Machine Learning Quarter 1 Project: Vehicle Damage Model

Ryan Ghimire and Aniketh Luchmapurkar

Thomas Jefferson High School for Science and Technology

Machine Learning 1

Dr. Yilmaz

September 22, 2024

Table of Contents

Machine Learning Quarter 1 Project: Vehicle Damage Model

Part 1: Project Statement

This project aims to utilize the Automated Crash Reporting System of the Maryland State Police to create a model that predicts the extent of damage caused by a traffic collision. The model aims to analyze various factors contributing to the damage of a vehicle after an accident. Specifically, we aim for the model to be able to classify the type of damage the vehicle involved in the collision sustained (Superficial, Functional, Disabling, Destroyed, and No Damage). As there are more than two possible outputs possible, this would be considered a multi-classification problem. In a real-world scenario, emergency services could use our model to predict the severity of vehicle damage before arriving at the scene of a traffic collision, allowing for them to prioritize more critical accidents. Along with this, our model could locate intersections or roads that are more prone to accidents, prompting authorities to implement better road designs, more traffic signals, or speed regulations at those locations.

Part 2: Dataset Description

The Montgomery County Police Department compiled a dataset of car accidents, with each instance being one accident. We are using the CSV format version of the dataset (it is called "Crash Reporting - Drivers Data"). A link to the dataset is present in the references section. There are 184,898 instances, 38 features (hence 38 dimensions), and one class attribute.

The first three features are for logistical purposes and keeping track of each separate incident (ASRC report number, local agency report number, and investigating local agency). The other features that come later are useful for a multi-classification problem, which we are doing. Most of these features are considered categorical data (with the key exceptions being IDs, report or case numbers, latitude, longitude, and location). We chose the "Vehicle Damage Extent" to be

our class attribute as it best corresponds to our research goal of analyzing damage caused by a traffic collision. The possible values of this attribute which we are to classify with our model are - as aforementioned - Superficial, Functional, Disabling, Destroyed, and No Damage. The full list of features and the class are as follows:

1. Report Number
    a. Automated Crash Report System (ACRS) report number of incident
2. Local Case Number
    a. Local investigating agency case number for incident
3. Agency Name
    a. Local investigating agency name
4. ACRS Report Type
    a. Classifies crash as property, injury, or fatal
5. Crash Date/Time
    a. Date and time of crash
6. Route Type
    a. Crash location’s roadway type
7. Road Name
    a. Crash location’s road name
8. Cross-Street Name
    a. Name of nearest street to crash location
9. Off-Road Description
    a. Description of location for off-road collisions

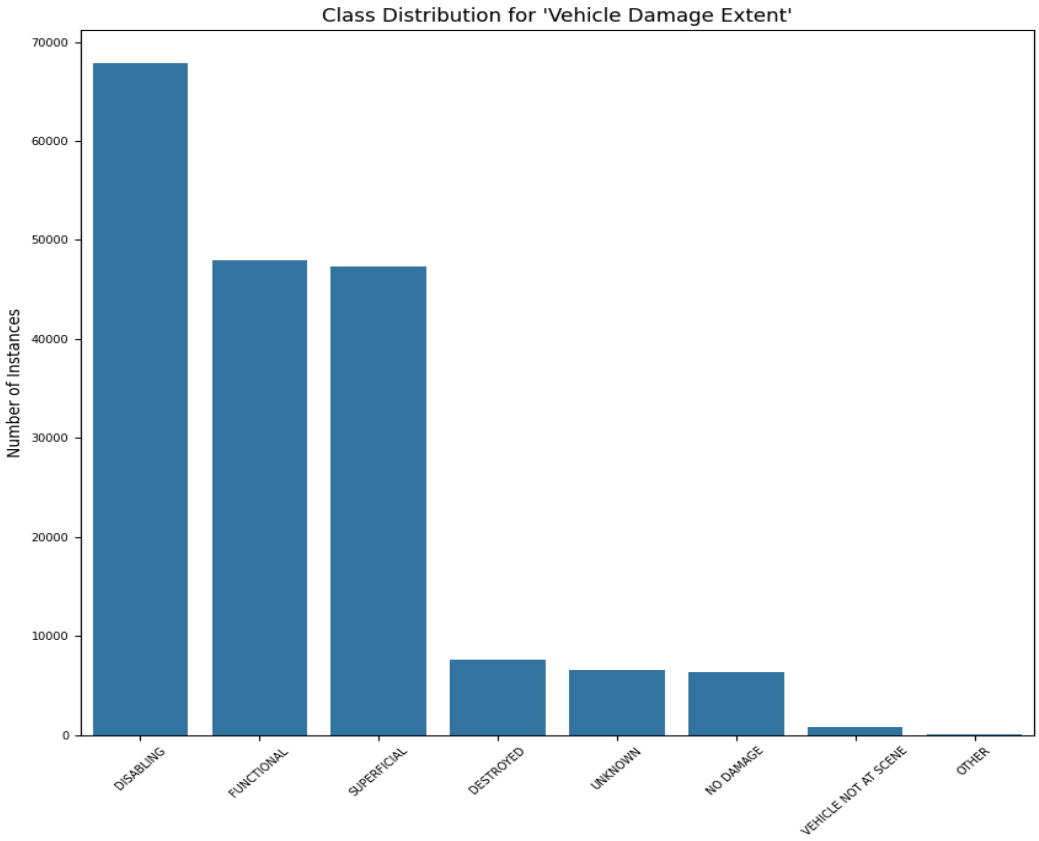
10. Municipality
    a. Crash location's jurisdiction
11. Related Non-Motorist
    a. Types of (if any) non motorists involved
12. Weather
    a. Weather at crash location
13. Surface Condition
    a. Roadway condition at crash location
14. Light
    a. Lighting conditions at crash location
15. Traffic Control
    a. Traffic control devices (ex: stop sign or traffic light) at crash location
16. Driver Substance Abuse
    a. Whether driver abusing substances
17. Non-Motorist Substance Abuse
    a. Whether bystander of collision abusing substances
18. Person ID
    a. Reporting party's ID
19. Driver At Fault
    a. Whether driver at fault for collision
20. Injury Severity
    a. Severity of injury to driver

21. Circumstance
    a. Special circumstances for this collision
22. Driver Distracted By
    a. Reason for distracted driving
23. Drivers License State
    a. State that issued driver's license
24. Vehicle ID
    a. Crashed vehicle license plate
25. Vehicle Damage Extent
    a. Severity of damage to vehicle
    b. Class attribute
26. Vehicle First Impact Location
    a. Location of first impact on vehicle
27. Vehicle Body Type
    a. Whether passenger car, SUV, truck, etc
28. Vehicle Movement
    a. Whether parked, reversing, driving forward, turning, etc.
29. Vehicle Going Dir
    a. Cardinal direction of vehicle movement before collision
30. Speed Limit
    a. Posted speed limit of roadway at collision
31. Driverless Vehicle
    a. Whether vehicle autonomous/driverless or not

32. Parked Vehicle
    a. Whether vehicle parked before collision
33. Vehicle Year
    a. Model year of vehicle (ex: 2024)
34. Vehicle Make
    a. Brand/make of vehicle (ex: Toyota)
35. Vehicle Model
    a. Model of vehicle (ex: Corolla)
36. Latitude
    a. Latitude of accident
37. Longitude
    a. Longitude of accident
38. Location
    a. (Latitude, Longitude) of accident

Figure 1

Histogram of output frequencies



Note. Histogram depicting frequency of each class output

As seen in Figure 1 above, the class distribution is significantly skewed, with the majority of crashes disabling the vehicle. The data has a skewness of 0.826. This is likely due to the fact that most traffic collisions where law enforcement is called are major enough to disable the vehicle; very minor accidents would not require having to call law enforcement in the first place. The specific frequencies of each output are below in Table 1.

Table 1

Table of output frequencies

Class Output	Frequency
DISABLING	67,822
FUNCTIONAL	47,906
SUPERFICIAL	47,297
DESTROYED	7,610
UNKNOWN	6,620
NO DAMAGE	6,378
VEHICLE NOT AT SCENE	846
OTHER	102
Blank/Missing	316

Note. Frequency of occurrences of each class output

As seen in Table 1 above, the class attribute has 316 blank values as well as 6620 values labeled as “UNKNOWN,” 846 labeled as “VEHICLE NOT AT SCENE,” and 102 labeled as “OTHER;” all of these values are considered missing values and have to be dealt with when preprocessing the data. Shown in Table 2, 35 out of the 37 other attributes also have missing values and will have to be dealt with the same way as the missing values in the class attribute.

Table 2

Table of missing values

Attribute	Number of Missing Values
Report Number	184897
Local Case Number	184829
Agency Name	0

ACRS Report Type	0
Crash Date/Time	184897
Route Type	18161
Road Name	19364
Cross-Street Name	28452
Off-Road Description	178336
Municipality	165771
Related Non-Motorist	179002
Collision Type	1388
Weather	14133
Surface Condition	21848
Light	3826
Traffic Control	26879
Driver Substance Abuse	44756
Non-Motorist Substance Abuse	180420
Person ID	158038
Driver At Fault	4686
Injury Severity	789
Circumstance	165594
Driver Distracted By	36683
Drivers License State	11432
Vehicle ID	158050
Vehicle First Impact Location	3238
Vehicle Body Type	9139
Vehicle Movement	4133

Note. Number of missing for each attribute

Part 3: Preprocessing

Firstly, we realized that our dataset had some formatting issues so we created a Python script (“cleaner.py”) to get rid of newline characters and “\” characters. This script, along with all others mentioned in this document, are located in the Google Drive folder for reference.

Attributes related to IDs and investigating agencies have no importance in classifying damage to vehicles. As such, we would get rid of these attributes entirely (in other words, we would delete that “column”). Furthermore, we delete the location attribute as it is derivable from latitude and longitude.

We also realized that a model trained on the crash date/time would not be able to gain much relevant, unique information from said attribute; all of the impactful information it could gain (i.e. time of day) is already accounted for in other attributes (i.e. light). In other words, it is yet another derivable attribute that can be removed.

Additionally, we would need to get rid of the instances with blank or missing attributes. We would also need to get rid of instances with outputs that are blank, Other, Unknown, or Vehicle not at Scene. These last three deletion steps were done by use of the "base_preprocessing.py" Python script.

Along with this, the "NO DAMAGE" label had extra quotes around it which we had to get rid of so it matched the other attributes, which we did using the fixNoDamage.py.

As most, if not all, of the data in our dataset is categorical, we would need to change them to numeric/quantitative forms. Most are True/False type data, meaning it can be easily changed to 1 and 0, respectively. The others which have several options will have to be converted to a 0-[number of options available] scale. This was done using the "encoder.py" script.

All these different scales (0-1 and 0-[number of options available]) would give unequal importance to different attributes, and as such, we would be required to normalize our data. This ensures a balanced importance for all available features. The specific normalization process we would do is z-score normalization as that is able to handle outliers better than other normalization methods and takes into consideration the mean and standard deviation to better maintain the overall structure of the dataset after normalization. In Weka, this is done using the "Standardize" filter.

Lastly, we validated the data by using k-fold cross validation with 10 folds, but also created an 85-15 train-test split to see if both methods produced similar accuracies.

Part 4: Attribute Selection

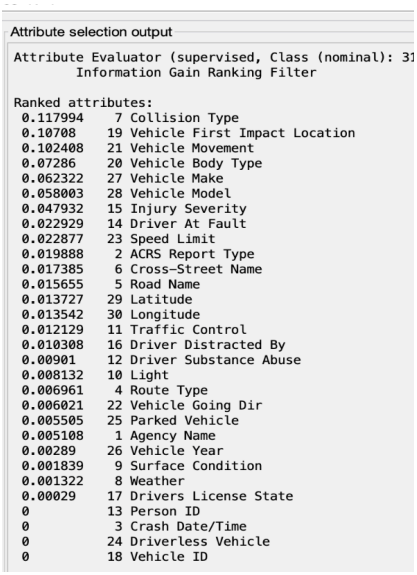
We used 4 attribute selection methods in Weka as well as manual analysis based attribute evaluation to choose which attributes to keep and remove from the dataset:

1. **Non-Weka:**

We analyzed which features would have no effect on the class attribute in order to choose which attributes to remove. Per our analysis, we realized that there wasn't a single crash with a driverless vehicle, so a model trained with the driverless feature would be able to learn nothing about the impact of a driverless vehicle on the vehicle damage extent. For these reasons, we removed the driverless vehicle attributes.

2. **Information Gain:**

We used Weka for this approach. Information Gain is an attribute selection algorithm that utilizes entropy (the randomness in data) to determine which attributes are most impactful on the class label. Specifically, the entropy of the cumulative dataset (all attributes) is calculated and then entropies of each attribute is also calculated. The entropies are subtracted, and the attributes that have a large (relative to other attributes) difference are considered important. Each difference is known as an Information Gain. For our dataset, we used a cutoff value of 0.008.



```
Attribute selection output

Attribute Evaluator (supervised, Class (nominal): 31
        Information Gain Ranking Filter

Ranked attributes:
 0.117994    7 Collision Type
 0.10708    19 Vehicle First Impact Location
 0.102408   21 Vehicle Movement
 0.07286    20 Vehicle Body Type
 0.062322   27 Vehicle Make
 0.058003   28 Vehicle Model
 0.047932   15 Injury Severity
 0.022929   14 Driver At Fault
 0.022877   23 Speed Limit
 0.019888    2 ACRS Report Type
 0.017385    6 Cross-Street Name
 0.015655    5 Road Name
 0.013727   29 Latitude
 0.013542   30 Longitude
 0.012129   11 Traffic Control
 0.010308   16 Driver Distracted By
 0.00901    12 Driver Substance Abuse
 0.008132   10 Light
 0.006961    4 Route Type
 0.006021   22 Vehicle Going Dir
 0.005505   25 Parked Vehicle
 0.005108    1 Agency Name
 0.00289    26 Vehicle Year
 0.001839    9 Surface Condition
 0.001322    8 Weather
 0.00029    17 Drivers License State
 0          13 Person ID
 0           3 Crash Date/Time
 0          24 Driverless Vehicle
 0          18 Vehicle ID
```

Note. Attributes ranked

3. Gain Ratio:

We used Weka for this approach. Gain Ratio calculates the information gain of each attribute in the dataset; however, to account for the bias caused by an attribute having many distinct values which normally unintentionally increases the information gain, Gain Ratio divides the information gain by the split information, which is a measure of the complexity of an attribute, to create a less biased output. For our dataset, we used a cutoff value of 0.005.

```
Attribute selection output

Attribute Evaluator (supervised, Class (nominal): 31 Vehicle Damage Extent):
        Gain Ratio feature evaluator

Ranked attributes:
 0.06031   25 Parked Vehicle
 0.0502    15 Injury Severity
 0.03963    7 Collision Type
 0.03941   20 Vehicle Body Type
 0.03646   21 Vehicle Movement
 0.03438   19 Vehicle First Impact Location
 0.02331   14 Driver At Fault
 0.02047    2 ACRS Report Type
 0.01824   27 Vehicle Make
 0.01743   12 Driver Substance Abuse
 0.01364   28 Vehicle Model
 0.01102   16 Driver Distracted By
 0.00938   23 Speed Limit
 0.00752   11 Traffic Control
 0.00627    6 Cross-Street Name
 0.00609   10 Light
 0.00538   29 Latitude
 0.00527   30 Longitude
 0.00426    5 Road Name
 0.00409    1 Agency Name
 0.00354    4 Route Type
 0.00274    9 Surface Condition
 0.00263   22 Vehicle Going Dir
 0.00155   26 Vehicle Year
 0.00129   17 Drivers License State
 0.0012     8 Weather
 0          3 Crash Date/Time
 0         13 Person ID
 0         18 Vehicle ID
 0         24 Driverless Vehicle
```

Note. Attributes ranked

4. **Correlation:**

We used Weka for this approach. The Correlation Ranking Filter utilizes the Pearson correlation with the class label to rank the attributes. The Pearson correlation is calculated through the following formula:

$$\textbf{Pearson Correlation = r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

For our dataset, we used a cutoff value of 0.0255.

Attribute selection output

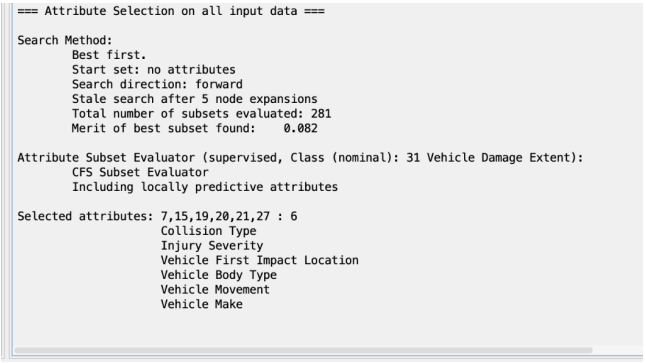
```
Attribute Evaluator (supervised, Class (nominal): 31 Vehicle Damage Extent):
        Correlation Ranking Filter
Ranked attributes:
 0.138985   21 Vehicle Movement
 0.126905   15 Injury Severity
 0.114866   14 Driver At Fault
 0.069776    2 ACRS Report Type
 0.064673   29 Latitude
 0.06223    23 Speed Limit
 0.057104   19 Vehicle First Impact Location
 0.048909   16 Driver Distracted By
 0.048255   25 Parked Vehicle
 0.045083   30 Longitude
 0.042913   10 Light
 0.033347   28 Vehicle Model
 0.026859    9 Surface Condition
 0.025994   27 Vehicle Make
 0.02516    11 Traffic Control
 0.025108   20 Vehicle Body Type
 0.024892    4 Route Type
 0.023447   12 Driver Substance Abuse
 0.019733    7 Collision Type
 0.017689    8 Weather
 0.011372    5 Road Name
 0.009428    1 Agency Name
 0.005355   26 Vehicle Year
 0.004213   22 Vehicle Going Dir
 0.002725    3 Crash Date/Time
 0.00214     6 Cross-Street Name
 0.001961   17 Drivers License State
 0.001354   13 Person ID
 0.000711   18 Vehicle ID
 0          24 Driverless Vehicle
```

Note. Attributes ranked

5. **Correlation-based Feature Selection:**

We used Weka for this approach. The Correlation-based Feature Selection (CFS) utilizes the same Pearson correlation of a feature to the class label as the previous Correlation ranking filter above. However, the distinction between the Correlation ranking filter and CFS is that CFS also analyzes the Pearson correlation between different attributes, prioritizing those with low

correlation. It will then select the attributes with the highest correlation with the label and lowest correlations with each other.



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

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 281
        Merit of best subset found:    0.082

Attribute Subset Evaluator (supervised, Class (nominal): 31 Vehicle Damage Extent):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 7,15,19,20,21,27 : 6
                     Collision Type
                     Injury Severity
                     Vehicle First Impact Location
                     Vehicle Body Type
                     Vehicle Movement
                     Vehicle Make
```

Note. Attributes selected

Part 5: Training and Analysis

Using each of the subsets created by the 5 attribute selection methods above, we ran 4 classification models and tested their accuracy:

1. **J48:**

   A decision tree algorithm that makes predictions by splitting the data into smaller groups based on key features. It works well with qualitative or discrete data (e.g., weather or road conditions) as it identifies patterns in vehicle damage and understands what features matter most, like how specific conditions increase the chance of severe damage.

a. Non-Weka:

```
    10279 5:'NO DAMAGE' 5:NO DAMAGE       0.923

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

Correctly Classified Instances       46161               44.9063 %
Incorrectly Classified Instances     56633               55.0937 %
Kappa statistic                          0.194
Mean absolute error                      0.2313
Root mean squared error                  0.4359
Relative absolute error                 83.4064 %
Root relative squared error            117.0631 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.611    0.323    0.560      0.611    0.584      0.285    0.651     0.503     DISABLING
                 0.412    0.199    0.410      0.412    0.411      0.213    0.611     0.340     SUPERFICIAL
                 0.328    0.242    0.347      0.328    0.337      0.087    0.538     0.314     FUNCTIONAL
                 0.094    0.022    0.147      0.094    0.114      0.089    0.659     0.079     DESTROYED
                 0.142    0.014    0.210      0.142    0.169      0.155    0.623     0.070     NO DAMAGE
Weighted Avg.    0.449    0.249    0.437      0.449    0.442      0.200    0.609     0.381

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 25240  5630  8726  1498   241 |     a = DISABLING
  6343 10674  7891   227   761 |     b = SUPERFICIAL
 10382  8253  9498   399   409 |     c = FUNCTIONAL
  2621   339   609   370    16 |     d = DESTROYED
   487  1107   671    23   379 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.62 seconds

=== Summary ===

Correctly Classified Instances        8160               44.9835 %
Incorrectly Classified Instances      9980               55.0165 %
Kappa statistic                          0.1965
Mean absolute error                      0.2315
Root mean squared error                  0.4361
Relative absolute error                 83.3733 %
Root relative squared error            116.978  %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.616    0.320    0.560      0.616    0.587      0.292    0.652     0.502     DISABLING
                 0.407    0.197    0.415      0.407    0.411      0.211    0.604     0.340     SUPERFICIAL
                 0.331    0.244    0.346      0.331    0.338      0.088    0.537     0.314     FUNCTIONAL
                 0.110    0.021    0.181      0.110    0.137      0.114    0.674     0.093     DESTROYED
                 0.132    0.016    0.176      0.132    0.151      0.134    0.609     0.059     NO DAMAGE
Weighted Avg.    0.450    0.248    0.438      0.450    0.443      0.203    0.607     0.380

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4447  956 1518  249   54 |    a = DISABLING
 1137 1889 1427   41  147 |    b = SUPERFICIAL
 1813 1453 1684   67   77 |    c = FUNCTIONAL
  465   59  120   80    2 |    d = DESTROYED
   78  197  115    5   60 |    e = NO DAMAGE
```

b. Information Gain:

ifier

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

Correctly Classified Instances       45682               44.4403 %
Incorrectly Classified Instances     57112               55.5597 %
Kappa statistic                          0.1882
Mean absolute error                      0.2322
Root mean squared error                  0.4394
Relative absolute error                 83.728  %
Root relative squared error            117.996  %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.323    0.556      0.600    0.577      0.275    0.646     0.499     DISABLING
                 0.410    0.201    0.407      0.410    0.409      0.209    0.610     0.336     SUPERFICIAL
                 0.327    0.246    0.343      0.327    0.335      0.083    0.535     0.312     FUNCTIONAL
                 0.099    0.023    0.148      0.099    0.119      0.093    0.652     0.077     DESTROYED
                 0.148    0.014    0.216      0.148    0.175      0.161    0.622     0.071     NO DAMAGE
Weighted Avg.    0.444    0.251    0.434      0.444    0.438      0.194    0.605     0.378

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 24805  5728  8970  1592   240 |     a = DISABLING
  6386 10615  7902   211   782 |     b = SUPERFICIAL
 10368  8269  9475   432   397 |     c = FUNCTIONAL
  2575   343   630   393    14 |     d = DESTROYED
   495  1111   640    27   394 |     e = NO DAMAGE
```

er

er

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.6 seconds

=== Summary ===

Correctly Classified Instances        8045               44.3495 %
Incorrectly Classified Instances     10095               55.6505 %
Kappa statistic                          0.1873
Mean absolute error                      0.2323
Root mean squared error                  0.4401
Relative absolute error                 83.6485 %
Root relative squared error            118.0415 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.609    0.325    0.554      0.609    0.580      0.280    0.647     0.498     DISABLING
                 0.401    0.200    0.408      0.401    0.405      0.202    0.600     0.335     SUPERFICIAL
                 0.322    0.246    0.338      0.322    0.330      0.077    0.533     0.312     FUNCTIONAL
                 0.120    0.021    0.190      0.120    0.147      0.123    0.671     0.095     DESTROYED
                 0.127    0.015    0.178      0.127    0.149      0.132    0.605     0.058     NO DAMAGE
Weighted Avg.    0.443    0.251    0.432      0.443    0.437      0.193    0.603     0.377

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4397  995 1534  247   51 |    a = DISABLING
 1146 1861 1442   47  145 |    b = SUPERFICIAL
 1857 1450 1642   74   71 |    c = FUNCTIONAL
  456   64  118   87    1 |    d = DESTROYED
   85  188  120    4   58 |    e = NO DAMAGE
```

c. Gain Ratio:

```
Time taken to build model: 24.83 seconds

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

Correctly Classified Instances       48536               47.2168 %
Incorrectly Classified Instances     54258               52.7832 %
Kappa statistic                          0.2215
Mean absolute error                      0.2311
Root mean squared error                  0.4145
Relative absolute error                 83.3147 %
Root relative squared error            111.3137 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.663    0.328    0.576      0.663    0.616      0.329    0.670     0.517     DISABLING
                 0.425    0.191    0.429      0.425    0.427      0.235    0.626     0.356     SUPERFICIAL
                 0.326    0.227    0.360      0.326    0.342      0.102    0.546     0.319     FUNCTIONAL
                 0.077    0.015    0.175      0.077    0.107      0.093    0.701     0.094     DESTROYED
                 0.137    0.012    0.235      0.137    0.173      0.163    0.630     0.082     NO DAMAGE
Weighted Avg.    0.472    0.245    0.454      0.472    0.460      0.228    0.624     0.393

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 27404  5044  7731   942   214 |     a = DISABLING
  6117 11013  7961   166   639 |     b = SUPERFICIAL
 10725  8131  9449   307   329 |     c = FUNCTIONAL
  2839   295   510   304     7 |     d = DESTROYED
   493  1195   591    22   366 |     e = NO DAMAGE
```

```
Time taken to test model on supplied test set: 0.63 seconds

=== Summary ===

Correctly Classified Instances        8638               47.6185 %
Incorrectly Classified Instances      9502               52.3815 %
Kappa statistic                          0.2283
Mean absolute error                      0.2296
Root mean squared error                  0.4129
Relative absolute error                 82.6851 %
Root relative squared error            110.7643 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.668    0.329    0.573      0.668    0.617      0.333    0.674     0.519     DISABLING
                 0.441    0.195    0.437      0.441    0.439      0.245    0.631     0.365     SUPERFICIAL
                 0.323    0.217    0.367      0.323    0.343      0.110    0.555     0.324     FUNCTIONAL
                 0.087    0.015    0.194      0.087    0.120      0.106    0.707     0.103     DESTROYED
                 0.125    0.010    0.239      0.125    0.165      0.158    0.613     0.072     NO DAMAGE
Weighted Avg.    0.476    0.243    0.457      0.476    0.464      0.234    0.629     0.397

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4828  890 1302  179   25 |    a = DISABLING
 1136 2047 1325   26  107 |    b = SUPERFICIAL
 1869 1478 1643   56   48 |    c = FUNCTIONAL
  496   65  101   63    1 |    d = DESTROYED
   92  199  106    1   57 |    e = NO DAMAGE
```

d. Correlation:

```
Time taken to build model: 20.85 seconds

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

Correctly Classified Instances       47046               45.7673 %
Incorrectly Classified Instances     55748               54.2327 %
Kappa statistic                          0.1932
Mean absolute error                      0.2407
Root mean squared error                  0.4112
Relative absolute error                 86.7861 %
Root relative squared error            110.4092 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.669    0.367    0.551      0.669    0.604      0.296    0.652     0.501     DISABLING
                 0.382    0.193    0.399      0.382    0.390      0.191    0.609     0.341     SUPERFICIAL
                 0.318    0.221    0.360      0.318    0.338      0.101    0.548     0.319     FUNCTIONAL
                 0.037    0.010    0.132      0.037    0.058      0.051    0.698     0.084     DESTROYED
                 0.061    0.010    0.142      0.061    0.086      0.078    0.561     0.050     NO DAMAGE
Weighted Avg.    0.458    0.259    0.432      0.458    0.441      0.199    0.611     0.381

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 27646  5585  7283   592   229 |     a = DISABLING
  7586  9882  7857   139   432 |     b = SUPERFICIAL
 11323  7874  9207   222   315 |     c = FUNCTIONAL
  2981   338   476   147    13 |     d = DESTROYED
   682  1082   727    12   164 |     e = NO DAMAGE
```

```
Time taken to build model: 20.04 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.15 seconds

=== Summary ===

Correctly Classified Instances        8362               46.097  %
Incorrectly Classified Instances      9778               53.903  %
Kappa statistic                          0.1985
Mean absolute error                      0.2409
Root mean squared error                  0.4103
Relative absolute error                 86.7571 %
Root relative squared error            110.0581 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.676    0.370    0.548      0.676    0.605      0.300    0.649     0.496     DISABLING
                 0.394    0.193    0.413      0.394    0.403      0.205    0.617     0.349     SUPERFICIAL
                 0.315    0.216    0.362      0.315    0.337      0.103    0.545     0.319     FUNCTIONAL
                 0.025    0.007    0.125      0.025    0.041      0.039    0.705     0.093     DESTROYED
                 0.055    0.011    0.114      0.055    0.074      0.063    0.562     0.045     NO DAMAGE
Weighted Avg.    0.461    0.258    0.433      0.461    0.442      0.204    0.612     0.381

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4887  973 1243   72   49 |    a = DISABLING
 1345 1828 1374   18   76 |    b = SUPERFICIAL
 2001 1384 1604   36   69 |    c = FUNCTIONAL
  553   60   94   18    1 |    d = DESTROYED
  137  182  111    0   25 |    e = NO DAMAGE
```

e. Correlation-based Feature Selection:

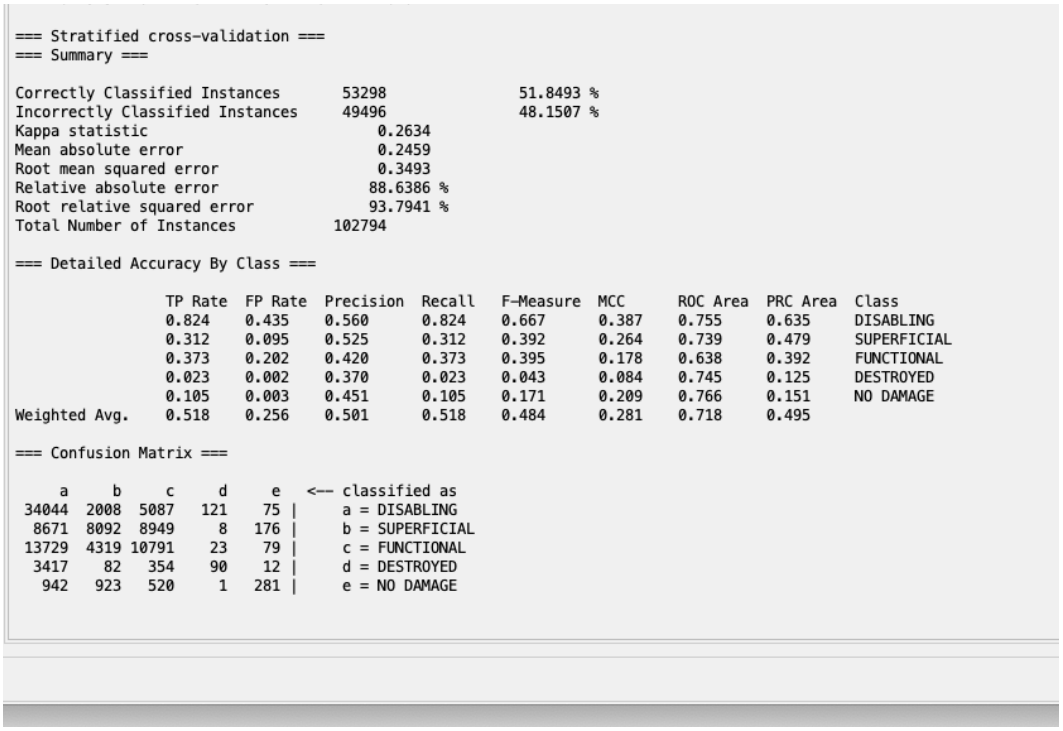
```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.6 seconds

=== Summary ===

Correctly Classified Instances        8407               46.3451 %
Incorrectly Classified Instances      9733               53.6549 %
Kappa statistic                          0.1895
Mean absolute error                      0.2587
Root mean squared error                  0.3622
Relative absolute error                 93.1565 %
Root relative squared error             97.1461 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.737    0.449    0.520      0.737    0.610      0.284    0.692     0.558     DISABLING
                 0.464    0.236    0.403      0.464    0.431      0.218    0.685     0.398     SUPERFICIAL
                 0.174    0.120    0.363      0.174    0.236      0.072    0.589     0.350     FUNCTIONAL
                 0.062    0.005    0.363      0.062    0.106      0.137    0.750     0.147     DESTROYED
                 0.000    0.000    ?          0.000    ?          ?        0.685     0.049     NO DAMAGE
Weighted Avg.    0.463    0.273    ?          0.463    ?          ?        0.663     0.430

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5321 1089  748   66    0 |    a = DISABLING
 1800 2153  688    0    0 |    b = SUPERFICIAL
 2346 1848  888   12    0 |    c = FUNCTIONAL
  568   54   59   45    0 |    d = DESTROYED
  190  200   64    1    0 |    e = NO DAMAGE
```

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

Correctly Classified Instances       47426               46.1369 %
Incorrectly Classified Instances     55368               53.8631 %
Kappa statistic                          0.1846
Mean absolute error                      0.2583
Root mean squared error                  0.3618
Relative absolute error                 93.1221 %
Root relative squared error             97.1514 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.731    0.454    0.520      0.731    0.608      0.274    0.691     0.564     DISABLING
                 0.464    0.235    0.399      0.464    0.429      0.219    0.687     0.390     SUPERFICIAL
                 0.171    0.122    0.356      0.171    0.231      0.065    0.591     0.349     FUNCTIONAL
                 0.059    0.004    0.355      0.059    0.101      0.132    0.758     0.134     DESTROYED
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.047     NO DAMAGE
Weighted Avg.    0.461    0.276    ?          0.461    ?          ?        0.664     0.430

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 30209  6376  4400   350     0 |     a = DISABLING
  9915 12020  3939    22     0 |     b = SUPERFICIAL
 13728 10198  4963    52     0 |     c = FUNCTIONAL
  3176   272   273   234     0 |     d = DESTROYED
  1067  1223   376     1     0 |     e = NO DAMAGE
```

2. **Decision Table:**

Decision table is a type of classifier that builds a table of rules based on training data and tries to match new test data according to the "rules" established in the aforementioned table. These models are useful as they can account for non-linear relationships and don't require an incredibly large amount of data, but they also have some drawbacks: namely, they tend to overfit.

a. **Non-Weka:**



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

Correctly Classified Instances       53298               51.8493 %
Incorrectly Classified Instances     49496               48.1507 %
Kappa statistic                          0.2634
Mean absolute error                      0.2459
Root mean squared error                  0.3493
Relative absolute error                 88.6386 %
Root relative squared error             93.7941 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.824    0.435    0.560      0.824    0.667      0.387    0.755     0.635     DISABLING
                 0.312    0.095    0.525      0.312    0.392      0.264    0.739     0.479     SUPERFICIAL
                 0.373    0.202    0.420      0.373    0.395      0.178    0.638     0.392     FUNCTIONAL
                 0.023    0.002    0.370      0.023    0.043      0.084    0.745     0.125     DESTROYED
                 0.105    0.003    0.451      0.105    0.171      0.209    0.766     0.151     NO DAMAGE
Weighted Avg.    0.518    0.256    0.501      0.518    0.484      0.281    0.718     0.495

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 34044  2008  5087   121    75 |     a = DISABLING
  8671  8092  8949     8   176 |     b = SUPERFICIAL
 13729  4319 10791    23    79 |     c = FUNCTIONAL
  3417    82   354    90    12 |     d = DESTROYED
   942   923   520     1   281 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 2.52 seconds

=== Summary ===

Correctly Classified Instances       45635               53.9083 %
Incorrectly Classified Instances     39018               46.0917 %
Kappa statistic                          0.2974
Mean absolute error                      0.242
Root mean squared error                  0.3441
Relative absolute error                 87.2649 %
Root relative squared error             92.4013 %
Total Number of Instances            84653

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.831    0.409    0.577      0.831    0.681      0.418    0.774     0.654     DISABLING
                 0.342    0.092    0.556      0.342    0.424      0.301    0.761     0.505     SUPERFICIAL
                 0.403    0.199    0.444      0.403    0.422      0.210    0.671     0.419     FUNCTIONAL
                 0.038    0.001    0.571      0.038    0.070      0.139    0.769     0.162     DESTROYED
                 0.162    0.004    0.552      0.162    0.250      0.289    0.813     0.231     NO DAMAGE
Weighted Avg.    0.539    0.244    0.533      0.539    0.508      0.316    0.742     0.520

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 28225  1547  4068    79    66 |     a = DISABLING
  6580  7292  7280     1   142 |     b = SUPERFICIAL
 10765  3455  9640    11    70 |     c = FUNCTIONAL
  2726    66   298   121    12 |     d = DESTROYED
   660   750   442     0   357 |     e = NO DAMAGE
```

b. Information Gain:

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

Correctly Classified Instances       53298               51.8493 %
Incorrectly Classified Instances     49496               48.1507 %
Kappa statistic                          0.2634
Mean absolute error                      0.2459
Root mean squared error                  0.3493
Relative absolute error                 88.6386 %
Root relative squared error             93.7941 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.824    0.435    0.560      0.824    0.667      0.387    0.755     0.635     DISABLING
                 0.312    0.095    0.525      0.312    0.392      0.264    0.739     0.479     SUPERFICIAL
                 0.373    0.202    0.420      0.373    0.395      0.178    0.638     0.392     FUNCTIONAL
                 0.023    0.002    0.370      0.023    0.043      0.084    0.745     0.125     DESTROYED
                 0.105    0.003    0.451      0.105    0.171      0.209    0.766     0.151     NO DAMAGE
Weighted Avg.    0.518    0.256    0.501      0.518    0.484      0.281    0.718     0.495

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 34044  2008  5087   121    75 |     a = DISABLING
  8671  8092  8949     8   176 |     b = SUPERFICIAL
 13729  4319 10791    23    79 |     c = FUNCTIONAL
  3417    82   354    90    12 |     d = DESTROYED
   942   923   520     1   281 |     e = NO DAMAGE
```

```
=== Summary ===

Correctly Classified Instances        9378               51.6979 %
Incorrectly Classified Instances      8762               48.3021 %
Kappa statistic                          0.2641
Mean absolute error                      0.2453
Root mean squared error                  0.349
Relative absolute error                 88.3605 %
Root relative squared error             93.6208 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.429    0.559      0.821    0.665      0.390    0.759     0.637     DISABLING
                 0.306    0.095    0.526      0.306    0.387      0.259    0.741     0.479     SUPERFICIAL
                 0.384    0.208    0.419      0.384    0.400      0.181    0.642     0.390     FUNCTIONAL
                 0.025    0.001    0.462      0.025    0.047      0.100    0.734     0.134     DESTROYED
                 0.119    0.004    0.443      0.119    0.187      0.220    0.778     0.167     NO DAMAGE
Weighted Avg.    0.517    0.254    0.504      0.517    0.483      0.282    0.721     0.495

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5933  350  904   18   19 |    a = DISABLING
 1549 1419 1640    0   33 |    b = SUPERFICIAL
 2360  763 1954    3   14 |    c = FUNCTIONAL
  605   17   84   18    2 |    d = DESTROYED
  171  147   83    0   54 |    e = NO DAMAGE
```

c. Gain Ratio:

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

Correctly Classified Instances       53298               51.8493 %
Incorrectly Classified Instances     49496               48.1507 %
Kappa statistic                          0.2634
Mean absolute error                      0.2459
Root mean squared error                  0.3493
Relative absolute error                 88.6386 %
Root relative squared error             93.7941 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.824    0.435    0.560      0.824    0.667      0.387    0.755     0.635     DISABLING
                 0.312    0.095    0.525      0.312    0.392      0.264    0.739     0.479     SUPERFICIAL
                 0.373    0.202    0.420      0.373    0.395      0.178    0.638     0.392     FUNCTIONAL
                 0.023    0.002    0.370      0.023    0.043      0.084    0.745     0.125     DESTROYED
                 0.105    0.003    0.451      0.105    0.171      0.209    0.766     0.151     NO DAMAGE
Weighted Avg.    0.518    0.256    0.501      0.518    0.484      0.281    0.718     0.495

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 34044  2008  5087   121    75 |     a = DISABLING
  8671  8092  8949     8   176 |     b = SUPERFICIAL
 13729  4319 10791    23    79 |     c = FUNCTIONAL
  3417    82   354    90    12 |     d = DESTROYED
   942   923   520     1   281 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.51 seconds

=== Summary ===

Correctly Classified Instances        9378               51.6979 %
Incorrectly Classified Instances      8762               48.3021 %
Kappa statistic                          0.2641
Mean absolute error                      0.2453
Root mean squared error                  0.349
Relative absolute error                 88.3605 %
Root relative squared error             93.6208 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.429    0.559      0.821    0.665      0.390    0.759     0.637     DISABLING
                 0.306    0.095    0.526      0.306    0.387      0.259    0.741     0.479     SUPERFICIAL
                 0.384    0.208    0.419      0.384    0.400      0.181    0.642     0.390     FUNCTIONAL
                 0.025    0.001    0.462      0.025    0.047      0.100    0.734     0.134     DESTROYED
                 0.119    0.004    0.443      0.119    0.187      0.220    0.778     0.167     NO DAMAGE
Weighted Avg.    0.517    0.254    0.504      0.517    0.483      0.282    0.721     0.495

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5933  350  904   18   19 |    a = DISABLING
 1549 1419 1640    0   33 |    b = SUPERFICIAL
 2360  763 1954    3   14 |    c = FUNCTIONAL
  605   17   84   18    2 |    d = DESTROYED
  171  147   83    0   54 |    e = NO DAMAGE
```

d. Correlation:

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

Correctly Classified Instances       51635               50.2315 %
Incorrectly Classified Instances     51159               49.7685 %
Kappa statistic                          0.2278
Mean absolute error                      0.2525
Root mean squared error                  0.3545
Relative absolute error                 91.0244 %
Root relative squared error             95.2011 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.854    0.515    0.527      0.854    0.652      0.348    0.722     0.592     DISABLING
                 0.275    0.090    0.507      0.275    0.356      0.234    0.705     0.448     SUPERFICIAL
                 0.316    0.169    0.422      0.316    0.361      0.162    0.615     0.373     FUNCTIONAL
                 0.005    0.000    0.297      0.005    0.009      0.034    0.728     0.106     DESTROYED
                 0.034    0.001    0.581      0.034    0.064      0.136    0.699     0.099     NO DAMAGE
Weighted Avg.    0.502    0.277    0.485      0.502    0.456      0.249    0.687     0.463

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 35280  1810  4178    39    28 |     a = DISABLING
 11204  7114  7549     5    24 |     b = SUPERFICIAL
 15714  4084  9132     0    11 |     c = FUNCTIONAL
  3506   117   311    19     2 |     d = DESTROYED
  1224   897   455     1    90 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.47 seconds

=== Summary ===

Correctly Classified Instances        8994               49.581  %
Incorrectly Classified Instances      9146               50.419  %
Kappa statistic                          0.2206
Mean absolute error                      0.2525
Root mean squared error                  0.355
Relative absolute error                 90.932  %
Root relative squared error             95.2191 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.849    0.518    0.520      0.849    0.645      0.339    0.719     0.589     DISABLING
                 0.269    0.090    0.507      0.269    0.351      0.228    0.707     0.448     SUPERFICIAL
                 0.312    0.173    0.413      0.312    0.356      0.153    0.618     0.368     FUNCTIONAL
                 0.007    0.000    0.455      0.007    0.014      0.052    0.743     0.137     DESTROYED
                 0.044    0.001    0.606      0.044    0.082      0.159    0.717     0.126     NO DAMAGE
Weighted Avg.    0.496    0.278    0.486      0.496    0.449      0.242    0.688     0.461

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 6131  340  743    6    4 |    a = DISABLING
 1998 1247 1390    0    6 |    b = SUPERFICIAL
 2801  699 1591    0    3 |    c = FUNCTIONAL
  643   14   64    5    0 |    d = DESTROYED
  211  161   63    0   20 |    e = NO DAMAGE
```

e. Correlation-based Feature Selection:

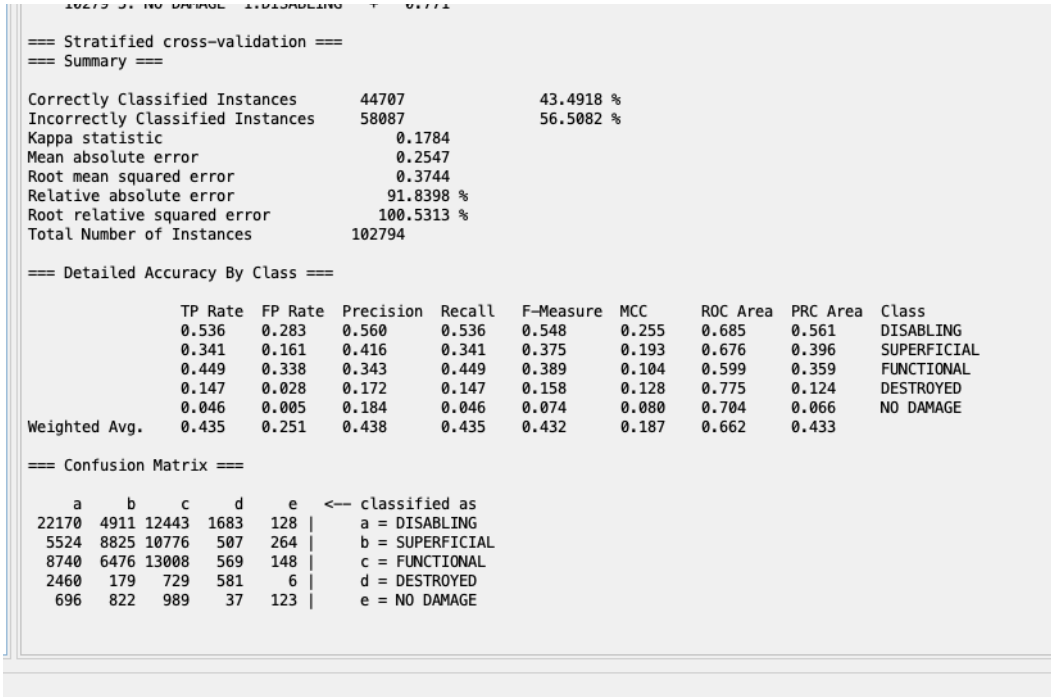
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53139               51.6947 %
Incorrectly Classified Instances     49655               48.3053 %
Kappa statistic                          0.2587
Mean absolute error                      0.2469
Root mean squared error                  0.3501
Relative absolute error                 89.0265 %
Root relative squared error             94.0145 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.830    0.446    0.556      0.830    0.666      0.384    0.748     0.625     DISABLING
                0.310    0.093    0.527      0.310    0.390      0.264    0.737     0.476     SUPERFICIAL
                0.368    0.200    0.419      0.368    0.392      0.175    0.637     0.391     FUNCTIONAL
                0.023    0.001    0.383      0.023    0.043      0.086    0.749     0.129     DESTROYED
                0.031    0.002    0.357      0.031    0.058      0.100    0.745     0.104     NO DAMAGE
Weighted Avg.   0.517    0.259    0.498      0.517    0.479      0.276    0.714     0.489

=== Confusion Matrix ===

    a     b     c     d     e   <-- classified as
 34291  1854  5036   119    35 |     a = DISABLING
  8918  8017  8873     5    83 |     b = SUPERFICIAL
 13913  4321 10657    21    29 |     c = FUNCTIONAL
  3436    75   350    90     4 |     d = DESTROYED
  1129   937   517     0    84 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.47 seconds

=== Summary ===

Correctly Classified Instances        9294               51.2348 %
Incorrectly Classified Instances      8846               48.7652 %
Kappa statistic                          0.2535
Mean absolute error                      0.2472
Root mean squared error                  0.3505
Relative absolute error                 89.0347 %
Root relative squared error             94.0074 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.832    0.448    0.551      0.832    0.663      0.384    0.749     0.623     DISABLING
                0.293    0.092    0.522      0.293    0.375      0.250    0.736     0.470     SUPERFICIAL
                0.371    0.205    0.414      0.371    0.391      0.172    0.640     0.389     FUNCTIONAL
                0.025    0.001    0.450      0.025    0.047      0.098    0.741     0.140     DESTROYED
                0.040    0.001    0.486      0.040    0.073      0.133    0.754     0.111     NO DAMAGE
Weighted Avg.   0.512    0.260    0.500      0.512    0.474      0.272    0.715     0.486

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 6011  310  879   19    5 |    a = DISABLING
 1645 1359 1629    0    8 |    b = SUPERFICIAL
 2432  766 1888    3    5 |    c = FUNCTIONAL
  611   14   82   18    1 |    d = DESTROYED
  205  152   80    0   18 |    e = NO DAMAGE
```

3. **Naïve-Bayes:**

   A classifier which utilizes Bayesian probability to determine the class label of an instance. A unique factor of this classifier is that it assumes each attribute to be independent, similar to the OneR classifier.

   a. **Non-Weka:**

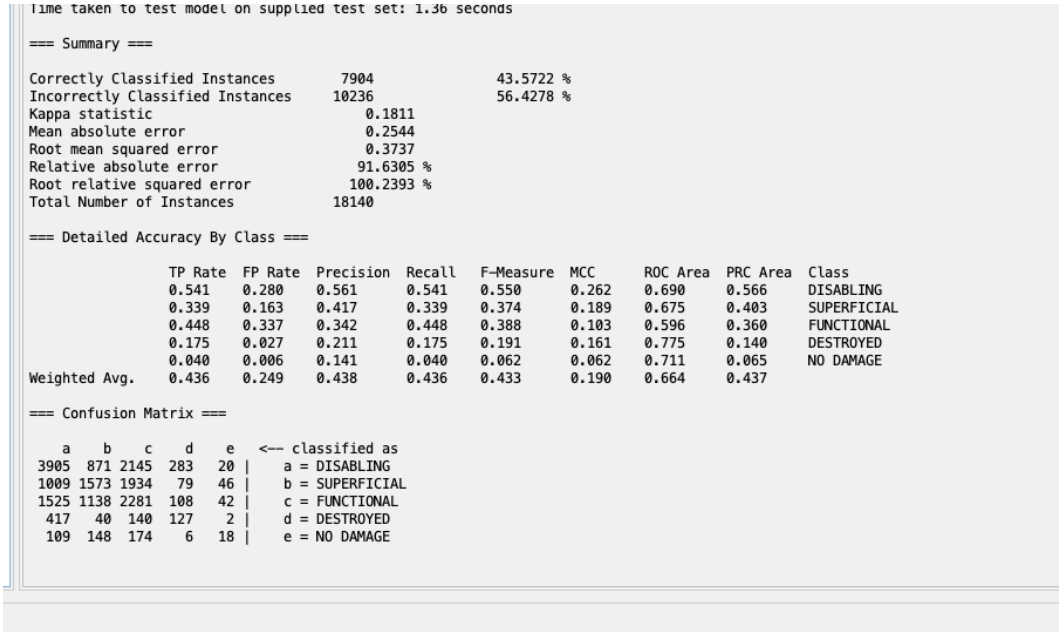


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       44707               43.4918 %
Incorrectly Classified Instances     58087               56.5082 %
Kappa statistic                          0.1784
Mean absolute error                      0.2547
Root mean squared error                  0.3744
Relative absolute error                 91.8398 %
Root relative squared error            100.5313 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.536    0.283    0.560      0.536    0.548      0.255    0.685     0.561     DISABLING
                 0.341    0.161    0.416      0.341    0.375      0.193    0.676     0.396     SUPERFICIAL
                 0.449    0.338    0.343      0.449    0.389      0.104    0.599     0.359     FUNCTIONAL
                 0.147    0.028    0.172      0.147    0.158      0.128    0.775     0.124     DESTROYED
                 0.046    0.005    0.184      0.046    0.074      0.080    0.704     0.066     NO DAMAGE
Weighted Avg.    0.435    0.251    0.438      0.435    0.432      0.187    0.662     0.433

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 22170  4911 12443  1683   128 |     a = DISABLING
  5524  8825 10776   507   264 |     b = SUPERFICIAL
  8740  6476 13008   569   148 |     c = FUNCTIONAL
  2460   179   729   581     6 |     d = DESTROYED
   696   822   989    37   123 |     e = NO DAMAGE
```

```
=== Summary ===

Correctly Classified Instances        7895               43.5226 %
Incorrectly Classified Instances     10245               56.4774 %
Kappa statistic                          0.1804
Mean absolute error                      0.2544
Root mean squared error                  0.3737
Relative absolute error                 91.6351 %
Root relative squared error            100.2431 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.540    0.280    0.561      0.540    0.550      0.262    0.690     0.566     DISABLING
                 0.339    0.163    0.416      0.339    0.374      0.189    0.675     0.403     SUPERFICIAL
                 0.447    0.337    0.341      0.447    0.387      0.103    0.596     0.360     FUNCTIONAL
                 0.172    0.028    0.207      0.172    0.188      0.158    0.774     0.140     DESTROYED
                 0.035    0.006    0.129      0.035    0.055      0.055    0.711     0.065     NO DAMAGE
Weighted Avg.    0.435    0.249    0.437      0.435    0.432      0.189    0.664     0.437

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 3903  873 2143  285   20 |    a = DISABLING
 1009 1574 1935   79   44 |    b = SUPERFICIAL
 1525 1141 2277  109   42 |    c = FUNCTIONAL
  417   41  141  125    2 |    d = DESTROYED
  109  151  173    6   16 |    e = NO DAMAGE
```

b. Information Gain:

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

Correctly Classified Instances       44694               43.4792 %
Incorrectly Classified Instances     58100               56.5208 %
Kappa statistic                          0.1783
Mean absolute error                      0.2547
Root mean squared error                  0.3744
Relative absolute error                 91.8353 %
Root relative squared error            100.5307 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.536    0.283    0.560      0.536    0.548      0.255    0.685     0.561     DISABLING
                 0.340    0.161    0.416      0.340    0.374      0.192    0.676     0.396     SUPERFICIAL
                 0.450    0.338    0.343      0.450    0.389      0.104    0.598     0.359     FUNCTIONAL
                 0.146    0.028    0.171      0.146    0.157      0.127    0.776     0.124     DESTROYED
                 0.046    0.006    0.181      0.046    0.073      0.079    0.704     0.066     NO DAMAGE
Weighted Avg.    0.435    0.251    0.438      0.435    0.432      0.187    0.662     0.433

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 22165  4920 12429  1692   129 |     a = DISABLING
  5509  8815 10797   509   266 |     b = SUPERFICIAL
  8741  6467 13014   569   150 |     c = FUNCTIONAL
  2464   176   731   578     6 |     d = DESTROYED
   696   822   990    37   122 |     e = NO DAMAGE
```



```
Time taken to test model on supplied test set: 1.36 seconds

=== Summary ===

Correctly Classified Instances        7904               43.5722 %
Incorrectly Classified Instances     10236               56.4278 %
Kappa statistic                          0.1811
Mean absolute error                      0.2544
Root mean squared error                  0.3737
Relative absolute error                 91.6305 %
Root relative squared error            100.2393 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.541    0.280    0.561      0.541    0.550      0.262    0.690     0.566     DISABLING
                 0.339    0.163    0.417      0.339    0.374      0.189    0.675     0.403     SUPERFICIAL
                 0.448    0.337    0.342      0.448    0.388      0.103    0.596     0.360     FUNCTIONAL
                 0.175    0.027    0.211      0.175    0.191      0.161    0.775     0.140     DESTROYED
                 0.040    0.006    0.141      0.040    0.062      0.062    0.711     0.065     NO DAMAGE
Weighted Avg.    0.436    0.249    0.438      0.436    0.433      0.190    0.664     0.437

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 3905  871 2145  283   20 |    a = DISABLING
 1009 1573 1934   79   46 |    b = SUPERFICIAL
 1525 1138 2281  108   42 |    c = FUNCTIONAL
  417   40  140  127    2 |    d = DESTROYED
  109  148  174    6   18 |    e = NO DAMAGE
```

c. Gain Ratio:

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

Correctly Classified Instances       45658               44.417  %
Incorrectly Classified Instances     57136               55.583  %
Kappa statistic                          0.189
Mean absolute error                      0.2543
Root mean squared error                  0.3713
Relative absolute error                 91.67   %
Root relative squared error             99.712  %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.592    0.321    0.554      0.592    0.572      0.269    0.689     0.568     DISABLING
                0.495    0.264    0.387      0.495    0.434      0.215    0.681     0.407     SUPERFICIAL
                0.267    0.189    0.357      0.267    0.305      0.087    0.601     0.359     FUNCTIONAL
                0.125    0.025    0.166      0.125    0.142      0.114    0.778     0.124     DESTROYED
                0.056    0.007    0.168      0.056    0.084      0.084    0.705     0.068     NO DAMAGE
Weighted Avg.   0.444    0.250    0.431      0.444    0.433      0.193    0.666     0.439

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 24480  8198  7075  1424   158 |     a = DISABLING
  6327 12808  5897   494   370 |     b = SUPERFICIAL
  9864 10613  7727   530   207 |     c = FUNCTIONAL
  2728   327   400   493     7 |     d = DESTROYED
   800  1130   551    36   150 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.01 seconds

=== Summary ===

Correctly Classified Instances        8081               44.548  %
Incorrectly Classified Instances     10059               55.452  %
Kappa statistic                          0.192
Mean absolute error                      0.2539
Root mean squared error                  0.3708
Relative absolute error                 91.4511 %
Root relative squared error             99.4727 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.600    0.319    0.555      0.600    0.576      0.278    0.693     0.570     DISABLING
                0.497    0.265    0.392      0.497    0.438      0.216    0.682     0.418     SUPERFICIAL
                0.258    0.186    0.352      0.258    0.297      0.080    0.598     0.358     FUNCTIONAL
                0.146    0.025    0.199      0.146    0.168      0.141    0.772     0.137     DESTROYED
                0.051    0.009    0.128      0.051    0.073      0.066    0.709     0.064     NO DAMAGE
Weighted Avg.   0.445    0.248    0.431      0.445    0.434      0.196    0.667     0.442

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4333 1415 1201  242   33 |    a = DISABLING
 1132 2306 1061   79   63 |    b = SUPERFICIAL
 1742 1881 1313  100   58 |    c = FUNCTIONAL
  469   72   77  106    2 |    d = DESTROYED
  138  206   82    6   23 |    e = NO DAMAGE
```

d. Correlation:

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

Correctly Classified Instances       45484               44.2477 %
Incorrectly Classified Instances     57310               55.7523 %
Kappa statistic                          0.1751
Mean absolute error                      0.2561
Root mean squared error                  0.3667
Relative absolute error                 92.3378 %
Root relative squared error             98.4643 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.627    0.369    0.533      0.627    0.576      0.253    0.679     0.548     DISABLING
                 0.509    0.304    0.360      0.509    0.422      0.186    0.667     0.388     SUPERFICIAL
                 0.207    0.136    0.374      0.207    0.266      0.088    0.602     0.363     FUNCTIONAL
                 0.073    0.007    0.293      0.073    0.117      0.130    0.774     0.144     DESTROYED
                 0.043    0.005    0.185      0.043    0.070      0.078    0.687     0.064     NO DAMAGE
Weighted Avg.    0.442    0.264    0.427      0.442    0.419      0.180    0.658     0.428

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 25906 10136  4599   577   117 |     a = DISABLING
  7678 13189  4744    30   255 |     b = SUPERFICIAL
 11029 11705  5985    88   134 |     c = FUNCTIONAL
  3000   413   249   288     5 |     d = DESTROYED
   975  1154   421     1   116 |     e = NO DAMAGE
```

```
=== Summary ===

Correctly Classified Instances        7997               44.0849 %
Incorrectly Classified Instances     10143               55.9151 %
Kappa statistic                          0.1744
Mean absolute error                      0.2564
Root mean squared error                  0.367
Relative absolute error                 92.3346 %
Root relative squared error             98.4541 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.629    0.363    0.534      0.629    0.577      0.260    0.680     0.548     DISABLING
                 0.505    0.307    0.361      0.505    0.421      0.180    0.666     0.399     SUPERFICIAL
                 0.203    0.138    0.364      0.203    0.260      0.080    0.598     0.356     FUNCTIONAL
                 0.085    0.007    0.330      0.085    0.136      0.151    0.764     0.155     DESTROYED
                 0.040    0.006    0.148      0.040    0.062      0.064    0.697     0.063     NO DAMAGE
Weighted Avg.    0.441    0.262    0.424      0.441    0.418      0.180    0.657     0.428

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4542 1769  787  104   22 |    a = DISABLING
 1358 2342  900    2   39 |    b = SUPERFICIAL
 1929 2072 1033   19   41 |    c = FUNCTIONAL
  517   90   55   62    2 |    d = DESTROYED
  163  209   64    1   18 |    e = NO DAMAGE
```

e. Correlation-based Feature Selection:

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

Correctly Classified Instances       47426               46.1369 %
Incorrectly Classified Instances     55368               53.8631 %
Kappa statistic                          0.1846
Mean absolute error                      0.2583
Root mean squared error                  0.3618
Relative absolute error                 93.1221 %
Root relative squared error             97.1514 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.731    0.454    0.520      0.731    0.608      0.274    0.691     0.564     DISABLING
                 0.464    0.235    0.399      0.464    0.429      0.219    0.687     0.390     SUPERFICIAL
                 0.171    0.122    0.356      0.171    0.231      0.065    0.591     0.349     FUNCTIONAL
                 0.059    0.004    0.355      0.059    0.101      0.132    0.758     0.134     DESTROYED
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.047     NO DAMAGE
Weighted Avg.    0.461    0.276    ?          0.461    ?          ?        0.664     0.430

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 30209  6376  4400   350     0 |     a = DISABLING
  9915 12020  3939    22     0 |     b = SUPERFICIAL
 13728 10198  4963    52     0 |     c = FUNCTIONAL
  3176   272   273   234     0 |     d = DESTROYED
  1067  1223   376     1     0 |     e = NO DAMAGE
```

```
=== Summary ===

Correctly Classified Instances        8407               46.3451 %
Incorrectly Classified Instances      9733               53.6549 %
Kappa statistic                          0.1895
Mean absolute error                      0.2587
Root mean squared error                  0.3622
Relative absolute error                 93.1565 %
Root relative squared error             97.1461 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.737    0.449    0.520      0.737    0.610      0.284    0.692     0.558     DISABLING
                 0.464    0.236    0.403      0.464    0.431      0.218    0.685     0.398     SUPERFICIAL
                 0.174    0.120    0.363      0.174    0.236      0.072    0.589     0.350     FUNCTIONAL
                 0.062    0.005    0.363      0.062    0.106      0.137    0.750     0.147     DESTROYED
                 0.000    0.000    ?          0.000    ?          ?        0.685     0.049     NO DAMAGE
Weighted Avg.    0.463    0.273    ?          0.463    ?          ?        0.663     0.430

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5321 1089  748   66    0 |    a = DISABLING
 1800 2153  688    0    0 |    b = SUPERFICIAL
 2346 1848  888   12    0 |    c = FUNCTIONAL
  568   54   59   45    0 |    d = DESTROYED
  190  200   64    1    0 |    e = NO DAMAGE
```

4. Random Forest:

A model that creates multiple decision trees and combines their results to make the most accurate model. It is good at handling all data types and it effectively works with high-dimension data like our dataset.

a. Non-Weka:

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

Correctly Classified Instances       55114               53.616  %
Incorrectly Classified Instances     47680               46.384  %
Kappa statistic                          0.2963
Mean absolute error                      0.2388
Root mean squared error                  0.3434
Relative absolute error                 86.0822 %
Root relative squared error             92.2182 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.384    0.589      0.817    0.685      0.428    0.783     0.666     DISABLING
                 0.483    0.162    0.501      0.483    0.492      0.325    0.772     0.509     SUPERFICIAL
                 0.298    0.156    0.428      0.298    0.352      0.161    0.646     0.397     FUNCTIONAL
                 0.006    0.000    0.532      0.006    0.012      0.055    0.801     0.171     DESTROYED
                 0.066    0.001    0.604      0.066    0.120      0.194    0.814     0.205     NO DAMAGE
Weighted Avg.    0.536    0.239    0.520      0.536    0.502      0.307    0.743     0.520

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 33785  3242  4267    22    19 |     a = DISABLING
  6756 12496  6575     0    69 |     b = SUPERFICIAL
 12660  7625  8631     0    25 |     c = FUNCTIONAL
  3566   124   237    25     3 |     d = DESTROYED
   610  1445   435     0   177 |     e = NO DAMAGE
```

```
Time taken to test model on supplied test set: 3.55 seconds

=== Summary ===

Correctly Classified Instances        9622               53.043  %
Incorrectly Classified Instances      8518               46.957  %
Kappa statistic                          0.289
Mean absolute error                      0.2387
Root mean squared error                  0.3437
Relative absolute error                 85.9827 %
Root relative squared error             92.1862 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.385    0.585      0.821    0.683      0.430    0.786     0.670     DISABLING
                 0.473    0.163    0.499      0.473    0.486      0.316    0.768     0.509     SUPERFICIAL
                 0.287    0.160    0.412      0.287    0.338      0.144    0.645     0.391     FUNCTIONAL
                 0.010    0.000    0.538      0.010    0.019      0.068    0.802     0.204     DESTROYED
                 0.055    0.001    0.595      0.055    0.101      0.176    0.821     0.212     NO DAMAGE
Weighted Avg.    0.530    0.240    0.513      0.530    0.495      0.299    0.743     0.520

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5930  535  750    6    3 |    a = DISABLING
 1213 2196 1224    0    8 |    b = SUPERFICIAL
 2236 1388 1464    0    6 |    c = FUNCTIONAL
  643   24   52    7    0 |    d = DESTROYED
  109  255   66    0   25 |    e = NO DAMAGE
```

b. Information Gain:

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

Correctly Classified Instances       55090               53.5926 %
Incorrectly Classified Instances     47704               46.4074 %
Kappa statistic                          0.2954
Mean absolute error                      0.2395
Root mean squared error                  0.3436
Relative absolute error                 86.3563 %
Root relative squared error             92.2563 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.387    0.587      0.817    0.683      0.425    0.782     0.667     DISABLING
                 0.480    0.159    0.504      0.480    0.492      0.326    0.771     0.508     SUPERFICIAL
                 0.301    0.157    0.430      0.301    0.354      0.164    0.646     0.398     FUNCTIONAL
                 0.004    0.000    0.412      0.004    0.007      0.035    0.803     0.172     DESTROYED
                 0.054    0.001    0.612      0.054    0.100      0.177    0.816     0.202     NO DAMAGE
Weighted Avg.    0.536    0.240    0.516      0.536    0.501      0.305    0.743     0.520

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 33786  3203  4313    19    14 |     a = DISABLING
  6858 12421  6563     0    54 |     b = SUPERFICIAL
 12732  7460  8724     1    24 |     c = FUNCTIONAL
  3584   115   242    14     0 |     d = DESTROYED
   638  1439   445     0   145 |     e = NO DAMAGE
```

```
Time taken to test model on supplied test set: 3.99 seconds

=== Summary ===

Correctly Classified Instances        9680               53.3627 %
Incorrectly Classified Instances      8460               46.6373 %
Kappa statistic                          0.2929
Mean absolute error                      0.2394
Root mean squared error                  0.3436
Relative absolute error                 86.2216 %
Root relative squared error             92.1713 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.824    0.391    0.582      0.824    0.682      0.427    0.784     0.670     DISABLING
                 0.476    0.159    0.507      0.476    0.491      0.324    0.770     0.512     SUPERFICIAL
                 0.293    0.156    0.424      0.293    0.347      0.156    0.643     0.396     FUNCTIONAL
                 0.008    0.000    1.000      0.008    0.016      0.089    0.805     0.208     DESTROYED
                 0.040    0.001    0.529      0.040    0.074      0.140    0.836     0.213     NO DAMAGE
Weighted Avg.    0.534    0.240    0.534      0.534    0.497      0.304    0.743     0.522

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5952  536  734    0    2 |    a = DISABLING
 1255 2210 1167    0    9 |    b = SUPERFICIAL
 2261 1334 1494    0    5 |    c = FUNCTIONAL
  642   16   62    6    0 |    d = DESTROYED
  109  259   69    0   18 |    e = NO DAMAGE
```

c. Gain Ratio:

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

Correctly Classified Instances       54942               53.4486 %
Incorrectly Classified Instances     47852               46.5514 %
Kappa statistic                          0.2979
Mean absolute error                      0.2347
Root mean squared error                  0.3443
Relative absolute error                 84.6226 %
Root relative squared error             92.4658 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.795    0.361    0.597      0.795    0.682      0.426    0.781     0.664     DISABLING
                 0.492    0.168    0.497      0.492    0.494      0.325    0.769     0.503     SUPERFICIAL
                 0.310    0.167    0.421      0.310    0.357      0.159    0.639     0.392     FUNCTIONAL
                 0.020    0.001    0.403      0.020    0.039      0.084    0.801     0.167     DESTROYED
                 0.105    0.003    0.524      0.105    0.175      0.227    0.811     0.201     NO DAMAGE
Weighted Avg.    0.534    0.235    0.513      0.534    0.505      0.307    0.740     0.516

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 32854  3437  4922    86    36 |     a = DISABLING
  6315 12741  6679    11   150 |     b = SUPERFICIAL
 11959  7913  8985    21    63 |     c = FUNCTIONAL
  3437   158   273    81     6 |     d = DESTROYED
   499  1411   474     2   281 |     e = NO DAMAGE
```

```
Time taken to test model on supplied test set: 3.95 seconds

=== Summary ===

Correctly Classified Instances        9575               52.7839 %
Incorrectly Classified Instances      8565               47.2161 %
Kappa statistic                          0.2894
Mean absolute error                      0.2343
Root mean squared error                  0.3443
Relative absolute error                 84.3832 %
Root relative squared error             92.3472 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.796    0.361    0.593      0.796    0.680      0.427    0.784     0.669     DISABLING
                 0.476    0.169    0.492      0.476    0.484      0.310    0.766     0.508     SUPERFICIAL
                 0.305    0.174    0.406      0.305    0.348      0.144    0.640     0.391     FUNCTIONAL
                 0.022    0.001    0.444      0.022    0.042      0.092    0.800     0.195     DESTROYED
                 0.101    0.003    0.469      0.101    0.166      0.209    0.828     0.213     NO DAMAGE
Weighted Avg.    0.528    0.236    0.506      0.528    0.498      0.299    0.741     0.519

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5752  598  855   13    6 |    a = DISABLING
 1114 2207 1288    3   29 |    b = SUPERFICIAL
 2125 1395 1554    4   16 |    c = FUNCTIONAL
  621   31   57   16    1 |    d = DESTROYED
   86  252   71    0   46 |    e = NO DAMAGE
```

d. Correlation:

```
Time taken to build model: 37.46 seconds

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

Correctly Classified Instances       51782               50.3745 %
Incorrectly Classified Instances     51012               49.6255 %
Kappa statistic                          0.2495
Mean absolute error                      0.2426
Root mean squared error                  0.3529
Relative absolute error                 87.4701 %
Root relative squared error             94.7627 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.764    0.395    0.565      0.764    0.650      0.363    0.748     0.627     DISABLING
                 0.433    0.169    0.463      0.433    0.448      0.270    0.729     0.454     SUPERFICIAL
                 0.305    0.180    0.398      0.305    0.345      0.136    0.616     0.373     FUNCTIONAL
                 0.016    0.002    0.285      0.016    0.030      0.060    0.771     0.137     DESTROYED
                 0.033    0.002    0.267      0.033    0.059      0.087    0.758     0.102     NO DAMAGE
Weighted Avg.    0.504    0.252    0.474      0.504    0.474      0.257    0.707     0.479

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 31586  4073  5511   119    46 |     a = DISABLING
  7676 11223  6876    12   109 |     b = SUPERFICIAL
 12488  7522  8821    25    85 |     c = FUNCTIONAL
  3394   177   317    63     4 |     d = DESTROYED
   733  1230   613     2    89 |     e = NO DAMAGE
```

```
=== Summary ===

Correctly Classified Instances        9032               49.7905 %
Incorrectly Classified Instances      9108               50.2095 %
Kappa statistic                          0.2427
Mean absolute error                      0.2428
Root mean squared error                  0.3536
Relative absolute error                 87.436  %
Root relative squared error             94.8434 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.763    0.396    0.561      0.763    0.646      0.361    0.749     0.627     DISABLING
                 0.424    0.171    0.460      0.424    0.442      0.260    0.726     0.453     SUPERFICIAL
                 0.299    0.184    0.388      0.299    0.338      0.126    0.614     0.368     FUNCTIONAL
                 0.022    0.002    0.308      0.022    0.041      0.073    0.765     0.150     DESTROYED
                 0.031    0.003    0.233      0.031    0.054      0.077    0.750     0.100     NO DAMAGE
Weighted Avg.    0.498    0.253    0.468      0.498    0.468      0.250    0.706     0.477

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5510  707  969   27   11 |    a = DISABLING
 1375 1970 1277    2   17 |    b = SUPERFICIAL
 2192 1356 1522    6   18 |    c = FUNCTIONAL
  604   32   74   16    0 |    d = DESTROYED
  149  214   77    1   14 |    e = NO DAMAGE
```

e. Correlation-based Feature Selection:

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

Correctly Classified Instances       50262               48.8958 %
Incorrectly Classified Instances     52532               51.1042 %
Kappa statistic                          0.2395
Mean absolute error                      0.2349
Root mean squared error                  0.367
Relative absolute error                 84.6872 %
Root relative squared error             98.5616 %
Total Number of Instances           102794

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.707    0.344    0.581      0.707    0.638      0.357    0.734     0.603     DISABLING
                 0.397    0.162    0.452      0.397    0.423      0.246    0.715     0.435     SUPERFICIAL
                 0.355    0.230    0.377      0.355    0.366      0.127    0.596     0.342     FUNCTIONAL
                 0.068    0.011    0.192      0.068    0.100      0.094    0.702     0.108     DESTROYED
                 0.082    0.009    0.199      0.082    0.116      0.113    0.721     0.098     NO DAMAGE
Weighted Avg.    0.489    0.244    0.466      0.489    0.473      0.248    0.688     0.455

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 29236  3971  7110   790   228 |     a = DISABLING
  6353 10274  8756   101   412 |     b = SUPERFICIAL
 11057  7168 10267   221   228 |     c = FUNCTIONAL
  2966   206   504   267    12 |     d = DESTROYED
   741  1096   601    11   218 |     e = NO DAMAGE
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 4 seconds

=== Summary ===

Correctly Classified Instances        8813               48.5832 %
Incorrectly Classified Instances      9327               51.4168 %
Kappa statistic                          0.2358
Mean absolute error                      0.2353
Root mean squared error                  0.3666
Relative absolute error                 84.7552 %
Root relative squared error             98.3498 %
Total Number of Instances            18140

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.710    0.350    0.573      0.710    0.634      0.352    0.732     0.599     DISABLING
                 0.386    0.166    0.445      0.386    0.413      0.231    0.710     0.431     SUPERFICIAL
                 0.352    0.224    0.381      0.352    0.366      0.132    0.603     0.348     FUNCTIONAL
                 0.081    0.012    0.223      0.081    0.119      0.114    0.717     0.129     DESTROYED
                 0.081    0.008    0.208      0.081    0.117      0.116    0.739     0.115     NO DAMAGE
Weighted Avg.    0.486    0.245    0.463      0.486    0.469      0.244    0.689     0.455

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 5131  727 1169  151   46 |    a = DISABLING
 1204 1793 1565   18   61 |    b = SUPERFICIAL
 1964 1273 1793   33   31 |    c = FUNCTIONAL
  523   48   93   59    3 |    d = DESTROYED
  133  192   90    3   37 |    e = NO DAMAGE
```

Table 3

K-fold accuracy

	Non-Weka	**InfoGain**	**GainRatio**	**Correlation**	**CFS**
J48	44.9063%	44.4403%	47.2168%	45.7673%	46.3451%
Decision Table	51.8493%	51.8493%	51.8493%	50.2315%	51.6947%
Naive-Bayes	43.4918%	43.4792%	44.417%	44.2477%	46.1369%
Random Forest	53.616%	53.5926%	53.4486%	50.3745%	48.8958%

Note. K-fold accuracies per model and attribute selection algorithm

Table 4

Test accuracy

	Non-Weka	**InfoGain**	**GainRatio**	**Correlation**	**CFS**
J48	44.9835%	44.3495%	47.6185%	46.097%	46.1369%
Decision Table	53.9083%	51.6979%	51.6979%	49.581%	51.2348%
Naive-Bayes	43.5226%	43.5722%	44.548%	44.0849%	46.3451%
Random Forest	53.043%	53.3627%	52.7839%	49.7905%	48.5832%

Note. Test accuracies per model and attribute selection algorithm

Part 6: Conclusion and Tasks Performed

In this project, we developed 20 models to predict the extent of vehicle damage from the Automated Crash Reporting System dataset. The accuracies were suboptimal (all ~50%), with Non-weka Random Forest being the highest with 53.616% accuracy, but that was due to our class attribute's inherent variability. The make and model of a car significantly impacts the damage it receives, and our dataset did not have enough data on every make and model to create

a proper relationship between that and vehicle damage. In conjunction with this, the slightest changes in angles and speed have a drastic effect on the damage a vehicle takes, and on top of this, the vehicle damage is partly subjective; for example, there is a fine line between functioning and superficial damage, and there is no objective scale for measuring vehicle damage. With our skewed dataset, there was bias towards the more present labels like "disabling." This is evident when looking at the True Positive (TP) rates of all our models; the TP rate of disabling (~0.7) is significantly higher than the TP rate of every other label.

In terms of tasks performed, Aniketh Luchmapurkar worked on creating the code (python scripts), running most of the Weka processes, and wrote a bit of the report; Ryan Ghimire also worked on some of the Weka processes and wrote the majority of the report as well as the presentation.

Part 7: Reproducibility of Best Model

1. Data setup
    a. Download the trainMerged_ND.csv under ML Q1 Project > Cleaned and Split Data > nonweka

    OR

    b. Manual data setup
        i. Download data from:

        https://catalog.data.gov/dataset/crash-reporting-drivers-data

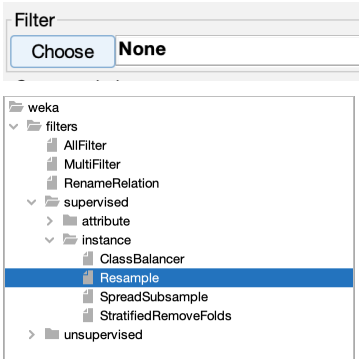
        ii. Run cleaner.py under "ML Q1 Project" Google Drive folder

        iii. Run base_preprocessing.py under "ML Q1 Project" Google Drive folder

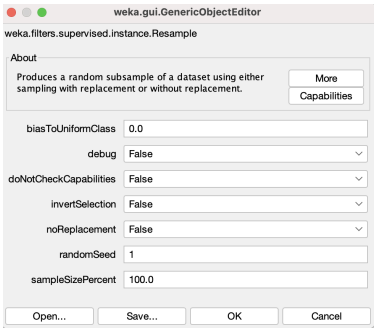
        iv. Run encoder.py under "ML Q1 Project" Google Drive folder

        v. Open csv outputted by previous step in Weka. Perform a train-test-validation split using stratified random sampling in Weka.

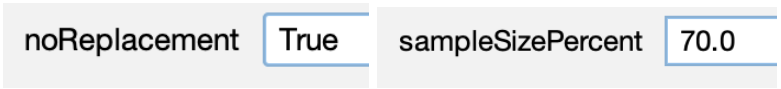
            1. Under Filter, press Choose and go to weka > filters > supervised > instance > Resample



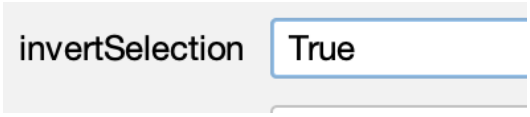
            2. Then, click on the box to the right of the Choose button (the box should say Resample if you chose the correct filter) to open the weka.gui.GenericObjectEditor where you can then edit the parameter of the Resample filter.



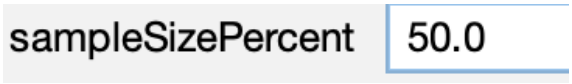
3. Then, set the noReplacement parameter to True, so each instance is only selected once. Set the sampleSizePercent to the proportion of the dataset to be in your training dataset (in our case, we are doing a 70/15/15 split so the training set has 70% of the data).



4. Press Ok and then Apply to run the filter. Congratulations, you have your training set. Now press Save… to save the newly made subset. Now, you have to run another Resample filter to get your testing and validation sets. Reopen your original dataset, and repeat the prior steps until you are on the screen where you edited all the parameters for the Resample filter. Use the exact same parameters as before, but now, this time set invertSelection to True. This will give you the other 30% of the dataset, which you will now split into your test and validation sets.



5. Repeat the same steps as before on this dataset, but now set your sampleSizePercent accordingly to how you want to split your dataset (We are making the test and validation sets both 15% of the dataset, so we will set our sampleSizePercent to 50.0 since we are working with 30% of the data right now). Run the filter and save your test set, then repeat that again, but setting invertSelection to True this time to get your validation set.



vi. Perform attribute selection algorithms as detailed in Part 4. As each arff file is being produced, place said file in a folder with its respective name (correlation, GainRatio, InfoGain, nonweka, CFS)

vii. Run merge.py to combine the training and validation sets

viii. Run fixNoDamage.py to clean up the class labels (this is to get rid of stray quotations added by Weka)

2. Open Weka and load the file trainMerged_ND.csv
    a. If this was downloaded from Google Drive, it should be in home directory
    b. If this was created from scratch, it should be under the “nonweka” folder
3. Navigate to the classify tab choose the RandomForest classifier under weka > classifiers > trees
4. Press start and view the results in the classifier output window

References

- dataMontgomery. (2024). Crash Reporting - Drivers Data. https://catalog.data.gov/dataset/crash-reporting-drivers-data
- Class Slides