Machine Learning 1 Project Report: Recidivism Predictive Model

Team Members:

James Beck, Jia Mody

10/23/24

	Table of Contents
Part 1 – Statement/Project Goal	3
Part 2 – Description of Dataset	3
Part 3 – Pre-Processing	9
Part 4 – Attribute Selection Algorithms & Model Classifiers Used	16
Part 5 – Results and Analysis	29
Part 6 – Conclusion	23

Part 1 – Statement/Project Goal

According to the Bureau of Justice Statistics, approximately 68% of convicted criminals commit a second crime within three years of release from jail. This trend highlights the necessity for strategies to address recidivism as repeat offenders not only overburden the legal system but also pose security risks to communities. In this project, our goal is to develop a comprehensive risk assessment model that identifies characteristics and behavioral patterns of criminals at high risk of reoffending. Specifically we looked at socio-economic, psychological, and environmental factors.

Leveraging data-driven insights from this study, legal institutions would have the tools necessary to identify potential repeat offenders and allocate targeted resources for their rehabilitation. This would invariably enhance public safety while ensuring high risk offenders receive adequate support throughout rehabilitation. This model could also provide valuable insights for policymakers, allowing them to design more effective parole conditions and post-release programs.

Part 2 – Description of Dataset

Our dataset was published on the "Data.gov" website by the U.S. Department of Justice. The dataset includes a total of 25,835 instances and 49 attributes excluding the class attributes. There are four potential class attributes for our model: "Recidvism_Arrest_Year1," "Recidvism_Arrest_Year2," "Recidvism_Arrest_Year3," and "Recidivism_Within_3years". Because a class variable that combines the information from each of these variables would make the most sense for our project, we will combine these four variables into a qualitative variable consisting of the four labels "1", "2", "3", and "Never". The label "Never" would correspond to

a criminal who was not arrested within 3 years after release from jail, "1" would correspond to a criminal arrested within one year of release, "2" a criminal arrested within two years, and so on. 10931 convicts did not commit a second crime within three years, 7724 committed a second crime within one year, 4567 within two, and 2613 within three. Therefore, convicts who commit a second crime appear more likely to do so soon after release from jail, creating a right skew. Below is an exhaustive list of the attributes:

	Attribute	Description	# Missing
			Values
1	ID	Number of person (1-25,825)	0
2	Gender	Male or Female	0
3	Race	Black or White	0
4	Age_at_Release	Age when released from jail	0
5	Residence_PUMA	Public Use Microdata Area: number	0
		correspond to a place of residence	
6	Gang_Affiliated	Is the person affiliated with a gang?	3,167
7	Supervision_Risk_	Supervision risk level in number format. Base	475
	Score_First	on severity of first crime committed	

8	Supervision_Level_ First	Supervision risk level in text format. Base on severity of first crime committed	1,720
9	Education_Level	Highest level of education	0
10	Dependents	Number of people dependent on inmate's income	0
11	Prison_Offense	Type of offense	3, 277
12	Prison_Years	Years in prison	0
13	Prior_Arrest_Episo des_Felony	Number of prior arrests	0
14	Prior_Arrest_Episo des_Misd	Number of arrests due to misdemeanors	0
15	Prior_Arrest_Episo des_Violent	Number of arrests due to violence	0
16	Prior_Arrest_Episo des_Property	Number of arrests due to property damages	0
17	Prior_Arrest_Episo des_Drug	Number of arrests due to drug usage	0

18	Prior_Arrest_Episo	Number of arrests due to protective order	0
	des_PPViolationCh	violations?	
	arges		
19	Prior_Arrest_Episo	Have there been arrests due to domestic	0
	des_DVCharges	violence?	
20	Prior_Arrest_Episo	Have there been arrests due to gun charges	0
	des_GunCharges		
21	Prior_Conviction_E	Number of convictions due to felony	0
	pisodes_Felony		
22	Prior_Conviction_E	Number of convictions due to misdemeanors	0
	pisodes_Misd		
23	Prior_Conviction_E	Number of convictions due to violence	0
	pisodes_Viol		
24	Prior_Conviction_E	Number of convictions due to property	0
	pisodes_Prop	damages	
25	Prior_Conviction_E	Number of convictions due to drug usage	0
	pisodes_Drug		

26	Prior_Conviction_E	Have there been convictions due to protective	0
	pisodes_PPViolatio	order violations?	
	nCharges		
27	Prior_Conviction_E	Have there been convictions due to domestic	0
	pisodes_DomesticV	violence?	
	iolenceCharges		
28	Prior_Conviction_E	Have there been convictions due to gun	0
	pisodes_GunCharge	charges?	
	S		
29	Prior_Revocations_	Have there been previous violations of parole	0
	Parole	terms?	
30	Prior_Revocations_	Have there been previous violations of	0
	Probation	probation terms?	
31	Condition_MH_SA	Has there been substance abuse/bad mental	0
		health?	
32	Condition_Cog_Ed	Is there a cognitive or educational condition?	0
33	Condition_Other	Are there any other conditions?	0

34	Violations_Electron	Have there been breaches of electronic	0
	icMonitoring	monitoring rules?	
25	X7. 1 I	TT 4 1 1 1 C: 4 C: 10	0
35	Violations_Instructi	Have there been breaches of instruction rules?	0
	on		
36	Violations_FailToR	Have there been failures to report as required?	0
	eport		
37	Violations_MoveWi	Has the inmate moved without permission?	0
	thoutPermission		
38	Delinquency_Repor	Number of delinquency reports	0
	ts		
39	Program_Attendanc	Number of programs attended	0
	es		
40	Program_Unexcuse	Number of unexcused absences from program	0
	dAbsences		
41	Residence_Changes	Number of place of residence changes	0
42	Avg_Days_per_Dru	Days between drug tests, on average	6,103
	gTest		

43	DrugTests_THC_P	Number of drug tests that testedTHC positive	5,172
	ositive		
44	DrugTests_Cocaine	Number of drug tests that tested cocaine	5,172
	_Positive	positive	
45	DrugTests_Meth_P	Number of drug tests that tested meth positive	5,172
	ositive		
46	DrugTests_Other_P	Number of drug tests that tested positive for	5,172
	ositive	other drugs	
47	Percent_Days_Emp	Percentage of 365/366 days employed	462
	loyed		
48	Jobs_Per_Year	Number of jobs worked per year	808
49	Employment_Exem	Is the inmate exempt from employment?	0
	pt		

Figure 1. Data Table

Part 3 – Pre-Processing

Pre-Processing was done using WEKA 3.8.6 and python scripts in Jupyter Notebook and VS Code.

Part 3.1 – Replace Missing Values

First, we began by filling in missing values. We did not remove any attributes because they all had <70% missing values. Because our quantitative attributes were heavily skewed, we used the attribute's median to fill in the missing values rather than the mean. This is because medians are more robust against outliers than means. Additionally, most of our quantitative attributes were discrete whole numbers, so it did not make sense to fill in missing values with decimal means. However, by default, WEKA fills in missing values with means for quantitative attributes. Furthermore WEKA lacks that ability to fill in missing values with median. Therefore, in order to fill in missing values with the median, we developed a Python script to compute the median for each quantitative attribute in our dataset and replace the missing values accordingly. This script ensures that the filled in missing values make sense with the distributions of their respective attributes.

The python script we used is shown in *Figure 2* below.

```
import csv

dataset = []

#Filling in missing values:

with open("NIJ_s_Recidivism_Challenge_Full_Dataset.csv", mode='r')as file:
    fileReader = csv.reader(file)

dct = {6:[], 41:[], 42:[], 43:[], 44:[], 45:[], 46:[], 47:[]}

for i, line in enumerate(fileReader):
    dataset.append(line)
    if i == 0: continue
    for key in dct:
```

```
if line[key] == "": continue
    dct[key].append(float(line[key]))
medianDct = {}

for key in dct:
    dct[key].sort()
    medianDct[key] = dct[key][len(dct[key])//2]

for rowNum, row in enumerate(dataset):
    for col, val in enumerate(row):
    if val == "" and col in medianDct:
        dataset[rowNum][col] = medianDct[col]
```

Figure 2. Python Script Demonstrating Missing Data Replacement

We then used the "ReplaceMissingValues" filter to fill in missing values for qualitative variables with the mode. All missing values being filled, we re-uploaded the dataset to WEKA for further analysis.

Part 3.2 – Create Derived Class

In our original dataset, there were four potential class attributes:

"Recidivism_Arrest_Year1," "Recidivism_Arrest_Year2," "Recidivism_Arrest_Year3," and

"Recidivism_Within_3Years." Each of these attributes represent specific arrest outcomes

following an offender's release. Because we cannot have multiple class variables, we decided to create a single qualitative class attribute that captures the information from each of these four variables.

We combined these variables into a qualitative variable with the following four labels:

- "1" Arrested within 1 year of release
- "2" Arrested within 2 years of release
- "3" Arrested within 3 years of release
- "Never" Not arrest within 3 years of release

The python script we used is shown in *Figure 3*.

```
#Combining classes:
with open('CombinedClass.csv', mode='w', newline='') as file:
   training = dataset[0].pop()
   Year3 = dataset[0].pop()
   Year2 = dataset[0].pop()
   Year1 = dataset[0].pop()
   within3 = dataset[0].pop()
    dataset[0].append("Years_Until_Recidivism")
    for i in range(len(dataset)-1):
        training = dataset[i+1].pop()
        Year3 = dataset[i+1].pop()
        Year2 = dataset[i+1].pop()
        Year1 = dataset[i+1].pop()
        within3 = dataset[i+1].pop()
        combinedVal = "Never"
        if Year1 == "true": combinedVal = "1"
        if Year2 == "true": combinedVal = "2"
        if Year3 == "true": combinedVal = "3"
        dataset[i+1].append(combinedVal)
```

```
writer = csv.writer(file)
writer.writerows(dataset)
```

Figure 3. Python Script Demonstrating Derived Class Creation

Part 3.3 – Normalize Data

To ensure that no attribute was over/underrepresented in our classification model, we needed to normalize the data. For this step, we used WEKA's "Normalize" filter which scales all quantitative attributes into the range [0, 1].

However, one attribute named "Residence_PUMA" did not make sense to normalize as it contained residency IDs. Because WEKA's "Normalize" filter normalizes all quantitative variables, we had to convert "Residence_PUMA" to a nominal attribute. To do so, we used WEKA's "NumericToNominal," which successfully converted "Residence_PUMA" to a nominal attribute. This step would successfully exclude "Residence_PUMA" from the normalization process.

We then applied the "NumericToNominal" filter, successfully normalizing the remaining quantitative attributes.

Part 3.4 – Changing Necessary Values from Nominal to Numeric

While preprocessing, we discovered that several quantitative attributes were marked "nominal" because they included values of the format "# or more" (e.g., "5 or more"). For example, "Prior_Arrest_Episodes_Felony" contained values from 0 to 9 and then values of "10 or more". This caused WEKA to interpret the entire column as categorical. For simplicity, we

decided to replace these values with numbers (i.e. "5 or more" would be replaced with "5") in order to make these attributes quantitative in WEKA.

To do this, we imported the dataset into Google Sheets and manually edited the affected columns. Specifically, we removed the "or more" suffix from each instance, leaving only the numeric portion (i.e., converting "5 or more" to "5"). This ensured that the entire column was properly recognized as numeric. With these changes made, we re-exported the modified dataset and uploaded it to WEKA.

Part 3.5 – Split final dataset into training and test dataset

After preprocessing our data, we needed to split it into train, validation, and test sets. Our dataset contained 25,825 instances, so we decided to create a 70/15/15 split (70% for training, 15% for testing, and 15% for validating). Additionally, because our dataset is unbalanced, we needed to ensure that each section had a sample that was representative of the class distribution. To do this we utilized a stratified random sampling method within Google CoLab. After this, however, we realized that our code had errors when creating the validation set. Since the train dataset was stratified, the number of instances in it did not match the number of instances in our dataset. We then decided to stratify the validation set based on the training set. We thought that this would work, but upon further inspection we noticed that this method does not result in an exact 70/15/15 split, since the validation set is stratified for 15% based on the 85% training set. We then decided to stratify the validation set based on the testing set. For this, we had to change the first split to be 70/30. From there, we stratified 50% of the 30% for validation and the other half for train.

The script used is shown in *Figure 4* Below

```
import pandas as pd

df = pd.read_csv('/content/drive/MyDrive/ML/Normalized.csv')

from google.colab import drive

drive.mount('/content/drive')

from sklearn.model_selection import train_test_split

train, remaining = train_test_split(df, test_size=0.30, stratify=df.iloc[:, -1])

val, test = train_test_split(remaining, test_size=0.50, stratify=remaining.iloc[:, -1])

train.to_csv('train.csv', index=False)

!cp train.csv /content/drive/MyDrive/ML

val.to_csv('val.csv', index=False)

!cp val.csv /content/drive/MyDrive/ML

test.to_csv('test.csv', index=False)

!cp test.csv /content/drive/MyDrive/ML
```

Figure 4: Python Script Demonstrating Train Test Validation Splits

Down the line we had issues utilizing the validation set in WEKA, so we decided to do a train-test split of 70/30 without validation. The updated script is seen in *Figure 5* below.

```
import pandas as pd

df = pd.read_csv('/content/drive/MyDrive/ML/Normalized.csv')

from google.colab import drive

drive.mount('/content/drive')

from sklearn.model_selection import train_test_split

train, test = train_test_split(df, test_size=0.30, stratify=df.iloc[:, -1])
```

train.to_csv('train.csv', index=False)

!cp train.csv /content/drive/MyDrive/ML

test.to csv('test.csv', index=False)

!cp test.csv /content/drive/MyDrive/ML

Figure 5: Python Script Demonstrating Train Test Splits

Part 4 – Attribute Selection Algorithms & Model Classifiers Used

Part 4.1 Attribute Selection Algorithms:

To select potentially useful attributes, we used the following 4 attribute selection algorithms in addition to self selection:

- CorrelationAttributeEval with cutoff 0.1
- InfoGainAttributeEval with cutoff 0.25
- OneRAttributeEval with cutoff 43.95
- WrapperSubsetEval

These 5 methods gave us the following attribute selections:

#	CorrelationAtt	InfoGainAttribut	OneRAttribute	WrapperSubse	Self Selection
	ributeEval	eEval	Eval	tEval	
1	Percent_Days_	Jobs_Per_Year	Jobs_Per_Year	Gang_Affiliat	DrugTests_Coc
	Employed			ed	aine_Positive

2	Prior_Arrest_E	Percent_Days_E	Percent_Days_	Prior_Arrest_	DrugTests_Met
	pisodes_PPVio	mployed	Employed	Episodes_PPV	h_Positive
	lationCharges			iolationCharge	
				S	
3	Prior_Arrest_E	Prior_Arrest_Ep	Gang_Affiliate	Prior_Convicti	Gang_Affiliate
	pisodes_Felon	isodes_PPViolati	d	on_Episodes_	d
	у	onCharges		PPViolationCh	
				arges	
4	Gang_Affiliate	Prior_Arrest_Ep	Prior_Arrest_E	Violations_Fai	Prison_Years
	d	isodes_Felony	pisodes_PPVio	lToReport	
			lationCharges		
5	Prior_Arrest_E	Gang_Affiliated	DrugTests_TH	Deliquency_R	Condition_Cog
	pisodes_Prope		C_Positive	eports	_Ed
	rty				
6	Supervision_R	Supervision_Ris	Prior_Arrest_E	Percent_Days	Education_Lev
	isk_Score_Firs	k_Score_First	pisodes_Proper	_Employed	el
	t		ty		
7	Prior_Arrest_E	DrugTests_THC	Prior_Arrest_E	Jobs_Per_Year	Dependent
	pisodes_Misd	_Positive	pisodes_Felon		
			у		

8	Prior_Convicti	Prior_Arrest_Ep	Age_at_Releas	Violations_Inst
	on_Episodes_	isodes_Property	e	ruction
	Misd			
9	Prior_Convicti	Age_at_Release	Prior Convicti	Percent Days
	Thoi_convicu	rige_at_Release	Thoi_convicti	1 creent_Days_
	on_Episodes_		on_Episodes_P	Employed
	Prop		rop	
10			Supervision_Ri	
			sk_Score_First	

Figure 6: Table Demonstrating Attribute Selection tests and their Results

Part 4.2 Classifier Models

To construct our classification models, on each combination of attributes selected by our attribute selection algorithms, we used the following 4 classification algorithms:

- J48
- NaiveBayes
- OneR
- RandomForest

In WEKA, we created and tested all 20 models using K-Fold cross validation with 10 folds. The outputs for these models is shown below:

J48 – CorrelationAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
Correctly Classified Instances
                                    8663
                                                       47.9042 %
Incorrectly Classified Instances
                                     0.2113
Mean absolute error
                                      0.3022
Root mean squared error
Relative absolute error
                                       0.4421
                                      87.5774 %
Root relative squared error
Total Number of Instances
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                        0.050 0.122
                                                    0.082
                0.061
                                            0.061
                                                               0.016
                                                                       0.517
                                                                                 0.110
                0.682
                        0.378
                                 0.570
                                            0.682
                                                    0.621
                                                               0.301
                                                                       0.667
                                                                                 0.546
                                                                                           Never
                0.539
                        0.258
                                 0.471
                                            0.539
                                                    0.503
                                                               0.271
                                                                       0.643
                                                                                 0.409
                0.131
                        0.094
                                 0.230
                                           0.131
                                                    0.167
                                                            0.046
                                                                       0.522
                                                                                 0.197
Weighted Avg.
               0.479
                        0.259
=== Confusion Matrix ===
                 d <-- classified as
  112 936 585 196 | a = 3
320 5218 1514 599 | b = Never
  298 1587 2915 607 |
  189 1416 1174 418 | d = 2
                                                                                                                     Log
```

Figure 7: J48 CorrelationAttributeEval Correlation Result

Naive Bayes - CorrelationAttributeEval Attributes

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances
                                                   49.3641 %
                               9157
Incorrectly Classified Instances
                                                   50.6359 %
                                  0.2051
Kappa statistic
                                    0.2938
Mean absolute error
Root mean squared error
                                    0.411
Relative absolute error
                                   85.1333 %
Root relative squared error
                                    98.942 %
Total Number of Instances
=== Detailed Accuracy By Class ===
               TP Rate FP Rate Precision Recall F-Measure MCC
                                                                    ROC Area PRC Area Class
               0.000 0.000 ?
0.717 0.446 0.541
                                         0.000 ? ?
0.717 0.617 0.270
                                                                    0.568 0.121
                                                                    0.699
                                                                             0.632
                                                                                     Never
                              0.279
               0.603
                      0.305
                                                                    0.716
                                                                             0.489
               0.056
                      0.043
                                                                    0.568
Weighted Avg.
             0.494
                      0.287
                                                                    0.667
                                                                           0.463
=== Confusion Matrix ===
   a b c d <-- class
0 1045 679 105 | a = 3
                d <-- classified as
   0 5488 1790 373 | b = Never
   0 1980 3259 168 | c = 1
0 1625 1392 180 | d = 2
                                                                                                                      × 0
                                                                                                              Log
```

Figure 8: NaiveBayes CorrelationAttributeEval Correlation Result

OneR – CorrelationAttributeEval Attributes

```
Classifier output
  == Stratified cross-validation ===
=== Summary =
                                                         47.9982 %
Correctly Classified Instances
Incorrectly Classified Instances
                                      0.198
0.26
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
                                         0.5099
                                        75.3499 %
Root relative squared error
                                      122.7611 %
                                  18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall F-Measure MCC
                                                                           ROC Area PRC Area Class
                                                       0.009 -0.006 0.499 0.101
                 0.005 0.007 0.077 0.005
                 0.623
                         0.368
                                   0.554
                                              0.623
                                                       0.587
                                                                  0.252
                                                                           0.628
                                                                                     0.505
                                                                                                Never
                                                       0.536
                                              0.693
                                                                  0.286
                                                                           0.656
                 0.693
                                   0.437
                         0.382
                                                      0.536 0.286
0.078 0.013
0.423 0.194
                 0.048
                         0.041
                                   0.200
                                             0.048
                                                                           0.503
                                                                                      0.178
Weighted Avg. 0.480
                                                                          0.601
                        0.278
                                             0.480
                                   0.408
=== Confusion Matrix ===
                  d <-- classified as
  9 981 730 109 | a = 3
49 4769 2578 255 | b = Never
28 1379 3748 252 | c = 1
   31 1482 1530 154 | d = 2
```

Figure 9: OneR CorrelationAttributeEval Correlation Result

RandomForest – CorrelationAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary =
                                                        48.8996 %
Correctly Classified Instances
Incorrectly Classified Instances
                                     9241
                                                        51.1004 %
                                     0.218
Kappa statistic
Mean absolute error
                                       0.3035
Root mean squared error
Relative absolute error
                                      87.9643 %
Root relative squared error
                                      98.1205 %
                                    18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                         ROC Area PRC Area Class
                                                    0.055 0.010
0.623 0.299
                0.693
                        0.390
                                 0.566
                                             0.693
                                                                0.299
                                                                         0.710
                                                                                   0.650
                                                                                             Never
                0.578
                         0.272
                                  0.475
                                             0.578
                                                     0.522
                                                                0.291
                                                                         0.715
                                                                                   0.470
                                                    0.147 0.037
0.451 0.221
                0.110
                        0.083
                                  0.222
                                            0.110
                                                                         0.561
                                                                                   0.204
               0.489
=== Confusion Matrix ===
                  d <-- classified as
 65 972 620 172 | a = 3
218 5300 1578 555 | b = Never
 167 1603 3125 512 |
   98 1495 1251 353 |
                                                                                                                               × 0
```

Figure 10: RandomForest CorrelationAttributeEval Correlation Result

J48 – InfoGainAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation =
=== Summary ===
Correctly Classified Instances
                                                             51.1391 %
Incorrectly Classified Instances
                                        8836
                                                             48.8609 %
                                          0.2666
Kappa statistic
Mean absolute error
                                           0.2814
Root mean squared error
Relative absolute error
Root relative squared error
                                          81.5542 %
                                          103.1964 %
Total Number of Instances
                                       18084
 === Detailed Accuracy By Class ===
                  TP Rate FP Rate Precision Recall F-Measure MCC
                                                                                 ROC Area PRC Area Class
                  0.072
                                     0.131
                                                           0.093
                            0.053
                                                 0.072
                                                                       0.024
                                                                                 0.540
                                                                                            0.119
                                                                                 0.705
                  0.671
                            0.317
                                     0.608
                                                 0.671
                                                           0.638
                                                                       0.351
                                                                                            0.607
                                                                                                       Never
                  0.146
                            0.090
                                     0.258
                                                0.146
                                                           0.186
                                                                      0.071
                                                                                 0.560
                                                                                            0 216
                                              0.511
                                                                    0.273
Weighted Avg. 0.511
                           0.234
                                                           0.484
                                                                                 0.669
 === Confusion Matrix ===
                    d <-- classified as
  a b c d <-- classific

131 894 582 222 | a = 3

882 5134 1503 632 | b = Never

242 1159 3517 489 | c = 1

243 1253 1235 466 | d = 2
                                                                                                                                            × 0
```

Figure 11: J48 InfoGainAttributeEval Correlation Result

Naive Bayes – InfoGainAttributeEval Attributes

```
Classifier output
  == Stratified cross-validation ===
=== Summarv ===
Correctly Classified Instances
                                                          51.3769 %
Incorrectly Classified Instances
                                       8793
                                                           48.6231 %
Kappa statistic
Mean absolute error
Root mean squared error
                                         0.4019
Relative absolute error
                                        83.8954 %
Root relative squared error
Total Number of Instances
                                     18084
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall F-Measure MCC
                                                                             ROC Area PRC Area Class
                                                                                     0.121
                 0.000
                         0.000 ?
0.486 0.544
                                              0.000
                                                                             0.571
                                                       0.645 0.311
                 0.792
                                               0.792
                                                                             0.730
                                                                                        0.665
                                                       0.562
                          0.248
                                    0.491
                                               0.562
                                                                             0.733
                                                                                        0.503
                                   0.254
                                            0.062
0.514
                 0.062
                          0.039
                                                                             0.586
                                                                                        0.221
=== Confusion Matrix ===
   a b c d <-- classified as
0 1170 570 89 | a = 3
0 6057 1363 231 | b = Never
0 2112 3037 258 | c = 1
0 1786 1214 197 | d = 2
                                                                                                                                      × 0
```

Figure 12: NaiveBayes InfoGainAttributeEval Correlation Result

One R – InfoGainAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances
                                                         47.9982 %
Incorrectly Classified Instances
                                                         52.0018 %
                                      0.198
0.26
Kappa statistic
Mean absolute error
Root mean squared error
                                        0.5099
                                        75.3499 %
Relative absolute error
Root relative squared error
                                      122.7611 %
                                    18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall F-Measure MCC
                                                                  ROC Area PRC Area Class
                 0.005 0.007 0.077 0.005
                                                      0.009
                 0.623
                         0.368
                                   0.554
                                             0.623
                                                      0.587
                 0.693
                         0.382
                                   0.437
                                              0.693
                                                      0.536
                                                                  0.286
                                                                           0.656
                                                                                     0.394
                                           0.048
                                 0.200
                                                                 0.194
Weighted Avg. 0.480
                         0.278
                                                      0.423
                                                                          0.601
                                                                                    0.373
=== Confusion Matrix ===
                  d <-- classified as
  a b c d <-- classifi

9 981 730 109 | a = 3

49 4769 2578 255 | b = Never

28 1379 3748 252 | c = 1

31 1482 1530 154 | d = 2
                                                                                                                          Log
                                                                                                                                  ×0
```

Figure 12: OneR InfoGainAttributeEval Correlation Result

RandomForest – InfoGainAttributeEval Attributes

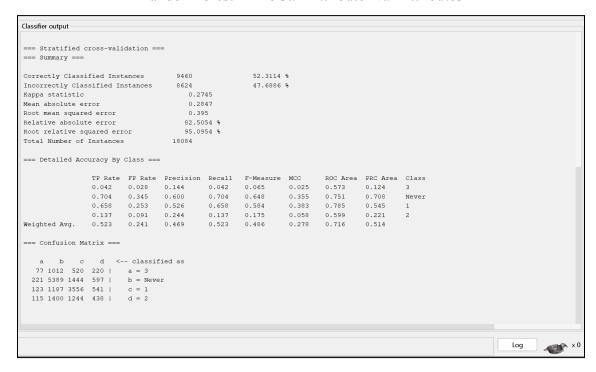


Figure 13: RandomForest InfoGainAttributeEval Correlation Result

J48 – OneRAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances
                                      9204
                                                         50.8958 %
                                                         49.1042 %
Incorrectly Classified Instances
                                      8880
                                       0.2642
Mean absolute error
                                         0.2814
                                         0.4314
Root mean squared error
Relative absolute error
Root relative squared error
                                      103.8502 %
                                     18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall
                                                       F-Measure MCC
                                                                           ROC Area PRC Area Class
                 0.084
                          0.058
                                   0.141
                                              0.084
                                                       0.105
                                                                  0.033
                                                                           0.538
                                                                                     0.119
                 0.671
                          0.319
                                   0.607
                                              0.671
                                                       0.637
                                                                  0.349
                                                                           0.700
                                                                                     0.600
                                                                                               Never
                                   0.519
                 0.640
                          0.253
                                              0.640
                                                       0.573
                                                                  0.367
                                                                           0.725
                                                                                     0.470
                 0.141
                          0.094
                                   0.244
                                                       0.179
                                                                  0.060
                                                                           0.559
                                                                                     0.212
Weighted Avg.
                0.509
                          0.233
                                  0.469
                                            0.509
                                                       0.483
                                                                 0.271
                                                                           0.666
                                                                                     0.444
=== Confusion Matrix ===
                   d <-- classified as
 154 887 557 231 | a = 3
418 5136 1439 658 | b = Never
252 1182 3462 511 | c = 1
  269 1259 1217 452 |
                                                                                                                                 × 0
                                                                                                                          Log
```

Figure 14: J48 OneRAttributeEval Correlation Result

Naive Bayes – OneRAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ==
Correctly Classified Instances
                                                       50 8958 %
Incorrectly Classified Instances
                                                        49.1042 %
                                    8880
Kappa statistic
Mean absolute error
                                       0.2881
Root mean squared error
                                       0.405
Relative absolute error
                                      83.4832 %
Root relative squared error
                                      97.5108 %
Total Number of Instances
                                    18084
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall F-Measure MCC
                                                                          ROC Area PRC Area Class
                                                                          0.567
                0.000
                         0.000
                                            0.000
                                                     ? ?
0.641 0.305
                                                                                   0.119
                         0.471
                                  0.546
                                             0.774
                                                                          0.725
                                                                                    0.661
                                                                                              Never
                                  0.486
                                                                 0.298
                                                     0.523 0.230
0.108 0.036
                0.070
                         0.049
                                  0.234
                                          0.070
0.509
                                                                          0.582
                                                                                   0.218
Weighted Avg.
               0.509
                        0.284
                                                                          0.685
                                                                                   0.480
=== Confusion Matrix ===
                  d <-- classified as
        b
             C
   0 1129 585 115 | a = 3
0 5925 1413 313 | b = Never
   0 2048 3055 304 | c = 1
0 1742 1231 224 | d = 2
                                                                                                                        Log
                                                                                                                                ×0
```

Figure 15: NaiveBayes OneRAttributeEval Correlation Result

OneR - OneR AttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
  == Summary ===
                                                            47.9982 %
Correctly Classified Instances
Incorrectly Classified Instances
                                        0.198
0.26
Kappa statistic
Mean absolute error
Root mean squared error
                                          0.5099
Relative absolute error
                                          75.3499 %
Root relative squared error
                                        122.7611 %
Total Number of Instances
                                      18084
=== Detailed Accuracy By Class ===
                  TP Rate FP Rate Precision Recall F-Measure MCC
                                    0.077
                                                                                         0.101
                                                0.005 0.009
                                                                     -0.006
                  0.005
                          0.007
                                                                               0.499
                                     0.554
                0.693 0.382 0.437
0.048 0.041 0.200
0.480 0.278 0.408
                                                0.693
                                                         0.536
                                                                     0.286
                                                                               0.656
                                                                                         0.394
                                            0.048
                                                                     0.013
                                                                  0.194 0.601
Weighted Avg.
                                                        0.423
=== Confusion Matrix ===
   a b c d <-- classific
9 981 730 109 | a = 3
49 4769 2578 255 | b = Never
   28 1379 3748 252 | c = 1
31 1482 1530 154 | d = 2
                                                                                                                                        ×0
                                                                                                                                Log
```

Figure 16: OneR OneRAttributeEval Correlation Result

RandomForest – OneRAttributeEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances
                                                       53.0248 %
Incorrectly Classified Instances
                                    8495
                                                       46.9752 %
                                     0.2835
Kappa statistic
                                       0.2859
Mean absolute error
Root mean squared error
                                       0.3928
Relative absolute error
                                      82.8537 %
Root relative squared error
                                      94.5644 %
Total Number of Instances
                                   18084
 === Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                         ROC Area PRC Area Class
                                                                         0.582
                0.037 0.025 0.144 0.037
                                                                                   0.126
                0.718
                        0.342
                                  0.607
                                             0.718
                                                     0.658
                                                                0.372
                                                                         0.754
                                                                                   0.710
                0.664
                        0.256
                                  0.525
                                             0.664
                                                     0.586
                                                                0.385
                                                                         0.787
                                                                                   0.548
                                                                0.066
                0.136
                        0.086
                                 0.254
                                           0.136
                                                     0.178
                                                                         0.595
                                                                                   0.221
                                                             0.286
Weighted Avg. 0.530 0.239
                                 0.473
                                           0.530
                                                     0.491
                                                                         0.718
                                                                                   0.516
=== Confusion Matrix ===
            c d <-- classified as
 8 989 547 225 | a = 3
185 5495 1442 529 | b = Never
111 1182 3590 524 | c = 1
108 1394 1259 436 | d = 2
                                                                                                                       Log
                                                                                                                               x 0
```

Figure 17: RandomForest OneRAttributeEval Correlation Result

J48 – WrapperSubsetEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances
                                                            54.4846 %
Incorrectly Classified Instances
                                       0.3035
0.2832
Kappa statistic
Mean absolute error
Root mean squared error
                                           0.3975
                                         82.0826 %
Relative absolute error
Root relative squared error
                                          95.6879 %
Total Number of Instances
=== Detailed Accuracy By Class ===
                  TP Rate FP Rate Precision Recall
                                                         F-Measure MCC
                                                                               ROC Area PRC Area Class
                                                                               0.570
                  0.039 0.022 0.167 0.039
0.704 0.330 0.610 0.704
                                                         0.064 0.034
0.654 0.370
                                                                                         0.132
                                                                               0.737
                                                                                         0.665
                                                                                                    Never
                                    0.530 0.745
0.299 0.115
0.486 0.545
                                    0.530
                                                0.745
                                                         0.619
0.166 0.087
                  0.115
                           0.058
                                                                               0.587
                                                                                          0.232
Weighted Avg.
                0.545
                          0.236
                                   0.486
                                                         0.498
                                                                               0.710
                                                                                         0.500
=== Confusion Matrix ===
                    d <-- classified as
  72 1009 582 166 | a = 3
159 5386 1670 436 | b = Never
82 1038 4027 260 | c = 1
  118 1393 1318 368 | d = 2
                                                                                                                                        x 0
```

Figure 18: J48 WrapperSubsetEval Correlation Result

Naive Bayes – WrapperSubsetEval Attributes

```
== Stratified cross-validation ===
  = Summary =
Correctly Classified Instances
                                                       50.9345 %
Incorrectly Classified Instances
                                                       49.0655 %
                                    0.2268
Kappa statistic
Mean absolute error
                                       0.3981
Root mean squared error
Relative absolute error
Root relative squared error
                                      95.8323 %
                                   18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                        ROC Area PRC Area Class
                                 0.179
                0.003 0.001
                                                                0.010
                                                                        0.596
                                                                                  0.132
                0.760
                        0.477
                                 0.539
                                            0.760
                                                    0.631
                                                                0.286
                                                                        0.702
                                                                                  0.628
                                                                                            Never
                0.579
                        0.242
                                 0.505
                                            0.579
                                                     0.539
                                                                0.325
                                                                        0.733
                                                                                   0.530
                0.082
Weighted Avg.
                0.509
                       0.284
                                0.440
                                          0.509
                                                    0.450
                                                               0.227
                                                                        0.680
                                                                                  0.477
 == Confusion Matrix ===
                  d <-- classified as
  5 1176 494 154 | a = 3
14 5816 1459 362 | b = Never
3 1979 3129 296 | c = 1
   6 1817 1113 261 | d = 2
                                                                                                                              x 0
```

Figure 19: NaiveBayes WrapperSubsetEval Correlation Result

OneR – WrapperSubsetEval Attributes

```
Classifier output
=== Stratified cross-validation ===
=== Summary ===
                                                            47.9982 %
Correctly Classified Instances
Incorrectly Classified Instances
                                        0.198
0.26
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
                                          0.5099
                                          75.3499 %
Root relative squared error
                                        122.7611 %
                                      18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                  TP Rate FP Rate Precision Recall
                                                        0.009
                 0.005 0.007 0.077 0.005
0.623 0.368 0.554 0.623
                                                                     -0.006
                                                                               0.499
                                                                                         0.101
                                                         0.587
                                                                     0.252
                                                                               0.628
                                                                                         0.505
                                                                                                    Never
                                                         0.536
                                                                                         0.394
                  0.693
                          0.382
                                     0.437
                                                0.693
                                                                     0.286
                                                                               0.656
Weighted Avg.
                 0.480
                          0.278
                                    0.408
                                              0.480
                                                         0.423
                                                                     0.194
=== Confusion Matrix ===
                   d <-- classified as
   9 981 730 109 | a = 3
49 4769 2578 255 | b = Never
   28 1379 3748 252 |
   31 1482 1530 154 |
                                                                                                                                Log
                                                                                                                                        x 0
```

Figure 20: OneR WrapperSubsetEval Correlation Result

RandomForest – WrapperSubsetEval Attributes

```
Classifier output
  = Stratified cross-validation =
Correctly Classified Instances
                                                     52.2506 %
Incorrectly Classified Instances
                                   0.2808
0.2799
Kappa statistic
Mean absolute error
                                      0.3957
Root mean squared error
                                    81.1266 %
Root relative squared error
                                     95.2679 %
                                  18084
Total Number of Instances
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                      ROC Area PRC Area Class
                                        0.057
                                                   0.082
                0.057 0.039 0.142
                                                            0.028
                                                                      0.579
                                                                               0.129
                0.659
                       0.323
                                 0.599
                                                              0.332
                0.698
                        0.260
                                 0.533
                                           0.698
                                                   0.605
                                                              0.411
                                                                      0.805
                                                                                0.578
               0.165
                        0.089
                                0.285
                                          0.165
                                                   0.209
                                                             0.096
                                                                      0.611
                                                                               0.233
                                                          0.283
                                        0.523
Weighted Avg. 0.523 0.234
                                0.478
                                                   0.492
                                                                      0.718
                                                                               0.522
=== Confusion Matrix ===
                 d <-- classified as
 105 938 558 228 | a = 3
370 5040 1558 683 | b = Never
  121 1093 3775 418 |
  141 1342 1185 529 | d = 2
                                                                                                                  Log
                                                                                                                         × C
```

Figure 21: RandomForest WrapperSubsetEval Correlation Result

J48 – Self Selected Attributes

```
Classifier output
 == Stratified cross-validation ===
Correctly Classified Instances
                                                      51.0064 %
Incorrectly Classified Instances
                                    0.2411
Kappa statistic
Mean absolute error
                                      0.3087
                                      0.4102
Root mean squared error
Relative absolute error
                                     89.4488 %
Root relative squared error
                                      98.7595 %
Total Number of Instances
                                  18084
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                        ROC Area PRC Area Class
                0.033 0.021 0.152 0.033
                                                    0.054 0.025
                                                                        0.545 0.117
                                                    0.550
                0.635
                         0.287
                                 0.485
                                           0.635
                                                               0.326
                                                                        0.696
                                                                                 0.454
                                          0.082
                0.082
                        0.047
                                 0.273
                                                              0.060
                                                                        0.534
                                                                                 0.206
                                                                                           2
                                          0.510
                                                             0.242
Weighted Avg. 0.510 0.266 0.448
                                                    0.460
                                                                       0.647
                                                                                 0.426
=== Confusion Matrix ===
                 d <-- classified as
 60 1053 594 122 | a = 3
121 5469 1782 279 | b = Never
131 1546 3432 298 | c = 1
  83 1587 1264 263 | d = 2
                                                                                                                    Log
```

Figure 22: J48 Self Selection Correlation Result

Naive Bayes – Self Selected Attributes

```
Classifier output —
  = Stratified cross-validation ==
=== Summary ===
Correctly Classified Instances
Incorrectly Classified Instances
                                                          53.257 %
                                        0.1121
0.2995
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
                                        86.8068 %
Root relative squared error
Total Number of Instances
                                       100.7237 %
                                     18084
=== Detailed Accuracy By Class ===
                 TP Rate FP Rate Precision Recall F-Measure MCC
                                                                            ROC Area PRC Area Class
                 0.001 0.000 0.250 0.001
                                                                            0.560
                 0.916
                          0.766
                                   0.467
                                               0.916
                                                       0.619
                                                                   0.197
                                                                            0.693
                                                                                       0.617
                                                                                                 Never
                 0.257
                          0.115
                                   0.489
                                               0.257
                                                        0.337
                                                                   0.179
                                                                            0.707
                                                                                       0.454
                 0.018
                          0.012
                                   0.235
                                              0.018
                                                       0.033
                                                                   0.018
                                                                            0.552
                                                                                       0.207
                 0.467
=== Confusion Matrix ===
                  d <-- classified as
    2 1516 275 36 | a = 3
5 7005 572 69 | b = Never
0 3940 1390 77 | c = 1
    1 2532 608 56 |
                                                                                                                                    x 0
                                                                                                                            Log
```

Figure 23: NaiveBayes Self Selection Correlation Result

OneR – Self Selected Attributes

```
Classifier output
  = Stratified cross-validation ===
Correctly Classified Instances
                                                     47.9982 %
Incorrectly Classified Instances
                                   9404
                                                     52.0018 %
                                    0.198
Kappa statistic
Mean absolute error
Root mean squared error
                                      0.5099
Relative absolute error
                                     75.3499 %
Root relative squared error
                                    122.7611 %
Total Number of Instances
                                  18084
=== Detailed Accuracy By Class ===
                TP Rate FP Rate Precision Recall F-Measure MCC
                                                                       ROC Area PRC Area Class
               0.005
                        0.007
                                0.077
                                          0.005
                                                   0.009
                                                              -0.006
                                                                      0.499
                                                                                0.101
                                                              0.252
                        0.368
                                0.554
                                           0.623
                                                   0.587
                                                                      0.628
                                                                                0.505
                0.623
                                                                                         Never
                0.048
                        0.041
                                0.200
                                           0.048
                                                   0.078
                                                              0.013
                                                                      0.503
                                                                                0.178
                               0.408
                                         0.480
Weighted Avg. 0.480
                                                                      0.601
                       0.278
                                                   0.423
                                                              0.194
                                                                                0.373
=== Confusion Matrix ===
                     <-- classified as
  9 981 730 109 | a = 3
49 4769 2578 255 | b = Never
  28 1379 3748 252 I
  31 1482 1530 154 | d = 2
                                                                                                                          x 0
                                                                                                                  Log
```

Figure 24: OneR Self Selection Correlation Result

Random Forest – Self Selected Attributes

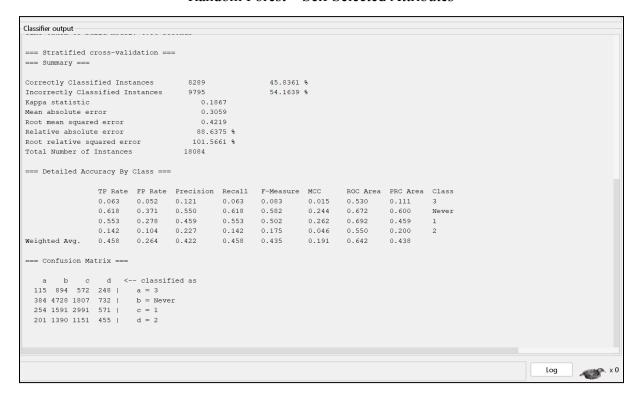


Figure 25: RandomForest Self Selection Correlation Result

Part 5 – Analysis & Conclusion

Part 5.1 - Analysis

Having built and tested our models, we then looked at their performance metrics to determine which model performed best. In the context of our dataset, a false positive denotes a criminal that is predicted to commit a second crime but will not, and a false negative denotes a criminal that is not predicted to commit a second crime but does. Because a false negative would be more detrimental or "costly," we decided to look at recall in addition to accuracy. This is because recall measures the proportion of true positive values to the total number of actually positive instances.

Below are five histograms, one for each set of attributes, that compare the accuracy of the four classification algorithms:

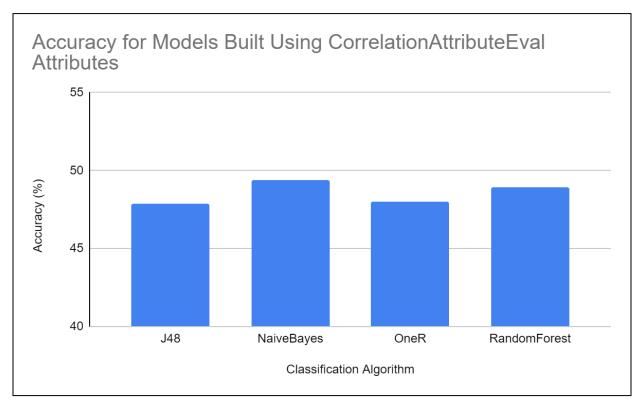


Figure 26: Accuracy for Models Using CorrelationAttributeEval

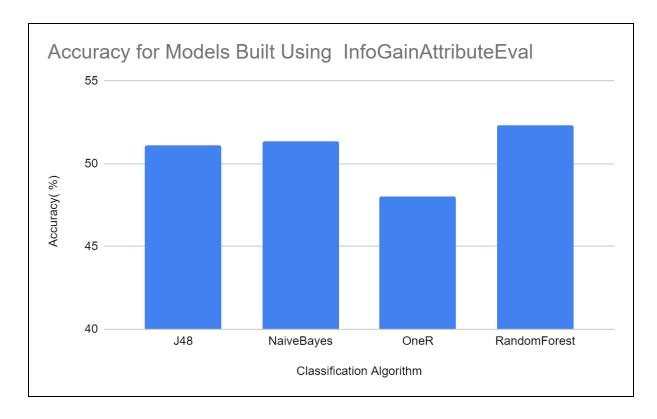
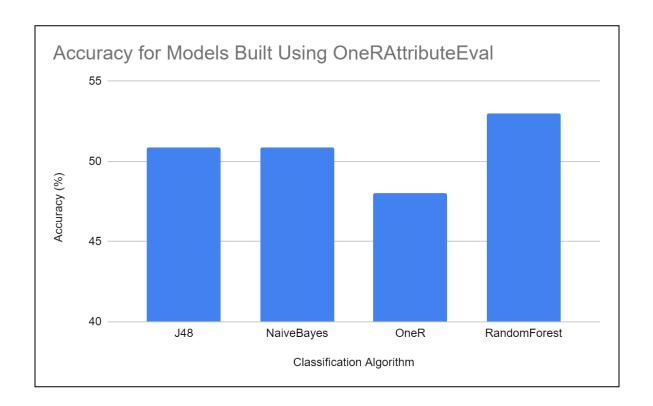


Figure 27: Accuracy for Models Using InfoGainAttributeEval



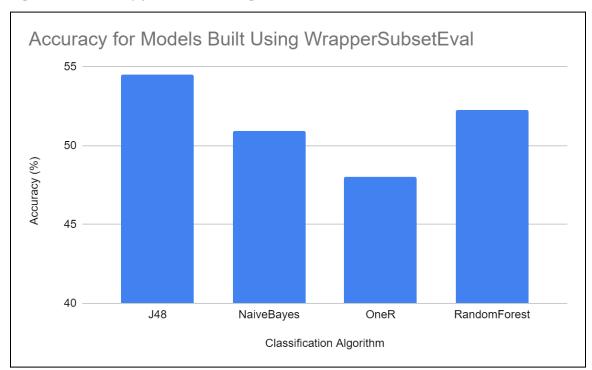


Figure 28: Accuracy for Models Using OneRAttributeEval

Figure 29: Accuracy for Models Using WrapperSubsetEval

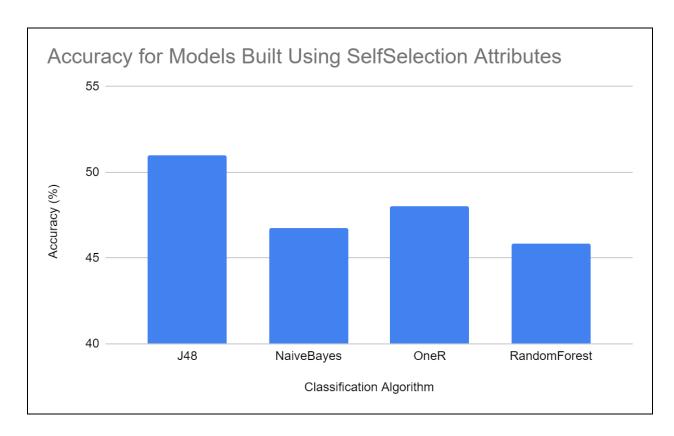


Figure 30: Accuracy for Models Using Self-Selection

In general, models built using attributes selected by CorrelationAttributeEval and SelfSelection had lower accuracy than models built using InfoGainEval, OneRAttributeEval, and WrapperSubsetEval. The model with the highest accuracy utilized attributes selected by the WrapperSubsetEval attribute selector, and used a J48 classification algorithm. This model had an accuracy of 54.4846.

Similarly, 5 histograms are depicted in the following *Figures*. These histograms depict the recall for each model. We calculated recall using Macroaveraging due to the class imbalance in our dataset.

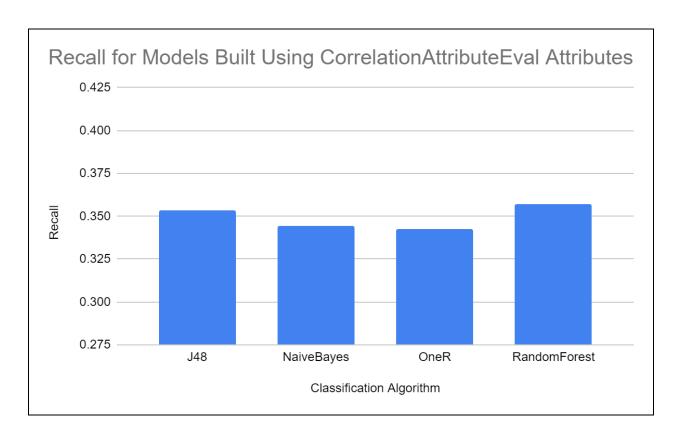


Figure 31: Recall for Models Using CorrelationAttributeEval

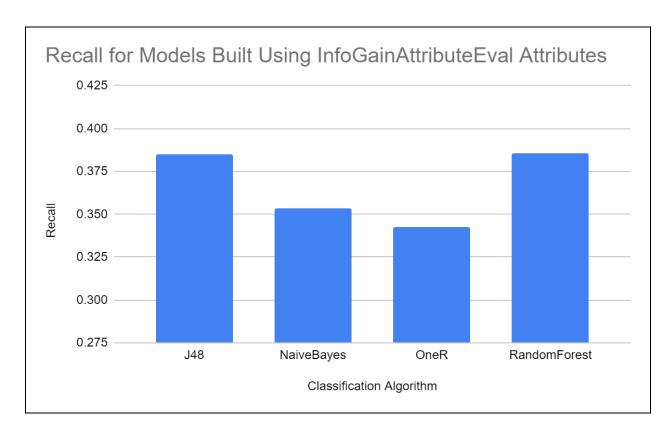


Figure 32: Recall for Models Using InfoGainAttributeEval

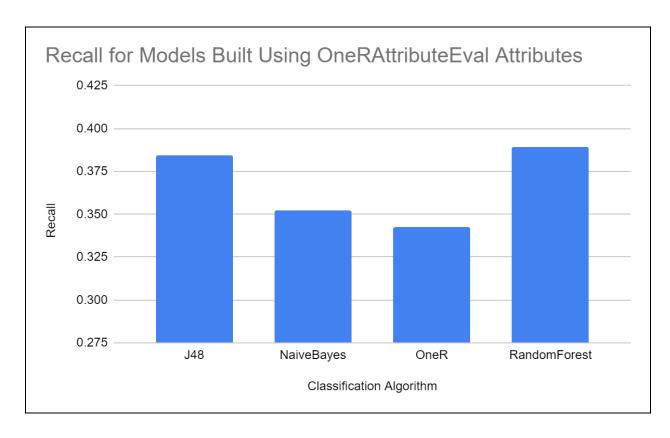


Figure 33: Recall for Models Using OneRAttributeEval

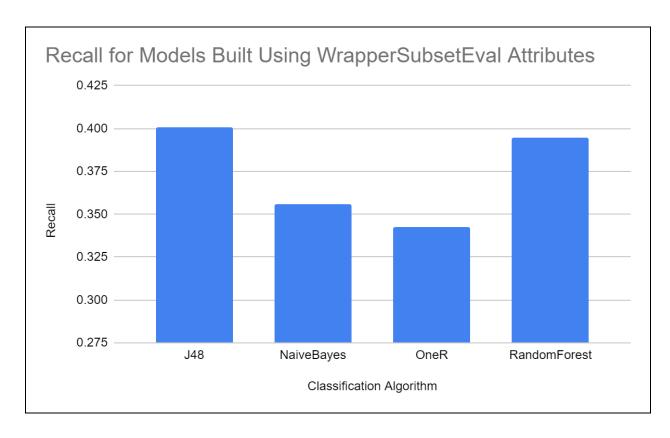


Figure 34: Recall for Models Using WrapperSubsetEval

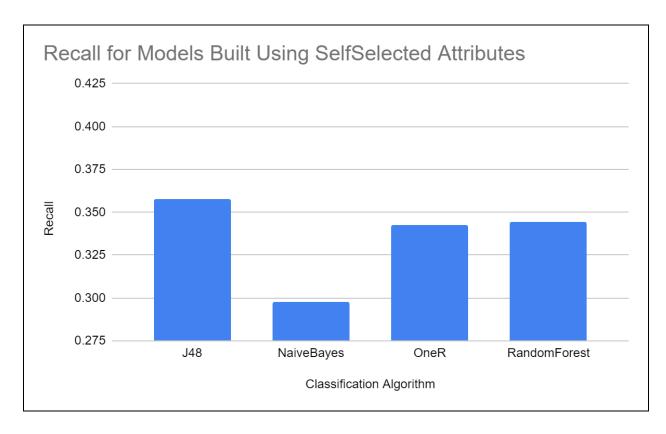


Figure 35: Recall for Models Using Self Selected Attributes

Once more, models built using attributes selected by CorrelationAttributeEval and SelfSelection have lower recalls than models built using InfoGainEval, OneRAttributeEval, and WrapperSubsetEval. The model with the highest recall, once again, utilized attributes selected by the WrapperSubsetEval attribute selector and used a J48 classification algorithm. This model had a recall of 0.400802311309.

Because the model built utilizing attributes selected by the WrapperSubsetEval attribute selector and a J48 classification algorithm had both the highest accuracy and the highest recall, this model is the best model for predicting recidivism.

Part 5.2 – Discussion of Error and Conclusion

While we identified some potentially pertinent factors in predicting the likelihood of recidivism, none of our models achieved an accuracy higher than 55%. One potential reason for this is that we separated out criminals who committed crimes after release from prison into three categories: within 1 year, 2 years, and 3 years. However, criminals that commit a second crime after release from jail likely share similar characteristics, making it difficult for our model to distinguish between these cases. In order to test this prediction, we removed all instances with the value "2" years or "3" years from the dataset. We then used our previous best model (WrapperSubsetEval Attributes with J48) to predict whether a criminal never committed a second crime, or committed a crime within one year. This model achieved an accuracy of 77.2% as seen below:

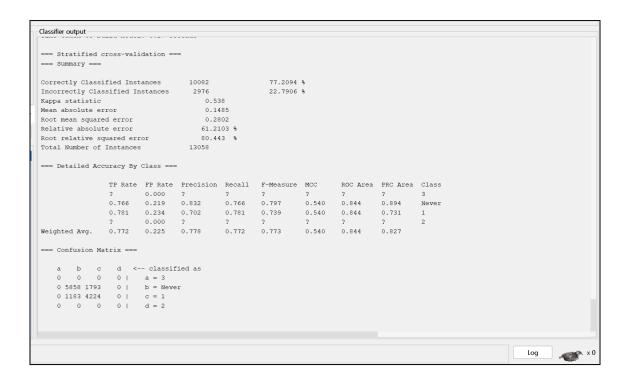


Figure 36: Accuracy of Classification Model for Instances Labeled "1" or "Never"

However, removing the label "Never" and running the model produced an accuracy of 59.8%, as seen below:

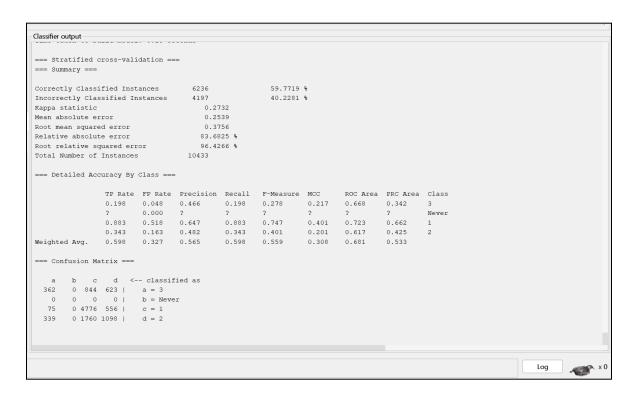


Figure 37: Accuracy of Classification Model for Instances Labeled "1", "2", or "3"

This was only slightly better than our best model with all labels. Like we predicted, these results potentially indicate that the model has a difficult time distinguishing between criminals that commit crimes 1, 2, or 3 years after release from jail.

Another potential source of error is the class imbalances in our dataset. Because so few criminals committed crimes within 2 or 3 years of release from prison, our models may have been unable to predict these class labels very well. Analyzing the correlation matrices of our 20 original models, we saw that models tended to inaccurately predict "Never" or within "1" year very frequently, while almost never predicting "2" or "3". This likely indicates that our class

imbalances (specifically, the abundance of the class label"Never" and within "1" year), were causing the model to predict "Never" and "1" most of the time.

While this risk assessment model would prove highly useful in a real world context, with such a low accuracy, more analysis is required to build a model accurate enough to be worth relying upon.