Project Report: Predicting Traffic Violation Penalty

Lavanya Krishna and Dylan Tallis
Machine Learning 1
Dr. Selma Yilmaz
October 23, 2024

Table of Contents

Project Goal

The county of Montgomery, Maryland keeps a daily updated record of all issued electronic traffic violations within the county. The dataset includes information about the nature of the violation, the description of the car and driver, the time and location, and the classification of the penalty. The penalties include: warnings, citations, safety equipment repair orders (SERO), and arrests.

Building a model to predict traffic violation penalties can identify the factors that contribute to violations that end in more significant penalties. Additionally, predicting violations based on objective factors, as opposed to personal factors like race or gender, can help reduce human bias in penalty assignment, making law enforcement practices more fair and transparent. Lastly, the model could be used to train new police officers by giving them the information about a traffic violation and comparing the violation type they would assign with what their coworkers typically did.

Description of Dataset

This dataset contains traffic violation information from all of the electronic traffic violations issued in Montgomery County, Maryland. The dataset was created in June of 2014 and since then has been updated daily. At the time of our download, the dataset had 1.96 million instances and 43 attributes. For the sake of relevancy, we decided to only consider the data from 2024, which reduced the number of instances. Our dataset now had 58,679 instances and a dimension of 42 with one class variable. That dataset contained 299,777 missing values. We chose Search Outcome as our class, with values that align with the traffic violation penalties.

Tools

The tools we used to assist us in this project included Visual Studio Code, Google Colaboratory, Notepad, and Weka. Visual Studio Code, or VS Code, is an integrated development environment that we used during preprocessing. We used the browser-based Python coding environment Google Colaboratory's (Colab) Jupyter Notebooks to create a stratified sample and split the sample into training and testing datasets. We used the Notepad text editor to change attribute data types for easier handling. Lastly, we used the Weka data platform to perform attribute selection algorithms and create and test our classification models.

Preprocessing

Initial Data Cleaning

We used VS Code's replace all tool to remove all apostrophes and double quotes from our dataset that prevented us from opening the data in Weka. We also removed all return characters placed in the middle of some instances. Instead of scrolling through and checking each of the 58,679 instances, we found it more efficient to open the file in Weka, as it would throw an error and specify the line, which was the line containing the return.

Missing Values

We found that four attributes, Search Disposition, Search Reason, Search Type, and Search Arrest Reason, had 97% of their values missing. This percentage exceeded our cutoff of 70% so we removed the four attributes. The attribute Commercial Vehicle's only value was "No"

so we removed it as well. We found that the majority of the rest of the missing values corresponded to instances that did not have a class value, and the attributes that contained the missing values were all related to a search. Our class describes the penalty outcome of the violation after conducting a search, so if a search was not conducted, there was no data for any of the search attributes. By choosing our class attribute as Search Outcome, we removed the 40% of the dataset that had missing class values. In doing this, we reduced the number of instances in our dataset to 35,208.

Distributions and Sampling

After dealing with missing values, our dataset had the following class distributions.

Class Distributions of Dataset

Class Label	**Number of Instances**	**Proportion of Dataset** (rounded)
Warning	20,434	58%
Citation	12,620	36%
Arrest	1,518	4%
SERO	636	2%

For easier handling, we used stratified random sampling in Colab to reduce the number of instances in our dataset to 1,000 with the following class distributions. The code can be found in the Appendix.

Class Distributions of Stratified Sample

Class Label	**Number of Instances**	**Proportion of Dataset** (rounded)
Warning	580	58%
Citation	359	36%
Arrest	43	4%
SERO	18	2%

Additional Data Cleaning

After creating our new dataset using stratified random sampling, we still had some missing values. The majority of the remaining missing values were in the Article attribute. We decided to delete the attribute because all of its missing values corresponded with SERO penalties and we could therefore not properly fill them without adding bias to the dataset and model. We discovered two disguised missing values, one in Make and one in Year, which we filled with the modes.

We removed the attributes Agency, Fatal, HAZMAT, Alcohol, and Work Zone because they all only had one value. We deleted the Geolocation attribute, as it was derivable data from the Longitude and Latitude attributes, and the Violation Type attribute as it was the same data as the Search Outcome. We also removed the SeqID attribute which contained a unique ID for each traffic violation.

We found that the attribute Model was saved as a string type so we used the filter "StringToNominal" to ensure that we could use the attribute selection algorithms. While doing so we found that the change would only save if you saved the dataset as an .arff and not as a .csv. We also took the advice from the Weka FAQs and changed the data type of the Date Of Stop and Time Of Stop attributes from nominal to date using Notepad.

Train - Validation - Test Split

To split the dataset into training, validation, and testing datasets, we used Scikit-learn's train_test_split. We used a stratified 70-15-15 split, with 700 instances in the training dataset and 150 instances in the validation and testing datasets each. To create the datasets we added a new attribute, index, which we removed after. Each dataset is representative of the original population. The code can be found in the Appendix.

Class Distributions of **Training** Dataset

Class Label	**Number of Instances**	**Proportion of Dataset** (rounded)
Warning	406	58%
Citation	251	36%
Arrest	30	4%
SERO	13	2%

Class Distributions of **Validation** Dataset

Class Label	**Number of Instances**	**Proportion of Dataset**
Warning	87	58%
Citation	54	36%
Arrest	6	4%
SERO	3	2%

Class Distributions of **Testing** Dataset

Class Label	**Number of Instances**	**Proportion of Dataset** (rounded)
Warning	87	58%
Citation	54	36%
Arrest	7	5%
SERO	2	1%

Description of Attribute Selection Process

We used Weka for all approaches.

Pearson's Correlation Evaluation

This approach evaluates the worth of each attribute by calculating the Pearson's correlation between the attribute and the class. We set the cutoff value at 0.1.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 28 Search Outcome):
        Correlation Ranking Filter
Ranked attributes:
 0.21601   12 Search Conducted
 0.15221    7 Accident
 0.15221   21 Contributed To Accident
 0.13588   10 Property Damage
 0.11972   23 Gender
 0.09767    9 Personal Injury
```

The attributes chosen are Search Conducted, Accident, Contributed To Accident, Property Damage, and Gender.

CFS Subset Evaluation

This approach evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the level of redundancy between them.

```
=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.334

Attribute Subset Evaluator (supervised, Class (nominal): 28 Search Outcome):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 4,9,10,12 : 4
                     Location
                     Personal Injury
                     Property Damage
                     Search Conducted
```

The attributes chosen are Location, Personal Injury, Property Damage, and Search Conducted

Information Gain Evaluation

This approach evaluated the worth of an attribute by calculating the information gain with respect to the class. We set the cutoff value at 0.2.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 28 Search Outcome):
        Information Gain Ranking Filter

Ranked attributes:
 1.207508    4 Location
 0.525144   18 Model
 0.509508   20 Charge
 0.427613   13 Search Reason For Stop
 0.213413   12 Search Conducted
 0.208574   24 Driver City
 0.199371   17 Make
```

The attributes chosen are Location, Model, Charge, Search Reason For Stop, Search Conducted, and Driver City.

Gain Ratio Evaluation:

This approach assesses the value of an attribute by calculating the gain ratio in relation to the class. For this we set the cut off value to 0.1

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 28 Search Outcome):
        Gain Ratio feature evaluator

Ranked attributes:
 0.68543   12 Search Conducted
 0.15778    9 Personal Injury
 0.15359   10 Property Damage
 0.12621    7 Accident
 0.12621   21 Contributed To Accident
 0.12446    4 Location
 0.0867    20 Charge
```

The attributes kept from Gain Ratio Evaluation are Search Conducted, Personal Injury, Property Damage, Accident, Contributed To Accident, Location.

OneR Evaluation:

This approach evaluates the worth of each attribute by using the OneR algorithm. We set the cutoff value at 60.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 28 Search Outcome):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
67                    20 Charge
64.7                  13 Search Reason For Stop
61.7                  12 Search Conducted
60                    22 Race
59.9                  21 Contributed To Accident
```

The chosen attributes are Charge, Search Reason For Stop, Search Conducted, and Race.

Description of Classifier Models

For each dataset made from the attribute selection algorithms, we used Naive Bayesian, OneR, J48, and Random Forest classifiers.

Naive Bayesian

The Naive Bayesian Classifier uses Bayes' Theorem to calculate the probability of each class based on input features, assuming the features are independent. It predicts the class with the highest probability, making it simple yet effective for supervised learning cases.

Accessed through bayes → NaiveBayes

OneR

The OneR classifier creates simple rules based on a single feature to make predictions. It evaluates each feature individually and selects the one with the lowest error rate to form a rule from. Despite its simplicity, it has high accuracy with certain cases.

Accessed through rules → OneR

J48

The J48 classifier builds a decision tree by recursively splitting the data based on the attribute that best separates the classes. It evaluates each feature to create branches, with the goal of minimizing classification error. The final decision tree is used to predict the class of new instances, offering a balance between simplicity and accuracy.

Accessed through trees → J48

Random Forest

The Random Forest Classification method uses the Weka Classifier Rules to build a forest of random trees, where each tree independently makes a class prediction. The final prediction is based on the class that receives the most votes across all trees, following a simple majority rule since the trees are equally weighted.

Accessed through tree → RandomForest

Result and Evaluation

Although we created training, validation, and testing datasets, we realized that Weka does not use validation datasets so we decided to use the cross-validation test option with 10 folds.

Correlation with Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         636               63.6    %
Incorrectly Classified Instances       364               36.4    %
Kappa statistic                          0.2009
Mean absolute error                      0.2311
Root mean squared error                  0.3425
Relative absolute error                 86.6378 %
Root relative squared error             93.8683 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.824    0.622      0.981    0.761      0.278    0.605     0.641     Warning
                 0.075    0.008    0.844      0.075    0.138      0.184    0.586     0.446     Citation
                 0.930    0.014    0.755      0.930    0.833      0.830    0.992     0.771     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.406     0.014     SERO
Weighted Avg.    0.636    0.481    ?          0.636    ?          ?        0.611     0.565

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 569   5   6   0 |   a = Warning
 325  27   7   0 |   b = Citation
   3   0  40   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Correlation with OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         617               61.7    %
Incorrectly Classified Instances       383               38.3    %
Kappa statistic                          0.149
Mean absolute error                      0.1915
Root mean squared error                  0.4376
Relative absolute error                 71.7883 %
Root relative squared error            119.9305 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.881    0.608      0.990    0.753      0.233    0.554     0.608     Warning
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.359     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.993     0.768     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.018     SERO
Weighted Avg.    0.617    0.512    ?          0.617    ?          ?        0.553     0.515

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 574   0   6   0 |   a = Warning
 352   0   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Correlation with J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         639               63.9    %
Incorrectly Classified Instances       361               36.1    %
Kappa statistic                          0.2103
Mean absolute error                      0.2359
Root mean squared error                  0.3436
Relative absolute error                 88.4229 %
Root relative squared error             94.18   %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.817    0.624      0.981    0.763      0.286    0.567     0.613     Warning
                 0.075    0.008    0.844      0.075    0.138      0.184    0.542     0.414     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.991     0.695     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.499     0.017     SERO
Weighted Avg.    0.639    0.477    ?          0.639    ?          ?        0.575     0.534

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 569   5   6   0 |   a = Warning
 325  27   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Correlation with Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         637               63.7    %
Incorrectly Classified Instances       363               36.3    %
Kappa statistic                          0.2052
Mean absolute error                      0.2321
Root mean squared error                  0.3415
Relative absolute error                 87.0247 %
Root relative squared error             93.5799 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.821    0.623      0.981    0.762      0.281    0.604     0.642     Warning
                 0.070    0.008    0.833      0.070    0.129      0.174    0.582     0.453     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.991     0.718     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.505     0.018     SERO
Weighted Avg.    0.637    0.480    ?          0.637    ?          ?        0.611     0.566

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 569   5   6   0 |   a = Warning
 327  25   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

CFS Subset with Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         648               64.8    %
Incorrectly Classified Instances       352               35.2    %
Kappa statistic                          0.2414
Mean absolute error                      0.2344
Root mean squared error                  0.3386
Relative absolute error                 87.8647 %
Root relative squared error             92.7901 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.769    0.633      0.962    0.764      0.294    0.630     0.677     Warning
                 0.131    0.025    0.746      0.131    0.223      0.209    0.608     0.485     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.993     0.831     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.514     0.019     SERO
Weighted Avg.    0.648    0.456    ?          0.648    ?          ?        0.635     0.603

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 558  16   6   0 |   a = Warning
 305  47   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

CFS Subset with OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               59.5    %
Incorrectly Classified Instances       405               40.5    %
Kappa statistic                          0.0697
Mean absolute error                      0.2025
Root mean squared error                  0.45
Relative absolute error                 75.9119 %
Root relative squared error            123.3269 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.917    0.593      0.966    0.734      0.106    0.524     0.592     Warning
                 0.081    0.025    0.644      0.081    0.144      0.129    0.528     0.382     Citation
                 0.140    0.002    0.750      0.140    0.235      0.313    0.569     0.142     Arrest
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.499     0.018     SERO
Weighted Avg.    0.595    0.541    0.607      0.595    0.488      0.121    0.527     0.487

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 560  16   2   2 |   a = Warning
 330  29   0   0 |   b = Citation
  37   0   6   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

CFS Subset with J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         638               63.8    %
Incorrectly Classified Instances       362               36.2    %
Kappa statistic                          0.2047
Mean absolute error                      0.2349
Root mean squared error                  0.3428
Relative absolute error                 88.0468 %
Root relative squared error             93.959  %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.829    0.622      0.988    0.763      0.292    0.566     0.612     Warning
                 0.061    0.002    0.957      0.061    0.115      0.191    0.540     0.413     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.991     0.695     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.494     0.017     SERO
Weighted Avg.    0.638    0.482    ?          0.638    ?          ?        0.573     0.533

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 573   1   6   0 |   a = Warning
 330  22   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

CFS Subset with Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               59.5    %
Incorrectly Classified Instances       405               40.5    %
Kappa statistic                          0.0739
Mean absolute error                      0.2454
Root mean squared error                  0.3515
Relative absolute error                 91.9906 %
Root relative squared error             96.3199 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960    0.910    0.593      0.960    0.733      0.105    0.628     0.681     Warning
                 0.084    0.030    0.612      0.084    0.147      0.120    0.613     0.468     Citation
                 0.186    0.002    0.800      0.186    0.302      0.375    0.991     0.763     Arrest
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.570     0.056     SERO
Weighted Avg.    0.595    0.538    0.598      0.595    0.491      0.120    0.637     0.597

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 557  19   2   2 |   a = Warning
 329  30   0   0 |   b = Citation
  35   0   8   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Info Gain with Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         704               70.4    %
Incorrectly Classified Instances       296               29.6    %
Kappa statistic                          0.4223
Mean absolute error                      0.1789
Root mean squared error                  0.3201
Relative absolute error                 67.0498 %
Root relative squared error             87.7345 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.828    0.433    0.725      0.828    0.773      0.411    0.768     0.792     Warning
                 0.524    0.148    0.664      0.524    0.586      0.400    0.760     0.683     Citation
                 0.744    0.010    0.762      0.744    0.753      0.742    0.995     0.885     Arrest
                 0.222    0.009    0.308      0.222    0.258      0.250    0.940     0.262     SERO
Weighted Avg.    0.704    0.305    0.697      0.704    0.696      0.419    0.778     0.748

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 480  91   4   5 |   a = Warning
 161 188   6   4 |   b = Citation
   7   4  32   0 |   c = Arrest
  14   0   0   4 |   d = SERO
```

Info Gain with OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               59.5    %
Incorrectly Classified Instances       405               40.5    %
Kappa statistic                          0.0697
Mean absolute error                      0.2025
Root mean squared error                  0.45
Relative absolute error                 75.9119 %
Root relative squared error            123.3269 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.917    0.593      0.966    0.734      0.106    0.524     0.592     Warning
                 0.081    0.025    0.644      0.081    0.144      0.129    0.528     0.382     Citation
                 0.140    0.002    0.750      0.140    0.235      0.313    0.569     0.142     Arrest
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.499     0.018     SERO
Weighted Avg.    0.595    0.541    0.607      0.595    0.488      0.121    0.527     0.487

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 560  16   2   2 |   a = Warning
 330  29   0   0 |   b = Citation
  37   0   6   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Info Gain with J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         580               58      %
Incorrectly Classified Instances       420               42      %
Kappa statistic                          0
Mean absolute error                      0.2663
Root mean squared error                  0.3649
Relative absolute error                 99.8197 %
Root relative squared error             99.9996 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.580      1.000    0.734      ?        0.500     0.580     Warning
                 0.000    0.000    ?          0.000    ?          ?        0.498     0.358     Citation
                 0.000    0.000    ?          0.000    ?          ?        0.474     0.041     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.455     0.017     SERO
Weighted Avg.    0.580    0.580    ?          0.580    ?          ?        0.497     0.467

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 580   0   0   0 |   a = Warning
 359   0   0   0 |   b = Citation
  43   0   0   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Info Gain with Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         619               61.9    %
Incorrectly Classified Instances       381               38.1    %
Kappa statistic                          0.1465
Mean absolute error                      0.2398
Root mean squared error                  0.3413
Relative absolute error                 89.9049 %
Root relative squared error             93.5416 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.945    0.819    0.614      0.945    0.745      0.200    0.718     0.774     Warning
                 0.178    0.056    0.640      0.178    0.279      0.195    0.708     0.566     Citation
                 0.163    0.001    0.875      0.163    0.275      0.368    0.837     0.417     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.945     0.597     SERO
Weighted Avg.    0.619    0.495    ?          0.619    ?          ?        0.724     0.681

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 548  31   1   0 |   a = Warning
 295  64   0   0 |   b = Citation
  32   4   7   0 |   c = Arrest
  17   1   0   0 |   d = SERO
```

Gain Ratio with Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         650               65      %
Incorrectly Classified Instances       350               35      %
Kappa statistic                          0.2493
Mean absolute error                      0.2305
Root mean squared error                  0.3386
Relative absolute error                 86.4009 %
Root relative squared error             92.8015 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.955    0.755    0.636      0.955    0.764      0.295    0.629     0.678     Warning
                 0.148    0.031    0.726      0.148    0.245      0.215    0.607     0.479     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.994     0.840     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.486     0.018     SERO
Weighted Avg.    0.650    0.450    ?          0.650    ?          ?        0.634     0.602

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 554  20   6   0 |   a = Warning
 299  53   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Gain Ratio with OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               59.5    %
Incorrectly Classified Instances       405               40.5    %
Kappa statistic                          0.0697
Mean absolute error                      0.2025
Root mean squared error                  0.45
Relative absolute error                 75.9119 %
Root relative squared error            123.3269 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.917    0.593      0.966    0.734      0.106    0.524     0.592     Warning
                 0.081    0.025    0.644      0.081    0.144      0.129    0.528     0.382     Citation
                 0.140    0.002    0.750      0.140    0.235      0.313    0.569     0.142     Arrest
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.499     0.018     SERO
Weighted Avg.    0.595    0.541    0.607      0.595    0.488      0.121    0.527     0.487

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 560  16   2   2 |   a = Warning
 330  29   0   0 |   b = Citation
  37   0   6   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Gain Ratio with J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         631               63.1    %
Incorrectly Classified Instances       369               36.9    %
Kappa statistic                          0.1881
Mean absolute error                      0.237
Root mean squared error                  0.3452
Relative absolute error                 88.8363 %
Root relative squared error             94.6024 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.840    0.618      0.984    0.759      0.268    0.564     0.610     Warning
                 0.047    0.005    0.850      0.047    0.090      0.146    0.538     0.409     Citation
                 1.000    0.014    0.768      1.000    0.869      0.870    0.991     0.695     Arrest
                 0.000    0.000    ?          0.000    ?          ?        0.496     0.017     SERO
Weighted Avg.    0.631    0.490    ?          0.631    ?          ?        0.571     0.531

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 571   3   6   0 |   a = Warning
 335  17   7   0 |   b = Citation
   0   0  43   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

Gain Ratio with Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         611               61.1    %
Incorrectly Classified Instances       389               38.9    %
Kappa statistic                          0.1213
Mean absolute error                      0.2414
Root mean squared error                  0.3474
Relative absolute error                 90.5004 %
Root relative squared error             95.2108 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.957    0.867    0.604      0.957    0.740      0.163    0.623     0.676     Warning
                 0.125    0.033    0.682      0.125    0.212      0.179    0.619     0.481     Citation
                 0.256    0.002    0.846      0.256    0.393      0.454    0.992     0.792     Arrest
                 0.000    0.002    0.000      0.000    0.000      -0.006   0.564     0.055     SERO
Weighted Avg.    0.611    0.515    0.631      0.611    0.522      0.178    0.636     0.600

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 555  21   2   2 |   a = Warning
 314  45   0   0 |   b = Citation
  32   0  11   0 |   c = Arrest
  18   0   0   0 |   d = SERO
```

OneR with Naive Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         717               71.7    %
Incorrectly Classified Instances       283               28.3    %
Kappa statistic                          0.4489
Mean absolute error                      0.1787
Root mean squared error                  0.312
Relative absolute error                 66.9742 %
Root relative squared error             85.5073 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.840    0.433    0.728      0.840    0.780      0.426    0.777     0.795     Warning
                 0.507    0.134    0.679      0.507    0.581      0.404    0.760     0.694     Citation
                 0.860    0.009    0.804      0.860    0.831      0.824    0.995     0.854     Arrest
                 0.611    0.006    0.647      0.611    0.629      0.622    0.992     0.707     SERO
Weighted Avg.    0.717    0.300    0.712      0.717    0.708      0.439    0.784     0.760

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 487  83   4   6 |   a = Warning
 172 182   5   0 |   b = Citation
   3   3  37   0 |   c = Arrest
   7   0   0  11 |   d = SERO
```

OneR with OneR

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         670               67      %
Incorrectly Classified Instances       330               33      %
Kappa statistic                          0.3091
Mean absolute error                      0.165
Root mean squared error                  0.4062
Relative absolute error                 61.8541 %
Root relative squared error            111.3236 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.917    0.590    0.682      0.917    0.782      0.389    0.663     0.674     Warning
                 0.345    0.108    0.642      0.345    0.449      0.289    0.619     0.457     Citation
                 0.023    0.006    0.143      0.023    0.040      0.041    0.508     0.045     Arrest
                 0.722    0.007    0.650      0.722    0.684      0.679    0.858     0.474     SERO
Weighted Avg.    0.670    0.382    0.644      0.670    0.629      0.344    0.644     0.565

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 532  42   1   5 |   a = Warning
 228 124   5   2 |   b = Citation
  16  26   1   0 |   c = Arrest
   4   1   0  13 |   d = SERO
```

OneR with J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         687               68.7    %
Incorrectly Classified Instances       313               31.3    %
Kappa statistic                          0.3491
Mean absolute error                      0.1976
Root mean squared error                  0.3312
Relative absolute error                 74.0841 %
Root relative squared error             90.7791 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.921    0.612    0.675      0.921    0.779      0.375    0.729     0.768     Warning
                 0.331    0.070    0.726      0.331    0.455      0.338    0.718     0.583     Citation
                 0.535    0.004    0.852      0.535    0.657      0.664    0.840     0.575     Arrest
                 0.611    0.007    0.611      0.611    0.611      0.604    0.946     0.523     SERO
Weighted Avg.    0.687    0.380    0.700      0.687    0.654      0.378    0.734     0.689

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 534  38   3   5 |   a = Warning
 237 119   1   2 |   b = Citation
  15   5  23   0 |   c = Arrest
   5   2   0  11 |   d = SERO
```

OneR with Random Forest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         684               68.4    %
Incorrectly Classified Instances       316               31.6    %
Kappa statistic                          0.3725
Mean absolute error                      0.1928
Root mean squared error                  0.3296
Relative absolute error                 72.2772 %
Root relative squared error             90.331  %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.836    0.488    0.703      0.836    0.764      0.371    0.754     0.800     Warning
                 0.462    0.153    0.629      0.462    0.533      0.337    0.731     0.618     Citation
                 0.442    0.005    0.792      0.442    0.567      0.579    0.925     0.675     Arrest
                 0.778    0.008    0.636      0.778    0.700      0.698    0.993     0.616     SERO
Weighted Avg.    0.684    0.338    0.679      0.684    0.671      0.374    0.758     0.726

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 485  87   3   5 |   a = Warning
 188 166   2   3 |   b = Citation
  14  10  19   0 |   c = Arrest
   3   1   0  14 |   d = SERO
```

Justification of Selection

The combination of OneR attribute selection and Naive Bayes produced the highest percent of correctly classified instances, 71.1%, closely followed by Info Gain with Naive Bayes with 70.4%. The OneR and Naive Bayes models also had among the lowest error rates, including mean absolute error, root mean squared error, relative absolute error, and root relative squared error, and the highest precision and F-measure. High accuracy and low error are good indicators of consistently satisfactory performance.

Discussions and Conclusions

A significant number of the models classified the majority or even all of the instances as Warning. This can be observed in the confusion matrices. Some models have zeros in all columns except the column classified as Warning. This can be explained by the skewed distribution of class values, with 58% of the instances classified as Warning. The class imbalance causes the models to perform well on the majority class value, Warning, and poorly on the minority class values.

Of the two best models, the shared attributes include Charge, Search Reason For Stop, and Search Conducted, with Charge being the highest rated by the attribute selection evaluations of OneR and Info Gain.

Although 71.1% accuracy was our best result, compared to an ideal model accuracy, our results were pretty low. There are multiple factors that could have contributed to this outcome. Because of the class imbalance, the models often had majority True Positives and False Positives. A way to mitigate the effects of the imbalance would be to balance the dataset by oversampling the minority class values or undersampling the majority class value instead of using stratified random sampling. Class weights could also be assigned to emphasize the misclassification of the minority class values. A factor that is out of our control is the inherent human bias within the data. Each traffic violation evaluation is conducted by a different police officer, at a different time, in a different place, and with different circumstances. The penalty assigned to each violation could vary depending on more trivial matters such as the officer's emotional state or fatigue, or more significant matters such as differences in seniority or experience. The data also reflects the bias of officer's towards the driver, if they tend to stop people and/or give harsher penalties to people of specific genders or ethnicities. A way to reduce some of the human bias is to take traffic violation data that was reported only by officers who have achieved a specific level of experience.

The model with the best results was the Naive Bayes model with OneR selection. We were able to find and preprocess a dataset with a real world application, and train and test twenty classification models to predict the penalty assigned to traffic violations in Montgomery County, Maryland. Considering the low accuracy, there is room for improvement in sampling to account for class value imbalances and filtering source data to include as little human bias as possible.

Steps to Reproduce our Chosen Model:

1. Download the Stratified Sample Dataset (found in the Appendix)
2. Open Weka and load updated_stratified_sample.arff
3. Click on the "Select Attributes" tab and choose "OneRAttributeEval"
    a. When prompted to select "Ranker" search, select yes
4. Select "Search Outcome" as class in drop down menu
5. Click Start
6. Use the cutoff value of 60 (inclusive) to select attributes
7. In the "Preprocess" tab select All in the Attributes window, unselect the chosen attributes and the class attribute, and press Remove
    a. This dataset can also be found in the Appendix, named OneR Evaluation Dataset
    b. Open Weka and load oner_eval.arff
8. Navigate to the "Classify" tab
9. Choose weka → classifiers → bayes → NaiveBayes
10. Ensure "Cross-validation" with 10 folds in chosen in Test options pane
11. Ensure class in correct
12. Click Start

Team Members and Tasks Performed

Finding the Data: Lavanya and Dylan
Proposal: Lavanya and Dylan
Goal, Dataset, Tools: Lavanya and Dylan
Preprocessing: Lavanya
Attribute Selection Algorithms and Classifiers: Lavanya
Results Output: Dylan
Discussions and Conclusions: Dylan
Appendix: Lavanya and Dylan
Final Report: Lavanya and Dylan
Slideshow: Lavanya

Appendix

Data Website

- Data Source Website
- Alternative Link

Datasets

- Cleaned Dataset With Violations From 2024
- Stratified Sample
- Splits (unused)

Attribute Selection Datasets

- Correlation Selection Dataset
- CFS Subset Selection Dataset
- Information Gain Selection Dataset
- Gain Ratio Selection Dataset

- OneR Selection Dataset

Code

- Code for Stratified Random Sampling:

```
import pandas as pd

df = pd.read_csv("/content/drive/MyDrive/ML_Project/traffic_data.csv")

from google.colab import drive
drive.mount('/content/drive')

condition_a = df['Search Outcome'] == "Warning"
strata_a = df[condition_a]

condition_b = df['Search Outcome'] == "Citation"
strata_b = df[condition_b]

condition_c = df['Search Outcome'] == "Arrest"
strata_c = df[condition_c]

condition_d = df['Search Outcome'] == "SERO"
strata_d = df[condition_d]
strata_a_sample = strata_a.sample(n = 580, random_state = 0)

strata_b_sample = strata_b.sample(n = 359, random_state = 0)

strata_c_sample = strata_c.sample(n = 43, random_state = 0)

strata_d_sample = strata_d.sample(n = 18, random_state = 0)

stratified_sample_df = pd.DataFrame()

stratified_sample_df = pd.concat([strata_a_sample, strata_b_sample, strata_c_sample,
strata_d_sample])

print(stratified_sample_df.head())

from google.colab import files
files.download('stratified_sample.csv')
```

- Code for Training - Validation - Test Splits:

```
import pandas as pd

df = pd.read_csv("/content/drive/MyDrive/machine_learning/d_stratified_sample.csv")

from sklearn.model_selection import train_test_split

cols = df.columns.tolist()
cols = cols[0:21] + cols[22:] + [cols[21]]
df = df[cols]

X = df.iloc[:, 0:-1]
y = df.iloc[:, -1]

X_train, X_val_test, y_train, y_val_test = train_test_split(X,y, test_size=0.3,
random_state=0, stratify=y)
X_val, X_test, y_val, y_test = train_test_split(X_val_test, y_val_test, test_size=0.5,
random_state=0, stratify=y_val_test)

X_train['index'] = X_train.index
y_train = y_train.reset_index()

train = pd.merge(X_train, y_train, on = 'index')

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

X_val['index'] = X_val.index
y_val = y_val.reset_index()

validation = pd.merge(X_val, y_val, on = 'index')

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

X_test['index'] = X_test.index
y_test = y_test.reset_index()

test = pd.merge(X_test, y_test, on = 'index')

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

Descriptions of Attributes

1. SeqID: unique traffic stop ID
2. Date Of Stop: date of traffic violation
3. Time Of Stop: time of traffic violation
4. Agency: agency issuing the traffic violation
5. SubAgency: court code that represents the officer's district assignment

6. Description:	text description of charge
7. Location:	address or intersection
8. Latitude:	latitude of violation location
9. Longitude:	longitude of violation location
10. Accident:	binary, yes if involved in an accident
11. Belts:	in accident cases, binary, yes in seat belts were used
12. Personal Injury:	binary, yes if involved personal injury
13. Property Damage:	binary, yes if involved property damage
14. Fatal:	binary, yes if involved a fatality
15. Commercial License:	binary, yes if driver has a CDL
16. HAZMAT:	binary, yes if involved hazardous materials
17. Commercial Vehicle:	binary, yes if the vehicle is a commercial vehicle
18. Alcohol:	binary, yes if included an alcohol related suspension
19. Work Zone:	binary, yes if in a work zone
20. Search Conducted:	binary, yes if a search was conducted
21. Search Disposition:	disposition of search
22. Search Reason:	reason for search
23. Search Reason For Stop:	reason for stop that lead to the search
24. Search Type:	person, property, both, etc.
25. Search Arrest Reason:	arrest reason from search
26. State:	state of vehicle registration
27. Year:	year of vehicle
28. Make:	manufacturer of vehicle
29. Model:	vehicle model
30. Color:	vehicle color
31. Charge:	numeric code for charge
32. Article:	article of state law
33. Contributed To Accident:	if violation contributed to accident
34. Race:	race of driver
35. Gender:	gender of driver
36. Driver City:	city of driver’s home address
37. Driver State:	state of driver’s home address
38. DL State:	state issuing driver’s license
39. Arrest Type:	marked, unmarked, etc.
40. Geolocation:	geo-coded location information
41. Violation Type:	warning, citation, ESERO
42. Search Outcome:	warning, citation, arrest, SERO

Sources

GeeksForGeeks. (2024, July 10). *Naive Bayes classifiers*. GeeksForGeeks. Retrieved October 22, 2024, from https://www.geeksforgeeks.org/naive-bayes-classifiers/

Khanna, N. (2021, August 18). *J48 classification (C4.5 algorithm) in a nutshell*. Medium. Retrieved October 22, 2024, from https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e

OneR. (n.d.). Saed Sayad. Retrieved October 22, 2024, from https://www.saedsayad.com/oner.htm

Random forest. (2024, October 2). Wikipedia. Retrieved October 22, 2024, from https://en.wikipedia.org/wiki/Random_forest