Predicting Lethality of Car Crashes in New York City Boroughs

Aaryan Sumesh, Sami Saleh

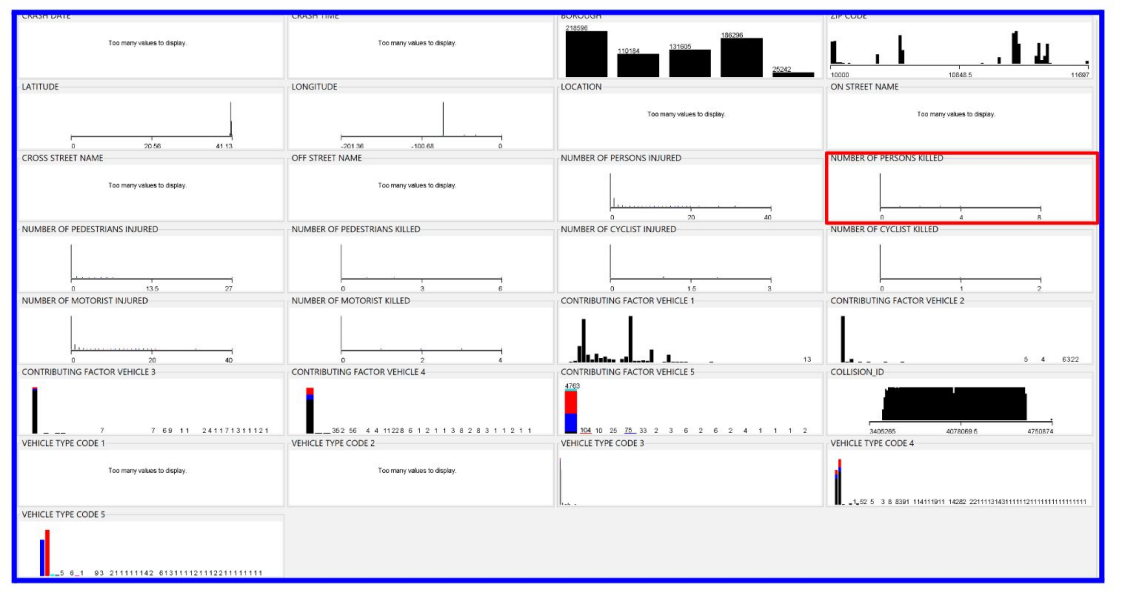
Background and Project Goal

Growing City Complexity : As NYC's traffic conditions become more complex, understanding road incidents is crucial.

Safety Insights : These classifications can reveal patterns that lead to preventative measures.

We will be trying to predict the **lethality of car crashes** in New York City.

Dataset Information



Key attributes:

NUMBER OF PERSONS KILLED: Total persons killed in car accident

CONTRIBUTING FACTOR VEHICLE 1, 2, 3 ...: A label of a possible reason that the driver was involved in the car accident, ex. Pavement slippery, Illegal drugs, ...

VEHICLE TYPE CODE 1, 2, 3...: A label for the form of transportation of the vehicle, ex. Sedan, bike, taxi, ...

Statistic	Value
Minimum	0
Maximum	8
Mean	0.001
StdDev	0.041

Preprocessing

Step 1: Remove all instances where class attribute is missing (RemoveWithValues)

Step 2: Remove unnecessary attributes - Collision_id

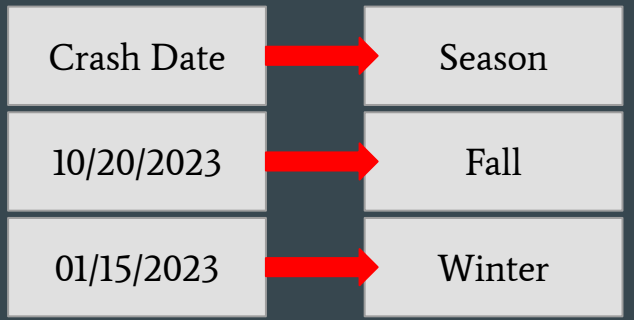
Step 3: Remove attributes with majority missing values - Cross street name, Off street name, Vehicle Type code 3, 4, 5.

Step 4: Remove redundant attributes - Number of pedestrians killed, Number of motorists killed, Number of cyclists killed, Location

Step 5: Alter crash date to Season

```
def get_season(month):
    if month in [12, 1, 2]:
        return 'Winter'
    elif month in [3, 4, 5]:
        return 'Spring'
    elif month in [6, 7, 8]:
        return 'Summer'
    else:
        return 'Fall'
```

Step 5



Preprocessing

Step 6: Alter crash time

Step 7: Alter contributing vehicle factors 3-5

Crash Time	Rush Hour	Time of Day
02:35	No	Night
16:07	Yes	Day

```python
def is_rush_hour(hour):
    if (6 <= hour <= 9) or (16 <= hour <= 19):
        return 'Yes'
    else:
        return 'No'

def time_of_day(hour):
    if 6 <= hour < 18:
        return 'Day'
    else:
        return 'Night'
```

Step 6

Contributing Vehicle Factors 3, 4, 5	More than two Vehicles Involved
Unspecified, ?, ?	Yes
?, ?, ?	No

```python
def more_than_two_vehicles(row):
    factors = [
        row['CONTRIBUTING FACTOR VEHICLE 1'],
        row['CONTRIBUTING FACTOR VEHICLE 2'],
        row['CONTRIBUTING FACTOR VEHICLE 3'],
        row['CONTRIBUTING FACTOR VEHICLE 4'],
        row['CONTRIBUTING FACTOR VEHICLE 5']
    ]
    if sum(factor != '?' for factor in factors) > 2:
        return 'Yes'
    else:
        return 'No'
```

Step 7

Preprocessing

Step 8: Replace missing values - ReplaceMissingValues filter in weka

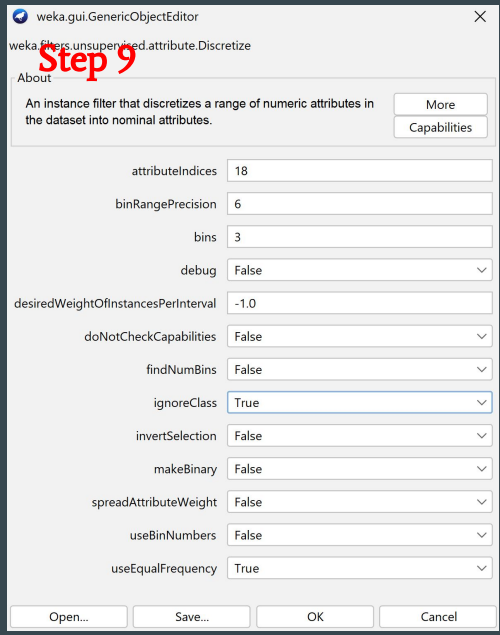
Step 9: Bin the class variable (-inf-0.5], (0.5-1.5], [1.5-inf)

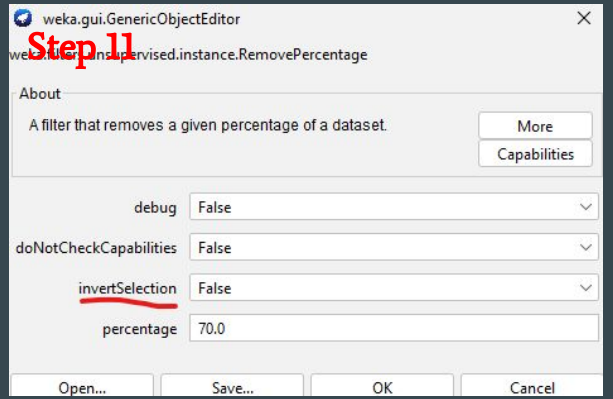
Step 10: Replace bin names with better names (non_lethal, somewhat_lethal, very_lethal)

Step 11: Train Test Split (70%, 30%) - stratified random sample

```
df["'NUMBER OF PERSONS KILLED'"] = df["'NUMBER OF PERSONS KILLED'"].replace({
    "'\\'(-inf-0.5]\\''": 'non_lethal',
    "'\\'(0.5-1.5]\\''": 'somewhat_lethal',
    "'\\'(1.5-inf)\\''": 'very_lethal'
})
```

Step 10





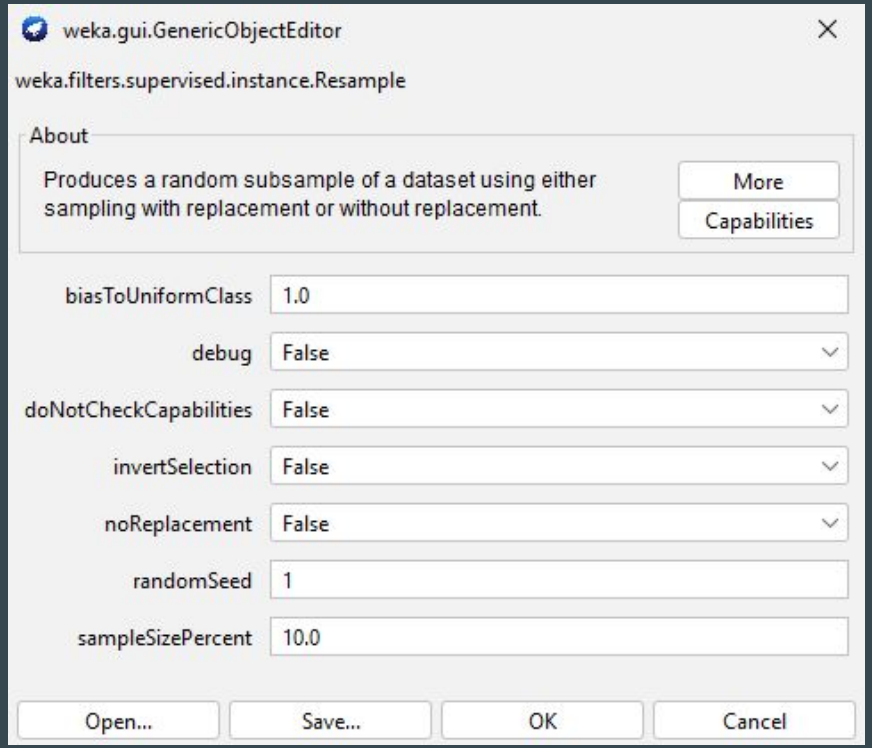
Preprocessing

No need for any Normalization

No.	Name
1	BOROUGH
2	ZIP CODE
3	LATITUDE
4	LONGITUDE
5	ON STREET NAME
6	NUMBER OF PERSONS INJURED
7	NUMBER OF PEDESTRIANS INJURED
8	NUMBER OF CYCLIST INJURED
9	NUMBER OF MOTORIST INJURED
10	CONTRIBUTING FACTOR VEHICLE 1
11	CONTRIBUTING FACTOR VEHICLE 2
12	VEHICLE TYPE CODE 1
13	VEHICLE TYPE CODE 2
14	SEASON
15	RUSH HOUR
16	TIME OF DAY
17	MORE THAN 2 VEHICLES INVOLVED
18	NUMBER OF PERSONS KILLED

Attribute Selection Algorithms and Model Classifiers

Due to the skewed distribution in the class attribute, with 732807 instances being classified as “non_lethal”, 1131 instances being classified as “somewhat_lethal”, and 38 instances being classified as “very_lethal”, we decided to take a stratified sample of the training data for attribute selection. To do this, we used the WEKA Resample filter with a sample size percent of 10% due to the dataset being very large.



Attribute Selection Algorithms and Model Classifiers

This left us with a stratified sample of the data, with each class label having 24465 instances:

Selected attribute

Name: NUMBER OF PERSONS KILLED Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	non_lethal	24465	24465
2	somewhat_lethal	24465	24465
3	very_lethal	24465	24465

Attribute Selection Algorithms and Model Classifiers

Attribute Selection Algorithm One: **GainRatioAttributeEval**

Cut-off value: **0.1**

Selected Attributes: LATITUDE, LONGITUDE, NUMBER OF PEDESTRIANS INJURED, NUMBER OF CYCLIST INJURED, CONTRIBUTING FACTOR VEHICLE 1, ON STREET NAME, CONTRIBUTING FACTOR VEHICLE 2

```
Attribute Evaluator (supervised, Class (nominal): 18 NUMBER OF PERSONS KILLED):
        Gain Ratio feature evaluator

Ranked attributes:
 0.14422    3 LATITUDE
 0.142      4 LONGITUDE
 0.13253    7 NUMBER OF PEDESTRIANS INJURED
 0.12548    8 NUMBER OF CYCLIST INJURED
 0.12393   10 CONTRIBUTING FACTOR VEHICLE 1
 0.11482    5 ON STREET NAME
 0.10769   11 CONTRIBUTING FACTOR VEHICLE 2
 0.09785    9 NUMBER OF MOTORIST INJURED
 0.09124   17 MORE THAN 2 VEHICLES INVOLVED
 0.08408    2 ZIP CODE
 0.08383    6 NUMBER OF PERSONS INJURED
 0.08249   16 TIME OF DAY
 0.07398   13 VEHICLE TYPE CODE 2
 0.06888   12 VEHICLE TYPE CODE 1
 0.06303   15 RUSH HOUR
 0.01096    1 BOROUGH
 0.00684   14 SEASON

Selected attributes: 3,4,7,8,10,5,11,9,17,2,6,16,13,12,15,1,14 : 17
```

Attribute Selection Algorithms and Model Classifiers

Attribute Selection Algorithm Two: **InfoGainAttributeEval**

Cut-off value: **0.1**

Selected Attributes: LONGITUDE, LATITUDE, ON STREET NAME, CONTRIBUTING FACTOR VEHICLE 1, ZIP CODE, VEHICLE TYPE CODE 1, NUMBER OF MOTORIST INJURED, VEHICLE TYPE CODE 2, NUMBER OF PERSONS INJURED

```
Attribute Evaluator (supervised, Class (nominal): 18 NUMBER OF PERSONS KILLED):
        Information Gain Ranking Filter

Ranked attributes:
 1.1871   4 LONGITUDE
 1.1859   3 LATITUDE
 0.9628   5 ON STREET NAME
 0.45    10 CONTRIBUTING FACTOR VEHICLE 1
 0.3458   2 ZIP CODE
 0.1734  12 VEHICLE TYPE CODE 1
 0.1454   9 NUMBER OF MOTORIST INJURED
 0.1378  13 VEHICLE TYPE CODE 2
 0.1298   6 NUMBER OF PERSONS INJURED
 0.086   11 CONTRIBUTING FACTOR VEHICLE 2
 0.0815  16 TIME OF DAY
 0.0625  17 MORE THAN 2 VEHICLES INVOLVED
 0.0547  15 RUSH HOUR
 0.0445   7 NUMBER OF PEDESTRIANS INJURED
 0.0173   1 BOROUGH
 0.0136  14 SEASON
 0.0127   8 NUMBER OF CYCLIST INJURED

Selected attributes: 4,3,5,10,2,12,9,13,6,11,16,17,15,7,1,14,8 : 17
```

Attribute Selection Algorithms and Model Classifiers

Attribute Selection Algorithm Three: **CfsSubsetEval w/ Greedy Stepwise Approach**

Selected Attributes: LATITUDE, LONGITUDE, NUMBER OF CYCLIST INJURED, CONTRIBUTING FACTOR VEHICLE 1

```
Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.272

Attribute Subset Evaluator (supervised, Class (nominal): 18 NUMBER OF PERSONS KILLED):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,4,8,10 : 4
                     LATITUDE
                     LONGITUDE
                     NUMBER OF CYCLIST INJURED
                     CONTRIBUTING FACTOR VEHICLE 1
```

Attribute Selection Algorithms and Model Classifiers

Attribute Selection Algorithm Four: **OneRAttributeEval**

Cut-off value: **40%**

Selected Attributes: LATITUDE, LONGITUDE, ON STREET NAME, CONTRIBUTING FACTOR VEHICLE 1, ZIP CODE, VEHICLE TYPE CODE 1, NUMBER OF MOTORIST INJURED, TIME OF DAY, NUMBER OF PERSONS INJURED, RUSH HOUR, VEHICLE TYPE CODE 2, MORE THAN 2 VEHICLES INVOLVED

```
Attribute Evaluator (supervised, Class (nominal): 18 NUMBER OF PERSONS KILLED):
        OneR feature evaluator.

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

Ranked attributes:
90.115   3 LATITUDE
89.678   4 LONGITUDE
81.76    5 ON STREET NAME
62.499  10 CONTRIBUTING FACTOR VEHICLE 1
54.769   2 ZIP CODE
48.105  12 VEHICLE TYPE CODE 1
46.936   9 NUMBER OF MOTORIST INJURED
46.738  16 TIME OF DAY
44.134   6 NUMBER OF PERSONS INJURED
42.868  15 RUSH HOUR
42.47   13 VEHICLE TYPE CODE 2
42.206  17 MORE THAN 2 VEHICLES INVOLVED
39.552  11 CONTRIBUTING FACTOR VEHICLE 2
38.482   1 BOROUGH
38.365  14 SEASON
35.752   7 NUMBER OF PEDESTRIANS INJURED
34.417   8 NUMBER OF CYCLIST INJURED

Selected attributes: 3,4,5,10,2,12,9,16,6,15,13,17,11,1,14,7,8 : 17
```

Attribute Selection Algorithms and Model Classifiers

Attribute Selection Algorithm Five: **Non-WEKA Approach**

Comparing all previous attribute selection algorithms, the only two attributes that were recommended to be removed in all algorithms were SEASON and BOROUGH

Attribute Selection Algorithms and Model Classifiers

Models Selected for Classification:

J48

Naive Bayes

OneR

DecisionTable

GainRatioAttributeEval

J48

```
Correctly Classified Instances       29414               93.5114 %
Incorrectly Classified Instances      2041                6.4886 %
Kappa statistic                          0.9027
Mean absolute error                      0.0455
Root mean squared error                  0.1615
Relative absolute error                 10.2483 %
Root relative squared error             34.2615 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.840    0.017    0.961      0.840    0.896      0.853    0.985     0.977     non_lethal
                 0.966    0.075    0.865      0.966    0.913      0.868    0.990     0.973     somewhat_lethal
                 1.000    0.005    0.990      1.000    0.995      0.993    1.000     1.000     very_lethal
Weighted Avg.    0.935    0.032    0.939      0.935    0.935      0.905    0.992     0.983

=== Confusion Matrix ===

     a     b     c   <-- classified as
  8803  1579   103 |     a = non_lethal
   359 10126     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Naive Bayes

```
Correctly Classified Instances       21332               67.8175 %
Incorrectly Classified Instances     10123               32.1825 %
Kappa statistic                          0.5173
Mean absolute error                      0.2145
Root mean squared error                  0.3876
Relative absolute error                 48.26   %
Root relative squared error             82.2132 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.712    0.148    0.707      0.712    0.709      0.563    0.889     0.823     non_lethal
                 0.322    0.009    0.948      0.322    0.481      0.466    0.923     0.861     somewhat_lethal
                 1.000    0.326    0.605      1.000    0.754      0.639    0.992     0.964     very_lethal
Weighted Avg.    0.678    0.161    0.753      0.678    0.648      0.556    0.935     0.883

=== Confusion Matrix ===

     a     b     c   <-- classified as
  7467   185  2833 |     a = non_lethal
  3100  3380  4005 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

GainRatioAttributeEval

OneR

```
Correctly Classified Instances       30174               95.9275 %
Incorrectly Classified Instances      1281                4.0725 %
Kappa statistic                          0.9389
Mean absolute error                      0.0271
Root mean squared error                  0.1648
Relative absolute error                  6.1087 %
Root relative squared error             34.9535 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.000    1.000      0.878    0.935      0.910    0.939     0.919     non_lethal
                 1.000    0.060    0.893      1.000    0.943      0.916    0.970     0.893     somewhat_lethal
                 1.000    0.001    0.998      1.000    0.999      0.998    0.999     0.998     very_lethal
Weighted Avg.    0.959    0.020    0.964      0.959    0.959      0.941    0.969     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9204  1260    21 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

DecisionTable

```
Correctly Classified Instances       29994               95.3553 %
Incorrectly Classified Instances      1461                4.6447 %
Kappa statistic                          0.9303
Mean absolute error                      0.0614
Root mean squared error                  0.1557
Relative absolute error                 13.8054 %
Root relative squared error             33.0319 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.861    0.000    1.000      0.861    0.925      0.897    0.984     0.976     non_lethal
                 1.000    0.069    0.879      1.000    0.935      0.904    0.984     0.953     somewhat_lethal
                 1.000    0.001    0.999      1.000    0.999      0.999    1.000     0.999     very_lethal
Weighted Avg.    0.954    0.023    0.959      0.954    0.953      0.933    0.989     0.976

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9024  1447    14 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

InfoGainAttributeEval

J48

```
Correctly Classified Instances       29917               95.1105 %
Incorrectly Classified Instances      1538                4.8895 %
Kappa statistic                          0.9267
Mean absolute error                      0.0312
Root mean squared error                  0.138
Relative absolute error                  7.0307 %
Root relative squared error             29.2666 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.856    0.002    0.996      0.856    0.921      0.891    0.989     0.987     non_lethal
                 0.997    0.067    0.882      0.997    0.936      0.905    0.994     0.979     somewhat_lethal
                 1.000    0.005    0.990      1.000    0.995      0.993    1.000     1.000     very_lethal
Weighted Avg.    0.951    0.024    0.956      0.951    0.951      0.930    0.994     0.988

=== Confusion Matrix ===

     a     b     c   <-- classified as
  8979  1404   102 |     a = non_lethal
    32 10453     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Naive Bayes

```
Correctly Classified Instances       25143               79.9332 %
Incorrectly Classified Instances      6312               20.0668 %
Kappa statistic                          0.699
Mean absolute error                      0.1481
Root mean squared error                  0.3136
Relative absolute error                 33.327  %
Root relative squared error             66.5352 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.908    0.204    0.690      0.908    0.784      0.669    0.935     0.891     non_lethal
                 0.490    0.011    0.956      0.490    0.648      0.600    0.946     0.904     somewhat_lethal
                 1.000    0.086    0.854      1.000    0.921      0.883    0.993     0.979     very_lethal
Weighted Avg.    0.799    0.100    0.833      0.799    0.784      0.717    0.958     0.925

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9519   235   731 |     a = non_lethal
  4279  5139  1067 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

InfoGainAttributeEval

OneR

```
Correctly Classified Instances       30174               95.9275 %
Incorrectly Classified Instances      1281                4.0725 %
Kappa statistic                          0.9389
Mean absolute error                      0.0271
Root mean squared error                  0.1648
Relative absolute error                  6.1087 %
Root relative squared error             34.9535 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.000    1.000      0.878    0.935      0.910    0.939     0.919     non_lethal
                 1.000    0.060    0.893      1.000    0.943      0.916    0.970     0.893     somewhat_lethal
                 1.000    0.001    0.998      1.000    0.999      0.998    0.999     0.998     very_lethal
Weighted Avg.    0.959    0.020    0.964      0.959    0.959      0.941    0.969     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9204  1260    21 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Decision Table

```
Correctly Classified Instances       29994               95.3553 %
Incorrectly Classified Instances      1461                4.6447 %
Kappa statistic                          0.9303
Mean absolute error                      0.0614
Root mean squared error                  0.1557
Relative absolute error                 13.8054 %
Root relative squared error             33.0319 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.861    0.000    1.000      0.861    0.925      0.897    0.984     0.976     non_lethal
                 1.000    0.069    0.879      1.000    0.935      0.904    0.984     0.953     somewhat_lethal
                 1.000    0.001    0.999      1.000    0.999      0.999    1.000     0.999     very_lethal
Weighted Avg.    0.954    0.023    0.959      0.954    0.953      0.933    0.989     0.976

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9024  1447    14 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

CfsSubsetEval with Greedy Stepwise Search Method

J48

```
Correctly Classified Instances       30596               97.2691 %
Incorrectly Classified Instances       859                2.7309 %
Kappa statistic                          0.959
Mean absolute error                      0.025
Root mean squared error                  0.1246
Relative absolute error                  5.614  %
Root relative squared error             26.433  %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.001    0.998      0.920    0.957      0.939    0.982     0.981     non_lethal
                 0.998    0.039    0.927      0.998    0.961      0.943    0.989     0.962     somewhat_lethal
                 1.000    0.001    0.998      1.000    0.999      0.999    1.000     0.999     very_lethal
Weighted Avg.    0.973    0.014    0.974      0.973    0.973      0.960    0.990     0.980

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9646   819    20 |     a = non_lethal
    20 10465     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Naive Bayes

```
Correctly Classified Instances       15669               49.814  %
Incorrectly Classified Instances     15786               50.186  %
Kappa statistic                          0.2472
Mean absolute error                      0.3447
Root mean squared error                  0.5415
Relative absolute error                 77.5494 %
Root relative squared error            114.8729 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.432    0.110    0.662      0.432    0.523      0.368    0.760     0.640     non_lethal
                 0.062    0.003    0.921      0.062    0.117      0.189    0.690     0.582     somewhat_lethal
                 1.000    0.640    0.439      1.000    0.610      0.397    0.867     0.757     very_lethal
Weighted Avg.    0.498    0.251    0.674      0.498    0.416      0.318    0.772     0.660

=== Confusion Matrix ===

     a     b     c   <-- classified as
  4531    56  5898 |     a = non_lethal
  2310   653  7522 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

CfsSubsetEval with Greedy Stepwise Search Method

OneR

```
Correctly Classified Instances       30174               95.9275 %
Incorrectly Classified Instances      1281                4.0725 %
Kappa statistic                          0.9389
Mean absolute error                      0.0271
Root mean squared error                  0.1648
Relative absolute error                  6.1087 %
Root relative squared error             34.9535 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.000    1.000      0.878    0.935      0.910    0.939     0.919     non_lethal
                 1.000    0.060    0.893      1.000    0.943      0.916    0.970     0.893     somewhat_lethal
                 1.000    0.001    0.998      1.000    0.999      0.998    0.999     0.998     very_lethal
Weighted Avg.    0.959    0.020    0.964      0.959    0.959      0.941    0.969     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9204  1260    21 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Decision Table

```
Correctly Classified Instances       15669               49.814  %
Incorrectly Classified Instances     15786               50.186  %
Kappa statistic                          0.2472
Mean absolute error                      0.3447
Root mean squared error                  0.5415
Relative absolute error                 77.5494 %
Root relative squared error            114.8729 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.432    0.110    0.662      0.432    0.523      0.368    0.760     0.640     non_lethal
                 0.062    0.003    0.921      0.062    0.117      0.189    0.690     0.582     somewhat_lethal
                 1.000    0.640    0.439      1.000    0.610      0.397    0.867     0.757     very_lethal
Weighted Avg.    0.498    0.251    0.674      0.498    0.416      0.318    0.772     0.660

=== Confusion Matrix ===

     a     b     c   <-- classified as
  4531    56  5898 |     a = non_lethal
  2310   653  7522 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

OneRAttributeEval

J48

```
Correctly Classified Instances       30360               96.5188 %
Incorrectly Classified Instances      1095                3.4812 %
Kappa statistic                          0.9478
Mean absolute error                      0.0288
Root mean squared error                  0.1322
Relative absolute error                  6.4713 %
Root relative squared error             28.0387 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.899    0.002    0.996      0.899    0.945      0.922    0.988     0.987     non_lethal
                 0.997    0.049    0.911      0.997    0.952      0.928    0.994     0.978     somewhat_lethal
                 1.000    0.002    0.996      1.000    0.998      0.997    1.000     1.000     very_lethal
Weighted Avg.    0.965    0.017    0.968      0.965    0.965      0.949    0.994     0.988

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9426  1020    39 |     a = non_lethal
    36 10449     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Naive Bayes

```
Correctly Classified Instances       25677               81.6309 %
Incorrectly Classified Instances      5778               18.3691 %
Kappa statistic                          0.7245
Mean absolute error                      0.1398
Root mean squared error                  0.3064
Relative absolute error                 31.4481 %
Root relative squared error             64.9907 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.921    0.186    0.712      0.921    0.803      0.699    0.943     0.905     non_lethal
                 0.528    0.013    0.954      0.528    0.680      0.626    0.945     0.904     somewhat_lethal
                 1.000    0.076    0.867      1.000    0.929      0.895    0.993     0.980     very_lethal
Weighted Avg.    0.816    0.092    0.844      0.816    0.804      0.740    0.960     0.930

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9653   268   564 |     a = non_lethal
  3906  5539  1040 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

OneRAttributeEval

OneR

```
Correctly Classified Instances       30174               95.9275 %
Incorrectly Classified Instances      1281                4.0725 %
Kappa statistic                          0.9389
Mean absolute error                      0.0271
Root mean squared error                  0.1648
Relative absolute error                  6.1087 %
Root relative squared error             34.9535 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.000    1.000      0.878   0.935      0.910    0.939     0.919     non_lethal
                 1.000    0.060    0.893      1.000   0.943      0.916    0.970     0.893     somewhat_lethal
                 1.000    0.001    0.998      1.000   0.999      0.998    0.999     0.998     very_lethal
Weighted Avg.    0.959    0.020    0.964      0.959   0.959      0.941    0.969     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9204  1260    21 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Decision Table

```
Correctly Classified Instances       30180               95.9466 %
Incorrectly Classified Instances      1275                4.0534 %
Kappa statistic                          0.9392
Mean absolute error                      0.0652
Root mean squared error                  0.1493
Relative absolute error                 14.6614 %
Root relative squared error             31.6714 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884    0.003    0.994      0.884   0.936      0.910    0.989     0.984     non_lethal
                 0.995    0.058    0.896      0.995   0.943      0.915    0.990     0.972     somewhat_lethal
                 1.000    0.000    0.999      1.000   1.000      0.999    1.000     0.999     very_lethal
Weighted Avg.    0.959    0.020    0.963      0.959   0.959      0.941    0.993     0.985

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9265  1213     7 |     a = non_lethal
    55 10430     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Non-WEKA Approach

J48

```
Correctly Classified Instances       30389               96.611  %
Incorrectly Classified Instances      1066                3.389  %
Kappa statistic                          0.9492
Mean absolute error                      0.0278
Root mean squared error                  0.1303
Relative absolute error                  6.2521 %
Root relative squared error             27.6335 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.899    0.000    0.999      0.899    0.946      0.925    0.988     0.988     non_lethal
                 0.999    0.049    0.911      0.999    0.953      0.931    0.994     0.979     somewhat_lethal
                 1.000    0.002    0.996      1.000    0.998      0.997    1.000     1.000     very_lethal
Weighted Avg.    0.966    0.017    0.969      0.966    0.966      0.951    0.994     0.989

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9428  1018    39 |     a = non_lethal
     9 10476     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Naive Bayes

```
Correctly Classified Instances       25851               82.1841 %
Incorrectly Classified Instances      5604               17.8159 %
Kappa statistic                          0.7328
Mean absolute error                      0.1323
Root mean squared error                  0.2991
Relative absolute error                 29.771  %
Root relative squared error             63.4448 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.940    0.196    0.706      0.940    0.806      0.706    0.953     0.923     non_lethal
                 0.525    0.012    0.955      0.525    0.678      0.625    0.949     0.909     somewhat_lethal
                 1.000    0.059    0.894      1.000    0.944      0.917    0.993     0.975     very_lethal
Weighted Avg.    0.822    0.089    0.852      0.822    0.809      0.749    0.965     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9857   261   367 |     a = non_lethal
  4106  5509   870 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Non-WEKA Approach

OneR

```
Correctly Classified Instances       30174               95.9275 %
Incorrectly Classified Instances      1281                4.0725 %
Kappa statistic                          0.9389
Mean absolute error                      0.0271
Root mean squared error                  0.1648
Relative absolute error                  6.1087 %
Root relative squared error             34.9535 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.878    0.000    1.000      0.878    0.935      0.910    0.939     0.919     non_lethal
                 1.000    0.060    0.893      1.000    0.943      0.916    0.970     0.893     somewhat_lethal
                 1.000    0.001    0.998      1.000    0.999      0.998    0.999     0.998     very_lethal
Weighted Avg.    0.959    0.020    0.964      0.959    0.959      0.941    0.969     0.936

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9204  1260    21 |     a = non_lethal
     0 10485     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Decision Table

```
Correctly Classified Instances       30180               95.9466 %
Incorrectly Classified Instances      1275                4.0534 %
Kappa statistic                          0.9392
Mean absolute error                      0.0652
Root mean squared error                  0.1493
Relative absolute error                 14.6614 %
Root relative squared error             31.6714 %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884    0.003    0.994      0.884    0.936      0.910    0.989     0.984     non_lethal
                 0.995    0.058    0.896      0.995    0.943      0.915    0.990     0.972     somewhat_lethal
                 1.000    0.000    0.999      1.000    1.000      0.999    1.000     0.999     very_lethal
Weighted Avg.    0.959    0.020    0.963      0.959    0.959      0.941    0.993     0.985

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9265  1213     7 |     a = non_lethal
    55 10430     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Results and Analysis

Accuracy

	Gain Ratio	**Info Gain**	Cfs Subset w/ Greedy Stepwise	**OneR**	**Personal**
J48	93.51%	95.11%	97.27%	96.52%	96.61%
Naive Bayes	67.82%	79.93%	49.81%	81.63%	82.18%
OneR	95.93%	95.93%	95.93%	95.93%	95.93%
Decision Table	95.36%	95.36%	49.81%	95.95%	95.95%

Results and Analysis

RMS-Error

	Gain Ratio	Info Gain	Cfs Subset w/ Greedy Stepwise	OneR	**Personal**
J48	0.162	0.138	0.125	0.132	0.130
Naive Bayes	0.314	0.314	0.542	0.306	0.299
OneR	0.165	0.165	0.165	0.165	0.165
Decision Table	0.156	0.156	0.542	0.149	0.149

Results and Analysis

Selected model: **CfsSubsetEval with J48**

Highest accuracy: **CfsSubsetEval with J48 -** **0.972691**

Lowest root mean squared error: **CfsSubsetEval with J48 -** **0.1246**

Highest TP Rate: **CfsSubsetEval with J48 -** **0.973**

Lowest FP Rate: **CfsSubsetEval with J48 -** **0.014**

Selected Attributes:
LATITUDE, LONGITUDE, NUMBER OF CYCLIST INJURED, CONTRIBUTING FACTOR VEHICLE 1

```
Correctly Classified Instances       30596               97.2691 %
Incorrectly Classified Instances       859                2.7309 %
Kappa statistic                          0.959
Mean absolute error                      0.025
Root mean squared error                  0.1246
Relative absolute error                  5.614  %
Root relative squared error             26.433  %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.920    0.001     0.998      0.920     0.957      0.939    0.982     0.981     non_lethal
               0.998    0.039     0.927      0.998     0.961      0.943    0.989     0.962     somewhat_lethal
               1.000    0.001     0.998      1.000     0.999      0.999    1.000     0.999     very_lethal
Weighted Avg.  0.973    0.014     0.974      0.973     0.973      0.960    0.990     0.980

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9646   819    20 |     a = non_lethal
    20 10465     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```


```
Correctly Classified Instances       30596               97.2691 %
Incorrectly Classified Instances       859                2.7309 %
Kappa statistic                          0.959
Mean absolute error                      0.025
Root mean squared error                  0.1246
Relative absolute error                  5.614  %
Root relative squared error             26.433  %
Total Number of Instances            31455

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.920    0.001     0.998      0.920     0.957      0.939    0.982     0.981     non_lethal
               0.998    0.039     0.927      0.998     0.961      0.943    0.989     0.962     somewhat_lethal
               1.000    0.001     0.998      1.000     0.999      0.999    1.000     0.999     very_lethal
Weighted Avg.  0.973    0.014     0.974      0.973     0.973      0.960    0.990     0.980

=== Confusion Matrix ===

     a     b     c   <-- classified as
  9646   819    20 |     a = non_lethal
    20 10465     0 |     b = somewhat_lethal
     0     0 10485 |     c = very_lethal
```

Conclusion and Further Steps

From our analysis above, we concluded that the CfsSubsetEval attribute selection algorithm combined with the J48 classifier model was most accurate in predicting the lethality of car crashes in New York City Boroughs.

In future projects the dataset could be used to focus on specific attributes, such as which streets or boroughs seem to be the most deadly or most prone to car crashes with injuries. Additionally, applying area-specific analyses could provide deeper insights. For example, identifying patterns in high-lethality zones across different boroughs or understanding factors contributing to higher crash rates on particular streets could lead to more targeted interventions, such as better traffic management, road design improvements, or stricter enforcement in high-risk areas.

Sources

"City of New York - Motor Vehicle Collisions - Crashes." Catalog, Publisher data.cityofnewyork.us, 19 Oct. 2024, catalog.data.gov/dataset/motor-vehicle-collisions-crashes.

Khanna, Nilima. "J48 Classification (C4.5 Algorithm) in a Nutshell." Medium, Medium, 18 Aug. 2021, medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e.

Zhang, Zixuan. "Naive Bayes Explained." Medium, Towards Data Science, 14 Aug. 2019, towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0.

Finding the Data & Building Proposal: Aaryan

Preprocessing Initial Attempt: Aaryan

Preprocessing & Project Update: Aaryan

Non-Weka Attribute Selection Algorithm: Sami

Attribute Selection Algorithms and Classifiers: Sami

Results Output: Sami

Results Analysis: Sami

Building Final Report: Aaryan and Sami