72	0.11	0.45	0.25	0.29
4	24	?	?	Aberdeent	1	0	0.42	0.49	0.56	0.17	0.04	0.39	0.47	0.28	0.32	0	0	0.3	0.58	0.19	0.39	0.38	0.4
5	34	5	81440	Willingbor	1	0.04	0.77	1	0.08	0.12	0.1	0.51	0.5	0.34	0.21	0.06	1	0.58	0.89	0.21	0.43	0.36	0.2
6	42	95	6096	Bethlehem	1	0.01	0.55	0.02	0.95	0.09	0.05	0.38	0.38	0.23	0.36	0.02	0.9	0.5	0.72	0.16	0.68	0.44	0.11
7	6	?	?	SouthPasa	1	0.02	0.28	0.06	0.54	1	0.25	0.31	0.48	0.27	0.37	0.04	1	0.52	0.68	0.2	0.61	0.28	0.15
8	44	7	41500	Lincolntow	1	0.01	0.39	0	0.98	0.06	0.02	0.3	0.37	0.23	0.6	0.02	0.81	0.42	0.5	0.23	0.68	0.61	0.21
9	6	?	?	Selmacity	1	0.01	0.74	0.03	0.46	0.2	1	0.52	0.55	0.36	0.35	0	0	0.16	0.44	1	0.23	0.53	0.97
10	21	?	?	Hendersor	1	0.03	0.34	0.2	0.84	0.02	0	0.38	0.45	0.28	0.48	0.04	1	0.17	0.47	0.36	0.34	0.55	0.48
11	29	?	?	Claytoncit	1	0.01	0.4	0.06	0.87	0.3	0.03	0.9	0.82	0.8	0.39	0.02	1	0.54	0.59	0.22	0.86	0.42	0.02
12	6	?	?	DalyCitycit	1	0.13	0.71	0.15	0.07	1	0.41	0.4	0.52	0.35	0.33	0.15	1	0.49	0.71	0.16	0.47	0.36	0.4
13	36	?	?	RockvilleC	1	0.02	0.46	0.08	0.91	0.07	0.1	0.34	0.36	0.22	0.57	0.04	1	0.72	0.53	0.23	0.74	0.63	0.17
14	25	21	44105	Needhamt	1	0.03	0.47	0.01	0.96	0.13	0.02	0.29	0.32	0.2	0.52	0.04	1	0.8	0.55	0.18	0.87	0.51	0.07
15	55	87	30075	GrandChut	1	0.01	0.44	0	0.98	0.04	0.01	0.35	0.53	0.32	0.23	0.02	0.77	0.46	0.77	0.41	0.73	0.28	0.1
16	6	?	?	DanaPoint	1	0.04	0.36	0.01	0.85	0.14	0.26	0.32	0.46	0.3	0.31	0.05	1	0.71	0.67	0.42	0.55	0.25	0.14
17	19	187	91370	FortDodge	1	0.03	0.34	0.06	0.93	0.03	0.03	0.39	0.41	0.28	0.58	0	0	0.18	0.42	0.81	0.49	0.62	0.37
18	36	1	1000	Albanycity	1	0.15	0.31	0.4	0.63	0.14	0.06	0.58	0.72	0.65	0.47	0.16	1	0.22	0.52	0.1	0.51	0.48	0.39
19	34	27	17650	Denvilleto	1	0.01	0.53	0.01	0.94	0.2	0.03	0.34	0.39	0.27	0.36	0.02	0.76	0.79	0.77	0.13	0.77	0.44	0.15
20	18	?	?	Valparaiso	1	0.02	0.47	0.01	0.97	0.07	0.02	0.7	0.67	0.63	0.37	0	0	0.33	0.56	0.28	0.62	0.43	0.21
21	42	129	66376	Rostravert	1	0	0.41	0.05	0.96	0.01	0.01	0.37	0.37	0.24	0.55	0.01	0.58	0.23	0.34	0.33	0.51	0.7	0.36
22	6	?	?	Modestoci	1	0.25	0.54	0.05	0.71	0.48	0.3	0.42	0.48	0.28	0.32	0.26	1	0.33	0.55	0.37	0.37	0.39	0.64
23	12	31	?	Jacksonvill	1	1	0.42	0.47	0.59	0.12	0.05	0.41	0.53	0.34	0.33	1	0.99	0.28	0.62	0.16	0.36	0.4	0.3
24	41	?	?	KlamathFa	1	0.01	0.34	0.02	0.87	0.07	0.11	0.49	0.56	0.43	0.47	0	0	0.13	0.4	0.26	0.42	0.52	0.41
25	19	193	93926	SiouxCityc	1	0.11	0.43	0.04	0.89	0.09	0.06	0.45	0.48	0.31	0.46	0.13	1	0.22	0.52	0.44	0.49	0.56	0.41
26	6	?	?	Delanocity	1	0.02	0.96	0.05	0	1	1	0.54	0.58	0.39	0.33	0	0	0.16	0.61	0.41	0.14	0.49	0.92
27	8	?	?	Goldencity	1	0	0.33	0.02	0.91	0.16	0.09	0.55	0.63	0.53	0.31	0.02	1	0.29	0.64	0.6	0.57	0.3	0.15
28	6	?	?	Gardenaci	1	0.06	0.49	0.46	0	1	0.43	0.35	0.5	0.32	0.34	0.08	1	0.35	0.68	0.16	0.36	0.36	0.38
29	39	29	61798	Perrytown	1	0.01	0.37	0.01	0.99	0.02	0.01	0.35	0.38	0.23	0.59	0	0	0.23	0.38	0.39	0.51	0.64	0.35
30	54	?	?	Beckleycit	1	0.01	0.27	0.43	0.64	0.08	0.01	0.36	0.33	0.22	0.74	0	0	0.12	0.1	0.11	0.39	0.84	0.48

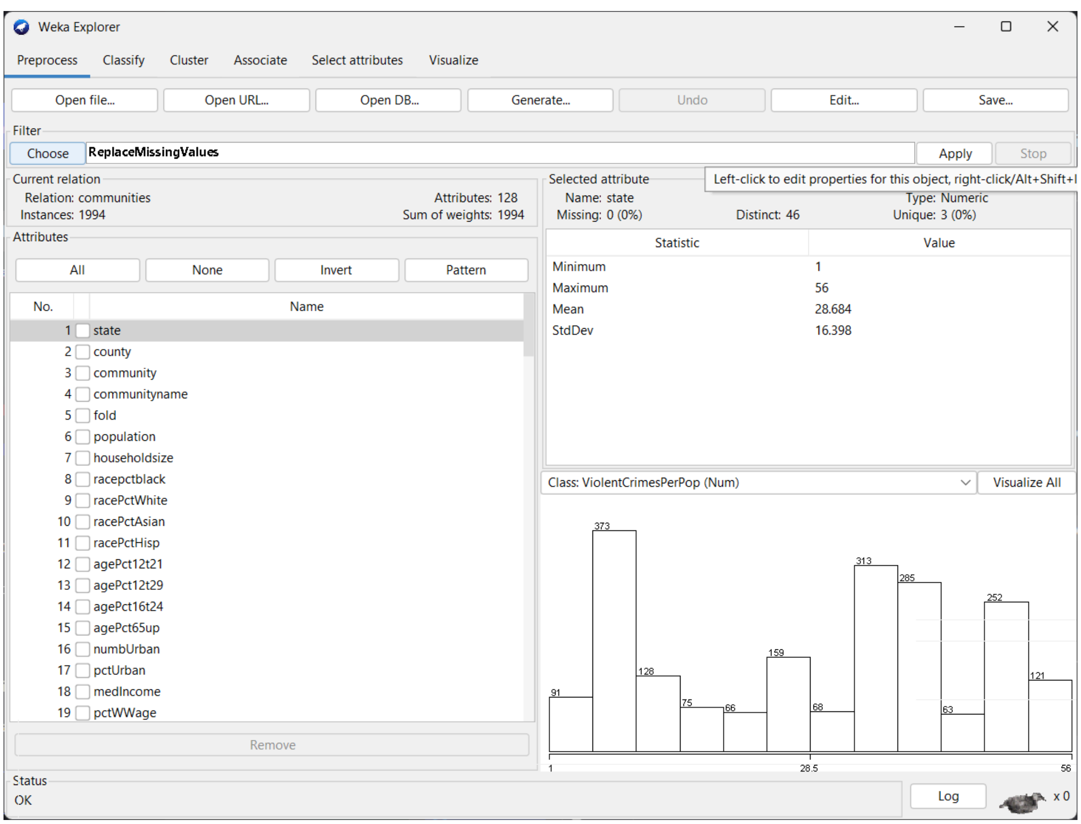
Snapshot of dataset

d. Implementation

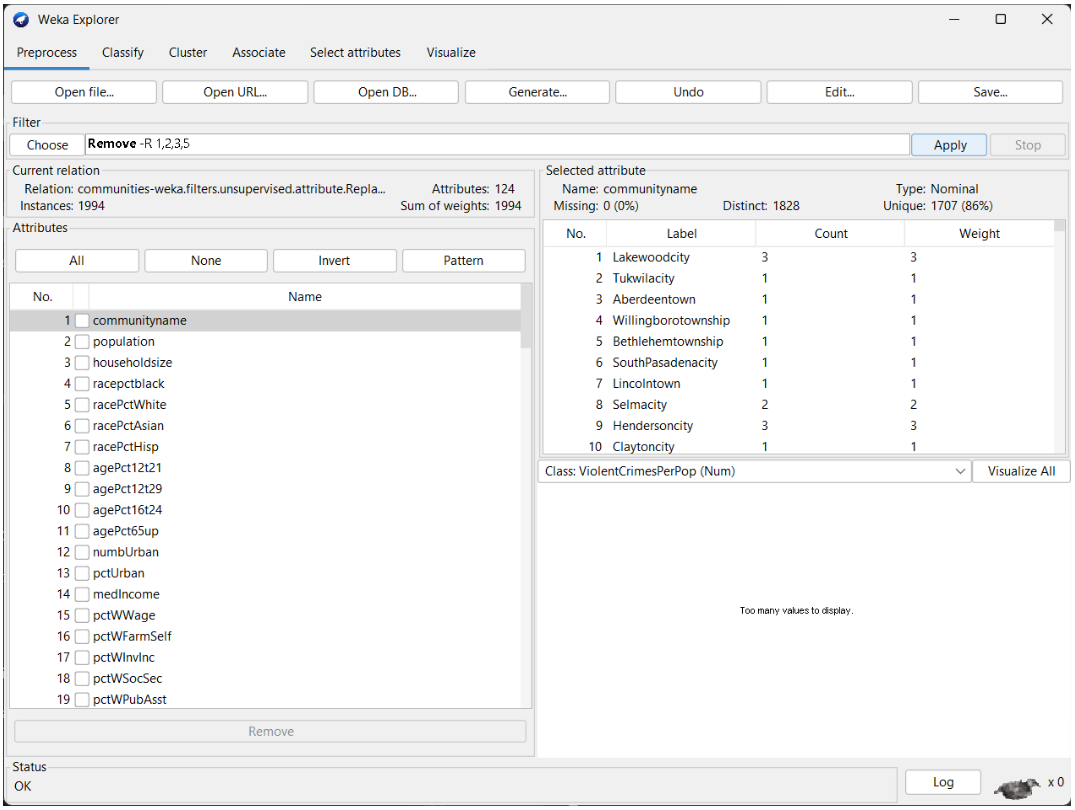
Prediction using WEKA



Loading the dataset in weka

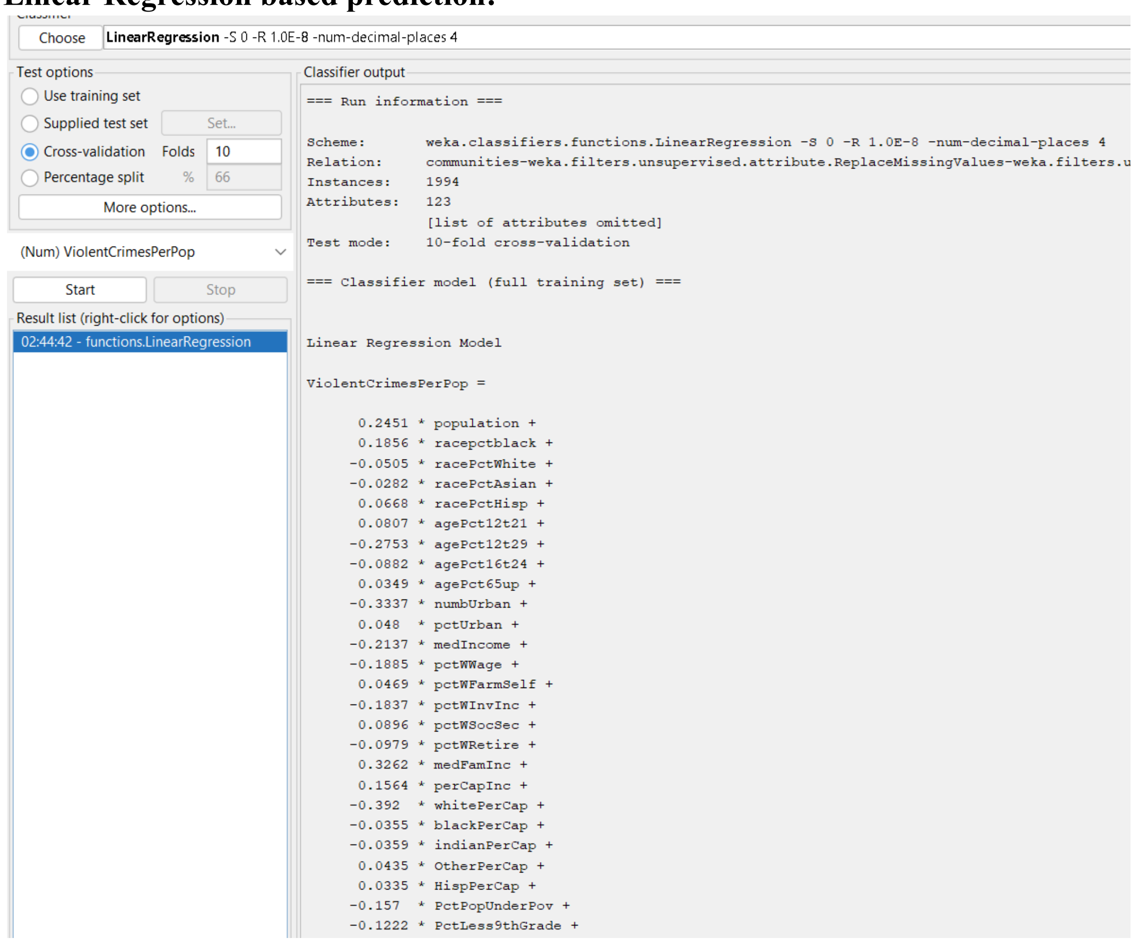


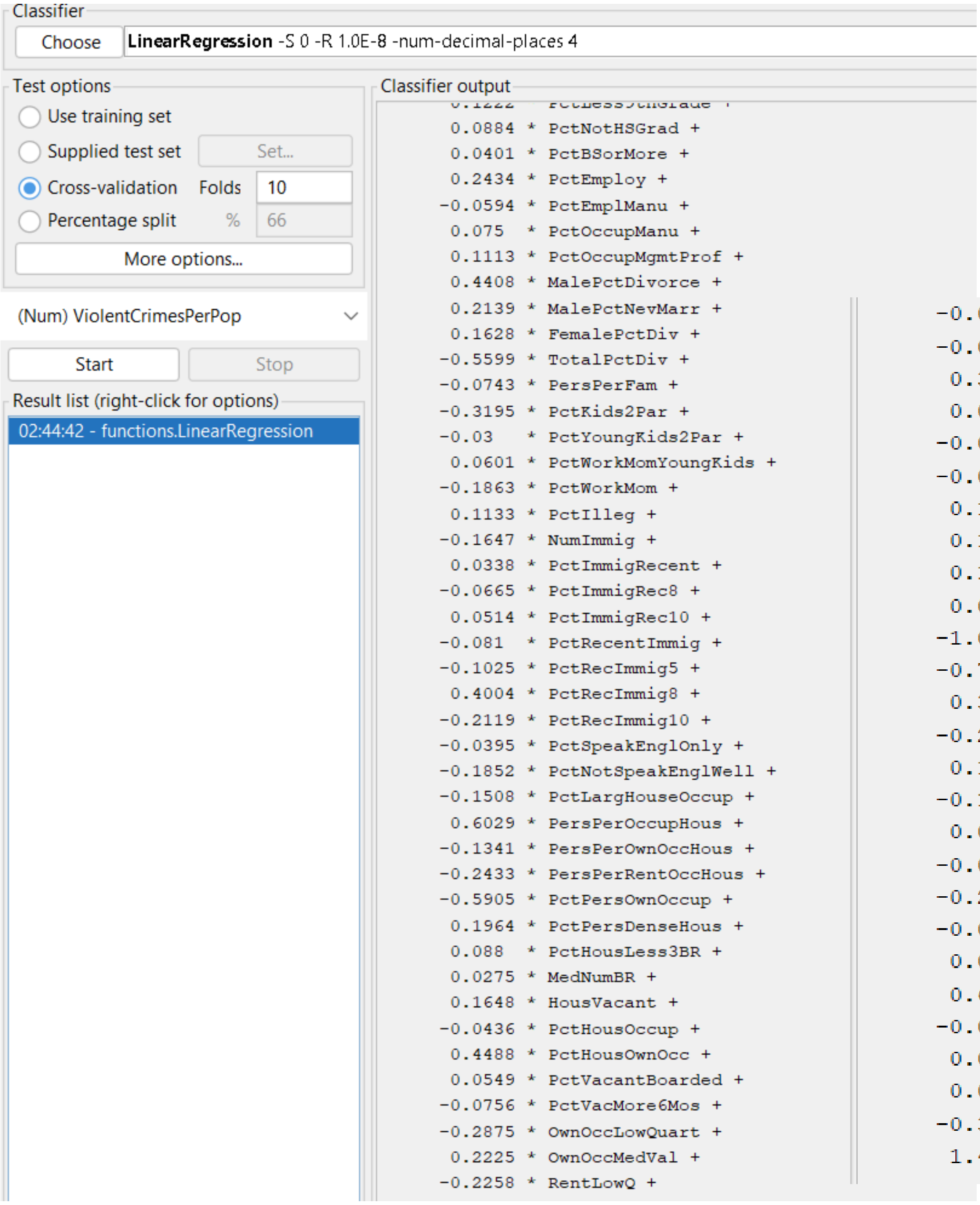
Replacing missing values with mean



non-predictive columns like state, county, community, and fold removed.

Linear Regression based prediction:






```
-0.034  * RentMedian +
-0.0641 * RentHighQ +
 0.3741 * MedRent +
 0.0405 * MedRentPctHousInc +
-0.0493 * MedOwnCostPctInc +
-0.0805 * MedOwnCostPctIncNoMtg +
 0.1104 * NumInShelters +
 0.1713 * NumStreet +
 0.139  * PctForeignBorn +
 0.0313 * PctSameCity85 +
-1.084  * LemasSwornFT +
-0.7676 * LemasSwFTFieldOps +
 0.3413 * LemasSwFTFieldPerPop +
-0.2191 * LemasTotalReq +
 0.1434 * PolicReqPerOffic +
-0.1379 * PctPolicWhite +
 0.0721 * PctPolicAsian +
-0.0858 * PctPolicMinor +
-0.2664 * OfficAssgnDrugUnits +
-0.0472 * PctUsePubTrans +
 0.0944 * PolicCars +
 0.6459 * PolicOperBudg +
-0.0769 * LemasPctPolicOnPatr +
 0.0352 * LemasGangUnitDeploy +
 0.0379 * LemasPctOfficDrugUn +
-0.3146 * PolicBudgPerPop +
 1.4413
```

```
Time taken to build model: 0.31 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8009
Mean absolute error                      0.0983
Root mean squared error                  0.1397
Relative absolute error                 55.1269 %
Root relative squared error             59.9784 %
Total Number of Instances             1994
```

OUTPUT

Output Meaning:
The linear regression model assigns weights to each feature to predict crime rates. A negative weight means the feature helps reduce crime, while a positive one indicates an increase. For example, a higher percentage of officers on patrol or a larger police budget per capita is linked with lower crime. On the other hand, more drug unit officers or gang unit deployment slightly increased predicted crime. The model starts with a base value (intercept) and adjusts based on these inputs.

SOME KEY INSIGHTS FOUND USING LINEAR REGRESSION

1) *High Divorce Rates → Higher Crime*
+0.4408 = MalePctDivorce, **+0.2139 = MalePctNevMarr**, **+0.1628 = FemalePctDiv**
Divorce and lack of marriage in males are strongly associated with higher predicted crime. This suggests a potential link between family instability and social outcomes.

2) *Two-Parent Households → Lower Crime*
-0.3195 = PctKids2Par, -0.03 = PctYoungKids2Par
A higher percentage of children living with two parents correlates with lower crime, reinforcing the role of stable family environments.

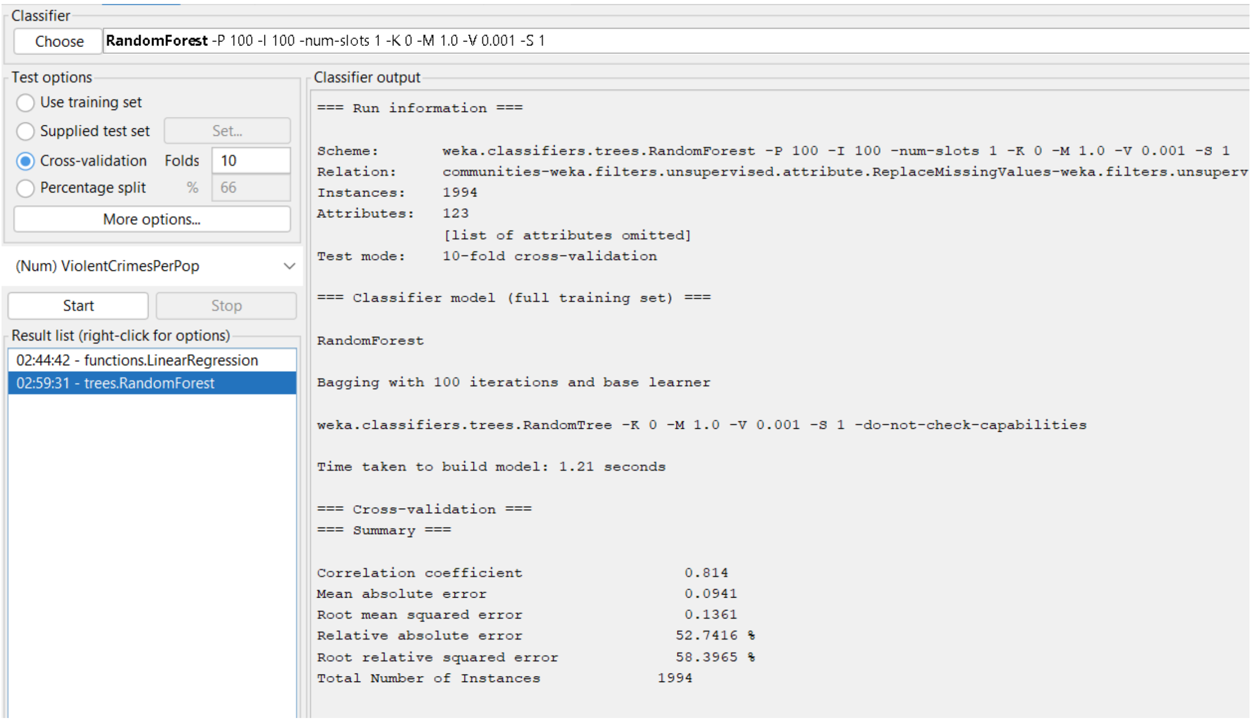
3) *More People per Household → Higher Crime*
+0.6029 = PersPerOccupHous
A high number of persons per occupied house is strongly linked to increased crime, which may reflect overcrowding and economic stress.

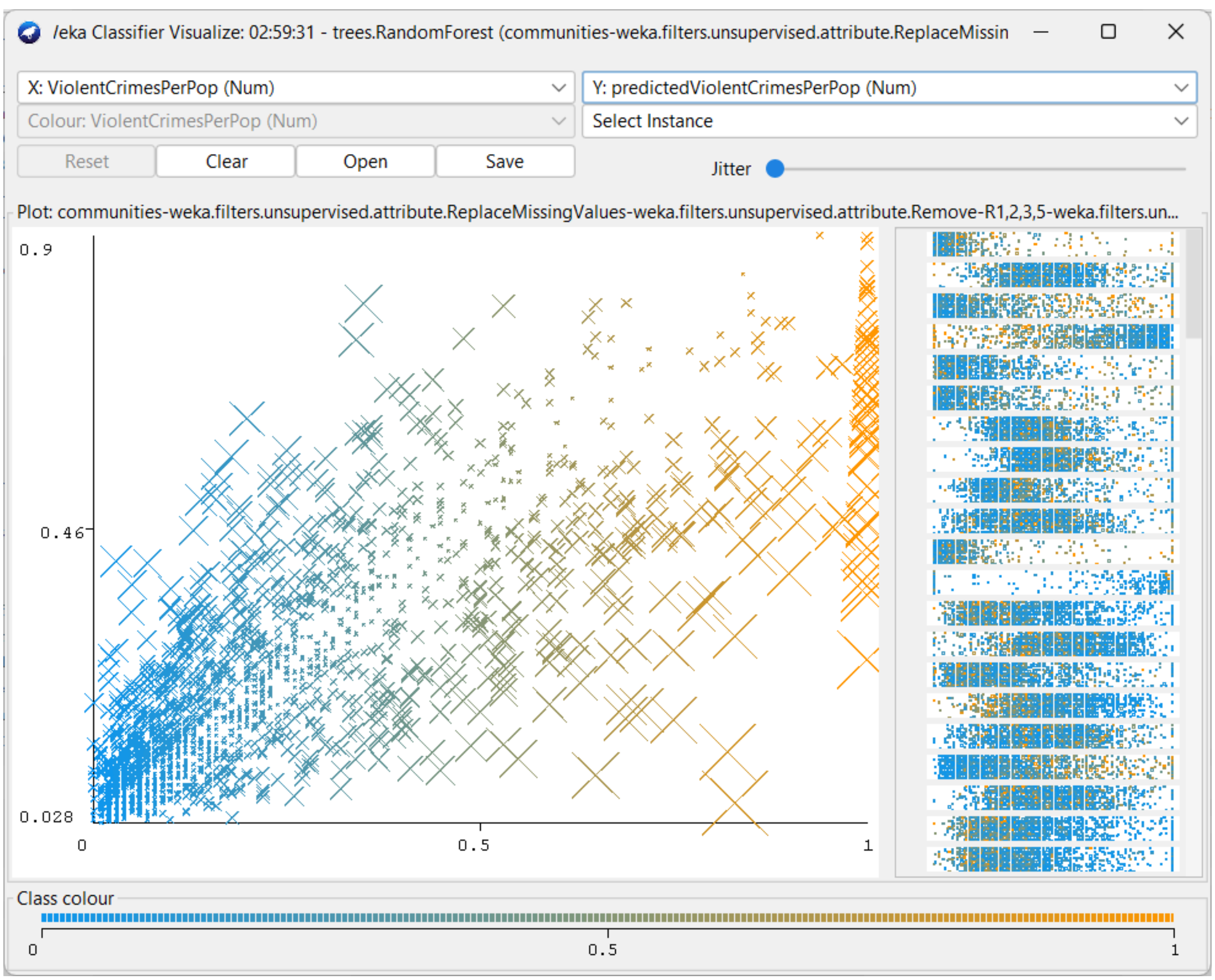
4) *Educational Attainment Shows Mixed Effects*
-0.1222 = PctLess9thGrade vs. +0.0884 = PctNotHSGrad and +0.0401 = PctBSorMore
The influence of education on crime is not linear—very low education correlates with less crime, which might reflect other underlying demographic traits, while some higher education levels slightly increase predicted crime, possibly due to collinearity or socioeconomic mixing.

5) *Higher White Per Capita Income → Lower Crime*
-0.392 = whitePerCap, -0.2137= medIncome
Higher median income and white per capita income show a strong negative association with crime, indicating that wealthier communities tend to be safer.

Metric	Meaning
Correlation coefficient (0.8009)	Strong correlation between predicted and actual values. Values close to 1 mean better fit.
Mean Absolute Error (0.0983)	On average, predictions are off by about 0.098 (in the scale of your target variable).
Root Mean Squared Error (0.1397)	Penalizes larger errors more. Lower is better.
Relative Absolute Error (55.13%)	MAE relative to just predicting the average. Lower % = better.
Root Relative Squared Error (59.98%)	RMSE relative to baseline. Below 60% = decent model.
Instances: 1994	Total records evaluated.

WEKA
(RANDOM FOREST)





```
Plot : weka.classifiers.trees.RandomForest (communities-weka.filters.unsupervi
Instance: 1267
                 population : 0.34
              householdsize : 0.43
               racepctblack : 0.86
               racePctWhite : 0.3
               racePctAsian : 0.1
                racePctHisp : 0.03
                agePct12t21 : 0.6
                agePct12t29 : 0.64
                agePct16t24 : 0.54
                 agePct65up : 0.36
                  numbUrban : 0.35
                   pctUrban : 1.0
                  medIncome : 0.17
                   pctWWage : 0.5
               pctWFarmSelf : 0.22
                 pctWInvInc : 0.36
                 pctWSocSec : 0.39
                pctWPubAsst : 0.47
                 pctWRetire : 0.45
                  medFamInc : 0.22
                  perCapInc : 0.25
                whitePerCap : 0.41
                blackPerCap : 0.16
               indianPerCap : 0.31
                AsianPerCap : 0.21
                OtherPerCap : 0.27
                 HispPerCap : 0.45
                NumUnderPov : 0.5
             PctPopUnderPov : 0.7
            PctLess9thGrade : 0.3
               PctNotHSGrad : 0.39
                PctBSorMore : 0.45
              PctUnemployed : 0.62
                  PctEmploy : 0.38
```

High Crime Correlations:

PctUnemployed (0.62): Unemployment is a strong predictor of crime. High unemployment (62%) is typically linked to higher crime rates due to social instability, lack of economic opportunities, and frustration among individuals who may resort to criminal activity as a means of survival.

PctPopUnderPov (0.7): 70% of the population living below the poverty line is another significant predictor. Poverty often correlates with crime, especially violent crime, as communities in poverty tend to face higher levels of desperation, social unrest, and lack of access to legal resources.

MedIncome (0.17) and MedFamInc (0.22): The low median income (both individual and family) indicates an economically disadvantaged community, which can lead to increased crime rates. Economic strain often fuels crime, particularly violent crimes.

PctBSorMore (0.45): While there is a significant portion of the population with a bachelor's degree or more (45%), the education disparity might contribute to unequal opportunities. Lack of access to quality education and well-paying jobs could drive some individuals to commit crimes.

PctUrban (1.0): Fully urbanized communities often have higher crime rates due to population density, anonymity, and greater opportunities for criminal activity. The urban environment might increase exposure to criminal elements or violence.

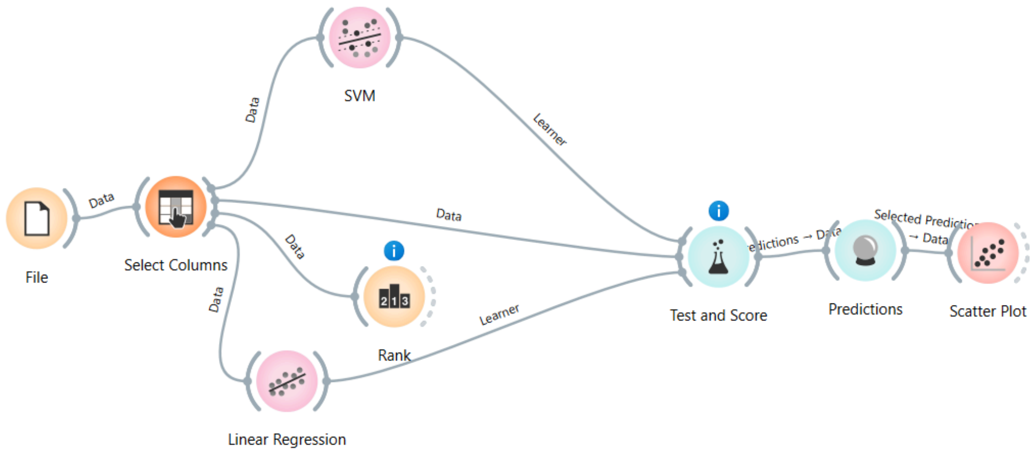
Specific Crime Predictions Based on Attributes:
Race Demographics:

With RacePctBlack (0.86) being 86%, the racial composition of the community may also correlate with crime patterns. Certain neighborhoods with racial homogeneity, especially if they face systemic challenges, can see higher crime rates due to historical and social factors.

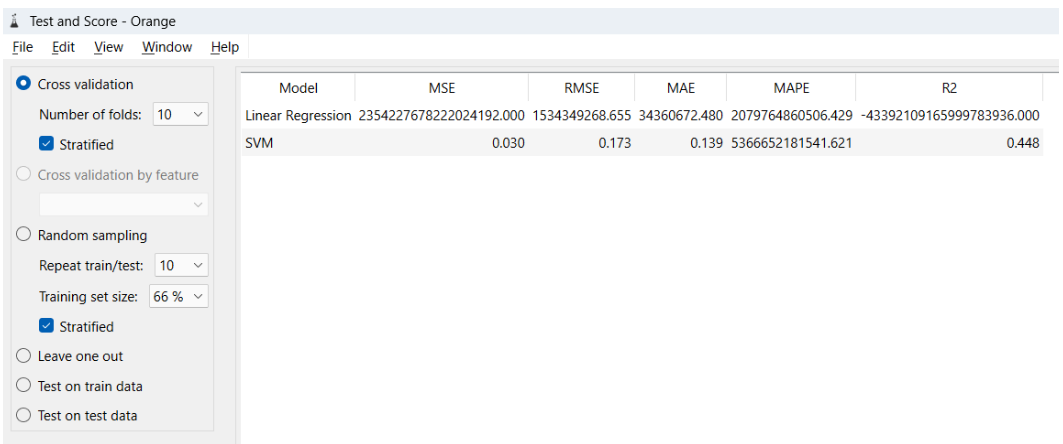
Age Distribution:
A young population (60% between 12-21) is a typical age range where crime rates, especially violent crime, tend to be higher. Young individuals are more likely to be involved in criminal activity, either as perpetrators or victims.

High Divorce Rate (MalePctDivorce: 54%): High divorce rates can contribute to family instability, which is often a contributing factor to violent crime in communities. Broken families can lead to a lack of supervision and support for young people, increasing the likelihood of delinquency.

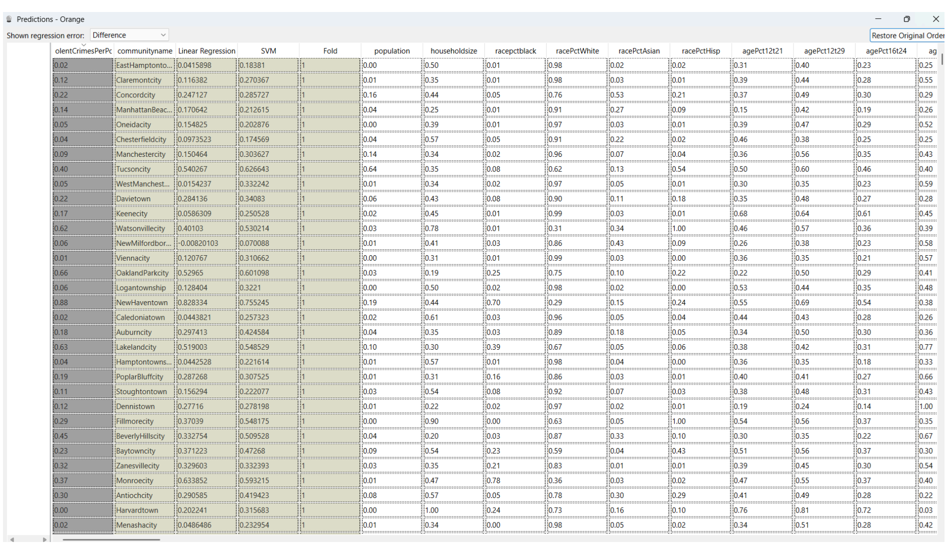
Orange



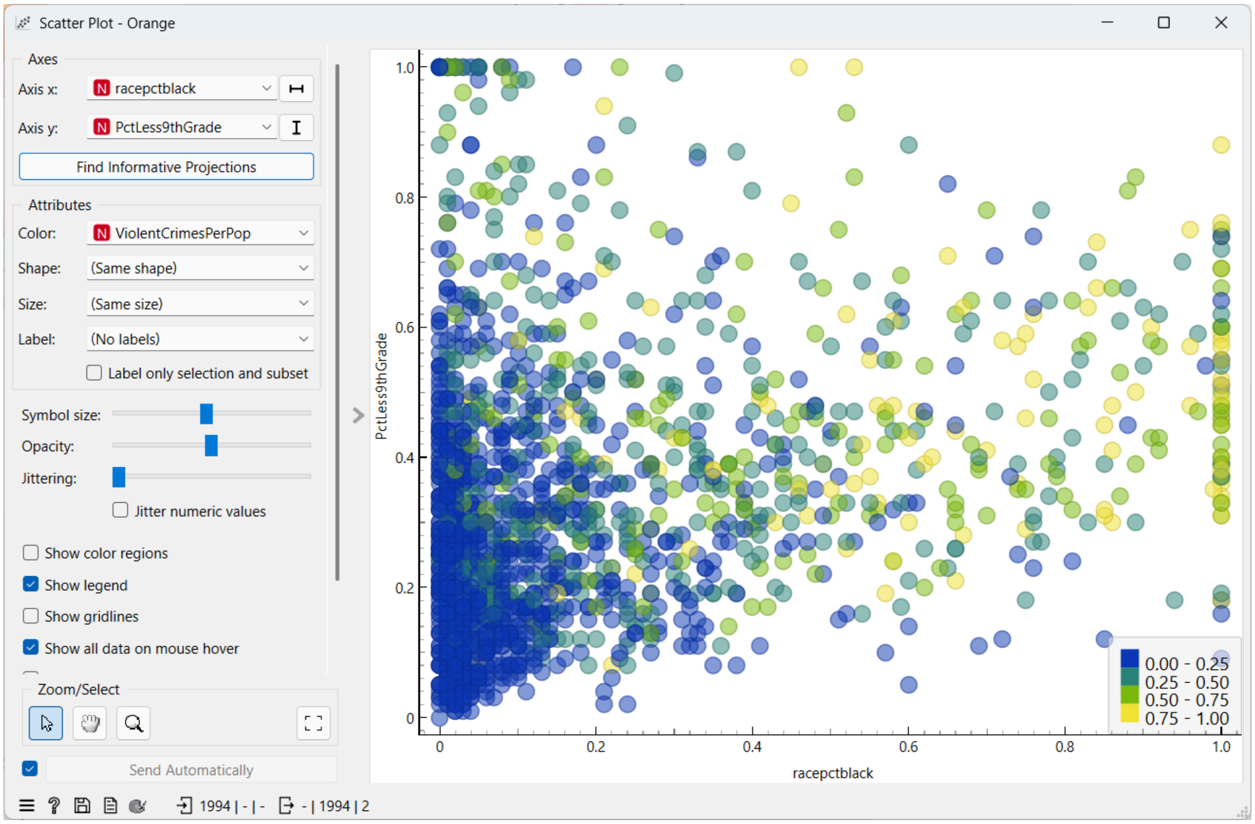
Setting Up workflow and adding my dataset



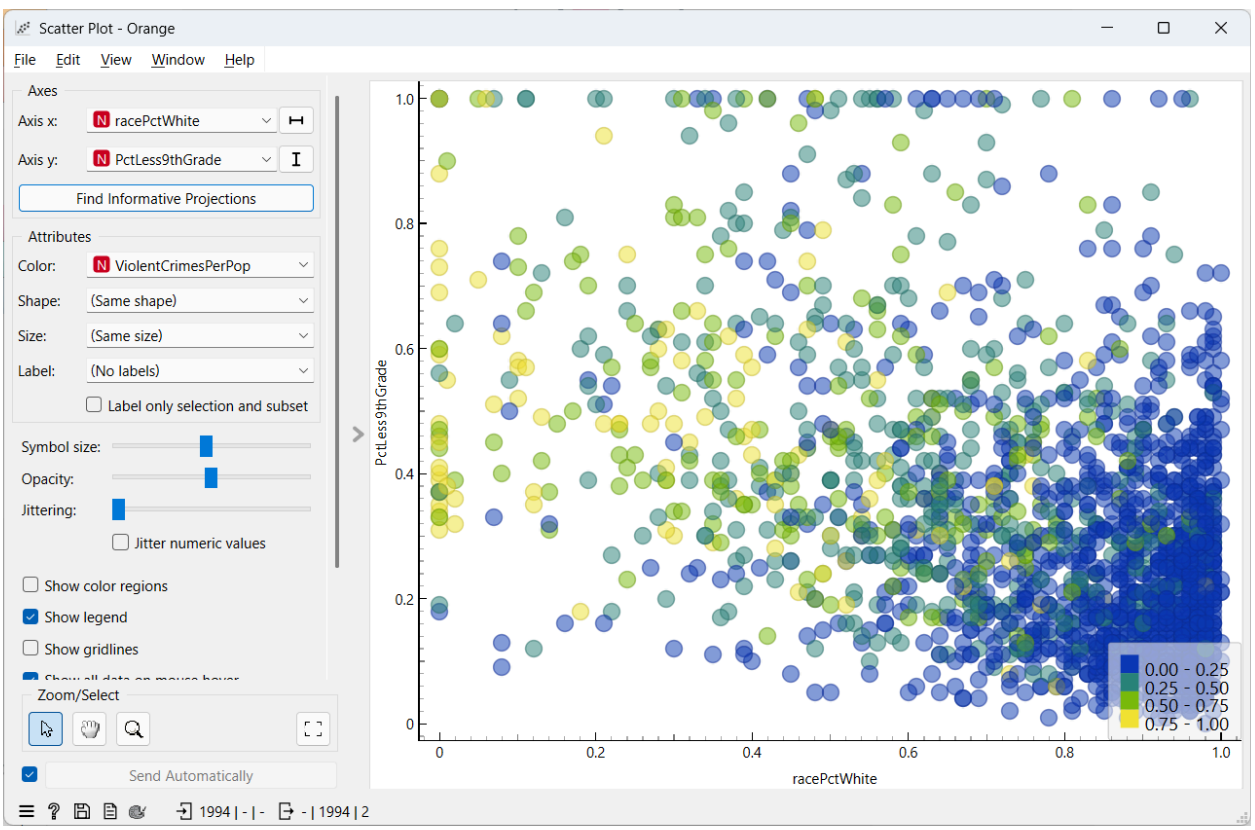
Comparing the metrics across both models



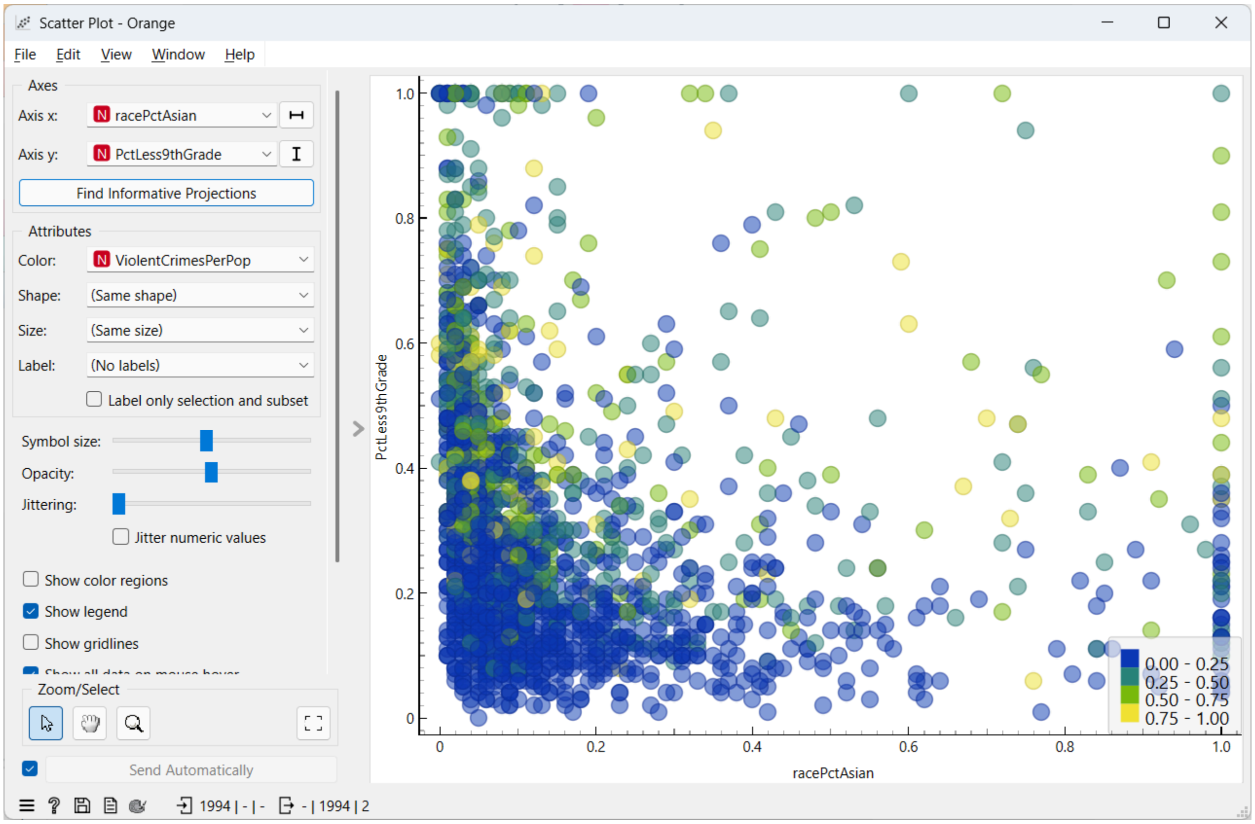
1st 4 columns showing actual crime vs crime predicted by both models in various communities



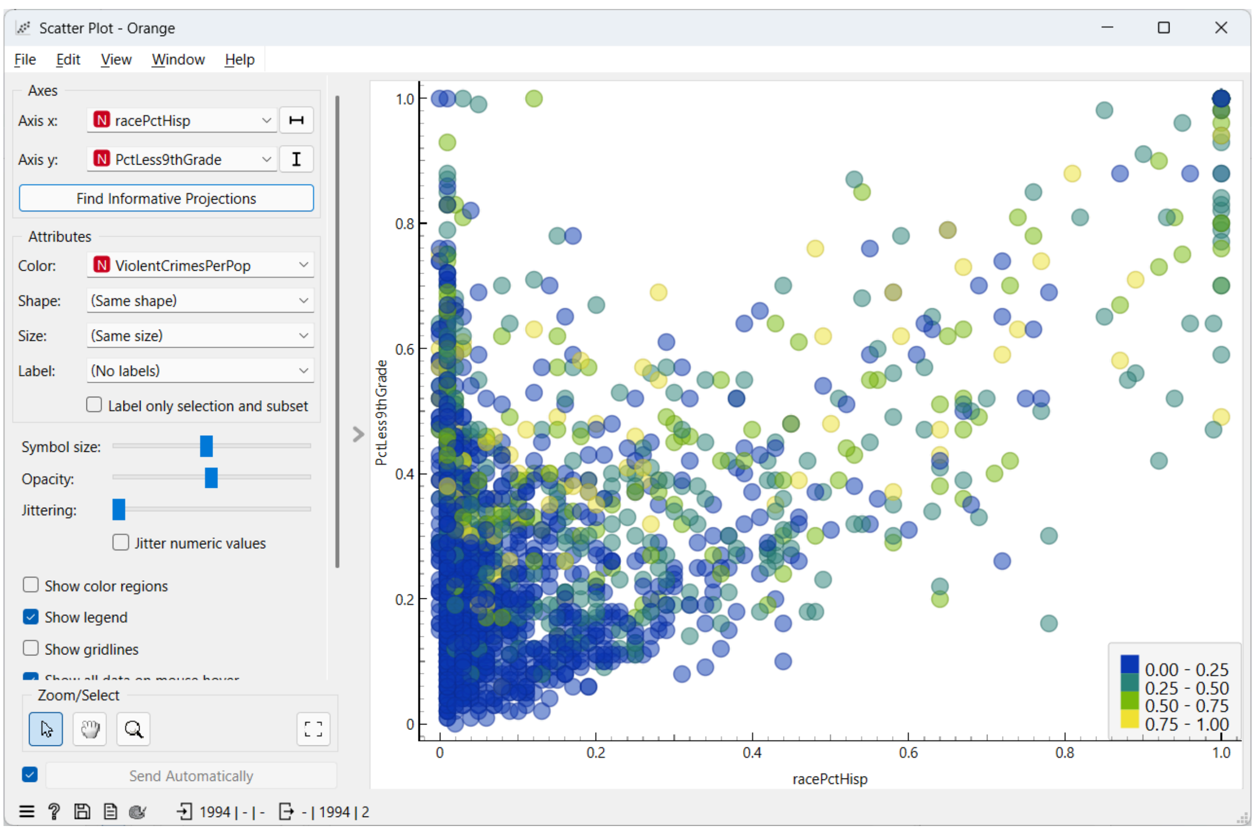
Race="Black" and Education less than 10th grade across Violent crime per population as color



Race="White" and Education less than 10th grade across Violent crime per population as color

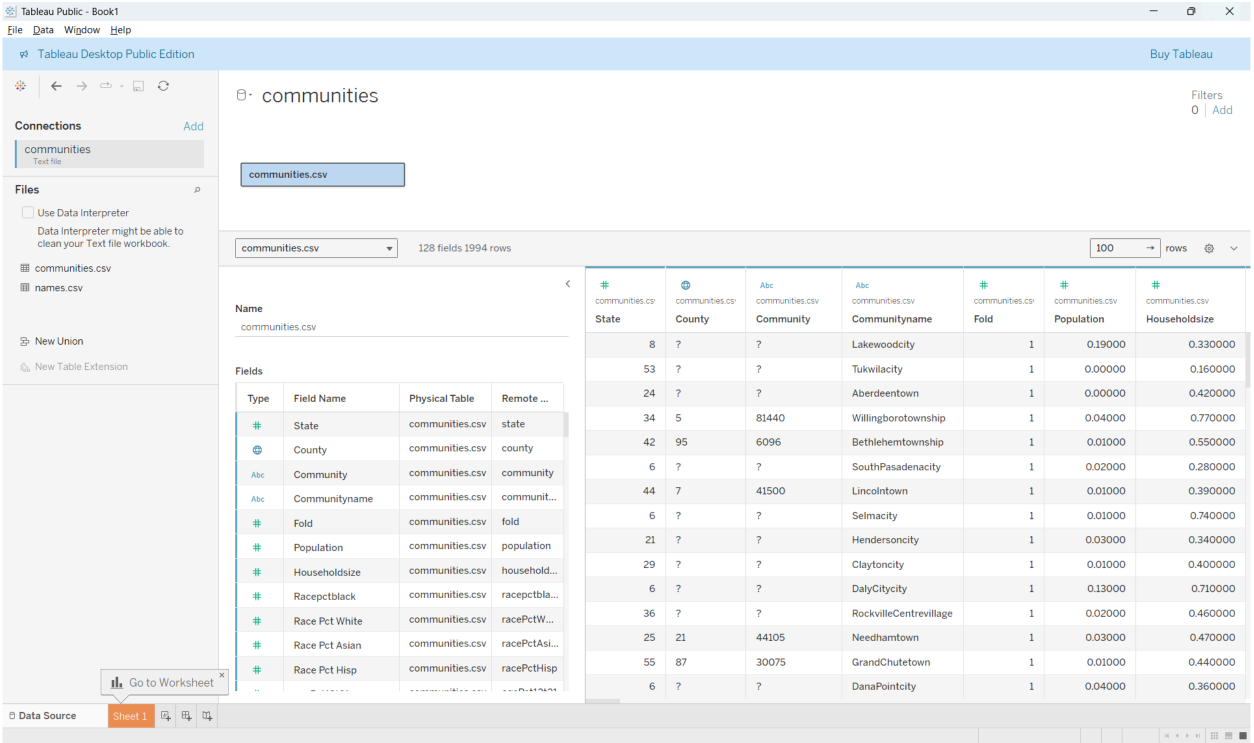


Race=”Asian” and Education less than 10th grade across Violent crime per population as color

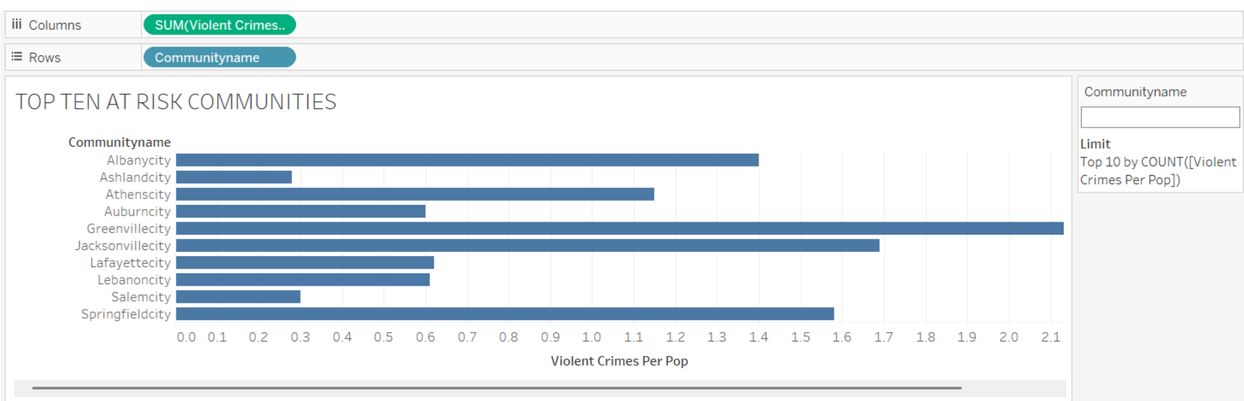


Race=”Hispanic” and Education less than 10th grade across Violent crime per population as color

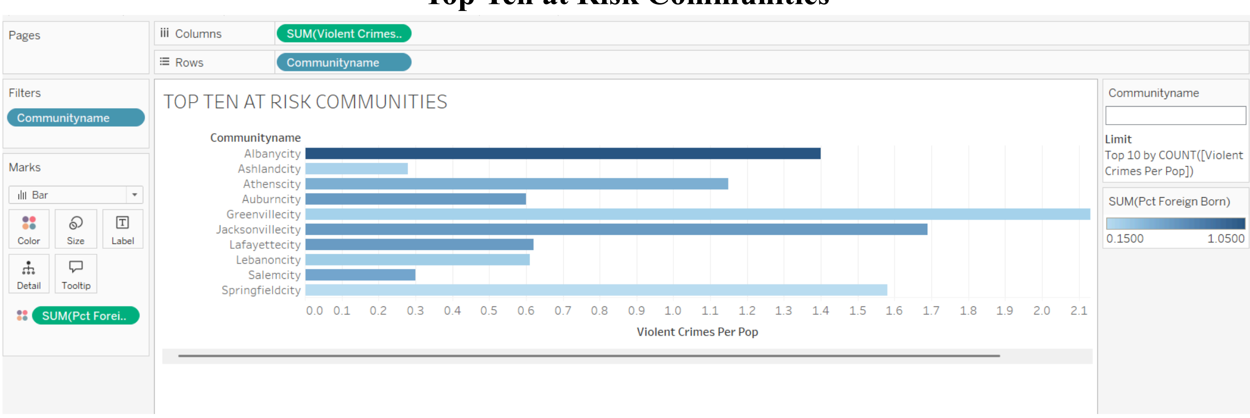
Tableau



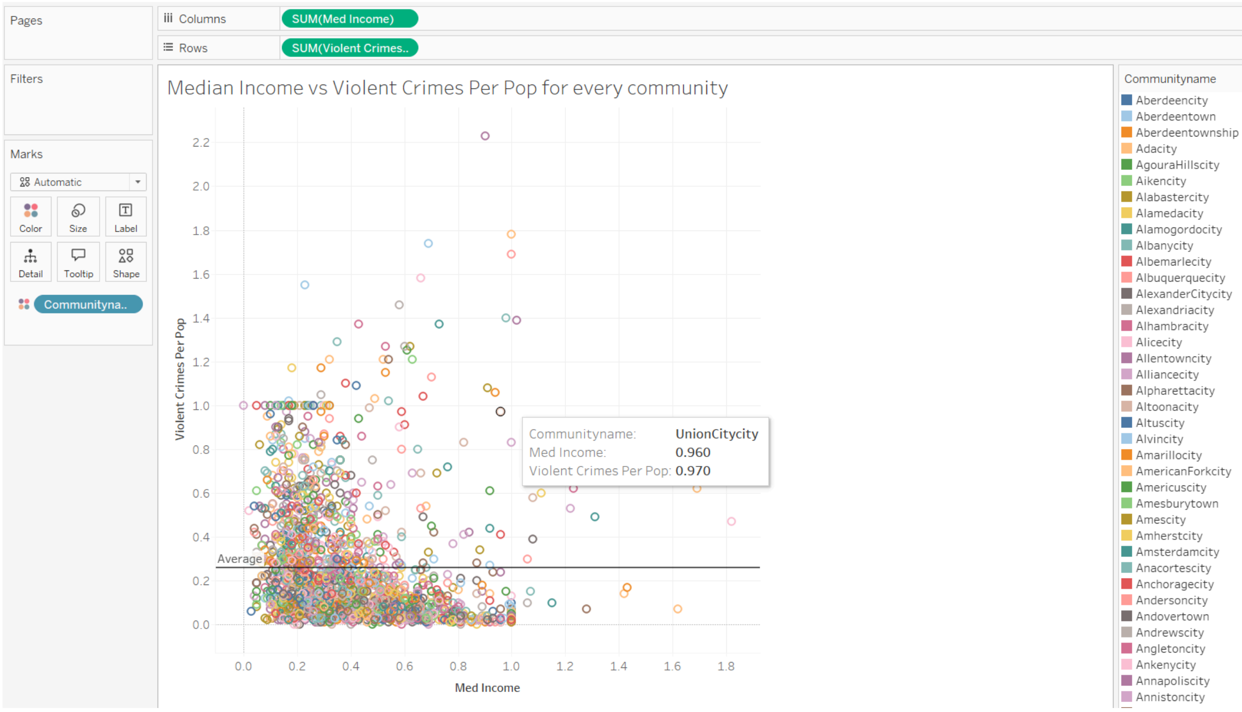
Loading my dataset in Tableau



Top Ten at Risk Communities



Top Ten at Risk Communities with crime committed by foreign born individuals



Median Income vs Violent Crimes Per Pop for every community

f. BI Decision
Resource Allocation:
Prioritize policing and social programs in communities with:

Poverty levels > 50%

Unemployment rates > 60%

Youth population (age 12–21) > 50%

Policy Recommendation:
Invest in youth-targeted job training and employment programs to address immediate unemployment and prevent youth involvement in violent crime.
Complement with family support initiatives and community-based engagement centers to strengthen social structure and reduce long-term crime risk.